**SI 618 Project 2 Report**

**Soobum Jun**

**Motivation**

I have been interested in baseball since childhood, so I memorized all the stats of players on my favorite baseball team. In college, after I became a captain in the college baseball club, I got more interested in the baseball strategy and stats. In particular, I found out that the professional baseball team I was cheering for had high salaries for individual players but did not perform well. At this time, I saw a movie called Moneyball. After read and watch at the Moneyball, each player's stats are vital, but I am also found out there are other hidden variables that can predict the game result with saving management cost of the team while finding out undervalued, but a high performed player. In this project, I aim to find out the hidden variable that can predict the game result.

Also, baseball is a game where data is more important than any other sport. Therefore, players having good stats are getting paid a good salary accordingly. However, there are many opinions on whether a player with a high salary does affect team victory. Accordingly, I will also explore the correlation between individual players’ salaries correlate with or affect team winning.

Question 1 : Which player have a higher batting average? Is top hitter always paid higher salary?

Question 2 : What type of variables has considerably predict the team winning?

Question 3 : Is the total salary of players in the team correlated with the winning ratio?

**Data Sources**

Lahman Dataset (<http://www.seanlahman.com/>, included in R as R package)

Lahman's data set is gathered by Sean Lahman, the famous sports reporter. This well-known Lahman data set is included in R as packages. The database contains 30 tables. It contains all the information on players’ and teams’ batting, salary, team and fielding performance, and other tables from 1871 through 2018, as recorded in the 2019 version of the database. I will use the recent data for all years for consistency of the result; then, I manipulate it depends on each question's characteristics. Likewise, I use this dataset to extract which features are essential to predict the team winning percentage, and exploring which features are considered important with a marketing and team management perspective.

\* Note that the Lahman data set is included in R as a package, you need to install packages in R using install.packages(‘Lahman’) and library(Lahman) instead, you download the whole dataset on the Lahman website.

1. Salaries : the dataframe has 26428 observation including players salary. It contains yearID, teamID, lgID, playerID, and salary.

yearID : Year

teamID : Team; a factor

lgID : League; a factor

playerID : Player ID code

salary : Salary

1. Team : teams table include 334 obesrvation with yearID, teamIDwinner, lgIDwinner, teamIDlosser, lgIdloser, wins, losses, ties.

yearID :Year

lgID : League; a factor with levels AA AL FL NL PL UA

teamID : Team; a factor

franchID : Franchise (links to TeamsFranchises table)

divID : Team’s division; a factor with levels C E W

Rank : Position in final standings

G : Games played

Ghome : Games played at home

W : Wins

L : Losses

DivWin : Division Winner (Y or N)

WCWin : Wild Card Winner (Y or N)

LgWin : League Champion(Y or N)

WSWin : World Series Winner (Y or N)

R : Runs scored

AB : At bats

H : Hits by batters

X2B : Doubles

X3B : Triples

HR : Homeruns by batters

BB : Walks by batters

SO : Strikeouts by batters

SB : Stolen bases

CS : Caught stealing

HBP : Batters hit by pitch

SF : Sacrifice flies

RA : Opponents runs scored

ER : Earned runs allowed

ERA : Earned run average

1. Batting : Batting table includes 105861 observations with 22 variables which is basic count of players offensive performance information and statistics(e.g. Games , At bat, Hits)

yearID :Year

stint : player’s stint (order of appearances within a season)

lgID : League; a factor with levels AA AL FL NL PL UA

teamID : Team; a factor

G : Games played

R : Runs scored

AB : At bats

H : Hits by batters

X2B : Doubles

X3B : Triples

HR : Homeruns by batters

RBI : Runs Batted In

BB : Base on Balls

SO : Strikeouts by batters

SB : Stolen bases

CS : Caught stealing

IBB : Intentional walks

HBP : Batters hit by pitch

SF : Sacrifice flies

SH: Sacrifice hits

GIDP : Grounded into double plays

1. People : People table includes demographic of players information with 19617 observations(e.g. birthyear, birthMonth, birthCity)

deathState: State where player died

deathCity : City where player died

nameFirst : Player’s first name

nameLast : Player’s last name

nameGiven : Player’s given name (typically first and middle)

