Final Writeup

Pipe It Up!: Nagaprasad Rudrapatna, Karen Deng, Jackson Muraika, Anna Zolotor

2020-04-30

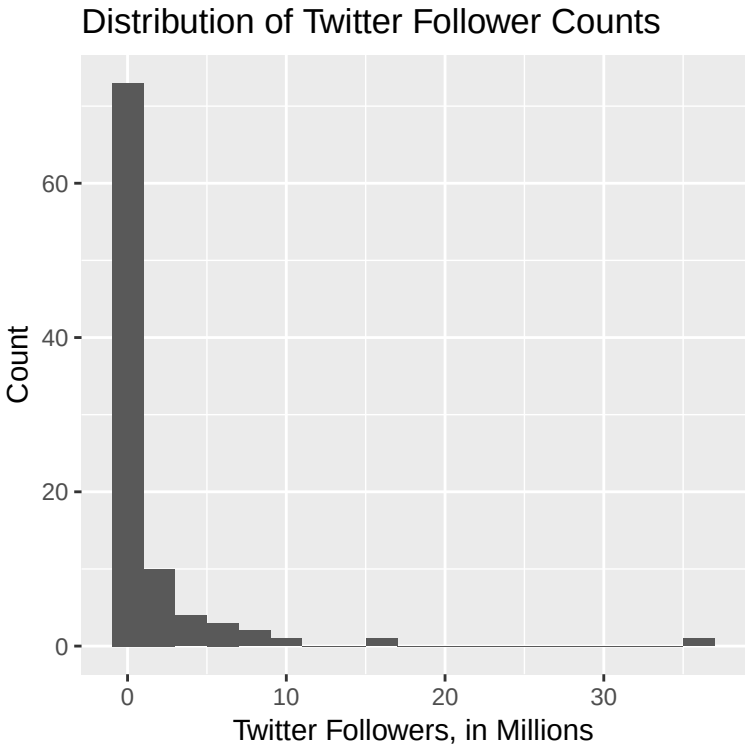
Section 1: Introduction

In this project, Pipe it Up! used the Kaggle dataset "Social Power NBA" to explore the following research question: Is there a relationship between measures of basketball success (such as win percentage and offensive/defensive ratings) and Internet popularity, measured in number of Twitter followers? The motivation for this analysis stems from the increasing relevance of social media in today's hyperconnected world. Social media provides a platform for people to share stories and opinions, and influential people, especially athletes, have considerable sway over large segments of the population. Additionally, since the NBA is a star-driven league due to its small team size and worldwide recognition, social media presence is significant for basketball players. The original creator of our Kaggle dataset, UC Davis and Northwestern Professor Noah Gift, conducted analyses to determine which factors relating to basketball success social media and internet popularity could accurately predict. He searched for correlation between those factors and a number of response variables, including arena attendance, endorsements, salary, and NBA performance. Our team was fascinated by the statistician's conclusions, but we wanted to take the reverse approach and determine which factors (relating to basketball success) can predict the social media popularity that he claims is so important.

Our dataset includes on-court performance data for NBA players in the 2016-2017 season, along with their salary and Twitter engagement. Because we are examining the relationship between player statistics and the number of Twitter followers, we decided to only consider players who had a Twitter accounts. After filtering out these players, we had 95 observations (we later removed one more observation because it influenced the regression line, leaving us with 94). The number of Twitter followers was our response variable, and we chose to include 15 explanatory variables in the actual analysis (excluding Twitter handles which were used to prepare the data): player name, player age, team name abbreviation, win percentage, offensive rating, defensive rating, field goals made, field goals attempted, assists-to-turnovers ratio, rebound percentage, usage percentage, player impact factor (a measure of the impact the player has on games), salary (in millions of dollars), whether they were active on Twitter in the 2015-2016 season, and points scored per game.

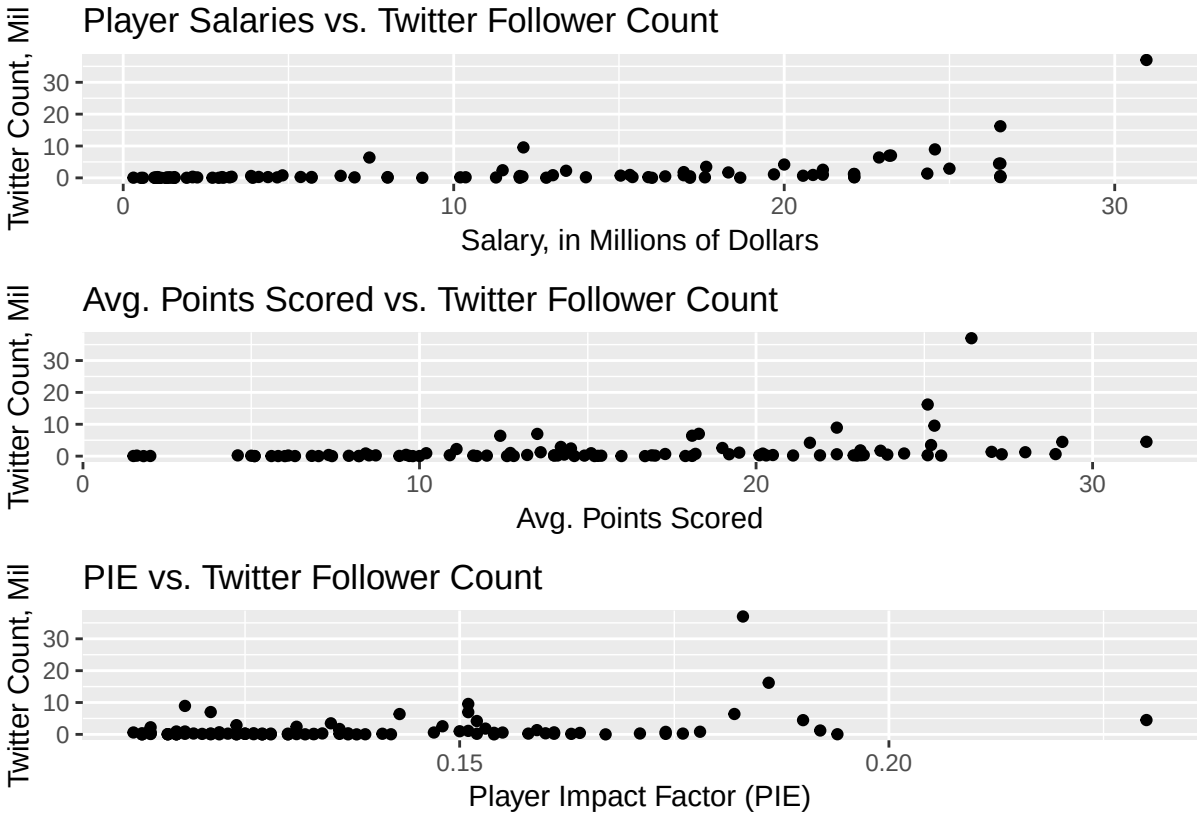
A few interesting/important points from our exploratory data analysis will be shown here, but the complete EDA can be found in additional analysis.

The distribution of Twitter follower count is shown here:



From the histogram, we can see that the distribution of Twitter follower counts is extremely right-skewed. The number of Twitter followers ranges from 2,000 to 37 million, with a mean of 1.6 million and a median of 246,000. There are two obvious outliers: Kevin Durant, with 16.2 million followers, and LeBron James, with 37 million followers.

Now, we'll include several of the bivariate plots that showed particularly interesting or stronger correlations between the explanatory variable and the response. Below, we can see scatterplots displaying the relationships between player salaries and Twitter follower count, average points scored in a game and Twitter follower count, and player impact factor and Twitter follower count.



These were three of the bivariate relationships that appeared to be particularly telling about what the model would look like; all three show weak positive correlations between the explanatory and response variables. None of the bivariate plots showed a strong or very linear relationship between the explanatory variable and Twitter follower count.

Section 2: Regression Analysis

Modeling Approach (all details included in Additional Analysis)

As our response, the number of Twitter followers (in millions), is a continuous numerical variable, we used a multiple linear regression model.

In regard to model selection, we began by fitting a multiple linear regression model with thirteen main effects (mean-centered age, mean-centered assists-to-turnovers ratio, mean-centered player offensive rating, mean-centered player defensive rating, mean-centered player impact factor (PIE), mean-centered rebound percentage, mean-centered usage percentage, mean-centered salary, mean-centered win percentage, mean-centered points scored, mean-centered field goals made, mean-centered field goals attempted, and whether the player has an active Twitter account in 2015-2016). We also considered interactions between mean-centered salary and mean-centered win percentage and mean-centered player impact factor (PIE) and mean-centered points scored because the multivariate EDA highlighted strong positive relationships between win percentage and player salary and points scored and PIE.

Next, we performed two iterations of backward selection on this initial model: (i) using AIC as the selection criterion and (ii) using adjusted R-squared as the selection criterion. We decided against trying BIC as the selection criterion because we would prefer more terms in the final model as our objective is to predict the Twitter follower counts of NBA players using measures of athletic success (and predictions are generally more accurate with more relevant predictor variables).

After completing the two iterations of backward selection, we compared the resulting models to see whether

certain terms were removed in both (which would suggest those terms are not statistically significant). We also reconciled the differences between the terms included in the selected model based on one of the selection criterion but not the other. Particularly, mean-centered win percentage was included in the model selected based on AIC but not in the model selected based on adjusted R-squared. We decided to keep mean-centered win percentage in the model so that the statistically significant interaction between mean-centered salary and mean-centered win percentage could remain in the model as well.

Additionally, we closely examined the p-values and confidence intervals for each of the remaining predictors in the model after reconciling the differences between the models produced from the two iterations of backward selection. We noticed mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in the previous season had high p-values and confidence intervals including zero, suggesting they were statistically insignificant. We compared the AIC and adjusted R-squared values for the model selected based on AIC as the selection criterion (first iteration of backward selection) and the same model without mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016. The results were: the model with all terms maximized adjusted R-squared and minimized AIC. Since adjusted R-squared penalizes for unnecessary predictors, the fact that the model with all terms had a higher adjusted R-squared value means that, despite the high p-values and the presence of zero in the confidence intervals associated with mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016, these predictors are valuable in predicting the response, the number of Twitter followers (in millions). Hence, we decided to keep mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016 in the model.

