Colin McGarry

STAT 5905 Final Report

**Introduction**

In this project, I will analyze baseball statistics taken from the 2021 and 2022 Major League Baseball seasons, and investigate relationships between different statistics. The statistics I am interested in pertain to starting pitchers. The primary research question that I will answer in this project is “What variables are best at predicting a starting pitcher’s ERA?” ERA stands for Earned Run Average, and it is a statistic that measures the average number of earned runs that a pitcher has allowed his opponent to score for every 9 innings he has pitched. As the number of runs scored determines who wins or loses a baseball game, ERA is one of the most important statistics used to evaluate MLB pitchers. Although ERA is my response variable for the multiple regression model that I will create using many explanatory variables, I will also explore relationships between pairs of variables.

I chose this topic because I am a huge baseball fan, and I have done other statistics projects in the past analyzing offensive statistics. For this project, I will analyze baseball from the defensive side of the game instead, pitching. In the last 10-15 years or so, more advanced pitching statistics have been developed and part of my analysis aims to examine how they differ from basic metrics. How much certain statistics really matter has also become debatable in baseball recently. I aim to see if these statistics still have significant relationships with ERA and other performance metrics.

The data I am using for this project comes from a database, Fangraphs. Fangraphs is a website that tracks baseball statistics during each MLB season, and gets updated after every game. Statistics are available from this website in a very simple table format, and these data tables are easy to export to an Excel File. In Fangraphs, there are 4 data tables I am using. The first, titled “Standard,” contains the most basic statistics, such as each player’s name, team, ERA, the number of games they played in, and how many innings they pitched. As I only wanted to examine data on starting pitchers, I filtered out players who did not pitch at least 100 innings in 2022. There were a total of 140 pitchers who pitched 100 innings or more. Setting this minimum filtered out relief pitchers who typically only pitch one inning each game, as well as starting pitchers who did not pitch a relatively full season.

The second data table, titled “Advanced” contains more advanced statistics that require some calculation. This table includes strikeout rate (K%), walk rate (BB%), opponents’ batting average against the pitcher (AVG), and more.

The third table, “Batted Ball” contains information on what direction balls were hit by batters. This includes the percentage of balls hit on the ground (GB%), as line drives (LD%), and as fly balls (FB%). It also shows what percentage of total pitches thrown were strikes. From this dataset, I would especially expect ground ball percentage to have a strong relationship with ERA as ground balls are typically easier for the team’s defense to field. My hypothesis is that a low ERA will be associated with a high ground ball rate.

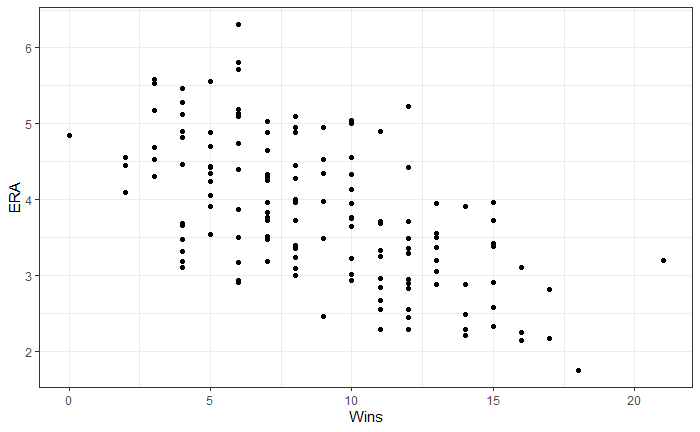
The last table, “Statcast,” has data on how hard baseballs were hit. This includes Exit Velocity, in miles per hour (EV), launch angle (LA), and HardHit%, which measures the percentage of balls hit faster than 95 miles per hour. After merging all tables and removing some duplicates, there are a total 74 variables in the dataset, although I will only be using about 20.

