Baseball Analysis Project Report

This project is for Udacity Data Analysis course. There are 28 data sets on the “Sean Lahman” website. In this project, I focus on Wins and Salaries and compare two different roles in Baseball Game.

Abstract

The main question I want to answer in this project is ‘What is the best model for predicting Wins for each team?’; Base on Baseball rules, wins related to Runs, so I will focus on modeling runs base on variables we already have. First I will look at data in different ways and try to understand how wins and salary relate to each other. For this reason, I will ask 9 questions and finally I will develop a model for predicting runs.

In this project, I used these data sets from the mentioned source:

Master, Batting, Pitching, Salaries, Teams.

Main Questions:

For being familiar with these datasets, I will look at these questions:

1. How players’ salaries are distributed? is it normal or skewed?

2. How to measure teams’ financial efficiency base on their payroll?

3. Who is the most paid pitcher in 2015?

4. Which key characteristics of the most paid pitcher is significantly different from other pitchers?

5. Who is the most paid batter in 2015?

6. Which key characteristics of the most paid batter is significantly different from other batters?

7. What is the trend of changing salaries over the time?

8. How does number of Wins change over the time?

9. Is there any difference on average between Batters and Pitchers Salary?

10. What's the best model for describing Wins for each team? Which features are best relate to Wins?

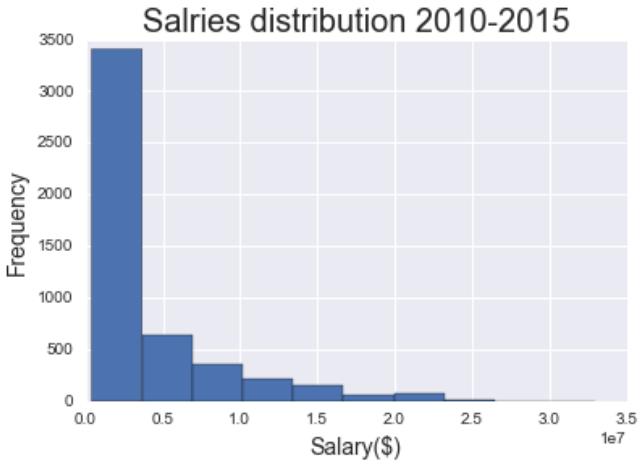
1. Salaries Distribution

For better understanding the salaries, I will look at salaries stats; I look at the whole salaries and found there are some $0.00 salaries; These should be wrong data; On the other hand, salaries is for the last 30 years, and it's not very reasonable to calculate the statistics of this variable over a very long period of time; So I sunset the data and limit it for the last 5 years. In this case I cleaned data from those $0.00 values and look at a more limited time frame; The results are here:

```
Salaries Statistics 2010 to 2015
count    4951
mean     $3,671,903
std      $4,909,905
min      $400,000
25%      $504,000
50%      $1,300,000
75%      $5,000,000
max      $33,000,000
```

And here is the histogram of Salaries from 1985 to 2015:

Maybe the reason of skewness is that we are looking at 30 years; for making it more clear, I will look at the most recent 5 year:



It seems the salaries distribution is skewed; again, for considering it and making it more clear, I will calculate Pearson's second skewness coefficient (median skewness):

```
Yearly Salaries Median Skewness is: 1.39159693472
```

It means that Salaries are Positively Skewed. Because the median skewness is positive.

2. Teams Payroll vs. number of their Wins

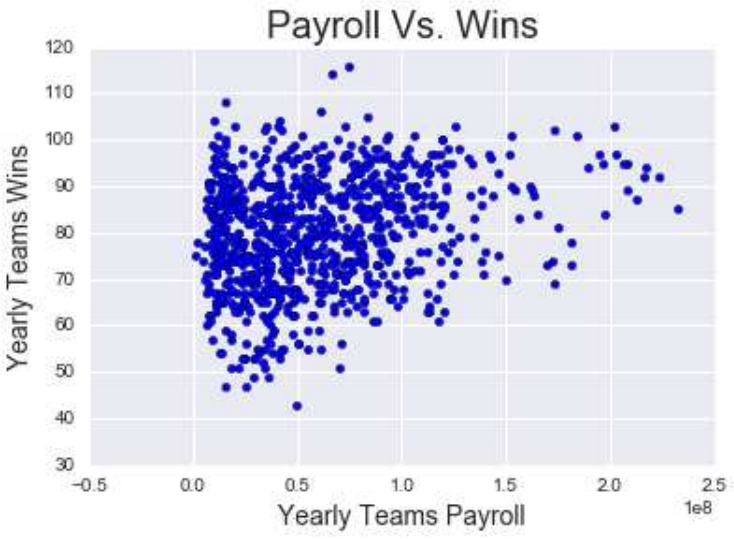
Baseball teams goal is to get Wins and on the other hand they should pay salary to player; At a financial point of view they try to get more Wins and pay the whole payroll less. There is a dilemma like other business situations.

