Project Report - Baseball Dataset Investigation

In this project, I will be performing Data Analysis on the **Baseball Dataset**. Following are the steps followed alongwith the questions, answers, detailed description and summary of the data analysis performed.

Step One - Choosing the Data Set : Baseball Data

The files chosen for Data Analysis in this project:

The dataset comprises of many different files, but since detailed analysis of every file can lead to infinite data wrangling possibilities, I am going to choose a few files and perform my analysis on them. Following are the tables I am most interested in:

Main tables:

1. Batting - batting statistics
2. Pitching - Pitching statistics
3. Fielding - Fielding statistics

Supporting Tables:

1. Salaries - player salary data
2. Teams - yearly stats and standings
3. Appearances - details on the positions a player appeared at
4. AwardsPlayers - awards won by players

Step Two - Questions posed on the Data Set

As per the selection of above dataset & tables, I will focus on answering the following questions, and see if the data is capable to provide some insight into relationships between different parameters (independent and dependent). The analysis conducted on the selected data set may/may not specifically and accurately answer all the above questions, but it might help to identify certain metrics which can be further used (or analyzed) to take some decisions (for example, selecting the best performers for future matches, creating a team with maximum award winners etc.)

Questions:

Team Table Analysis

1. Analyzing the teams dataframe for extracting relevant information. We will analyze how the wins are distributed over all the years? Also, we will identify the teams with maximum wins over all the years and calculate ERA per game per year. Lastly, we will look if any correlation exists between Runs per game, Runs allowed per game and Home runs per game?

Batters and Salary Association

2. Considering Batting table as the main table here, we will try to identify the metric that helps in evaluating batters (using the details provided in the Batting and Salaries tables):
    a. First we will calculate three major metrics (AVG, OPS and RC) to measure a batter's performance. How to calculate these metrics has been explained in the solution section. We will try to identify which is the best metric amongst them which reveals how many homeruns will a batter hit?
    b. Also, we will find out which of these statistic has the highest correlation with player salary?

Awards, Appearances, Pitching and Fielding DataFrame Association

3. Using data and stats collected from these 4 dataframes, we will create a consolidated dataframe. We will also clean the data by removing duplicates and handling nulls and NaNs. We will then analyze this data to find if there is any association between total games played and the chance of winning and losing a match? That is, if a player plays many matches, how does his chance of winning/losing varies? We will analyze this data for the players from two sections - one who have won more than 10 total awards and others who have won less than 10.

Step Three - Solution Methodology(Analysis) and Synthesis of all Data Wrangling Performed

Please Note : Explanation of handling missing values all throughout the project:

Step 1 :
We have executed the statement : *<df_name>.isnull().sum(axis=0)*
to identify if there are already any null or missing values in the dataset. This step provides the sum of null values corresponding to each column in the dataframe as output. Please refer the output of cells in ipynb.

Step 2 :
The above statement will provide how many columns and amongst them how many rows have null values.

Now we will handle these values manually by filling them with zeros using the function:

fillna(value='0',inplace=True)
inplace = True will change the original dataframe

For columns such which matter to the average/mean calculation, we have replaced null values with the median of the column values. We have done this to ensure standardization of these metrics. We don't want a metric as low as zero to affect our column average, analysis calculations and visualizations, hence we will use following function for these columns:

fillna(record.median())

Please Note:

We did not use dropna here since we want to retain the details of each team irrespective of whether the value is missing or available. We wanted to standardize data and therefore used the above steps

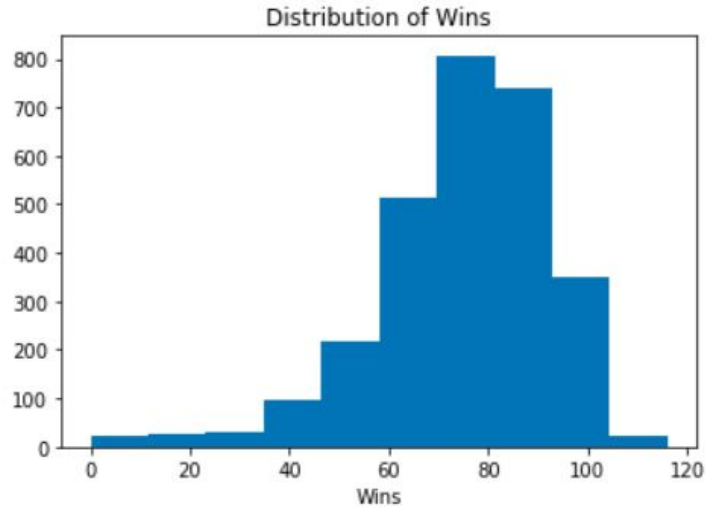
Solution to Problem 1:

In this problem, we will be analyzing teams dataframe for relevant information. We will be performing following steps to draw some inferences:

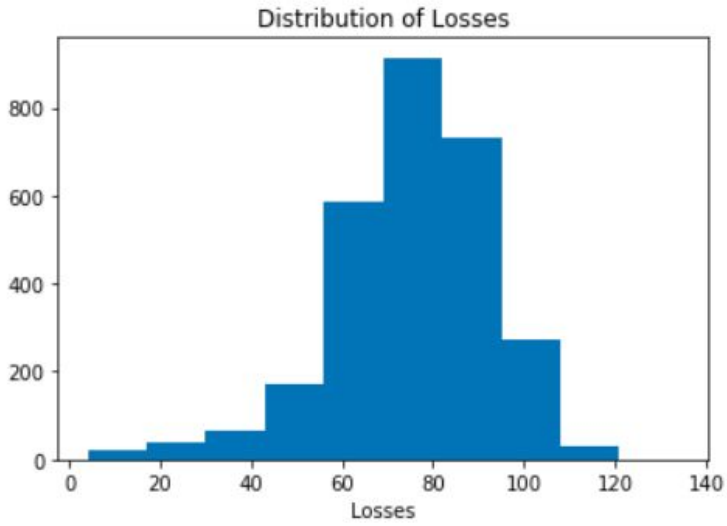
Step 1: Collecting team stats and remove any unwanted columns from the dataframe and handle null or NaNs. Then we will have a cleaned data ready for further data analysis.

Step 2: We will look into the distribution of Wins and losses and identify the mean wins and losses in the above cleaned dataset

```
Min of Wins in teams table: 0
Mean of Wins in teams table: 74.81410934744268
Max of Wins in teams table: 116
```

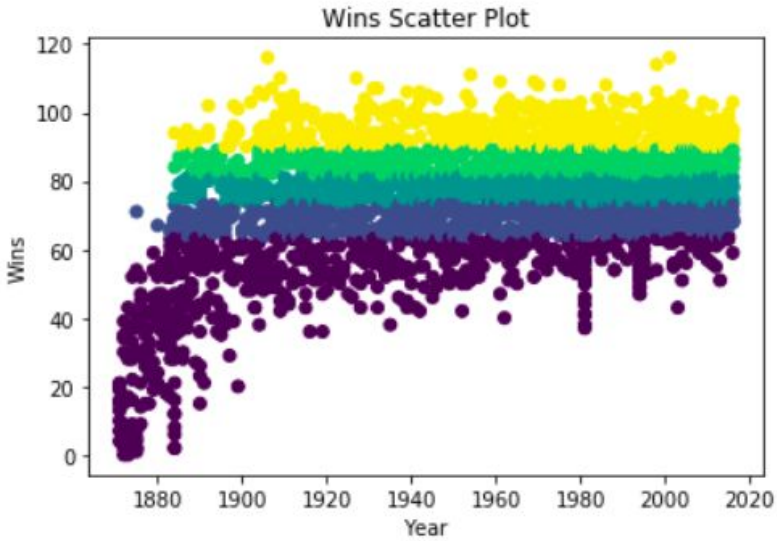


```
Min of Losses in teams table: 4
Mean of Losses in teams table: 74.81410934744268
Max of Losses in teams table: 134
```

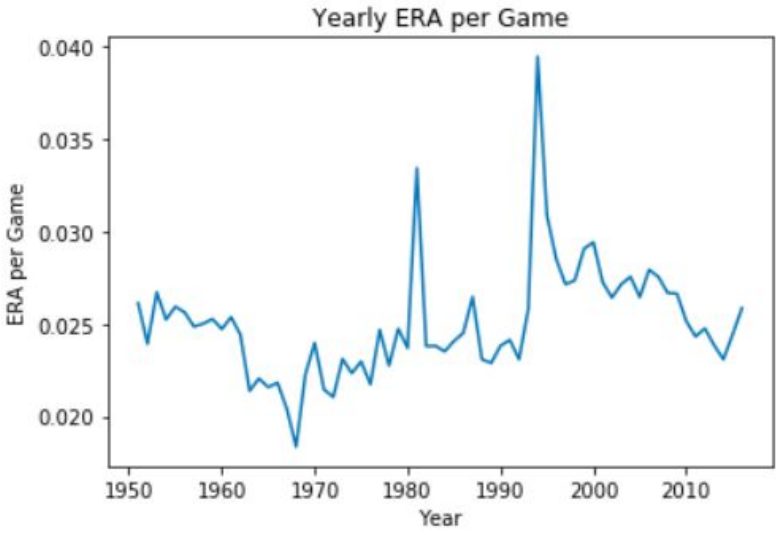


On comparison, we observed that mean of wins and losses in a team are almost same but the minimum loss by any team is 4. That means there is a team which never won and only lost that too 4 times.

Step 3: Analyzing the scatter plot for wins over the years. We observe that the wins have been increasing significantly from 1910 onwards. Before that teams have lost badly. This could be due to improved techniques over time and careful selection of players.



Step 4: On analyzing ERA per game per year, we observed this has been increasing till 1995 and achieved its peak there. But soon after that it started to decrease and is now almost matching the rate at which it was at the beginning of the stats collection period.



Step 5: Analyzing the correlation between Runs per Game, Runs allowed per game and Home runs per game v/s wins. The plots have been shared in summary section.

