UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA DI SCIENZE SOCIALI
DIPARTIMENTO DI ECONOMIA

Corso di Laurea Magistrale in Economics and Data Science

Elaborato scritto per la Prova finale in Statistical Models

Moneyball: how Statistical Models have changed Baseball

Docente di riferimento: Prof. Fabio Rapallo

Candidato: Manuel Delfino

Anno accademico 2023-2024

Contents

Abstract **5**

Introduction **6**

1 The *Moneyball* approach and Data Analytics **8**

1.1 The history behind *Moneyball* approach 8

1.2 Differences between Oakland Athletics and New York Yankees 14

1.3 *Moneyball* and competitive advantage 22

2 Statistical analysis in baseball **28**

2.1 Fundamentals of Statistical Analysis 28

2.1.1 Linear Regression 28

2.1.2 Pythagorean Formula 31

2.1.3 Logistic Regression 32

2.2 First application: baseball dataset 39

2.2.1 Dataset and Data manipulation 39

2.2.2 Hakes and Sauer data simulation 40

2.2.3 Linear and Logistic regression to make Playoffs 43

2.3 Second application: Teams and Salaries dataset 57

2.3.1 Exploratory Data Analysis 57

2.3.2 Pythagorean Expectation on Wins 61

2.3.3 Logistic Regression with different variables than Wins 62

3 An analysis on salaries in MLB **70**

3.1 Salary background and history 70

3.2 Z-score on Salary and Multiple Regression 73

3.2.1 Z-score on Salaries 74

3.2.2 Multiple Regression and Model Selection 77

3.3 Player Replacement 81

Conclusion **83**

Glossary **84**

A Appendix A **93**

A.1 Simple Linear Regression 95

A.2 Expected value of LS estimator 98

List of Figures

1 Wins Comparison between Athletics and Yankees in 1998 and 2002 . . . 19
2 Runs and Runs Allowed Comparison between Athletics and Yankees in 1998 and 2002 . . . 21
3 Relationship between W_{pct}, OBP and SLG . . . 41
4 Teams and Wins relationship to access the Playoffs . . . 44
5 Wins and Runs Differential relationship to access the Playoffs . . . 44
6 Wins Density by Playoffs . . . 45
7 Decision Tree for Baseball Playoffs Prediction . . . 49
8 ROC curve for False and True positive rate in Training set . . . 53
9 ROC curve for Precision and Recall in Training set . . . 54
10 ROC curve for False and True positive rate in Test set . . . 56
11 ROC curve for Precision and Recall in Test set . . . 56
12 Histogram of Win Counts with mean and median line . . . 58
13 Boxplots of Teams and Wins . . . 58
14 Correlation Plot . . . 59
15 Relationship between Runs and other variables . . . 60
16 Relationship between Runs Against and other variables . . . 61
17 Decision Tree for Playoff Prediction on Training set . . . 65
18 Relative Influence of Predictors . . . 67
19 ROC curve on Test Set . . . 69
20 Salary Analysis between Teams in 2002 . . . 72
21 Salary Analysis between Teams in 2002 . . . 73
22 Relationship between Salary and Winning Percentage if the team is a Division Winner . . . 75
23 Average Wins with Z-Score in Salary . . . 76
24 Relationship between Salary and other variables . . . 82

List of Tables

1 Oakland Athletics' Win Ratio and Payroll, 1997-2006 15
2 Pay and Performance, Oakland A's versus New York Yankees, 1997-2006 17
3 Example of a 2x2 Confusion Matrix 36
4 Regression Results 42
5 Linear Regression Results between W and RD 46
6 Linear Regression Results for RS and RA 47
7 Logistic Regression Results for Complete Model 50
8 Results from Anova 51
9 Confusion Matrix with threshold 0.5 in Training set 52
10 Confusion Matrix with threshold set at 0.7 in Training set 52
11 Confusion Matrix with threshold set at 0.2 in Training set 52
12 Confusion Matrix with threshold 0.5 in Test set 54
13 Confusion Matrix with threshold set at 0.7 in Test set 54
14 Confusion Matrix with threshold set at 0.2 in Test set 55
15 Results for Linear Regression between W_{pct} and RD 62
16 Accuracy table for Logistic prediction in Training set 64
17 Area Under the Curve Accuracy for Logistic prediction in Training set . . 64
18 Decision Tree Confusion Matrix on Training Set 66
19 Gradient Boosting Machine Confusion Matrix on Training set 67
20 Logistic Regression Results on Training set 68
21 Confusion Matrix on Test set 69
22 Regression results deriving from AIC 78
23 Regression results deriving from BIC 80
24 Replacement players in 2001 82

Abstract

At the end of 2001 season, Oakland Athletics are defeated by New York Yankees, losing the opportunity to access to the World Series of the Major League Baseball. After the defeat, Billy Beane was denied the budget increase to upgrade the team. Then, he decided to implement Sabermetric, a statistical analysis used to improve the Athletics' results subject to budget constraints, resuming the theory of Billy James. The aim of this thesis is to examine the impact of 'Moneyball' on Athletics and compare its results to those of the New York Yankees. Moreover, different statistical models - linear and logistic regression - are used in order to make better predictions using datasets deriving from the famous Lahman database. Firstly, historical background of Moneyball will be discussed, then a more appropriate analysis will be argued on Athletics results. Finally, also salaries trend and the theory of player replacement will be studied.

Introduction

"If you challenge conventional wisdom, you will find ways to do things much better than they are currently done" - Bill James [19]

These are words coming from Bill James, father of sabermetrics. In the world of baseball, a game deeply rooted in tradition and instinct, a revolution has been unfolding over the past few decades. This transformation has been driven by the statisticians and data analysts working behind the scenes. The dawn of the "*Moneyball*" era, a term popularized by Michael Lewis's 2003 book, has announced a new age where data and statistical models are at the base of baseball strategy and decision-making.

At its core, *Moneyball* sums up the idea that unconventional and often overlooked statistics can be leveraged to build competitive teams, even with limited financial resources. The Oakland Athletics' 2002 season, under the guidance of General Manager Billy Beane, is a significant case study in this approach. Faced with one of the smallest budgets in Major League Baseball (MLB), the Athletics avoided traditional scouting wisdom in favor of sabermetrics, a form of baseball analytics. This strategy led them to remarkable success, achieving playoff and fighting against the giants in baseball field.

"The pleasure of rooting for Goliath is that you can expect to win. The pleasure of rooting for David is that, while you don't know what to expect, you stand at least a chance of being inspired." - Bill James [19]

This thesis explores the impact that statistical models have had on the game of baseball. It delves into the historical context of baseball analytics, the key principles of sabermetrics and how teams have utilized these insights to gain competitive advantages.

Moreover, this thesis investigates the broader implications of the *Moneyball* revolution. Linear and Logistic regression are the main characters in the second Chapter. After a brief description of these two methodologies, analysis on different datasets will be done looking also at a specific literature review.

Baseball is a model of the broader societal shift towards analytics in a world where data is becoming more important in decision-making across industries. Baseball's transition from gut instinct to data-driven precision not only redefines field strategies but also

provides valuable lessons for other domains where information and insight can challenge established norms.

Finally, in the third Chapter, the impact of statistical learning is applied also in salary management. It will be shown how salaries are evolving over time looking at their historical backgrounds. Then, player replacement, coming from *Moneyball* idea, will be studied in order to demonstrate how Oakland Athletics could have managed the loss of three important players.

I The *Moneyball* approach and Data Analytics

I.1 The history behind *Moneyball* approach

Michael Lewis' *New York Times* best seller, *Moneyball* (Lewis, 2004)[21], is a book about baseball. In fact, Lewis shows how Billy Beane's reliance on theoretically relevant statistics and on a scientific approach to baseball allowed him to achieve winning season despite being burdened by severe budget constraints [8]. Billy Beane was the General Manager (GM) of the Oakland Athletics, a MLB (Major League Baseball) team. Before starting his career as a General Manager, Billy Beane was a player who has milited in MLB in the 80's. He was drafted by the New York Mets at the first round and his first debut in the major team has taken place in 1985. Then, he played in other different teams as Minnesota Twins, Detroit Tigers and Oakland Athletics.

However, Beane was remembered not so much for his sporting career, but for his career General Manager. Lewis' book has been described also in the famous film *Moneyball* directed by Bennett Miller, with Brad Pitt as Billy Beane [66] which gave further prestige to the phenomenon of statical approach in sport world.

What is very important to remember about Beane's history is the way by which he revolutionized the scouting system in MLB using statistical rules and how he challenged the vision on data analysis in sport contexts. Beane exploited the inefficiency by implementing a player performance measurament and feedback system that allowed him to field a highly competitive team while having one of the lowest payrolls in MLB [21]. He decided to use rigorous statistical analysis - sabermetrics - to determine the best set of players with the best value for his team.

Over the years, baseball was depicted as being guided by wise traditions and by clinical expertise derived from years of experience in dealing with the unique situations and irreducible complexities inherent in the game of baseball [8]. With Beane's approach a new paradigm has emerged. However, sabermetric was found by Bill James, an iconoclastic figure from Kansas, who produced a lot of data and essays those questioned widely accepted baseball knowledge and practices [15]. Bill James main ideas are written in his famous abstract entitled *The new Bill James historical baseball abstract* [19]. James was not the first to use evidence to provide new visions of the baseball enterprise. However,

over the years, the use of evidence to inform baseball decisions proved sporadic and overwhelmed by the traditional insider paradigm [8]. The traditional baseball paradigm was based on two main beliefs [8].

1. A player's talent is most accurately appraised by having "baseball man" scouts look at the individual in person.

2. The statistics that have long been collected, such as batting average and runs batted were considered more than adequate to assess a player's performance and value to a team.

James started his work in the late 1970s and his scope was to alterate the intellectual landscape of professional baseball. Even if at the beginning he was not well known, in 2002 he would be employed as an advisor to the Boston Red Sox, one of the best team in MLB's history[68]. At the Red Sox he won the first World Series (annual championship series in MLB [67]) in 2004 and the second one in 2007. These two successes gave to his theory further importance. However, three other main considerations contributed to increasing his status and influence.

First, he gave the new paradigm a name: "sabermetrics". In fact, the term sabermetrics derives from the acronym for the Society for American Baseball Researc ("SABR") and the Latin Suffix for measurement ("metrix")[55]. The creation of a new word able to classify his approach has been necessary in order to let other baseball researchers to join an identifiable movement. James has defined sabermetrics as "objective knowledge" about baseball. Sabermetric reasoning often involves trying to find out how many wins a player is worth to a team above an average replacement player based on their fielding, hitting and pitching [7].

Second, James has used the Hagan's theory of the "sociology of the interesting" [8]. This theory is based on taking a common belief and subsequently proving that it is incorrect [17].

Third, he was able to provide rigorous data to stir up his claims. As Gray analysed in his book [15], the appeal of James's analyses wanted to reveal that the traditional paradigm has misunderstood the nature of baseball and thus led to irrational practices [8].

James was interested in the basic question of what constitutes a good player. He began to look at previously undervalued categories in baseball and he wanted to know the additive value of a player who could steal bases for a team [7]. In 1971, the Society for American Baseball Research (SABR) was founded by Leonard Robert Davids with four guiding principles that encouraged the study of baseball, education and historical preservation. SABR facilitated the creation of new ways of contextualizing baseball.

Taking up James' ideas the *Moneyball* theory places emphasis on the body of the athlete or the physical tools that the athlete possess. This theory illustrates the simplicity of baseball by asking two questions [63]:

Does this player get on base? Can he hit?

James believed that hitter's job was not to compile a high batting average, neither was to maintain a high on-base percentage, nor to create a high slugging percentage. The job of a hitter was to create runs. So, he developed a formula (see Equation 1 below) that allows one to establish created runs [19]:

$$\frac{(\text{Hits + Walks}) \times \text{Total Bases}}{\text{At - bats} + \text{Walks}} \tag{1}$$

From this philosophy, Beane has developed his theory through a new formula that took into account more aspects of meaningful baseball statistics. Beane considered that the only way to score runs is to get on base and since walks [1] are a vital part of the created runs formula, on-base percentage should be closely monitored [21]. However, additional steps can be taken to improve the accuracy and other meaningful baseball statistics could be inserted in a new simple formula, as Equation 2 shows:

$$\frac{A \times B}{C} \tag{2}$$

where:

1. The A variable adjusts the "on-base" aspect of baseball:

   A = hits + walks + hit batsmen - caught stealing - ground into double play

$$A = H + BB + HBP - CS - GIDP \tag{3}$$

[1] A base on balls (BB), also known as a walk, occurs in baseball when a batter receives four pitches during a plate appearance [64]

From Equation 3, it is useful to define the variables that have been used. A hit (H) occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice [29]. A hit-by-pitch (HBP) takes place when a batter is struck by a pitched ball without swinging at it [30]. A caught stealing arises when a runner attempts to steal but is tagged out before reaching second base, third base or home plate [25].Then, a GIDP occurs when a player hits a ground ball that results in multiple outs on the bases [28].

2. The variable B is taking into account the advancement of the player:
   B = total bases plus 0.26 times hit batsmen and non-intentional walks, plus 0.52 times stole bases, sacrifice hits, and flies.

$$B = TB + 0.26(TBB - IBB + HBP) + 0.52(SB + SH + SF) \tag{4}$$

   From Equation 4, total bases refer to the number of bases gained by a batter through his hits [40]. Instead, TBB stands for Total Base on Ball and it represents the total of Walks [64]. An intentional walk (IBB) occurs when the defending team elects to walk a batter on purpose, putting him on first base instead of letting him try to hit [33]. A stolen base (SB) takes place when a baserunner advances by taking a base to which he isn't entitled. This generally occurs when a pitcher is throwing a pitch [39]. A sacrifice bunt (SH) arises when a player is successful in his attempt to advance a runner (or multiple runners) at least one base with a bunt [36]. Then, a sacrifice fly (SF) occurs when a batter hits a fly-ball [2] out to the outfield or foul territory that allows a runner to score [37].

3. The C variable accounts for opportunity:
   C = at-bats + total walks + sacrifice hits and flies + hit batsmen

$$C = AB + TBB + SF + HBP \tag{5}$$

   From Equation 5, an official at-bat (AB) comes when a batter reaches base via a fielder's choice, hit or an error (not including catcher's interference) or when a batter is put out on a non-sacrifice [24].

[2]For statistical purposes, MLB uses the term "fly ball" for such balls that go into the outfield, and a separate term (pop-up, below) for such balls that stay in the infield [65].

In this way, James believed that looking at the number of runs created would be a great tool to evaluate hitters from the moment that hitter's job is to create runs [19].

Since 2002, sabermetrics has changed the way baseball teams are constructed, pushing away old techniques of assessing talents through eye tests and intuition. Sabermetrics and the *Moneyball* experiment started the analytics movement by promoting two important but undervalued statistics, on-base percentage (OBP) and slugging percentage (SLG).

In baseball jargon, "on base" means occupy one of the bases, which are commonly three and they are denoted respectively by base one, two and three [48]. On one side, on-base percentage refers to how frequently a batter reaches base per plate appearance. Times on base include hits, walks and hit-by-pitches, but do not include errors, times reached on a fielder's choice or a dropped third strike [22]. The full-detailed formula is described in Equation 6:

$$\begin{aligned} \text{OBP} &= \frac{(\text{Hits + Base on Balls + Hit by Pitch})}{(\text{At-bats + Base on Balls + Hit by Pitch + Sacrifice Flies})} = \\ &= \frac{(\text{H + BB + HBP})}{(\text{AB + BB + HBP + SF})} \end{aligned} \tag{6}$$

On the other side, slugging percentage or average, called SLG, represents the total number of bases a player records per at-bat. Unlike on-base percentage, slugging percentage deals only with hits and does not include walks and hit-by-pitches in its equation [23]. Equation 7 describes how SLG is calculated:

$$\text{SLG} = \frac{\text{Total Bases}}{\text{At-bat}} = \frac{\text{TB}}{\text{AB}} \tag{7}$$

Slugging percentage differs from batting average in that all hits are not valued equally. While batting average is calculated by dividing the total number of hits by the total number of at-bats, the formula for slugging percentage is described by Equation 8:

$$\text{SLG} = \frac{(1\text{B} + (2\times 2\text{B}) + (3\times 3\text{B}) + (4\times \text{HR}))}{\text{AB}} \tag{8}$$

where:

- 1B represents the single. A single occurs when a batter hits the ball and reaches first base without the help of an intervening error or attempt to put out another baserunner. Singles are the most common type of hit in baseball [38].

- 2B is related to doubles. A batter is credited with a double when he hits the ball into play and reaches second base without the help of an intervening error or attempt to put out another baserunner [26].

- 3B stands for triple. a triple occurs when a batter hits the ball into play and reaches third base without the help of an intervening error or attempt to put out another baserunner [41].

- HR is the acronym for home run. A home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error. In almost every instance of a home run, a batter hits the ball in the air over the outfield fence in fair territory. In that situation, the batter is awarded all four bases, and any runners on base score as well [31].

Then, the two stats were combined to form a new statistic called on-base plus slugging (OPS). These statistics were considered important because they correlated well with a team's ability to score runs, which is a key determinant of a team's success. The Athletics also looked at a player's salary, as they sought to find undervalued players who were being paid less than their performance would warrant [47]. In his approach Beane did not consider power, even if he believed that power could be developed but he thought that patience at the plate and the ability to get on base could not. James' idea was focused on the philosophy of hitters and it was different from the draft process of Beane. Therefore, in Beane point of view, managers must decide the best order in which the teams has the best chance of winning. To win a game one must score more runs than the opposing team [63].

Another game-changing statistic that has been introduced into sabermetrics is Walks and Hits Per Innings Pitched (WHIP). WHIP is one of the most commonly used statistics for evaluating a pitcher's performance. The statistic shows how well a pitcher has kept runners off the basepaths, one of his main goals. The formula is simple enough: it is the sum of a pitcher's walks and hits, divided by his total innings pitched [3] [42]. A inning is a very important component in a baseball game. Indeed, it is defined as the division of a

[3]Innings pitched measures the number of innings a pitcher remains in a game. Because there are three outs in an inning, each out recorded represents one-third of an inning pitched [32].

baseball game consisting of a turn at bat for each team [49]. A Major League Baseball game consists of nine scheduled innings, in which each team has an opportunity to score runs on offense in its half of each inning [18]. WHIP contrasts Earned Run Average, that is by definition the number of earned runs a pitcher allows per nine innings – with earned runs being any runs that scored without the aid of an error or a passed ball. ERA is the most commonly accepted statistical tool for evaluating pitchers [27]. Statistics like on-base percentage, on-base plus slugging, wins above replacement, and walks plus hits-per-innings pitched have created a new foundation and perspective on evaluating baseball players that have been proven effective.

Moneyball introduced to an era of advanced analytics in baseball and beyond, popularizing stats like OBP, OPS, and WHIP while paving the way for new metrics. An example can be found in wins above replacement (WAR), that becomes a significant statistics that major league teams value above all others. WAR measures a player's value in all facets of the game by deciphering how many more wins he's worth than a replacement-level player at his same position [43].

I.2 Differences between Oakland Athletics and New York Yankees

Billy Beane's work has started in 1998, when the Athletics had one of the lowest budget in the league. As Table 1 shows, the turnaround has been truly remarkable. In the analysis, also 1997 has been taken into account in order to better demonstrate how *Moneyball* works.

Beane's first season as general manager wasn't a winning one. The A's finished with a losing record (0.457 win ratio), ranking 22^{nd} out of 30 teams. However, there was a silver lining: Oakland had one of the lowest payrolls in baseball (third lowest in 1998).

Following that initial season, the Athletics under Beane never had a losing record again. Remarkably, they achieved this success while consistently being one of the lowest spenders in the league. Only in 2004 did their payroll climb above the bottom third.

The book *"Moneyball"* [21] dives deep into the 2001 and 2002 seasons. In both years, Oakland boasted the second-highest win ratio in the regular season, exceeding 100 wins (out of a 162-game schedule). However, they defied expectations by having the second-lowest payroll in 2001 and the third lowest in 2002. This impressive feat highlights the

effectiveness of Beane's strategy for building a winning team on a limited budget.

Table 1: Oakland Athletics' Win Ratio and Payroll, 1997-2006

Year	Win Ratio	Win Ratio Ranking	Payroll	Payroll Ranking
1997	0.401	30^{th}	$21.9m	26^{th}
1998	0.457	22^{nd}	$20.1m	28^{th}
1999	0.537	10^{th}	$24.2m	26^{th}
2000	0.565	6^{th}	$32.1m	25^{th}
2001	0.630	2^{nd}	$33.8m	29^{th}
2002	0.636	2^{nd}	$40.0m	28^{th}
2003	0.593	4^{th}	$50.3m	23^{rd}
2004	0.562	9^{th}	$59.4m	16^{th}
2005	0.543	9^{th}	$55.4m	22^{nd}
2006	0.574	5^{th}	$62.2m	21^{st}

Sources: Baseball Refernece; Bill Gerrard: "Is the Moneyball Approach Transferable to Complex Invasion Team Sports?" [14]; MLB Standings

To truly understand the scale of Oakland's achievement, a comparison with the Major League Baseball's giant, the New York Yankees, needs to be done. While the Athletics were defying budget constraints, the Yankees were known for their high-spending approach [14]. In this sense, Table 2 highlights the differences in payroll and performance between Athletics and Yankees. In order to provide a better comparison between the two teams, it is fair to remember that the Yankees, in their history, have won 27 World Series, 40 League Titles and 19 Division Titles.

Moreover, the Yankees can boast of great support from the public, since the Yankee Stadium can accommodate more than 47 thousand people [34]. Even if the number of possible spectators is similar for the two teams [4], the Oakland Athletics has got in their palmarès only 9 World Series, 15 League Title and 17 Division Titles. The relationship between Athletics and Yankees could be seen as a similarity to the history of David and Goliath, surely in terms of payroll [14].

The Oakland Athletics consistently finished behind the New York Yankees in terms of wins over an eight-year period, but the gap between the two teams was relatively small. From 1999 to 2006, the Yankees averaged just under six more wins per season than Oakland, translating to a mere 3.9% win advantage [14]. Interestingly, despite this, the Yankees spent significantly more money on players during this time. Their payroll was 3.22 times higher than Oakland's, a difference of 216.7% [14]. However, this substantial spending edge only yielded a minor win advantage for the Yankees. In fact, the Athletics even managed to outperform the Yankees in win percentage during two of those seasons (2000 and 2002).

It should be notice how both payrolls have increased their value. From 1997 to 2006, Athletics payroll increased by more than 40 million dollars goring from 21.9 millions dollars to 62.2 millions dollars, almost tripling the starting figure, 21.9 million. The rise in Yankees salaries has been even more noticeable. Their payroll amount is almost fourfold, from 59.1 million dollars to 194.7 million dollars, peaking at 208.3 million in 2006.

[4]The Athletics have 46.847 seats at the Oakland Coliseum, their stadium [35].