weight : Player’s weight in pounds

height : Player’s height in inches

bats : a factor: Player’s batting hand (left (L), right (R), or both (B))

throws : a factor: Player’s throwing hand (left(L) or right(R))

debut : Date that player made first major league appearance

finalGame : Date that player made first major league appearance (blank if still active) retroID : ID used by retrosheet, <http://www.retrosheet.org/>

bbrefID : ID used by Baseball Reference website, http://www.baseball-reference.com/ birthDate :Player’s birthdate, in as.Date format

deathDate : Player’s deathdate, in as.Date format

1. Pitching : Pitching table includes stats of pitcher with 46699 observations on 30 variables (saves, earned runs, walks)

Player ID : Player ID code

yearID :Year

stint : player’s stint(order of appearance within a season)

lgID : League; a factor with levels AA AL FL NL PL UA

teamID : Team; a factor

G : Games played

W : Wins

L : Losses

GS : Games Started

CG : Complete Games

SHO : Shutouts

SV : Saves

IPouts : Outs Pitched(inning pitched \*3)

H : Hits by batters

ER : Earned Run

HR : Homeruns

BB : Walks

SO : Strikeouts

BAOpp : Opponent’s Batting Average

ERA : Earned run average

IBB : Intentional Walks

WP : Wild Pitches

HBP : Batters hit by pitch

BK : Balks

BFP : Batters faced by Pitcher

GF : Games Finished

R : runs Allowed

SH : Sacrifices by opposing batters

SF : Sacrifice flies

GIDP : Grounded into double plays by opposing batter

**Method & Analysis**

The project proceed based on the R program. Since the database is already in R as a package, the data are clean, and the database is well organized. Therefore, I did not have many processes for pre-processing. I used ‘Lahman’, ‘dplyr’, ‘data.table’, ‘ggplot2’ and 'Picante' for my package. As mentioned earlier, the salary, team, batting, and people tables were loaded using the data () format and tables needed for each question were selectively used.

**Question 1 : Which player have a higher batting average? Is top hitter always paid higher salary?**

**Method**

For this part, my purpose is finding out comparing the Batting average of top players and their salary in each season to find out whether a player who has higher salaries always worth the money. To do this task, I needed a salary table and a batting table. Also, the People table has imported so that the player demographics can be seen.

First, I loaded the Salary and people table and saved it as an R data frame. The batting data table is saved using data () format. To remove unnecessary variables from the salaries table and retrieve the required variables, we used the **dplyr** select function to select the playerID, yearID, teamID, and salary variables and store them in salaries.

In the same way, we saved playerID, birthYear, birthMonth, nameLast, nameFirst, and bats from the people tables. Also, by using batting stats () in layman package, basic stat information is converted into aggregated data such as Batting Average and merged with batting data. Also, left join the people and salaries table to the batting stat. At this time, the basic information of the player in the people dataframe is replaced with age. Finally, I sorted it by playerID, yearID, and stint and make a base table for the analysis.

To select eligible hitter from the base table, I sort player who has more than 520 counts of PA(plate appearance of player), which considered as Regular Plate in MLB. Then, I sort top hitters in the league in each year based on the batting average after the year 2000 and saved table as topHitters. Then I left joined salaries to topHitters dataframe to check whether there is a correlation between salary and batting average among top hitters in the season. After I got the table, there is missing value for the salary of top hitters in 2017, 2018. I googled and imputed the missing value manually for the analysis.

The challenge for Question 1 is there are some missing values of the salary of players. I have tried to impute those missing values after googling, the information of players salary differs from the information resources. Finally, I calculated the average amount of salary that I found on google search then imputed those value