Finally, we chose to analyze the impact of prominent athletes at the end of the model selection phase because prominent athletes are also included in the population we want to understand when fitting the multiple linear regression model (it is important to explore this topic since our objective is to design a model with the best predictive accuracy). We determined whether prominent athletes were influential points by looking at standardized residuals, leverage, and Cook's Distance. We identified LeBron James as an influential point in the data and decided to remove him to avoid overestimating the number of Twitter followers (in millions) for less prominent athletes.

However, as a result of this decision, the model coefficients changed considerably – mean-centered field goals made and mean-centered field goals attempted became statistically insignificant. So, we conducted another iteration of backward selection using AIC as the selection criterion and eliminated mean-centered assists-to-turnovers ratio, mean-centered usage percentage, mean-centered field goals made, and mean-centered field goals attempted from the model. But, mean-centered age and whether the player had an active Twitter account in 2015-2016 – insignificant predictors (based on p-values and confidence intervals) – remained in the model. To determine whether these predictors should be removed, we compared the AIC and adjusted R-squared values for the final model selected based on AIC as the selection criterion (without LeBron) and the same model without mean-centered age and whether the player had an active Twitter account in 2015-2016.

Based on the AIC and adjusted R-squared values, the model with all five terms minimized AIC and maximized adjusted R-squared. Therefore, we chose to move forward with the model which includes mean-centered salary, mean-centered win percentage, mean-centered age, whether the player had an active Twitter account in 2015-2016, and the interaction between mean-centered salary and mean-centered win percentage.

Model and Relevant Fit Statistics (Prior to Assumptions)

Thus, our model is:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.109	0.226	4.904	0.000	0.660	1.559
ageCent	0.109	0.059	1.834	0.070	-0.009	0.227
salary_millionsCent	0.097	0.027	3.552	0.001	0.043	0.152
w_pctCent	4.533	1.550	2.924	0.004	1.452	7.614
active_twitter_lyear0	-2.379	1.508	-1.578	0.118	-5.376	0.618

term	estimate	std.error	statistic	p.value	conf.low	conf.high
salary_millionsCent:w_pctCent	0.513	0.171	2.995	0.004	0.173	0.853

To get a better sense of the model fit, we will calculate the R-squared and adjusted R-squared values:

```
## [1] 0.3545028
```

```
## [1] 0.3178268
```

The proportion of the variation in the number of Twitter followers (in millions) explained by the regression model is roughly 35.5%. Although this might suggest the model fit is relatively poor, it is important to remember we have removed many explanatory variables from the model so that predominantly significant variables remain (and R-squared increases as more explanatory variables are included). Since the adjusted R-squared value is close to the R-squared value, we conclude the variables in the model are significant in understanding the variation in the number of Twitter followers (in millions).

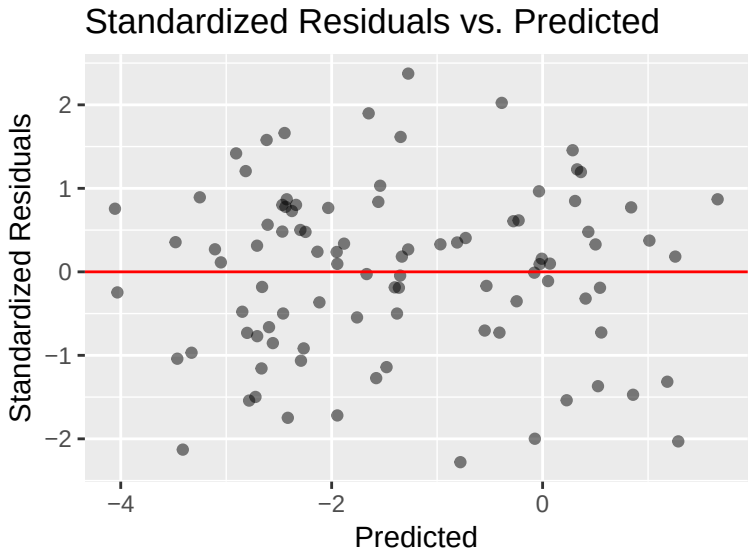
Discussion of Assumptions

Next, we checked the linearity, constant variance, normality, and independence assumptions for multiple linear regression.

When we checked the standardized residuals vs. predicted values, we found that our model violated constant variance because the height of the cloud of points varied as you move from left to right. Points were clustered at the very left, but they were sparser as you moved along the graph. Therefore, as constant variance was not satisfied, we made some adjustments to our model. We decided to log-transform the response variable because the plot of standardized residuals vs. predicted values violated constant variance.

After log-transforming the response (the number of Twitter followers, in millions), we discovered the interaction between mean-centered salary and mean-centered win percentage and mean-centered age had very large p-values (> 0.05), so we removed them from our model. We will see if the log transformation on our response variable and the removal of insignificant predictors are enough to support our assumptions:

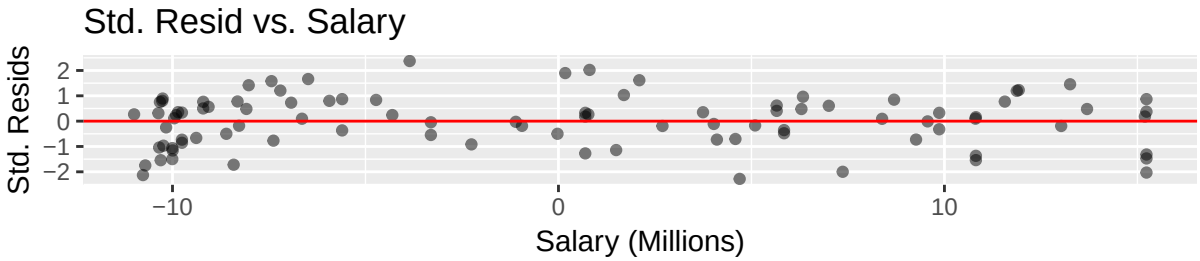
First, we checked for linearity (i.e. whether the response variable had a linear relationship with the predictor variables in the model) by plotting the standardized residuals vs. predicted values:

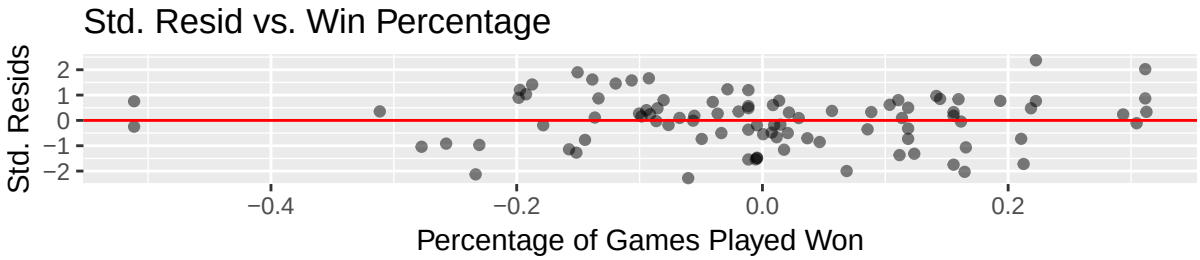


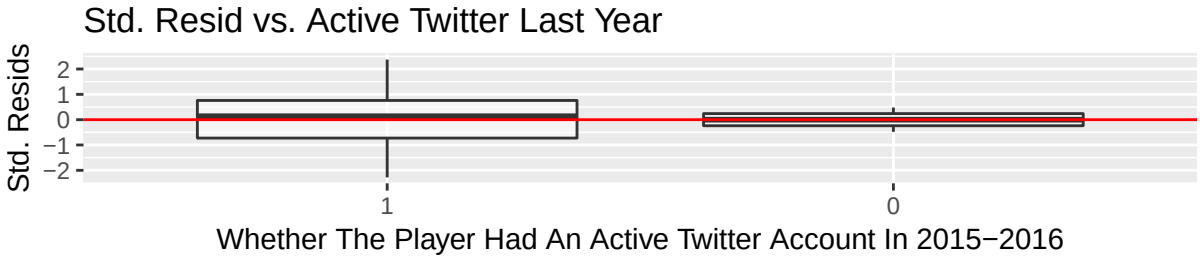
When observing for constant variance, the height of the cloud of points seems to be constant as you move from left to right. Therefore, constant variance is satisfied.

There is no obvious pattern in the plot of standardized residuals vs. predicted values. Hence, this plot presents no issues with the linearity assumption.

Next, we individually assess the plots of standardized residuals vs. predictors:

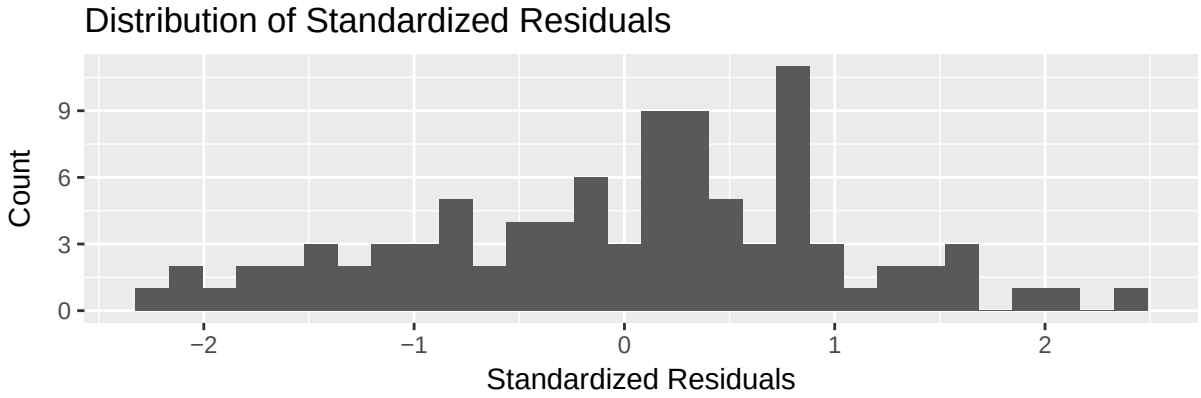


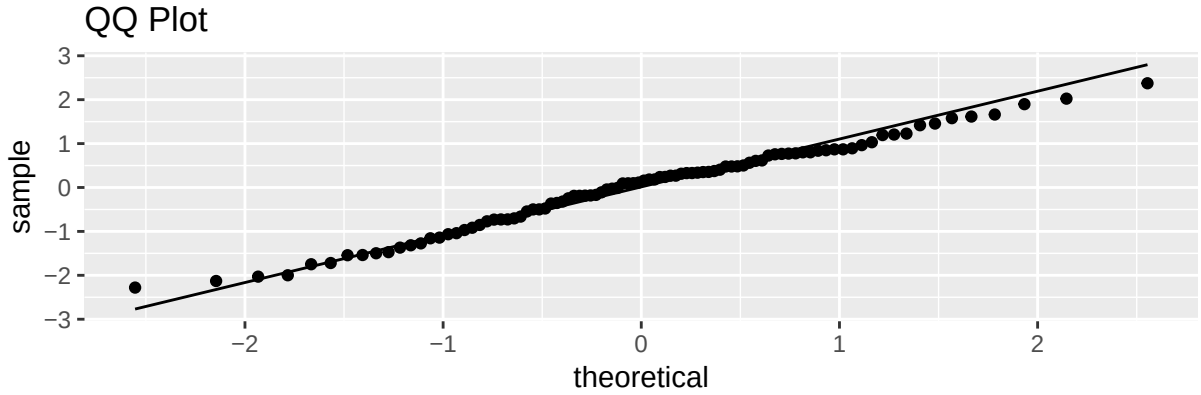




There is no distinguishable pattern in any of the plots because there are no discernible curves. The boxplot also shows little difference in the median standardized residuals between players with active Twitter accounts in 2015-2016 and those without active Twitter accounts in the previous season (both roughly 0). Therefore, linearity seems to be satisfied.

