Multivariate Analysis of NFL Quarterbacks

Brandon Domash

12/16/2018

Abstract

Throughout this report, I analyze various factors of quarterback play. First, I graphically analyze the difference in quarterback play based on a quarterback's age. Older quarterbacks tend to be better at everything relating to throwing ability, whereas younger quarterbacks tend to run the ball more often. In addition, older quarterbacks make far fewer mistakes than young quarterbacks, an indication that quarterbacks become much better when gaining professional experience. In terms of passing ability, Patrick Mahomes is a clear outlier in this data. The 23-year-old quarterback is primed to be the league's MVP this year, and his numbers are truly remarkable when considering how young he his.

I later assess ways to reduce dimensionality of the data; I am able to account for nearly 70% of the variation in the data using just three principal components. However, when segmenting the columns that are naturally highly correlated, I am able to establish one variable for all passing stats, and one variable for all rushing stats, in which you can clearly categorize each NFL quarterback by both their passing and running abilities. I also used principal component analysis to conclude that both Total QBR and QB Rating are good summary tools in analyzing each quarterback's performance.

I finally used cluster analysis to group quarterbacks into three tiers of skill. Quarterbacks in the top tier, meaning the league's best quarterbacks, are basically the only quarterbacks capable of leading their team to a Super Bowl victory. These quarterbacks are primely composed of older quarterbacks, although there are a few young quarterbacks in this tier such as Patrick Mahomes and Jared Goff. However, teams with middle-tier quarterbacks are still capable of leading their team to a strong record, and possible a playoff run. The lower tier quarterbacks are primarily young quarterbacks who are currently not capable of leading their team to much success, but rather need time to develop and improve their skills.

Introduction

The data for this analysis contains stats for 34 quarterbacks through 10 weeks of play of the 2017 NFL season. The quarterbacks contained in this dataset are the QBs who have thrown enough passes to be qualified for the NFL's passing leader title (in this case, that means QBs who have attempted at least 150 passes at this point in the season). The features included for the dataset are most of the commonly used statistics for analyzing quarterback play. For those not familiar with football, a quarterback can help his team by either throwing the ball or running the ball. Quarterbacks can hurt their team by either

turning the ball over or taking a sack. Thus, the features can generally be grouped into four groups:

- Stats relating to a quarterback's passing
    - Yds: Passing yards
    - TD: Passing touchdowns
    - Cmp.PCT: Percentage of throws for completions
    - First.DownPCT: Percent of completions that gained a first down
    - Passes20Yd: Number of completed passes that gained at least 20 yards
- Stats relating to a QB's rushing
    - Rush.Att: Number of rushing attempts
    - Rush.Yds: Number of rushing yards
    - Rush.TD: Number of rushing touchdowns
- Stats relating to a QB's turnovers and negative plays (ie interceptions)
    - Int: Number of interceptions
    - Fumbles: Number of fumbles
    - Sacks: Number of sacks taken
- Stats relating to a QB's general ability
    - QB.Rating: Quarterback Rating
    - QBR: Total QBR
    - SB: Whether or not the quarterback has won a Super Bowl

The final feature that is included in the dataset is the quarterback's age. The data comes from www.pro-football-reference.com, and these statistics are some of the most commonly used stats when analyzing a quarterback's play. There are a total of 34 QBs and 16 features for each QB in the dataset.

Goals

There are a few goals of the following analysis.

- How does a quarterback's age relate to his performance? I am particularly interested in seeing whether quarterbacks change their style of play over time, or if quarterbacks get better or worse as they get older.
- Can we describe a QB's play with fewer statistics? There are so many statistics to evaluate quarterback play, but most people do not know which variables are most

important and how to summarize quarterback play. It would be a lot simplier if there were a few stats that analysts could use to describe quarterback performance.

- Do the statistics Total QBR and QB rating accurately measure a quarterback's performance? The Total QBR statistic was created by ESPN in 2011 as a measure of performance of quarterbacks in American football, based on the quarterback's "expected points added" for each play. The calculations for this stat are rather convoluted and intricate, and thus many people are skeptical whether or not this stat bears any significane. QB rating, also known as passer rating, has been the standard measure of quarterback play since the NFL's inception, which is a calculation based on a quarterback's basic statistics. With this analysis, I aim to see if these statistics are indicative of a QB's overall performance.
- Are there subgroups or tiers of quarterback play in the NFL? This topic accounts for the largest percent of discourse for NFL fans and analysts alike. Specifically, people love to discuss which quarterbacks are "elite". I hope to answer this question by seeing which quarterbacks can be grouped together.
- How are Super Bowl winning quarterbacks different from the non-superbowl winning QBs? Are there specific traits that Super Bowl winning quarterbacks have?

Main Results

To analyze how a quarterback's age relates to his performance, I first explore the data by breaking the data into 3 groups: "old quarterbacks", which are quarterbacks over the age 30. This is a common age at which analysts begin to question if a quarterback is past their prime. The second group is "young quarterbacks", which are quarterbacks who are 25 and under. These quarterbacks are still relatively unproven and for the most part are still in their first contract. The final group is "middle-age quarterbacks", which are the quarterbacks who are neither young nor old, rather they should be in the prime of their career. I first compare the three subsets of quarterbacks in *Table 1.*

Table 1

```
## [1] "Old QBs"

##         W.PCT       Cmp.PCT           Yds            TD           Int
##          0.53         66.52       2961.18         19.64          6.73
##         Sacks First.DownPCT    Passes20Yd      Rush.Att      Rush.Yds
##         21.73         37.67         39.55         23.36         79.18
##       Rush.TD       Fumbles     QB.Rating           QBR            SB
##          1.18          4.18        100.33         65.09          0.45

## [1] "Middle QBs"

##         W.PCT       Cmp.PCT           Yds            TD           Int
##          0.44         66.68       2512.36         17.18          7.55
##         Sacks First.DownPCT    Passes20Yd      Rush.Att      Rush.Yds
##         24.55         35.54         30.91         30.91        137.00
```

```
##       Rush.TD       Fumbles     QB.Rating           QBR            SB
##          0.73          5.73         95.80         54.11          0.09

## [1] "Young QBs"

##         W.PCT       Cmp.PCT           Yds            TD           Int
##          0.50         62.69       2157.00         15.00          8.33
##         Sacks First.DownPCT    Passes20Yd      Rush.Att      Rush.Yds
##         23.58         36.13         29.42         37.67        177.67
##       Rush.TD       Fumbles     QB.Rating           QBR            SB
##          1.75          5.83         89.48         55.40          0.00
```

By *Figure 1*, it is evident simply by subsetting the data that there is a clear distinction between QB performance and age group. Old QBs have the highest average in nearly every positive statistic, and the lowest average in nearly every negative statistic. However, I am still interested in how age and other variables are related. I first analyze how age is related to the passing yards and the percentage of passes he completes.

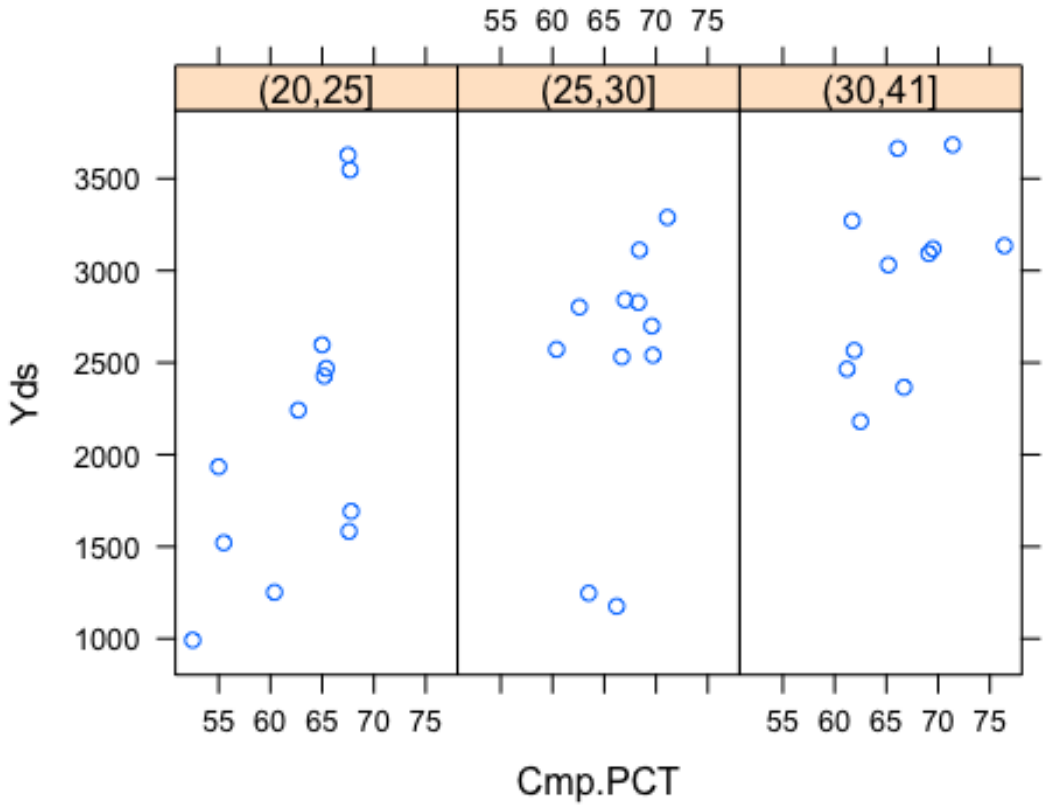

Figure 1

Figure 1 shows that difference in distribution amongst the 3 age groups is clear. Young QBs have much lower completion percentages and yards passing than the other 2 groups. The middle group has two quarterbacks who are low in yards as well, while the oldest group has only quarterbacks that are both high in yards and completion percentage. However, these are not the only passing statistics for a quarterback. To get a better idea of how age relates to passing in total, I look at how age is related to each of passing statistic.

