Analysis of MLB Pitching/Batting from 2013-2019

I have spent the past year completing courses in Machine Learning and Data Analysis in the R language. Once I felt solid in both coding fundamentals and statistics, I began creating projects showcasing my passions. I included all pertinent code in my github repository.

For this project, I used a dataset titled ‘Pitching’ and one titled ‘Batting’ from the Sean Lahman Baseball Database (Lahman package) in R. I used these datasets because they included extensive statistics for pitchers and batters alike. Next, I imported a csv file I obtained from mlb.com which I had cleaned in order to only focus on players’ salaries from 2016. My goal was to look at both hitting and pitching data from 2013-2015 and 2017-2019 in order to see if there was a statistical correlation between output and salary paid. Specifically, I had multiple questions about this data that I hoped to concretely answer.

1. Of the pitching statistics, (Earned Run Average, Strikeouts, Home Runs, Wins, Saves, Opponent’s Batting Average), which (if any) had a clear correlation with a pitcher’s salary?

I hypothesized that Strikeouts and Wins would show to be directly correlated with a pitcher’s salary. Fundamentally, it makes sense to pay pitchers a higher salary who produce more outs as well as overall wins for their team. That being said, receiving a Win for pitching a game is directly related to the remainder of the team’s performance (batting, fielding, relief pitching).

2. Of the batting statistics, (Runs Batted In, Home Runs, Hits, Doubles, Triples, Walks, Strikeouts, Ground into a Double Play, Hit by Pitch, Stolen Bases, Caught Stealing, Games Played, At Bats), which (if any) had a clear correlation with a batter’s salary?

Initially, I believed that a batter’s number of home runs would be clearly correlated with his salary. Upon further thought, I hypothesized that RBIs, GIDPs, and Home runs would be related to a batter’s salary. Home runs are obvious because of team owners wanting to secure players who tend to hit balls out of the park. GIDPs also are a reasonable choice because those aforementioned ‘hard home run hitters’ statistically are more likely to ground into double plays. In the past decade with the wide usage of sabermetrics, I hypothesize that RBIs will be directly correlated with salary. More team owners have begun to value the daily contribution of players (hitting runs in) just as much and maybe even more so than the occasional long ball (and the outs that come as a result). An apt real life example of these two phenomena is Joe Mauer vs Adam Dunn.

3. Are these correlations greater in the three years prior (2013-2015) to the 2016 salary or the three years afterward (2017-2019)?

This question is especially prevalent with the MLB lockout currently taking place. An idealist might believe that a player who had three phenomenal seasons (2013-2015) would be paid a high salary (2016) and then continue to excel (2017-2019). As we’ve seen in recent history, much of baseball is random and MLB has shown to pay high performing players a higher salary after a couple of seasons even though it is statistically very likely they will drift down to the average in the years that follow.

4. Can these correlations help to assess a batter/pitcher’s value relative to others in the league?

These hypothetical correlations tell me that even though it is common practice to pay a player a massive salary after a couple of stellar seasons, doing so isn’t very logical. It makes more sense to avoid handing out those massive contracts and instead, to focus on players that might be playing 3A or 2A ball. This is because if you hone in on these players when they’re near a phenomenal rookie season, then you end up paying them what they deserve rather than spending massive amounts of money on under performing veterans.

5. Can I construct a machine learning model using a batter/pitcher’s performance data from 2013-2015 to predict if said batter/pitcher will receive an above average salary in 2016?

I hypothesize that this is possible because as I’ve already explained some of my theory above, I believe that players who have a couple of phenomenal seasons will likely be well compensated even though it stands to reason that their performance regresses back to average in the seasons to follow.

**Note

In order to perform this analysis, I used the 2020 version of salary, batting, and pitching datasets. I was able to obtain the batting and pitching datasets directly from R and used mlb.com to obtain the salary dataset. Below is the documentation for the Lahman R package where this data was found.

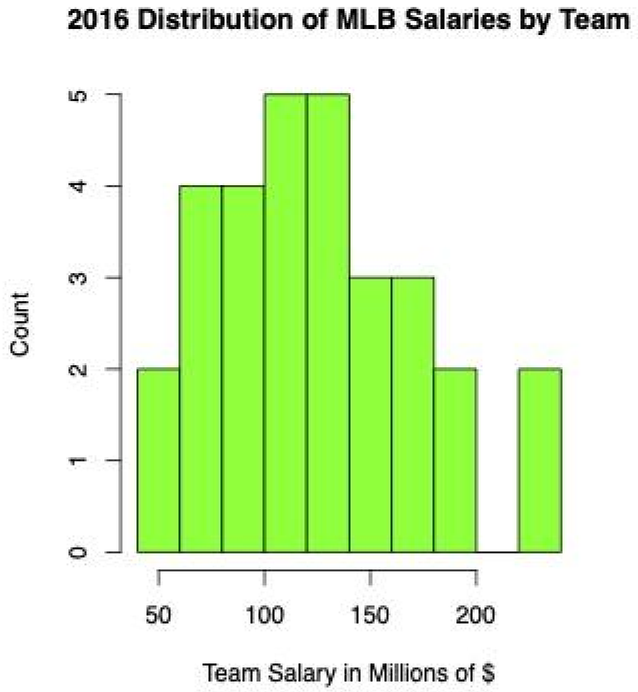
https://cran.r-project.org/web/packages/Lahman/Lahman.pdf

A. Preliminary Analysis

I decided to focus on salaries paid in 2016 because the version of MLB data available in the Lahman package was through the end of the 2019 season. This provided a straightforward metric as I could look at the three seasons after 2016 (2017-2019) and well as the three that preceded it (2013-2015).

I began by focusing on the Salary dataset. I filtered by year (2016) and created a histogram focusing on each team’s salary to see the distribution of payroll across both leagues.

```
teamSalary <- Salaries %>%
      filter(yearID == 2016) %>%
      group_by(lgID, teamID, yearID) %>%
      summarize(Salary = sum(as.numeric(salary))) %>%
      group_by(yearID, lgID) %>%
      arrange(desc(Salary))
hist(teamSalary$Salary/1e6, main="2016 Distribution of MLB Salaries by Team",
    ylab="Count", xlab="Team Salary in Millions of $", col="green")
```

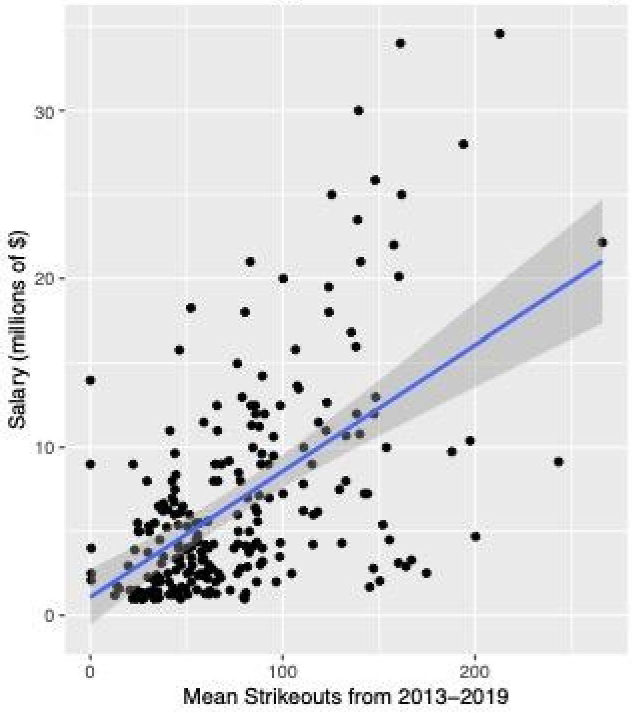


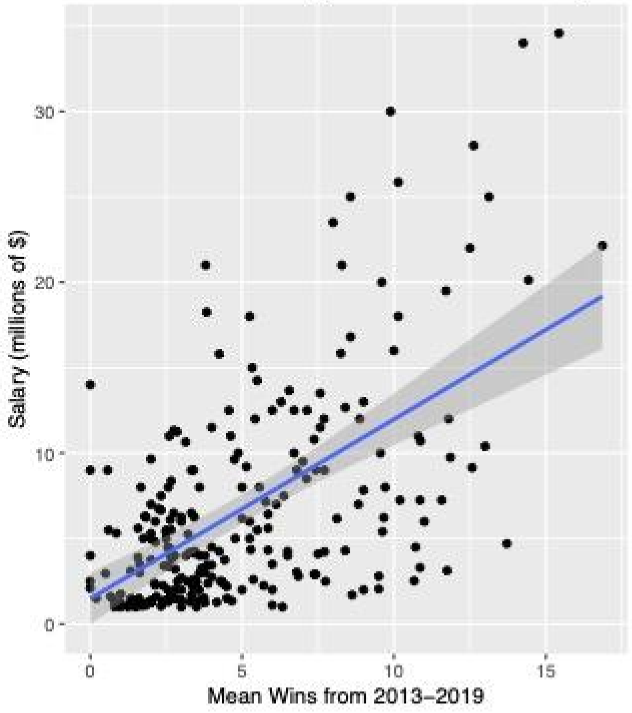
I chose to import my cleaned salary dataset (.csv) only containing information from 2016. This was more thorough as it consisted of lengthy individual player data, which I wanted to use in order to merge the Batting and Pitching datasets with the salary dataset by playerID.

