CSE 5544: Introduction to Data Visualization

Analysis and Visualization of MLB Baseball Datasets

Chandrasekar Swaminathan (.42)

Introduction

We want to analyze the Major League Baseball (MLB) statistics of all teams, players and pitchers for the year of 2015 to determine the answers to the following questions:

- In the year of 2015, Is a team's win-loss record related to its payrolls?
- In the year of 2015, Is a player's batting performance related to his team's win-loss record?
- In the year of 2015, is a team's win-loss record related to its pitching performance?

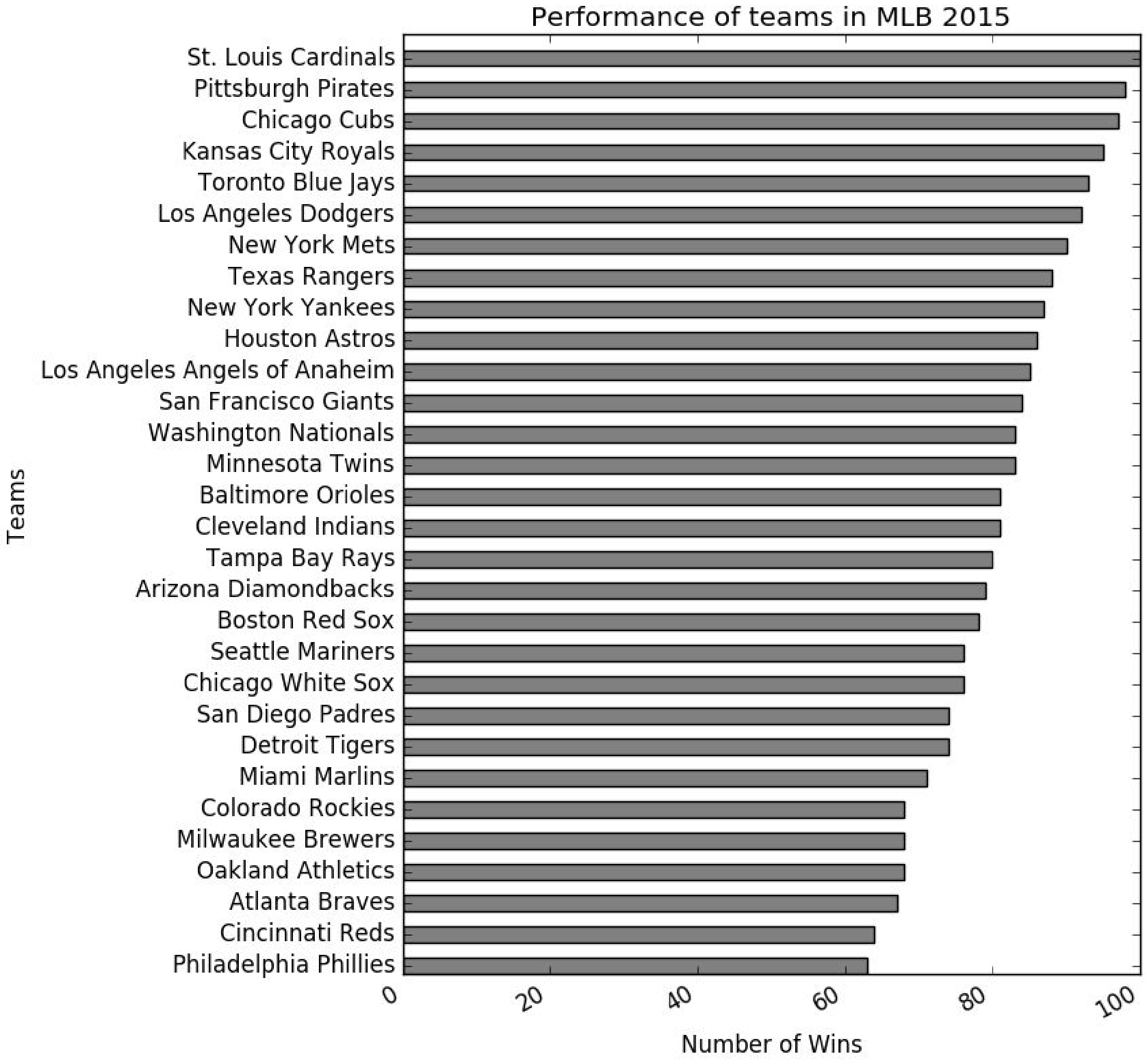
Relationship between payroll and team performance

We can first look at the performance of the teams, the payroll of the teams and then look at the correlation between the two features to identify if there is any relationship between the payroll and the team performance.