For looking at salaries related to teams, I join Teams data frame and Salaries data frame. It creates some NAN values and I found out the reason is the Salaries data set has the data for years after 1985 and Teams data set years start from 1871; For cleaning this join table, first I subset the Teams data frame to years 1985 and after, so these two data sets being compatible.

Another data cleaning step is to summarize the data I need for this part; So, I choose only the columns I need from Teams data frame and put it in my new data frame. I didn't change the main Teams data frame and I create a new data frame as a subset of the main one: *teams_after_1985*.

I want to look at each team in an arbitrary year as a point in my data. So, I changed the index to a compound index that contains yearID and teamID; Again, I didn't change the main data frame and only change the sunset I made for this analysis.

Let's look at the relationship between Payroll and number of Wins each team has in each year after 1985:



Base on this scatter plot, most teams spend an average money and get an average Wins. Efficient teams are ones which spend less money and get more Wins. For modeling this concept, I am defining a criteria base on Wins and Payroll. The criteria can be the ratio of number of Wins to total

Payroll and multiply it to 10^6 to have more reasonable numbers. Because the ratio without this multiplication has a factor of 10^-6;

Then I define a function to find the most efficient team in each year base on this criterion. In the code, I run this function for 2000 to 2015 years and the result is hereunder:

```
2000    (MIN, 4.17688186688)
2001    (MIN, 3.52258599254)
2002    (OAK, 2.57473177732)
2003    (TBA, 3.20937340805)
2004    (MIL, 2.43384129175)
2005    (TBA, 2.25748336361)
2006    (FLO, 5.31642981290)
2007    (TBA, 2.73592140444)
2008    (FLO, 3.85117942370)
2009    (FLO, 2.36194820003)
2010    (SDN, 2.38099647348)
2011    (TBA, 2.21661594311)
2012    (OAK, 1.69759357082)
2013    (HOU, 2.85064307154)
2014    (HOU, 1.99337629534)
2015    (ARI, 1.27761425753)
```

Base on the above results, we can find out which teams are the most efficient in each year, refer to their total payroll and number of wins they have in that year. As an example, we can see in year 2002 the Oakland A's is the most efficient team; It was in the Billy Beane era and starting using analysis for recruitment. In 2002, the Athletics became the first team in the 100 plus years of American League baseball to win 20 consecutive games. The nice point is that they got this record with a very low payroll. And it means exactly the efficiency I mean.

It's not necessary that a team with a high efficiency ratio, has a high rank in a championship. It only means that base on the money they spent, they got relatively high wins.

3. The most paid Pitcher in 2015

I want to know who is the most paid pitcher first and then find out which characteristics he has for being in this level of salary.

First I merged Pitching data frame and Salaries data frame. The base for this merge should be the Pitching data frame, and it should be done on 4 variables: ‘yearID’, ‘playerID, ‘teamID’, ‘lgID’. There are many players that not exist in Salaries data frame for specific year and I got many NAN values in Salary column. These NAN don’t effect this part of analysis, because the most paid pitcher salary will be find between existing numbers. Then I found the most salary in the merged data frame.

Base on code I wrote, here is the most Pitcher:

Clayton Kershaw.