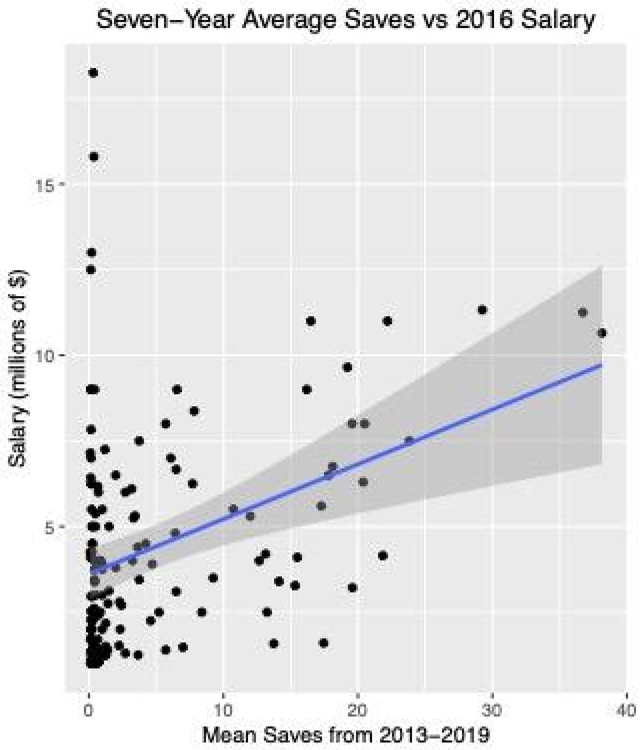
My goal is for both Batting and Pitching to construct three different data frames. The first will be a seven year average (2013-2019) of the statistics I choose to analyze. The second will be a three year average (2013-2015) prior to 2016 and the third will be a three year average post 2016 (2017-2019).

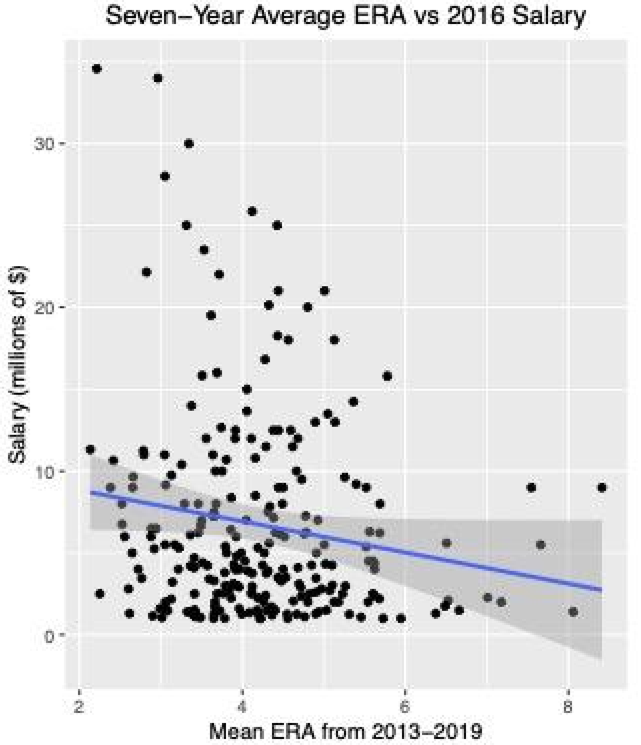
I began by reading the salary dataset csv into R studio. Then, I filtered the Pitching dataset in order to examine data from 2013-2019. Next, I merged Pitching with my imported salary dataset by the parameter “playerID”. I continued cleaning the combined dataset and removed columns I previously thought were important, but realized they would no longer be pertinent in this analysis. I filtered the data to only contain complete cases (rows) as well as no duplicate players. Then, I removed statistics from the tibble that I wouldn’t be analyzing. I created six graphs showing the relationship between average ERA, SO, HR, W, SV, BAOpp from 2013 to 2019 vs each pitcher’s 2016 salary. I filtered to focus on (0 < meanERA < 9), (meanSV > 0), and (salary_2016 >= 1,000,000).

I included a data table listing the correlations, but the three most positive correlations with salary were Strikeouts, Wins, and Saves.This aligns with my hypothesis. It makes sense that Saves would show a positive correlation as this statistic is related to Wins. There is also a clear negative correlation between ERA and salary. This relationship is expected as a pitcher would receive a hefty contract for producing a low ERA meaning the higher the ERA, the lower the salary.