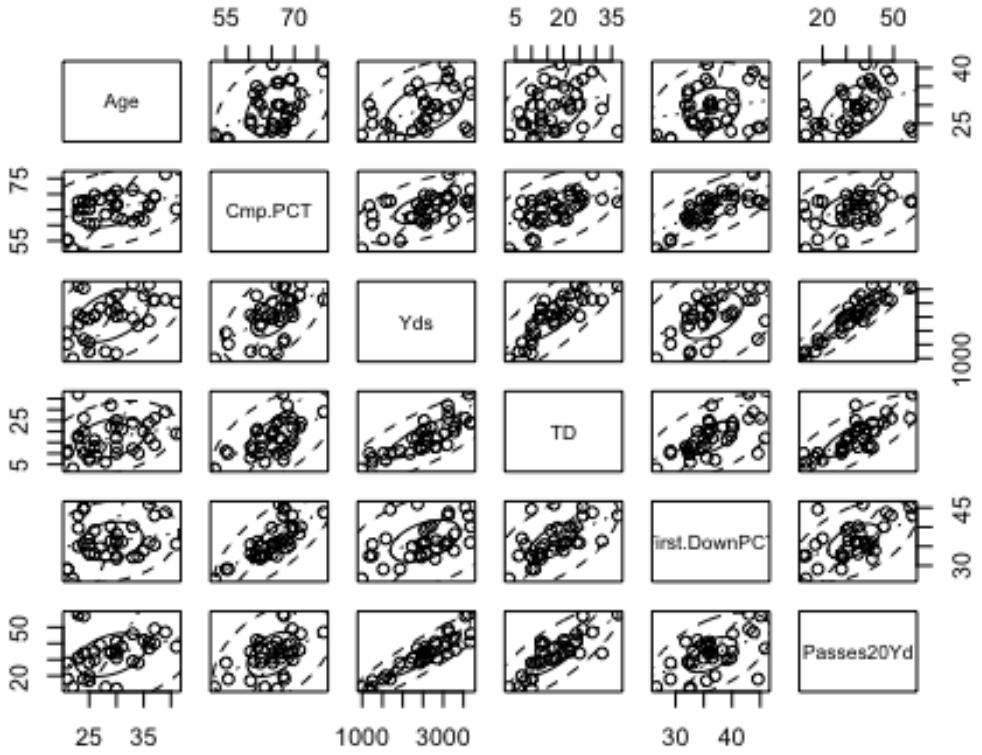

Figure 2

In *Figure 2* above, the first column shows the relationship between age and each other variable. Each of these variables are measured so that higher amounts indicate a performance. We see a positive, rather linear relationship between each of these variables and the age of the quarterback. Thus, quarterbacks who are older seem to have more passing ability than younger quarterbacks. The one interesting thing to note is the outlier

in the TD vs Age plot, which is plotted below.

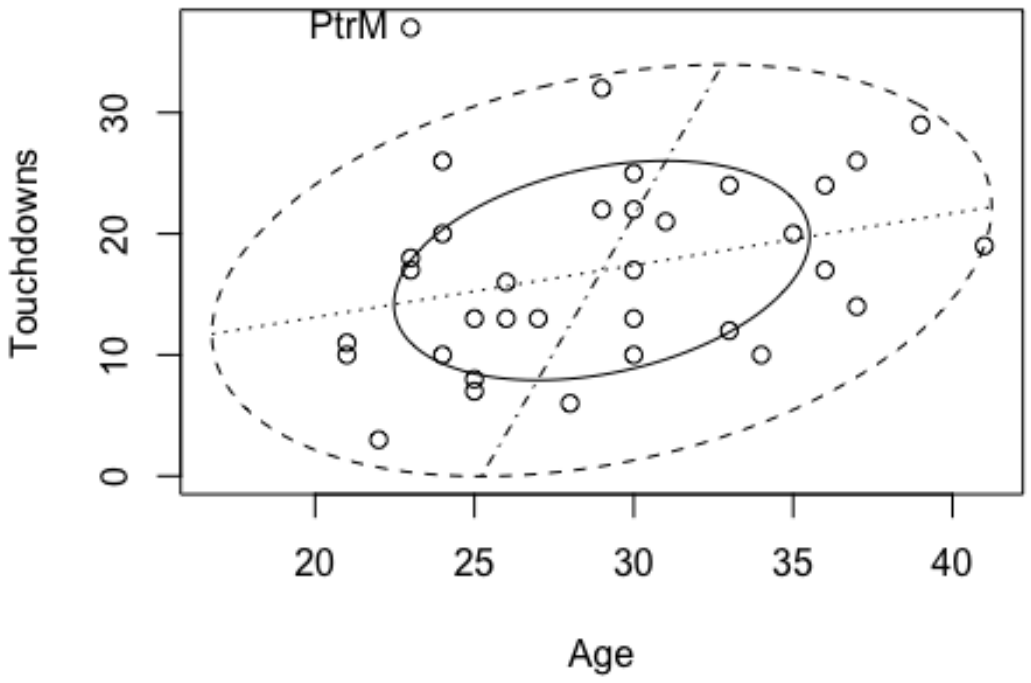

Figure 3

Figure 3 shows a clearly apparent positive relationship between touchdowns thrown and a quarterback's age. The one quarterback of interest here is Patrick Mahomes, who is the quarterback of the Kansas City Chiefs. The second year 23-year-old quarterback has taken the league by storm this season and is currently the favorite to win the Most Valuable Player award in the NFL. Having that many touchdowns for a quarterback that young is very rare, as seen by the data, which is a large factor into why he is viewed as the most valuable player to many around the league.

While older quarterbacks tend to have higher amounts of positive passing stats, these stats do not account for negative passing plays; interceptions for example are one of the main indicators of bad throws for a quarterback. I thus inspect how a quarterback's touchdowns relate to their interceptions, with a focus on a quarterback's age. This relationship is shown in *Figure 4* below, where the color of each point corresponds to the quarterback's age.

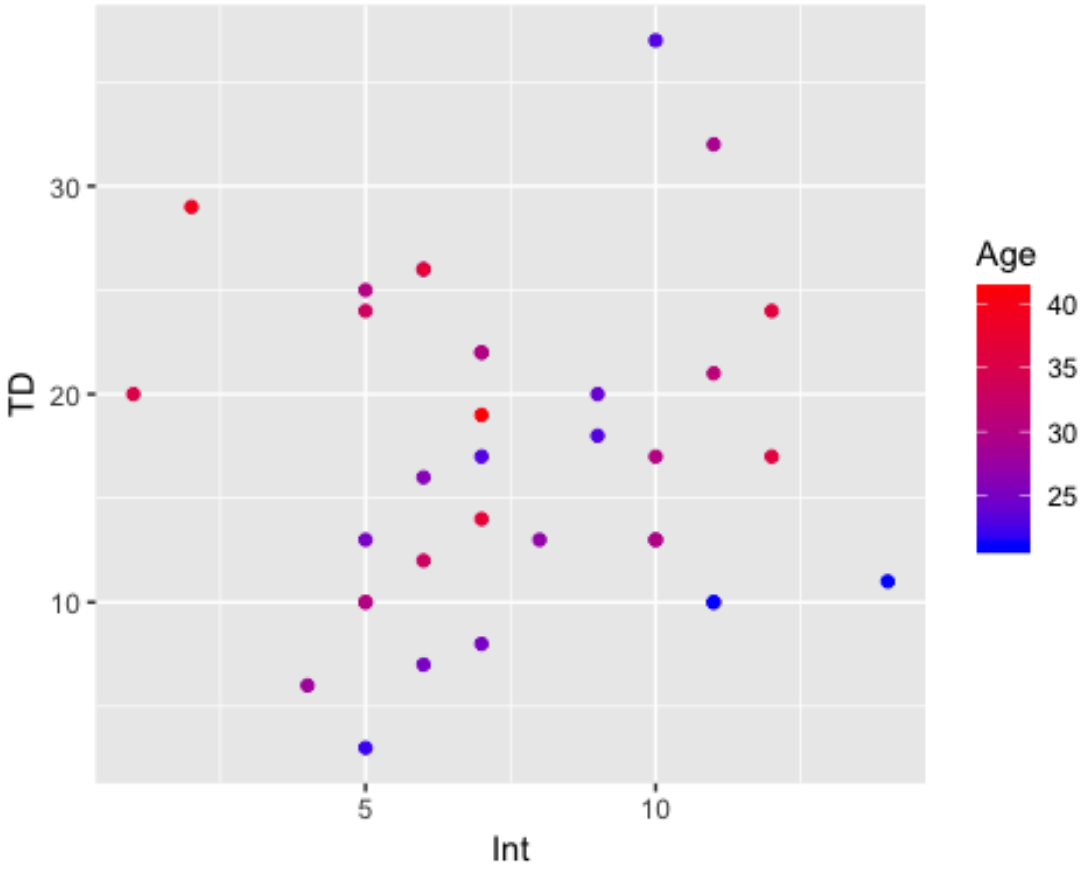

Figure 4

Surprisingly, there is a positive relationship between the number of touchdowns thrown and the number of interceptions a quarterback throws. This could possibly be explained by the number of passing attempts a quarterback throws. However, the upper left region, indicating quarterbacks that have thrown many touchdowns and few interceptions, appears to be dominated by older quarterbacks. This supports the hypothesis that older quarterbacks are better passers and make less mistakes.