****

Solution to Problem 2, Part a:

Step 1: Based on certain relevant theories, (stated in the Reference Section), we will be calculating certain parameters, as explained below and then continue with the analysis:

a) Calculate AVG - A statistic used long earlier to evaluate performance. To calculate player's the batting average, divide the number of hits by the number of at-bats.

$$AVG = H/AB$$

b) Calculate OBP (on-base percentage) - It is a statistic generally measuring how frequently a batter reaches base. Specifically, it records the ratio of the batter's times-on-base (the sum of hits, walks, and times hit by pitch) to their number of plate appearances. To find a player's on-base percentage, or OBP, add his hits, walks and hit-by-pitch totals and divide that sum by the combined total of his at-bats, walks, hit-by-pitch and sacrifice flies.

$$OBP = [(H+BB+HBP)/(AB+BB+HBP+SF)]$$

c) Calculate TB (Total bases) - Sum of one for each single, two for each double, three for each triple, and four for each home run

$$TB = [H + 2B + (2 \times 3B) + (3 \times HR)]$$

d) Calculate SLG (slugging percentage) - It is a measure of the batting productivity of a hitter. It is calculated as total bases divided by at bats, through the following formula, where AB is the number of at-bats for a given player, TB is the Total Bases calculated above. To find SLG, divide his total bases by his at-bats.

$$SLG = TB/AB$$

e) Calculate TA (Total average) - total bases, plus walks, plus hit by pitch, plus steals, minus caught stealing divided by at bats, minus hits, plus caught stealing, plus grounded into double plays

$$TA = [(TB + BB + HBP + SB – CS)/(AB – H + CS + GIDP)]$$

f) Calculate OPS (on-base plus slugging) - An even better measure of performance than slugging percentage or on-base percentage is their sum.

$$OPS = SLG + OBP$$

g) Calculate RC (Runs Created Formula) - In the late 1970s, a remarkable statistic was discovered for measuring a batter's performance.

$$RC = (H + BB) \times (\text{Total bases})/[AB + BB]$$

Step 2: In this step, we will try to find if any correlation exists between the homeruns a batter might hit and the above three metrics. If we find any strong positive correlation, we can identify which of these (TA, OPS and RC) is the best metric to identify the batter hitting maximum homeruns.

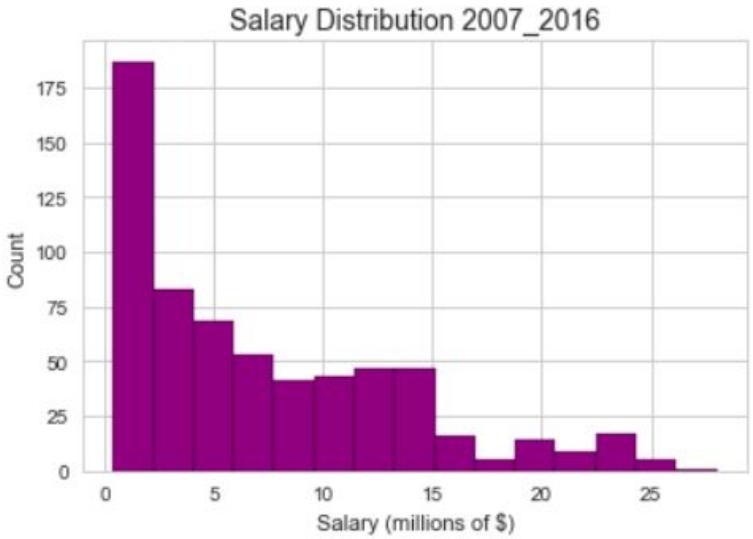
Step 3: In this step, we will plot the graphs to show the correlation calculated in step 2. The plots have been shared in summary section.

Solution to Problem 2, Part b:

Step 1: First off, we are selecting Batting data for period from 2007 to 2016. Then we will further filter records by including players who have played more than 25 games. This will enable us to standardize all other calculations. After that, we will merge this data with the corresponding salary data to identify how batting stats affect salary of a player.

Step 2: Now we will Calculate correlation between salary and important batting metrics(AVG, OPS and RC). Also plotting a graph for visualization. The stats are discussed in the summary section.

Step 3: Analyzing the distribution of Salary and as we expect, it's clear from the visualization that very few players have salary above $15million.



****

Solution to Problem 3:

Step 1: We will first fetch pitching, fielding, awards and appearances data into dataframes