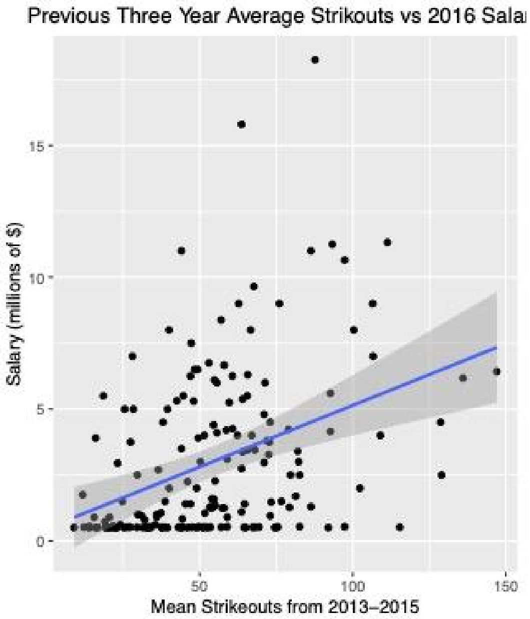

Seven−Year Average Strike Outs vs 2016 Salary

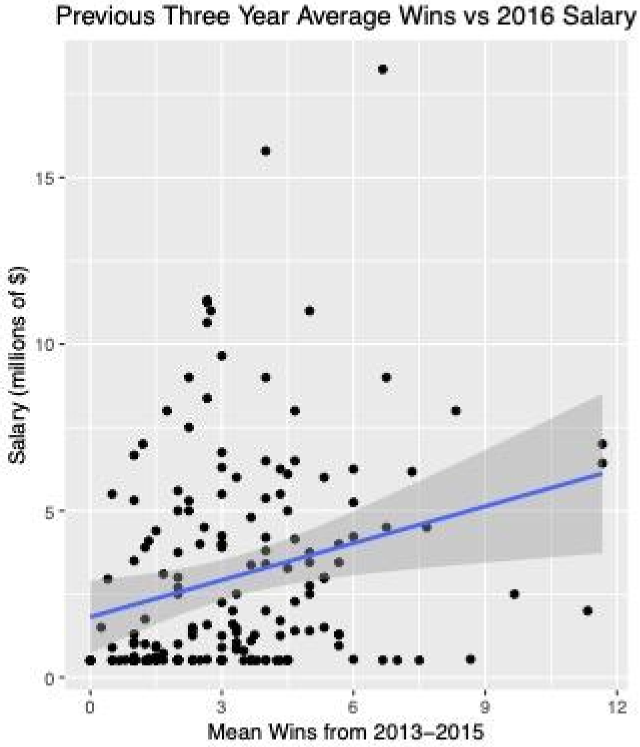

Seven−Year Average Wins vs 2016 Salary

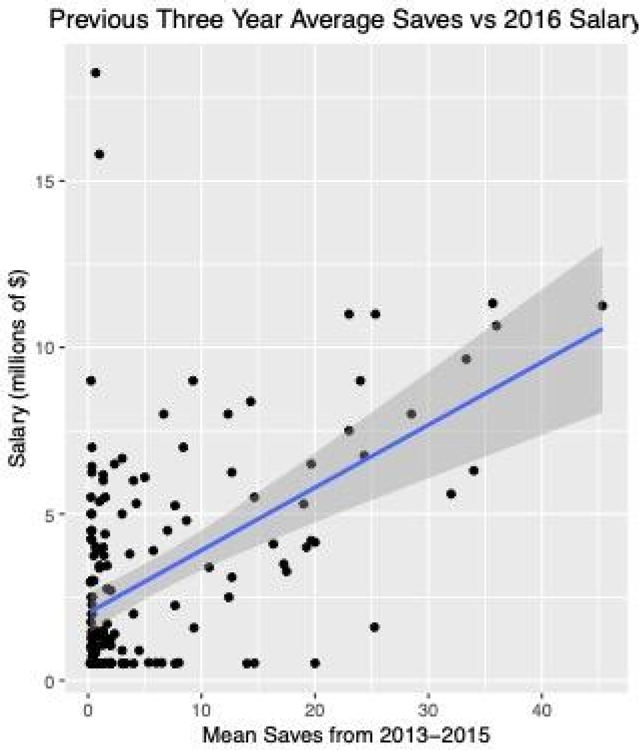


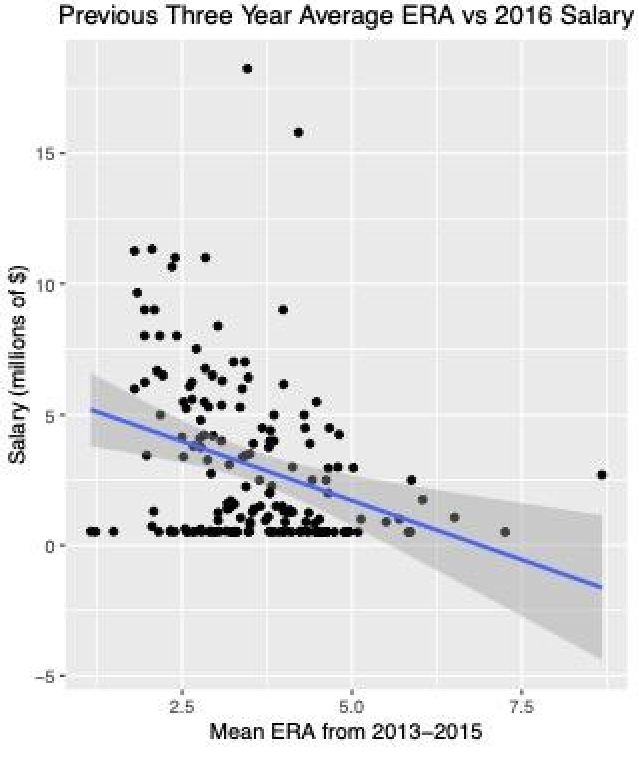


Next, I cleaned this pitching and salary merged dataset similar to how I did previously when looking at seven year averages. However, here I filtered in order to look at data from 2013 through 2015. In order to have a similar sized dataset (~150 entries), I only focused on limiting the data to meanERA > 0 and meanSV > 0. I understand that this inherently skews my data, but my goal was to be consistent regarding the number of rows being examined. Again, the result was that the strongest positive correlations with salary were Strikeouts, Wins, and Saves. Conversely, there was an even more negative correlation between ERA and salary, which aligns with my hypothesis that high performing pitchers will post a lower ERA prior to their salary obtained in 2016.









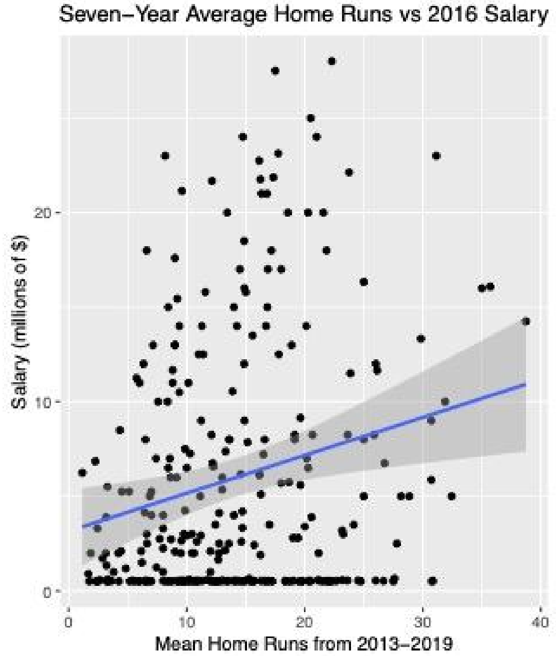
Lastly, I cleaned this pitching and salary merged dataset similarly, but instead focusing on 2017-2019. However, here I filtered in order to look at data from 2013 through 2015. In order to have a similar sized dataset (~150 entries), I focused on limiting the data to (meanERA > 0, (meanSV > 0), and salary_2016 > 600,000. Once again, I understand that this inherently skews my data, but my goal was to be consistent regarding the number of rows being examined. *Here, as I anticipated, the correlations have all substantially decreased.* There was a clear linear relationship between a pitcher's performance in the three seasons prior and the salary he received in 2016, but this dependence was not observed in the three seasons following.

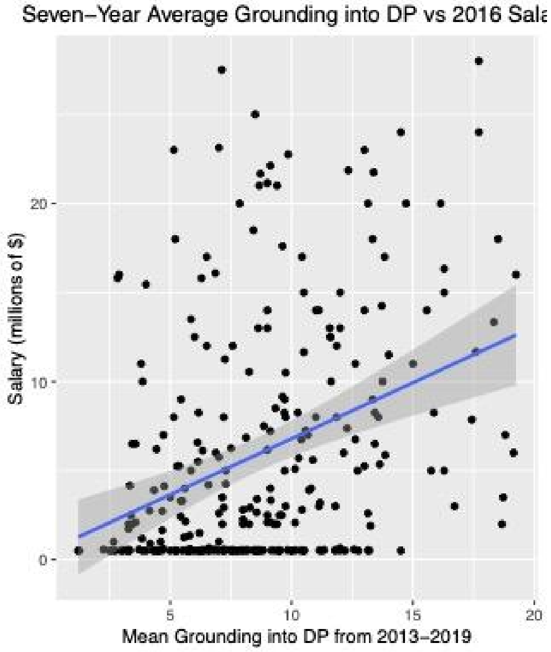
Correlation (cor)	**Previous Three Seasons**	**Seven Year Average**	**Latter Three Seasons**
ERA	-0.321	-0.152	0.0787
Strikeouts	0.390	0.228	0.153
Home Runs	0.183	0.116	0.0971
Wins	0.256	0.167	-0.0782
Saves	0.521	0.381	0.292
BAOpp	-0.266	-0.206	-0.0377

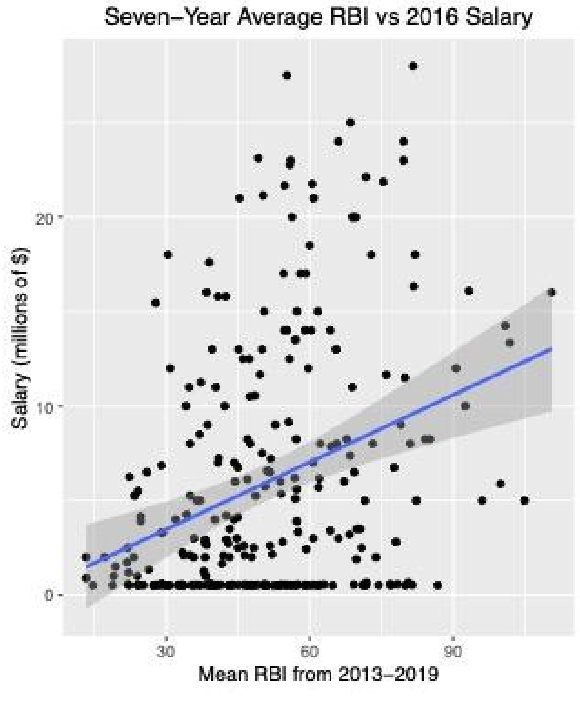
This goes hand in hand with what I've already discussed in this analysis, meaning, in any random experiment, an observed outlier will eventually over time regress back toward the average. In this case, pitchers who made above average appearances from 2013-2015 will receive large contracts (2016) even though statistically they are more likely to return back to the mean from 2017-2019. It stands to reason that those pitchers who were paid more will not be of any substantial value in the seasons to follow. A simpler example is that pitchers who posted more wins from 2013-2015 will post fewer wins following their salary paid in 2016.