However, passing is not the only part of the game. Quarterbacks can also help their team by running the ball, and a quarterback's ability to run the ball can add another dimension to a team's offense. Thus, I see if there is a relationship between a quarterback's running ability

and age. Figure 5 compares a quarterback's running stats to his age.

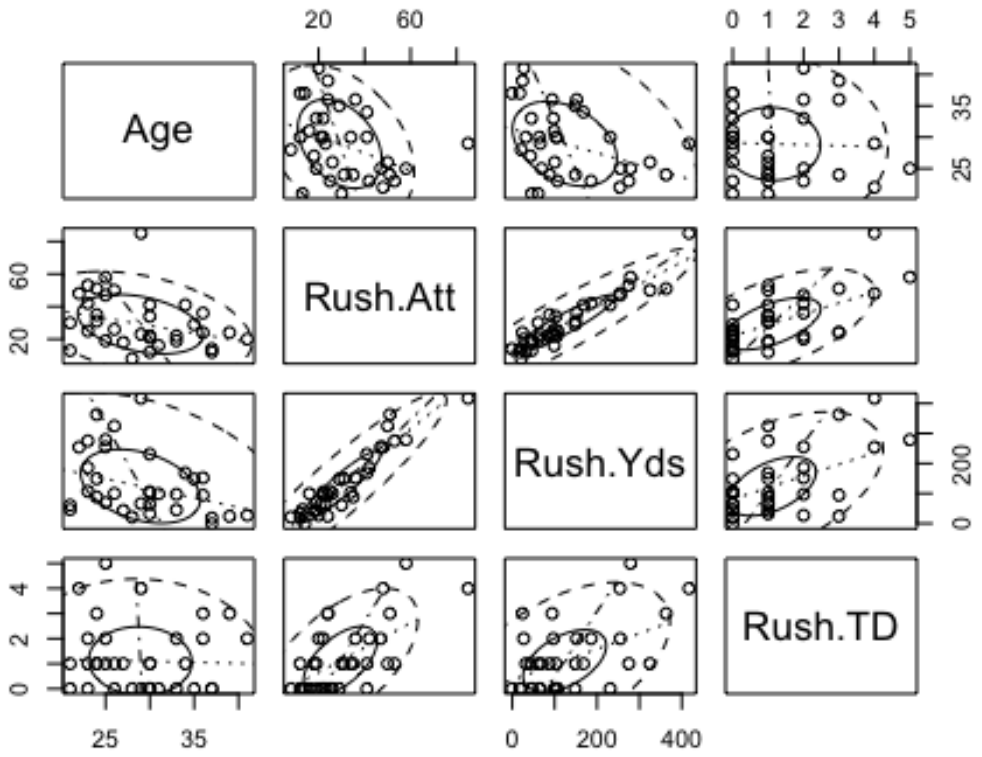

Figure 5

Here we see the opposite results to the trend found with passing stats. There appears to be a negative relationship between a quarterback's age and the amount of times he runs the ball, as well as his age and his number of rushing yards. There does not seem to be much of a relationship between rushing touchdowns and age, however. This could be explained as older quarterbacks are likely well beyond their athletic prime, and thus rely on their throwing abilities rather than their running abilities. The one major outlier in this data can be seen in the age vs rushing attempts plot. The outlier is Carolina Panther's quarterback Cam Newton, who is commonly known throughout the league as a "running QB". It will be interesting to see if as he gets older, his volume of rushes will decrease. Based on the data, my hypothesis is that it will.

Finally, I look at each quarterback's passing yards vs rushing yards, to get a sense whether there are clear cut "passing quarterbacks" and "running quarterbacks".

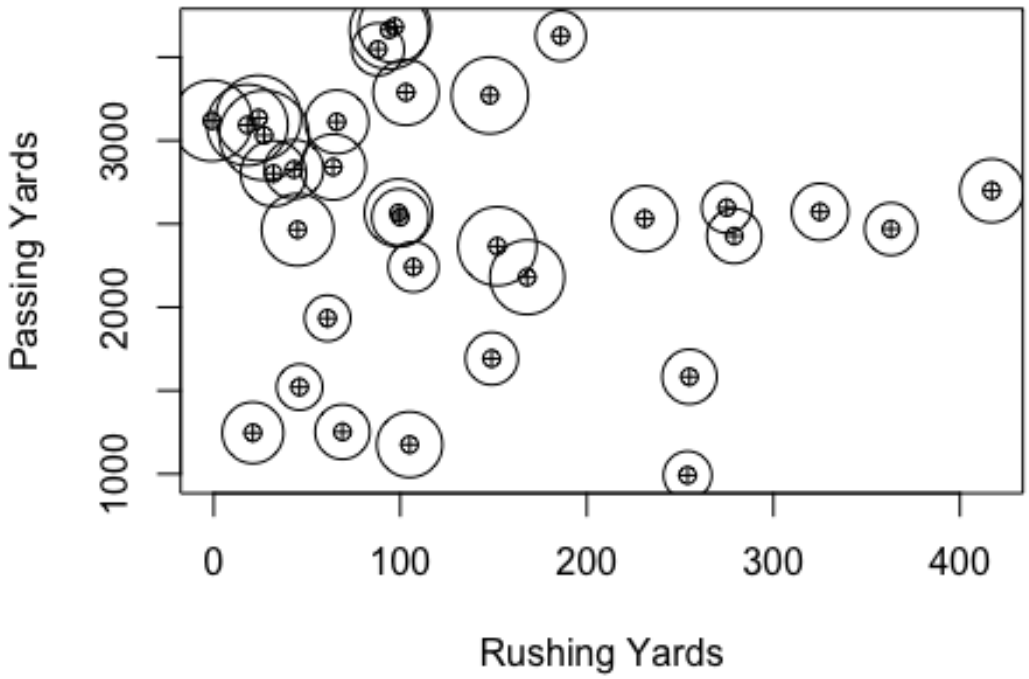

Figure 6

The size of the circles in *Figure* 6 correspond to the quarterback's age. The plot supports the hypothesis that younger quarterbacks tend to run the ball for more yards and throw the ball for fewer yards than older quarterbacks, as seen by the size of the circles for each data-point. After analyzing the all of the data through these plots, I can confidently say that older quarterbacks tend to be more passing oriented, they make fewer throwing mistakes, and have less of a desire to run the ball.

Next, I seek to reduce dimensionality in order to describe quarterbacks using fewer than 16 variables. To do so, I use principal component analysis (PCA) in order to find linear combiations of the variables that account fot the greatest amount of variation in the data. Because the scale of each variable of the data is much different, it is necessary to use the correlation matrix in computing the principal components rather than the using the covariance matrix. Below is a summary of the principal components of the data.