For visualizing the team performance, it makes sense to use a bar chart with the number of wins for each team, as shown in ***Figure 1***. From this visualization we can infer that the team with the maximum number of wins is **St. Louis Cardinals** and the team with the least number of wins is **Philadelphia Phillies**.

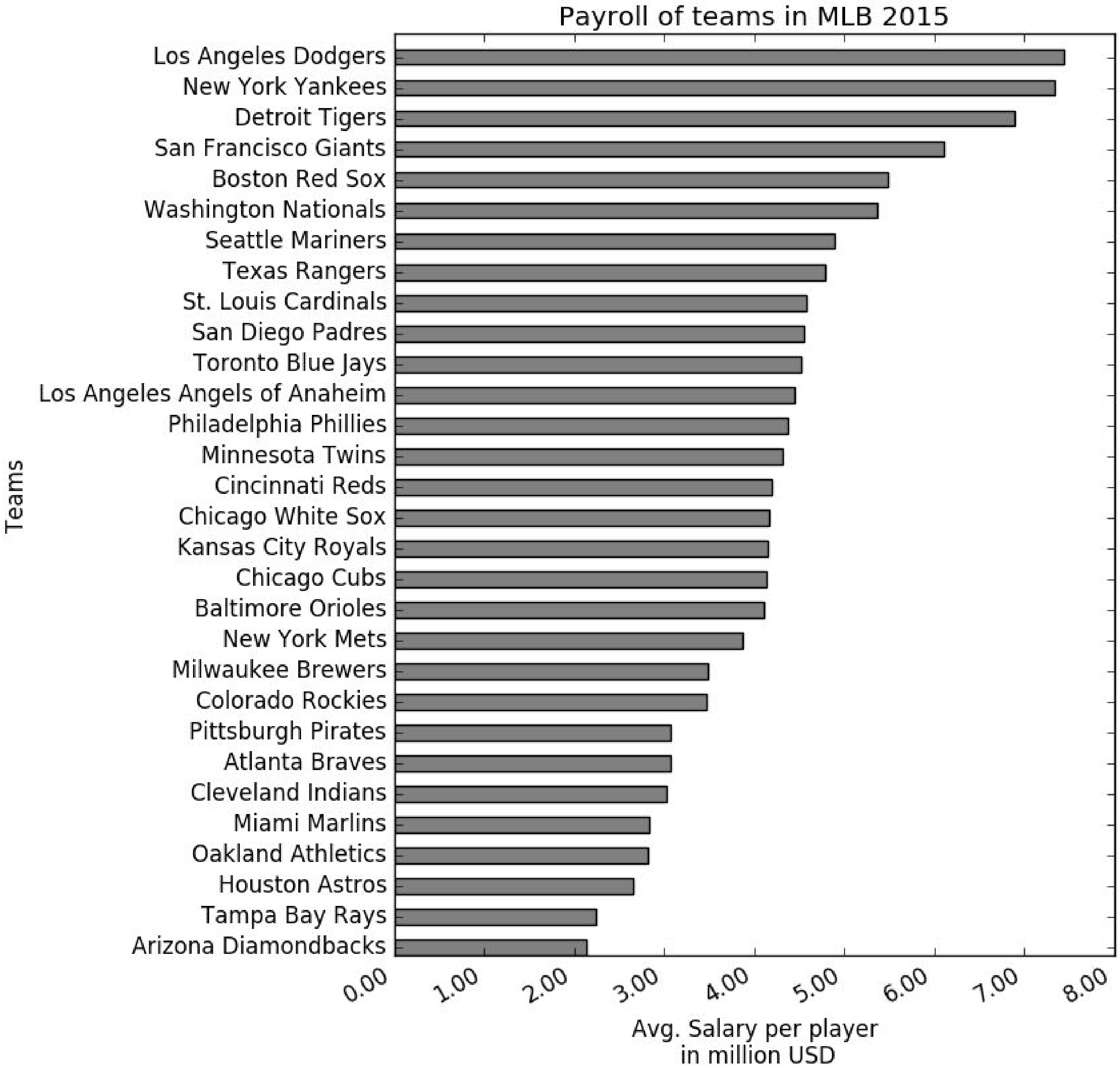
A similar visualization can be created to compare the payroll of the team. The average salary per player in the team is considered as a measure of how much the team has spent towards players' salaries. This visualization is shown in ***Figure 2***. From the visualization we can quickly infer that the team spending the most is **Los Angeles Dodgers** and the team spending the least towards payroll is **Arizona Diamondbacks**. Records for which the salary was either **missing or NULL** were not taken into consideration for computing the average salary per player.