Next, we will check for the normality assumption by creating a histogram of the standardized residuals and a normal QQ-plot of the residuals:





Normality might be concerning because the histogram of standardized residuals is not perfectly normal and is slightly right-skewed. Points in the QQ plot also don't fall exactly along the diagonal line. Nevertheless, most inference methods for regression are robust to some departures from normality, so we will continue with our analysis and assume normality is satisfied.

Independence is also satisfied because data was not taken over time, so we know there is no temporal correlation. There is also no evidence of spatial correlation between the observations and no purposeful order to how the dataset was collected according to Kaggle, so there is no structure/order to the dataset according to observation number. The number of Twitter followers of one player will not affect the number of Twitter followers of another player (there is no rule saying if a person, for example, follows Derrick Rose that he/she cannot also follow James Harden). Therefore, independence is satisfied.

In sum, linearity, constant variance, normality, and independence are all reasonably satisfied in this model.

Lastly, we will also check for multicollinearity in our model. If two or more predictor variables are highly correlated in our model, our regression may change eratically in response to small changes in our data. We will check the variance inflation factor (VIF) for every predictor variable to check for concerns with multicollinearity:

```
## # A tibble: 3 x 2
##   names                     x
##   <chr>                 <dbl>
## 1 salary_millionsCent    1.09
## 2 w_pctCent              1.09
## 3 active_twitter_lyear0  1.02
```

None of the variables have VIFs greater than or equal to 10, so there are no issues with multicollinearity.

Final Model and Relevant Fit Statistics

So, our revised final model is:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.292	0.140	-9.212	0.000	-1.571	-1.013
salary_millionsCent	0.142	0.017	8.514	0.000	0.109	0.175
w_pctCent	2.538	0.901	2.816	0.006	0.747	4.328
active_twitter_lyear0	-2.404	0.969	-2.480	0.015	-4.329	-0.478

To get a better sense of the model fit, we will calculate the R-squared and adjusted R-squared values for this revised model:

```
## [1] 0.53055
```

```
## [1] 0.5149017
```

The proportion of the variation in the number of Twitter followers (in millions) explained by the revised regression model is roughly 53.1%. Comparing this to the R-squared value from the final model prior to assumptions (approximately 0.355), we conclude our model has significantly improved (especially considering R-squared increases with more explanatory variables and this revised model actually has fewer predictors)! Since the adjusted R-squared value (approximately 0.515) is close to the R-squared value, we conclude the variables in the model are significant in understanding the variation in the number of Twitter followers (in millions). This corroborates the interpretation of the p-values: the p-values for each of the predictor variables are less than 0.05, so we conclude these variables are indeed significant predictors of our response.

Section 3: Discussion

Interpretations

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.292	0.140	-9.212	0.000	-1.571	-1.013
salary_millionsCent	0.142	0.017	8.514	0.000	0.109	0.175
w_pctCent	2.538	0.901	2.816	0.006	0.747	4.328
active_twitter_lyear0	-2.404	0.969	-2.480	0.015	-4.329	-0.478

```
## # A tibble: 1 x 3
##   mean_salary mean_winpct mean_followers
##         <dbl>       <dbl>          <dbl>
## 1        11.1       0.510           1.22
```

The intercept of the model is -1.292, so a player with an average salary of 11.1 million dollars, an average win percentage of 0.510, and a Twitter account during the 2015-16 season is expected (median) to have exp(-1.292) million, or approximately 274,721 Twitter followers.

The slope estimate associated with mean-centered player salary is 0.142, which means that for every 1 million increase in player salary, the median number of Twitter followers is expected to multiply by a factor of exp(0.142) = 1.153, holding all else constant.

The slope estimate associated with mean-centered win percentage is 2.538; however, due to the way win percentage is calculated, this estimate is misleading. Win percentage is computed as if it was win proportion, which means that for every 10% increase in win percentage, a player's median number of Twitter followers is expected to multiply by a factor of exp(0.2538) = 1.289, holding all else constant.

The slope estimate associated with the variable that indicates whether players were active on Twitter in the 2015-2016 is -2.404, so if an NBA player had a Twitter account during the 2015-16 season, his median number of Twitter followers is expected to multiply by a factor of exp(-2.404) = 0.786, holding all else constant. Of course, it is important to note that the vast majority of players in the dataset were indeed active on Twitter in 2015-2016.

Our objective in creating the model was predicting Twitter follower counts, so we did not prioritize ensuring that the interpretations of coefficients made sense.

Predictions

We carefully selected six NBA players from the 2016-2017 season who were not included in the dataset and tested the final model's predictive accuracy by comparing predicted Twitter follower counts to actual Twitter data from 2017.

Salary data was taken from ESPN (www.espn.com), and win percentage data was calculated based on win-loss ratios (stats.nba.com). The number of Twitter followers was collected from various articles written in 2017. All players were active on Twitter during the previous season.

Derrick Rose:

```
##          fit       lwr      upr
## 1 -0.1073978 -2.826623 2.611827
```

During the 2016-2017 season, Derrick Rose made 21.3 million dollars and had a winning percentage of 40.6%. Derrick Rose is expected to have exp(-0.1074) = 898,817 Twitter followers. He actually had 2.49 million Twitter followers, so this is an extreme underprediction. Despite being on a poor team during this season, Rose was previously a member of very competitive teams and actually won the NBA MVP award, which explains his popularity.

Wesley Matthews:

```
##          fit       lwr      upr
## 1 -0.7259807 -3.429891 1.977929
```

Wesley Matthews made 17.1 million dollars and had a winning percentage of 39.7% during the 2016-2017 season. Based on our model, Wesley Matthews is expected to have exp(-0.726) = 483,850 Twitter followers. He actually had 241,000 Twitter followers, so this is actually an overprediction. His salary of 17.1 million dollars may be the reason for this overprediction despite his poor record.

Boris Diaw:

```
##         fit       lwr      upr
## 1 -1.602879 -4.300245 1.094486
```

From his 7 million dollar salary and 61.6% winning percentage, Boris Diaw is expected to have exp(-1.603) = 201,316 Twitter followers. He actually had 462,000 Twitter followers (gross underestimate). Diaw's relatively low salary of 7 million certainly contributed to this underprediction. Also, Diaw was previously an NBA champion and he is of French origin (international appeal), which may further explain the underprediction.

Tony Parker:

```
##          fit      lwr      upr
## 1 -0.3206892 -3.03336 2.391982
```

Tony Parker is expected to have exp(-0.321) = 725,423 followers based on his 14 million dollar salary and 73.0% winning percentage. He actually had 2.12 million Twitter followers, another gross underestimate. Despite having an above average salary, the model did not account for the fact that he is a four-time NBA champion. Beyond that, he is from France and has international appeal that other players may not have.

James Harden:

```
##        fit      lwr      upr
## 1 1.297407 -1.43839 4.033204
```

During the 2016-2017 season, James Harden made 26.54 million dollars and had a winning percentage of 66.7%. James Harden is expected to have exp(1.297) = 3.66 milion followers. He actually had 4.8 million Twitter followers, so this is a slight underprediction. During this season, Harden entered the first year of a four-year contract extension. Based on the NBA salary cap, his contract was backloaded, meaning he made more in each additional contract year. For this reason, his 26.54 million dollar salary in 2016-17 is not indicative of his average salary over the four-year extension (29.5 million), leading to an underprediction.

Serge Ibaka:

```
##         fit       lwr      upr
## 1 -1.226665 -3.913073 1.459743
```

Serge Ibaka is expected to have exp(-1.227) = 293,171 Twitter followers based on his 46.8% winning percentage and salary of 12.3 million dollars. He actually had around 848,000 Twitter followers, so this is another extreme underprediction. Just the previous year, Ibaka's winning percentage was 69.2%, but he was traded after that season. This fluctuation in winning percentage can explain this underprediction, in addition to Ibaka's origin (the Democratic Republic of the Congo), which contributes to his international appeal.

The final model has decent predictive capabilities but also has many flaws. The most likely reason for inaccuracies is that the predictor variables (on-court performance statistics) are only based on one season and there few stars in the NBA who have large amounts of Twitter followers (in millions).

It is important to note the dates at which these Twitter follower counts were recorded vary and do not perfectly coincide with the data collection of this dataset. However, the dataset itself only specifies that the data was collected during the 2016-2017 season and the articles which were used to obtain Twitter follower counts for players outside the dataset were written during this season as well.

Section 4: Limitations

We should have recoded win percentage so that it more accurately reflects how it is computed. For instance, we could have relabeled win percentage as win proportion to better reflect that it is calculated as a win-loss ratio. This way, our interpretations may have been more intuitive; however, as our objective was prediction, we did not think it was necessary to do so in this analysis.