Table 2

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     2.4680622 1.7545830 1.3428547 1.2219702 1.01987227
## Proportion of Variance 0.3807082 0.1924101 0.1127037 0.0933257 0.06500872
## Cumulative Proportion  0.3807082 0.5731183 0.6858219 0.7791476 0.84415636
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.81364746 0.69833427 0.61605882 0.50486285
## Proportion of Variance 0.04137639 0.03047942 0.02372053 0.01593041
## Cumulative Proportion  0.88553275 0.91601217 0.93973270 0.95566310
##                           Comp.10    Comp.11     Comp.12     Comp.13
## Standard deviation     0.48114619 0.44536318 0.361258070 0.261091783
## Proportion of Variance 0.01446885 0.01239677 0.008156712 0.004260557
## Cumulative Proportion  0.97013196 0.98252873 0.990685443 0.994946001
##                            Comp.14     Comp.15      Comp.16
## Standard deviation     0.220842823 0.162891689 0.0745569375
## Proportion of Variance 0.003048222 0.001658356 0.0003474211
## Cumulative Proportion  0.997994223 0.999652579 1.0000000000
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Age            0.211  0.287  0.138  0.347  0.167  0.401         0.289
## W.PCT          0.277 -0.103 -0.222  0.174  0.175 -0.562        -0.306
## Cmp.PCT        0.326                       -0.421  0.318        -0.240
## Yds            0.341         0.222 -0.226  0.251         0.217
## TD             0.359                -0.200  0.149 -0.207 -0.150  0.154
## Int                        -0.264 -0.578  0.418  0.402 -0.214 -0.136
## Sacks                -0.183  0.636 -0.174 -0.141         0.218
## First.DownPCT  0.311        -0.207 -0.145 -0.354  0.298 -0.103
## Passes20Yd     0.339         0.130 -0.241  0.231 -0.103  0.323  0.236
## Rush.Att             -0.534         0.129                       0.301
## Rush.Yds             -0.522         0.114               -0.220  0.453
## Rush.TD              -0.369 -0.151  0.318  0.296  0.288  0.538 -0.383
## Fumbles              -0.391  0.403 -0.177               -0.279 -0.421
## QB.Rating      0.379                       -0.254
## QBR            0.367        -0.204
## SB             0.163  0.144  0.325  0.383  0.389  0.128 -0.527 -0.161
```

Only 38% of the variance in the data can be accounted for by its first principal component, not nearly enough to fully summarize the data. However, it is intersting to note that the first principal component disregards all rushing data for the quarterbacks, as well as all data on negative plays. Thus, the first principal component can be seen as a measure of a quarterback's passing ability.

In order to get close to 70% of the variation in the data, 3 principal components are necessary, which are plotted below in *Figure 7.*

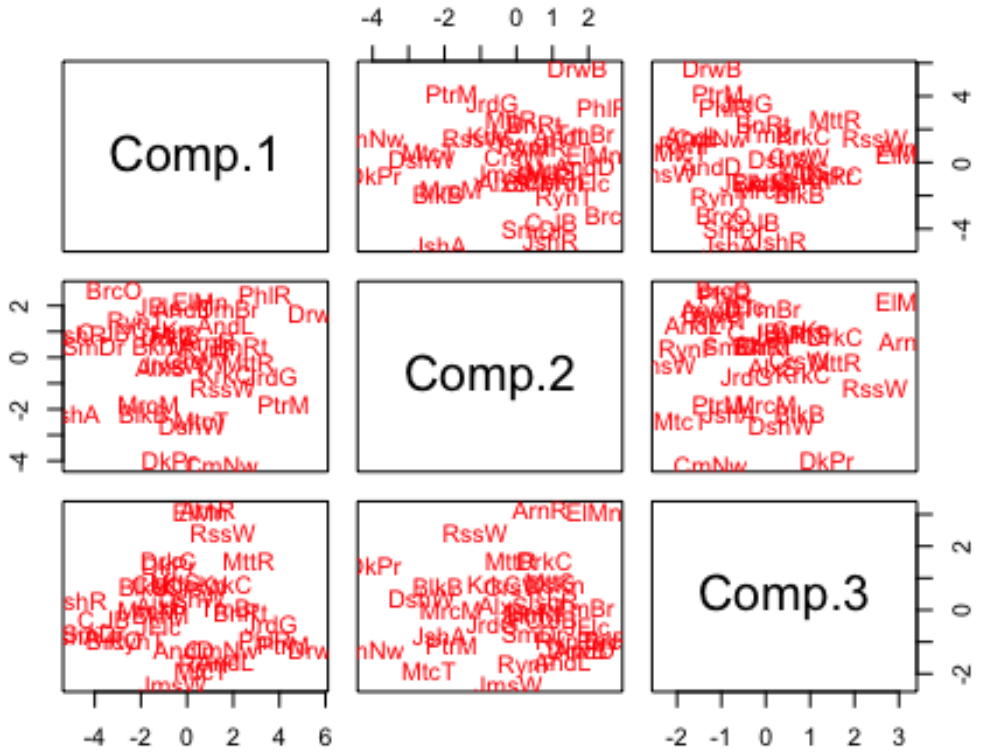

Figure 7

From looking at component scores 1 and 2, a low amount of the first principal component seems to generally correspond to passing ability, as I can note that some of the league's worst passers are the farthest left on the plot, such as Brock Osweiler, Josh Allen and Sam Darnold. Low amounts of the second principal component seem to correspond to a quarterback's running ability, as notable running quarterbacks Cam Newton and Dak Prescott are at the bottom in the second component. It is hard to assess exactly what the third principal component corresponds to based on the loadings and the graphs.

Another way to reduce dimensionality is to reduce each statistical category into principal components. That is, have principal components for throwing varibles, running variables, and turnover variables. Below are our passing variables, which are highly correlated with each other.

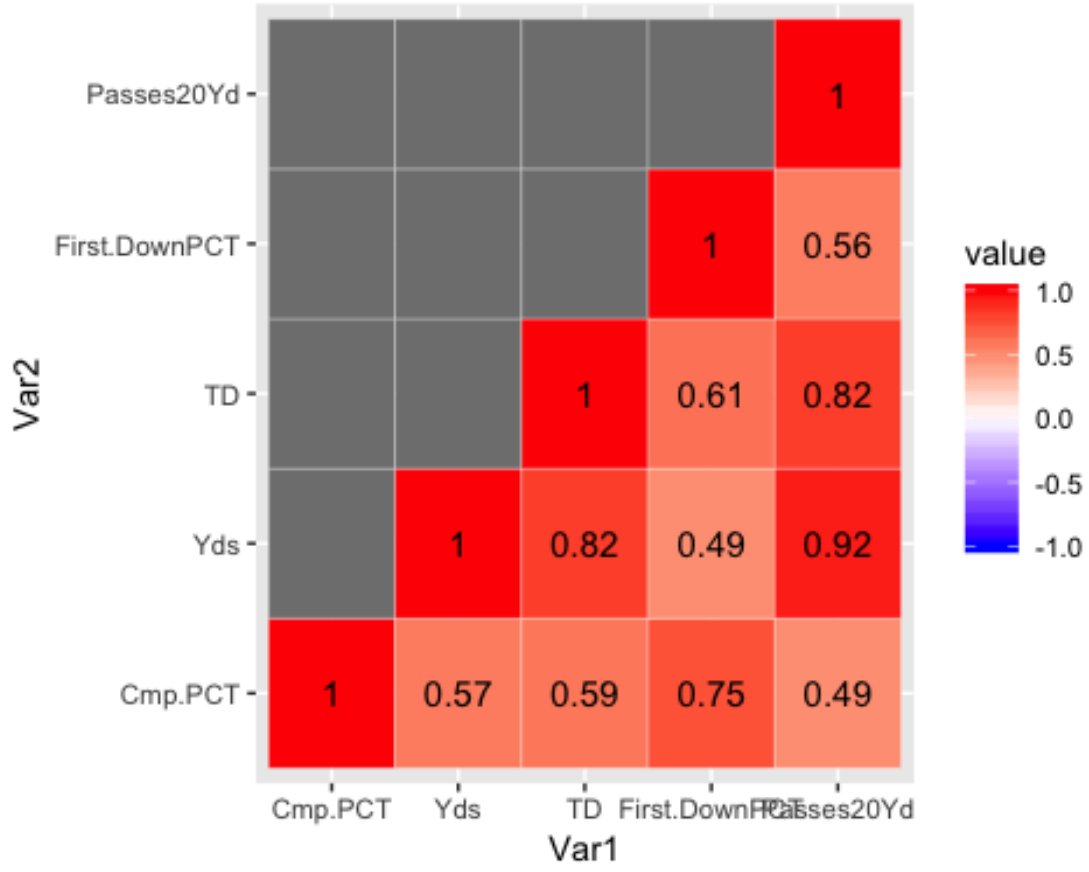

Figure 8

Ideally, I would be able to reduce these five highly correlated passing variables into one or two variables. A summary of the principal component analysis below tells the following:

Table 3

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     1.912017 0.9029779 0.52715995 0.44718551 0.22571742
## Proportion of Variance 0.731162 0.1630738 0.05557952 0.03999498 0.01018967
## Cumulative Proportion  0.731162 0.8942358 0.94981535 0.98981033 1.00000000
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Cmp.PCT         0.409  0.567  0.668         0.252
## Yds             0.471 -0.391  0.318 -0.258 -0.677
## TD              0.473 -0.203 -0.192  0.835
## First.DownPCT   0.409  0.573 -0.625 -0.225 -0.254
## Passes20Yd      0.470 -0.396 -0.162 -0.428  0.642
```

As seen in *Table 3*, nearly 75% of the variation can be explained using one principal component, which is a significant amount of the variation in the data. A biplot of this data also shed some insight into how quarterbacks relate to each other in terms of passing. The the distance between each point in *Figure* 9 represents the distance between the units, and the difference in angles corresponds to the correlation between the variables.

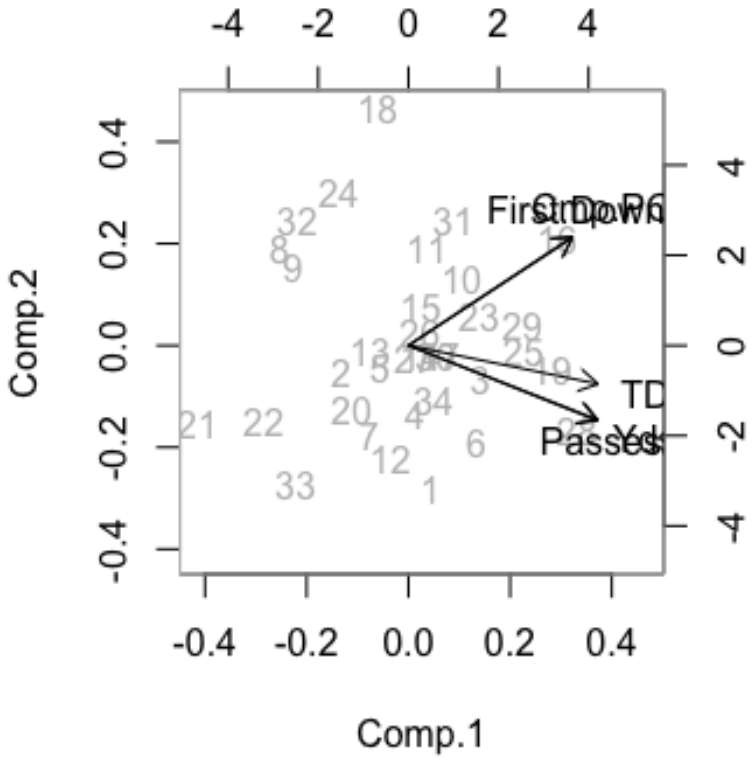

Figure 9

The number of long passes and passing yards are clearly significantly correlated from *Figure 9*, as are first down percentage and completion percentage. The plot also that quarterback 28, Patrick Mahomes, is particularly unique in terms of his long passes and passing yards. This is an interesting contrast from quarterback 16, Drew Brees, who excels in completion percentage and first down percentage. Interestingly enough, these are the

two quarterbacks who are the top two candidates to win the most valuable player award this season.