Table 2: Pay and Performance, Oakland A's versus New York Yankees, 1997-2006

	Oakland Athletics		New York Yankees	
Year	**Regular Season Win Ratio**	**Payroll**	**Regular Season Win Ratio**	**Payroll**
1997	0.401	$21.9m	0.593	$59.1m
1998	0.457	$20.1m	0.704	$63.5m
1999	0.537	$24.2m	0.605	$85.0m
2000	0.565	$32.1m	0.540	$92.5m
2001	0.630	$33.8m	0.594	$109.8m
2002	0.636	$40.0m	0.640	$125.9m
2003	0.593	$50.3m	0.623	$149.7m
2004	0.562	$59.4m	0.623	$182.8m
2005	0.543	$55.4m	0.586	$208.3m
2006	0.574	$62.2m	0.599	$194.7m

Sources: Baseball Refernece; Bill Gerrard: "Is the Moneyball Approach Transferable to Complex Invasion Team Sports?" [14]; MLB Standings

Figure 1 shows in a more intuitive way how the win rate of the Athletics increased thanks to the *Moneyball* approach. The two graphs represent the wins comparison between Athletics and Yankees. In Figure 1a it is taken into account the data for 1998, whereas, in Figure 1b the data for 2002.

This kind of graph is useful to get the differences between the teams in two periods of time. In this sense, 1998 can be assumed as a *pre-Moneyball* period, instead 2002 as a *post-Moneyball* period.

In Figures 1a and 1b, the x-axis is labeled "Game" and it goes from 0 to 162, likely representing the total number of games played in the 1998 season. The y-axis labeled "Wins" goes from 0 to 100.

The data series for the Oakland Athletics is plotted with a blue line. There is a steady increase in wins for the Athletics until approximately game 40. The line then begins to flatten out, with the number of wins gradually increasing until the end of the season. They have finished the season with 88 wins.

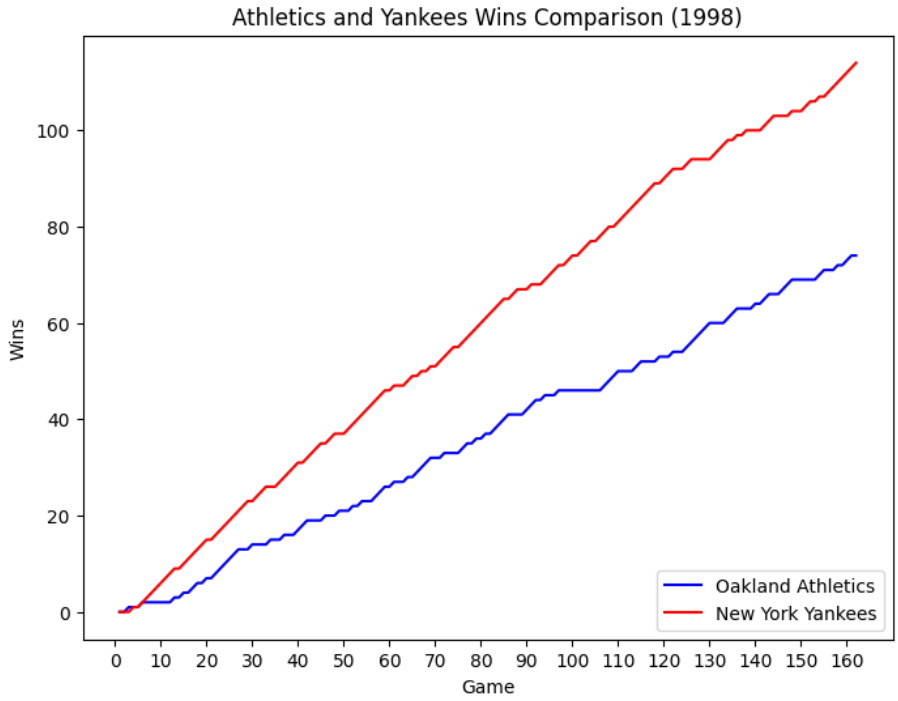
Instead, the data series for the New York Yankees is plotted with a red line. The Yankees line shows a gradual increase in wins throughout the season. It continues to increase at a steeper rate than the Athletics line, indicating they won more games later in the season. In fact, during 1998 season they won 114 games, almost 60 more than the Athletics. Overall, the graph suggests that before *Moneyball* there was a huge difference between the two teams.

Indeed, looking at Figure 1b it can be noted that both teams have approximately followed the same trend. The Yankees were more constant, whereas the Athletics had a period, between Game 30 and 40, in which the wins struggled to get. However, even if during in the middle of the season the Athletics had always have a lower number of wins, in the final part of the season they grind wins, catching up the Yankees. In fact, both teams ended the season with a total of 103 wins.

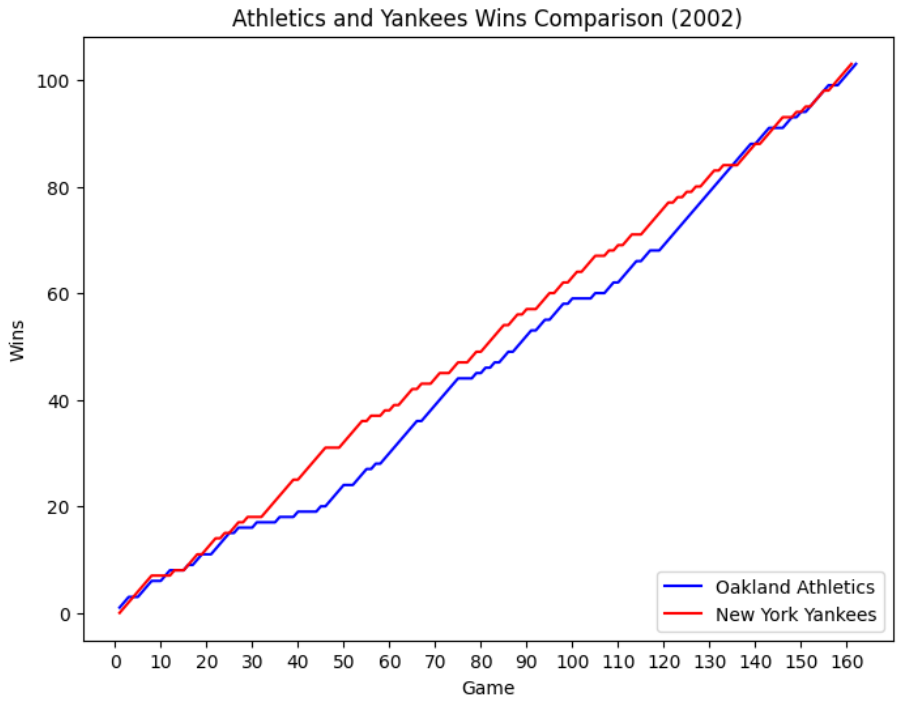
So, this graph shows how Beane managed to make the Athletics competitive. It can be stated, looking at the salaries in Table 2, that the Athletics, in relative terms, have outperformed the Yankees during the 2002 season.

Figure 1: Wins Comparison between Athletics and Yankees in 1998 and 2002

(a) *Athletics and Yankees Wins Comparison (1998)*



(b) *Athletics and Yankees Wins Comparison (2002)*



Source: own computation using Python algorithm based on Baseball Refernece data

Since Bill James has considered the creation of runs very important for the application of the *Moneyball* approach, comparing the runs created and allowed by both the Athletics and the Yankees during the same period as before, namely 1998 and 2002, may be advantageous.

Looking at the graphs in Figure 2a and 2b related to 1998, it can be seen that during the year the Yankees scored more runs than the Athletics in most games throughout the season, whereas the Yankees allowed less runs during all the season. In fact, data in hand, the Athletics have created 804 runs, while the Yankees have created 965, providing a difference of -161 between the two franchises. Furthermore, the Athletics allowed to score to opponents 866 runs, much more compared to 656 allowed by the Yankees, providing a difference of -210. So, during 1998, the Yankees have created more runs and have allowed less runs to the opponents [3, 5].

Instead, during 2002, in Figure 2c for the first 40 games, the two teams have created more or less the same amount of runs (around 200). Then, the Yankees prevailed over the Athletics in this statistic. However, the difference between the total runs is lower than the one of 1998. The Athletics created 800 runs while the Yankees created 897. This provides a difference of -97 runs [4, 6].

Nevertheless, Figure 2d shows that during 2002 the Athletics have allowed less runs than the Yankees. They allowed 654 runs, 43 less than the Yankees allowed, which were 697.

So, citing Bill James:

"The numbers don't lie, they tell the story of the game. We just need to know how to interpret them [19]"

This phrase refers to the importance of statistics in understanding baseball.

- *The number don't lie:* statistics accurately reflect what happened on the field. Batting averages, earned run averages, strikeouts, etc. are all objective measurements of a player or team's performance.

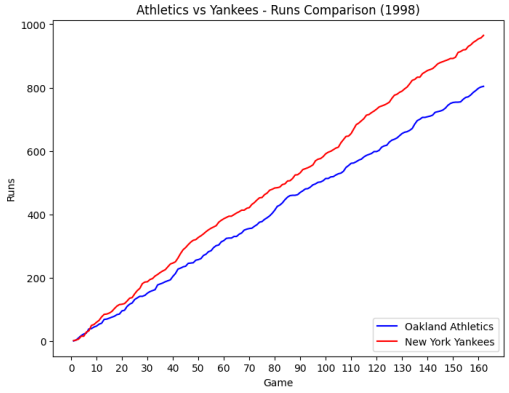
- *They tell the story of the game:* statistics go beyond the final score. They reveal underlying trends, strengths, and weaknesses. For example, a high strikeout pitcher might be dominant, but also struggle with control.

- *We just need to know how to interpret them:* not all statistics are created equal, and some require context to be meaningful. A high batting average for a hitter who only hits singles might not be as valuable as a lower average for someone who hits home runs.

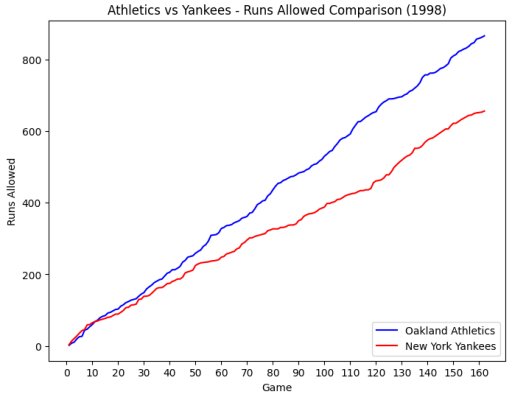
Doing an interpretation of Figure 1 and 2 might demonstrate how effectively *Moneyball* has worked in very short time.

Figure 2: Runs and Runs Allowed Comparison between Athletics and Yankees in 1998 and 2002

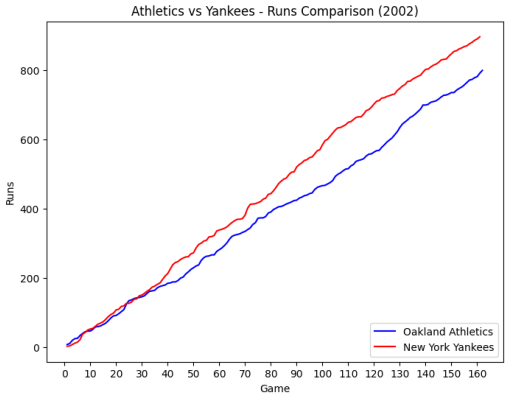
(a) *Runs comparison in 1998*



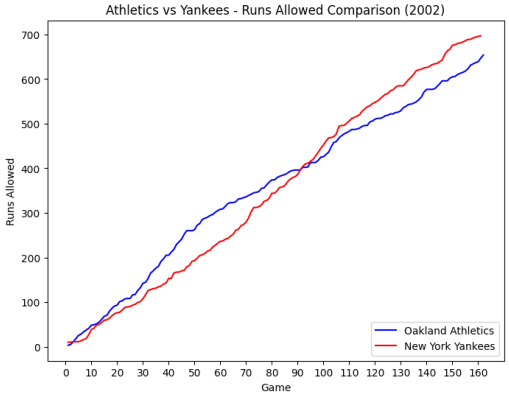
(b) *Runs Allowed comparison in 1998*



(c) *Runs comparison in 2002*



(d) *Runs Allowed comparison in 2002*



Source: own computation using Python algorithm based on Baseball Refernece data

I.3 *Moneyball* and competitive advantage

The 2002 Athletics had one of the lowest budgets in the league, along with some of the lowest expectations for success. When Billy Beane was hired, he knew he had to switch things up. He was able to take the risks that nobody else wanted to, which meant going against the wishes of the coaching staff, recruiters, players and media. In order to avoid stagnating, sometimes thinking differently is required.

From this point of view, when the New York Yankees incorrectly judge a player's talent, it is a disappointment but not a tragedy. With a salary budget in excess of $200 million, it is very simple for the owner to purchase the next star free agent. However, Billy Beane's Oakland Athletics are not so fortunate due to his limited resources [8].

Sabermetrics were successfully adopted and implemented by the Athletics. Billy Beane was the key to the successful implementation of sabermetrics. In fact, Beane was (almost) all-powerful, reporting only to team ownership. Ownership was supportive, as the small-market Athletics had to be innovative to compete with large-market teams [69]. As described by Lewis, Beane was the prototypical innovation champion. He had the necessary energy, commitment, and organizational power to propel the adoption and implementation of sabermetrics [21]. The threat of sabermetrics to extant skills and to livelihoods resulted in considerable resistance to the innovation. As suggested in the innovation literature, in such situations, innovation implementation is dependent upon the innovation champion having considerable organizational power [70].

A seminal lesson in the *Moneyball* story involves the length of time it took for sabermetrics to be adopted, its slow diffusion, and the considerable competitive advantage it has provided the Athletics [69].

The adoption of sabermetrics has contributed to remarkable results for the Athletics, which consisted in a significantly improved winning per-cent with significantly decreased salaries. Billy Beane took a chance on players who nobody wanted, changed the player evaluation standards, and pushed the envelope to challenge conventional wisdom [62]. *Moneyball* results have been sustained while consistently allowing the team's most acclaimed players to move on; from 1998, Beane has traded or simply not re-signed eight All-Stars [69]. This line of thought has allowed the Athletics to maintain lower salaries

with respect to the average of the MLB. However, this did not affect the winning per-cent, indeed the Athletics have been among the top third of MLB teams in this statistic, whereas they have been among the bottom third in salaries.

From the results described by *Moneyball* approach, it should be confirmed that sabermetrics have worked. The Athletics' approach to identifying hitters with superior skills at reaching base without paying a market premium for them has resulted in winning games at a discount relative to the competition [70].

A more economic analysis can be implemented. It concerns the concept of competitive advantage. Competitive advantage can be achieved through resource-picking and capability-building, as identified by strategy literature [69]. From the study of Richard Makadok, a strategic management professor, managers gather information and analysis to outsmart the resource market in picking resources, similar to the way that a mutual fund manager tries to outsmart the stock market in picking stocks [45]. So, the resource-picking mechanism creates economic rents when the firm purchases resources for less than their marginal productivity, when used in combination with its stock of other resources [45].

If the Ricardian perspective is taken, then resource-picking is the main mechanism for the creation of economic rent. This perspective is based on the "resource-based view" or RBV. One important implication of Ricardian thesis is that this mechanism for creating economic rent actually takes place before the acquisition of resources. This will imply that firms with superior resource-picking skill will apply that skill in order to distinguish which resources are winners and which ones are losers, so that they will bid for the former while avoiding the latter [45]. Competitive advantage through resource-picking is only possible when the firm has superior information [69].

In order to assess how sabermetrics might provide a competitive advantage, the RBV of the firm can be used. This perspective has emerged as a major strategic paradigm from the study of Jay Barney [2]. In his view, "a firm is said to have a competitive advantage when it is implementing a value creating strategy not simultaneously being implemented by any current or potential competitors" [2]. According to the RBV, firms are given heterogeneous bundles of resources, and competitive advantage only occurs if and only if a resource is *valuable* and *rare* [69]. From the baseball point of view, it is

reasonable to conclude that, according to Beane idea, by identifying players with superior skills that were undervalued, the Athletics resource-picking mechanism meets the value criterion. Furthermore, sabermetrics fulfilled the rare criterion.

In addition, Barney gives a definition for the *sustained* competitive advantage. Indeed, by definition a sustained competitive advantage is created when it is implementing, by a firm, a value creating strategy not simultaneously being implemented by any current or potential competitors and when these other firms are unable to duplicate the benefits of this strategy [2]. Furthermore, other criteria must be met by a resource in addition to the RBV value and criteria. Stand at what Barney has argued in his work, a resource must be imperfectly imitable (or substitutable) in the sense that competing organizations face cost and/or quality disadvantages in developing an appropriate substitute for it [70]. Thus, the firm must be organized such that it can realize a competitive advantage based on resources which will add value, are rare, are imperfectly imitable and there cannot be strategically equivalent substitutes for this resource that are valuable but neither rare or imperfectly imitable [2, 51, 69]. If a resource can be copied or trumped by a strategically equivalent substitute, then the market imperfection is unlikely to be sustained.

In order to provide a more relevant analysis for RBV, the article entitled *"What is strategy"* written by Michael E. Porter should be useful. In fact, Porter argues that operational effectiveness and strategy are both essential to superior performance, which is the primary goal of any enterprise. In his view, a company can outperform rivals only if it can establish a difference that it can preserve. In one hand, operational effectiveness is intended to carry out similar activities better than rivals. It is a term used to describe practices that enable a company to utilize its input more effectively, such as decreasing product defects or creating better products more quickly. In the other hand, strategic positioning means performing different activities from rivals or performing similar activities in different ways [52]. Porter argues that competitive advantage occurs from an organization's choice of unique activities more efficiently than competitors [70]. Porter asserts that in today's environment, a heightened diffusion of best practices often results in a temporary competitive advantage. Furthermore, he states that firms' advantages are based on unique, tailored sets of congruent activities [52]. Moreover, strategic fit among activities is crucial in order to reduce costs and/or increase differentiation and to create sustainabil-

ity of competitive advantage. In fact, it is more difficult to match an array of interlocked processes and activities than it is to imitate one or two processes or activities [69].

The Oakland Athletics created a competitive advantage through the implementation of sabermetrics. However, the Athletics' competitive advantage maybe sustained by performing sabermetrics more efficiently than their competitors. The key factors to consider in maintaining the Athletics' competitive advantage through sabermetrics lies in the sustainability of their social complex resources. These resources include the collaborative front-office culture boosted by Beane's leadership, as well as the control and reward structures he implemented. If these elements are difficult to replicate, then following the Athletics' sabermetric approach may put imitators at a disadvantage [69].

In this sense, it is useful to come back to Makadok thesis. It should be noted the role of capability-building. In accordance with Makadok, capability is defined "as a special type of resource - specifically, an organizationally embedded nontransferable firm-specific resource whose purpose is to improve the productivity of the *other* resources possessed by the firm" [45]. Standing at this definition, Makadok has argued that capabilities cannot easily be bought. Instead, they must be built. So, capability-building implies developing and building internal capabilities. Given the current context, achieving this requires cultivating and integrating functionalities centered around utilizing sabermetric principles. If the Athletics competitive advantage is to be sustained, it should be associated to the capabilities that Beane has built in the "working culture" which has improved the productivity of other resources possessed by the firm [69].

However, during and following the 2003 season, some other MLB teams started to move in the way of innovation. Indeed, two senior managers from the Athletics front-office were hired as general managers by the Toronto Blue Jays and the Los Angeles Dodgers. Moreover, the Boston Red Sox hired the father of sabermetrics, Bill James, in an advisory capacity [18, 70].

At this point, it is appropriate to ask two main questions:

1. Is sabermetrics, as implemented by the Oakland Athletics, imitable?

2. Since some managers, who were very involved in the "Athletics culture", have been hired by other teams, has the Athletics' competitive advantage been lost?

According to the analysis provided by Hakes and Sauer (2006), it seems that the competitive advantage offered by sabermetrics may not be sustainable [18]. They argues that sabermetrics diffused rapidly, leading to a correction in baseball's labor market. This suggests the *"Moneyball anomaly"*, which consists in the disappearance of market inefficiency, once Athletics' managers have been hired by competing franchises. Even if before there could be some doubts about this anomaly, nowadays it seems that baseball's market is more efficient compared to the *Moneyball* era. However, inefficiencies can still emerge, substantially in under-explored analytical areas. Surely, the Athletics' unique historical circumstances and related organizational structure and systems developed and improved by Beane have been crucial in order to create a disadvantage for those attempting to imitate sabermetrics.

Finally, the historical success of the Oakland Athletics can be attributed to a confluence of strategic factors:

- Michael Porter's concept of interlocked activities highlights the A's ability to tightly integrate their data-driven approach (sabermetrics) into all aspects of player evaluation and acquisition.

- Building on this, Jay Barney's framework suggests that the A's organizational structure, specifically their culture of innovation and talent development, allows them to leverage these capabilities effectively.

- Richard Makadok's ideas on resource-picking and capability building shed light on how the Athletics have outsmarted the market by gathering superior information (sabermetrics) and embedding these analytical skills within the organization, making them difficult for competitors to replicate.

As said, the implementation of sabermetrics does not provide a sustainable competitive advantage. In MLB team, the Cleveland Indians, under the leadership of General Manager Mark Shapiro and his top management team, has adopted an innovative and interlocked systems approach in implementing sabermetrics. The Indians have realized this approach using two proprietary programs: DiamondView and PlayerPlan [70]. The first one is a comprehensive player database system that is updated electronically on a daily basis. It includes scouting reports, player statistics, biographical information, injury reports,

video footage, player contract, team payroll information and notes from trade discussion for the nearly 6.000 major- and minor-league professional baseball players [70]. This system aims to provide more accurate assessments of player performance and worth. DiamondView has facilitated the action of recruiting and selection, but also it has facilitated the determination of team salary distributions. Instead, PlayerPlan is a detailed program for player training and development, based on the objective of evaluating and improving each player's skill. Through evaluation, coaches and instructors identify a player's areas for improvement, encompassing physical conditioning, baseball fundamentals, and mental game. They then collaborate with the player to design a personalized development plan that fosters player ownership. This plan is documented in the player's DiamondView profile.

The approach taken by the Indians is coherent with Porter's arguments concerning unique, tailored sets of interlocked activities that reinforce one another [70].

II Statistical analysis in baseball

In this section, some principles for statistics and sabermetrics will be applied. In the first part there will be an introduction to the regression analysis. First of all, it will be taken into account the linear regression. Then, a logistic regression will be implemented. These two statistical tools are going to be used in order to outperform the Pythagorean formula, which was implemented by Bill James to measure actual or projected runs scored against runs allowed and projects a team's won-loss percentage.

II.1 Fundamentals of Statistical Analysis

II.1.1 Linear Regression

Linear regression plays a fundamental role in statistical modeling. Linear regression is a useful tool for predicting a quantitative response. In this sense, consider the regression problem in which a continuous response Y is to be regressed on a number of predictors $X_1,\ldots,X_p$. The aim of the linear regression is to provide the simplest model form model the regression function as a linear combination of predictors [60]. This is a very straightforward approach for predicting Y on the basis of a single predictor variable X.

If there is only one X, which is approximately linearly related to Y, then the regression is called simple linear regression.

Mathematically, this relationship can be written as in Equation 9:

$$Y = \beta_0 + \beta_1 X \tag{9}$$

where β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, the two constants are called coefficients or parameters of the model. Once the estimation of the two parameters has been done, calling them respectively $\hat{\beta_0}$ and $\hat{\beta_1}$, then the prediction for the Y will be done [20].

