Major League Baseball Exploratory Analysis

SI370

Winter 2017

Viviana Hernandez

Motivation

Baseball has always been a huge part of my life. Ever since I was little girl I have been going to games with my dad. Over the years there has been an increasing release of public baseball data. There is even a whole field dedicated to the analysis of this data called Sabermetrics. Using what I have learned in SI370 and what I have observed and know about baseball I want to see if I can find answers to questions that I have thought about throughout the years.

Data Source

For this project, I used a dataset that I found on Sean Lahman's website: http://www.seanlahman.com/baseball-archive/statistics/. The data is comprised of different baseball statistics and player, team, and manger information. The data was originally an SQL database, but fortunately there were CSV's that could be downloaded for use. Out of 24 CSV's I downloaded 5. Master csv contained player's name biographical information; Salaries csv contained player salary data; Teams csv contained yearly salary stats and standing; Batting csv contained batting statistics; Appearances csv contained positions a player appeared as. I can't give an exact number of rows/columns because my dataset was very dirty I had to do a lot of cleaning for several of the questions. However, here are some (there are many more) of the variables my data contained:

- Player weight (in pounds)
- Player height (in inches)
- TeamID (corresponded to Team in league)
- YearID (year)
- Salary (USD player salary)
- Team Wins (W)
- Team Losses (L)
- G_1B(games as 1st base) several like this but for other positions
- Home Runs
- Team Runs
- Team Runs Allowed
- World Series Winners

Research Questions and Methods For Report

Question 1: What is the distribution of weight in the MLB?

- Distributional analysis: Histogram
- Distributional analysis: QQ plot

Question 2: Have players' weight increased over time?

- Time Series Analysis
- Time Series (checked rolling average in python notebook, but trend is clear)

Question 3: *Can we classify players' position by weight?*

- Classification
- Classification related evaluation: Confusion matrix
- Classification related evaluation: Histogram (evaluating error in classification)

Question 4: *Can we predict team wins from their run differential?*

- Regression
- Outlier Analysis

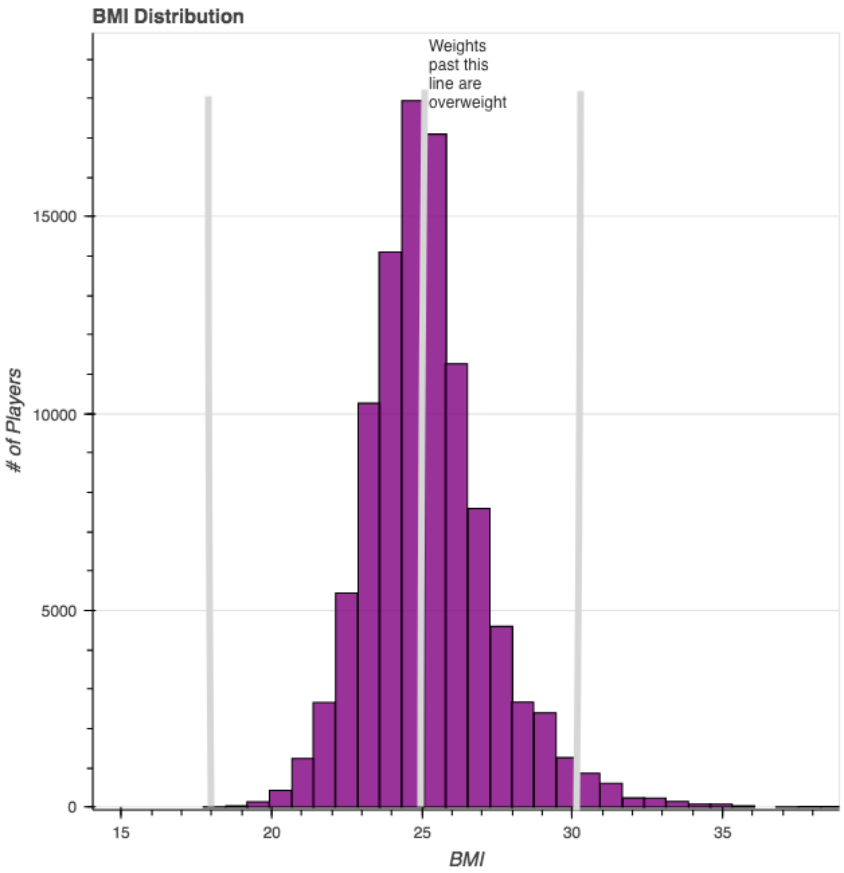
Questions 5: *Do teams with higher payrolls win more world series?*