**Relationships Between Variables**

One such relationship that I will explore is the relationship between Wins and ERA. For every game played, there is a winning pitcher and a losing pitcher. For most games, the Win and the Loss are given to the starting pitchers, but there are cases where a relief pitcher receives a decision. The only requirements that a starting pitcher must meet to earn a win are that they must complete 5 innings, and their team must be winning when they exit the game. There are no restrictions on a pitcher’s performance, or a limit to how many runs they can allow and still earn a win. In theory, pitchers who give up fewer runs should receive more Wins, but receiving a win also relies on a pitcher’s offense to score enough runs to put their team in the lead. This makes Wins an imperfect statistic, and I am interested to see how strong the association actually is between Wins and ERA. Wins are imperfect because it is not uncommon for a starter to pitch well, but fail to earn a win because his team did not score enough runs. This is an important relationship too, as Wins often play a part in a pitcher’s salary, and whether they receive votes to be an All-Star or win a Cy Young Award.

**Separate Analyses**

**Wins and Winning Percentage**

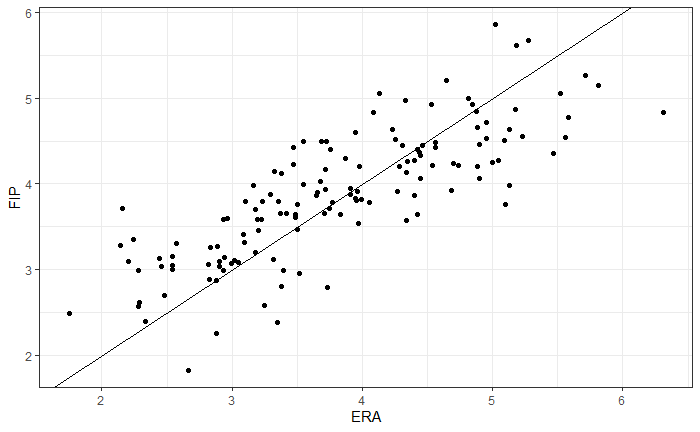
The first alternate analysis I did is in Wins, and Win Percentage. Win percentage is a variable I calculated, and is equivalent to . It is essentially the proportion of decisions a pitcher received that were wins. The most relevant relationship involved is that between ERA and Wins. This relationship is shown in this scatterplot below. There is a decent amount of scatter, but there is a clear negative trend. This is what I expected to see, as it makes sense that the higher a pitcher’s ERA is, the less wins they will accumulate. To see if this is truly a statistically significant relationship, I also ran a simple linear regression model on the two variables. The result of this model was the following equation: . The p-value of this model is less than 0.001, so there is clearly a significant relationship. That Wins is a significant variable is not necessarily surprising, but I was not expecting the p-value to be as low as it is.

As wins are also highly dependent on the pitcher’s team scoring runs, I investigated the relationship between Wins (and also winning percentage) and run support (RS/9). RS/9 is essentially the number of runs that a team scores while that pitcher is in the game, standardized to a 9-inning scale (in the same way ERA is standardized to a 9-inning scale). The linear regression model for these variables has the equation . This model also has a p-value of less than 0.001, so there is definitely a significant relationship.