**Analysis and Result**

A screenshot of text

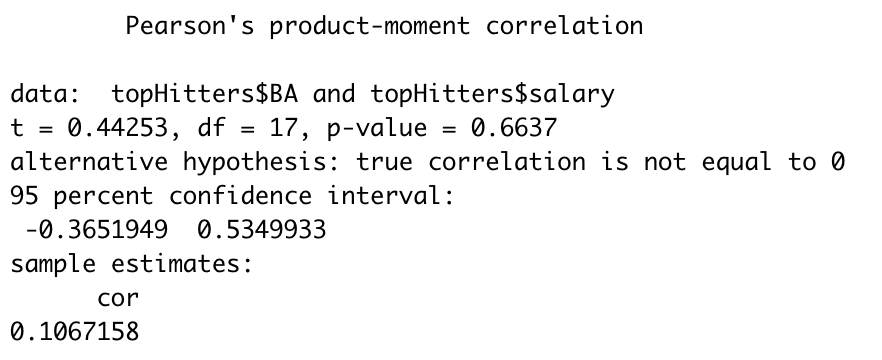
Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 1. Top hitters salary Figure 2. Top hitters Batting Average

For Question 1, I was first plotting top hitters salary and top hitters batting average by each year using geom\_line() in ggplot. I have used a scatter plot and cor.test to verify the result.

Figure 1 shows the top hitters' salary in each season. In this graph, the salary of the top players shows a fluctuating pattern, which means that the batting average for the year is the highest, but the salary of such a player is very different. This can easily be compared to the Batting average graph shown in Figure 2. For example, Nomar Garciaparra, who was the top hitter in 2000, received a lower salary, although the batting average was significantly higher than in other players in different years. At a glance, we could see them there is no trend between Batting average and player salary. To check whether it is true, I used scatter plot and cor.test() to find out statistical evidence.

A close up of a piece of paper

Description automatically generated

Figure 3. Top hitters BA vs Salary and cor.test() result

As you saw in the scatter plot in Figure 3., the scatter plot does not show the trends. Also, the correlation test result indicates the number of 0.106, which stands for there is no or less correlation between the top hitters’ salary and batting average. In particular, we could assume that the player with the least salary performed better in some seasons, and the salary is not always a good predictor for players’ performances.

**Question 2 : What type of variables has considerably predict the team win?**

**Method**

For Questions 2, I load the Teams table to perform team level analysis using data(Teams). Then I filter the data using dplyr packages that data after 1920, the reason for filtering data only after 20. The reason for filtering data only after the year 1920 is that MLB has a different trend in whether the pitcher outperforms or the attack outperforms based on the official game ball difference. Before 1920, the repulsive force was not good on baseball, so it is called the dead-ball age. After the repulsive power has changed, the batters batting average is skyrocketed, and we called it is the "liveball" era. Also, I calculated variables such as Batting average, Total base, and Win percentage, which are synthetic variables that did not exist in the existing Teams table, and muted those synthetic variables and saved them in 'teams' data frames. After, dplyr's select() function was used to drop unnecessary columns such as franchised, divID, teamIDBR, etc. We will discuss the detailed analysis process in the Q2 analysis part.

The challenge I encountered is I have to understand the history of the data, especially the data which has over 100 years, so it has to remove the data which considered as outdated. Although I like baseball, I had less knowledge of the history of MLB. Fortunately, After I have study about MLB history, I could select the right standard for choosing a range of years.

**Analysis and Result**

In baseball, hit(H) is one important factor in determining the quality of team performances. Hits include single hits, doubles, triples, and home runs. Because of the more hits you have, the more likely you are to score. Also, the Hit Against(HA) feature is another crucial feature that shows how many hit the opponent team made from our pitchers.

To find out how these two features correlate with the winning percentage, I calculated the correlation score using cor (). As a result, Hit vs. WinPct (Winning Percentage) was 0.312. HA vs. WinPct also scored slightly higher than H with -0.38.

A screenshot of a cell phone

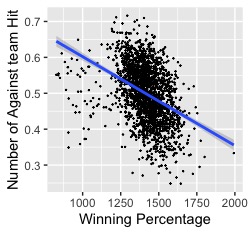
Description automatically generated

Figure 4.1 WinPct vs. Team Hit Figure4.2 WinPct vs. Against Team hit

To verify this, I analyzed it with linear regression using R's lm () function. Both H and HA showed a p-value of less than 0.05. It looks like the results show that HA is a better predictor for predicting H and team wins. However, it looks like the result is meaningful, but the R-squared values ​​were 0.09 and 0.14, respectively. And, both variables showed low scores of 0.4 or less on a correlation test, so I judged it was not possible to predict the winning percentage accurately with each H and HA.