Next, plotting the correlation matrix between running variables also shows that the variables are very highly correlated, as seen in *Figure 10*. Thus I use principal component analysis again, hoping that the first principal component will likely account for much of the variation.

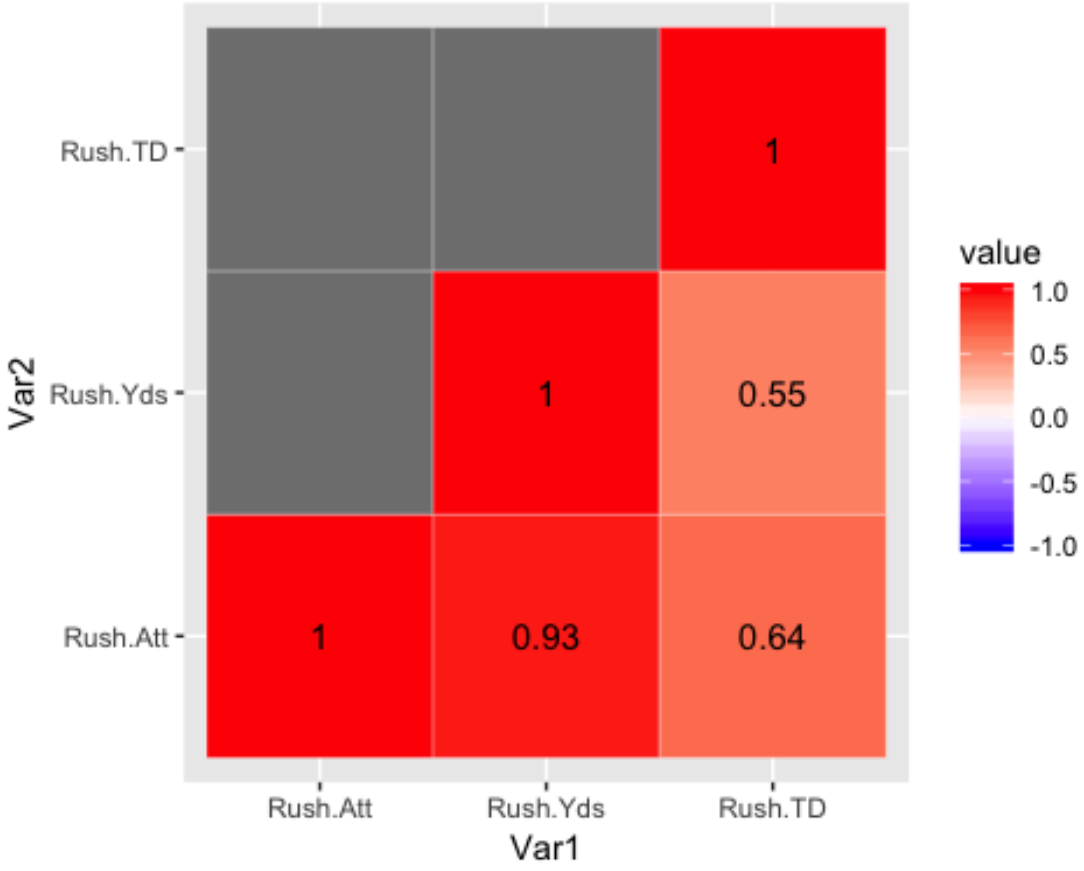

Figure 10

Table 4

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     1.5598114 0.7127471 0.24285804
## Proportion of Variance 0.8110039 0.1693361 0.01966001
## Cumulative Proportion  0.8110039 0.9803400 1.00000000
##
## Loadings:
##          Comp.1 Comp.2 Comp.3
## Rush.Att  0.619  0.270  0.738
```

```
## Rush.Yds  0.599  0.446 -0.665
## Rush.TD   0.509 -0.853 -0.115
```

Table 4 above shows that the three rushing variables can be reduced into 1 principal component while still accounting for over 80% of the variation in the data.

Finally, I look at the correlation matrix for features related to negative plays for each quarterback, as seen in *Figure 11.*

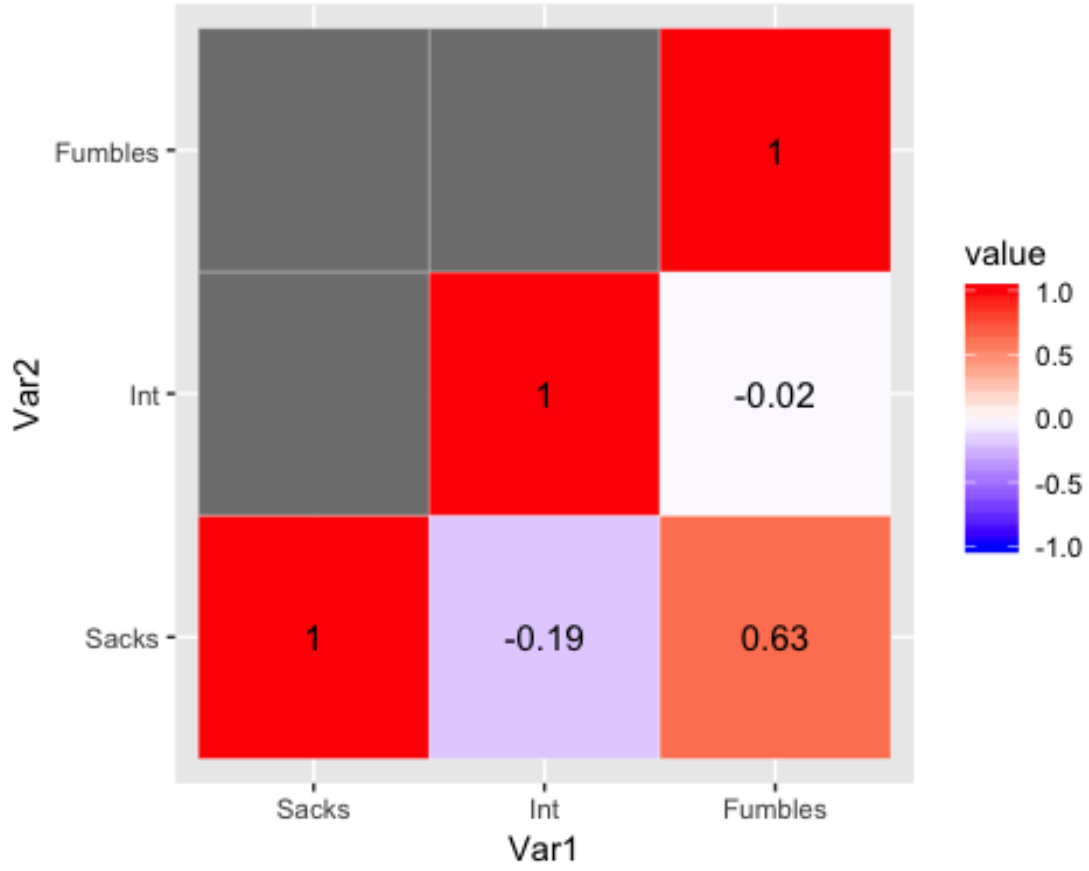

Figure 11

Table 6

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3
## Standard deviation     1.2915117 0.9944844 0.5856605
## Proportion of Variance 0.5560008 0.3296664 0.1143327
## Cumulative Proportion  0.5560008 0.8856673 1.0000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3
## Sacks    0.704         0.710
## Int     -0.220  0.958  0.185
## Fumbles  0.675  0.286 -0.680
```

Here the variables are not that correlated, and thus only 55% of the variation is explained with one principal component, as seen in *Table 6.* I choose to ignore that principal component and can still graph the two first principal components for passing stats and rushing stats to help better visualize each quarterback's style.

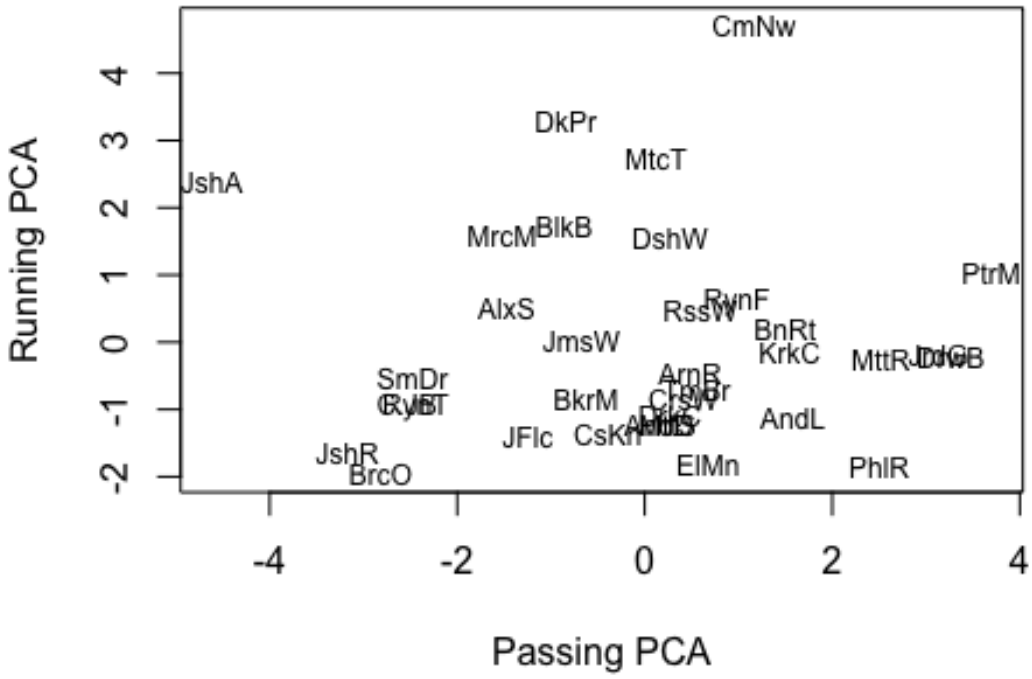

Figure 12

Figure 12 allows us to conceptualize which quarterbacks are good passers vs which quarterbacks are good runners, while accounting for most of the variation in both sets of variables. The two outliers seen earlier are shown clearly here: Cam Newton for his running abilities and Patrick Mahomes for his passing abilities. While there are not clearly defined clusters by this graph, you can compare which quarterback's styles are relatively similar. For example, Josh Rosen and Brock Osweiler, in the lower left corner, both lack passing and running abilities. We also see how terrible of a passer Josh Allen has been this year.