- Distribution analysis
- Chi-square Test

Analysis and Results

Question 1: *What is the distribution of weight among players in the MLB? More specifically, their BMI (Body Mass Index)?*

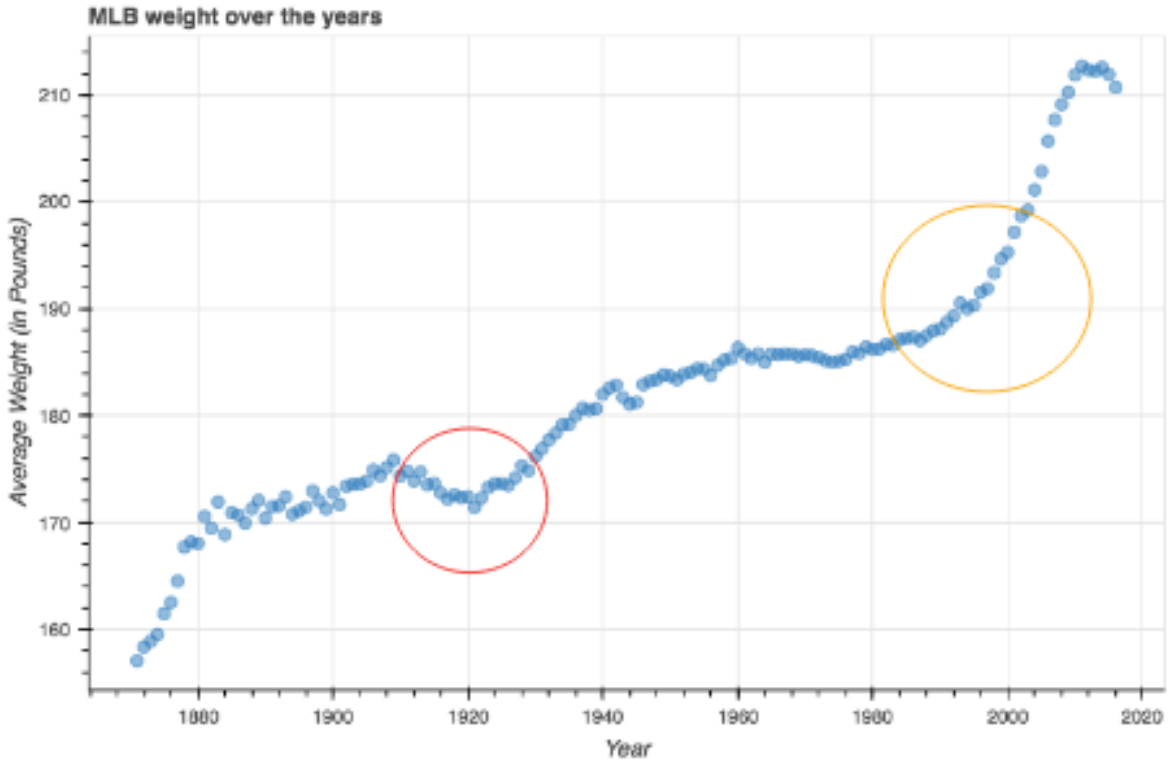
Before calculating BMI, I thought it was important to look at the variables that would comprise BMI and what their distributions looked like. I first examined weight's distribution by plotting a histogram, the histogram appeared to have a bell shaped curve, so I deemed it as normal. To confirm normality, I ran a QQ plot on weight and when I examined it there was no strange pattern and all the dots followed the line appropriately which confirmed my histogram's normality. Similar to weight, I performed the same analyses on height and found normality to be present. I then went on to calculate BMI (can see BMI calculation in notebook). I then created a histogram of my BMI. Interestingly, I found that 27% of



players were considered overweight (BMIs past 25 are overweight) , 2% of players were considered obese(BMIs under 18.5 are obese), and .025% were underweight(BMIs under 18.5 are underweight). This was shocking to me and led me to question the weight of players over time.