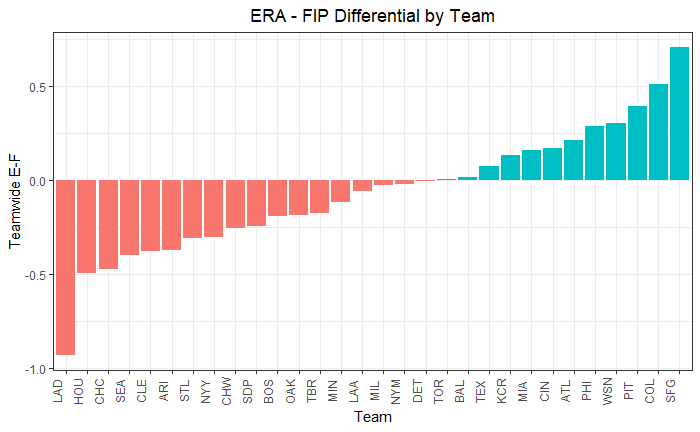
**Ground Ball Rate**

Another hypothesis I had was that a high ground ball rate would be associated with a low ERA, but this is not necessarily the case. This relationship can also be seen in a scatterplot and simple linear regression model. The scatterplot in Appendix H shows that there seems to be very little association between ground ball rate and ERA, and the correlation coefficient between the two variables confirms this, as the value is -0.0867. This shows there is almost no correlation. The simple linear regression model also confirms there is not a significant relationship as the corresponding p-value is 0.3082, well above the accepted threshold of 0.05. This was surprising to me, as a high ground ball rate is typically equated to inducing weak contact from the batter. More ground balls should result in less overall hits allowed, which I thought would result in fewer runs allowed as well. However, there appears to be no statistically significant relationship.

**FIP and E-F**

FIP is an interesting statistic to analyze, as it is designed to be a number comparable to ERA, but it ignores lots of data. FIP is based only on home runs, walks, hit by pitches, and strikeouts, and ignores all balls put in play for the defense to field. E-F stands for ERA minus FIP, and this value can be either positive or negative. If a pitcher’s FIP is lower than his ERA, this value will be positive. This case suggests that the pitcher was somewhat unlucky, or underperformed to a degree, as his ERA might have been lower if his team played better defense on balls put in play. If a pitcher’s FIP is higher than his ERA, then E-F is negative. This case suggests that the pitcher may have benefited from his team’s defense, and the overall quality of his pitching may not have been as good as his ERA implies. The goal of the FIP statistic is for it to be on a similar scale to ERA, so I plotted ERA and FIP in this scatterplot to the right just to see how they compare. Also included on this plot is the simple line , in this case, . In the scatterplot, the relationship between ERA and FIP follows this line pretty closely, showing the calculation of FIP is effective at creating a scale similar to ERA. There is a fair amount of scatter on the plot, but many values are close to the line, showing that the ERA-FIP differential is not very large for most pitchers.

I also examined the E-F statistic by team to see if there are any specific teams that stand out with consistently high or low E-F values. This is visualized in the bar graph below.



In this graph, the bar is shaded red if E-F is negative, and the bar is shaded blue if E-F is positive. It stands out that there are only 12 teams with positive E-F values, and 18 with negative E-F values. This means that the majority of teams have higher FIPs than ERAs, and suggests that they may have stronger defensive players. It is also possible that the 12 teams with lower FIPs are more effective at striking batters out and limiting walks and home runs. It also stands out that the Los Angeles Dodgers have an E-F of nearly -1, meaning that on average, their pitchers have FIPs that are almost an entire run higher than their ERA. Since the Dodgers’ E-F value is so much lower than the next lowest team, it is possible they rely on strong defense, or that their pitchers do not necessarily focus on strikeouts, or limiting home runs and walks.

**Full Linear Regression Model and Multicollinearity**

Multicollinearity is a major concern for this project as many variables go hand-in-hand, explaining the same concepts. For example, in a correlation matrix I found that ground ball percentage and fly ball percentage have a strong negative correlation (-0.9503). This makes sense as when a ball is hit on the ground, it is obviously not simultaneously hit into the air. The same goes for Strike Percentage (Strike\_pct) and Walk Rate (BB/9). If a pitcher throws a lot of his pitches for strikes, then it makes sense that he might have a lower walk rate. This is reflected in the correlation coefficient of -0.8144. To counteract this multicollinearity, the multiple regression model must be built stepwise, one variable at a time, to avoid including explanatory variables in the model that have strong associations with each other. When running a full model, including every variable that may go into the final model, the variables AVG, BB/9, and WHIP all had incredibly high variance inflation factors, meaning that they should not be used in combination in the final model. The variance inflation factors for those three variables were 158.08, 60.81, and 240.71 respectively. Like the other examples, this makes sense. WHIP stands for Walks plus Hits per Inning Pitched. If a pitcher has a higher batting average against, allowing more hits, and walks more batters they will naturally have a higher WHIP.