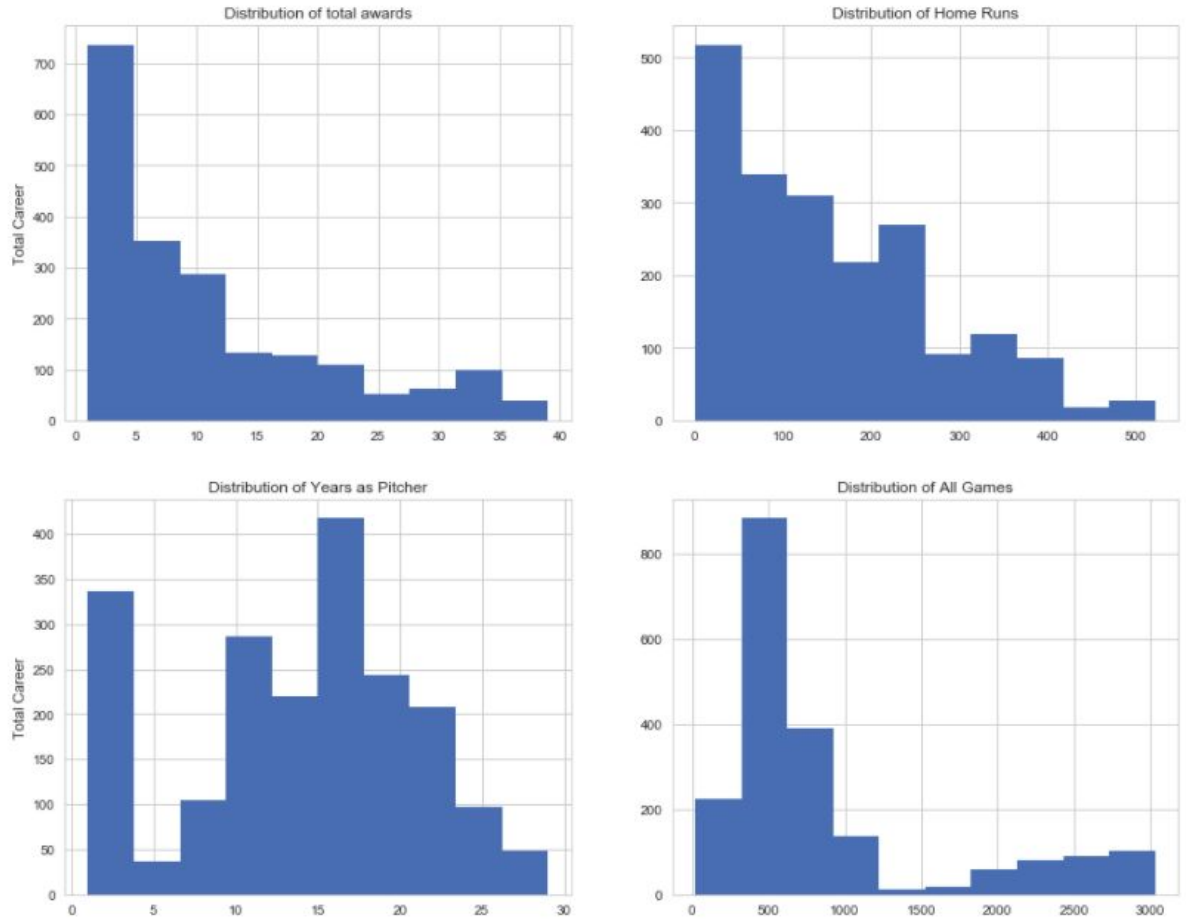
Step 2: We will then create consolidated statistics for pitching and fielding dataframes by summing up stats such as H, HR, 2B and so on. Also, we will sum up the years played as pitcher and fielder. We will then create a consolidated dataframe 'player_stats_df with all relevant data.

Step 3: Similarly we will compute total_awards and then add relevant columns to player_stats_df

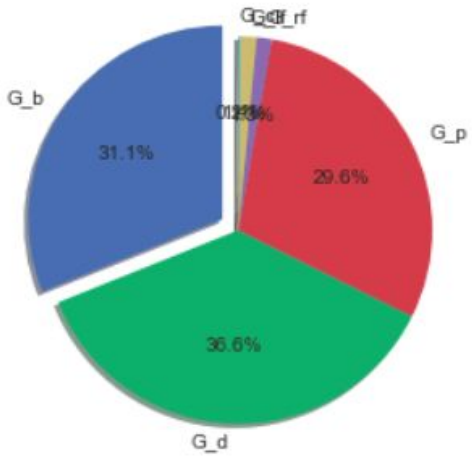
Step 4: We will perform calculations on appearances data and then finally merge with player_stats_df and create a final consolidated dataframe : final_stats

Step 5: We will then handle missing values and NaNs

Step 6: We will then compare the plots for total_awards, total_years_pitcher, home runs and all games played by a player.



Step 7: We will then identify the mean appearances of all players using the visualization and identify which roles are played most by any player. We observed that a player appears (mean value) most in defense, followed by batting and then in pitching. The other appearances are quite insignificant.



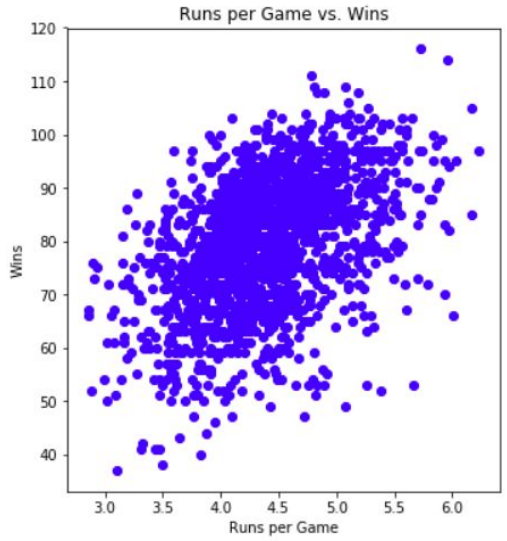
Step 8: Lastly we will compare wins and losses against all games played by the players who won awards. Results are shared in the summary section.