Figure 1



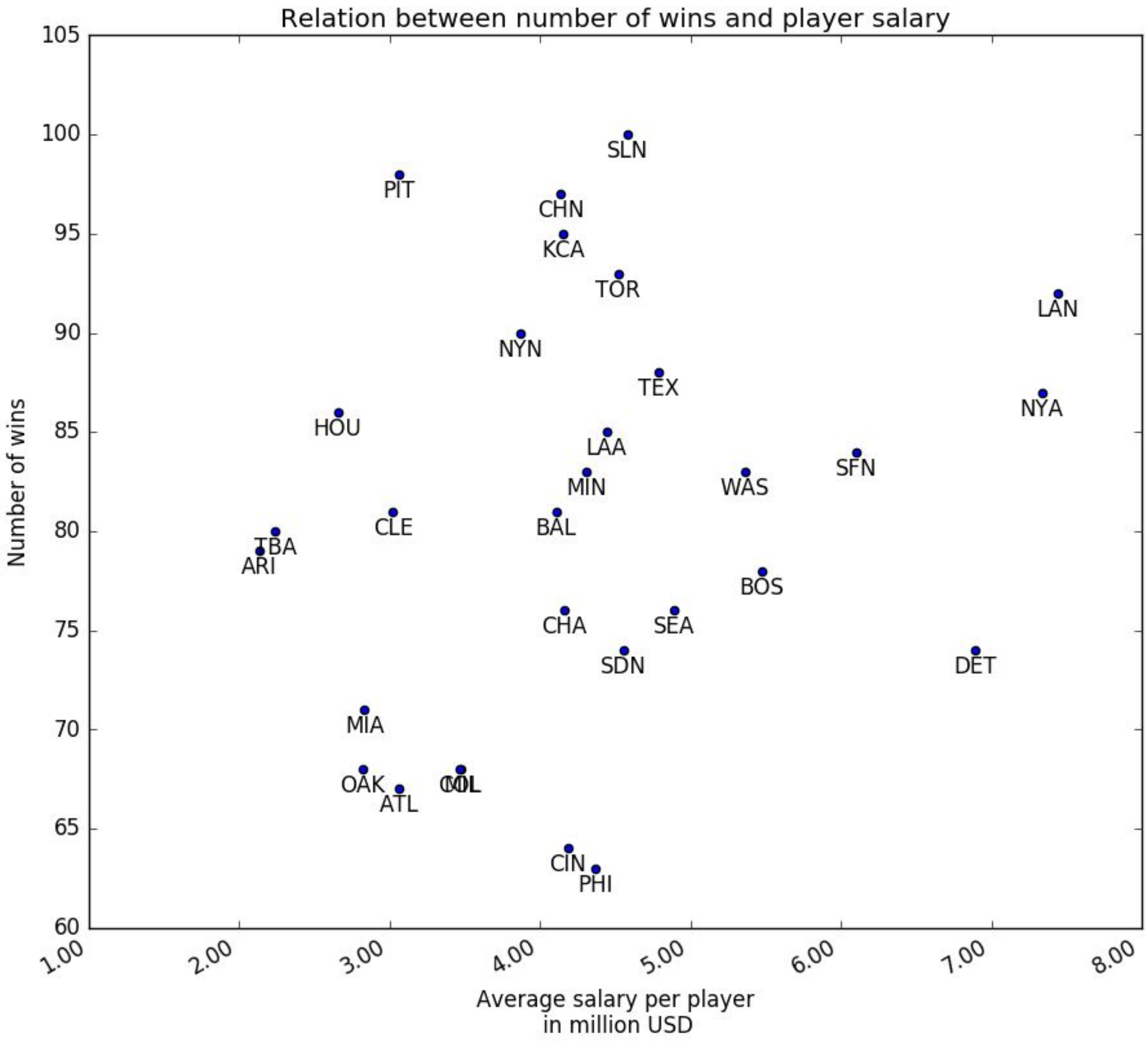
Performance of teams in MLB 2015
St. Louis Cardinals
Pittsburgh Pirates
Chicago Cubs
Kansas City Royals
Toronto Blue Jays
Los Angeles Dodgers
New York Mets
Texas Rangers
New York Yankees
Houston Astros
Los Angeles Angels of Anaheim
San Francisco Giants
Washington Nationals
Minnesota Twins
Baltimore Orioles
Cleveland Indians
Tampa Bay Rays
Arizona Diamondbacks
Boston Red Sox
Seattle Mariners
Chicago White Sox
San Diego Padres
Detroit Tigers
Miami Marlins
Colorado Rockies
Milwaukee Brewers
Oakland Athletics
Atlanta Braves
Cincinnati Reds
Philadelphia Phillies
Teams
0
20
40
60
80
100
Number of Wins