According to our dataset, five players were missing Twitter handles. These values are missing at random because the missingness depends on other observed variables (e.g. the person's social media usage). Hence, the probability that a variable is missing depends on information not included in our dataset. We decided to remove these five players from our dataset because there were very few observations with missing values relative to the sample size (after removing these players, we still had 95 observations). We also determined, since the observations with missingness are random, the resulting analysis will not be biased because the missingness does not differ systematically from the complete observations. In addition, since our objective is predicting the number of Twitter followers for NBA athletes, players without Twitter accounts are outside of the model's predictive capabilities and hence do not belong in the dataset. One other modification we made to our original dataset was removing LeBron James. We decided, based on Cook's distance, standardized residuals, and leverage, LeBron was an influential point with a significant effect on the regression line. Our model was limited by the fact that the predictor variables only came from one season, and that one year of on-court statistics may not be indicative of a player's career statistics. Both salary and win percentage for individuals fluctuate from year to year because of changes in the NBA salary cap and player trades between teams, respectively. The NBA salary cap limits how much each team can pay players over a certain year.

To improve our model, we could use a nested F-test to determine whether interactions between mean-centered salary and mean-centered win percentage, as well as the interactions between mean-centered player impact factor and mean-centered points scored were significant rather than relying on backwards selection with AIC

and adjusted R squared as the selection criteria. One other consideration might be to revisit the full Kaggle dataset and randomly select a different set of explanatory variables, as there were 63 choices.

In order to make significant progress, we would have to add additional observations. One smaller addition could be to add more players who were not active on Twitter last year. If we had the ability to make larger changes, we could make the model stronger by to building it using on-court statistics from more than just one season, ideally three at minimum. Also, the inclusion of more players in the dataset would allow us to use k-fold cross validation to create a new model after splitting the observations into training and test sets. Lastly, we would like to test our model's predictive accuracy on a greater number of randomly selected NBA players.

Section 5: Conclusion

In conclusion, our model exhibited the ability to make decent, albeit relatively conservative predictions of Twitter follower count based on mean-centered salary, mean-centered win percentage, and whether the player had an active Twitter account in the previous season. Its R-squared value was 0.531, so the proportion of variation in Twitter followers explained by the model was 53.1%. Because this R-squared value is similar to the adjusted R-squared value of 0.515, and our predictors were significant, we can conclude that on-court performance statistics and salary have an impact on the number of Twitter followers for NBA players.

From our test cases (six players outside the dataset), we noticed that the certain factors of player popularity were not accounted for in the model – international appeal, prior success (i.e. previously winning an NBA championship), player trades, contract extensions, and playing for large-market teams as opposed to small-market teams. Thus, in future analyses, we could try to incorporate some of these variables to improve the predictive accuracy of our model.

Overall, we noticed an interesting trend in our model's predictive capabilities. The model tended to overestimate the number of Twitter followers of higher-paid players who played on more competitive teams (based on collective win percentage). Conversely, the model tended to underestimate the number of Twitter followers of higher-paid players who played on less competitive teams. This trend is logical because higher-paid players on more competitive teams (e.g. James Harden) are also some of NBA's best players. Furthermore, we believe win percentage is one of the most factors in accurately predicting the number of Twitter followers for NBA athletes.

To stimulate interesting discussion, we could attempt to use social media power/social media popularity to predict on-court statistics, mirroring Professor Noah Gift's initial analysis. We could also use the data to try to predict other measures of social media popularity, like Instagram followers or Facebook page likes.

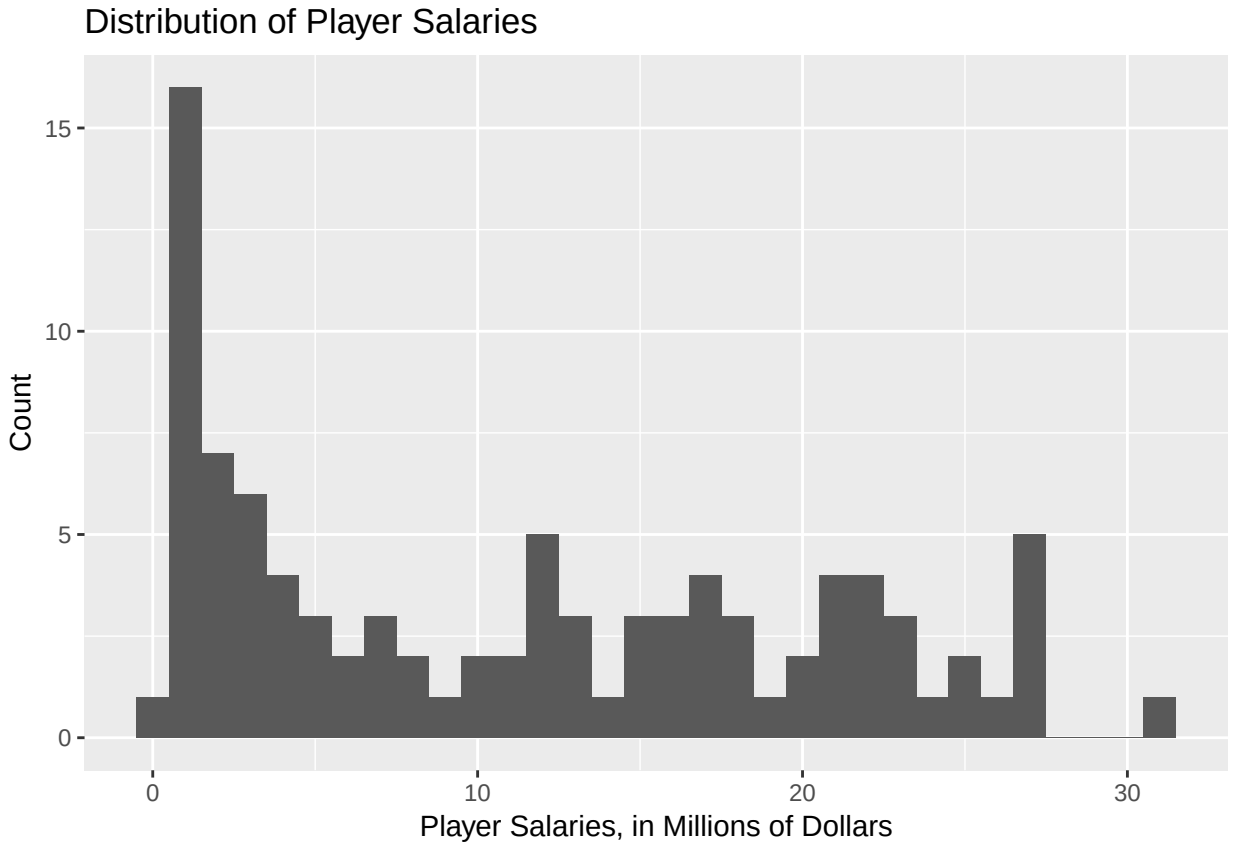
Section 6: Additional Work

Complete EDA

Univariate

First, we will do univariate EDA on the dataset. Player name will be used to refer to observations in our dataset, but since each player name is distinct we do not need to do EDA on this variable.

Here, we'll look into players' salaries, in millions of dollars:

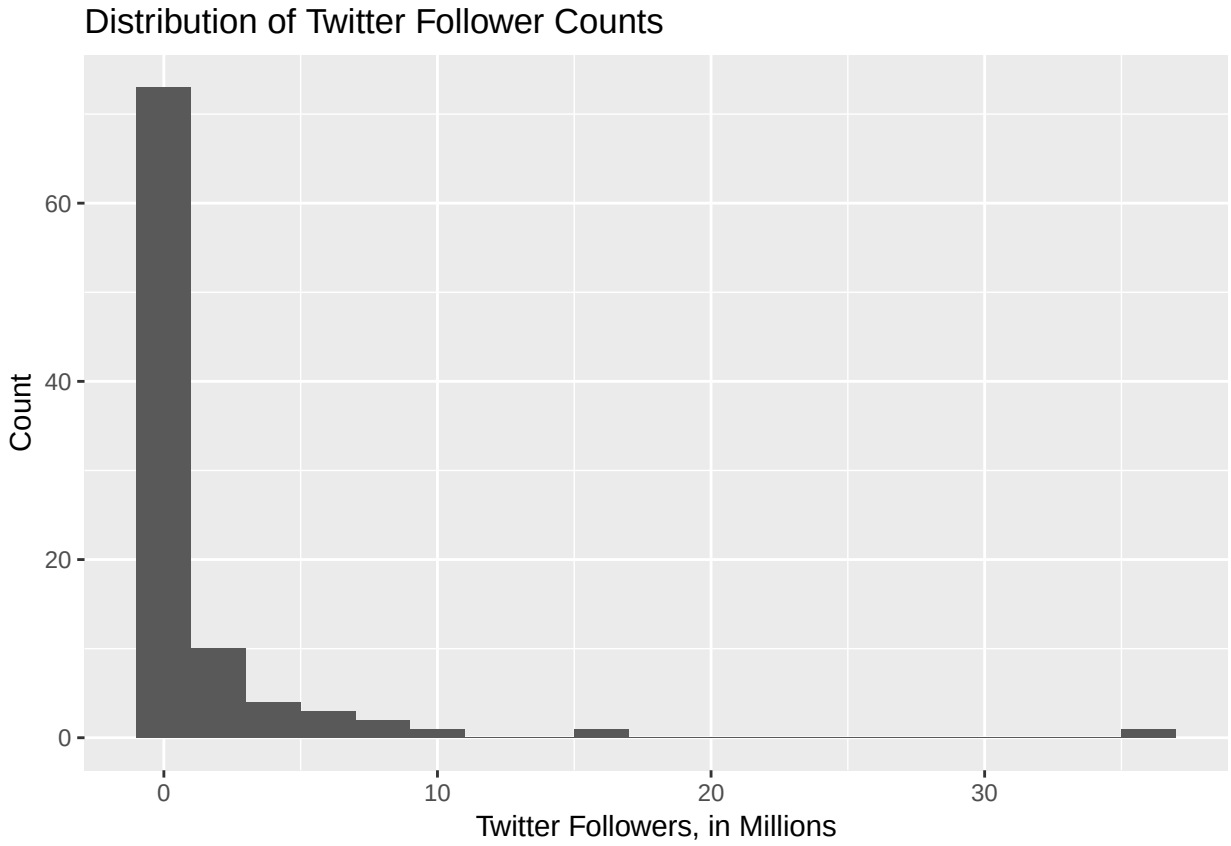



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  11.3  0.31  2.47   11.3  18.5  31.0
```