When constructing the multiple regression model one variable at a time, WHIP and FIP where the most significant variables by a fairly wide margin. When picking a third explanatory variable to include in the model, it was very close between Barrel% and BB/9. The 3-variable model including WHIP, FIP, and Barrel% had an R2 equal to 0.8306, and the Barrel% variable had a p-value of 0.000904. The 3-variable model with BB/9 instead of Barrel% had an R2 equal to 0.8304 and the BB/9 variable had a p-value of 0.000979. I chose Barrel% as the third variable, as it had marginally better results than BB/9.

The final model that resulted from this process included 6 significant variables and had an R2 value of 0.8596. The adjusted R2 was 0.8532. Those 6 variables were WHIP, FIP, Barrel%, BB/9, AVG, and LD%. The equation of this model is:

The Residuals vs. Fitted and Normal Q-Q plots can be seen in Appendix A after the glossary of statistics. The Residuals vs. Fitted plot shows that this model has fairly consistent residuals, that are not skewed above or below zero. The Normal Q-Q plot shows the model fits fairly well, as the data points are close to the line, except for a few outlier points at the end of the line.

|  |  |
| --- | --- |
| Variable | VIF |
| WHIP | 179.86 |
| FIP | 3.42 |
| Barrel% | 1.67 |
| BB/9 | 43.49 |
| AVG | 110.83 |
| LD% | 1.48 |

To see if this stepwise creation of a model fixed the issues with multicollinearity, I checked the variance inflation factors on this model as well. The results of this are in the table to the right. We can see that WHIP, AVG, and BB/9 still have significant multicollinearity concerns. FIP, Barrel% and LD% all have acceptable VIFs, but the values for WHIP, FIP, and AVG make this model effectively unusable. To combat this problem, I remade the ERA model, but chose BB/9 as the third variable instead of Barrel%. The p-values and R2 values were incredibly close anyway, so this may produce a better model.

Another reason to remake the model is to see if the K/9 variable is included. When constructing the first model, K/9 was significant in several models, but including it caused problems. Even though K/9 was significant, its inclusion was causing other variables to no longer be significant, so it was not included in the final six variable model. But when I kept BB/9 as the third variable, and finished building the model the same way, the process produced a new seven variable model that did include K/9. This second model has the equation:

WHIP, FIP, BB/9 and LD% carry over from the first model, but in this second model, K/9, EV, and Win% have been added, and AVG and Barrel% have been removed. The Residual and Noral Q-Q plots for this second model are shown in Appendices C and D. The Residual vs. Fitted plot shows a decent amount of scatter, but again the residuals are pretty evenly distributed above and below zero. The Normal Q-Q plot has a slight curve towards the bottom but overall, the points are close to the line and follow it linearly. These plots indicate the model is a good fit.

|  |  |
| --- | --- |
| Variable | VIF |
| WHIP | 4.78 |
| FIP | 3.12 |
| BB/9 | 2.34 |
| K/9 | 2.14 |
| LD% | 1.50 |
| EV | 1.32 |
| Win% | 1.97 |

To see if this second model solved the problem of multicollinearity, I found the variance inflation factors again, with much better results. The VIFs of this model are shown in the table to the left. In this new model, the highest VIF is 4.78, which is an acceptable value to use in a final model. This means it is safe to say multicollinearity is no longer a big concern.

This second model was also better in explaining variability. The R2 value for the second model is 0.8844 and the adjusted R2 is 0.8782. In the first model those values were 0.8596 and 0.8532 respectively. Since it eliminates multicollinearity and explains more variability in ERA, I chose the second model as my final model.