Next, I seek if the derived principal component scores relate to the measures of quarterback play in the data, specifically Total QBR and QB rating. As a reminder, Total QBR is a stat created by ESPN using play-by-play data and complex modeling in order to quantify the total value of a quarterback, between 0 and 100. QB Rating, or passer rating, is a simple formula using a transformation of basic stats to come up with a number between 0 to 158.3, which rates how good a quarterback was in terms of passing during a particular game or season. In theory, the principal component scores should measure both of these statistics fairly well. Because a high QBR and QB Rating indicates a good performance, I first need to make sure that for all of the variables, a higher value is indicitive of a good performance. Thus, I must first transform three statistics: fumbles, interceptions, and sacks, as a lower number for each of these stats indicates a better performance. I will then recompute the principal component scores using a dataset without QBR and QB Rating.

Table 7

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.0497 1.6917 1.3054 1.14000 0.97310 0.74720
## Proportion of Variance 0.3232 0.2202 0.1311 0.09997 0.07284 0.04295
## Cumulative Proportion  0.3232 0.5433 0.6744 0.77439 0.84723 0.89018
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.68906 0.59016 0.47202 0.45126 0.30655 0.21487
## Proportion of Variance 0.03652 0.02679 0.01714 0.01566 0.00723 0.00355
## Cumulative Proportion  0.92670 0.95349 0.97063 0.98630 0.99353 0.99708
##                           PC13
## Standard deviation     0.19492
## Proportion of Variance 0.00292
## Cumulative Proportion  1.00000

##         W.PCT       Cmp.PCT           Yds            TD           Int
##   0.336158990   0.371019762   0.431886754   0.438561343   0.056255426
##         Sacks First.DownPCT    Passes20Yd      Rush.Att      Rush.Yds
##  -0.017690804   0.361271982   0.428273104   0.074132503   0.006456221
##       Rush.TD       Fumbles            SB
##   0.092813354  -0.074853685   0.182514693
```

Table 7 once again shows that not much of the variation explained by the first principal component. Regardless, the variables with the most weight are winning percentage, completion percentage, passing yards, first down percentage, touchdowns, and long passes. I compute the first principal component score for each quarterback by taking the linear combination of the explanatory variables with the weights given by the first principal component, and then see how it relates to both QBR and QB rating.

Figure 13 shows the relationship between PC1 Score and Total QBR on the left, and the relationship between PC1 Score and QB rating on the right. The correlation between PC Score and Total QBR is 0.87, while the correlation between PC1 and QB Rating is 0.9. Both are very high, meaning the first principal component of these throwing, running, and turnover stats are a fairly accurate measure of QB performance, even though the first component accounts for a relatively low amount of the total variation in the data. Whereas

ESPN uses complex analysis and play-by-play data, they could instead use principal component scores and not be too far off.

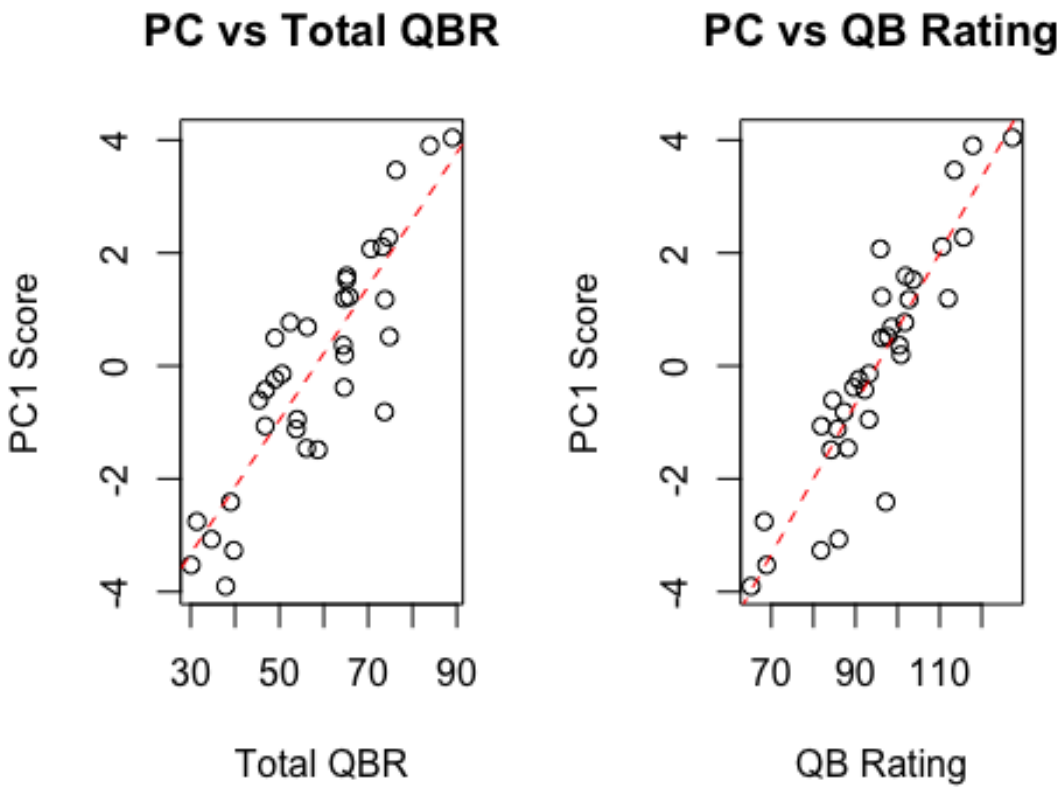

Figure 13

After analyzing the residuals, the highest residuals for the QBR plot includes Jameis Winston, Josh Allen, and Cam Newton, all of whom were vaguely classified as rushing quarterbacks, stylistically (from *Figure 12*). Amongst the highest residuals for the QB Rating plot is Ben Roethlisberger and Sam Darnold, two of the players with the most interceptions. After looking further into the calculations for QBR and QB Rating, QBR takes all plays into account (including rushing plays), while QB Rating only takes passing plays (including interceptions). Thus is makes sense why some players for both plots had large residuals; the first principal components do not take rushing plays or interceptions into account, respectively.

Next, I wish to cluster quarterbacks in order to find tiers of quarterback rankings. I aim to cluster the data using the hierarchical method. I do not want to use k-means clustering as this will result in different clusters every time I use the algorithm, depending on the starting point. The point of this analysis is to find definitive tiers, thus I want definitive clusters. However, this method does have the drawback that I must pick the "correct" number of clusters. In deciding this, I look at the dendrogram for complete linkage. I use complete linkage because I want to avoid chaining clusters, where observations are added to a cluster due to being close to a single observation, rather than being close to the cluster as a whole. This is the case with single linkage, which does not give a useful description of the data, as seen with the dendrogram below (*Figure 14*). Rather, I want to have compact clusters as seen in the dendrogram for complete linkage. For the sake of comparisons, the dendrograms for single, complete, and average linkage are all below.

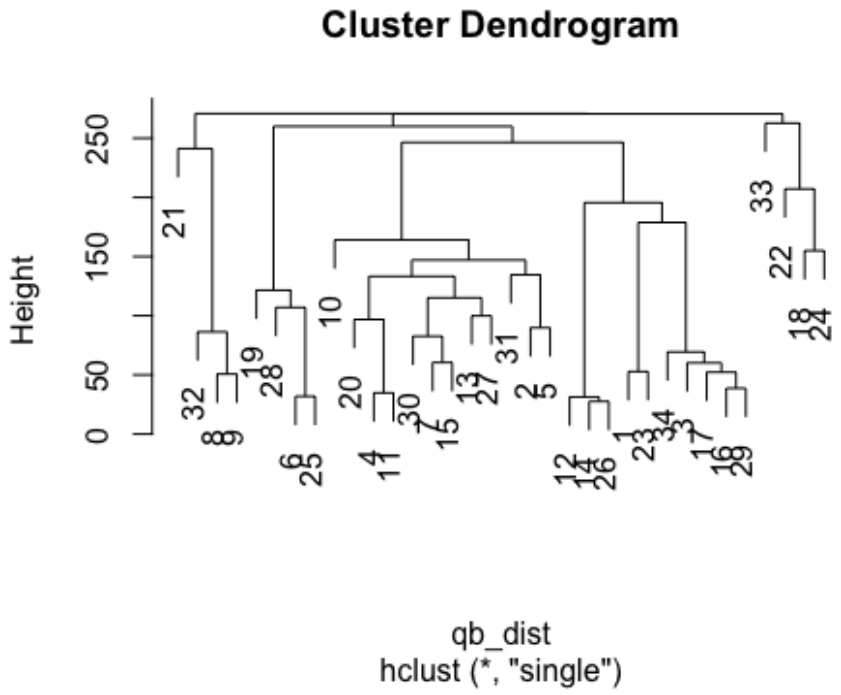

Figure 144

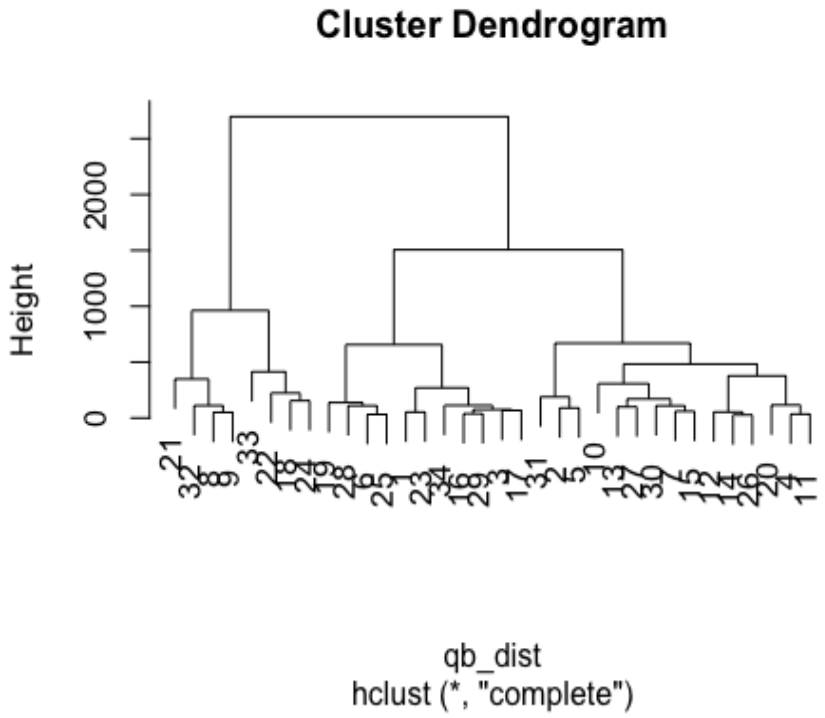

Figure 15

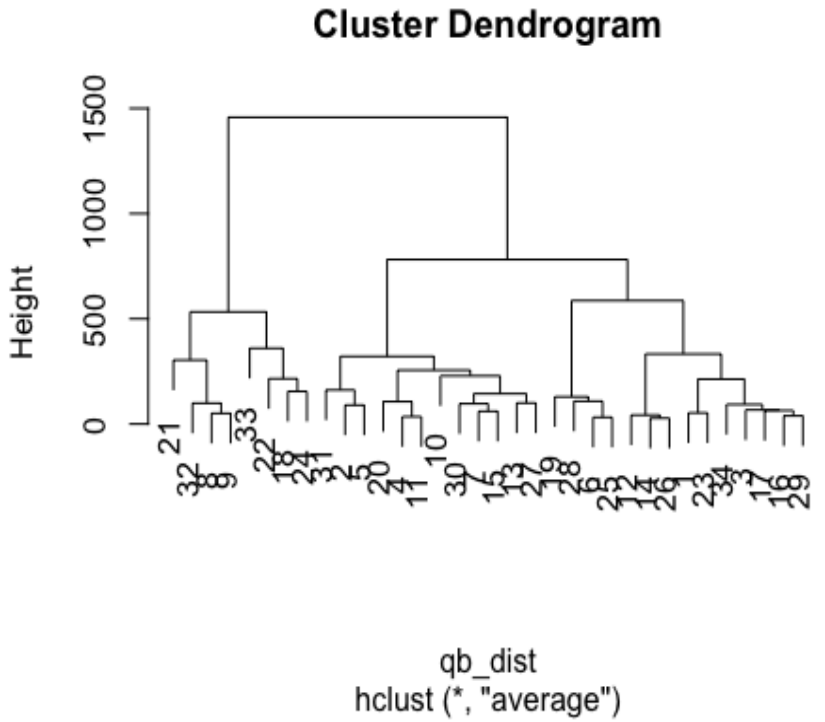

Figure 16

Now I run into the issue of choosing the "correct" amount of clusters, that is, where should I cut the dendrogram? One method for solving this issue is to look at how many groups a model-based clustering technique chooses, such as maximum-likelihood based clustering. However, upon further analysis, the maximum -likelihood model based clustering chooses only one cluster (*Table 8*), which does not help.

Table 8