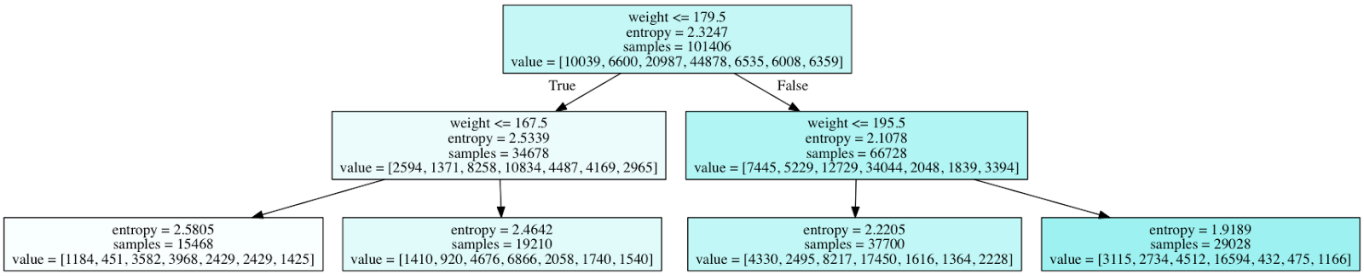
Question 2: Have players' weight increased over time?

In order to see the increased weight over time I knew it would be beneficial to use a time series. The data stems from 1871 to 2016, so to better interpret and plot my data I created a data frame with the average weight of baseball players by year. My first plot showed a trend and to confirm what I saw I plotted a rolling average which clearly followed my time series (this can be seen in my python notebook). My time series visualization is interactive and I can see the data from each point, but for this report I have add a static image below. I found that the average weight had increased from 157 pounds in 1871(the first point) to 210 pounds in 2016 (the last point). There are interesting points to note in my visualization. The dips in the 1920 (red circle) coincides with World War I, the rationing of food and limited supplies, may be a reason for this dip. Then there is the increase in the 1980's to 2000's (orange circle) this is considered the Steroid Era of baseball and I would attribute such an increase in weight to steroids.

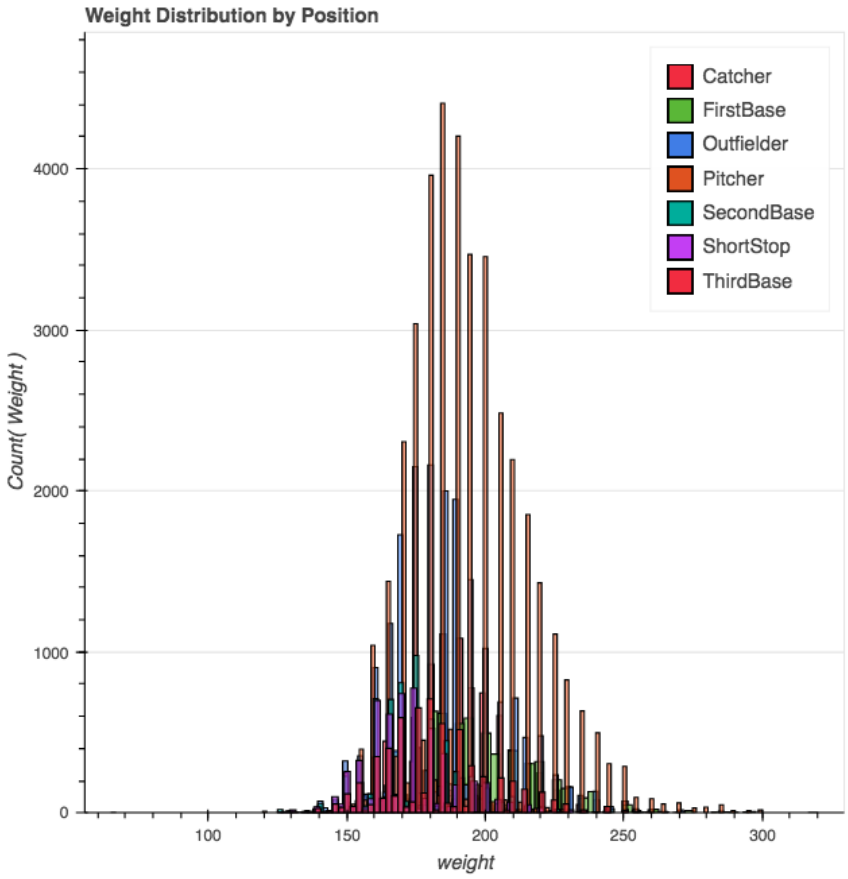


After examining this time series, I wondered if weight played any sort of factor in the position a player played.