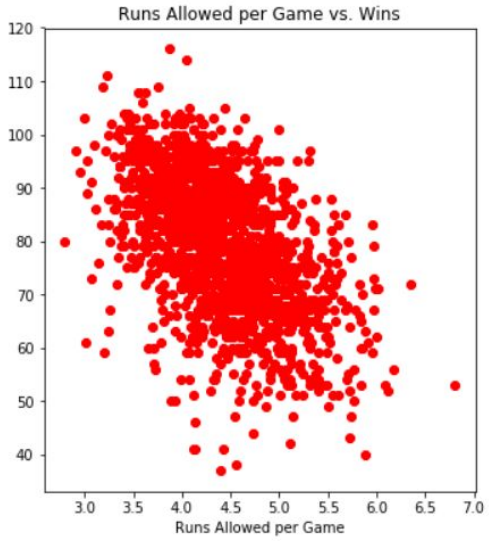
Step Four - Summary Statistics and Plots - All Findings

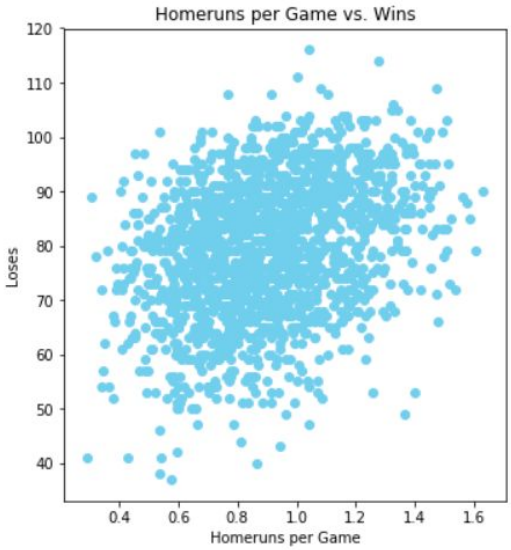
Here we will be listing all the correlation we found in the above problems. Rest of the graphs have been explained in the solution section itself.

Summary 1:
On analyzing the teams table, and finding the correlation between the three metrics :
Runs per Game, Runs allowed per game and Home runs per game v/s wins, we observed that: R_per_game and HR_per_game are positively correlated with wins but Runs allowed per game is negatively correlated. Though none of them is strongly correlated, but we can see from the below plot that Runs per game is highest correlated statistic with wins than any other metric. Hence, we expect that if runs_per_game is high for a team, its chance for winning the game increases by almost 0.5.

```
R_per_game      0.471139
RA_per_game    -0.524900
HR_per_game     0.327289
```