However, a generalization and a specification need to be applied. Consider a set of data $\mathscr{D} = \{(y_i,x_i) : i = 1,\ldots,n\}$ where y_i is the ith response and $\mathbf{x_i} = (x_{i1},\ldots,x_{ip}) \in \mathbb{R}^p$ is the associated predictor vector and n ($\gg p$) is the sample size [60]. Then, the linear model is specified as in Equation 10:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{10}$$

with $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$, for $i = 1, \ldots, n$. The term ε_i represents the error term, which is general for what is missed in the model. In fact, the true relationship may not be linear, there may be other variables that cause variation in y and there may be measurement errors [20]. The error term is said to be independent and identically distributed or iid and it is symmetric zero mean random variable. Identically distributed means that the distribution does not fluctuate and all its terms have the same probability distribution. Instead, independent means that the items of his sample are all independent events.
The Equation 11 shows the matrix form of Equation 10:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad \text{with} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{11}$$

where $\mathbf{y} = [y_i]_{n\times 1}$ is the n-dimensional response; $\mathbf{X} = (x_{ij})_{n\times(p+1)}$ with $x_{i0} = 1$ is often called the design matrix; and $\varepsilon = [\varepsilon_i]_{n\times 1}$ is the error term.

Looking at the work of Su et al. [60], four major statistical assumptions can be involved in order to specify Equations 10 and 11. They are listed below:

1. Linearity: $\mu \equiv \left[\mathbb{E}(y_i|x_i)\right]_{n\times 1} = \mathbf{X}\beta$;
2. Independence: ε_i's are all independent;
3. Homoscedasticity: ε_i's have equal variance σ^2;
4. Normality: ε_i's are normally distributed.

The parameters $\beta_0, \ldots, \beta_p$ are unknown and must be estimated from the data. In vector notation we can write $\beta = (\beta_0, \ldots, \beta_p)$. It is remarkable the fact that, in regression analysis, the predictors are considered as deterministic. The random variable is only on the residuals, and thus on the response variable.
In order to estimate the vector β, predicted values must be computed as Equation 12 describes:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \ldots + \hat{\beta}_p x_{i,p} \tag{12}$$

The aim of the linear regression is to choose the $\hat{\beta}$ that minimizes the sum of the squared residuals. The residual sum of squares is the sum of the difference between the true response variable y_i and the predicted variable $\hat{y_i}$. In a mathematical way, the residual

sum of squares could be written as in Equation 13:

$$\text{RSS} = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

In the case of the simple regression, it can be easily demonstrated the way by which the two expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained (see Appendix A). Both are reported respectively in the following Equations 14 and 15:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x_1} \tag{14}$$

where $\bar{y}$ represents the mean of y and $\bar{x_1}$ is the mean of the predictor.

$$\hat{\beta}_1 = \frac{\text{COV}(X_1, Y)}{\text{Var}(X_1)} \tag{15}$$

where the term $\text{COV}(X_1, Y)$ explains the covariance between the predictor X_1 and the response variable Y. The term $\text{VAR}(X_1)$, instead, represents the variance of X_1.

However, for multiple regression (see Equation 10), the expression for β is usually expressed in terms of the design matrix. As said before, the design matrix can be defined as $X = \left[1; X_1; X_2; \cdots; X_p\right]$. It is formed by the p columns of the predictors plus a column of 1's for the intercept, β_0. Next, let $\mathbf{y} = \left(y_1, \ldots, y_n\right)^t$ be the vector of the observed responses. Then, the minimization problem of the RSS is given by the Equation 16:

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t \mathbf{y} \tag{16}$$

This is called the Least Square (LS) estimate of the parameters. Now, in order to do inference, it is necessary to fix the distribution of the residuals. In this sense, a Gaussian linear model can be applied. The Gaussian linear model can be written as Equation 17 suggests:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \varepsilon_i = \left(\mathbf{X}\beta\right)_i + \varepsilon_i \tag{17}$$

with iid $\varepsilon_i \sim \mathcal{N}((\mathbf{X}\beta)_i, \sigma^2)$. Then, also $Y_i \sim \mathcal{N}((\mathbf{X}\beta_i), \sigma^2)$. It means that Y_i are independent but they do not have the same distribution, so the expected value is different at each point. Now, let denote with B the LS estimator of β, as Equation 18 shows:

$$\text{B} = \left(X^t X\right)^{-1} X^t \mathbf{Y} \tag{18}$$

It can be proven that:

- the LS estimator B is unbiased: $\mathbb{E}(\mathrm{B}) = \beta$ (the proof can be seen in Appendix A)
- if the error terms are independent with equal variance σ^2, the variance/covariance matrix of B is shown by Equation 19:

$$\mathrm{COV}(\mathrm{B}) = \sigma^2 \left(X^t X\right)^{-1} \tag{19}$$

So, under the Gauss-Markov theorem, if these assumptions are verified, then the LS estimator is called to be BLUE (Best Linear Unbiased Estimator). In fact, it will be unbiased and will have the lowest variance at all. All the others estimator will have a higher variance than the LS estimator [57].

II.1.2 Pythagorean Formula

Bill James, regarded as the father of Sabermetrics, empirically derived the following non-linear formula to estimate winning percentage, called the Pythagorean expectation.

$$W_{pct} = \frac{R^2}{R^2 + RA^2}$$

where W_{pct} is the Winning percentage and R are the Runs scored, whereas RA are the Runs Allowed.

From here, some computations can be done in order to find an exponent which would give a better fit relative to the originally proposed exponent value of 2 [46]. Starting with the replacement of the value 2 with an unknown value k, the previous formula can be written as:

$$W_{pct} = \frac{R^k}{R^k + RA^k}$$

Using some algebra, this equation can be rewritten as:

$$\frac{W}{L} = \frac{R^k}{RA^k}$$

Now, applying the logarithm on both sides of the equation, a linear relationship can be obtained:

$$\log\left(\frac{W}{L}\right) = k \cdot \log\left(\frac{R}{RA}\right)$$

The value for k is estimated using a linear regression, taking as dependent variable $\log\left(\frac{W}{L}\right)$ and the predictor is $\log\left(\frac{R}{RA}\right)$ [46].

Doing computations with R software, it can be shown that the result of the linear regression is approximately near 1.903, which is significantly smaller than value 2 [46].

II.1.3 Logistic Regression

Regression Analysis is a multivariate statistical methodology to investigate relationship and predict outcome. One type of regression analysis is known as logistic regression. Logistic regression is used when the predicted outcome is a binary variable. For binary variable is intended a variable which can take just two values. For example, on/off, infected/not infected or 0/1. Logistic regression techniques resolve inconsistencies associated with dichotomous dependent data and the assumptions of Ordinary Least Squares regression methods.

It is fair to note that, in logistic regression, the independent variables that are used for outcome prediction may be dichotomous, categorical or continuous. This feature give to this model the chance to be used in any application where binary outcomes can be predicted.

Let consider the response variable as a Bernoulli variable, in this sense it will be distributed as:

$$Y_i \sim \text{Bern}(p_i)$$

where $p_i = \mathbb{P}(Y_i = 1)$ is the probability of success of the i-th trial. Now, imagine to use a Generalized Linear Model of the form:

$$g(p_i) = \beta_0 + \beta_1 x_i$$

then the canonical link function for this model is described as:

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

This kind of canonical link function is called logit function:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

Then:

$$\text{logit}(p_i) = \eta_i = \beta_0 + \beta_1 x_i$$

where η_i is the systematic component of the regression and it is $\eta_i = x_i^t \beta$.

Logistic regression is based on the *logit* transformation of the dependent variable. The logit transformation generates a continuous logarithmic curve from non-continuous data so that a regression model can be developed.

The outcome probabilities for each dependent variable value are the basis for the model. The logit transformation is necessary since dichotomous dependent data violates ordinary least squares assumptions. Another issue with dichotomous data is that the error terms are not normally distributed, thus ordinary sum of squares regression and all normality tests are invalid [16].

Logistic regression is more flexible than ordinary least squares regression. It does not necessitate normally distributed dependent variables or equal variance. Predictions in ordinary least squares regression rely on the observed variations in the independent variables. In contrast, logistic regression is founded on the logarithm of the odds of a specific event occurring given a set of observations. The core principles of logistic regression are grounded in probabilities and the characteristics of the logarithmic curve.

The assumptions of logistic regression are that the resulting logit transformation is linear, the dependent variable is binary, and the resultant logarithmic curve is free of outliers. Both discriminant analysis and logistic regression yield similar results when dealing with dichotomous dependent variables; however, discriminant analysis is more restrictive and complex. Unlike discriminant analysis, logistic regression imposes no restrictions on the nature of the independent variables, allowing for categorical independent variables. Discriminant analysis requires strict adherence to assumptions of normality and equal variance, whereas logistic regression does not have these requirements [16].

The "problem" now is how to read the values of the parameters. The interpretation of β_0 and β_1 is easy since β_0 is the value of the intercept and β_1 is the increase of the response variable caused by an unitary increase of the predictor. This means that, in standard linear regression, the interpretation of the β_1 is very straightforward. But now it is a different framework and the information contained in the link-function is needed. On one side, there is the linear predictor η_i. On the other side, the expected value of the response variable μ_i (or p_i for the Bernoulli distribution).

To plot the regression line and to obtain the fitted values for p_i, the inverse of the link function g^{-1} must be used:

$$p_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}$$

where $\eta_i = \beta_0 + \beta_1 x_i$ is the linear predictor for observation i. For any given value of "x"

(not necessarily one of the x_i's in the data set), the predicted value is computed as:

$$\hat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}$$

In this way, it is obtained the predicted expected value for each value of x.
From the inverse of the link function it can be denoted that:

- When $\hat{\beta}_1$ is positive: as X gets larger, $\hat{p}_i$ goes to 1, as X gets smaller, $\hat{p}_i$ goes to 0;
- When $\hat{\beta}_1$ is negative: as X gets larger, $\hat{p}_i$ goes to 0, as X gets smaller, $\hat{p}_i$ goes to 1.

The intercept $\hat{\beta}_0$ is the estimate of the linear predictor when $x_i = 0$, and thus:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}}$$

is the estimate of the mean response when the predictor is set to 0.

The foundational step in logistic regression analysis involves applying a logit transformation to the dependent variable using maximum likelihood estimation, which leverages the odds ratio. The odds of an event is the ratio between the probability of the event and its complementary, so:

$$\text{odds}(E) = \frac{\mathbb{P}(E)}{1-\mathbb{P}(E)}$$

In logistic case this ratio can be written as

$$\text{odds}(Y_i = 1) = \frac{\mathbb{P}(Y_i = 1)}{1-\mathbb{P}(Y_i = 1)} = \frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)}$$

Consequently, the odds ratio for two events E_1 and E_2 can be defined as:

$$\text{OR}(E_1, E_2) = \frac{\text{odds}(E_1)}{\text{odds}(E_2)}$$

In logistic regression, taking two probabilities $p_0 = \mathbb{P}(Y_i = 1|X = x)$ and $p_1 = \mathbb{P}(Y_i = 1|X = x+1)$ and applying the logarithm to the odds ratio, the log-odds-ratio is obtained and precisely it is like:

$$\log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) = \log\left(\frac{e^{\beta_0}e^{\beta_1(x+1)}}{e^{\beta_0}e^{\beta_1 x}}\right) = \log(e^{\beta_1}) = \beta_1$$

So, an interpretation of the β_1 in terms of its absolute value can be done:

- The β_1 is the change in the log-odds-ratio when the predictor increases for "1" unit.

- The e^{β_1} is the odds-ratio comparing responses with "1" unit of difference.

The maximum likelihood estimation (MLE) is now used to estimate the coefficients from the logit transformation. MLE is similar to the ordinary least squares used in multiple regression analysis. The likelihood is the probability that the observed values of the dependent variable will be predicted by the observed independent variable data. Instead, the log likelihood (LL) is the logarithm of the likelihood, and it ranges from negative infinity to positive infinity. The logistic curve simplifies the estimation of coefficients. The maximum likelihood estimate aims to maximize the LL value and estimate the coefficients at that maximum point.

It's important to note that MLE is highly accurate for large sample sizes. Since the LL represents the log probability that the dependent variables will be predicted by the observed independent variables, our goal is to maximize that probability. The coefficient estimate at which the log likelihood is maximized will represent the highest probability that the observed dependent variable is predicted by the observed independent variables.

Another important measure used in logistic regression is deviance. Deviance is used in order to assess the goodness of fit of a model. It is derived from the likelihood function and compares the fit of the current model to a perfect model. In logistic regression, deviance can be understood as a measure of discrepancy between the observed outcomes and the outcomes predicted by the model [16].

The deviance for the logistic regression is computed from the individual contributions:

$$\begin{aligned} -2(\ell(y_i,\hat{p}_i)-\ell(y_i,y_i)) = \\ -2((y_i\log(\hat{p}_i)+(1-y_i)\log(1-\hat{p}_i))-(y_i\log(y_i)+(1-y_i)\log(1-y_i))) = \\ -2(y_1\log(\hat{p}_i)+(1-y_i)\log(1-\hat{p}_i)) = d_i \end{aligned}$$

II.1.3.1 Confusion Matrix

In this section some important features for the logistic regression are discussing. All of them are used in the following sections in order to evaluate the performance of the logistic model. The first one is the confusion matrix. It is a tool used to evaluate the performance of a classification algorithm, such as logistic regression. It provides a visual

representation of the performance by comparing the actual target values with the values predicted by the model [10, 59].

For a binary classification problem, the confusion matrix is a 2x2 table:

Table 3: Example of a 2x2 Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negatives (TN)	False Positives (FP)
Actual Positive	False Negatives (FN)	True Positives (TP)

Source: "A systematic analysis of performance measures for classification tasks" [59]

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a confusion matrix shown in Table 3 for the case of the binary classification.

Then, there are some metrics derived from a Confusion Matrix in order to better evaluate the correctness of a classification. They are listed below:

- Accuracy: the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}}$$

- Precision: the ratio of correctly predicted positive instances to the total predicted positive instances.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

- Recall: the ratio of correctly predicted positive instances to the total actual positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

- Specificity: the ratio of correctly predicted negative instances to the total actual negative instances.

$$\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}}$$

- F1-Score: relations between data's positive labels and those given by a classifier.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This tool is useful in logistic analysis because it permits to not just how many predictions were correct, but also the types of errors the model makes.

II.1.3.2 Gradient Boosting Machine

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The idea of gradient boosting originated in the observation by Leo Breiman that boosting can be interpreted as an optimization algorithm on a suitable cost function. Boosting is one of the most important recent developments in classification methodology. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced [12].

A general gradient descent "boosting" paradigm is developed for additive expansions based on any fitting criterion. Specific algorithms are presented for least-squares, least absolute deviation, and Huber-M loss functions for regression, and multiclass logistic likelihood for classification [11].

Gradient boosting of regression trees produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than clean data [11].

II.1.3.3 Receiver operating characteristic curve

The Receiver Operating Characteristic (ROC) curve is a graphical tool used to assess the performance of binary classification systems, especially in diagnostic tests. These tests categorize outcomes into two distinct groups, such as detecting the presence or absence of a disease. Often, the test results are continuous or ordinal, requiring a specific

threshold (cut-off value) to be established for making diagnostic decisions. This threshold helps determine the presence of a disease based on the test outcome. The ROC curve is instrumental in evaluating how well different cut-off values distinguish between these two categories [50].

Originating during World War II, the ROC curve was first designed to differentiate between true signals (true positives) and noise (false positives) in radar signal detection. Originally developed for signal detection and discrimination in radar systems, the ROC curve found its way into psychology to analyze perceptual and decision-making processes. Over time, its utility expanded significantly into the medical field, where it became a crucial tool for assessing the performance of diagnostic tests. Today, the ROC curve is widely applied in diverse fields including bio-informatics and machine learning, where it helps evaluate the effectiveness of classification algorithms and predictive models [50].

The ROC (Receiver Operating Characteristic) curve is a valuable tool with various advantages and disadvantages for evaluating diagnostic methods.

One of its main strengths lies in its ability to provide a comprehensive visualization of how well a test can distinguish between normal and abnormal results across the full spectrum of test outcomes. This graphical representation includes all possible sensitivity and specificity values, avoiding the need to group data into categories like in a histogram. Additionally, since the ROC curve plots sensitivity (true positive rate) against specificity (false positive rate), it remains unaffected by the prevalence of a condition in the population being sampled. This means that the evaluation can be applied consistently, regardless of how common or rare the disease is within the test population.

However, the ROC curve has its limitations. It does not directly show the specific cut-off values used to differentiate between normal and abnormal results, nor does it indicate the number of samples from which the curve is derived. These omissions can make it challenging to pinpoint the optimal threshold for decision-making. Furthermore, while smaller sample sizes often produce a jagged curve, simply increasing the sample size does not necessarily result in a smoother curve. This can sometimes lead to difficulties in interpreting the overall performance and reliability of the diagnostic test based on the visual appearance of the ROC curve.

Thus, while the ROC curve is highly effective for assessing the performance of di-

agnostic tests and classification models, these nuances must be carefully considered to ensure accurate and meaningful interpretation.

II.2 First application: baseball dataset

II.2.1 Dataset and Data manipulation

In the first application the "baseball.csv" was used. This dataset is available on Kaggle and it includes baseball data from 1962 to 2012. The dataset contains a series of variables, such as:

- Team: the name of the team.
- League: the selected league joined by the team.
- Year: as said from 1962 and 2012.
- W: number of wins during the season.
- RS: number of runs scored during the season.
- RA: number of runs allowed during the season.
- OBP: on base percentage during the season.
- SLG: slugging percentage in the season.
- BA: batting average in the season.
- Playoffs: binary variable for accessing the playoffs.
- RankSeason: the ranking of each team at the end of the season.
- RankPlayoffs: the ranking of each team that has done the Playoffs.
- G: number of games played during the season.
- OOBP: opponent on base percentage.
- OSLG: opponent slugging percentage.

This dataset contains 1232 observation with 15 variables. In order to do a good analysis, the dataset has been reduced taking in consideration the data until 2004. Now the dataset presents 992 observations with 15 variables. As the good practice of a data analyst requires, some other data manipulation was needed. In this sense, the NA values coming from the variable "RankPlayoffs" have not influenced so much the work. In fact, it is true that "RankPlayoffs" could be directly canceled out. Then, the variable relative to the difference in runs, called RD, was created by taking the difference of "R" and "RA" and it was added to the dataset.

II.2.2 Hakes and Sauer data simulation

The first aim of this work is to try to replicate the evaluation approach done by Jahn K. Hakes and Raymond D. Sauer in their paper entitled *"An Economic Evaluation of the Moneyball Hypothesis"*.

The two researchers have explored the efficacy of the Moneyball approach which emphasizes the use of statistical analysis to identify undervalued players and improve team performance cost-effectively [18].

The authors analyze Major League Baseball (MLB) data from the late 1990s to the early 2000s. They assess whether teams that adopted sabermetric principles, like the Athletics, achieved better performance relative to their payroll compared to teams relying on conventional scouting.

The study finds empirical support for the Moneyball hypothesis. Teams utilizing sabermetric strategies achieved greater wins per dollar spent on player salaries. This efficiency translated to competitive success despite smaller payrolls.

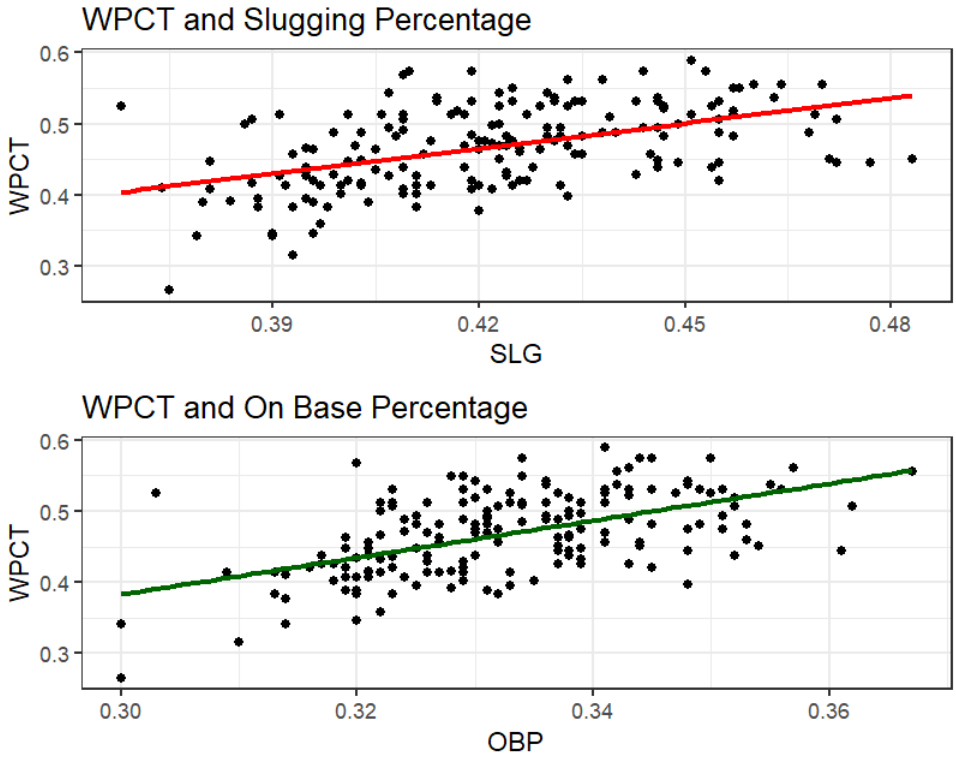
The study observes a shift in the baseball labor market as the insights from Moneyball spread. Initially, teams employing these strategies gained a significant competitive edge. However, as more teams adopted similar approaches, the market corrected, reducing the early adopters' advantage.

The integration of sabermetrics in MLB led to a more efficient market for player talent, where player salaries better reflected their true contributions to team performance. This evolution suggests that advanced analytics can sustainably improve decision-making processes in professional sports.

The scope of this section is to focus on the first analysis that the two researchers have implemented in their paper [54]. In order to replicate all the passages, a subset from "baseball" dataset has been created taking in consideration the same years, i.e. from 1999 to 2006. In addition, only the teams those have not participated to the Playoffs during the post-season have been taking into account. After that, a new variable called W_{pct} to define the Winning percentage. It is given by the ratio between Wins and the number of games played by the team.

Since Hakes and Sauer has implemented a work based on the relationship between W_{pct} and OBP and SLG, it is useful to show these kind of relationships through Figure 3:

Figure 3: Relationship between W_{pct}, OBP and SLG



Source: own computation using R software

This visualization clearly demonstrates a positive correlation between W_{pct} (winning percentage) and SLG (slugging percentage) throughout MLB history. A similar, and potentially stronger, correlation is evident with OBP (on-base percentage).

However, the magnitude of the difference cannot be determined from this perspective

alone. Therefore, a regression analysis is needed. In addition, three different regressions are ran:

1. Impact of SLG and Opponent SLG on Winning
2. Impact of OBP and Opponent OBP on Winning
3. A Mix of SLG and OBP on Winning

Results deriving from these analysis are reported in Table 4:

Table 4: Regression Results

	Winning Percentage (W_{pct})		
	SLG Regression (1)	OBP Regression (2)	OBP+SLG Regression(3)
	(1)	(2)	(3)
SLG	1.468^{***} (0.112)		0.719^{***} (0.114)
OSLG	-1.515^{***} (0.111)		-0.740^{***} (0.126)
OBP		2.995^{***} (0.189)	2.102^{***} (0.227)
OOBP		-2.730^{***} (0.162)	-1.821^{***} (0.215)
Constant	0.506^{***} (0.061)	0.400^{***} (0.078)	0.406^{***} (0.067)
Observations	176	176	176
R^2	0.634	0.734	0.811
Adjusted R^2	0.630	0.731	0.807
Residual Std. Error	0.035 (df = 173)	0.030 (df = 173)	0.026 (df = 171)
F Statistic	150.126^{***} (df = 2; 173)	238.248^{***} (df = 2; 173)	184.004^{***} (df = 4; 171)

*p<0.05; **p<0.01; ***p<0.001

Then, some comments must be done:

- SLG and OSLG are significant to explain W_{pct};

- OBP and OOBP are also significant;
- OBP and OOBP have much higher coefficients which means they have an higher impact on W_{pct};
- The R^2 value of the OBP Regression is much higher than the R^2 value of the SLG Regression. OBP can explain more of the variance of W_{pct}.