Figure 2


Payroll of teams in MLB 2015
Teams
Los Angeles Dodgers
New York Yankees
Detroit Tigers
San Francisco Giants
Boston Red Sox
Washington Nationals
Seattle Mariners
Texas Rangers
St. Louis Cardinals
San Diego Padres
Toronto Blue Jays
Los Angeles Angels of Anaheim
Philadelphia Phillies
Minnesota Twins
Cincinnati Reds
Chicago White Sox
Kansas City Royals
Chicago Cubs
Baltimore Orioles
New York Mets
Milwaukee Brewers
Colorado Rockies
Pittsburgh Pirates
Atlanta Braves
Cleveland Indians
Miami Marlins
Oakland Athletics
Houston Astros
Tampa Bay Rays
Arizona Diamondbacks
0.00
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
Avg. Salary per player
in million USD

We can create a scatterplot of Number of Wins against the team payroll to identify if there is any relationship between the team performance and payroll. If there is a strong correlation, then there should be an increase or decrease in number of wins as the payroll increases.

Figure 3

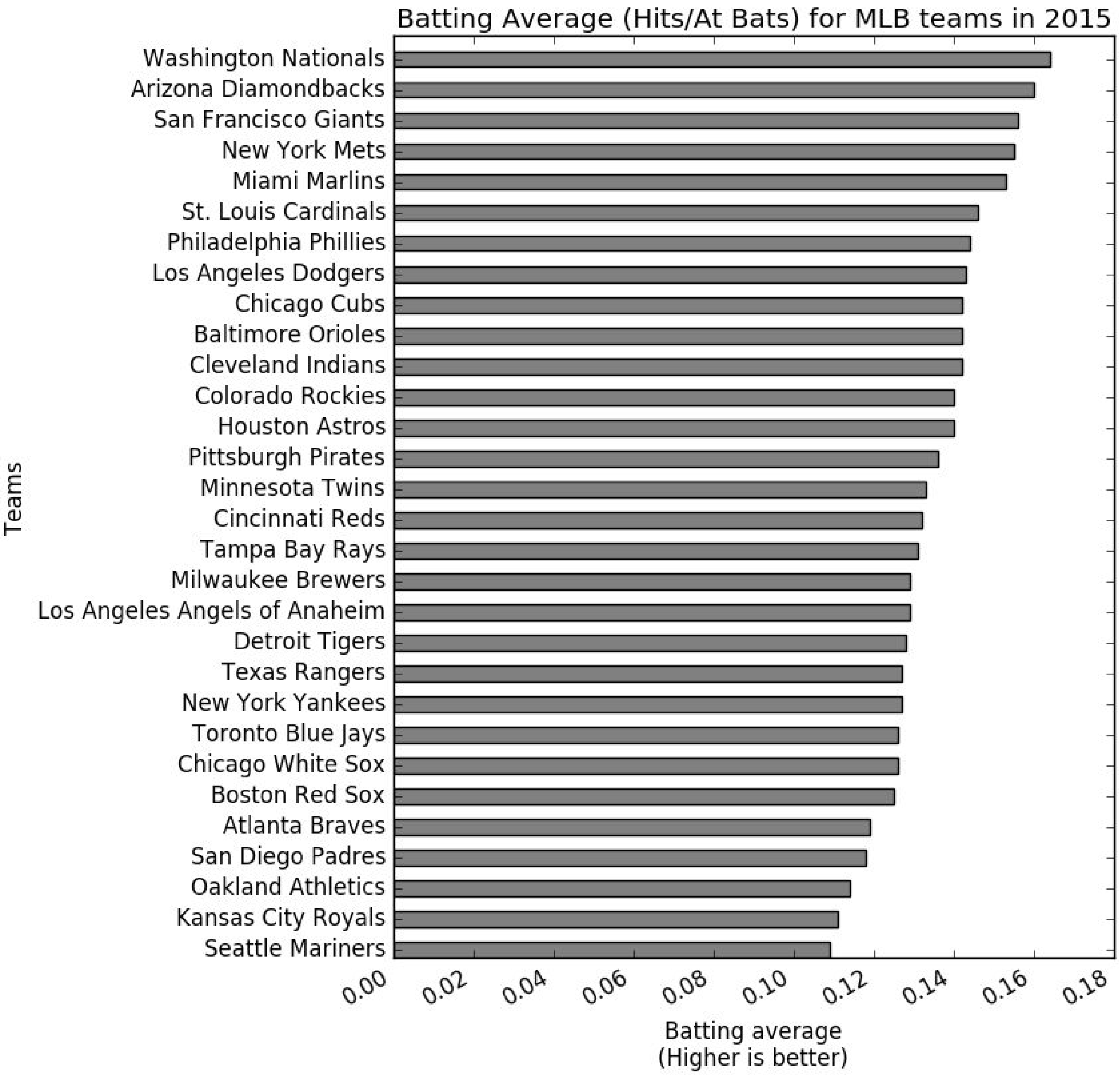