Batting:
To start, I filtered the Batting dataset to examine the data from 2013-2019. As I did previously, I merged Batting with my cleaned salary dataset. I continued cleaning the combined dataset and removed columns I previously thought were important, but realized they would no longer be pertinent in this analysis. I filtered the data to only contain complete cases (rows) as well as no duplicate players. Then, I removed statistics from the tibble that I wouldn't be analyzing. I created thirteen graphs showing the relationship between average RBI, HR, H, 2B, 3B, BB, SO, GIDP, HBP, SB, CS, G, AB from 2013 to 2019 vs each batter's 2016 salary. I filtered to focus on position players (removing pitchers who had batting data) as well as meanG > 81 in order to restrict the number of rows between 250-300 and to look at batters who played at least half of the 162 game season.

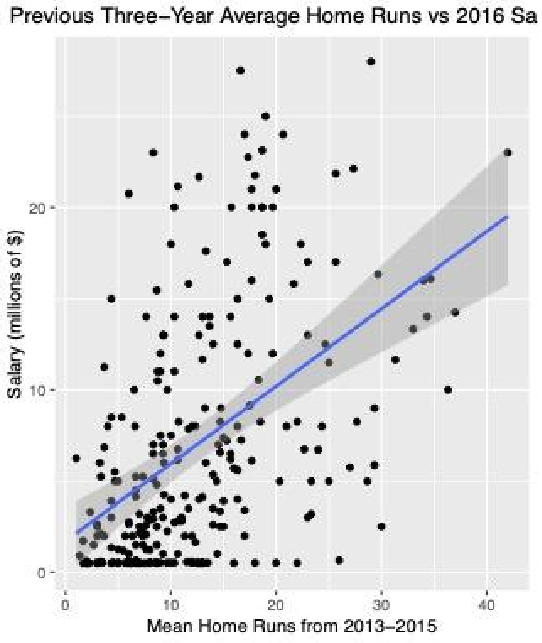
I included a data table listing the correlations, but the three most positive correlations with salary were GIDP (Grounding into Double Play), HR (Home Runs), and RBI (Runs batted in).This aligns with my hypothesis. As I previously explained, team management is looking for the player who can regularly 'hit it out of the park'. More home runs equate to more hard-hit outs (GIDP). There is also a strong positive relationship between the seven year average RBIs and the batter's 2016 salary, which makes sense because consistently performing players are valuable to the team as a whole.

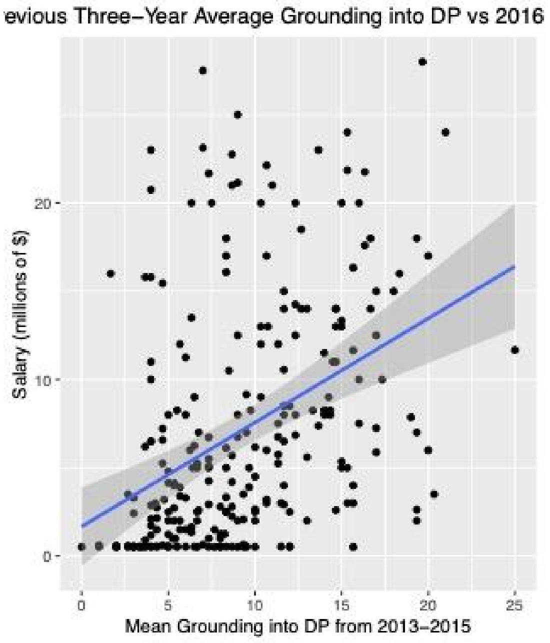


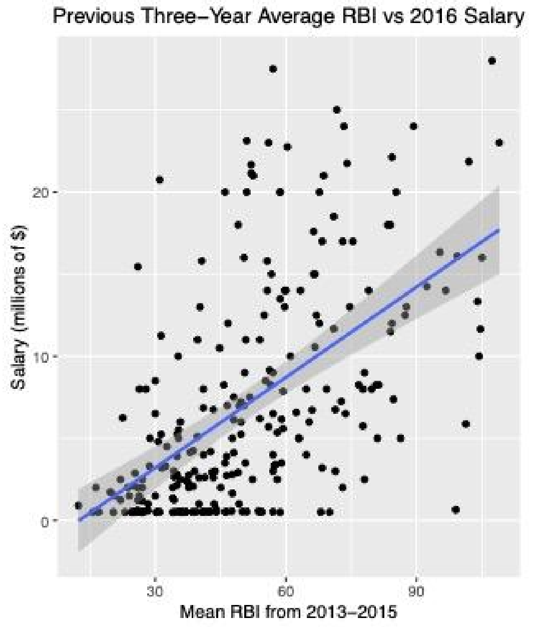


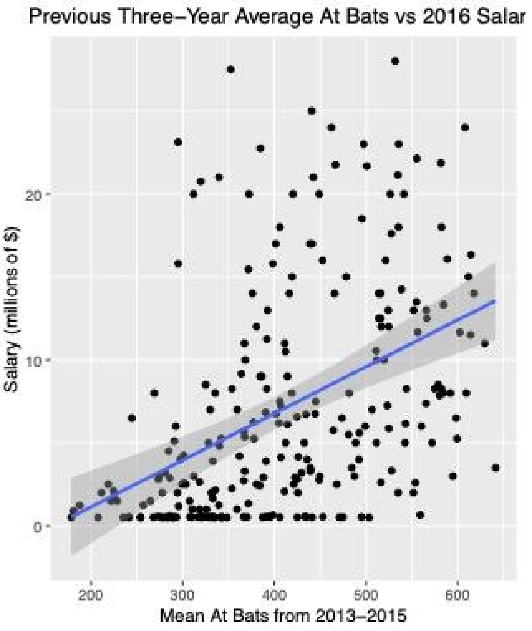


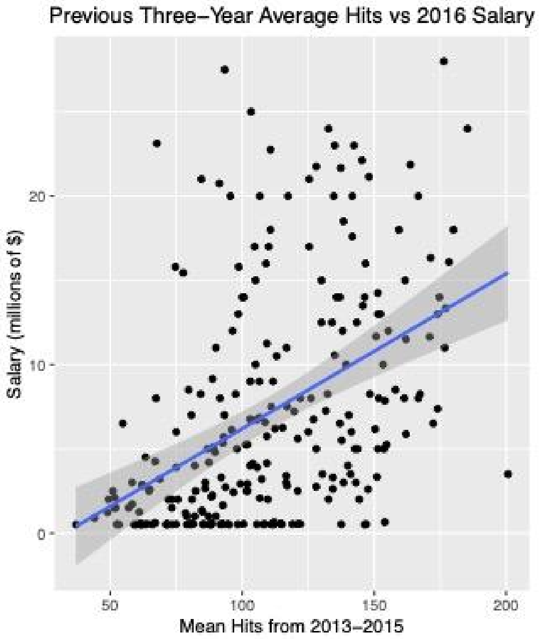
Next, I cleaned this pitching and salary merged dataset similar to how I did previously when looking at seven year averages. However, here I filtered in order to look at data from 2013 through 2015. In order to have a similar sized dataset (~250 entries), I only focused on limiting the data to meanG > 81. I understand that this inherently skews my data, but my goal was to be consistent regarding the number of rows being examined. In this case, the strongest correlations with 2016 salary were RBI, HR, H, 2B, BB, GIDP, and AB. In detail, I have explained the expected positive relationship between salary and HR, GIDP, and RBI. The other positive relationships observed in the dataset from 2013-2015 are reasonable considering the large number of at-bats young players can have. More specifically, when referring to players receiving large contracts in 2016, it stands to reason that they performed above-average offensively - meaning more hits, extra-base hits, and walks.