By combining both metrics in one regression the impact of OBP is shown clearly: OBP has an higher impact on W_{pct} than SLG.

II.2.3 Linear and Logistic regression to make Playoffs

In this section, an analysis related to the access to the Playoffs has been done. Starting from the data manipulation describe above, a dataset containing 992 observation was used.

Firstly, taking in consideration the idea deriving from Bill James and subsequently by Billy Beane, some data analysis was done in order to describe how the access to the Playoffs depends on the Wins of each teams. As Figure 4 shows, in general if a team want to do Playoffs (orange point in the graph) has to win surely more than 90 games in a season. It can be noted that teams with a wins counter under 90 historically had not participated to the post-season games.

To be more precise, the dashed line in the graph is located at 95. In fact, if a team reaches this threshold is very difficult to not participate to post-season. There are just few exceptions and they may be related to season in which the overall wins for each team has been very large.

In order to reinforce the thesis of good correlation between Wins and Runs Difference, it is useful to study the relationship between the two variables. Looking at Figure 5, it can be noted that run differential is a strong indicator of a team's performance over a season. Teams aiming to increase their chances of making the playoffs should focus on improving their run differential, either by boosting their offensive capabilities to score more runs or by strengthening their defense and pitching to allow fewer runs.

Teams can use this plot to benchmark their performance against others. For example, if a team has a positive RD but fewer wins, they might investigate specific games or

situations where they lost despite a strong overall performance.

Figure 4: Teams and Wins relationship to access the Playoffs

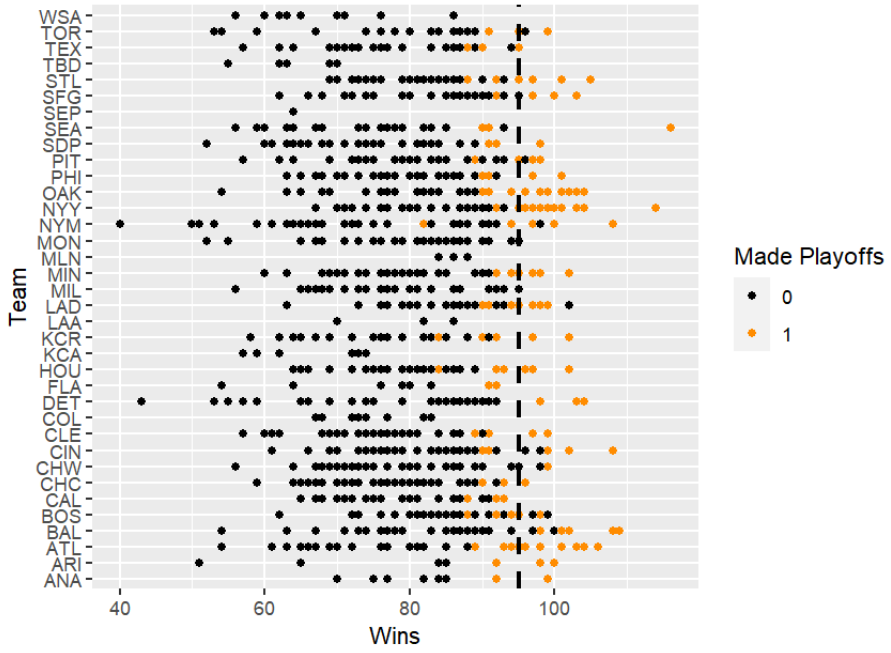
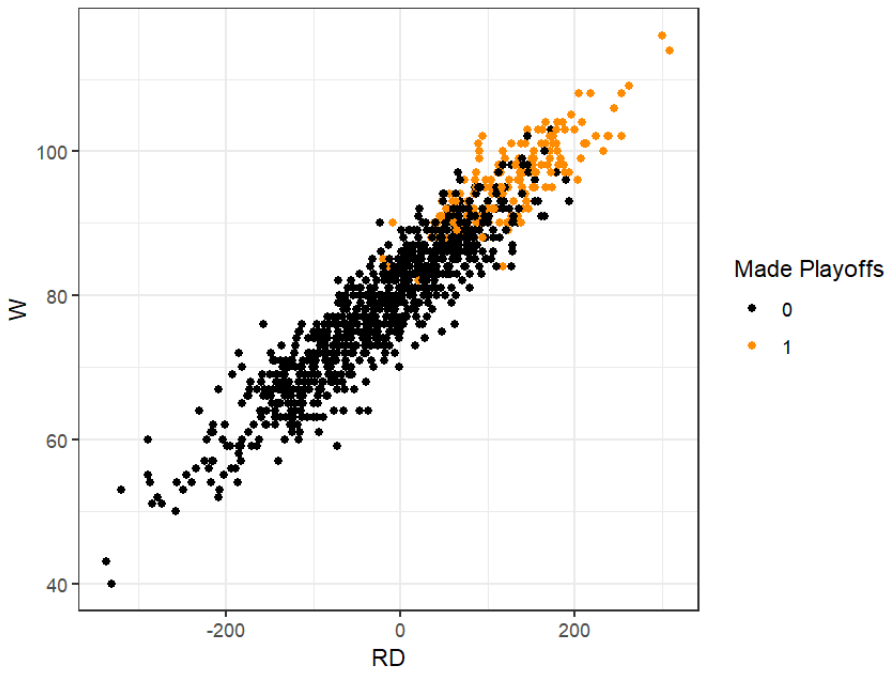


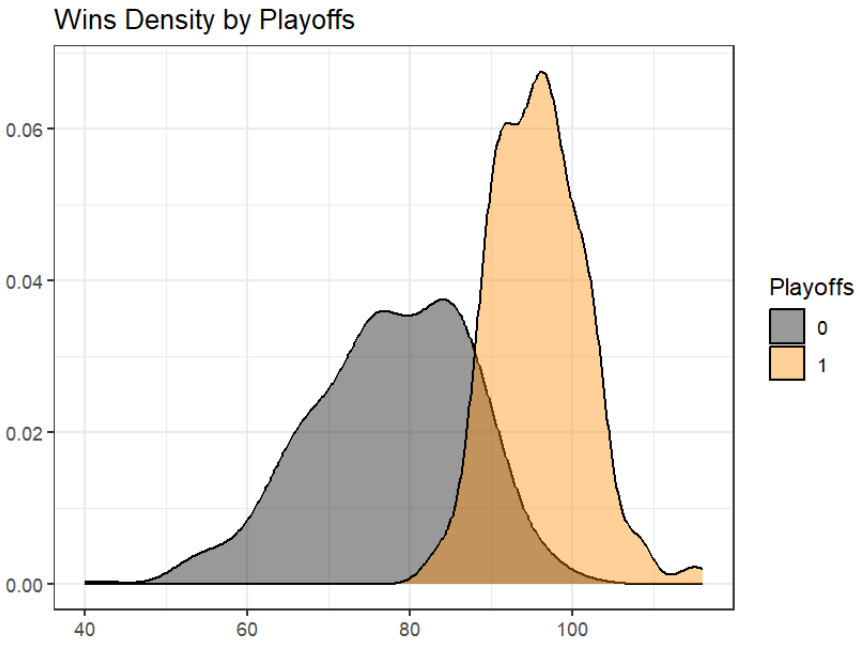
Figure 5: Wins and Runs Differential relationship to access the Playoffs



In Figure 6, the density plot clearly delineates the typical win ranges for playoff versus non-playoff teams. It highlights the crucial threshold around 85-90 wins, which often

differentiates between making and missing the playoffs. This visualization is particularly useful for baseball teams and analysts to understand how the number of wins correlates with playoff qualification and to set strategic goals accordingly.

Figure 6: Wins Density by Playoffs



II.2.3.1 Linear Regression

In this section, the aim is to try to predict the wins for Oakland Athletics in 2002, using a linear regression based on the relationship between "W" and "RD". So, the first linear model is taken considering "W" as the response variable and with "RD" as a regressor. In the previous section, the Figure 5 has already shown the relationship between the two variables and it has proven that there is positive correlation between them.

Furthermore, the regression analysis performed is summarize in Table 5. So the model taken into account suggest an equation that can be written as:

$$W = 80.8841 + 0.1057 \cdot RD$$

Then, taking for example a Win threshold set at 95, the number of RD predicted by the model will be:

$$RD = \frac{95 - 80.8841}{0.1057} = 134$$

Table 5: Linear Regression Results between W and RD

	Dependent variable:
	W
RD	0.1057***
	(0.001)
Constant	80.8841***
	(0.126)
Observations	992
R^2	0.885
Adjusted R^2	0.884
Residual Std. Error	3.953 (df = 990)
F Statistic	7,582.688*** (df = 1; 990)

*p<0.1; **p<0.05; ***p<0.01

Now, it is necessary to study the relationship between Runs Scored and On Base Percentage and Slugging Percentage, using as before linear regression model. In a similar way, a regression must be taken between Runs Allowed and Opponent On Base Percentage and Opponent Slugging Percentage. The results are stored in Table 6.
Then, the two relationship can be written following the linear model:

$$\text{RS} = -808.921 + 2768.392 \cdot \text{OBP} + 1567.943 \cdot \text{SLG}$$

$$\text{RA} = -872.583 + 2,701.761 \cdot \text{OOBP} + 1755.441 \cdot \text{OSLG}$$

All the coefficients are statistically significant in this framework. Moreover, R^2 values are very high for both models.

Table 6: Linear Regression Results for RS and RA

	Dependent vari0able:	
	RS	RA
	(1)	(2)
OBP	2,768.392***	
	(87.392)	
SLG	1,567.943***	
	(39.821)	
OOBP		2,701.761***
		(233.833)
OSLG		1,755.441***
		(131.813)
Constant	−808.921***	−872.583***
	(18.365)	(46.033)
Observations	992	180
R^2	0.929	0.906
Adjusted R^2	0.929	0.905
Residual Std. Error	25.005 (df = 989)	27.515 (df = 177)
F Statistic	6,518.222*** (df = 2; 989)	850.325*** (df = 2; 177)

*p<0.1; **p<0.05; ***p<0.01

Lets try now to predict the wins for the Oakland Athletics in 2002, using 2001 data. In 2001, the Athletics have performed an OBP = 0.345, SLG = 0.439, OOBP = 0.308 and OSLG=0.38. Taking this values and putting them into the two equation that are specified above, one can obtain a value for RS = 835 and RA = 627 and so a RD = 835-627 = 208.

Then, the Wins for the Athletics in 2002 can be predicted as:

$$W = 80.8841 + 0.1057 \cdot 208 = 102.87 \approx 103$$

Looking at the data for the season in 2002, the Athletics have won 103 games, so the

model is good in order to predict the winning based on the runs differential.

II.2.3.2 Decision Tree

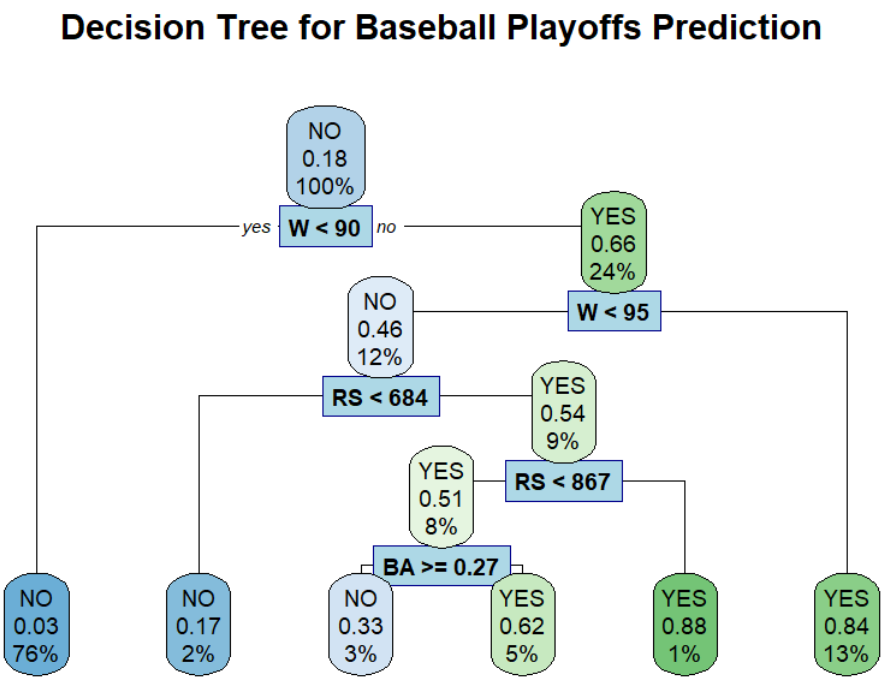
Since Playoffs variable is a binary variable because it can take values 1 if the team accesses to the Playoffs and 0 otherwise, it could be useful to try to apply a decision tree.

Decision trees are versatile tools for prediction and classification that have a long history in computational statistics. They were among the earliest statistical algorithms to be digitized with the advent of electronic computing in the latter part of the 20th century. Over time, they have evolved into essential methods across various disciplines, serving key roles in prediction, classification, artificial intelligence (AI), machine learning, and knowledge discovery. They are now fundamental to many data mining and AI applications [9].

The core feature of decision trees is their ability to recursively partition a dataset based on the values of input variables, known as predictors. This process creates a hierarchical structure of nodes, where each node represents a decision point. These nodes split the data into subsets that become increasingly homogeneous with respect to the target variable as you move down the tree. Each subset, or leaf, groups together data points with similar target values, while ensuring distinct separation between groups at each level of the tree. This hierarchical subsetting allows for progressively refined predictions or classifications within the data structure [9].

After the creation of a labeled variable for Playoffs, assigning "YES" to the chance to go to Playoffs and, on the contrary, the value "NO", a decision tree has been created taking in consideration three main variables: Wins (W), Runs Score (RS) and Batting Average (BA). The result coming from the tree is represented in Figure 7.

Figure 7: Decision Tree for Baseball Playoffs Prediction



Following the tree, it can be noted that teams with wins under 90 will definitely not go to Playoffs. They represent the 76% of the dataset. The remaining 24% now have to overcome the next threshold set at 95 wins in a season. If a team has a number of wins larger than 95, then it will go directly to Playoffs.

In the other case, a next step must be taken into account. In fact, looking at Runs Score, if the team has reached a number of Runs under the threshold set at 684 then it will be out for the post-season. Again, if the team scores runs between 684 and 867, then it can go to the Playoffs only if its Batting Average is greater or equal to 0.27; otherwise it will be out. If the team overcome the Runs scored threshold, then it will access to the Playoffs with a probability of 88%.

II.2.3.3 Logistic Regression

After the study of linear regression, logistic regression has been applied to the variable Playoffs [1]. Starting with some data manipulation, the summary statistics show that there are quite many NAs in RankSeason, RankPlayOffs, OOBP and OSLG. Then the idea was

to impute the null values so that these have some values per the imputation and our model has some robust and complete data to deal with.

After the session of data manipulation, a logistic regression with substantially all the variable has been run in order to catch the most significant variables to predict Playoffs appearance. The results of the regression are stored in Table 7.

Table 7: Logistic Regression Results for Complete Model

	Dependent variable:
	Playoffs
RS	0.001
	(0.006)
RA	0.0003
	(0.004)
W	0.353***
	(0.043)
OBP	26.988
	(23.091)
SLG	3.771
	(11.431)
BA	−7.371
	(20.580)
OOBP	−2.002
	(16.057)
OSLG	3.255
	(8.471)
Constant	−43.206***
	(7.542)
Observations	992
Log Likelihood	−189.233
Akaike Inf. Crit.	396.466
	*p<0.1; **p<0.05; ***p<0.01

From the table it can be noted that, taking this kind of regression, only "W" is statistically significant and has an impact on Playoffs. Then, using the vif() command from R, it is true that this model presents a problem of multicollinearity.

So, two other models have been run. The first one is the so called "null model" and it is composed only by the constant term. Instead, the second one has as predictors "W",

"OBP" and "SLG". Again, this model presents statistical significance only for "W", but in some sense, it can be considered better than the previous model.

Table 8 compares two nested logistic regression models to evaluate if adding predictor variables improves the model's ability to explain the dependent variable, which in this case is whether a team makes the playoffs.

Table 8: Results from Anova

Resid. Df	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
991	933.54			
988	378.90	3	554.64	2.16e-16***

The first row is related to the null model, including only the intercept. It assumes that no predictors are used to explain the outcome.

Null model has 991 residual degrees of freedom. This is calculated as the number of observations (992) minus the number of parameters estimated (1 for the intercept). Instead, the second model has 988 residual degrees of freedom, reflecting the estimation of three additional parameters ("W", "OBP", "SLG"), reducing the degrees of freedom by 3.

The deviance is the difference between the residual deviance of null model and full model (933.54 - 378.90 = 554.64). A large deviance indicates that the full model significantly reduces the residual deviance compared to the null model.

A very low p-value indicates an extremely significant difference. This means that the reduction in deviance when adding the predictors is highly significant.

Then, the Mc-Fadden R^2 has been calculated using the following formula:

$$R^2_{\text{Mc-Fadden}} = 1 - \frac{\text{log-likelihood of the full model}}{\text{log-likelihood of the null model}}$$

This procedure leads to a R^2 equal to 0.5941, which is quite good for logistic regression.

Next, a division into training and test sets to the dataset has been applied. The division has followed the classic 80% partition for training set and 20% for test set. The logistic regression has been run on the training set and it has provided similar results as in the previous case. Then, taking into account various thresholds for the sensitivity, the confusion matrix are reported in the next Tables 9, 10 and 11.

Table 9: Confusion Matrix with threshold 0.5 in Training set

	FALSE	TRUE
0	626	25
1	43	99

Table 10: Confusion Matrix with threshold set at 0.7 in Training set

	FALSE	TRUE
0	637	14
1	66	76

Table 11: Confusion Matrix with threshold set at 0.2 in Training set

	FALSE	TRUE
0	576	75
1	14	128

The values of the threshold will be selected based on the tolerance for precision and specificity. For the 0.5 threshold, the overall accuracy for the model - using formulas in Section 2.1.3.1 - is 0.9143. So, it is quite high. The precision has a value of 0.7984 and a specificity of 0.9616.

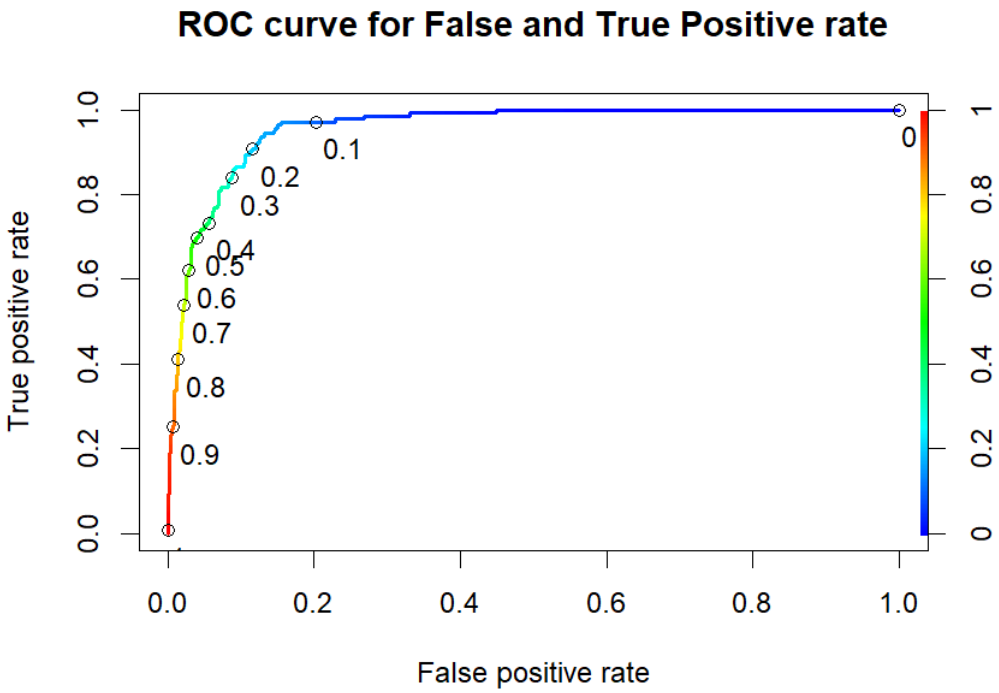
For the 0.7 threshold , the overall accuracy is equal to 0.8991, with a precision of 0.8444 and a sensitivity of 0.9622. Hence, there is a trade-off between sensitivity and specificity if 0.7 is selected as the value of the threshold. If the threshold increases, the value of precision drops and the value of the specificity increases.

Instead, for the 0.2 threshold, the overall accuracy is equal to 0.8878. In this case the precision is 0.6305 and a specificity of 0.8848. If the threshold decreases, the value of precision decreases too and the value of specificity drops. Surely, the threshold set at 0.2 must not be used.

Then, using Receiver Operating Characteristic (ROC) curves some analysis can be made. In Figure 8, the plot evaluates the performance of a binary classification model.

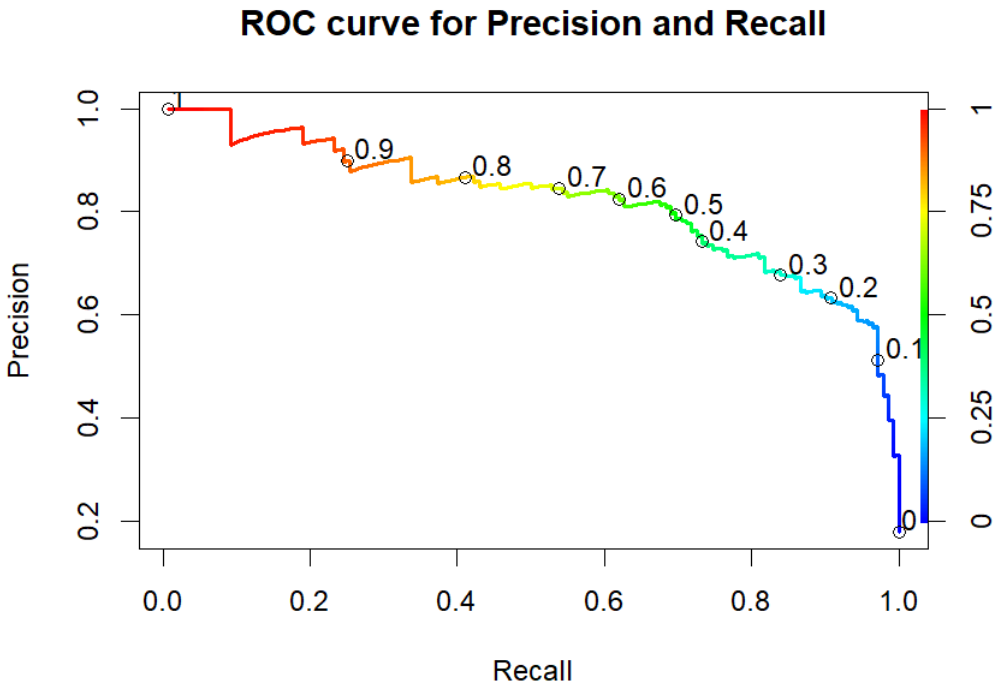
The curve itself is a plot of the true positive rate against the false positive rate for various threshold values. It starts at (0,0), where both the true positive rate and false positive rate are zero, and ideally, moves towards (1,1) as the true positive rate increases to 1. The curve quickly ascends to a high true positive rate, indicating that the model is effective in correctly identifying positives. For a substantial range, the false positive rate remains low, which means the model also avoids incorrectly classifying negatives as positives.