```
## # A tibble: 10 x 2
##    PLAYER_NAME       SALARY_MILLIONS
##    <chr>                       <dbl>
##  1 LeBron James                 31.0
##  2 Russell Westbrook            26.5
##  3 Kevin Durant                 26.5
##  4 Mike Conley                  26.5
##  5 DeMar DeRozan                26.5
##  6 Al Horford                   26.5
##  7 James Harden                 26.5
##  8 Dirk Nowitzki                25
##  9 Carmelo Anthony              24.6
## 10 Damian Lillard               24.3
```

As we can see from the histogram, the distribution of salaries is somewhat right-skewed, with most of the players making less than 20 million dollars a year. The mean salary is 11.3 million dollars a year. The player who earns the most, at 30.96 million dollars per year, is LeBron James (LeBron could be an influential point, so we will revisit this after the model selection phase). On the other hand, Russell Westbrook, Kevin Durant, Mike Conley, DeMar DeRozan, and Al Horford each earn 26.54 million dollars per year.

Here, we'll look into the response variable, the number of Twitter followers (in millions):




```
## # A tibble: 1 x 6
##    mean   min     Q1 median    Q3   max
##   <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1  1.60 0.002 0.0595  0.246 0.912    37
```

```
## # A tibble: 2 x 29
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING AST_RATIO
##   <chr>       <chr>            <dbl> <dbl>      <dbl>      <dbl>     <dbl>
## 1 LeBron Jam~ CLE                 32 0.689       115.       107.      25.6
## 2 Kevin Dura~ GSW                 28 0.823       117.       101.      18.4
## # ... with 22 more variables: REB_PCT <dbl>, USG_PCT <dbl>, PIE <dbl>,
## #   SALARY_MILLIONS <dbl>, ACTIVE_TWITTER_LAST_YEAR <dbl>,
## #   TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>, PTS <dbl>, FGM <dbl>, FGA <dbl>,
## #   ageCent <dbl>, ast_ratioCent <dbl>, off_ratingCent <dbl>,
## #   def_ratingCent <dbl>, fgaCent <dbl>, fgmCent <dbl>, PIECent <dbl>,
## #   reb_pctCent <dbl>, usg_pctCent <dbl>, salary_millionsCent <dbl>,
## #   w_pctCent <dbl>, ptsCent <dbl>, active_twitter_lyear <fct>
```

From the histogram, we can see that the distribution of Twitter follower counts is extremely right-skewed. The number of Twitter followers ranges from .002 million to 37 million, with a mean of 1.6 million and a median of.246 million. There are two obvious outliers: Kevin Durant, with 16.2 million followers, and LeBron James, with 37 million followers.

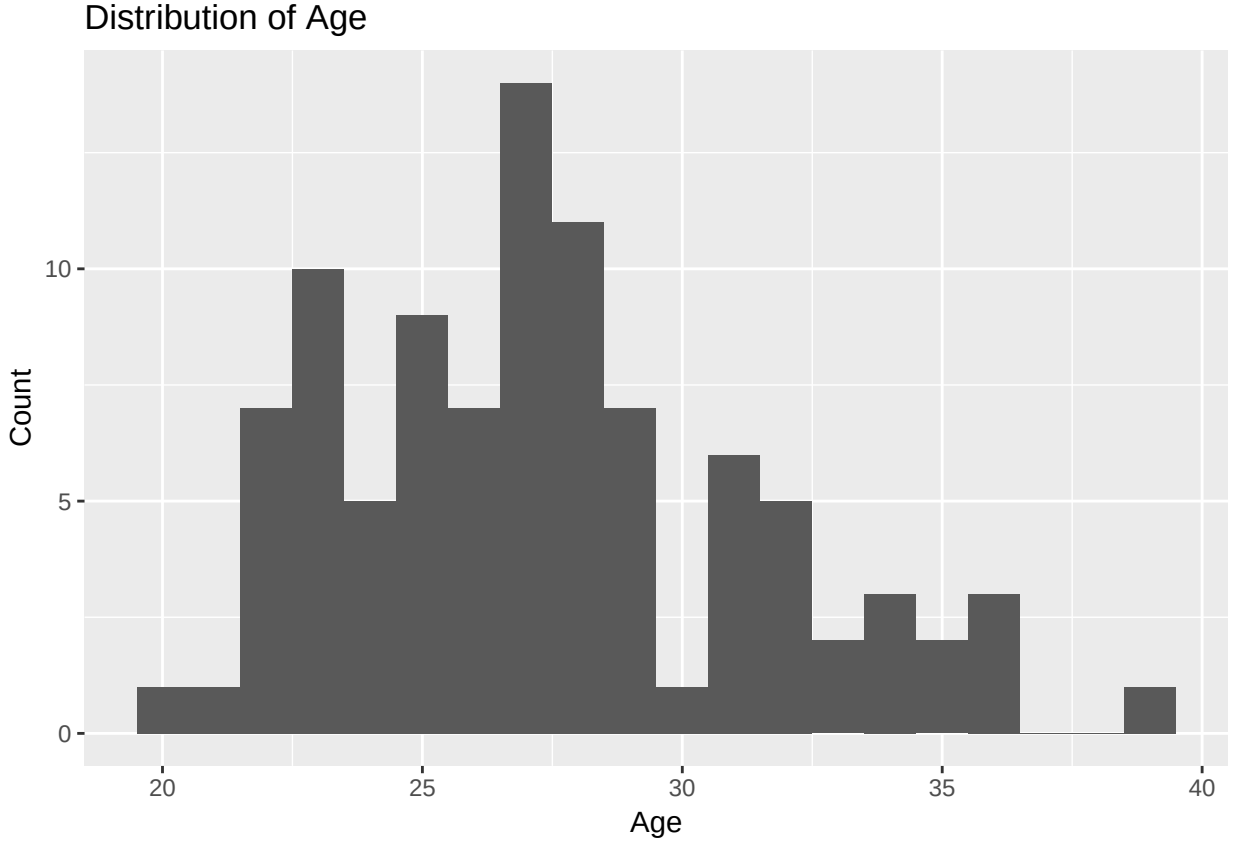
Here, we'll take a look at how many players there are from each team in the dataset:

```
## # A tibble: 30 x 2
##    TEAM_ABBREVIATION     n
##    <chr>             <int>
##  1 SAC                   1
```

```
##  2 CHI                   2
##  3 IND                   2
##  4 LAL                   2
##  5 MIA                   2
##  6 MIN                   2
##  7 ORL                   2
##  8 WAS                   2
##  9 ATL                   3
## 10 BKN                   3
## # ... with 20 more rows
```

As we can see from the output, there is only one team that is represented just once in the dataset: SAC, the Sacramento Kings. The greatest number of times teams are represented in the dataset is 5. GSW (Golden State Warriors), LAC (Los Angeles Clippers), and SAS (San Antonio Spurs) are all represented 5 times.

Now, we'll explore the distribution of player age in the dataset:



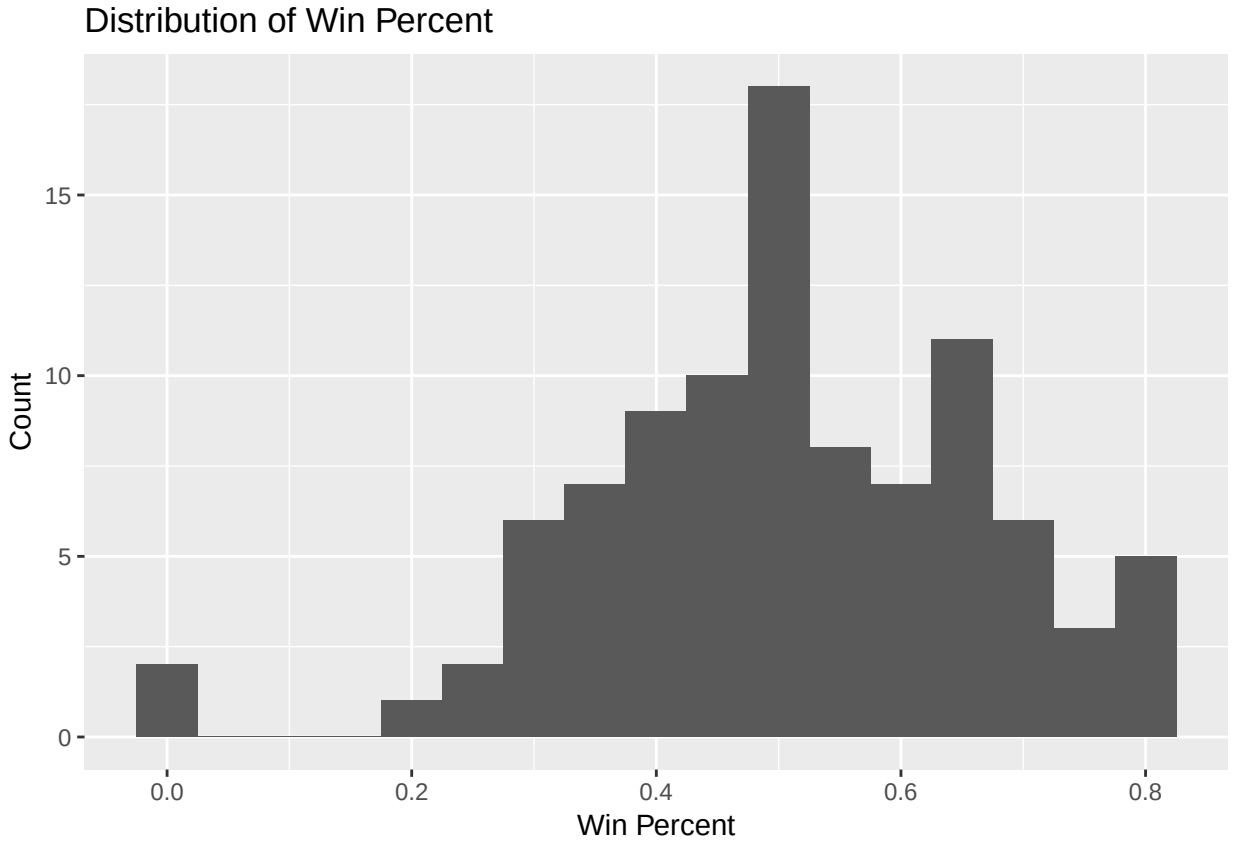