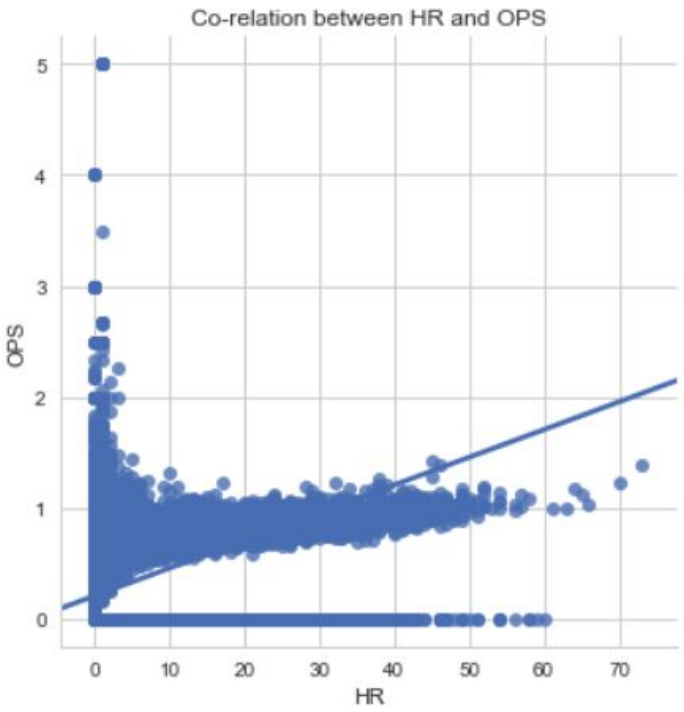


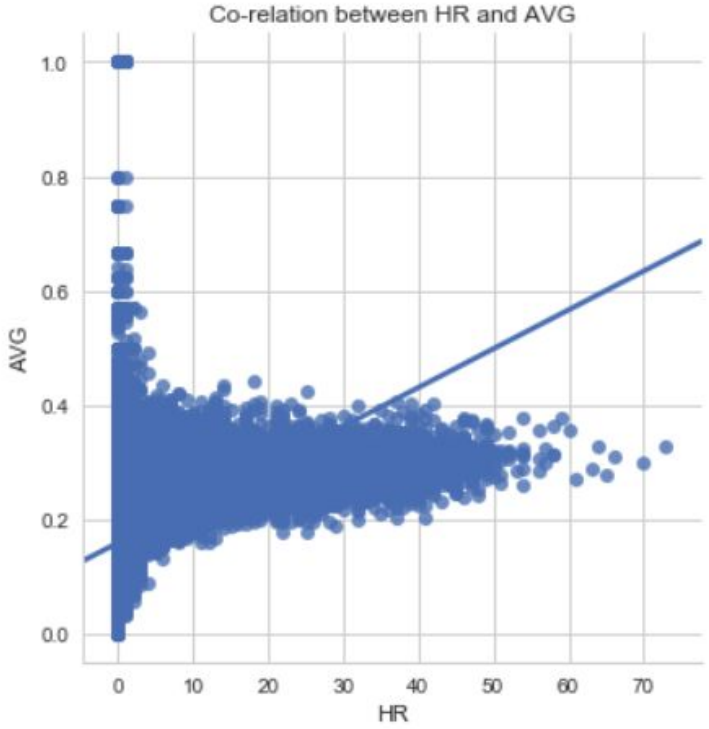


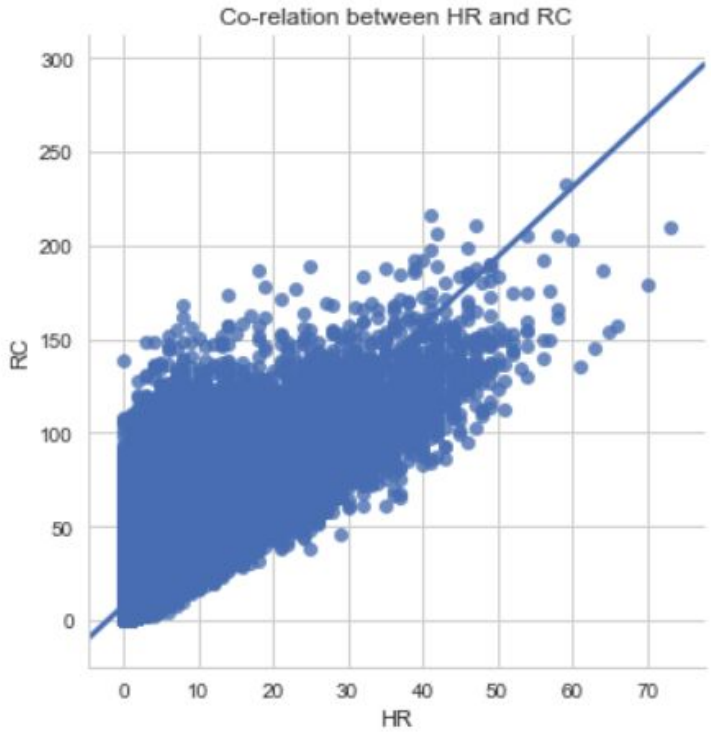
Summary 2: Part A

As observed from the stats below and from the correlation between Home Runs and three metrics ‘AVG’, ‘OPS’ and ‘RC’, ‘RC’ is the factor which is strongly positively correlated with Home Runs. It has also been proved by number of baseball experts that Runs created is the basis for identifying a better performer in batting. Hence, the data supports this fact.

```
Correlation between homeruns and average: 0.31667851920192336
Correlation between homeruns and on_base_plus_slug: 0.4381493419248805
Correlation between homeruns and runs_created: 0.8149748818841528
```


Co-relation between HR and OPS
OPS
HR
0
1
2
3
4
5
0
10
20
30
40
50
60
70


Co-relation between HR and AVG
AVG
HR
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70