Figure 8: ROC curve for False and True positive rate in Training set



In Figure 9 is shown the ROC curves for Precision-Recall. The Precision-Recall curve shows a high initial precision, then a decline as the recall increases. This suggests that the model can initially identify positive cases (e.g., making Playoffs) with high accuracy, but as it tries to capture more positive cases, it starts to include more false positives.

Figure 9: ROC curve for Precision and Recall in Training set



Now, the same procedure must be applied to the test set in order to verify the correctness of results. In this case, test set is composed by 199 observations. The following Tables are representing the Confusion Matrix (as for training set), obviously computed on 199 observations.

Table 12: Confusion Matrix with threshold 0.5 in Test set

	FALSE	TRUE
0	158	5
1	5	31

Table 13: Confusion Matrix with threshold set at 0.7 in Test set

	FALSE	TRUE
0	162	1
1	11	25

Table 14: Confusion Matrix with threshold set at 0.2 in Test set

	FALSE	TRUE
0	150	13
1	3	33

For the 0.5 threshold, the overall accuracy for the model is 0.9498. So, it is quite high. The precision has a value of 0.8611 and a specificity of 0.9693.

For the 0.7 threshold , the overall accuracy is equal to 0.9397, with a precision of 0.9615 and a sensitivity of 0.9939. In the test set, the threshold set at 0.7 could be a very interesting alternative to the one set at 0.5. In fact, both precision and sensitivity are higher with respect to the threshold used before.

Instead, for the 0.2 threshold, the overall accuracy is equal to 0.9196. In this case the precision is 0.7174 and a specificity of 0.9202. Being the worst, this threshold must not be taken into consideration.

Looking at ROC curves shown in Figures 10 and 11. The two plots show a trend very similar to the one in the training set. Then, it is assumed that the model made a good performance in order to predict logistic results.

Figure 10: ROC curve for False and True positive rate in Test set

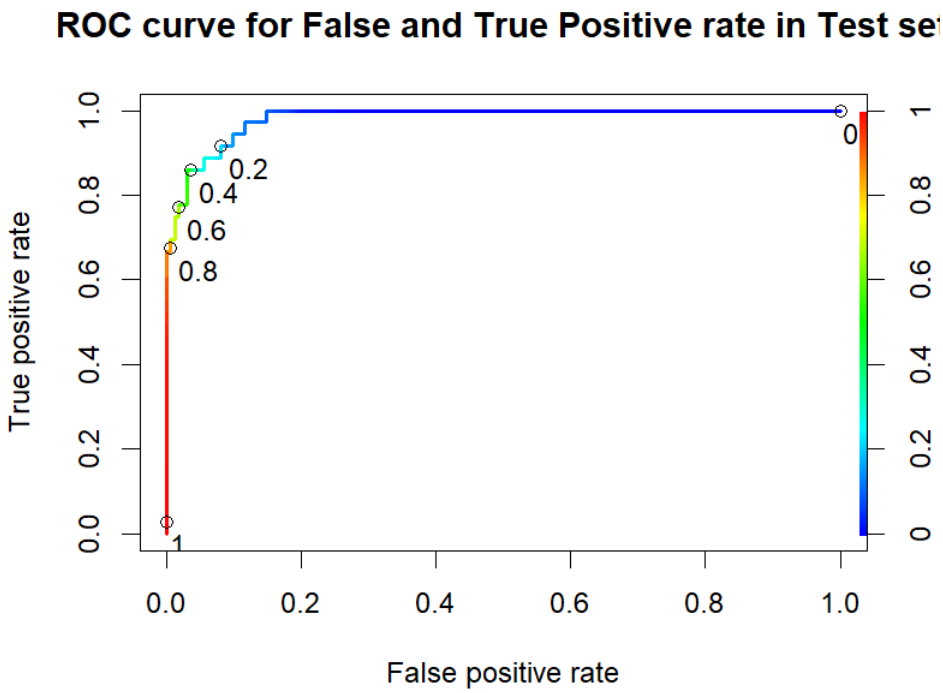
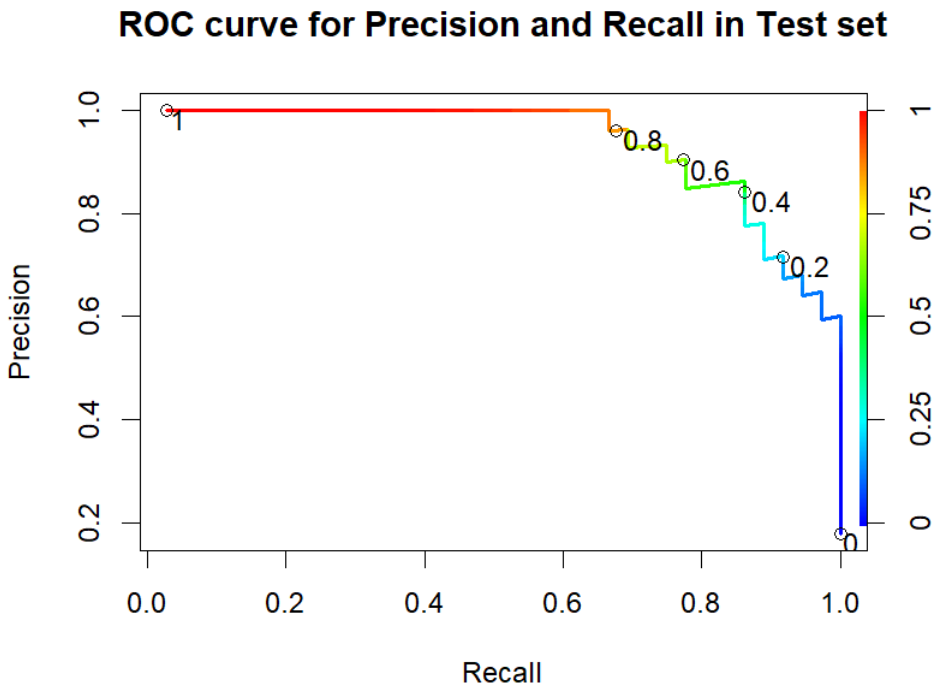


Figure 11: ROC curve for Precision and Recall in Test set



II.3 Second application: Teams and Salaries dataset

In the third part of this work the Lahman database was used [56]. Lahman database is one of the most important database related to baseball. In fact, it is reach of datasets from which one can choose the best dataset for his eventuality. Specifically, the analysis is concentrated on "Teams" and "Salaries" datasets. The first dataset does not differ too much from "baseball" used in Subsection 2.2.1.

II.3.1 Exploratory Data Analysis

In this section will be presented an exploratory data analysis (EDA) of the chosen datasets [53]. Since in this section the logistic regression will be applied on other variables which are not the evergreen wins and run differential, a most specific analysis can be applied in order to describe relationship between variables. However, before starting some general plots may be visualized.

Since the aim of this work is to eventually predict Playoff teams, first one should want to look into what makes a playoff team.

The factor the relates to Playoff appearances is Wins, so is necessary to look at their distribution and discover on average how many wins a team need to make their way into October. As discussed before, a threshold of more or less 95 Wins would lead to the access to Playoffs. However, in Figure 12 is represented an histogram in which are present also two dashed lines: the green one is for the median, whereas the red one is for the mean. It can be noted that the two are very close and they are set around 95.

Figure 13 presents a boxplot. This plot better represents the data. In this boxplot one can see how rare it is for some teams to actually make the playoffs in some certain years and how common it was for some teams to make the playoffs. A black dashed line is plotted at the 94 wins where a team were 99.8% likely to make the playoffs and a red line is plotted at 88 wins where the team had a 94% chance of making the playoffs.

Figure 12: Histogram of Win Counts with mean and median line

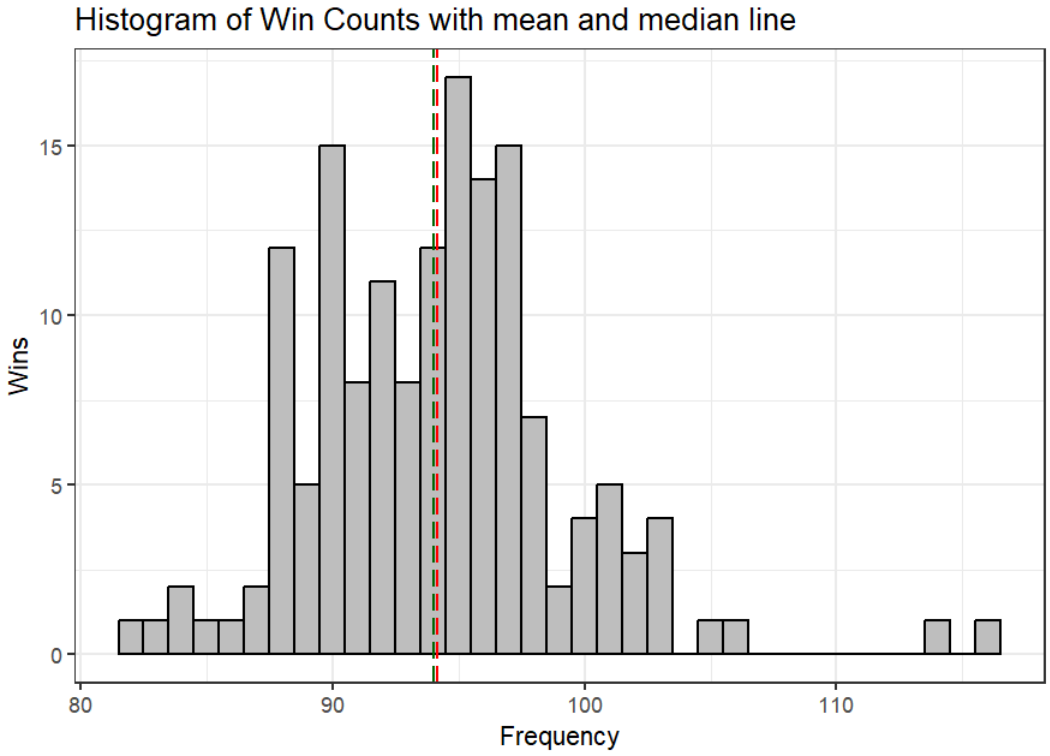
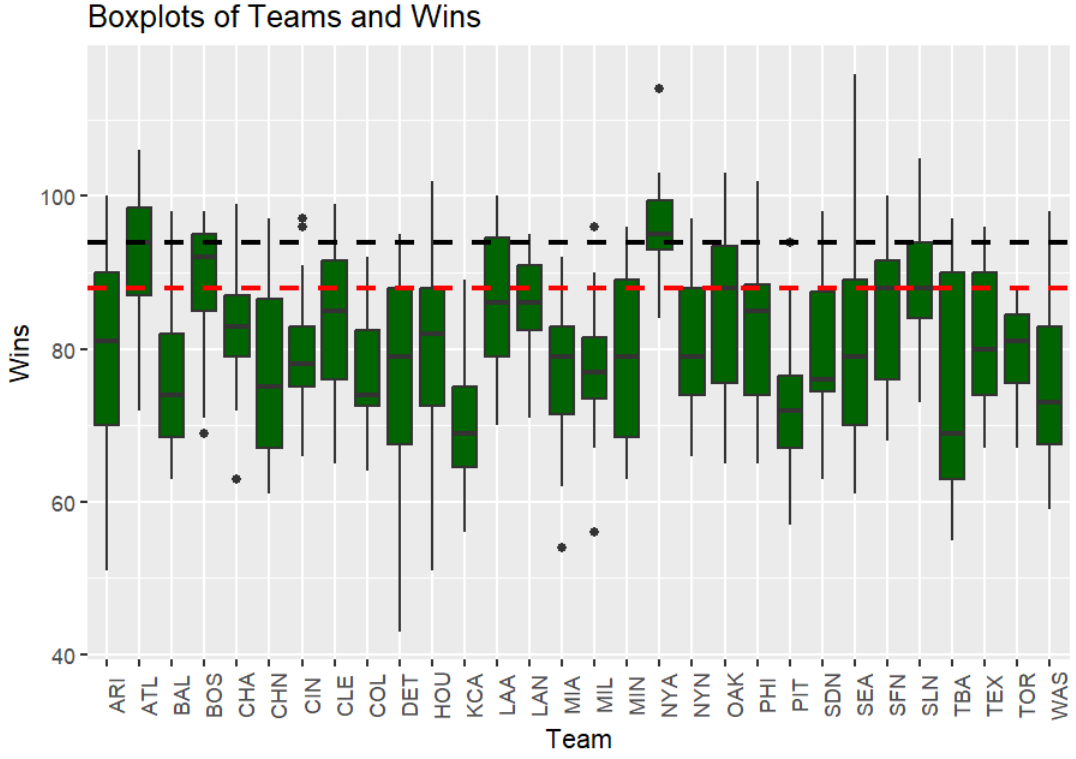


Figure 13: Boxplots of Teams and Wins

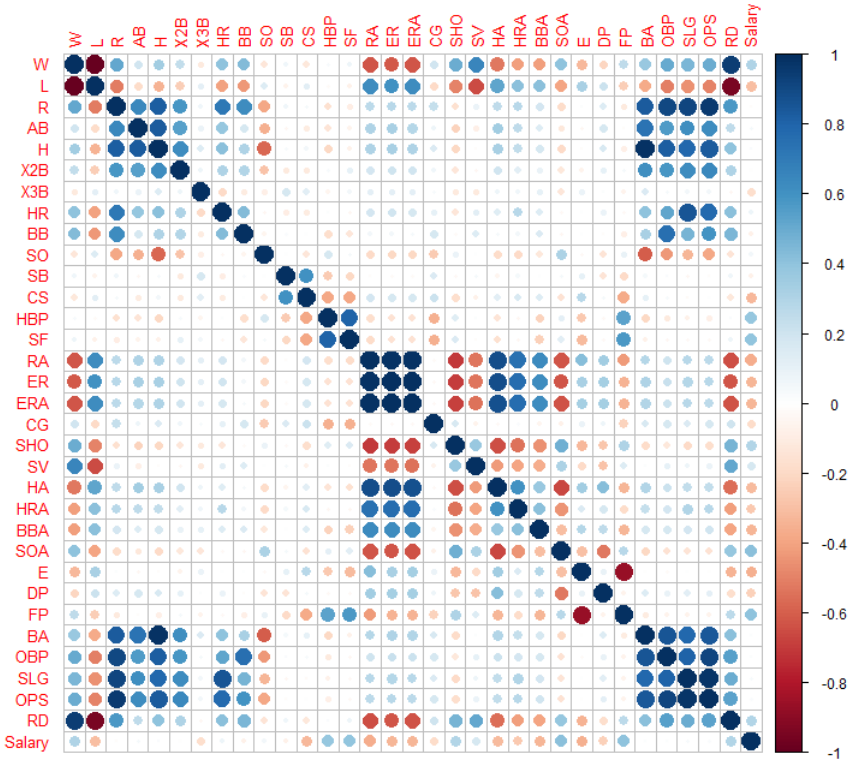


There are obviously a lot of independent variables that are correlated with one another since baseball stats have a lot to do with one another. In Figure 14 is represented the correlation plot between all the variables in the merged dataset.

This visualization displays the correlation coefficients between a large number of variables. Each cell in the grid represents the correlation between two variables, with red tones indicating positive correlation, blue tones indicating negative correlation, and white representing no correlation. Darker colors indicate stronger correlations.

In order to forecast Playoff outcomes, a team must accumulate numerous victories. Winning necessitates surpassing opponents in run scoring versus run conceding, which in turn relies on strong hitting, pitching, and defensive capabilities. Hence, nearly every variable holds the potential to forecast Playoff appearances. This will be the field for Subsection 2.3.3,

Figure 14: Correlation Plot

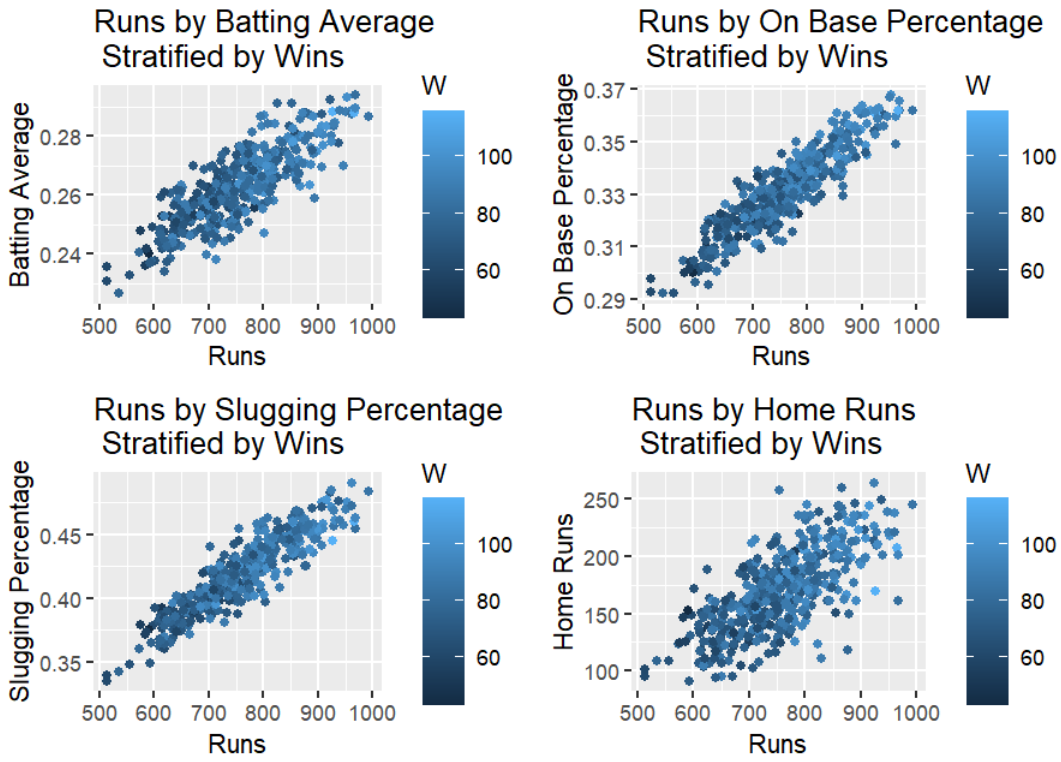


Now, since Wins and Runs are very obvious in predicting Playoffs, it is compulsory to look at real baseball statistics that predict Runs, so we can pick out some variables to explore them further. Starting from hitting plots, display in Figure 15, it is clear that

Runs is positively correlated with respectively Batting Average (BA), On Base Percentage ("OBP"), Slugging Percentage ("SLG") - which have already been demonstrated - and Home Runs ("HR"). Correlations are all positive and they have this values:

- 0.832 between Runs and "BA";
- 0.905 between Runs and "OBP";
- 0.912 between Runs and "SLG";
- 0.712 between Runs and "HR".

Figure 15: Relationship between Runs and other variables

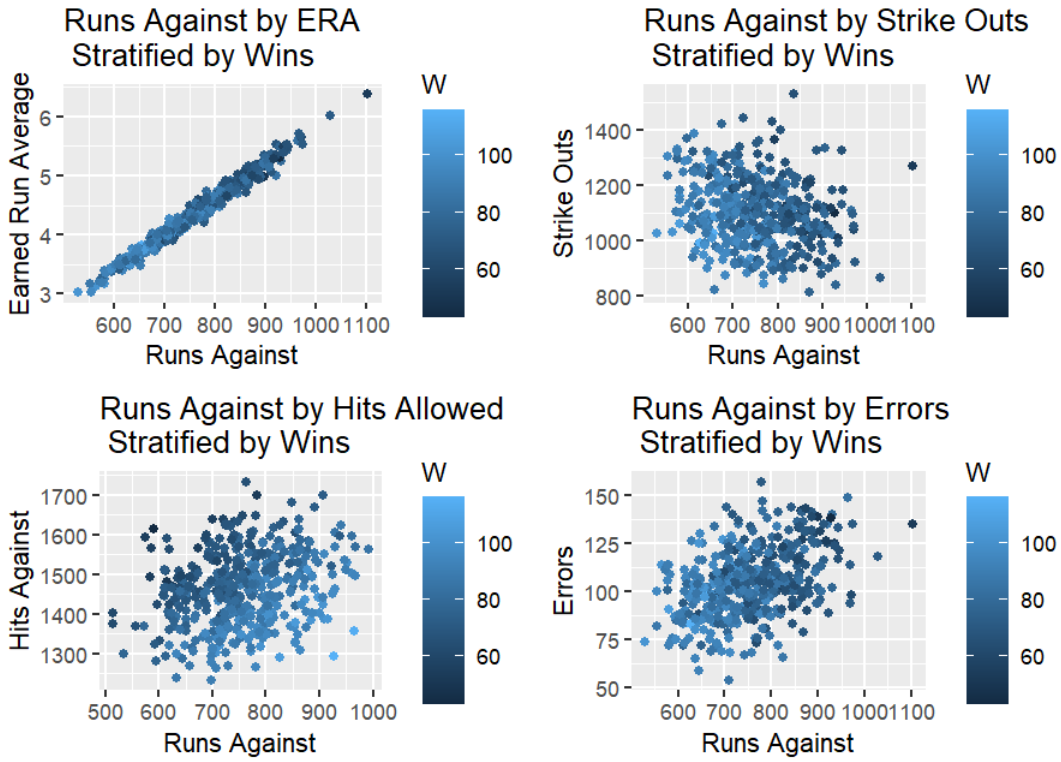


As one can evaluate offensive statistics, an analysis can be made taking into account the most important defensive statistics. Figure 16 presents this concept, basing attention to Runs Against (or Allowed) with respectively Earned Runs Average ("ERA), Strike Outs ("SO"), Hits Allowed ("HA") and Errors ("E"). In this case only the correlation between Runs Against and "ERA" is extremely positive, with a value of 0.989. In fact from the plot, the distribution of the points is similar to the bisector at 45 degrees, which represents

perfect positive correlation. In all the other cases a precise distribution cannot be specified. In fact, the correlation values for the other three variables are respectively:

- -0.198 with "SO";
- 0.883 with "HA";
- 0.422 with "E".

Figure 16: Relationship between Runs Against and other variables



II.3.2 Pythagorean Expectation on Wins

Since Pythagorean formula was used by Bill James to predict Wins given by Runs scored and Runs allowed, for the sake of completeness, it can useful to demonstrate a bit his computation.

After the creation of the Win percentage variable and of the Runs differential, taking the year starting from 1996, a comparison between Pythagorean formula and the one derived from the linear regression model can be done [46]. The results of the linear regression are reported in the Table 15.

Table 15: Results for Linear Regression between W_{pct} and RD

	Dependent variable:
	W_{pct}
RD	0.001^{***}
	(0.00001)
Constant	0.500^{***}
	(0.001)
Observations	566
R^2	0.883
Adjusted R^2	0.883
Residual Std. Error	0.024 (df = 564)
F Statistic	$4{,}247.428^{***}$ (df = 1; 564)
	*p<0.1; **p<0.05; ***p<0.01

As Hakes and Sauer have demonstrated in their paper [18], the Run Differential is statistically significant for W_{pct}.