<u>Question 3:</u> Can we classify players' position by weight?



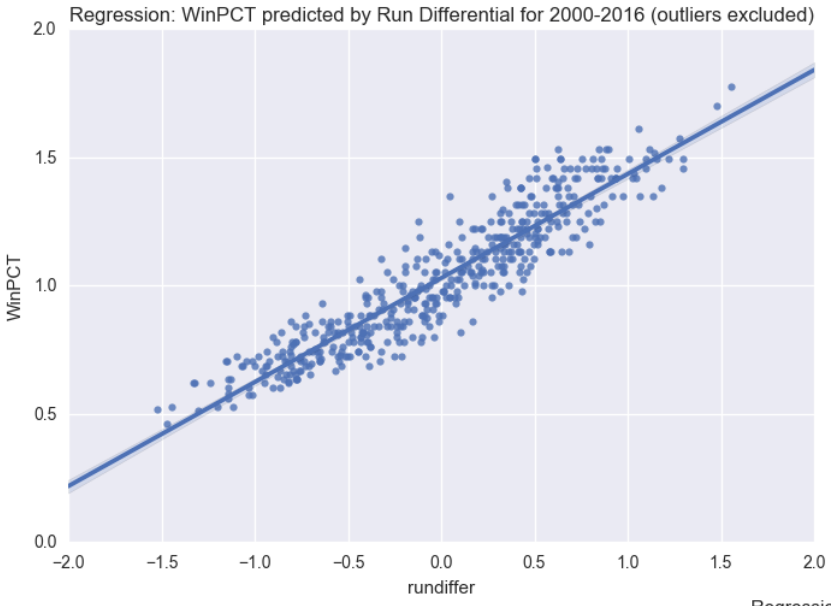
In order to tackle this classification problem, I had to do some data cleaning. Using the appearances csv I created a column of labels for position. For each year, there was data on how many games a player had played a given position. For every given year, I assigned the player the position they had played the most in that season. While there are players who play more than one position, most players specialize in one position. Once I had created my labels I decided to run a decision tree classifier using a player's weight to predict their position. The accuracy score of my model was .443, meaning I had a poor model. When I examined my confusion matrix for the model it was apparent that my model was guessing every time. After examining the predicted labels, which were mostly "pitcher" I decided to do more investigative work. I made a histogram of weight distribution by position and it was clear to me why my classifier failed. The positions' weight distributions all heavily overlapped each other, thus my classifier was having trouble

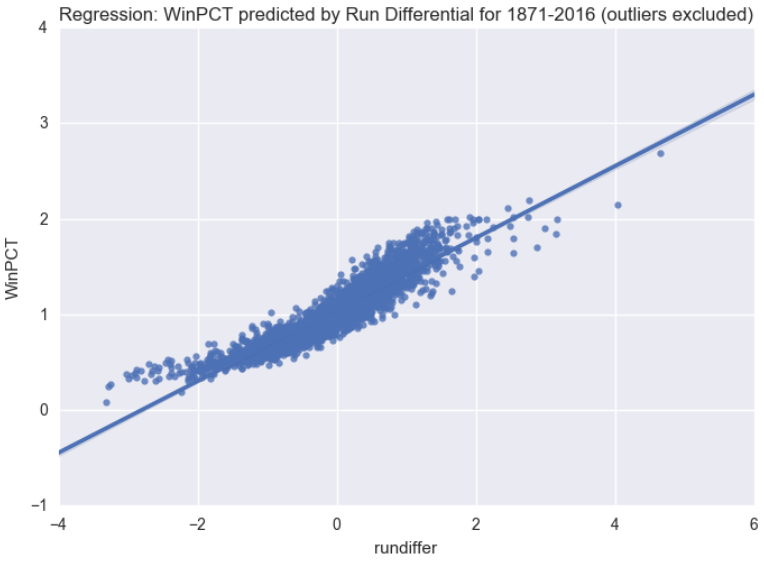


finding meaningful ways to cut the data by weight.

Question 4: Can we predict teams' winning percentage from their run differential?