As we can see from the above visualization, there is no strong correlation between payroll and team performance. We can see that the team with the highest number of wins **SLN** has almost the same payroll value as the team with the least number of wins **PHI**. We can see that majority of the teams have a payroll value between 4 million USD and 5 million USD, and there is no clear distribution of the number of wins in this payroll range. We can also find many examples where teams with similar payroll have contrasting performances. **PIT**, **HOU**, **CLE**, **MIA**, **OAK** and **ATL** have payroll around 3 million USD but their number of wins range all the way from 65 to 100.

Relationship between batting performance and team performance

The batting average of the team is used as a measure of the team's batting performance. The batting average is computed by taking an average of the batting averages of all the players in the team, which is computed as **Hits/At bat** for each player.

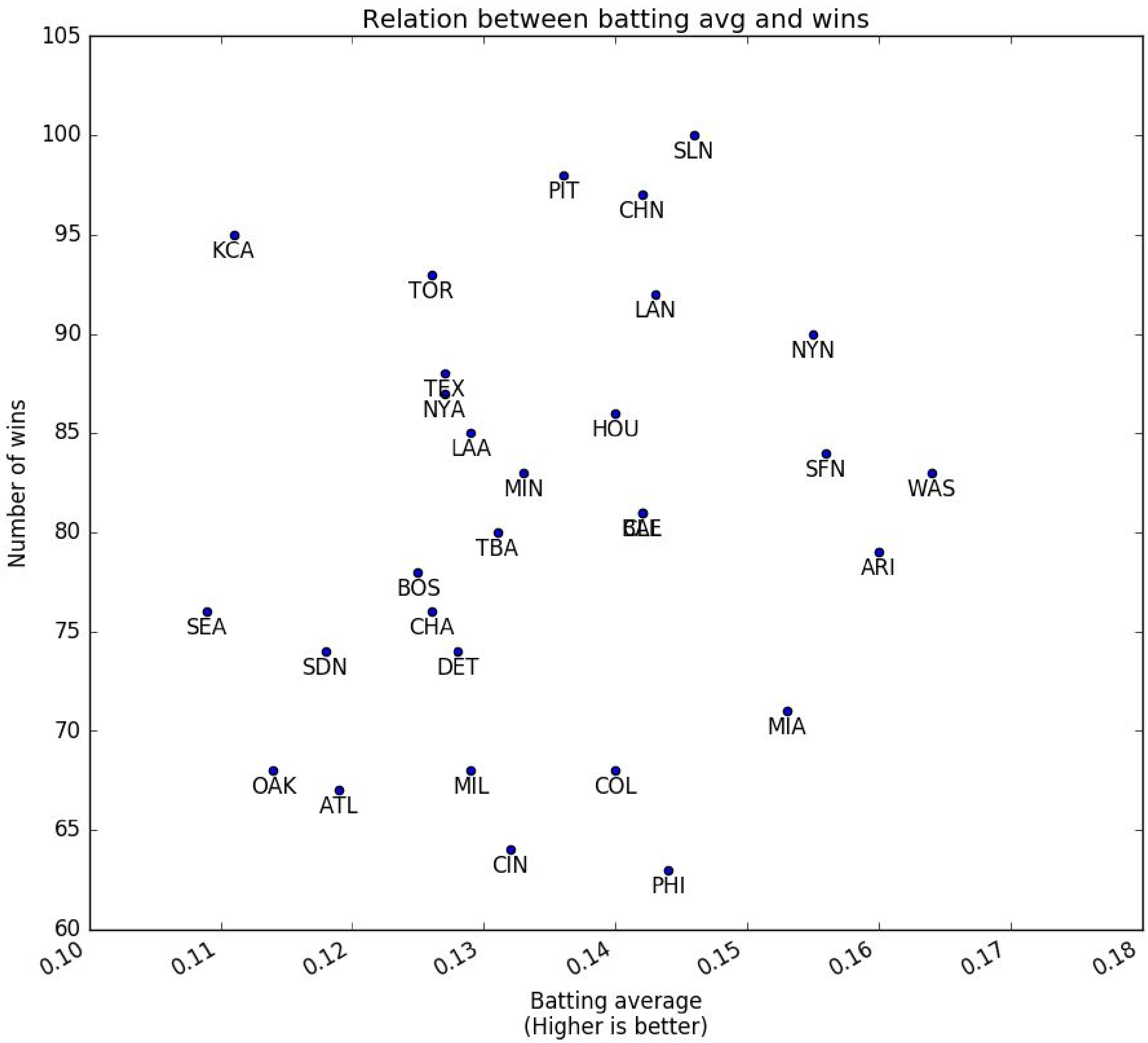
Figure 4



From ***Figure 4***, we can see that the team with the highest batting average is **Washington Nationals** and the team with the lowest batting average is **Seattle Mariners**.