Instead, using the Pythagorean formula, it will lead to a value for the residuals equal to 0.02435155. The RMSE calculated on the Pythagorean predictions is similar in value to the one calculated with the linear predictions [46].The Pythagorean formula is very similar to the linear regression model. It also has the added benefit of reacting better at the extremes. Thus it does not seem justifiable using a more complex model. However, the Pythagorean expectation has several desirable properties missing in the linear model [46].

Then, following the steps in Subsection 2.1.2, the Pythagorean formula con be optimized. The value that optimizes the formula is equal to 1.889. While it is significantly different from a statistical point of view, it also makes sense intuitively since it is close to 2.

II.3.3 Logistic Regression with different variables than Wins

II.3.3.1 Data Manipulation and Different Models

In this part, the aim is to find the best logistic regression, that could predict the appearance to the Playoffs, using other variables which are not Wins and Run Differential. First of all, after the loading of the two datasets, there was the phase of clean up the data,

in this case in order to make previous team names and cities match the newer one.

Furthermore, since the two datasets do not have the Playoffs variable it must be created looking at the Division Wins of each team. Playoffs is a binary variable and it can assume two different label: "Y" if the team has reached the Playoffs; "N" otherwise.

In this case, the subset data only include wild card teams. The wild card was introduced in 1995. In baseball, a "wildcard" refers to a team that qualifies for the playoffs despite not winning its division [44]. In 1995 every team played less than 146 games, so year 1996 was used as the cut off.

Then, after the merge of the two main datasets, 530 observations were ready to be used. Obviously the dataset was split in training set (at 80%) and in test set. The training set was composed by 424 observations, whereas the test was composed by 106 observations.

Now, after all the phase of data manipulation and data splitting, the dataset is ready to be used for logistic regression. Firstly, five different logistic regressions were run. They are summarized below, taking always binary variable Playoffs as response:

- Model 1 $\rightarrow$ just one predictor: "OPS".
- Model 2 $\rightarrow$ two predictors: "OPS" and "ERA".
- Model 3 $\rightarrow$ predictors: "OPS", "ERA", "E".
- Model 4 $\rightarrow$ predictors: "OPS", "ERA", "E", "Salary".
- Model 5 $\rightarrow$ predictors: "OPS", "SF", "SO", "HA", "RA", "SV", "BBA", "DP", "HRA", "HR", "AB", "R", "ER", "X2B", "H", "SB", "BB", "HBP", "SOA", "E", "SHO", "OBP", "Salary", "SLG", "X3B", "CG", "BA", "CS", "ERA"

The last model substantially all the variables have been put inside the model to just look at the result, however the research is not really interested in this model as it is not very interpretable model and need to only select certain subset of variables.

Then, in order to evaluate the best model, the best accuracy, in Table 16, and the best area under the curve (AUC), in Table 17, has been analyzed. AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.

Table 16: Accuracy table for Logistic prediction in Training set

Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
0.7217	0.8703	0.8726	0.8750	0.9198

Table 17: Area Under the Curve Accuracy for Logistic prediction in Training set

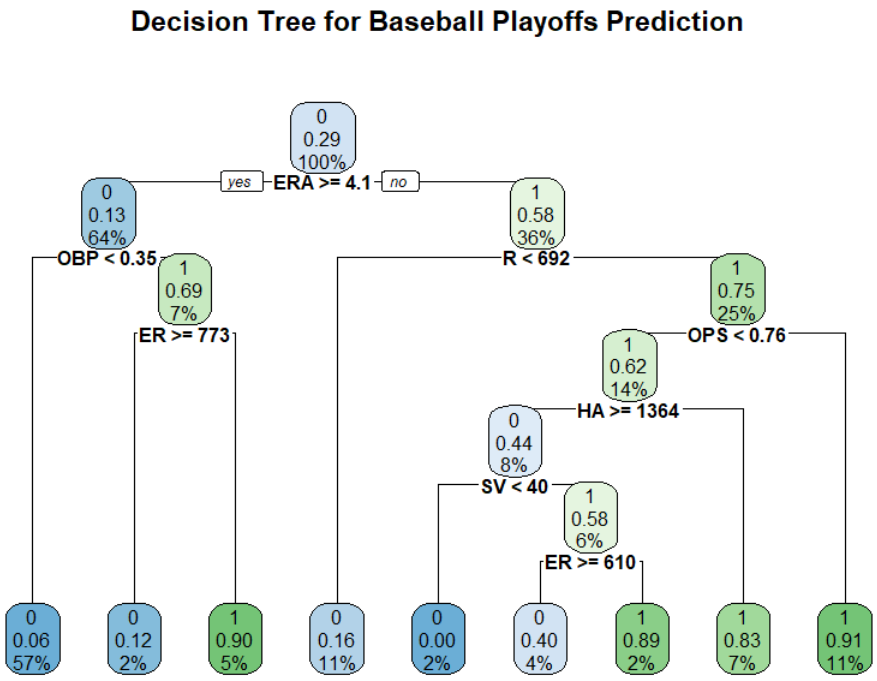
Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
0.7039	0.9355	0.9351	0.9357	0.9684

Even when all of variables are put into our model, there was only a slight increase in predictive performance, so simpler is better. Therefore, the best model for logistic regression would be Model 2.

II.3.3.2 Decision Tree

A decision tree can then be used. Decision trees are typically not highly predictive but are excellent for easy interpretability, and the plot can be comprehended even by non-analysts. Cross-validation is first used to set the complexity parameter to better create the tree. The decision tree is represented in Figure 17 below.

Figure 17: Decision Tree for Playoff Prediction on Training set



At the root of the tree is the metric ERA (Earned Run Average), with a threshold of 4.1. This initial split indicates that teams with an ERA greater than or equal to 4.1 are generally less likely to make the playoffs. If a team's ERA is below 4.1, they follow the right branch, increasing their chances of playoff qualification.

For teams with an ERA above 4.1, the next important metric is OBP (On-Base Percentage), with a critical value of 0.35. Teams with an OBP less than 0.35 are unlikely to make the playoffs, as evidenced by the lower probability in this leaf. For those with a higher OBP, the decision tree evaluates the number of earned runs (ER). A team allowing 773 or more earned runs has a reduced probability of making the playoffs. Conversely, teams with fewer earned runs, even with a high ERA, have a somewhat higher chance of playoff success.

On the other hand, if a team's ERA is less than 4.1, the tree examines their run production, specifically if they score fewer than 692 runs. Teams that score less than this threshold typically do not make the playoffs. For teams scoring more, further splits based on hits allowed (HA) and on-base plus slugging (OPS) are considered. Teams allowing

more than 1364 hits are analyzed further for their OPS; those with an OPS lower than 0.76 are less likely to make the playoffs. For those with fewer hits allowed, additional splits consider the number of saves (SV) and the total earned runs (ER).

Each leaf node in the tree gives a final decision. Green nodes indicate teams likely to make the Playoffs (class "1"), and blue nodes represent teams unlikely to make it (class "0"). These nodes also provide the probability of the prediction and the percentage of total samples that fall into each category.

From this decision tree, several key insights emerge. A lower ERA is critically associated with making the Playoffs, highlighting the importance of pitching strength. Offensive metrics like OBP and OPS are also significant, as teams with higher values in these categories are more likely to qualify for the postseason. Defensive metrics, particularly the number of earned runs and hits allowed, further influence Playoff chances. This indicates that a balanced performance across both offensive and defensive metrics is essential for a team aiming to secure a Playoff spot. Overall, this decision tree clearly outlines the interplay between various performance factors and their impact on a team's Playoff prospects.

Furthermore, the prediction accuracy and AUC on training set have been evaluated. The results are stored in Table 18.

Table 18: Decision Tree Confusion Matrix on Training Set

	FALSE	**TRUE**
0	289	12
1	29	94

In addition, the accuracy of this model is equal to 0.9033, whereas the AUC is 0.9029.

II.3.3.3 Gradient Boosting Machine

Taking again Model 5, a gradient boosting machine (gbm) was used to predict whether a baseball team will make the playoffs based on various performance metrics. After the setting of the random seed for generating random number, the random processes involved in training the model will produce the same results each time the code is run.

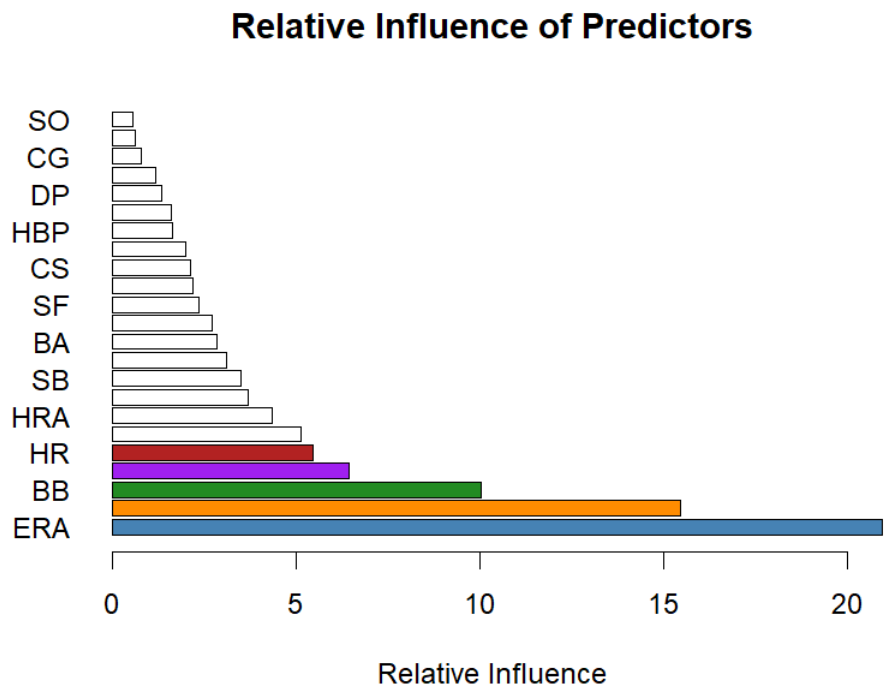
Since the Playoffs variable is binary, then the bernoulli distribution was set, with a number of trees equal to 5000. This means the model will consist of 5000 decision trees, each contributing to the final prediction. Running the algorithm, gbm has lead to the results reported in Table 19:

Table 19: Gradient Boosting Machine Confusion Matrix on Training set

	FALSE	**TRUE**
0	301	0
1	0	123

Since in Table 19 there are not value for false positives and false negatives, the gbm model has perfectly predict the results for the regression, even if there were a great number of variables. One could look at the importance of each variable from the boosting model. This is shown by Figure 18:

Figure 18: Relative Influence of Predictors



II.3.3.4 New Logistic Model

Using the level of influence of the predictor, looking at Figure 18, then another logistic model can be created trying to add the top 5 predictors form the Gradient Boosting Machine. Then, the new logistic regression will take this form:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{OPS} + \beta_2 \text{ERA} + \beta_3 \text{BB} + \beta_4 \text{HR} + \beta_5 \text{BBA} + \varepsilon_i$$

Results from this regression are stored in Table 20:

Table 20: Logistic Regression Results on Training set

	Dependent variable:
	Playoff
OPS	52.832^{***}
	(7.775)
ERA	-5.529^{***}
	(1.045)
BB	0.007^{**}
	(0.003)
HA	0.003
	(0.004)
BBA	-0.008^{**}
	(0.004)
Constant	-21.992^{***}
	(5.768)
Observations	424
Log Likelihood	-117.741
Akaike Inf. Crit.	247.483

*p<0.1; **p<0.05; ***p<0.01

Except "HA", the other variables are at least all statistically significant at $\alpha = 5\%$. For example, the coefficient for "BB" indicates that for each unit increase in "BB", the log-odds of making the playoffs increase by 0.007. Transforming to an odds ratio, each unit increase in "BB" multiplies the odds of making the playoffs by $e^{0.007} \approx 1.007$. This suggests a positive association between higher "BB" and the likelihood of making the Playoffs.

II.3.3.5 Results on Test set

Finally, the results coming from the latest model are applied into the Test set, which is composed by 106 observations. In this brief section, the analysis of the ROC curve and of the confusion matrix will be made.

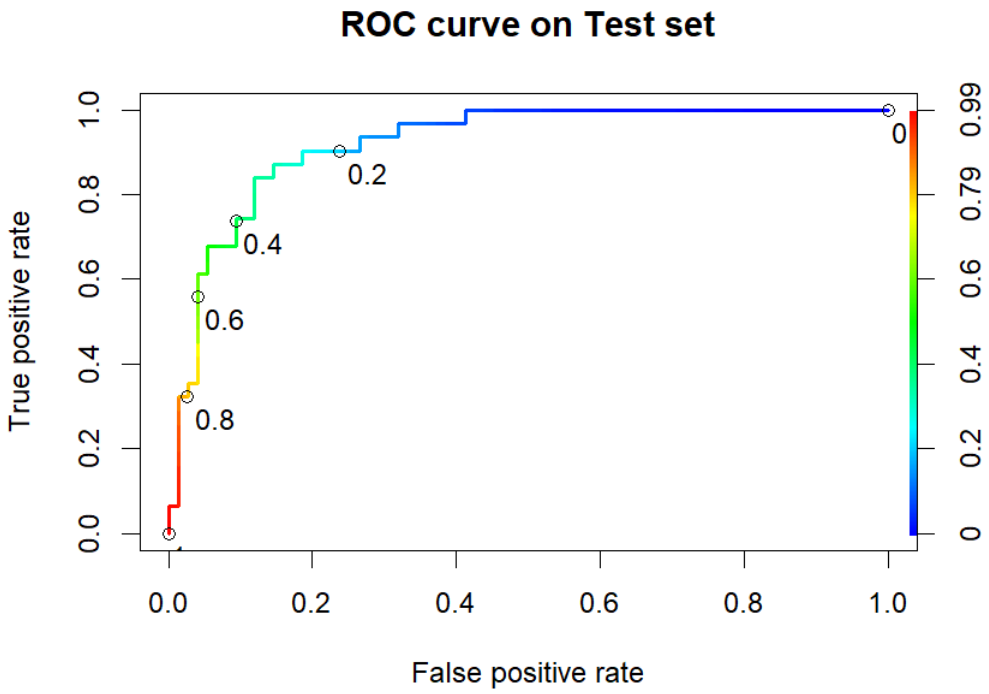
Starting from the confusion matrix, the results can be described by looking at Table 21.

Table 21: Confusion Matrix on Test set

	FALSE	**TRUE**
0	71	4
1	11	20

Here the model get a test accuracy of 0.8585 and AUC of 0.9213, only a couple of percentage point lower than our training set predictions, meaning that the latest model did not overfit the data.

Figure 19: ROC curve on Test Set



From the ROC curve point of view, the situation is plotted in Figure 19 above. From the graph it can be noted that the lack of accuracy is depicted by the curve. If the threshold is set between 0.4 and 0.6, then there will be a higher true positive rate and a lower false positive rate than looking at other thresholds.

III An analysis on salaries in MLB

In this section it will be implemented an analysis relative to salaries in the MLB. In general, salary's theme is important in sport. Baseball is not an exception, it is sufficient to remember how the Athletics have decided to reduce their salary cap because of the change in their team management.

III.1 Salary background and history

While the on-field statistics of baseball, such as batting averages, pitching records and fielding percentages, are well known and easily accessible, the sport's financial history has been less transparent. Nowadays. multi-year, multi-million dollar contracts are commonly reported s searchable as the game's traditional statistics. However, this transparency is recent.

The financial side of baseball was not well-known before 1985, when the Major League Baseball Players Association (MLBPA) began to regularly disclose player salaries. Salary information was not readily available until a trove of financial documents was discovered in recent years.

These documents have dramatically changed the understanding of baseball's financial past. In fact, there were discovered team financial documents for the New York Yankees and Philadelphia Phillies. Nowadays, one can have access to thousands of observation of salary information and detailed financial documents [58].

The concept of a player's salary is not as straightforward as it might initially seem. In fact, a player's compensation may include, in addition to the base salary, various bonus payments. So, a player's salary can be complex to define, particularly when considering historical contracts.

Historically, the notion of salary was unclear because of the lack of guaranteed contracts. Until players gained sufficient bargaining power in the 1970s, most contracts

weren't guaranteed. Instead, this kind of contracts often included ten-day clauses that allowed team owners to waive a player with just ten days' notice, effectively nullifying the remainder of the contract.

For instance, a player might have signed a contract for a $6,000 salary, but if he was waived halfway through the season, he would only receive $3,000. This raises the question: should the player's salary be considered as the total amount stipulated in the contract, or the actual amount he was paid?

In contrast, today's contracts are guaranteed, meaning players are entitled to their full contracted salary even if they are released by the team. This shift simplifies the definition of a player's salary in modern terms but underscores the complexities involved in understanding historical player compensation [58].

Stand on what USA Today depicts in one of their articles, in 2024 the New York Mets opened the season with the highest-player payroll, which consists of $305.6 million [61]. The Mets, followed by the New York Yankees, with a payroll of $303 million, has a payroll which is three times more than the six franchises in baseball and nearly $245 million more than the Oakland Athletics [61].

Figure 20 illustrates the trends in average player salaries in Major League Baseball from the mid-1980s to the 2010s. It compares the salaries in the American League (AL), represented in blue, with those in the National League (NL), represented in green.

Over the span of this period, a general trend of increasing salaries is evident in both leagues. Starting from the mid-1980s, player salaries were relatively low and grew modestly through the early 1990s. However, the mid-1990s to 2000 marked a period of significant salary escalation. This rapid increase could be attributed to changes in the economic landscape of baseball, such as lucrative television deals, increased sponsorships, or impactful labor agreements.

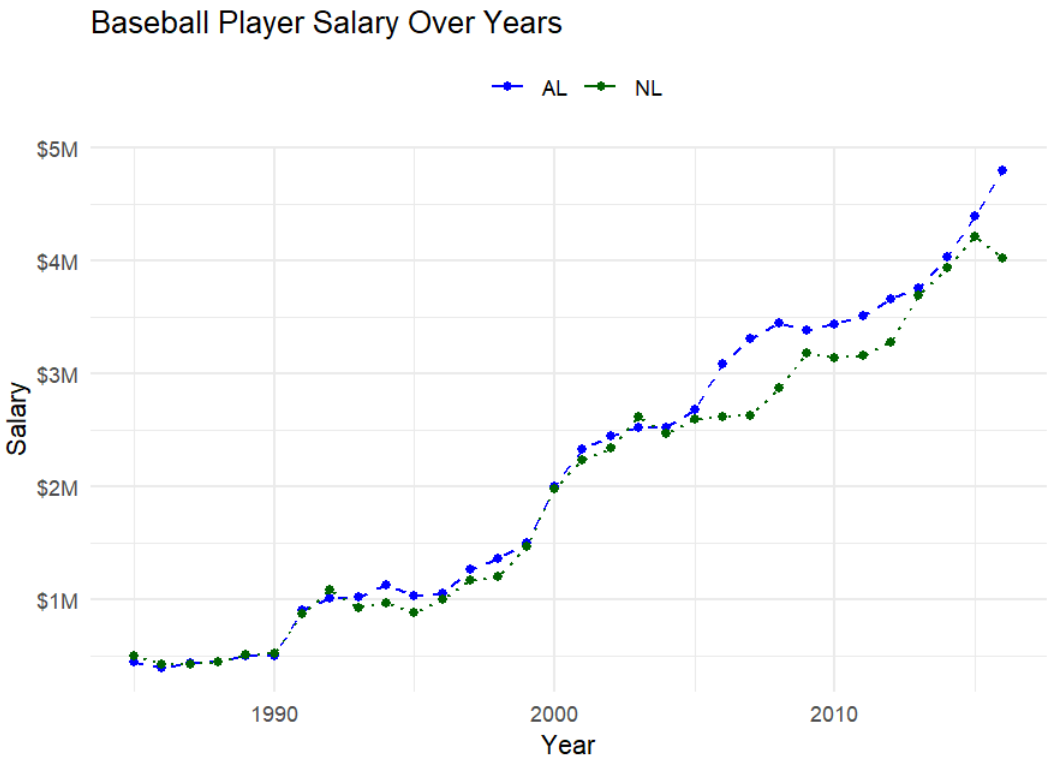
Going to the early 2000s, the trend of rising salaries continues, though with some variability. The American League consistently maintains higher average salaries compared to the National League throughout this period. Around the mid-2000s, there is a noticeable dip in salaries, particularly in the NL, which suggests temporary economic adjustments or fluctuations within the league.

From the late 2000s into the 2010s, salaries in both leagues resume their upward tra-

jectory, reaching new heights. The AL, in particular, shows a more pronounced increase, with average salaries peaking just above $4.5 million by the early 2010s, while the NL peaks slightly below $4 million. This consistent rise in the AL salaries compared to the NL highlights potential differences in market size, revenue generation, and financial strategies between the leagues.

Overall, the chart underscores a huge increase in player salaries over the decades, reflecting broader economic shifts within MLB. The significant salary growth, especially in the AL, suggests evolving dynamics and the financial health of the league. This visualization provides a clear picture of how MLB player compensation has escalated over time, driven by various economic and structural factors within the sport.

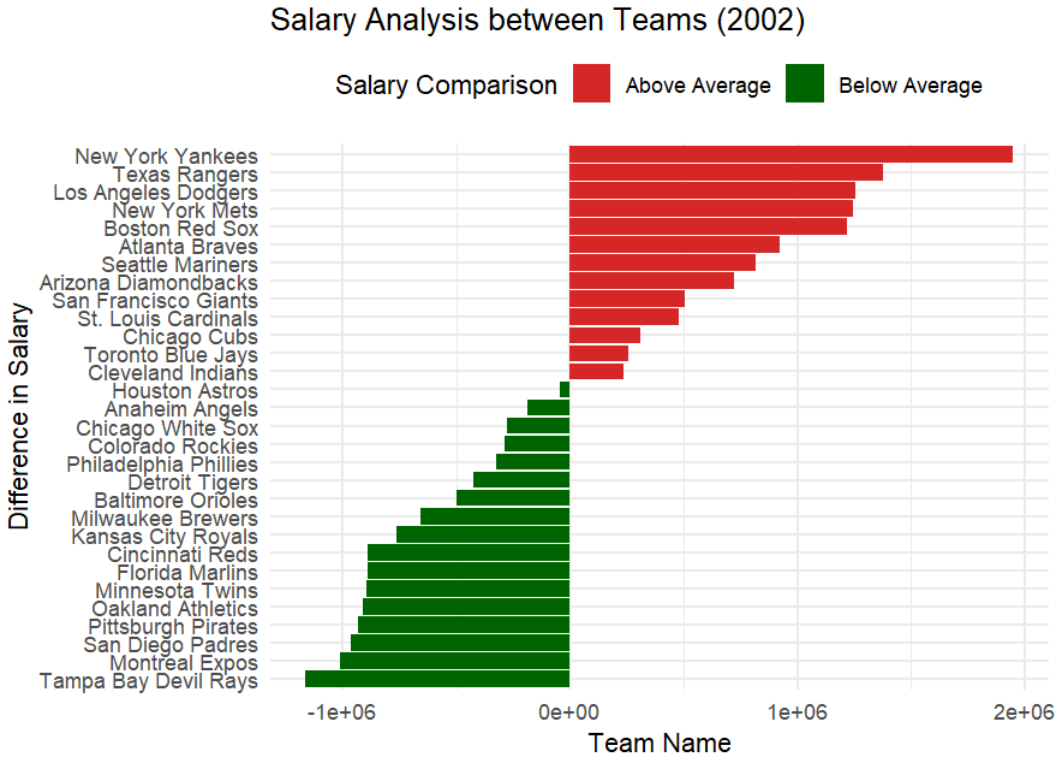
Figure 20: Salary Analysis between Teams in 2002



Since this work is focus on the *Moneyball* theory and obviously on the Oakland Athletics, it is useful to show how the salaries were distributed in 2002, when the Athletics had cut three of the best players in their roster. Figure 21 shows, for each team, if the difference in salary, which is calculated as the difference between the mean of the salaries for each team and the average in the Leaugue, is above or below the League's average.