```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  27.4    20  24.5     27    29    39
```

```
## # A tibble: 1 x 29
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING AST_RATIO
##   <chr>       <chr>            <dbl> <dbl>      <dbl>      <dbl>     <dbl>
## 1 Dirk Nowit~ DAL                 39 0.426       105.       106.       9.5
## # ... with 22 more variables: REB_PCT <dbl>, USG_PCT <dbl>, PIE <dbl>,
## #   SALARY_MILLIONS <dbl>, ACTIVE_TWITTER_LAST_YEAR <dbl>,
## #   TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>, PTS <dbl>, FGM <dbl>, FGA <dbl>,
```

```
## #   ageCent <dbl>, ast_ratioCent <dbl>, off_ratingCent <dbl>,
## #   def_ratingCent <dbl>, fgaCent <dbl>, fgmCent <dbl>, PIECent <dbl>,
## #   reb_pctCent <dbl>, usg_pctCent <dbl>, salary_millionsCent <dbl>,
## #   w_pctCent <dbl>, ptsCent <dbl>, active_twitter_lyear <fct>
```

As we can see from the histogram, age is somewhat normally distributed in the dataset, with a mode around 27 and a surprisingly low number of 30-year olds. The mean age, 27.39, and median age, 27, are very close together, indicating little skew. The lowest age is 20 and the highest is 39. The oldest player by far, at 39, is Dirk Nowitzki.

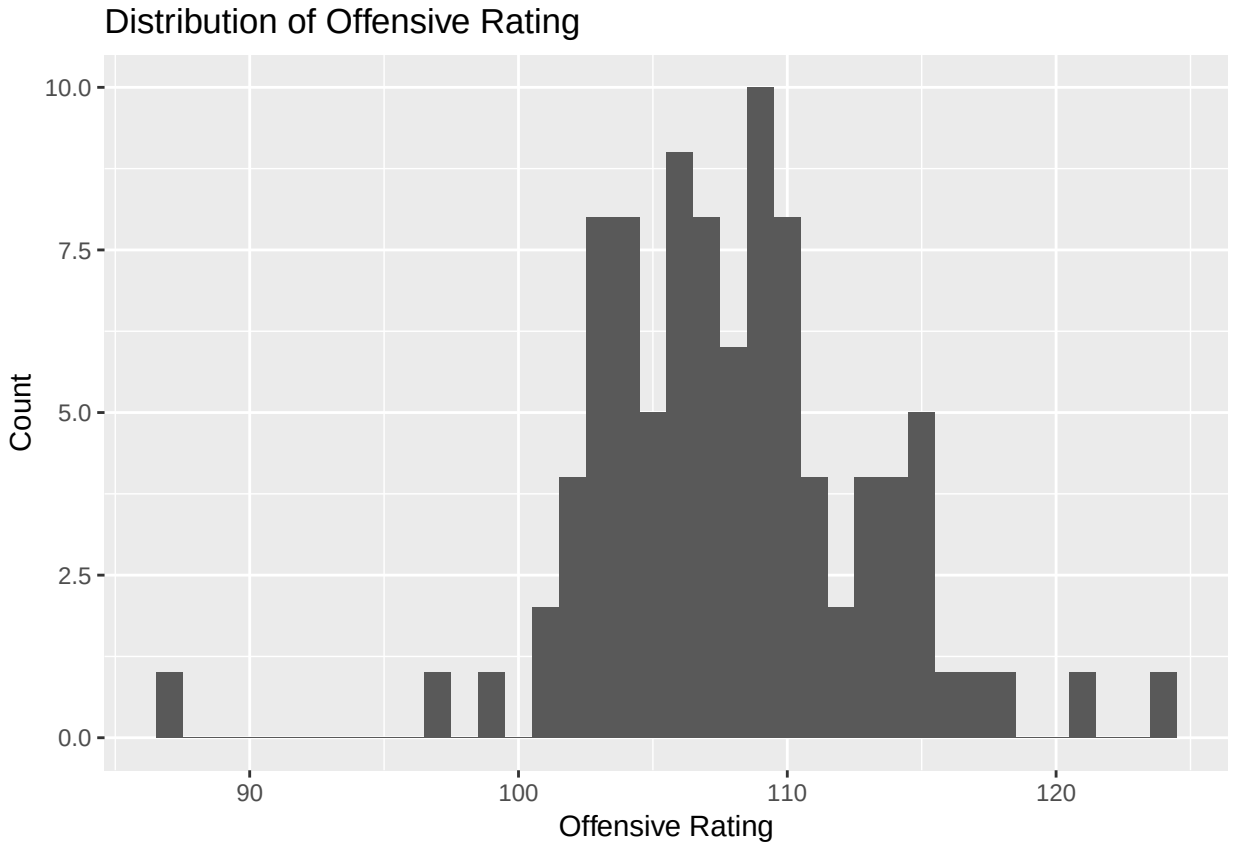
Now, we'll examine the distribution of win percentage:



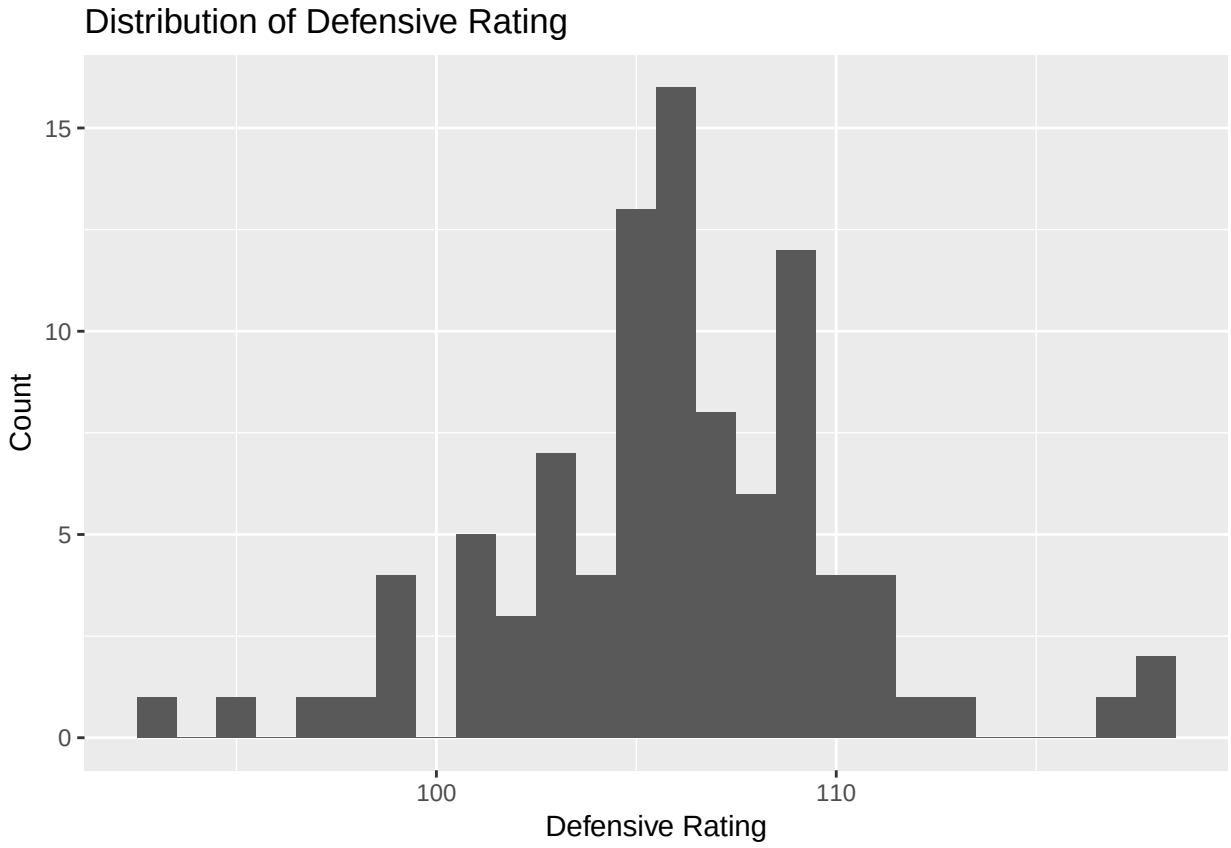

```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.511     0 0.418  0.507  0.63 0.824
```

As we can see from the histogram, win percent is also somewhat normally distributed, with a mode around 50 percent. The minimum win percent in the dataset is 0, while the maximum is 82.4. The median of 50.7 is very similar to the mean of 51%. The fact that the mean and median win percents in the dataset fall so close to 50% indicate good randomness in the dataset, b/c the mean and median win percents for all nba players are 50%.

Next, we'll look at the distributions for offensive rating and defensive rating:



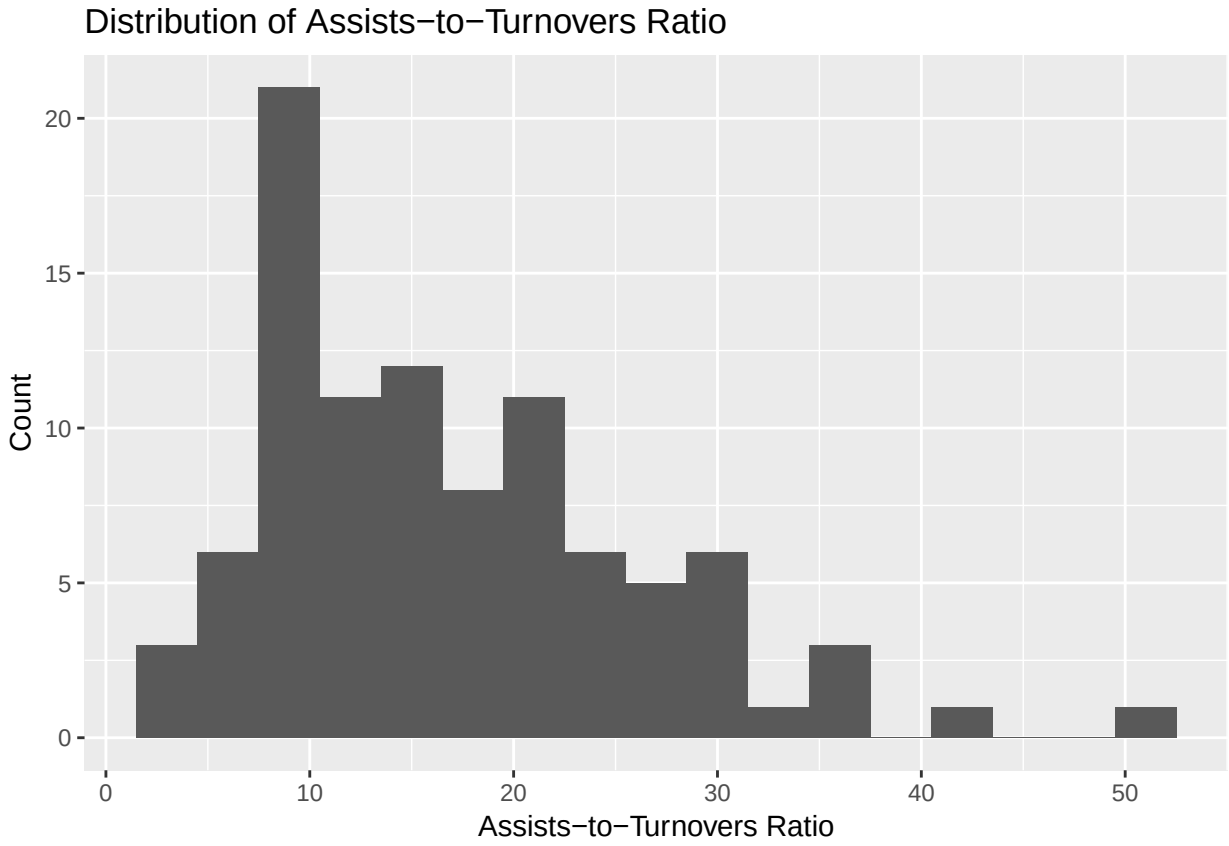