```
## The following object is masked from 'package:purrr':
##
##     map

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Instead I simply look at the dendrogram and see that there are three fairly equal sized clusters at a height close to 750. *Table* 9 below displays how the 34 quarterbacks are clustered into three groups.

Table 9

```
"Cluster 1"

##  [1] "Aaron Rodgers"      "Andrew Luck"        "Ben Roethlisberger"
##  [4] "Drew Brees"         "Eli Manning"        "Jared Goff"
##  [7] "Kirk Cousins"       "Matt Ryan"          "Patrick Mahomes"
## [10] "Philip Rivers"      "Tom Brady"

"Cluster 2"

##  [1] "Alex Smith"        "Andy Dalton"       "Baker Mayfield"
##  [4] "Blake Bortles"     "Cam Newton"        "Carson Wentz"
##  [7] "Case Keenum"       "Dak Prescott"      "Derek Carr"
## [10] "Deshaun Watson"    "Joe Flacco"        "Matthew Stafford"
## [13] "Mitchell Trubisky" "Russell Wilson"    "Ryan Fitzpatrick"

"Cluster 3"

## [1] "Brock Osweiler" "C.J. Beathard"  "Jameis Winston" "Josh Allen"
## [5] "Josh Rosen"     "Marcus Mariota" "Ryan Tannehill" "Sam Darnold"
```

Below, *Figure* 17 shows a plot of each quarterback listed by their cluster, with the Super Bowl winning quarterbacks shown in red. From *Table 9, c*luster 3 contains the quarterbacks who are struggling mightily, and is comprised mostly of young quarterbacks and quarterbacks who are on the verge of losing their starting job. Cluster 2 contains the middle of the pack Quarterbacks, most of which have seen some levels of success in the NFL, but not as much as the top tier of quarterbacks.

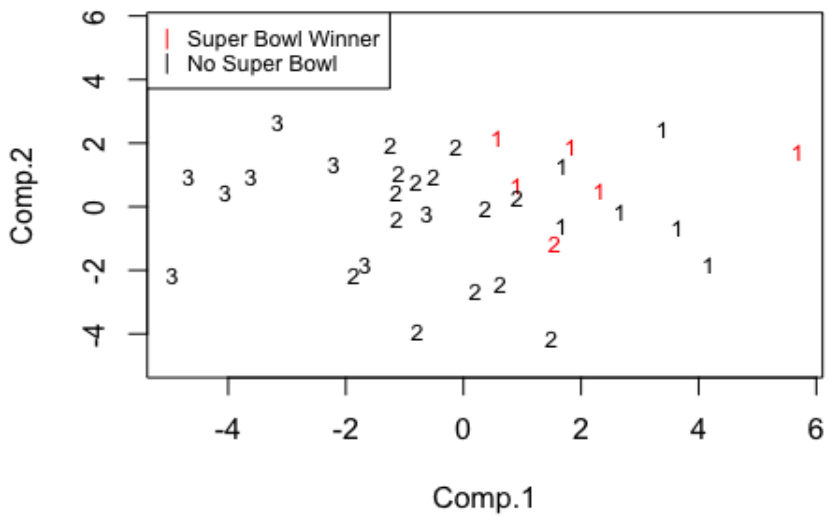

Figure 17

Cluster 1 contains nearly every quarterback in the dataset that has won a Super Bowl. A reminder that principal component one generally measures passing ability helps us undertand that cluster 1 is full of the NFL's best passing QBs, in terms of both passing ability and winning a Super Bowl. In addition, most quarterbacks in cluster 1 have high

amounts of PC2, indicating that they are not quarterbacks that run the ball (PC2 generally measures running ability).

Next, I look at he clusters by the quarterback's team's record (*Figure 18)*. Quarterbacks with more wins than losses are now colored in red in the figure below, whereas quarterbacks with an even (.500) or losing record are displayed in black.