**New Statistical Method - PCA**

It was clear in both ERA models that WHIP was the predictor most responsible for the multicollinearity challenges. It consistently had the highest variance inflation factor, even though that value was drastically lowered in the final model. Two other variables that seemed to be a large cause of multicollinearity were BB/9 and AVG. One method to see what variables have high correlations with each other is Principal Component Analysis (PCA). The goal of PCA is to analyze large datasets by transforming the dataset into a lower-dimensional space and grouping together variables that are highly correlated.

The results of a principal component analysis on the dataset can be seen in Appendices E and F. Appendix E shows that about 5/8 of the variability in the data can be represented by the first principal component, as there is a cumulative proportion of about 0.625. 99% of the variability in the data can be represented by the first 8 principal components, although there are 18 variables being tested. Appendix F shows how each individual variable is related to each principal component.

In the first principal component, there are high positive values for the variables ERA, WHIP, FIP, and AVG (all above 0.3), and high negative values for Win%, Strike% and IP (all below -0.25). This means that these 7 variables have the highest influence on the first principal component. The biplot in Appendix G also shows how each variable affects the first two principal components.

We see that Strike%, Win%, GS, IP and K/9 are grouped together on the left side of the biplot. We also can see that ERA, WHIP, AVG, K/9, BB/9, FIP, Barrel%, Pull% and LD% are grouped close by on the right side. This is not a coincidence as all of those variables were seen to have significant relationships with ERA in the multiple regression models.

**Conclusions**

There are a number of variables that have a significant relationship with ERA. WHIP and FIP seem to have the strongest association, although several others have varying degrees of significance as well.

In the dataset, there are also several relationships between other variables, such as Wins/ERA and Win%/Run Support. I was surprised to see that ground ball percentage, GB%, was not highly correlated with ERA. However, in the multiple regression model, line drive percentage was significant. This makes sense as line drives are typically the most well-hit balls, and LD% had a positive coefficient in the final model. As more balls are hit for line drives, the pitcher will have a higher ERA.

A principal component analysis showed that the majority of the variance in the dataset can be represented by very few components. The first component represented 62.5% of variability, the first three represented about 89%, and the first 8 represented 99%. This shows that the variability can be explained without using all of the variables in the dataset.

This project helped illustrate a lot of information in the data, and it helped that I had a strong knowledge of baseball, helping me interpret the results, and knowing whether a statistical result was surprising.

Glossary of Stats

* ERA: Earned Run Average. Calculated:
* WHIP: Walks plus hits per inning pitched.
* FIP: Fielding Independent Pitching. Calculated with the formula:

FIP is a number similar to ERA, except calculated using only outcomes within a pitcher’s sole control (home runs, walks, hit by pitches, and strikeouts).

* LA: Launch Angle: angle above the ground at which the ball comes off the bat
* Barrel: Definition from Fangraphs: A batted ball with comparable hit types (in terms of exit velocity and launch angle) have led to a minimum 0.500 batting average and 1.500 slugging percentage.
* Hard%: Percentage of balls hit at least 95 MPH
* AVG: Batting average opposing hitters had against a pitcher