```
## # A tibble: 1 x 3
##     min median   max
##   <dbl>  <dbl> <dbl>
## 1  86.8   108.  124.
```




```
## # A tibble: 1 x 3
##     min median   max
##   <dbl>  <dbl> <dbl>
## 1    93    106  118.
```

Defensive rating and offensive rating do not stray far from being normally distributed. Offensive rating varies from 86.8 to 124.2, with a median of 107.6. Defensive rating varies from 93 to 118.3, with a median of 106. Thus, the dataset contains a larger range in terms of offensive rating, and the median is also slightly higher for defensive rated players.

Next, we'll look at the distribution of the assists-to-turnovers ratio:

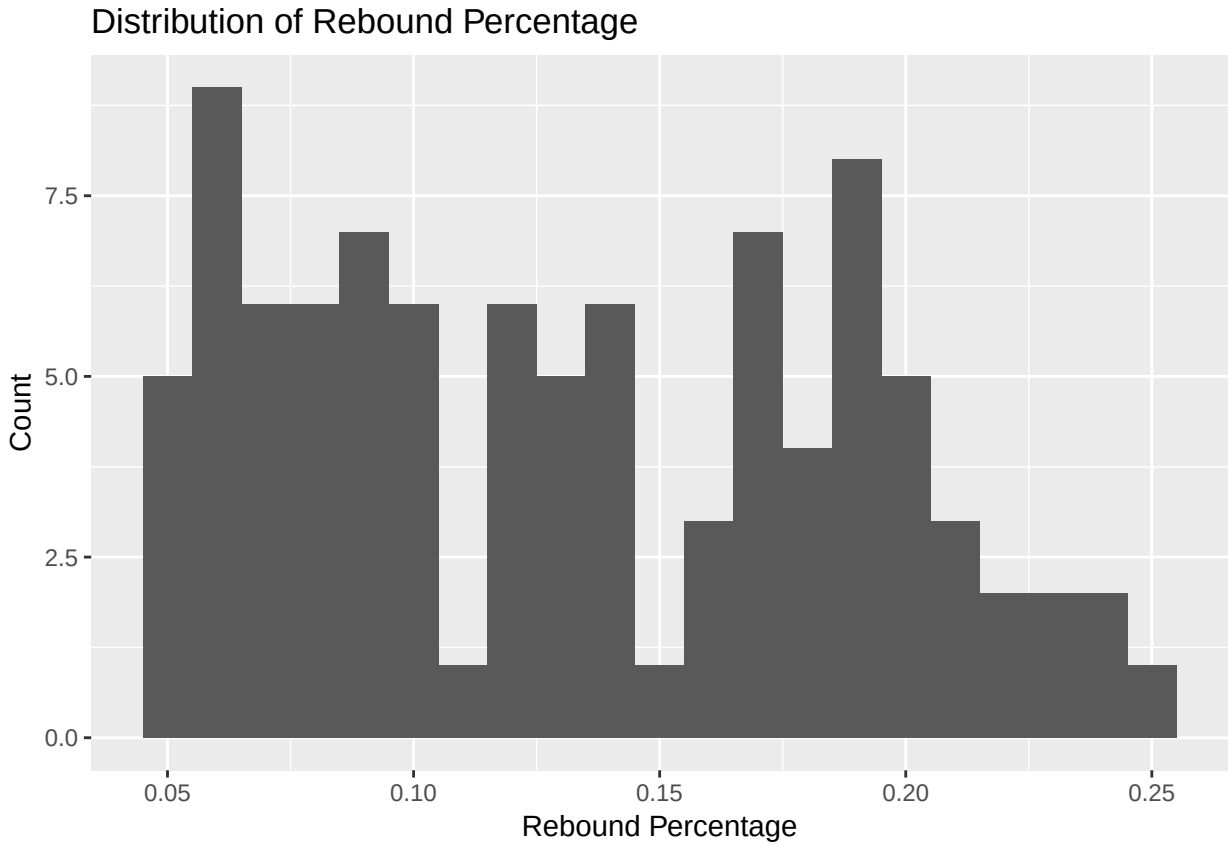



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  17.1     4  9.75     15  22.2  51.5
```

```
## # A tibble: 2 x 29
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING AST_RATIO
##   <chr>       <chr>            <dbl> <dbl>      <dbl>      <dbl>     <dbl>
## 1 Jarnell St~ DEN                 23 0           115.       118.      51.5
## 2 Ricky Rubio MIN                 26 0.373       109.       110.      41.3
## # ... with 22 more variables: REB_PCT <dbl>, USG_PCT <dbl>, PIE <dbl>,
## #   SALARY_MILLIONS <dbl>, ACTIVE_TWITTER_LAST_YEAR <dbl>,
## #   TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>, PTS <dbl>, FGM <dbl>, FGA <dbl>,
## #   ageCent <dbl>, ast_ratioCent <dbl>, off_ratingCent <dbl>,
## #   def_ratingCent <dbl>, fgaCent <dbl>, fgmCent <dbl>, PIECent <dbl>,
## #   reb_pctCent <dbl>, usg_pctCent <dbl>, salary_millionsCent <dbl>,
## #   w_pctCent <dbl>, ptsCent <dbl>, active_twitter_lyear <fct>
```

As we can see from the histogram, the assists-to-turnovers ratio is very right skewed. The mode is at around 10, even though the median is at 15, and the mean is 17.12526, all of which are summary statistics that emphasize the right skew. This means that while most players in the dataset had a very high assists-to-turnovers ratio (meaning they had many more assists than turnovers), there is a wider variation among players with a high ratio and the players with lower ratios are concentrated around a few numbers. The dataset minimum ratio of 4 means that there were no players with more turnovers than assists. Notably, this is the first variable we've examined so far with a significantly non-normal distribution. The two players with very high assists-to-turnovers ratios, 51.5 and 41.3, are Jarnell Stokes and Ricky Rubio, respectively.

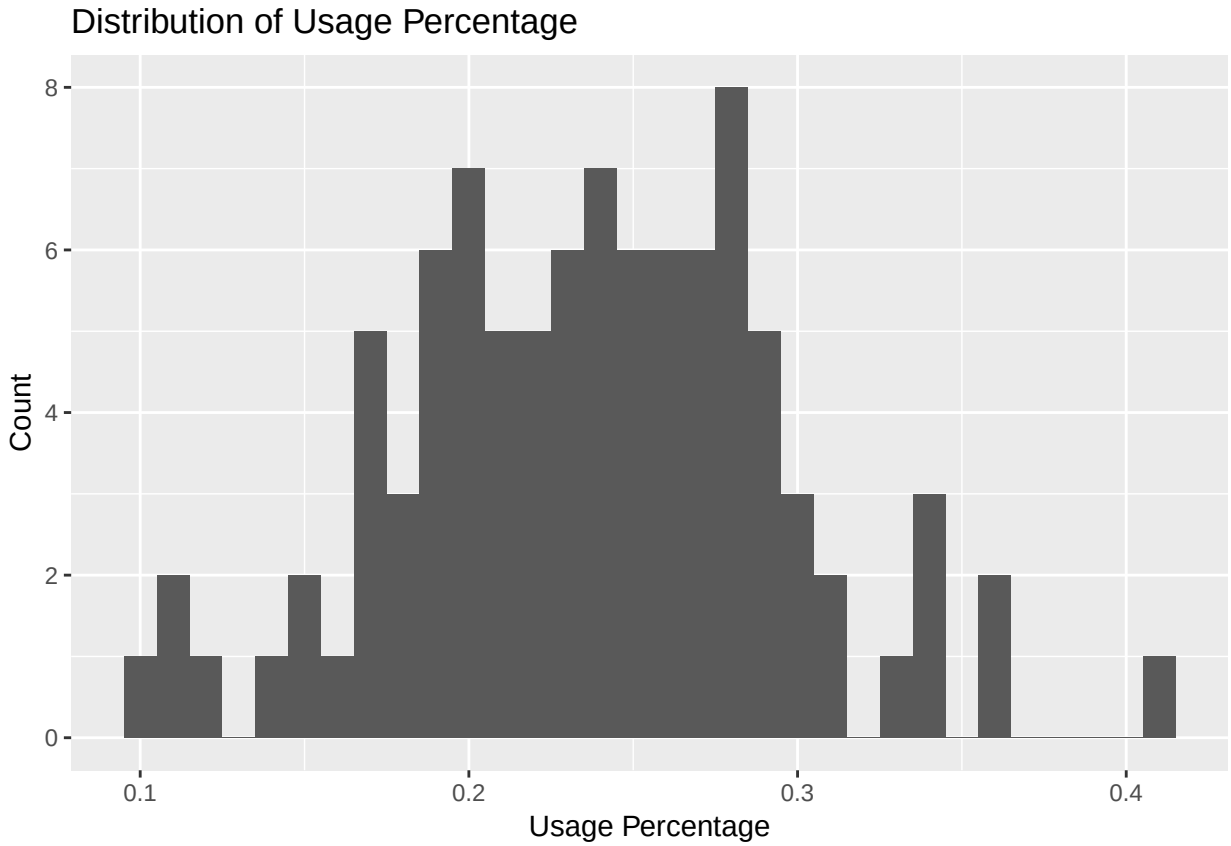
Next, we'll examine the variation in the percent of rebounds a player grabs:




```
## # A tibble: 1 x 6
##    mean   min     Q1 median    Q3   max
##   <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 0.133 0.045 0.0825  0.127 0.180 0.252
```

The distribution of rebound percentage has a minimum of 0.045 and a maximum of 0.252. The distribution is not very skewed one way or another, as supported by the similar mean of .133 and median of 0.127. However, the distribution is not normal in that it does not resemble a bell curve; with exceptions, the data is somewhat evenly distributed from the minimum to near the maximum (although there is some trail-off towards the right side of the distribution). This non-normal spread is likely partially an indication of the fact that the dataset contains both offensive and defensive players, because whether a player is on offense or defense has a significant effect on their rebound percentage.