To answer this question, I ran a simple linear regression. However, before running the regression I had to create the winning percentage and run differential. To create a team's winning percentage, I divided the teams wins by its losses. To create a team's run differential, I subtracted the runs scored by a team minus the runs allowed by the team and divide that by the number of games played. I divided by the number of games to normalize for seasons in which there may have been less games or perhaps missing game data. My first regression was a simple linear regression with rundiffer as a regressor and WinPct as the regressand. My adjusted r-squared was .84, which means that 84% of the variation in WinPct was explained by the model; also run differ was a significant predictor in this model. I did do outlier analysis on this model and when I removed outliers my adjusted r square increased to .86. This initial model only examined the 2000-2016 seasons. Frankly, I had no particular choice for those years expect for my understanding of teams and winnings, so I decided to run it again with all the data, that being 1871 through 2016. Interestingly, when I ran this regression and once again preformed outlier analysis using dffits, and removed outliers from the regression, I found similar results to my subset (2000-2016) regression. The adjusted r-squared of this model was .84, once again 84% of the variance in winning percentage could be explained by the model. This makes sense, a team that

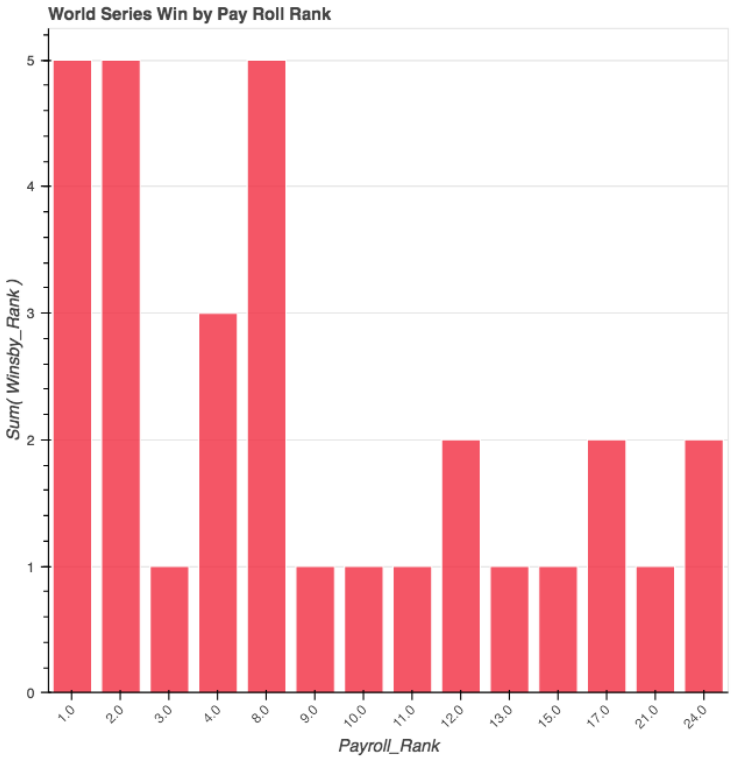




scores more runs than the runs they allow is more likely than not to win the game then lose the game, thus resulting in a higher winning percentage.

Question 5: Do teams with higher payrolls win more world series?

In order to perform this analysis, I had to find the payroll of each team for every year. Using the salary data and team data I created a new data frame which held the payroll of each team for every year and the payroll rank. I calculated the payroll rank by summing players' salary by team for a given year, then used rank to rank the payrolls in that year. This data frame only contained world series winners. Limited to the salary data I had, the data frame I made only included the years 1985 to 2016. I took counts for each payroll rank that was a world series win. These counts can be found in Rank_Wins data frame. It contains "payroll_rank"(rank of payroll) and wins_byrank (the number of world series winners who had that payroll_rank). Visualizing this data frame resulted in a distribution analysis. Looking at the bar chart we see that the higher payroll ranks (1,2,3,4,5) have a greater sum of world series wins. So, the closer your payroll rank is to 1 the better your chances are at a world series win. I really wanted to solidify this finding so I ran a chi-squared test. I got a p-value of .34, which is greater than .05. Thus, I fail to reject the null that payroll rank is independent upon world series wins. This reduces the validity of my findings, however after more examination, this chi-square finding may be due to the small sample size of world series winners. However, I do believe that the bar chart is still meaningful because it does give a better understanding of world series wins in terms of payroll.



Future Steps

If I had more time to clean the data I believe I could have performed a more in-depth analysis. This data set was messier than I expected. I would have liked to look into

salary and homeruns, both of these variables needed to be normalized in order to be deemed useful for the regression analyses I wanted to run. I found the classification aspect to be really interesting, I would have focused on predicting positions with more features. Also, I would have looked into my time series analyses more and preformed cross correlation between position time series.