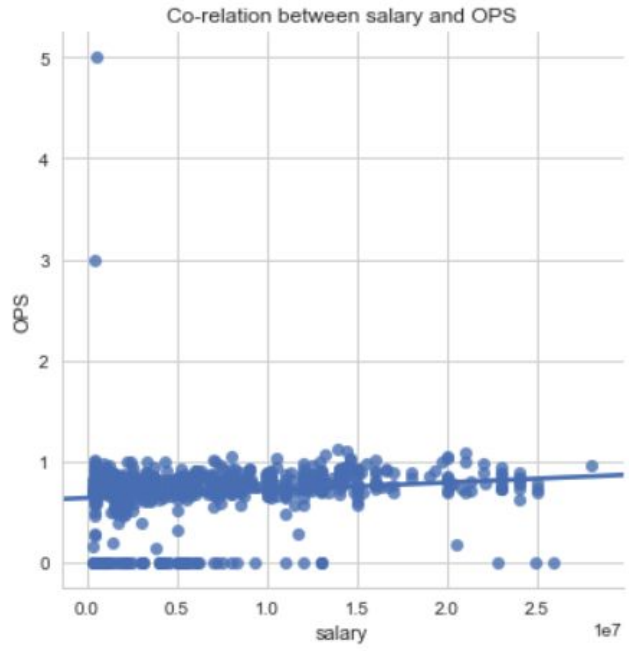


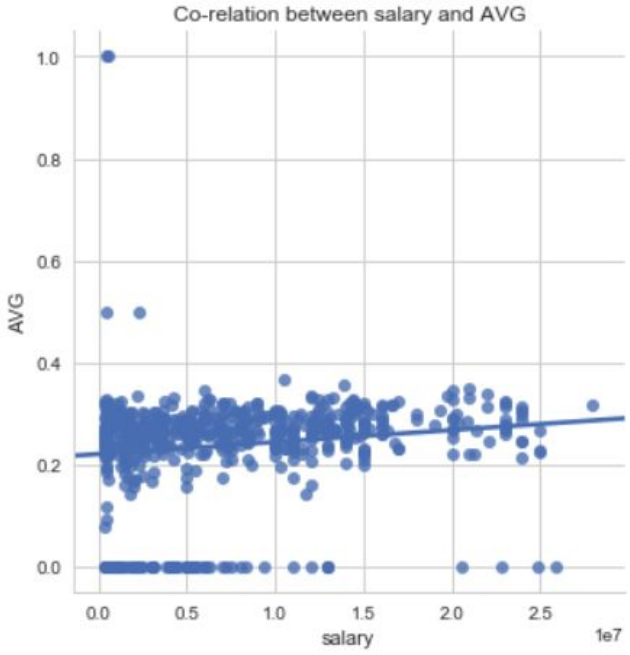
Summary 2: Part B

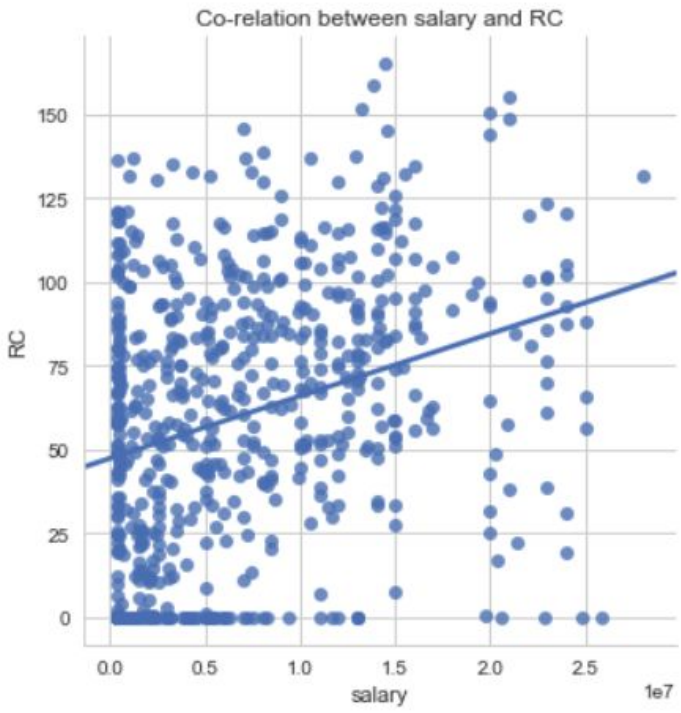
The three metrics - ‘AVG’, ‘OPS’ and ‘RC’, when compared with salary provide the result:

AVG	0.144177
OPS	0.145258
RC	0.300589

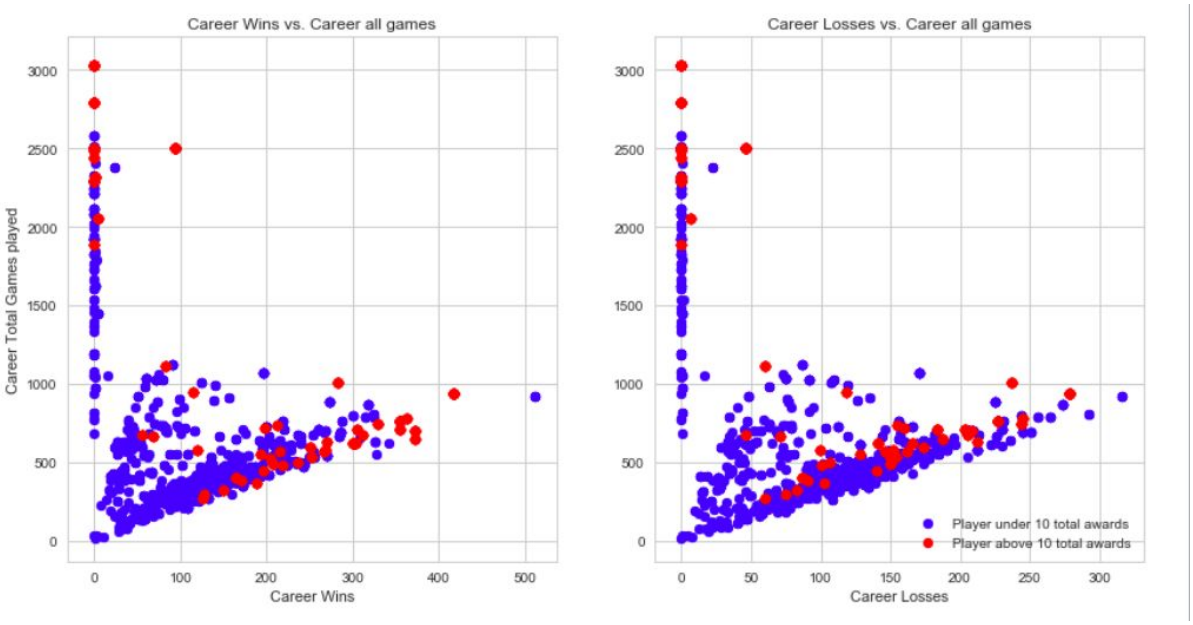
This data shows that though salary is positively correlated with all three metrics, there is no single metric that affects salary strongly. Amongst the three, however, RC is the one metric that is highest correlated statistic with salary. However, there may be other factors which affect salary as well.