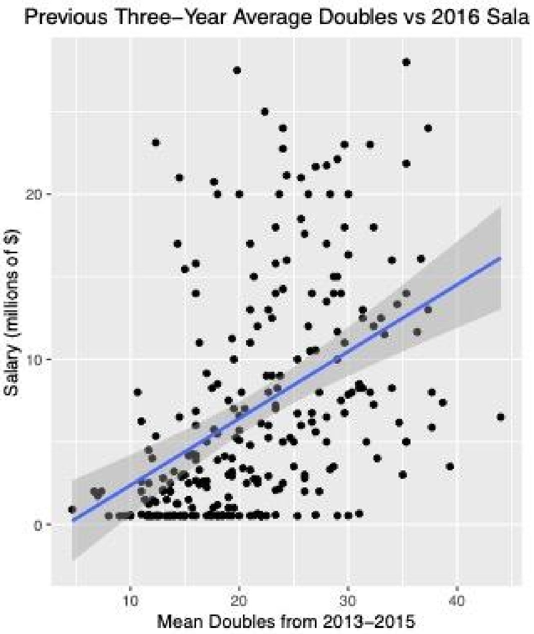


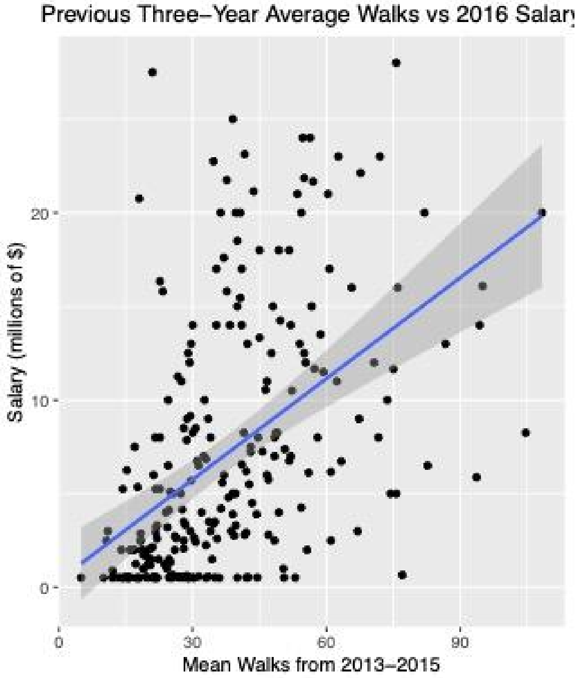












Lastly, I cleaned this batting and salary merged dataset similarly, but instead focusing on 2017-2019. However, here I filtered in order to look at data from 2013 through 2015. In order to have a similar sized dataset (~250 entries), I again focused on limiting the data to meanG >81. I understand that this inherently skews my data, but my goal was to be consistent regarding the number of rows being examined. *Here, as I anticipated, the correlations have all substantially decreased.* There was a clear linear relationship between a batter's performance in the three seasons prior and the salary he received in 2016, but this dependence was not observed in the three seasons following.

If teams can hone in on a player before he reaches his optimal performing season, they acquire the player for less money.

Correlation (cor)	**Previous Three Seasons**	**Seven Year Average**	**Latter Three Seasons**
RBI (Runs Batted In)	0.564	0.327	0.0514
HR (Home Runs)	0.484	0.219	-0.0187
H (Hits)	0.460	0.271	0.0513
2B (Doubles)	0.448	0.231	0.0162
3B (Triples)	-0.169	-0.211	-0.216
BB (Walks)	0.490	0.295	0.106
SO (Strikeouts)	0.272	0.101	-0.104
GIDP (Ground into Double Play)	0.410	0.362	0.208
HBP (Hit By Pitch)	0.150	0.0441	-0.00818
SB (Stolen Bases)	-0.0348	-0.0883	-0.112
CS (Caught Stealing)	-0.110	-0.152	-0.184
G (Games)	0.370	0.214	-0.0338
AB (At Bats)	0.447	0.272	0.0373

B-1 Further Analysis : T-Test (Pitching)

Again using 2016 as my year of interest (since I am looking at salaries from that year), I wanted to see if there was a measurable difference between a pitcher's performance if he was paid above or below the mean. I needed to use an independent sample t-test due to the unrelatedness of the two means. I've included two tables showing the means and standard deviations and the t-scores and standard errors for all pertinent pitching statistics. I am going to delve into performing this t-test looking at the mean of SV (Saves) vs 2016 salary. Specifically, I was looking at the change in SV from the three seasons (2013-2015) preceding the 2016 salary and the latter three seasons (2017-2019). Logic shows that if this were a positive number, it would mean that the pitcher threw better in the years following the 2016 salary whereas a negative value would mean he threw worse. If my hypothesis is correct - meaning that over time, a pitcher's performance falls to the average - then pitchers paid above the mean 2016 salary would have a smaller △SV . If I am incorrect, then there would be no observable difference in △SV between pitchers who were paid above average in 2016 and those who were paid below average. I believe it is appropriate to use a one-tailed t-test because my question to answer is if the means are different in one direction (not two).

△ SV = (Avg SV 2017-2019) - (Avg SV 2013-2015)

To perform any statistical test, I need to create a null and an alternative hypothesis. The purpose of the null hypothesis is to effectively state there is no difference between the two groups being compared, meaning it's used to either certify or denounce the statistical claim I'm making. Conversely, the purpose of the alternative hypothesis is to state that there is a relationship (in this case an observed difference) between the two groups being compared, meaning that my analysis was statistically significant.

My null hypothesis (H_0) is: pitchers who were paid above average in 2016 ($\mu_{\text{above avg salary}}$)will have an average number of saves (△SV) equivalent or greater than the (△SV) recorded by pitchers who were paid below average ($\mu_{\text{below avg salary}}$) in 2016.

$H_0 : \mu_{\text{above avg salary}} - \mu_{\text{below avg salary}} >= 0$

My alternative hypothesis (H_a) is: pitchers who were paid above average ($\mu_{\text{above avg salary}}$) in 2016 will have an average number of saves (△SV) less than the (△SV) recorded by pitchers who were paid below average ($\mu_{\text{below avg salary}}$) in 2016.

$H_a : \mu_{\text{above avg salary}} - \mu_{\text{below avg salary}} < 0$

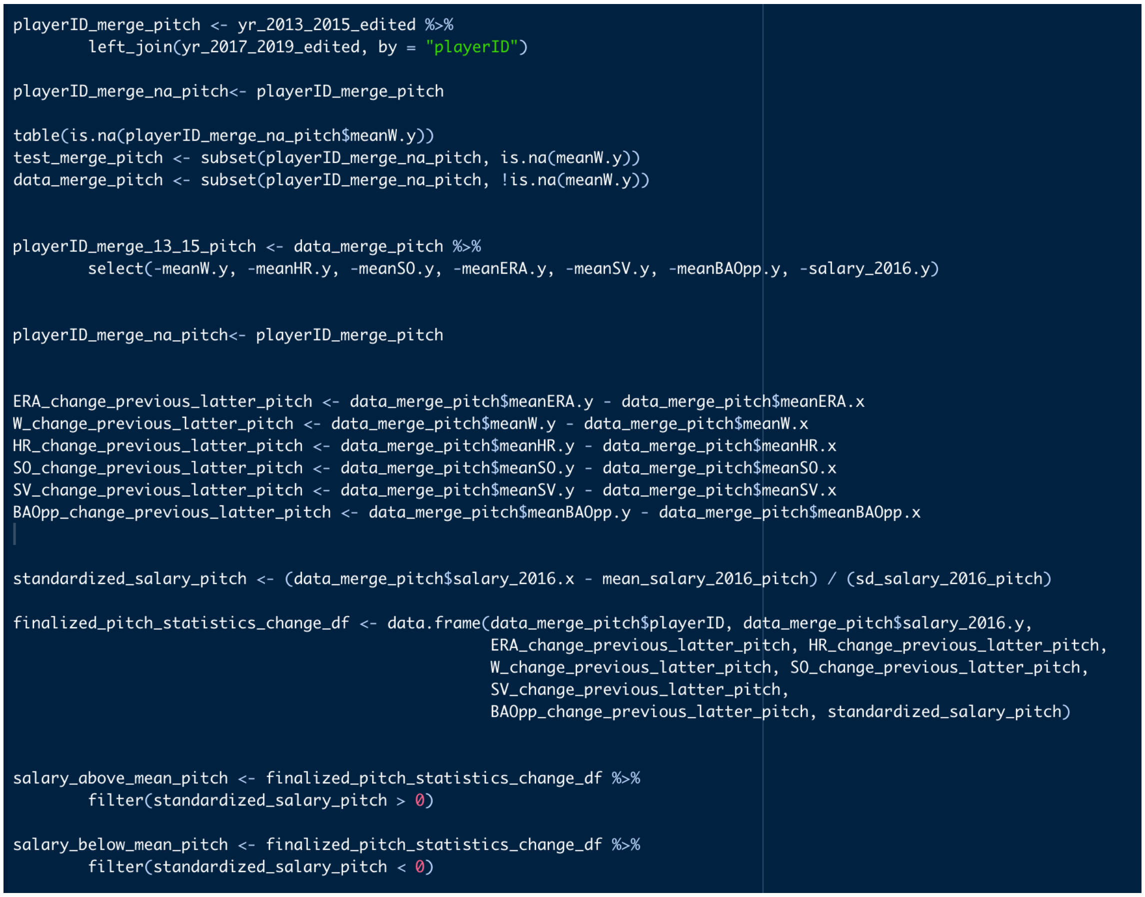
To perform a t-test, I began by establishing a confidence level (α) of 0.05. Next, I had to calculate the degrees of freedom value (df) for my one-tailed t-Test in order to find the t-critical value.