4. Key characteristics of the most paid Pitcher

For understanding the key characteristics, I calculate the z-value for all columns and base on what we know about normal distribution, I will look at the z-values related to the most paid pitcher that are more than 3. It means that the most paid Pitcher has a better performance than 99.9% of the population. Base on the code, I think the key factors are these ones:

SHO, CG, BK, SO, W

5. The most paid Batter in 2015

The explanation is like topic 3; And the answer of this question is:

Alex Rodriguez

6. Key characteristics of the most paid Batter

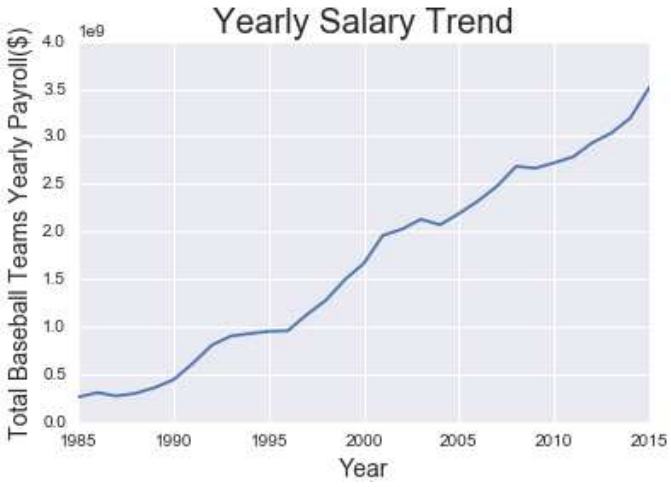
Again, the explanation is like topic 4 and the key characteristic for him are:

HR, RBI, BB, SO, SF, GIDP

I should mention that in the current topic and topic 4 I don't mean these characteristics are the cause of big salaries, I only realize there is a positive correlation between big salaries and these performance values.

7. Yearly Salary Trend

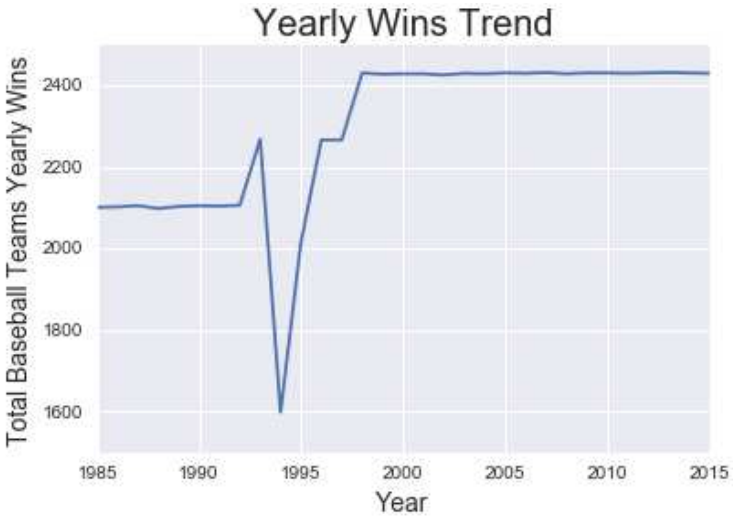
In this question, I am looking at the total salary paid to players over the past 30 years. This question can be considered as part of baseball economics. I mean how the whole money that paid to all players changed over the last 30 years.



It means that the whole baseball payroll folded 13 times over the last 30 years.

8. Number of Wins Changes Over Time

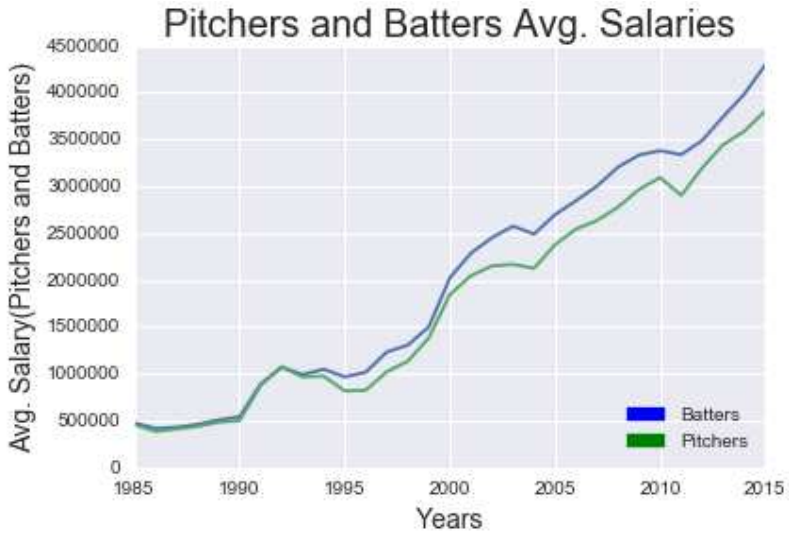
Now I am looking at the total number of wins in the last 30 years;



It means total number of wins has an absolute growth from 2100 to more than 2400; There is an interesting year in this graph:
Year 1994. There should be a reason for dropping total number of wins in that year.