Co-relation between salary and OPS
OPS
salary
1e7
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5


Co-relation between salary and AVG
AVG
salary
1e7
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5



Summary 3: Though there are not quite strong correlations but we can surely conclude that the if a player plays n games, his chances of winning or losing do not vary significantly. Even there are certain data points that prove that though the players have won more than 10 awards, it's not necessary that they will win as compared to players who have on less than 10 awards. Therefore, there might be other factors deciding wins for a player. Since we did not take into account the batting stats, probably hitting more runs would proportionally result in more wins.



Step Five - Final Conclusions and Limitations

Conclusion

There are several conclusions that can be drawn from this analysis but there are also numerous caveats that must be mentioned that prevent these conclusions from being accepted as fact. I also must state the ever-present rule that correlation does not imply causation. In other words, players statistics may be correlated with salary, but that does not mean the statistics caused a higher or a lower salary.

The main conclusions of all three problems are as follows:

1. On analyzing the teams table, we observed that : Runs per game is highest correlated statistic with wins than any other metric amongst Runs per Game, Runs allowed per game and Home runs per game v/s wins. Hence, we expect that if runs_per_game is high for a team, its chance for winning the game increases by almost 0.5. For other two metrics, HR_per_game is positively correlated with wins but Runs allowed per game is negatively correlated.

2. As observed from the correlation between Home Runs and three metrics ‘AVG’, ‘OPS’ and ‘RC’, ‘RC’ is the factor which is strongly positively correlated with Home Runs. It has also been proved by number of baseball experts that Runs created is the basis for

identifying a better performer in batting. Also in part b, the analysis shows that though salary is positively correlated with all three metrics, there is no single metric that affects salary strongly. Amongst the three, however, RC is the one metric that is highest correlated statistic with salary. However, there may be other factors which affect salary as well.

3. Though there are not quite strong correlations but we can surely conclude that the if a player plays n games, his chances of winning or losing do not vary significantly. Even there are certain data points that prove that though the players have won more than 10 awards, it's not necessary that they will win as compared to players who have on less than 10 awards. Therefore, there might be other factors deciding wins for a player. Since we did not take into account the batting stats, probably hitting more runs would proportionally result in more wins.

Limitations

Besides the ever-important rule that **correlation does not equal causation**, there were several other limitations to the data analysis. The main caveats I was able to identify are as follows:

1. The size of the data. After consolidating four tables in problem 3, we had only summed up and counted data in the table. We also did not consider batting table in join which could have affected the Hits significantly and therefore we might have been able to identify a strong correlation between chances of winning and losing.
2. As a player gets older and accumulates more years spent in the league, their salary tends to increase even if their performance does not. I did not control for the age of the players in the analysis.
3. Due to randomness in baseball stats from season to season, some years are better pitching years and some years are better for batters. Statistics can also vary widely between ballparks due to the geometry of the park or even the density of the air. I did not control for the such metrics in which the player accumulated their statistics.
4. Post-season performances may also factor heavily into a player's salary than during the regular season. A more thorough analysis would look at the relative impacts on salary of post-season performance compared to the regular season. Moreover, there could be more effective measures of player performance, as opposed to the

ones we chose for analysis in problem 2. One of the more effective measures of performance is known as Wins Above Replacement (WAR) which assigns a number to a player representing how many wins they contributed to their team versus what an average player would have generated. WAR for batters takes into account at least 8 separate statistics!

Step Six - Future Research

For future research, I would like to focus on the prediction stuff I stated in the beginning of this report. As to can we predict managers to the team who have won more awards for their teams in the past. Similarly predictions can be made for batters performance, though a lot of variable controlling would be required for that. I am hoping to get answers for alike questions in the future by using machine learning techniques.

Links and References used for performing data analysis:

Below are the links referred to in data analysis:

http://pandas.pydata.org/pandas-docs/stable/dsintro.html#dataframe
https://docs.scipy.org/doc/numpy/user/basics.html
http://pandas.pydata.org/pandas-docs/stable/visualization.html
http://pandas.pydata.org/pandas-docs/stable/groupby.html
https://www.maa.org/external_archive/devlin/devlin_09_04.html
http://matplotlib.org/api/pyplot_api.html
https://seaborn.pydata.org/generated/seaborn.lmplot.html

Baseball rules and statistics calculation:

http://www.wikihow.com/Read-Baseball-Statistics
http://www.csgnetwork.com/baseballoffensestatscalc.html
http://www.baseball-almanac.com/bstatmen.shtml

https://www.baseball-reference.com/about/bat_glossary.shtml
https://en.wikipedia.org/wiki/On-base_percentage