The red bars represent teams which in 2002 had, in average, a higher payroll amount with respect to the mean calculated on all the League. Instead, the green bars represent teams with a mean of payroll amount lower than the League's average. As it has been described in the first part of this work, one can notice the difference between the New York Yankees, on the top of the graph with the larger difference, and the Oakland Athletics, fifth last bar in the green part. This is another point in favor for *Moneyball*.

Figure 21: Salary Analysis between Teams in 2002



III.2 Z-score on Salary and Multiple Regression

In this section, two main concepts are treated. The first one is a representation of a specific model selection in order to construct the best multiple regression for salaries. Instead, the second one refers to the concept of player replacement. The Athletics' replacement of the three main players due to economic reasons will be examined.

III.2.1 Z-score on Salaries

Small-market teams frequently express concerns about the disadvantages they face when competing against large-market teams like the New York Yankees and Los Angeles Dodgers. In this analysis, the extent to which team salaries impact success in Major League Baseball (MLB) is investigated. Success is measured using two key metrics: the winning percentage of a team and whether they won their division [13].

To ensure a fair comparison across different seasons and account for salary inflation over time, the standard score (z-score) of each team's salary for each season is calculated. This approach normalizes salaries, allowing to effectively compare the financial capabilities and successes of teams from different eras, thereby providing a clearer understanding of the relationship between team spending and on-field performance [13].

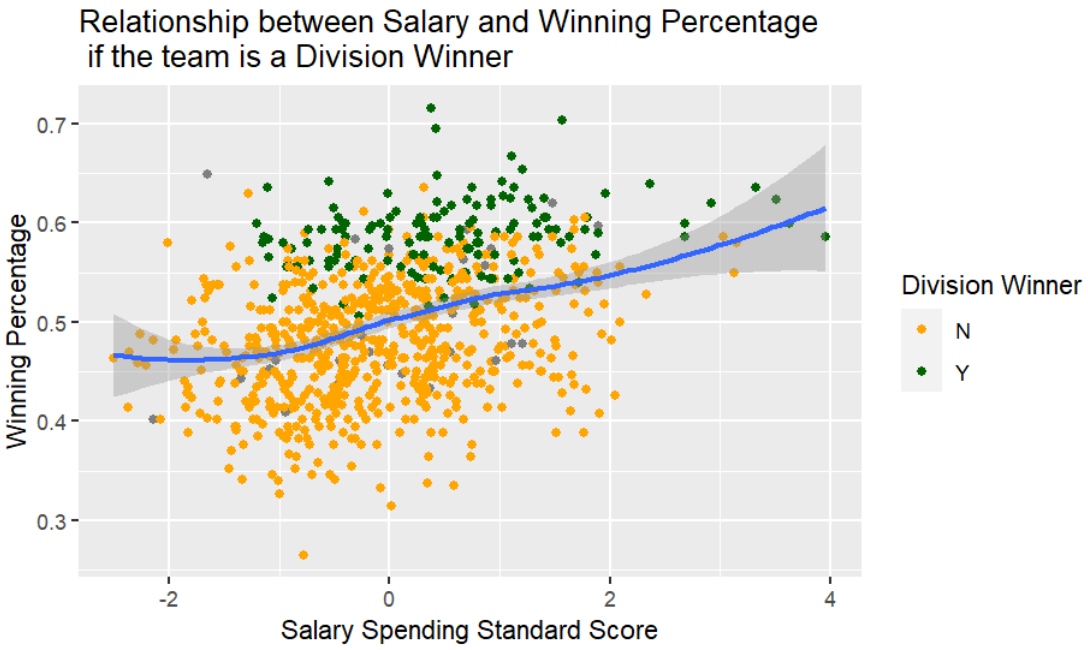
The datasets used in this kind of analysis are inside Lahman package in R and they are respectively: Batting, Pitching, Salary and Teams. After data manipulation, each dataset was merged with the Salary dataset.

Then, it was computed the salary ratio for each team. This ratio relates to the salary deriving from batters and the one deriving from pitching. After, the Z-score was computed. In statistics, the Z-score refers to a measure that describes a value's position relative to the mean of a group of values. This calculation was useful because it permitted to normalize salaries and to compare them season by season.

Before starting, it is useful to look at Figure 22. The scatter plot provided offers an examination of the relationship between team salary spending and winning percentage in MLB, while also highlighting whether teams became division winners. Each dot on the plot represents a team's performance in a given season. The color of the dots signifies whether the team was a division winner (green) or not (orange). The difference in color allows for a quick visual comparison of the spending habits between teams that won their division and those that did not. The scatter plot describes a positive correlation between salary spending and Winning percentage. So, it means that as spending on salary increases, then also the winning percentage increases as well. Furthermore, there is a larger possibility to become a Division Winner. If, for example, one looks to the right, he can note that teams with a standard score larger than 2.5 are becoming, most of the times,

Division Winner.

Figure 22: Relationship between Salary and Winning Percentage if the team is a Division Winner



However, there were teams with high spending which were not Division Winner. In order to distinguish teams that have achieved good results by spending a lot and teams that have achieved excellent results having a lower level of expenditure, it is advisable to divide the teams into two groups [13]. Teams were divided on the base of their Z-score. Teams with a Z-score larger than 2 were called "Big Spenders", whereas teams with a negative Z-score were called "Overachievers".

The results coming from this separation is shown in Figure 23. It offers a visual analysis of the performance of MLB teams over a specified period. The plot shows the comparison between the average number of wins per team, accompanied by a color-coded Z score, which measures each team's performance relative to the league average.

In this chart, the teams are listed in descending order of their average wins, from the New York Yankees at the top to the Kansas City Royals at the bottom. The X-axis quantifies the average wins, spanning from around 65 to 95 wins. This is useful to provide a clear understanding of each team's success.

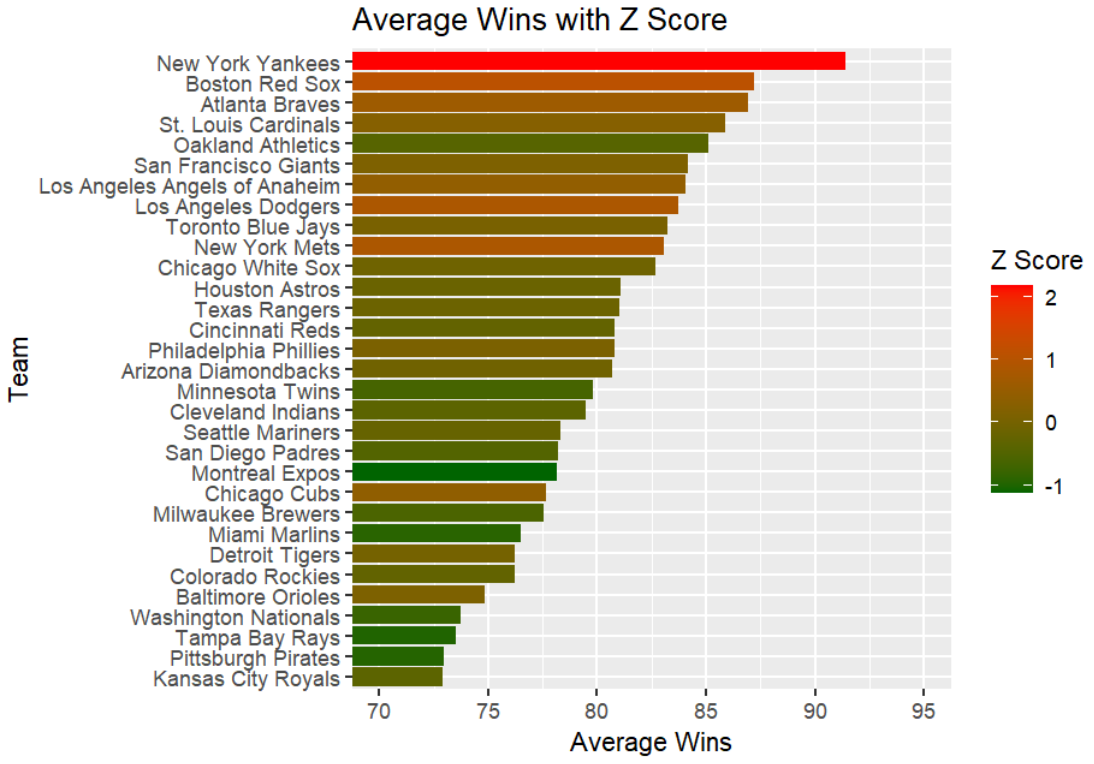
The Z score is represented by a color gradient where red indicates a very high Z score,

showing teams that perform significantly above average, and green indicates a low Z score, representing teams that perform below average. Green and brown shades, situated around zero on the Z score scale, denote teams performing near the league average.

Looking at the chart, the New York Yankees were the team with the highest average wins and a Z-score exceeding 2, marked in red. This highlights their dominance and consistent performance well above the league average. Following closely are the Boston Red Sox, Atlanta Braves and St. Louis Cardinals, all exhibit high average wins with Z-scores around or above 1, signified by darker orange shades. These teams have also been strong performers, regularly achieving success. On the other end of the spectrum, the Kansas City Royals, Pittsburgh Pirates and Tampa Bay Rays are at the lower end of the average wins scale, with Z scores nearing -1 or lower, highlighted in green.

Special attention must be payed to the situation of the Oakland Athletics. In fact, the chart displays that they are the fifth team in order of average wins, with a bar that is in a gradient of green. This confirms again the importance of the approach used by Beane in order to fight the masterclass of spending in MLB.

Figure 23: Average Wins with Z-Score in Salary



III.2.2 Multiple Regression and Model Selection

From the same merged dataset of Subsection 3.2.1, it was taken only data starting from 1996. Then, some other variables were added to the dataset, such as "OBP", "SLG" and "BA".

The aim of this section is to use model selection techniques. In fact, starting from a complete model, one wants to do a backward stepwise model selection to improve its model. In model selection the idea is to find the smallest set of predictors which provide an adequate description of the data.

In this case, two different model selection techniques were used. The first one is the so called Akaike information criterion (AIC), which can be explained by the following formula:

$$\text{AIC} = -2\ell(y,\hat{y}) + 2(p+1)$$

The second one is called Bayesian information criterion (BIC), described by this formula:

$$\text{BIC} = -2\ell(y,\hat{y}) + (p+1)\log(n)$$

Both of them are measuring the goodness of the model, using penalty functions to favour smaller ones.

The starting regression, considered as the complete model, was a multiple linear regression with the logarithm of salaries as response variable and all the numerical variables inside the dataset as predictors. They can be synthesized in this way: "OBP", "SLG", "R", "BA", "AB", "H", "HR", "BB", "SO", "SB", "CS", "HBP", "SF", "RA", "ER", "ERA", "CG", "SHO", "SV", "HA", "HRA", "BBA", "SOA", "E", "DP", "FP" [5]. From this regression, the results show that only few of this multiple variables are significant for the logarithm of the salary.

Then, firstly it was used the AIC algorithm. The results are stored in Table 22. Variables include on-base percentage (OBP), slugging percentage (SLG), batting average (BA), at-bats (AB), home runs (HR), walks (BB), strikeouts (SO), stolen bases (SB), caught stealing (CS), hit by pitches (HBP), sacrifice flies (SF), runs allowed (RA), earned runs (ER), complete games (CG), home runs allowed (HRA), batters faced (BBA), and strikeouts for opposing batters (SOA).

[5] All the variables' names can be founded in the Glossary at the end of the work.

Table 22: Regression results deriving from AIC

	Dependent variable:
	log_salary
OBP	-309.842^{***}
	(71.793)
SLG	-14.876^{***}
	(3.795)
BA	300.564^{***}
	(64.302)
AB	-0.004^{***}
	(0.001)
HR	0.010^{***}
	(0.002)
BB	0.034^{***}
	(0.008)
SO	-0.001^{***}
	(0.0002)
SB	0.001^{**}
	(0.001)
CS	-0.008^{***}
	(0.002)
HBP	0.031^{***}
	(0.008)
SF	-0.022^{***}
	(0.005)
RA	-0.004^{***}
	(0.001)
ER	0.005^{***}
	(0.002)
CG	-0.014^{**}
	(0.006)
HRA	-0.003^{***}
	(0.001)
BBA	-0.002^{***}
	(0.0003)
SOA	0.001^{***}
	(0.0002)
Constant	49.711^{***}
	(6.980)
Observations	390
R^2	0.415
Adjusted R^2	0.388
Residual Std. Error	0.345 (df = 372)
F Statistic	15.536^{***} (df = 17; 372)
	*p<0.1; **p<0.05; ***p<0.01

Key findings show significant positive relationships between salary and several met-

rics such as "OBP", "SLG", "BA", "HR", "BB", "HBP" and "SF", while others like "AB", "SO", "CS", "RA", "ER", "CG", "HRA" and "BBA" show negative relationships. The statistical significance is strong for many of these variables, as indicated by the significance levels marked with one, two, or three asterisks. The model represents an adjusted R^2 of 0.388.

Instead, the results using the BIC stepwise algorithm are summarized in Table 23. Among the metrics analyzed, "OBP" has a negative coefficient, while slugging percentage "SLG" and batting average "BA" exhibit positive coefficients. The analysis also shows the impact of "AB", "HR" and "BB". Specific variables such as "SO", "CS", "RA", "HRA" and "BBA" have negative coefficients, suggesting an adverse effect on salary. In contrast, metrics like hit by pitches "HBP" and earned runs "ER" have positive coefficients, indicating a positive relationship with salary. All the coefficients are statistically significant. The adjusted R^2 is equal to 0.376. The model is not strong in explain the variability of the logarithm of player salaries.

Table 23: Regression results deriving from BIC

	Dependent variable:
	log_salary
OBP	-311.391^{***}
	(72.468)
SLG	-16.307^{***}
	(3.801)
BA	304.637^{***}
	(64.914)
AB	-0.004^{***}
	(0.001)
HR	0.010^{***}
	(0.002)
BB	0.034^{***}
	(0.008)
SO	-0.001^{***}
	(0.0002)
CS	-0.006^{***}
	(0.002)
HBP	0.031^{***}
	(0.008)
SF	-0.022^{***}
	(0.005)
RA	-0.005^{***}
	(0.001)
ER	0.005^{***}
	(0.002)
HRA	-0.003^{***}
	(0.001)
BBA	-0.001^{***}
	(0.0003)
SOA	0.001^{***}
	(0.0002)
Constant	48.942^{***}
	(7.044)
Observations	390
R^2	0.400
Adjusted R^2	0.376
Residual Std. Error	0.349 (df = 374)
F Statistic	16.622^{***} (df = 15; 374)
	*p<0.1; **p<0.05; ***p<0.01

III.3 Player Replacement

In this last section the aim is to analyze the player replacement, id est which players could replace the three main guys delivered by the Athletics in 2001. In fact, in 2001 the Athletics have decided to trade three main players:

1. baseman Jason Giambi traded to the New York Yankees. He has been MVP (Most Valuable Player) in 2000.

2. outfielder Johnny Damon traded to Boston Red Sox.

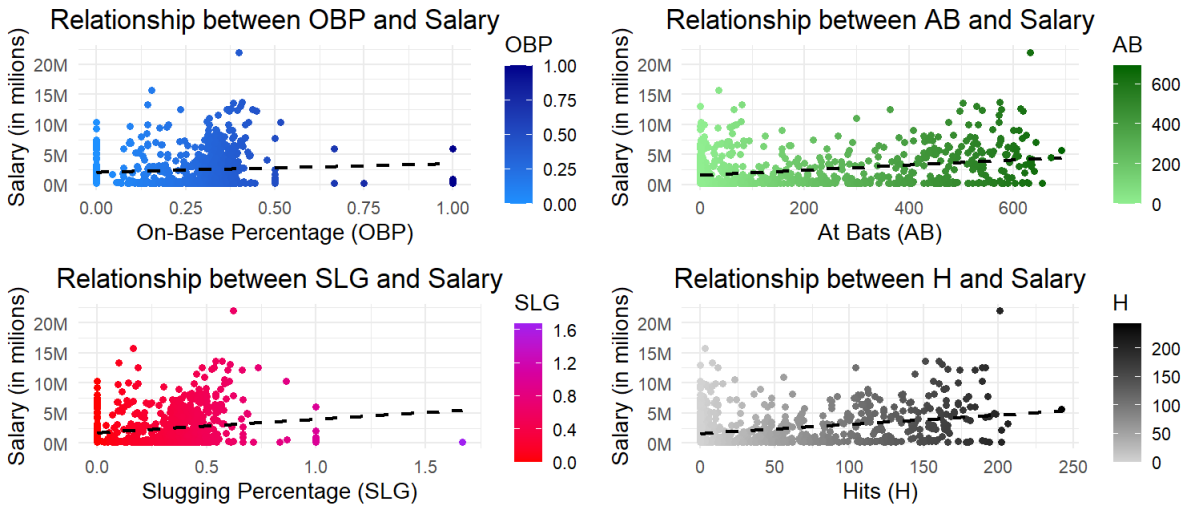
3. infielder Olmedo Sáenz traded to Los Angeles Dodgers.

The datasets used for this kind of analysis were "Salary" and "Batting" from Lahman database. First of all, it was necessary to merge the two datasets in order to add the column of salaries for each players. Because of their absences, "BA", "SLG", "OBP", "1XB" (singles) were created and added to the merged dataset. Data starting from 1985 were used. Then, a subset was created using the tag names of the three players lost by the Athletics ('giambja01','damonjo01','saenzol01'). Since these players were lost during the 2001 off-season, then data taking from 2001 were used. Then, another data frame was created taking into account only available players. From them, three players must be chosen to replace the lost player of the Athletics.

Before this, Figure 24 contains four scatter plots, which are used to visually analyze the relationships between two numeric variables. Each scatter plot has salary (in millions) on the vertical axis and a different baseball statistic on the horizontal axis. The four baseball statistics are on-base percentage (OBP), at bats (AB), slugging percentage (SLG), and hits (H). In general, there is a positive correlation between salary and all four baseball statistics. This means that as a player's OBP, AB, SLG, or H increases, their salary also tends to increase.

It is important to note that correlation does not necessarily equal causation. Just because a player has a high OBP, AB, SLG, or H doesn't necessarily mean they will have a high salary. There could be other factors that influence a player's salary, such as their position, experience, and other intangibles

Figure 24: Relationship between Salary and other variables



Then, three players whose combined salaries are not exceeding 15 million dollars must be selected from the available players. Furthermore, are chosen players with a combined number of AB equal or greater than the lost players. Moreover, their SLG must be larger than 0.5 and their hits must be at least larger then the its average.

Following the R algorithm [54], the results are reported in Table 24. The three players the Oakland Athletics could have hired are Todd Helton, Lance Berkman and Luis Gonzalez.

Table 24: Replacement players in 2001

Player ID	**Name**	**OBP**	**AB**	**SLG**	**H**	**Salary**
heltoto01	Todd Helton	0.4316	587	0.6848	197	4,950,000
berkmla01	Lance Berkman	0.4302	577	0.6204	191	305,000
gonzalu01	Luis Gonzalez	0.4286	609	0.6880	198	4,833,333

Conclusion

This study has provided that Oakland Athletics proved that they could take competitive advantage from an inefficient market using Sabermetrics. Advanced statistics and Sabermetrics were not utilized in the early 2000s like they are today, therefore, the Athletics have capitalized on the opportunity to evaluate undervalued players to complete with large market organizations like the New York Yankees. These concepts are all confirmed by the study and it is important to underlying how *Moneyball* has had the role of game-changer in the history of baseball. In fact, from 2010s all 32 Major League Baseball teams have started to use the same tools.

However, *Moneyball* did not just stop at baseball.The lessons learned from baseball's analytics revolution have demonstrated the power of data to drive innovation and improve outcomes across diverse fields. In recent years this approach is also being used in other sports that are characterized by a high economic environment: National Basketball Association (NBA) and in football. For instance, in Italy, the famous football society A.C. Milan has started to use this approach. With their new management, they are trying to discover undervalued players in order to beat the market appropriating their sporting performances before others.

Analytics have reshaped the strategic landscape of baseball, and teams now use data-driven approaches to optimize every aspect of the game, from lineup construction to defensive shifts and bullpen management. Managers can use statistical models to simulate game scenarios and make more informed decisions during matches. The game is now more precise and analytically driven, with decisions based on probabilities and historical data rather than gut feelings or tradition.

In the future, the role of analytics in baseball is expected to increase. Advancements in technology, like machine learning and artificial intelligence, have the potential to uncover deeper insights and enhance predictive capabilities. With the continue innovation of these tools, the strategic complexity and enthusiasm of baseball will continue to evolve.

Glossary

Batting

AB = At-bats

BA = Batting Average = H / AB

BB = Bases on balls (Walks)

BBA = Walks allowed

X1B = Singles

X2B = Doubles

X3B = Triples

CG = Complete games

CS = Caught stealing

DP = Double Plays

FP = Fielding percentage

G = Games played

H = Hits

HA = Hits Allowed

HBP = Batters hit by pitch

HRA = Homeruns allowed

HP = Hit by pitch

HR = Home runs

IBB = Intentional bases on balls

OBP = On-base percentage = (H + BB + HP) / (AB + BB + HP)

OPS = On-base plus slugging = OBP + SLG

R = Runs scored

RA = Opponent Runs scored

RBI = Runs batted in

SB = Stolen bases

SF = Sacrifice flies

SH = Sacrifice hits (Bunts)

SHO = Shutouts

SLG = Slugging percentage = TB / AB

SO = Strikeouts by batters

SOA = Strikeouts by pitchers

TB = Total bases = 1(1B) + 2(2B) + 3(3B) + 4(HR)

Pitching

BB = Bases on balls (allowed)

E = Errors

ER = Earned Runs allowed

ERA = Earned run average = 9 × ER / IP

IP = Innings pitched

IPouts = Outs Pitched (IP × 3)

K = Strikeouts

L = Losses

SV = Saves

W = Wins

W_{PCT} = Winning percentage = W / (W + L)

References

[1] Abdul Qureshi. *Using Logistic Regression Analysis to Predict Baseball World Series Champion using R*. [Online; accessed 08-June-2024]. URL: https://medium.com/@aqureshi/using-logistic-regression-analysis-to-predict-baseball-world-series-champion-using-r-3d43e5a03068.

[2] Jay Barney. "Firm resources and sustained competitive advantage". In: *Journal of management* 17.1 (1991), pp. 99–120.

[3] Baseball Reference. *1998 New York Yankees Statistics*. [Online; accessed 07-April-2024]. 2024. URL: https://www.baseball-reference.com/teams/NYY/1998.shtml.