$df = \#samples_{\text{above avg salary}} + \#samples_{\text{below avg salary}} - 2$

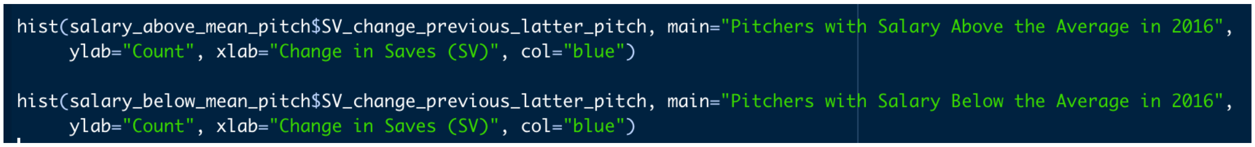
Since there were 38 pitchers paid above average and 27 paid below, my df value was 63. When looking at the t-table, 63 corresponded to a t-critical value of -1.669. The t-critical value is negative because the null hypothesis states that the means will change negatively.

I then had to calculate the sample mean and standard deviation of △SV for pitchers paid above average in 2016. I had to perform the same calculation for pitchers paid below average in 2016. Next, I had to find the difference in means between pitchers paid above average and those paid below (the two sample groups). Afterwards, I used the standard deviations from each of my sample groups and normalized them using the number of pitchers in each respective sample in order to find the standard error. Then I found the t-statistic, which was the difference in means between the two sample groups divided by the standard error. The next step was comparing the calculated t-statistics to the previously found t-critical value in order to either reject or accept the null hypothesis.

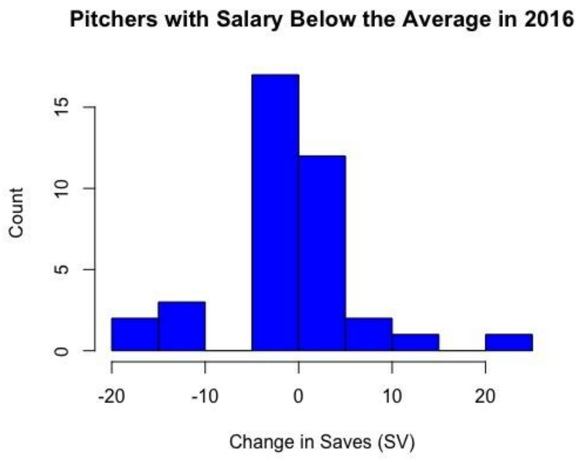
I used my same altered pitching data frame that I have already discussed, but added a standardized column so that I could easily observe if a pitcher was paid above or below the average in 2016.



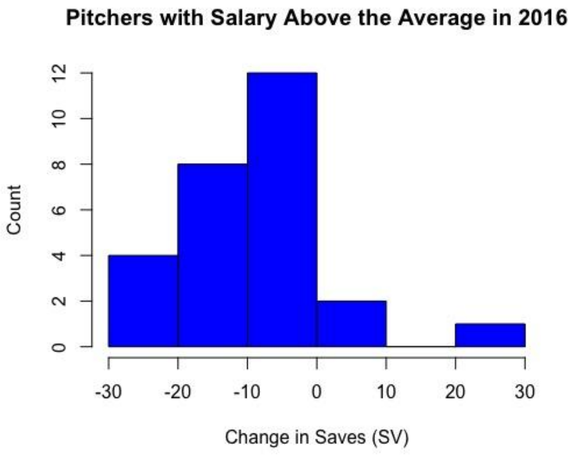
```
playerID_merge_pitch <- yr_2013_2015_edited %>%
        left_join(yr_2017_2019_edited, by = "playerID")

playerID_merge_na_pitch<- playerID_merge_pitch

table(is.na(playerID_merge_na_pitch$meanW.y))
test_merge_pitch <- subset(playerID_merge_na_pitch, is.na(meanW.y))
data_merge_pitch <- subset(playerID_merge_na_pitch, !is.na(meanW.y))


playerID_merge_13_15_pitch <- data_merge_pitch %>%
        select(-meanW.y, -meanHR.y, -meanSO.y, -meanERA.y, -meanSV.y, -meanBAOpp.y, -salary_2016.y)


playerID_merge_na_pitch<- playerID_merge_pitch


ERA_change_previous_latter_pitch <- data_merge_pitch$meanERA.y - data_merge_pitch$meanERA.x
W_change_previous_latter_pitch <- data_merge_pitch$meanW.y - data_merge_pitch$meanW.x
HR_change_previous_latter_pitch <- data_merge_pitch$meanHR.y - data_merge_pitch$meanHR.x
SO_change_previous_latter_pitch <- data_merge_pitch$meanSO.y - data_merge_pitch$meanSO.x
SV_change_previous_latter_pitch <- data_merge_pitch$meanSV.y - data_merge_pitch$meanSV.x
BAOpp_change_previous_latter_pitch <- data_merge_pitch$meanBAOpp.y - data_merge_pitch$meanBAOpp.x


standardized_salary_pitch <- (data_merge_pitch$salary_2016.x - mean_salary_2016_pitch) / (sd_salary_2016_pitch)

finalized_pitch_statistics_change_df <- data.frame(data_merge_pitch$playerID, data_merge_pitch$salary_2016.y,
                                                   ERA_change_previous_latter_pitch, HR_change_previous_latter_pitch,
                                                   W_change_previous_latter_pitch, SO_change_previous_latter_pitch,
                                                   SV_change_previous_latter_pitch,
                                                   BAOpp_change_previous_latter_pitch, standardized_salary_pitch)


salary_above_mean_pitch <- finalized_pitch_statistics_change_df %>%
        filter(standardized_salary_pitch > 0)

salary_below_mean_pitch <- finalized_pitch_statistics_change_df %>%
        filter(standardized_salary_pitch < 0)
```



```
hist(salary_above_mean_pitch$SV_change_previous_latter_pitch, main="Pitchers with Salary Above the Average in 2016",
     ylab="Count", xlab="Change in Saves (SV)", col="blue")

hist(salary_below_mean_pitch$SV_change_previous_latter_pitch, main="Pitchers with Salary Below the Average in 2016",
     ylab="Count", xlab="Change in Saves (SV)", col="blue")
```

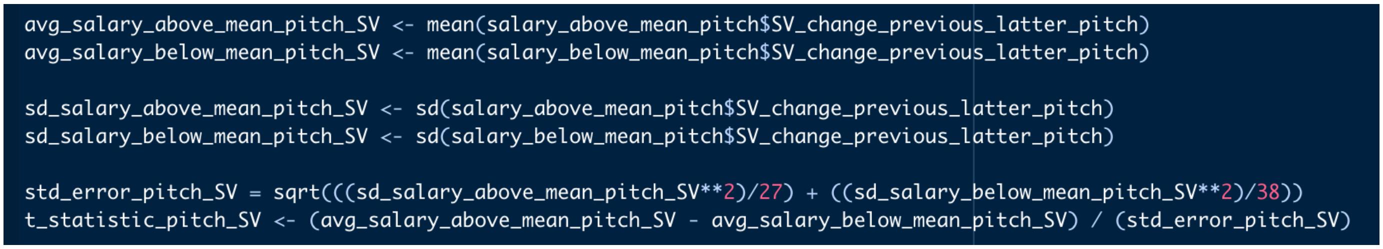


Pitchers with Salary Below the Average in 2016



Pitchers with Salary Above the Average in 2016

Both histograms loosely follow a normal distribution, but the above average salary graph is skewed more negative. The mean for pitchers paid below average in 2016 was -0.7842 and the mean for those paid above average was -8.219. The negative change in saves (△SV) demonstrates that the pitchers didn't perform as well during the years post 2016 as they did in the years preceding 2016.