We can identify at the relationship between the team's performance and its batting performance using a scatterplot.

Figure 5

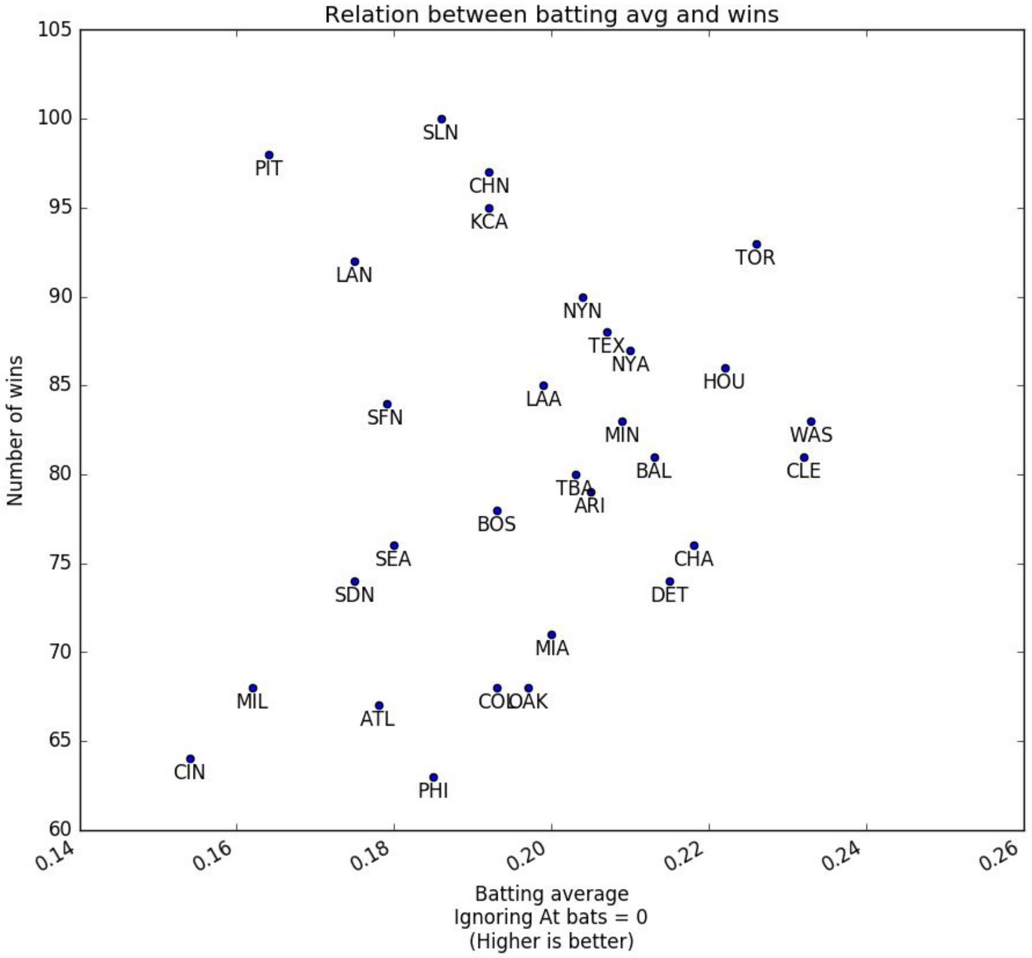


The scatterplot in ***Figure 5*** is very similar to the scatterplot we saw comparing team's performance and payroll. There doesn't seem to be any strong relationship between the batting average and Number of wins.

Ignoring At Bats = 0

In the above case, whenever we saw a zero value for At Bats, we assumed the batting average also to be zero as **Hits/At Bats** is not defined when **At Bats** is zero. This has the side effect of bringing down the overall batting average of the team. Instead of assuming the batting average as zero, if we simply ignore that player while computing the overall batting average for the team, we get the following scatterplot.