9. Difference between Batters and Pitchers Salaries

Another question worth to think about is the difference between Batters and Pitchers salary on average. I mean is there any difference at all and if yes, what is this difference?



Base on this plot, it seems like the average salary for batters are more than Pitchers. For testing this arbitrary conclusion, we can calculate the ratio of these two series and look at the average ratio over the years; The average ratio of these two salaries is: 1.11;
It means that the average Batters salary is 11% more than the average Pitchers salary in this 30 years.

10. Modeling Team Wins

Base on baseball rules, the winning team has more runs. So, the most important criteria for modeling Wins is Runs. I am looking at correlation between Runs and other features for teams and find out the most correlated characteristics. I need these features to model runs correctly. I will use a linear regression model with those features.

Base on correlation coefficients after running the code, those numbers suggest that Runs has a strong correlation with R and H; R is Runs itself and it's obvious any variable has a 100% correlation with itself. H is Hits and again it's an obvious correlation; Because Hits are the major cause of Runs. The other correlations are not very high, so I will define some new variables and look at their correlation with Runs; These new variables are:

BA: Batting average

OBP: On Base Percentage

SLG: Sluggish Percentage

And do the correlation coefficient calculation between Runs and these new variables;

Based on these correlation numbers, we conclude that all three new variables have a high correlation with Runs. Now is time for modeling. For linear regression modeling, I will use *statsmodels* library in python.

We know that the best model would be the model with maximum R-Squared;

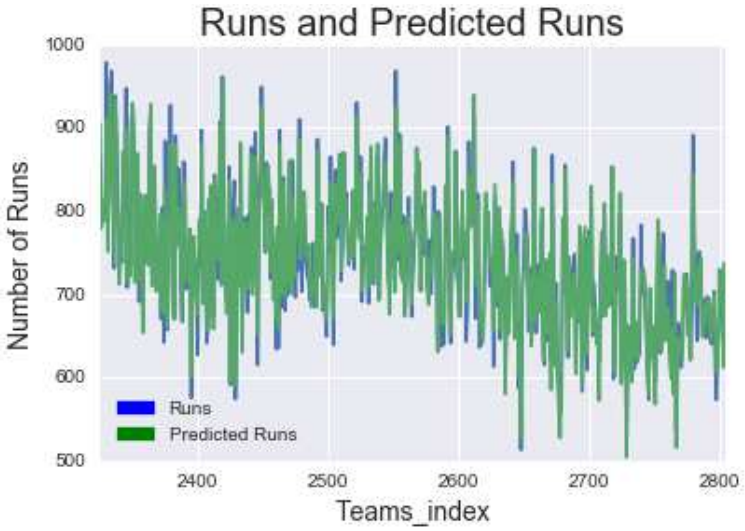
The first model for Runs is using only BA as feature. Because BA was one of the most powerful predictors for Runs. After this simple model, I tried several other combinations of most correlated features with Runs. The second model uses three features: OBP, SLG and BA; The third model uses two features: OBP and SLG;

The R-Squared for the second model is less than two other models, So I think it’s the best model for predicting Runs and there after predicting Wins.

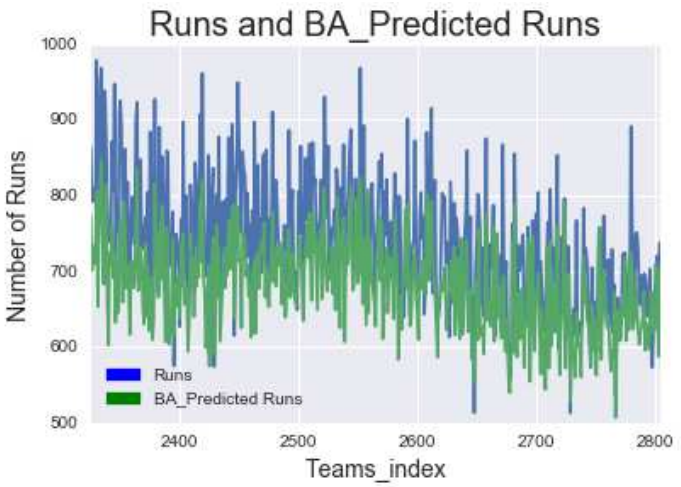
The Second Model for Runs is as follows:

*R = -864.83 + 2676.15 * OBP + 1741.78 * SLG + 14.08 * BA*

For visually testing this model, I plot the real Runs and Predicted Runs base on this model; The result is as follows:



The First Model for Runs is as follows: $R = -627.94 + 5010.09 * BA$
The prediction vs. values we have for Runs are as follows:



Based on these graphs and overlaps between Runs and Predicted Runs, it's clear that the first model is far better predictor for Runs.