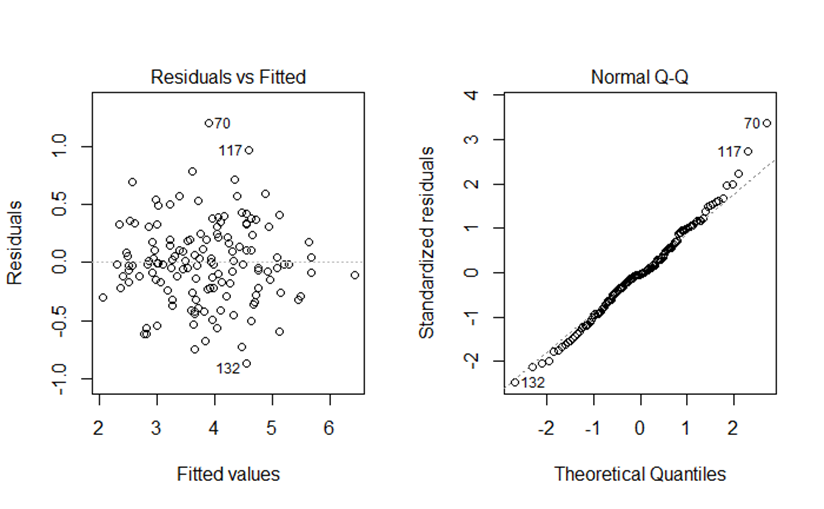
Data Source: Fangraphs.com

Link: <https://www.fangraphs.com/leaders/major-league?pos=all&stats=pit&lg=all&type=8&month=0&ind=0&qual=100&startdate=&enddate=&season1=2022&season=2022>

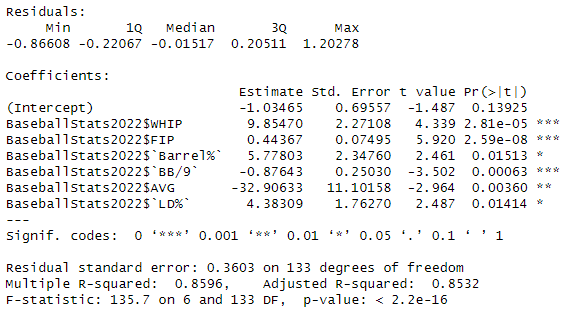
Source on PCA:

Keita, Zoumana. “Principal Component Analysis (PCA) in R Tutorial.” *DataCamp*, DataCamp, 13 Feb. 2023, www.datacamp.com/tutorial/pca-analysis-r.

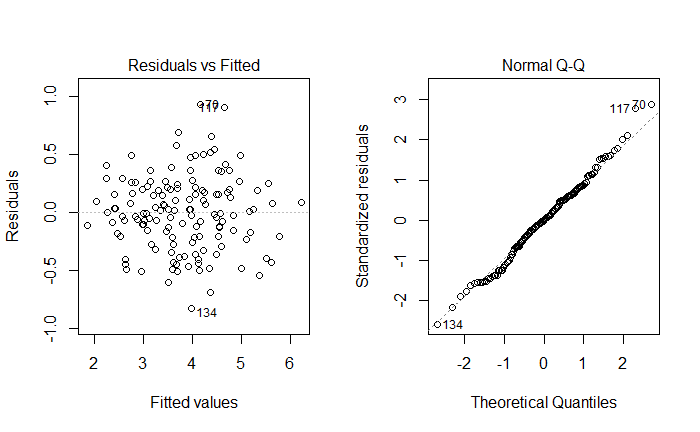
**Appendix A** **– Residual and Normal Q-Q Plots for ERA Model #1**

****

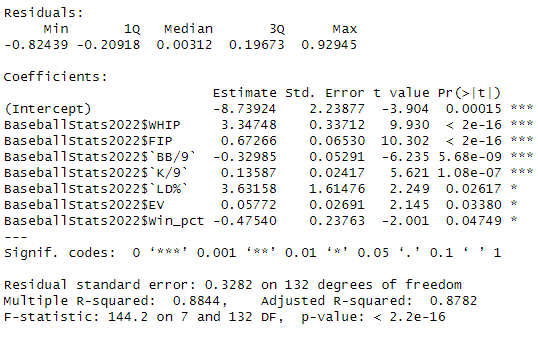
**Appendix B – Summary of ERA Model #1**

****

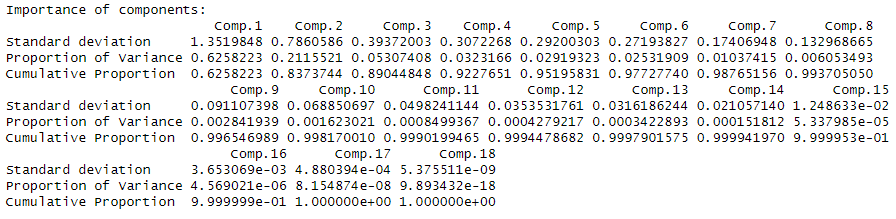
**Appendix C – Residual and Normal Q-Q Plots for ERA Model #2**