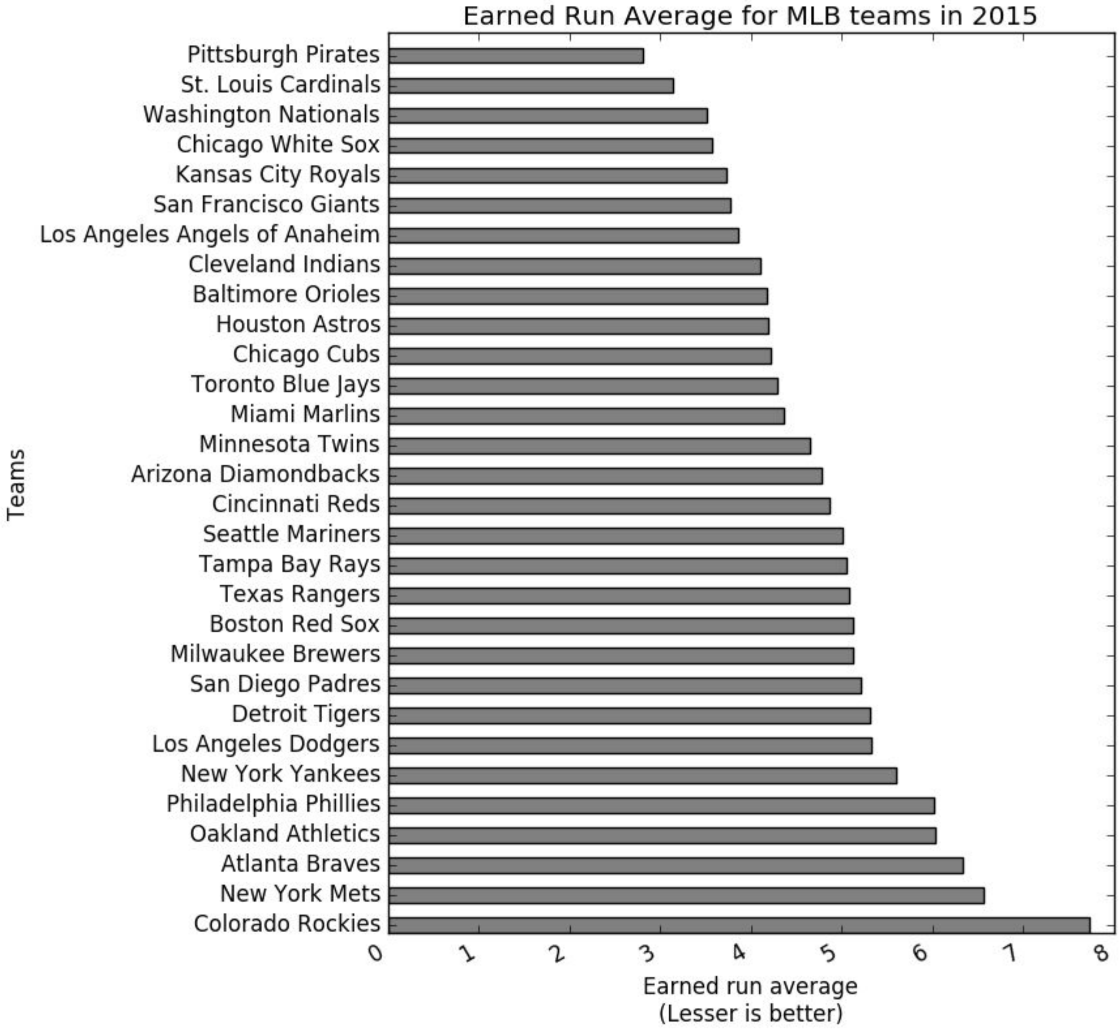
Figure 6



Now we can see a weak relationship between batting average and the number of wins for a majority of the teams. We can see a trend that the number of wins increases along with the team's batting performance. However, this is true only up to a certain level, if we look at the teams with more than 85 wins, we cannot see any meaningful trend. This agrees with conventional wisdom to an extent. A baseball team cannot win a championship just by having a good batting performance, the batting performance of a team does influence the number of games it can win but a team that only relies on batting performance cannot win all the games.