Conclusion:

In this project, I looked at Wins and Salaries of Baseball teams. I tried to get a sense of data and explore it in different aspects at the first part of project and at the second part I analyzed some questions and I elaborate two of them here:

1. *Is there any difference between Batters and Pitchers salaries?*
   - Hypothesis: These two salaries are equal on average.
   - Analysis: I calculate the average salaries for both group over the last 30 years. Now I have two series: one is average salaries for Batters and the other one is average salaries for Pitchers. For each year, I calculate the ratio of these two averages and create a new series that contain a ratio for each year that shows average of Batters salaries to average of Pitchers salaries in that year. Finally, I calculate the average of these ratios.
   - Result: This ratio is more than one and it means I reject the null hypothesis and there is a difference between Batters and Pitchers salaries. On average Batters salaries are 11% more than Pitchers salaries.

   Note that I didn't use a sample and I used all the data for the last 30 years. So, there is no need to use t-test for testing the hypothesis. Base on this data set I can say for the last 30 years there was a 11% difference between Batters and Pitchers salaries on average. I don't say anything about before 1985 or after 2015. Because I have no data in these time frames.

2. *What are the key characteristics of the most paid Pitcher in 2015?*
   - Hypothesis: The most paid Pitcher have some extra ordinary features correlate with big salary
   - Analysis: I calculate the z-score of all features related to Pitchers included their salary. Next I looked at z-scores of the most paid Pitcher. The z-score related to his salary is more

than 6; For finding important features I consider all features z-scores and found out there are some features related to him that has the value of more than 3.

- Results: The key characteristics that correlate with a very high salary in pitchers are:
  SHO, CG, BK, SO, W
  And between them, SHO (Shout Outs) has the most significant correlation. Its z-score for the most paid Pitcher is 7.9; And the probability of getting this number by chance is less than 0.01%;

- Limitation: For all the results, I discussed above, there are two kinds of limitation;
  - one is related to datasets itself:
    1. The Salaries dataset dates are after 1985 and all the result are limited to this time frame: 1985 to 2015.
    2. There are some zero salaries in Salaries dataset. It shows potential anomalies in the dataset. So maybe we have this kind of wrong data entries in other datasets and they will provide wrong results.
    3. The variables in all data sets are the variables that have a role in measuring player's performance. Maybe with the new era of sensors we can measure some other aspects of performance very soon. But now we are limited to the characteristics that are accepted as performance factors and measured.

  - The other is related to analysis process:
    1. First limitation is the domain knowledge. I am not a professional person in Baseball domain. So, it's possible that some important characteristics are absent in my analysis.
    2. In modeling part, I research and fount the features that are well known for modeling runs. Maybe there are some other features for this purpose that I am not aware of them.
    3. In second question, I focused on the payroll vs. wins and I defined a financial efficiency ratio. It has its own limitations like every other parameter. It doesn't show the most successful team in each year. Maybe one team has a good financial efficiency in a specific year and it is the last one in championship ranking in that year. By this criterion I only compare they money they spent vs. number of Wins they got.
    4. In the last topic, I modeled Runs with some parameters. This model has an error like every other model. The best model I choose between those three ones, has the least modeling error; By it doesn't mean it's the best model overall. Maybe a nonlinear model can describe the data we have better. But I use a linear regression model.
    5. And finally, the important notice: "correlation doesn't mean causation".
       All the results are mentioned above are show correlation between two or more parameters and I didn't use a Causal Experimental approach for finding cause of the questioned features. All result showed correlation only and there is no cause and effect in none of the results and answer to questions I mentioned.

Future research direction

1. Modeling salaries base on features as performance keys.
2. Modeling the value of a player regardless of his salary.
3. Predicting likelihood of winning a game by one team based on the performance keys of all players in both teams in that game.