****

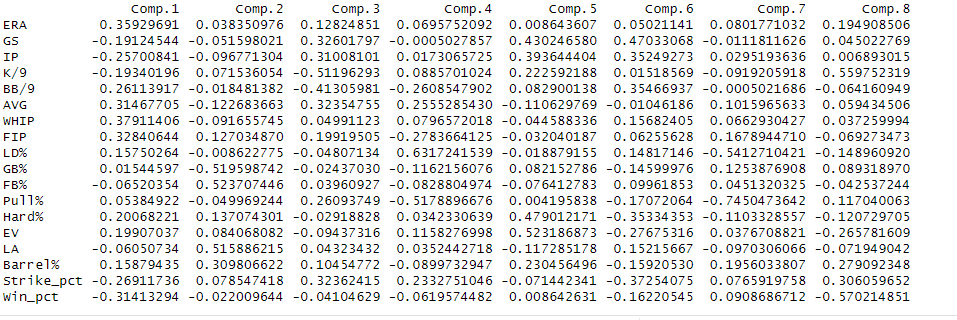
**Appendix D – Summary of ERA Model #2**

****

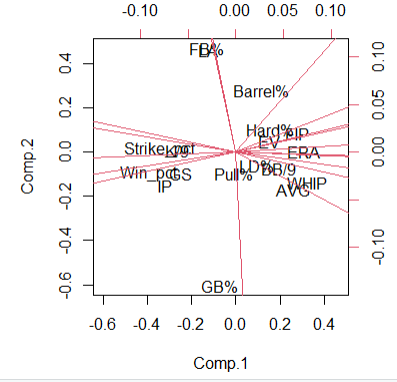
**Appendix E – Principal Component Analysis Summary**

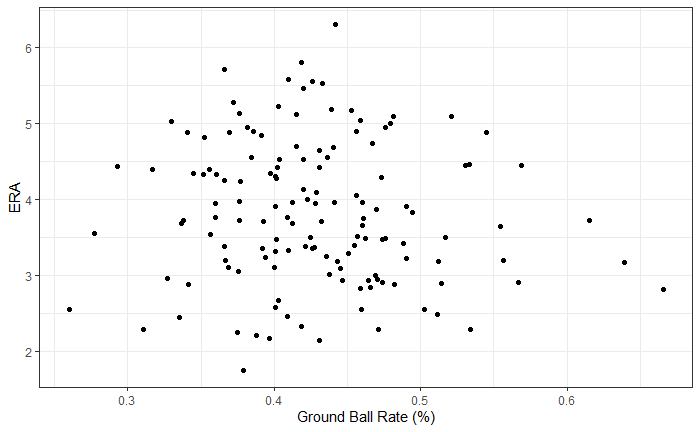
****

**Appendix F – Principal Component Loadings**

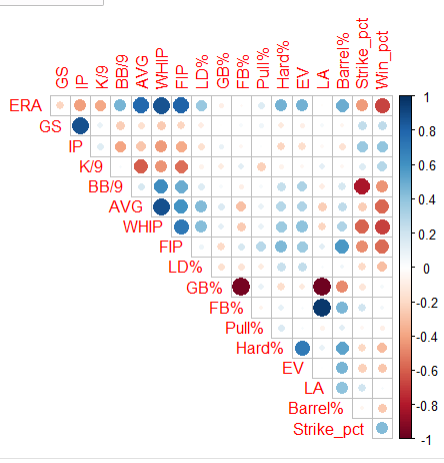
****

**Appendix G – Biplot of Principal Components**

****

**Appendix H – Scatterplot of Ground Ball Rate and ERA**

**Appendix I – Correlation Matrix**

****