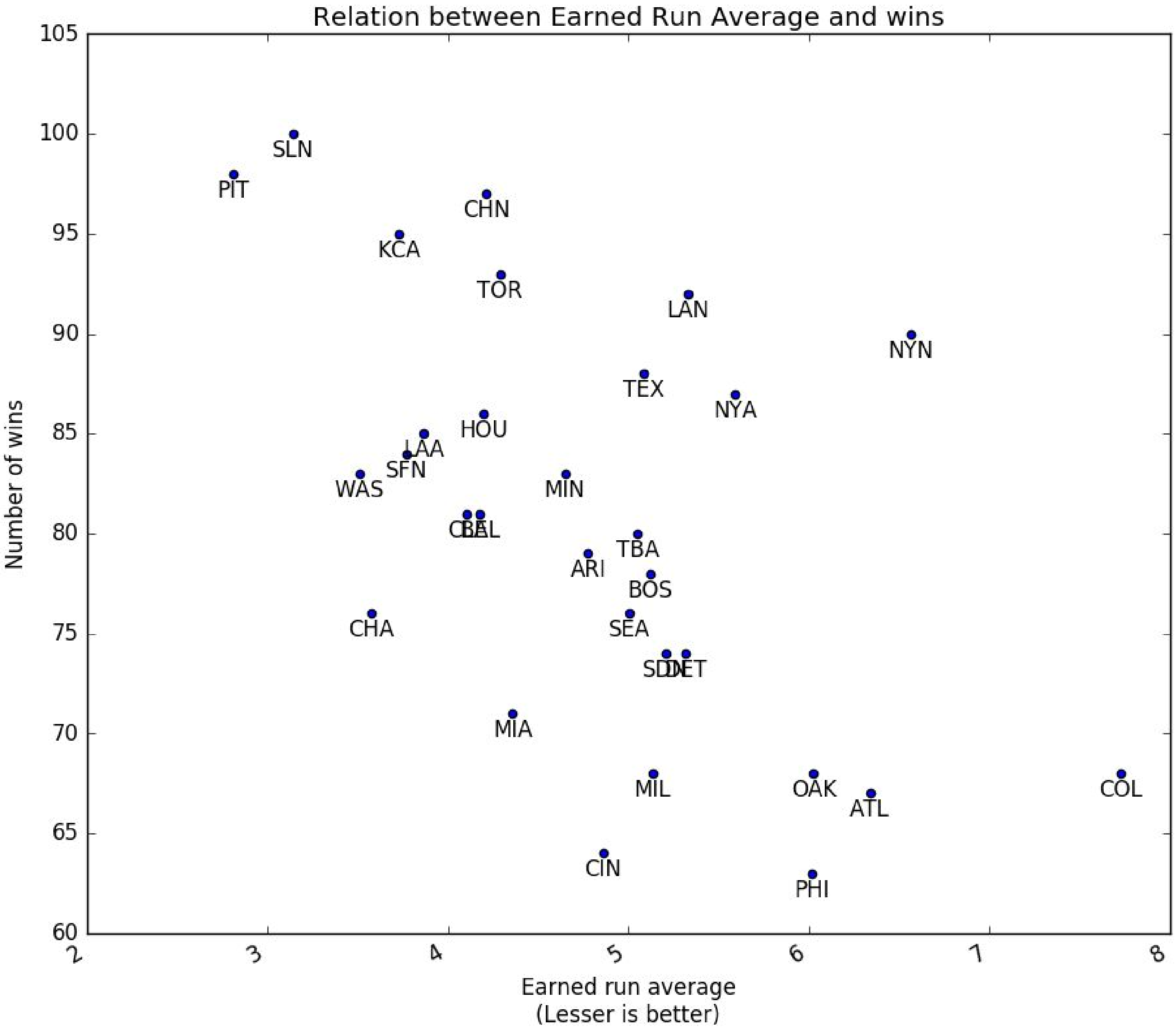
Relationship between pitching performance and team's performance

Figure 7



The team’s **Earned Run Average (ERA)** was used as a measure of the pitching performance. Lower ERA indicates better pitching performance and higher ERA indicates otherwise. Similar to the previous two cases, we can plot a bar chart (***Figure 7***) and a scatter plot (***Figure 8***) to learn more about ERA.

Figure 8



Similar to the batting performance, we can see a general trend in the pitching performance. The team performance seems to improve along with the pitching performance.

There are a few teams, **MIA** & **CIN**, whose pitching performance is as good as the top performing teams but their number of wins is not anywhere close. This could mean that there are other factors that seem to affect their performance.

How to become a successful team in MLB?

Not by increasing the payroll. From our analysis, we can clearly see that there is no direct relation between the payroll and the performance of the team. Therefore, with the given data we can conclude that increasing the payroll of the team will not make them more successful.

Would buying the best batters in MLB guarantee success for the team? Not necessarily, we can see from our analysis that the top performing MLB teams don't really have an exemplary batting average. Better batting performance can definitely lead to more wins but it might not be sufficient to clinch the championship

We can look at the stats of the top performing teams in MLB to identify the winning combination. Most of the top performing teams have less than average batting performance and exceptionally good pitching performance. Therefore, we can conclude there should be more focus on pitching rather than batting if you are trying to reach the top of the standings. However, if your team's current number of wins is less than 70, then focusing on both batting and pitching is necessary.