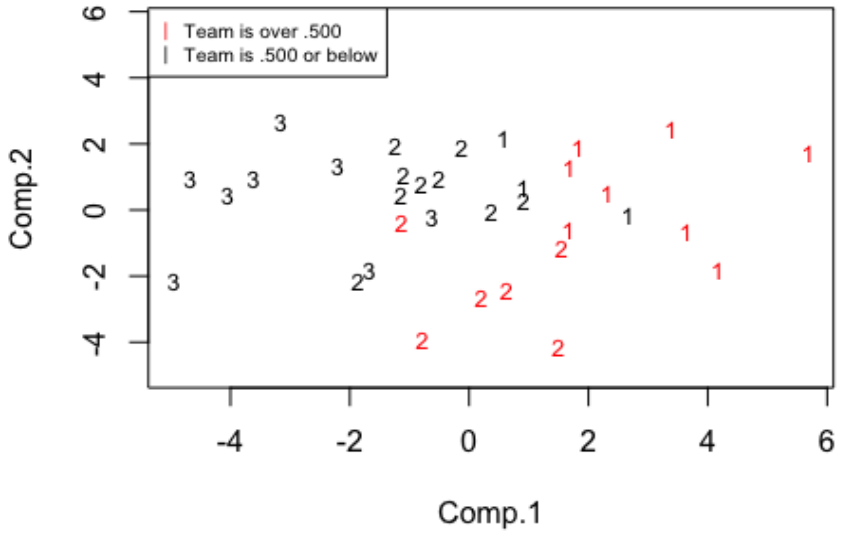

Figure 18

Figure 18 shows how every QB on a winning team is in either cluster 1 and 2. Cluster 2 has more quarterbacks with low amounts of PC2, or a high running ability. This also displays information regarding cluster 3, as cluster 3 contains only quarterbacks who have a losing record. Thus, there is a clear hierarchical ranking of the 3 clusters. Cluster 1 contains the top-tier quarterbacks that can win a team the Super Bowl. These quarterbacks all tend to be of the pocket passer type, meaning they like to stay in the pocket and throw the ball, and are reluctant to run the ball. Cluster 2 contains the middle-of-the-road quarterbacks, that can still lead a team to successful, and perhaps the playoffs, but it is much rarer for these quarterbacks to win a Super Bowl. These quarterbacks also are generally more mobile than the top-tier quarterbacks. Cluster 3 contains the bottom-tier quarterbacks, who likely will not be able to lead a team to the playoffs, and need to improve their overall quarterback play. This severe disparity in performance for teams with low-tier quarterbacks versus top and middle-tier quarterbacks makes a strong case as to why quarterbacks are commonly seen as the most important position in sports.

Finally, to conceptualize how age is distributed by tier, I look at the clusters colored by age (*Figure 19)*. To be consistent with earlier analysis, age is broken into 3 categories: young quarterbacks (25 and under), middle-aged quarterbacks (26-30), and veteran quarterbacks (31+).

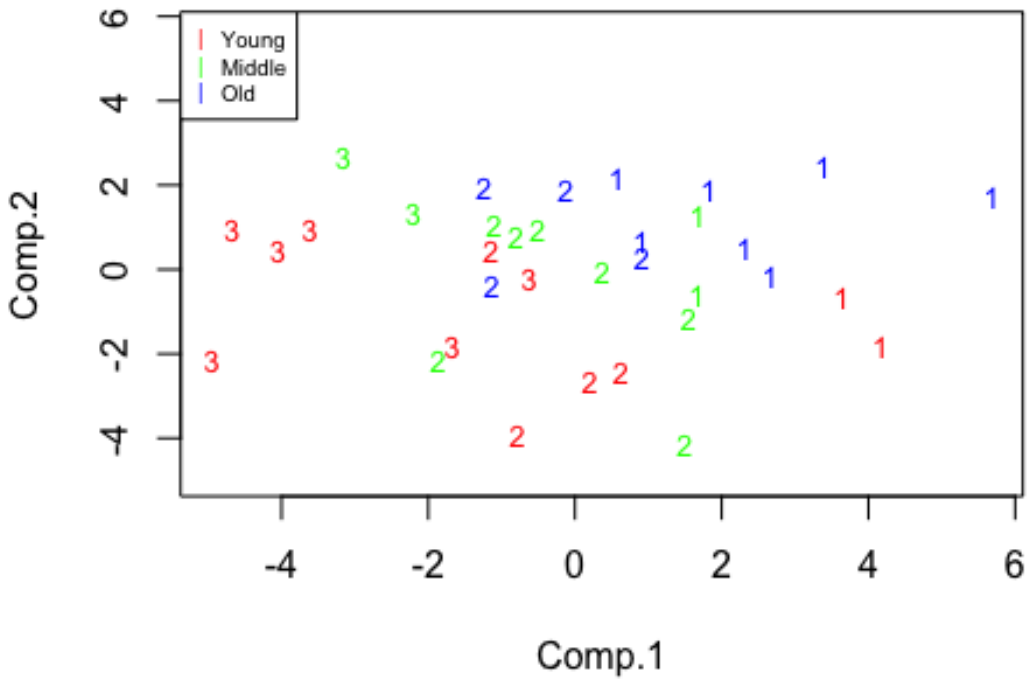

Figure 19

Figure 19 shows that old quarterbacks exclusively belong to tiers 1 and 2, which are the top tier and mid-tier quarterbacks. Young quarterbacks dominate tier 3, signifying that these quarterbacks might need time to develop and improve their skills before they will be able to lead a team to success. There are also some young quarterbacks in the middle and upper tiers (Patrick Mahomes and Jared Goff are the two young quarterbacks in the upper tier), signifying that these quarterbacks are up-and-coming and will be successful in the league for years to come. In addition, we see a few middle-aged quarterbacks in tier 3, which signifies that these quarterbacks might soon be out of the league unless their play drastically improves.

The final technique I use to analyze the data is multi-dimensional scaling, used to get a clear visualization of the proximity of quarterbacks in two dimensional space (*Figure 20)*.

The one weakness of this method is that the choice of distance equation is not clear. In this case, I simply use Euclidian distances.

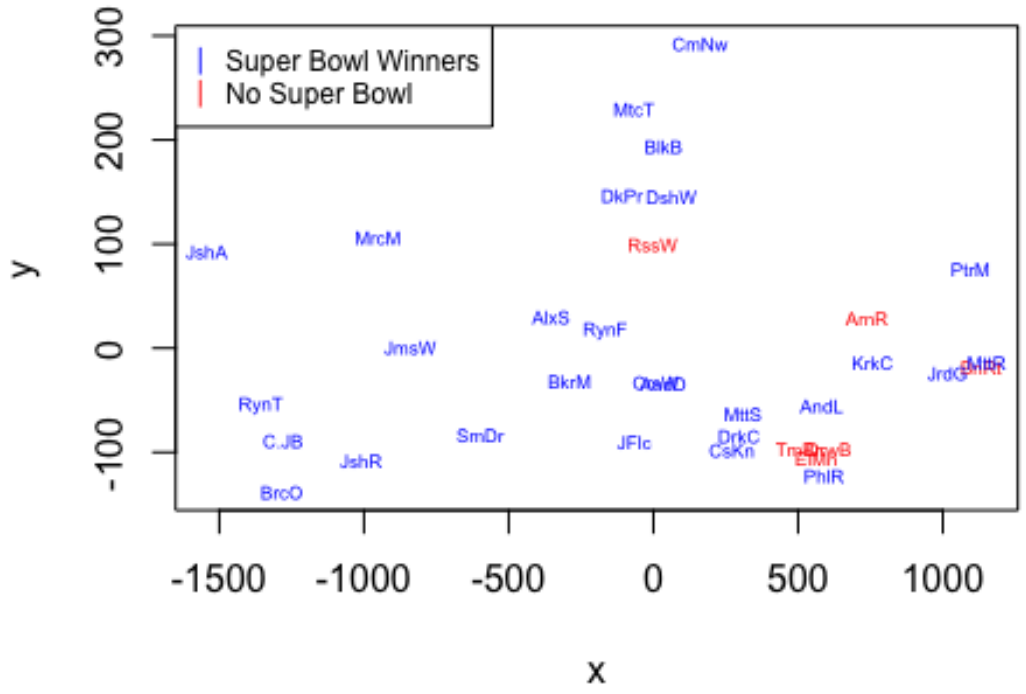

Figure 2015

I have highlighted the Super Bowl winners to see which quarterbacks are playing like the former super bowl winners, possibly an indication of future Super Bowl winners. As shown by *Figure 20,* quarterbacks like Phillip Rivers and Andrew Luck (lower right), are very close in proximity to a group of three former super bowl winners. Perhaps this indicates that these QBs are primed to win a super bowl at some point in the near future.

Conclusions

In conclusion, I was able to directly answer most of the goals originally stated throughout this report. I was able to see that a quarterback's age strongly relates to the style of that quarterback, their passing abilities, rushing abelites, as well as the tier of that quarterback.

I was also able to use principal component analysis to describe quarterback play with fewer statistics, although it still did take quite a few principal components in order to reach a substantial amount of the variance in the data. In particular I saw that both passing stats

and rushing stats could be summarized primarily using one statistic for each, which greatly simplifies the data.

In addition, I analyzed whether Total QBR and QB rating are accurate measures of a quarterback's performance by using principal component analysis. By accounting for all variables, I saw a very strong correlation between principal component scores and both of these statistics, indicating that they do a good job measuring quarterback performance.

Another aspect of my analysis was determining that there are in-fact tiers of quarterbacks within the NFL, as seen through cluster analysis. The clusters had interesting patterns, such as Super Bowl winning quarterbacks primarily coming from the top tier, while no team with a winning record has a quarterback from the lowest tier. This makes a strong argument for why the quarterback is commonly seen as the most important position in sports. In addition, I saw that these Super Bowl winning quarterbacks mostly have similar playing styles, as they are the best at throwing the ball and do not focus on running the ball nearly as much as other quarterbacks. Teams that want to win a Super Bowl should focus on getting quarterbacks who can stand in the pocket and throw the ball well, as these are the types of quarterbacks that not only win games, but win championships.