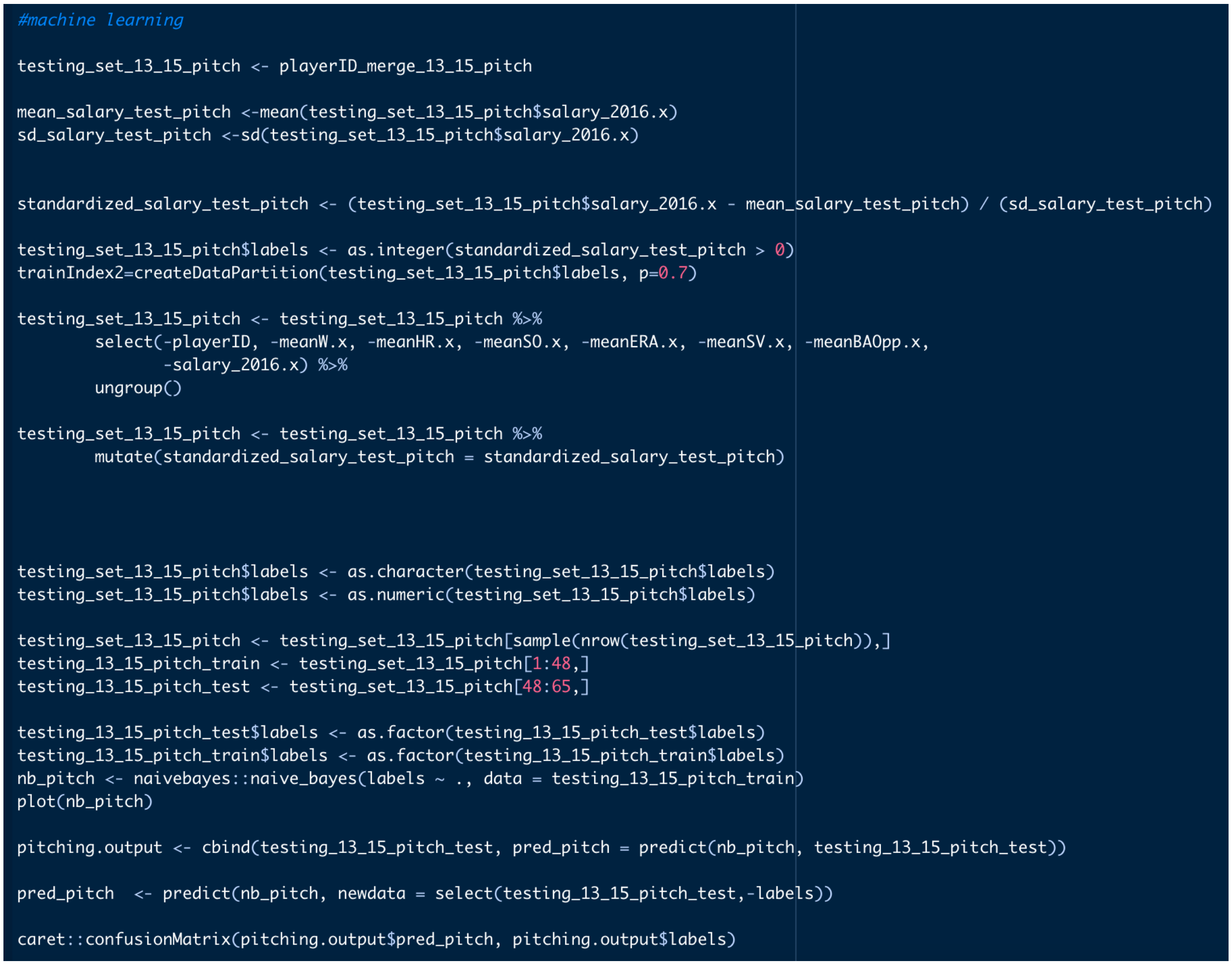
```
avg_salary_above_mean_pitch_SV <- mean(salary_above_mean_pitch$SV_change_previous_latter_pitch)
avg_salary_below_mean_pitch_SV <- mean(salary_below_mean_pitch$SV_change_previous_latter_pitch)

sd_salary_above_mean_pitch_SV <- sd(salary_above_mean_pitch$SV_change_previous_latter_pitch)
sd_salary_below_mean_pitch_SV <- sd(salary_below_mean_pitch$SV_change_previous_latter_pitch)

std_error_pitch_SV = sqrt(((sd_salary_above_mean_pitch_SV**2)/27) + ((sd_salary_below_mean_pitch_SV**2)/38))
t_statistic_pitch_SV <- (avg_salary_above_mean_pitch_SV - avg_salary_below_mean_pitch_SV) / (std_error_pitch_SV)
```

The calculated t-statistic (listed in the table below) was -3.156, which is less than -1.669 (the t-critical value) using the confidence level 0.05. In other words, I'm able to confidently reject H_0. I've used statistics to prove that pitchers who were paid above average in 2016 performed worse compared to pitchers who were paid below average from the years preceding 2016 (2013-2015) to the years post 2016 (2017-2019). Regression to the mean indeed occurred because pitchers who threw better from 2013-2015 were paid above average in 2016, but their output fell from 2017-2019 **more** than the group of pitchers who were paid below average in 2016. Statistically speaking, I used this t-test to see how a pitcher's output changed three years prior to his 2016 salary and the three years afterwards. The change in saves was demonstrated by the more negative mean from 2017-2019 than the one found from 2013-2015. The calculated t-statistic being less than the t-critical value imply that pitchers who were paid above average in 2016 will perform worse from 2017-2019 than those paid below average in 2016. This means that over a lengthy period of time, an indicator of a pitcher's performance (saves) will eventually fall towards the mean. This follows my claim that MLB pays pitchers hefty salaries even though they will inevitably throw worse in the seasons that follow.

I also performed independent one-sided t-tests for ERA, W, HR, SO, and BAOpp with respect to pitchers and their 2016 salaries. Then, I performed independent one-sided t-tests for RBI, H, BB, and HR with respect to batters and their 2016 salaries. As previously stated, these were all one-tailed t-test because I wanted to determine if the means differed in one direction (negative) and not two. As expected, the statistics I chose to analyze demonstrated a negative correlation between 2016 salary and a batter's/pitcher's performance in the years that followed. The only positive correlation was between ERA and a pitcher's 2016 salary - meaning that a pitcher who was paid above average tended to post a lower ERA.