A screenshot of text

Description automatically generated A screenshot of text

Description automatically generated

Figure 5.1.WinPct vs Team Hit regression summary Figure 5.2. WinPct vs Against Team Hit Regression summary.

To find a better model, I manipulated two variables. Since both are essential features, we created a new variable called HD (Hit difference = Hit-Hit against). The correlation test and linear regression were performed with this.

A close up of a map

Description automatically generated

A screenshot of text

Description automatically generated

Figure 6. HD vs WinPct and regression summary

The correlation score of HD and WinPct was 0.787, which is much better than the model I explained above(H vs. WinPct and HA vs. WinPct). Also, as you can see in Figure 6., linear regression resulted in a p-value of less than 0.05. In addition, the R-squared value was found to be 0.619, which makes it a good model.

A close up of a mans face

Description automatically generated

Figure 7. Residual plot of Hit Difference

Next, based on the regression results, the prediction was performed using the predict () function and then fitted values ​​were calculated. As you seen in Figure 7.,The residuals were also computed using residuals () from the regression results of HD, and then plots and root mean square error values ​​(RMSE) were computed. As a result, the RMSE value is 0.048, and the number shows Hit Difference is one of the good predictors to predict the team winning.

**Question 3 : Is the team total salary correlated with the winning ratio?**

**Method**

First, I merged the pithing table using the batting table we created in Q1. The reason for combining the pitching table is that it doesn't exist in the batting table because American League pitchers are only on defense. So I left join pithing table to the batting table to create a data frame called players. Then I joined the salary information to this player table. Then, I group by each team in the players' table, the total salary of each team was calculated by summating salary by the player, and then salary.cap data frame was created. Then, I inner join the winning percentage in the team we calculated for the analysis.

In the analysis process, a correlation test was performed between salary caps and winning ratios of teams for each year using a for a loop. I tried to use the for loop for each year, but there is error kept came out. I have not yet fully mastered the syntax of R and dplyr. However, I was able to solve this problem by finding a solution in the stack overflow.

**Analysis and Result**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| year | Correlation score |  | year | Correlation score |
| 1985 | 0.1562385 |  | 2001 | 0.1780343 |
| 1986 | 0.1921632 |  | 2002 | 0.2695943 |
| 1987 | -0.2556442 |  | 2003 | 0.1333688 |
| 1988 | 0.02174955 |  | 2004 | 0.3637566 |
| 1989 | 0.06140128 |  | 2005 | 0.3821037 |
| 1990 | -0.01104522 |  | 2006 | 0.3486422 |
| 1991 | 0.01949574 |  | 2007 | 0.3834081 |
| 1992 | -0.1553252 |  | 2008 | 0.146725 |
| 1993 | 0.08233727 |  | 2009 | 0.20969 |
| 1994 | 0.329456 |  | 2010 | -0.02169068 |
| 1995 | -0.1198194 |  | 2011 | 0.2977321 |
| 1996 | 0.3633152 |  | 2012 | -0.2766616 |
| 1997 | 0.158912 |  | 2013 | 0.02593012 |
| 1998 | 0.4287351 |  | 2014 | 0.221098 |
| 1999 | 0.3654035 |  | 2015 | -0.3194242 |
| 2000 | 0.05145455 |  | 2016 | 0.3037889 |

Table. 1. Correlation score by year (Winning ratio Vs. Team Total Salary)

To analyze the data, I have used for loop to extract the correlation between a team winning ratio and team total salary using cor.test(). Within the process, I extract the unique value of years to put it in for loop.

Table 3 shows the correlation scores between the winning percentage and the total salaries of the players by year. Looking at the table, 1998, 2005, and 2007 were 0.428, 0.382, and 0.383, respectively, which showed some correlation between total salary of a team and their winning ratio. However, there is no correlation in other years since the scores are low. Even in some seasons, there is a minus correlation. Judging by this, it was found that the entire salary did not affect the team winning ratio.

In the meantime, baseball has paid astronomical signup bonuses and salary to players to the star players who released as Free Agent status. However, based on my analysis, it was easily observed that a higher total salary of the team does not mean a high winning ratio. In other words, instead of paying a higher salary to the star player. Building a team based on the players have the ability, but undervalued players could help to improve the team's winning ratio.