[4] Baseball Reference. *1998 New York Yankees Statistics*. [Online; accessed 07-April-2024]. 2024. URL: https://www.baseball-reference.com/teams/NYY/2002.shtml.

[5] Baseball Reference. *1998 Oakland Athletics Statistics*. [Online; accessed 07-April-2024]. 2024. URL: https://www.baseball-reference.com/teams/OAK/1998.shtml.

[6] Baseball Reference. *2002 Oakland Athletics Statistics*. [Online; accessed 07-April-2024]. 2024. URL: https://www.baseball-reference.com/teams/OAK/2002.shtml.

[7] Benjamin Burroughs. "Statistics and baseball fandom: Sabermetric infrastructure of expertise". In: *Games and Culture* 15.3 (2020), pp. 248–265.

[8] Francis T Cullen, Andrew J Myer, and Edward J Latessa. "Eight lessons from Moneyball: The high cost of ignoring evidence-based corrections". In: *Victims and Offenders* 4.2 (2009), pp. 197–213.

[9] Barry De Ville. "Decision trees". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (2013), pp. 448–455.

[10] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[11] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)”. In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[13] Gary Garrison. *Are Big Spenders Big Winners?* [Online; accessed 08-June-2024]. URL: `https://www.kaggle.com/code/garrison/are-big-spenders-big-winners/report`.

[14] Bill Gerrard. “Is the Moneyball approach transferable to complex invasion team sports?” In: *International Journal of Sport Finance* 2.4 (2007), p. 214.

[15] Scott Gray. *The mind of Bill James: How a complete outsider changed baseball*. Crown, 2006.

[16] Joel B Greenhouse, Judith A Bromberg, and Davida Fromm. “An introduction to logistic regression with an application to the analysis of language recovery following a stroke”. In: *Journal of communication disorders* 28.3 (1995), pp. 229–246.

[17] John Hagan. “Labelling and deviance: a case study in the “sociology of the interesting””. In: *Social Problems* 20.4 (1973), pp. 447–458.

[18] Jahn K Hakes and Raymond D Sauer. “An economic evaluation of the Moneyball hypothesis”. In: *Journal of Economic Perspectives* 20.3 (2006), pp. 173–185.

[19] Bill James. *The new Bill James historical baseball abstract*. Simon and Schuster, 2010.

[20] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. “Linear regression”. In: *An introduction to statistical learning: With applications in python*. Springer, 2023, pp. 69–134.

[21] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[22] MBL. *On-Base Percentage (OBP)*. [Online; accessed 30-October-2023]. 2023. URL: `https://www.mlb.com/glossary/standard-stats/on-base-percentage`.

[23] MBL. *Slugging Percentage (SLG)*. [Online; accessed 30-October-2023]. 2023. URL: https://www.mlb.com/glossary/standard-stats/slugging-percentage.

[24] MBL. *At-Bat (AB)*. [Online; accessed 26-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/at-bat.

[25] MBL. *Caught Stealing (CS)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/caught-stealing.

[26] MBL. *Double (2B)*. [Online; accessed 29-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/double.

[27] MBL. *Earned Run Average (ERA)*. [Online; accessed 31-March-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/earned-run-average.

[28] MBL. *Ground Into Double Play (GIDP)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/ground-into-double-play.

[29] MBL. *Hit (H)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/hit.

[30] MBL. *Hit-by-pitch (HBP)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/hit-by-pitch.

[31] MBL. *Home Run (HR)*. [Online; accessed 29-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/home-run.

[32] MBL. *Innings Pitched (IP)*. [Online; accessed 31-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/innings-pitched.

[33] MBL. *Intentional Walks (IBB)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/intentional-walk.

[34] MBL. *Official New York Yankees Website*. [Online; accessed 04-April-2024]. 2024. URL: https://www.mlb.com/yankees.

[35] MBL. *Official Oakland Athletics Website*. [Online; accessed 04-April-2024]. 2024. URL: https://www.mlb.com/athletics.

[36] MBL. *Sacrifice Bunt (SH)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/sacrifice-bunt.

[37] MBL. *Sacrifice Fly (SF)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/sacrifice-fly.

[38] MBL. *Single (1B)*. [Online; accessed 29-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/single.

[39] MBL. *Stolen Base (SB)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/stolen-base.

[40] MBL. *Total Bases (TB)*. [Online; accessed 25-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/total-bases.

[41] MBL. *Triple (3B)*. [Online; accessed 29-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/triple.

[42] MBL. *Walks And Hits Per Inning Pitched (WHIP)*. [Online; accessed 31-January-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/walks-and-hits-per-inning-pitched.

[43] MBL. *Wins Above Replacement (WAR)*. [Online; accessed 31-March-2024]. 2024. URL: https://www.mlb.com/glossary/standard-stats/earned-run-average.

[44] MLB. *Wild Card History*. [Online; accessed 08-June-2024]. 2024. URL: https://www.mlb.com/postseason/history/wild-card.

[45] Richard Makadok. "Toward a synthesis of the resource-based and dynamic-capability views of rent creation". In: *Strategic management journal* 22.5 (2001), pp. 387–401.

[46] Max Marchi and Jim Albert. *Analyzing baseball data with R*. CRC Press, 2013.

[47] Medium. *Linear Regression In Moneyball: Explained*. [Online; accessed 12-November-2023]. 2023. URL: https://medium.com/@krishnapiryakm/linear-regression-in-moneyball-explained-b27554fccc8c#:~:text=Linear%20regression%20is%20a%20statistical,slugging%20percentage)%20and%20their%20salary..

[48] Merriam-Webster. *On base. In Merriam-Webster.com dictionary.* [Online; accessed 30-October-2023]. 2023. URL: https://www.merriam-webster.com/dictionary/on%20base.

[49] Merriam-Webster. *Definition: Inning*. [Online; accessed 31-January-2024]. 2024. URL: https://www.merriam-webster.com/dictionary/inning.

[50] Francis Sahngun Nahm. "Receiver operating characteristic curve: overview and practical use for clinicians". In: *Korean journal of anesthesiology* 75.1 (2022), p. 25.

[51] Laura Poppo and Keith Weigelt. "A test of the resource-based model using baseball free agents". In: *Journal of Economics & Management Strategy* 9.4 (2000), pp. 585–614.

[52] Michael E Porter et al. "What is strategy?" In: (1996).

[53] RStudio. *Predicting Playoff Teams*. [Online; accessed 08-June-2024]. URL: https://rpubs.com/mevanoff24/107772.

[54] Saurabh Srivastava. *Baseball Replacement Players Prediction Analysis*. [Online; accessed 08-June-2024]. URL: https://www.kaggle.com/code/saurabh3089/baseball-replacement-players-prediction-analysis#Replacement-Players.

[55] Alan Schwarz. *The numbers game: Baseball's lifelong fascination with statistics*. Macmillan, 2004.

[56] Sean Lahman. *Sean Lahman Database*. [Online; accessed 08-June-2024]. 2024. URL: http://www.seanlahman.com/.

[57] Juliet Popper Shaffer. "The Gauss—Markov theorem and random regressors". In: *The American Statistician* 45.4 (1991), pp. 269–273.

[58] Society for American Baseball Research (SABR). *Baseball's Major Salary Milestones*. [Online; accessed 08-June-2024]. 2024. URL: https://sabr.org/journal/article/baseballs-major-salary-milestones/.

[59] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information processing & management* 45.4 (2009), pp. 427–437.

[60] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. “Linear regression”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.3 (2012), pp. 275–294.

[61] USA Today. *MLB payrolls 2024: Full list of every baseball team from highest to lowest*. [Online; accessed 08-June-2024]. 2024. URL: https://eu.usatoday.com/story/sports/mlb/2024/04/03/mlb-team-payrolls-2024-highest-lowest-mets/73139425007/.

[62] Anthony G Vito and Gennaro F Vito. “Lessons for policing from Moneyball: The views of police managers–a research note”. In: *American journal of criminal justice* 38 (2013), pp. 236–244.

[63] Ehren Wassermann, Daniel R Czech, Matthew J Wilson, A Barry Joyner, et al. “An examination of the Moneyball theory: a baseball statistical analysis.” In: *The Sport Journal* 8.1 (2005).

[64] Wikipedia contributors. *Base on balls — Wikipedia, The Free Encyclopedia*. [Online; accessed 12-November-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Base_on_balls&oldid=1178990251.

[65] Wikipedia contributors. *Batted ball — Wikipedia, The Free Encyclopedia*. [Online; accessed 29-January-2024]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Batted_ball&oldid=1177622616.

[66] Wikipedia contributors. *Moneyball (film) — Wikipedia, The Free Encyclopedia*. [Online; accessed 30-October-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Moneyball_(film)&oldid=1181201573.

[67] Wikipedia contributors. *World Series — Wikipedia, The Free Encyclopedia*. [Online; accessed 1-November-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=World_Series&oldid=1182901820.

[68] Wikipedia. *Boston Red Sox — Wikipedia, L'enciclopedia libera*. [Online; in data 1-novembre-2023]. 2023. URL: http://it.wikipedia.org/w/index.php?title=Boston_Red_Sox&oldid=135415572.

[69] Richard Wolfe, Kathy Babiak, Kim Cameron, Robert E Quinn, Dennis L Smart, James R Terborg, Patrick M Wright, et al. "Moneyball: A business perspective". In: *International Journal of Sport Finance* 2.4 (2007), pp. 249–262.

[70] Richard Wolfe, Patrick M Wright, and Dennis L Smart. "Radical HRM innovation and competitive advantage: The Moneyball story". In: *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in Alliance with the Society of Human Resources Management* 45.1 (2006), pp. 111–145.

I Appendix A

Since the estimation of the coefficients is crucial in the regression analysis, it should be useful to explain how effectively significant they are. In this sense, two main question could be done:

1. Is at least one of the predictors useful in predicting the response?

2. Do all the predictors help to explain the response variable?

The first question recalls the the concept of the Fisher Test, or F-Test. To better explain this concept, let consider two nested models. For nested, it means that there is a large model with a smaller model inside of it. On one side there is a so-called complete model with predicted values that takes the form: $\hat{y_i}^{(c)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$. On the other side, there is an empty or null model, which has the following form: $\hat{y_i}^{(r)} = \beta_0 = \bar{y}$.

Ideally, a comparison between the residuals of the two models has to be made. So, let define the residual sum of squares for both the complete and the reduced model:

$$\text{RSS}_C = \sum_{i=1}^{n} \left(y_i - \hat{y_i}^{(c)}\right)^2 \qquad \text{RSS}_R = \sum_{i=1}^{n} \left(y_i - \hat{y_i}^{(r)}\right)^2 = SS_Y$$

Obviously, the complete model is wanted to be more precise than the reduced model since there is a larger number of degrees of freedom:

$$\text{RSS}_R > \text{RSS}_C$$

Then, the hypothesis test method is used in order to test the validity of the reduced model. If the reduced model is accepted, it means that there is no significant predictor in the model. It will be a model with only the intercept. Let define the two hypothesis:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H1 : \exists \beta_j \neq 0$$

The test statistic is:

$$F = \frac{(RSS_R - RSS_C)/p}{RSS_C/(n-p-1)}$$

The test statistic is based exactly on the difference between the RSS of the empty and of the complete model. If this difference is sufficiently small than H_0 (there are no significant predictors) is accepted. If the errors ε_i are iid normally distributed, then F has a known

distribution under H_0, named as "*the Fisher's distribution with "p" degrees of freedom in the numerator and "$(n-p-1)$" degrees of freedom in the denominator*". The test is a one-tailed right test.

The F-test can be used in general for testing the nullity of a set of "q" coefficients. In such a case let consider two nested models:

- The complete model with predicted values:

$$\hat{y}_i^{(c)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- A reduced model obtained by deleting "q" predictors, with predicted values:

$$\hat{y}_i^{(r)} = \beta_0 + \beta_{q+1} x_{i,q+1} + \beta_{q+2} x_{i,q+2} + \cdots + \beta_p x_{ip}$$

To test the validity of the reduced model, let define:

$$H_0 : \beta_1 = \cdots = \beta_q = 0 \qquad H_1 : \exists \beta_j \neq 0$$

The test statistic is:

$$F = \frac{(RSS_R - RSS_C)/q}{RSS_C/(n-p-1)}$$

with Fisher distribution with "q" degrees of freedom in the numerator and "$(n-p-1)$" degrees of freedom in the denominator.

In order to answer to the second question, it is useful to introduce the t-test. In the multiple regression, the coefficient β_j can be read off as:

$$\beta_j = \frac{\partial Y}{\partial X_j}$$

so β_j accounts for the variation of Y with respect to X_j, all other quantities being fixed. The parameter β_j is estimated with B_j and both the expected value and the variance of the estimator B_j are known, so we a standardization and a definition of the test statistic can be applied.

For a given $j = 1, \ldots, p$, the hypotheses on the single parameter can be defined:

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0$$

The test statistic is:

$$T_j = \frac{B_j}{\sqrt{COV(B)_{jj}}}$$

Under H_0 the statistic T_j has a Student's t distribution with "$(n-p-1)$" degrees of freedom. The test is (usually) a two-tailed test.

However, another coefficient is useful in regression analysis. In fact, in order to get how well the model fits the data, it is used the R^2. It is defined as:

$$R^2 = 1 - \frac{RSS_C}{SS_Y} \qquad 0 \leq R^2 \leq 1$$

where the term RSS_C is also called error variance and the term SS_Y is called total variance. There are also other possible choices:

- The adjusted R^2, which takes into accounts the number of predictors:

$$R^2_{adj} = 1 - \frac{RSS_C/(n-p-1)}{SS_Y/(n-1)}$$

- The residual standard error RSE:

$$RSE = \sqrt{\frac{1}{(n-p-1)} RSS_C}$$

This section is set to explain and proof the way by which the coefficient of the the simple and multiple linear regression may be found.

I.1 Simple Linear Regression

Let's start by defining the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Then, some assumptions have to be apply. They are listed below:

1. Linearity: the parameters are constant over time

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 x$$

2. Variability:

$$\exists i \neq j \quad \text{s.t.} \quad x_j \neq x_i$$

This will imply two other concepts:

(a) Positive sample variance: $\frac{1}{(n-1)} \sum \left(x_i - \bar{x}\right)^2 > 0$

(b) No multicollinearity: a variable cannot be a linear combination of another one.

3. Sample Randomness:

$$\forall i \neq j : COV\big(y_i, y_j | x\big) = 0$$

It means that there is no correlation between individuals.

4. Esogeneity: $\mathbb{E} = 0$. There no correlation between the variables and the error. Consequently:

$$\mathbb{E}\big(y|x\big) = \beta_0 + \beta_1 x$$

Then, the first order condition (F.O.C.) can be computed:

- $\beta_0 = \mathbb{E}\big(y|x\big)$
- $\beta_1 = \frac{\partial \mathbb{E}(y|x)}{\partial x}$

5. Homoscedasticity:

$$Var(\varepsilon|x) = \sigma^2 = \mathbb{E}(\varepsilon^2|x)$$

The variance is constant among the individuals.

6. Random errors:

$$\varepsilon|x \sim \mathscr{N}(0; \sigma^2)$$

which implies that

$$y|x \sim \mathscr{N}(\beta_0 + \beta_1 x, \sigma^2)$$

Once the assumption are defined, the objective function must be found. It is the minimization of the residuals sum of squares:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} e_i^2 \quad \text{with} \quad e_i = y_i - \hat{y}_i$$

Since $\hat{y}_i = \beta_0 + \beta_1 x_i$ then a substitution can be made:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

Let call $\hat{\beta}$ the vector which minimize the objective function $D(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$ then let compute the F.O.Cs:

$$\text{FOC 1} : \frac{\partial D(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

Multiplying both parts by $1/n$, it can be obtained:

$$-2 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

Since $\frac{1}{n} \cdot y_i = \bar{y}$ (and the same speech can be applied to x_i), the equation above becomes:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \quad \Longleftrightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Moving to the second foc:

$$\text{FOC 2} : \frac{\partial D(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

It gives the following equation:

$$-2 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) x_i = 0$$

Doing some computations, it will be obtained:

$$\frac{1}{n} \cdot \sum (y_i x_i) = \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x} \cdot x_i \quad \text{where} \quad \hat{\beta}_0 \bar{x} = \frac{1}{n} \cdot \sum (x_i) \cdot \beta_0$$

Then

$$\frac{1}{n} \cdot \sum (y_i x_i) = \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \cdot \frac{1}{n} \sum x_i^2$$

Now, let subtract FOC 2 $-$ FOC 1 $\cdot \bar{x}$:

$$\frac{1}{n} \cdot \sum (y_i x_i) - \bar{y} \cdot \bar{x} = \hat{\beta}_1 \big(\frac{1}{n} \sum x_i^2 - \bar{x}^2 \big)$$

$$\frac{1}{n} \cdot \sum (y_i x_i) - \frac{1}{n} \cdot \bar{y} \cdot \sum x_i = \hat{\beta}_1 \big(\frac{1}{n} \sum x_i^2 - \bar{x} \cdot \sum x_i \big)$$

Now, $\frac{1}{n}$ can be deleted and the factor x_i can be collected:

$$\sum x_i \cdot (y_i - \bar{y}) = \hat{\beta}_1 \sum x_i (x_i - \bar{x})$$

So, $\hat{\beta}_1$ will be:

$$\hat{\beta}_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\text{COV}(x, y)}{\text{Var}(x)}$$

I.2 Expected value of LS estimator

As said, the LS estimator is unbiased since its expected value is constant over time and it it equal to β.

$$\mathbb{E}(B) = \beta$$

The proof of this property can be found below.

$$\mathbb{E}((X^T X)^{-1} X^T Y) = \mathbb{E}((X^T X)^{-1} X^T (\underbrace{X\beta + \varepsilon}_{Y}))$$

$$\mathbb{E}((X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \varepsilon) = \mathbb{E}((X^T X)^{-1} X^T X\beta) + \mathbb{E}((X^T X)^{-1} X^T \varepsilon)$$

So the first part is a constant (a number) since neither the Y nor the "ε" are known. In the second term, there is ε and so it can be obtained:

$$= \beta + (X^T X)^{-1} X^T \underbrace{\mathbb{E}(\varepsilon)}_{=0} = \beta$$

but since the mean of the error is 0 the last component disappears.

Ringraziamenti

Eccoci qui! Siamo arrivati alla parte più importante della mia Tesi. I famosissimi, dispendiosissimi e mai banali ringraziamenti.

Partiamo con i ringraziamenti istituzionali. Innanzitutto vorrei ringraziare il professor Rapallo per il sostegno e l'appoggio durante la stesura di questo lavoro che, bene o male, sarà il mio primo grande progetto. Ho sempre pensato che avere un relatore che infonda tranquillità e serenità sia un aiuto in più per poter lavorare senza ulteriore dispendio di energie.

Passiamo poi alla mia famiglia. In primis, ringrazio i miei genitori Cristina e Valerio che mi hanno permesso, grazie ai loro sacrifici e al loro gigantesco amore, di ottenere tutti i risultati che ho raggiunto e sono sicuro che saranno un'ottima spalla su cui contare anche per i progetti futuri. Ringrazio, poi, mio zio Fabio e tutti i miei nonni: Vincenzo, Rita, Dino e Fernanda. Nonostante la vita vada avanti per tutti siete sempre stati un punto di riferimento sia per papà e la mamma che per me e mio fratello.

Ringrazio mio fratello Samuele, grande fonte d'ispirazione e mente brillante. La tua serietà ed il tuo modo di essere mi hanno aiutato a diventare la brava persona che sono ora. Non è possibile cancellare la distanza che, per motivi lavorativi, ci separa, ma sei sempre qui con me. Le passioni che abbiamo in comune hanno stretto il nostro rapporto, rendendolo indissolubile.

Volevo poi ringraziare Marta. Sei entrata come un ciclone all'interno della mia vita e come un ciclone hai rovesciato tutto. Dopo un periodo in cui facevo fatica a provare emozioni sei arrivata ed hai portato freschezza, divertimento, felicità e amore. Non dimenticherò ogni momento passato insieme e gli sforzi che stiamo facendo per creare qualcosa di bello insieme anche nel futuro. Sarò sempre pronto a supportarti nelle tue prossime decisioni, che saranno fondamentali per la tua carriera. Sei una persona fantastica e sono sempre convinto che la tua carta d'identità menta! Ci tengo a ringraziare anche Eugenio, Monica, Andrea e Sara che mi hanno accolto con grande gentilezza e non mi hanno mai fatto mancare niente. Non posso non citare Lilla e Nina, le tue splendide gatte che fanno parte della nostra quotidianità.

Iniziamo la cerchia degli amici ora. Ringrazio Luca, amico e fratello da una vita. La

vita ultimamente ci ha messo davanti a scelte diverse, ma ciò non può far dimenticare tutte le belle esperienze che abbiamo vissuto, i pomeriggi in due sostanzialmente a non fare niente, le prime birre, le prime cotte, i primi litigi. Sei e sarai sempre importante per me perché riesci a compensare il mio carattere rigido con una bella dose di spensieratezza.

Ringrazio inoltre Thomas, Davide, Paolo, Alberto, Tommaso, Giulia, Filippo, Matteo, Sebastiano, Federico, Giovanni e, un altro, Federico. Con ognuno di voi ho passato momenti bellissimi e sono contento di aver contribuito a creare un'amicizia bella e duratura in un piccolo paese sulla riviera ligure. Grazie ad ognuno di voi per avermi mostrato la propria parte buona dove poter attingere per migliorare giorno dopo giorno.

Siccome il tema dello sport è stato ricorrente anche nella stesura di questo lavoro, voglio ringraziare la Juvenilia Varazze, la mia attuale squadra di basket. Una seconda famiglia, composta da persone brave, disponibili e affiatate che ti fanno sentire di essere all'interno di un ambiente protetto e sicuro. Ringrazio ogni singolo elemento della dirigenza (tranne il custode del palazzetto) e della mia squadra per avermi concesso l'opportunità di poter giocare e divertirmi insieme ai miei amici. A settembre inizierà una nuova stagione, l'ennesima per tutti noi. Con determinazione e cervello possiamo toglierci tante soddisfazioni, nel caso si può sempre applicare l'approccio del Moneyball!

Ringrazio i miei ormai datati compagni di classe: Alberto, Nicolò, Matteo, Federica, Kimberli e Margherita. Ovviamente, per questioni di studio, ci siamo dovuti separare ma è sempre bello vedervi.

Passando al lato universitario, ringrazio sicuramente Alberto, mio compagno di viaggio e di studio per questi cinque anni. Spero di averti aiutato così come hai fatto te. Ringrazio Simone per il supporto e per il divertimento offerto sui treni giocando come i quindicenni a Brawl Stars all'urlo di "GA STA NIII". Ringrazio poi Giorgio aka Peerz2PeerzLovers. Sei stato fondamentale per il tuo altruismo e per i tuoi stupendi appunti che mi hanno permesso di faticare meno! Ringrazio Ben, studente di altissimo livello. Spero che vada tutto per il meglio per il tuo prossimo lavoro, te lo meriti. Ringrazio, inoltre, tutti gli altri miei compagni di corso e vi auguro di poter finire gli studi nel migliore dei modi e di poter intraprendere una carriera brillante.

Ringrazio poi i miei compagni di triennale, i cosiddetti "Disperatix". Anche con voi abbiamo intrapreso strade diverse, ma volevo ringraziarvi per l'aiuto che mi avete dato

durante tutto il terzo anno. Ringrazio, inoltre, Lorenzo per essere un fratello doriano e per la bella persona che ha dimostrato di essere.