B-2: Create a Model with Machine Learning

As stated in the beginning of this report, my fifth objective was to construct a machine learning model using a batter/pitcher's performance data from 2013-2015 in order to predict if said batter/pitcher would receive a salary above the average in 2016. I used the same filter - meaning pitchers who posted a meanERA > 0, meanSV >0, and were paid a salary in 2016, I used their data from 2013-2015 to make the model. Continuing to build the model, I used the same pitching metrics (meanERA, meanW, meanHR, meanSO, meanSV, meanBAOpp).



```
#machine learning

testing_set_13_15_pitch <- playerID_merge_13_15_pitch

mean_salary_test_pitch <-mean(testing_set_13_15_pitch$salary_2016.x)
sd_salary_test_pitch <-sd(testing_set_13_15_pitch$salary_2016.x)


standardized_salary_test_pitch <- (testing_set_13_15_pitch$salary_2016.x - mean_salary_test_pitch) / (sd_salary_test_pitch)

testing_set_13_15_pitch$labels <- as.integer(standardized_salary_test_pitch > 0)
trainIndex2=createDataPartition(testing_set_13_15_pitch$labels, p=0.7)

testing_set_13_15_pitch <- testing_set_13_15_pitch %>%
        select(-playerID, -meanW.x, -meanHR.x, -meanSO.x, -meanERA.x, -meanSV.x, -meanBAOpp.x,
               -salary_2016.x) %>%
        ungroup()

testing_set_13_15_pitch <- testing_set_13_15_pitch %>%
        mutate(standardized_salary_test_pitch = standardized_salary_test_pitch)




testing_set_13_15_pitch$labels <- as.character(testing_set_13_15_pitch$labels)
testing_set_13_15_pitch$labels <- as.numeric(testing_set_13_15_pitch$labels)

testing_set_13_15_pitch <- testing_set_13_15_pitch[sample(nrow(testing_set_13_15_pitch)),]
testing_13_15_pitch_train <- testing_set_13_15_pitch[1:48,]
testing_13_15_pitch_test <- testing_set_13_15_pitch[48:65,]

testing_13_15_pitch_test$labels <- as.factor(testing_13_15_pitch_test$labels)
testing_13_15_pitch_train$labels <- as.factor(testing_13_15_pitch_train$labels)
nb_pitch <- naivebayes::naive_bayes(labels ~ ., data = testing_13_15_pitch_train)
plot(nb_pitch)

pitching.output <- cbind(testing_13_15_pitch_test, pred_pitch = predict(nb_pitch, testing_13_15_pitch_test))

pred_pitch  <- predict(nb_pitch, newdata = select(testing_13_15_pitch_test,-labels))

caret::confusionMatrix(pitching.output$pred_pitch, pitching.output$labels)
```

In this data frame, my inputs were meanERA, meanW, meanHR, meanSO, meanSV, meanBAOpp averaged from 2013-2015. To carry out machine learning, I also needed an output. The output was if a pitcher with those inputs was paid a salary in 2016 that was above the average. To do this, I standardized the 2016 salaries as I did when performing the t-test and then alter them to be label of 1 or 0 with the 1 referencing above average salaries and 0 referring to below average salaries.

I used a Gaussian Naive Bayes classifier because it follows a Gaussian (normal) distribution, which is what I wanted since I was dealing with data that was continuous. My confusion matrix produced an accuracy of 1, which means that it can predict with better than random accuracy that a player will be paid above the average in 2016 based off the meanERA, meanW, meanHR, meanSO, meanSV, meanBAOpp from 2013-2015. It's obvious that this classifier would be more effective with the data from more pitchers. In addition, these statistics (inputs) weren't significantly tied with the salaries pitchers were paid. It's possible that other statistics could better demonstrate with more accuracy if a pitcher will mandate a salary above the average. The machine learning model I built looking at position players' salaries vs the specified batting metrics (meanG, meanAB, mean2B, mean3B, meanSB, meanCS, meanHBP, meanGIDP) included more entries (154) than the 65 used with respect to pitchers. The batting model still produced an accuracy rate of 83 percent. This usage of machine learning showed that even without a lengthy amount of data, this mechanism can make predictions with a higher accuracy than guessing randomly.

C. Summary

This analysis showed that a pitcher/batter's statistics might be related to his salary, but this does not equate to said statistics leading to a less than or greater than average salary.

These were the conclusions I made:

1. Pitching data from 2013-2019 demonstrated that SV (saves) and SO (strikeouts) were the statistics most highly correlated to 2016 salary. These correlations were certainly stronger (larger) from 2013-2015 than they were from 2017-2019.

2. Batting data from 2013-2019 demonstrated that GIDP (grounding into a double play) and RBI (runs batted in) were the statistics most highly correlated to 2016 salary. These correlations were certainly stronger (larger) from 2013-2015 than they were from 2017-2019.

3. Using a t-Test, I showed that pitchers who were paid above average in 2016 performed worse (measured by Saves) from 2017-2019 than they did from 2013-2015. This statistically shows that regarding performance, pitchers paid above the average regress towards the mean over a lengthy period of time.

4. Using a t-Test, I showed that batters who were paid above average in 2016 performed worse (measured by Hits, Walks, Home Runs, and RBIs) from 2017-2019 than they did from 2013-2015. This statistically shows that regarding performance, batters paid above the average regress towards the mean over a lengthy period of time. Therefore, team owners should try to find batters/pitchers who are undervalued prior to them having strong performing seasons. There is a huge disparity in payrolls of MLB teams and by adopting this approach, I believe more teams will be successful. This analysis showed that once a pitcher/batter has gotten to the point where they're paid a heftier salary, they statistically will not perform as well in the seasons that follow. This is because his previous performance was mathematically an outlier and will eventually fall towards the mean.

These were the limitations I ran into while performing this analysis:

1. I was not able to set a control for the age of the players when analyzing this data. The age of the pitchers/batters was an issue in that as a player ages and spends more time in the majors, his salary increases even if his performance likely falters.

2. The number of entries (pitchers/batters) in the datasets. When filtering the pitching dataset - I was only left with around 65 pitchers and when filtering the batting dataset, I was only left with 154 batters. In order to look at more data, I could have looked at a longer span of years (maybe 14) or I could have removed filters (each player batting in at least 81 games).

3. Some teams inevitably make it to the post-season and it's entirely possible that a batter/pitcher's salary could increase based off their performance. It would be interesting to conduct this analysis again, but to include post-season data in order to see its effect vs that of regular season play on one's salary. In addition, I didn't account for the phenomena that a pitcher's number of wins/saves is very related to the defense of the team (their ability to field effectively). If I were to conduct another analysis, I would use these basic metrics (RBIs, Hits, HRs, etc) to look at more in-depth statistics in order to pin down all of the factors that might contribute to a recorded number of home runs. An example of this is WAR (wins above replacement) which gives each batter a number that indicates how many wins they helped their team achieve vs the number that any random batter would have created.

4. The randomness in baseball statistics meaning that some years batters do better and some pitchers perform better. This is obviously in conjunction with the different sizes of MLB stadiums as well as the climate found at each (precipitation, air quality, etc).

Overall, I chose MLB pitching and batting as the focus of my analysis because I am very passionate about baseball and have been for years. This was a great topic for generating hypotheses about the topic, cleaning the data, and then using statistics to point out pertinent discoveries. I even attempted using machine learning to see if I could construct a model that could predict more accurately than if I guessed at random!

D. Data Tables

Pitching Statistic	Standard Deviation/Mean	Below the Mean Salary 2016	Above the Mean Salary 2016
ERA	Mean	0.5678	1.650
	Std. Dev	2.023	1.902
W	Mean	-0.7338	-1.258
	Std. Dev	2.010	2.760
HR	Mean	0.4737	0.5420
	Std. Dev	3.755	2.940
SO	Mean	-13.82	-18.55
	Std. Dev	23.47	23.01
SV	Mean	-0.7842	-8.219
	Std. Dev	6.890	10.78

BAOpp	Mean	0.005304	0.02638
	Std. Dev	0.03954	0.03383

Pitching Statistic	**Standard Error**	**T-Statistic**
ERA	0.4915	2.201
W	0.6233	-0.8411
HR	0.8314	0.08214
SO	5.840	-0.8114
SV	2.356	-3.156
BAOpp	0.009140	2.306

Batting Statistic	**Standard Deviation/Mean**	**Below the Mean Salary 2016**	**Above the Mean Salary 2016**
RBI	Mean	12.88	-9.660
	Std. Dev	19.84	18.26
H	Mean	5.898	-28.16
	Std. Dev	30.67	27.48
BB	Mean	9.850	-6.642
	Std. Dev	15.42	11.80
HR	Mean	6.213	-1.165
	Std. Dev	6.940	6.629

Batting Statistic	**Standard Error**	**T-Statistic**
RBI	3.072	-7.338
Hit	4.689	-7.263
BB	2.210	-7.464

HR	1.093	-6.748