Next, we'll look at usage percentage, which is an estimate of how often a player makes team plays:

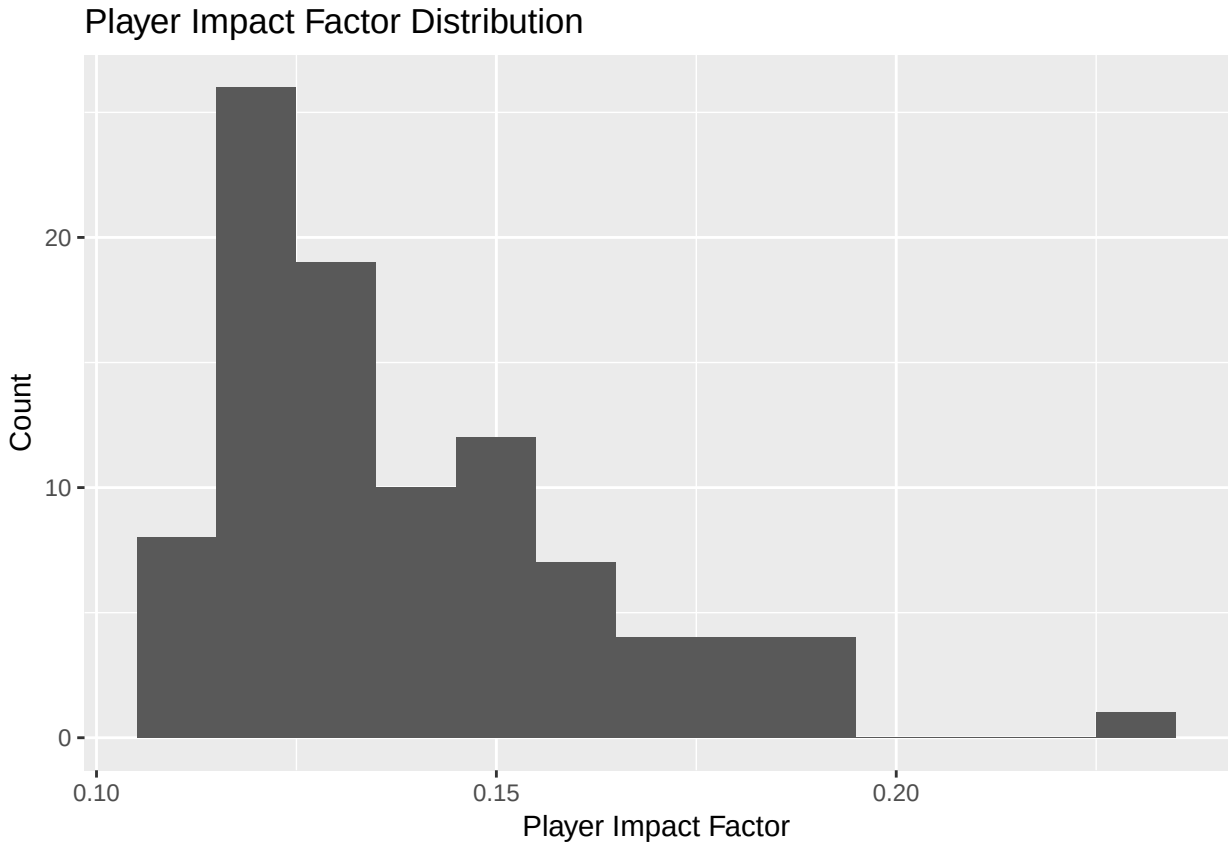



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.238 0.101   0.2  0.242 0.276 0.408
```

```
## # A tibble: 1 x 29
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING AST_RATIO
##   <chr>       <chr>            <dbl> <dbl>      <dbl>      <dbl>     <dbl>
## 1 Russell We~ OKC                 28 0.568       108.       105.      23.4
## # ... with 22 more variables: REB_PCT <dbl>, USG_PCT <dbl>, PIE <dbl>,
## #   SALARY_MILLIONS <dbl>, ACTIVE_TWITTER_LAST_YEAR <dbl>,
## #   TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>, PTS <dbl>, FGM <dbl>, FGA <dbl>,
## #   ageCent <dbl>, ast_ratioCent <dbl>, off_ratingCent <dbl>,
## #   def_ratingCent <dbl>, fgaCent <dbl>, fgmCent <dbl>, PIECent <dbl>,
## #   reb_pctCent <dbl>, usg_pctCent <dbl>, salary_millionsCent <dbl>,
## #   w_pctCent <dbl>, ptsCent <dbl>, active_twitter_lyear <fct>
```

The distribution of usage percentage, with a minimum of .101 and a maximum of .408, is fairly normally distributed. The mean, .238, and median, .242, are similar. The fairly wide spread may indicate that the dataset contains a decent sampling of players- some 'star player' types and others that are not the centerpieces of their teams.The maximum of .408, while perhaps not quite an outlier, is separated from most of the other points; this usage percentage belongs to Russell Westbrook.

Next, we'll examine player impact factor (PIE), a statistic roughly measuring a player's impact on the games that they play that's used by nba.com:

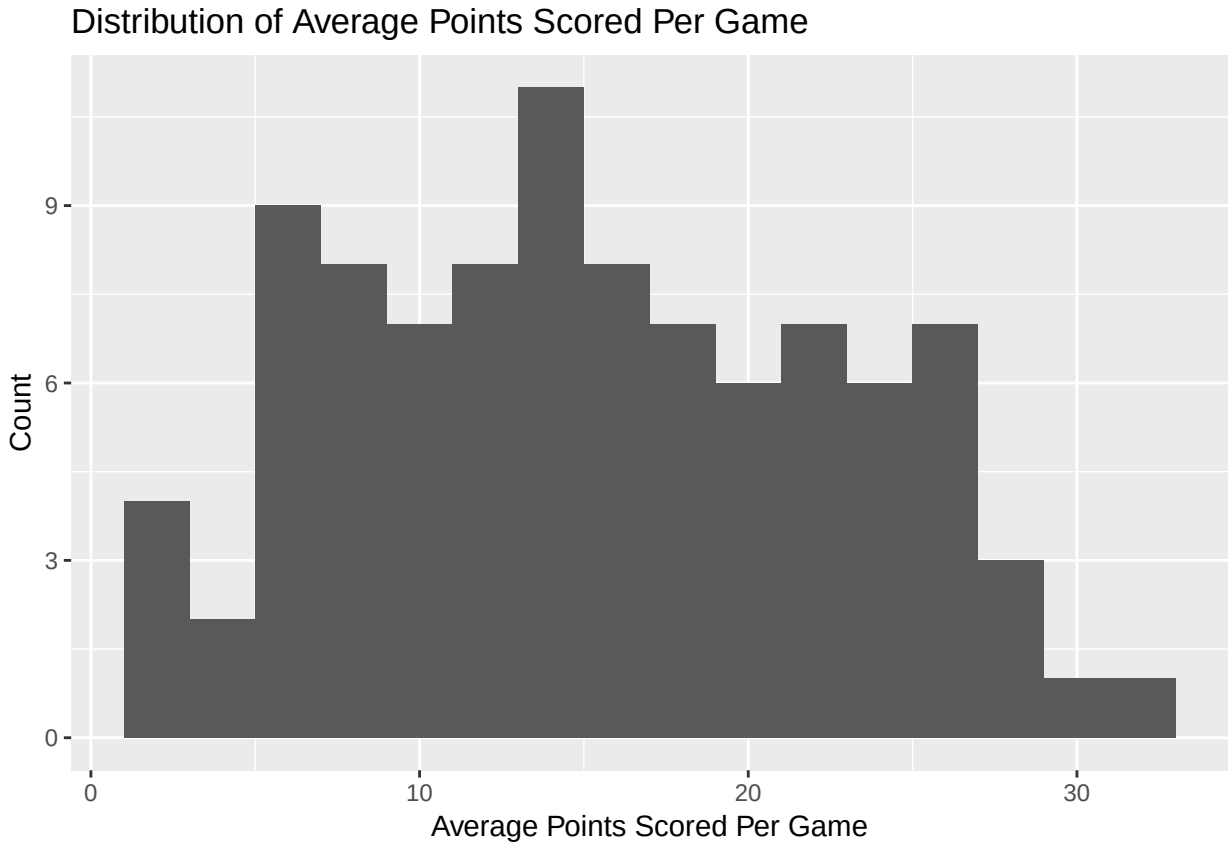



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.139 0.112 0.122  0.131 0.152  0.23
```

```
## # A tibble: 10 x 2
##    PLAYER_NAME             PIE
##    <chr>                 <dbl>
##  1 Russell Westbrook     0.23
##  2 Demetrius Jackson     0.194
##  3 Anthony Davis         0.192
##  4 James Harden          0.19
##  5 Kevin Durant          0.186
##  6 LeBron James          0.183
##  7 Chris Paul            0.182
##  8 DeMarcus Cousins      0.178
##  9 Giannis Antetokounmpo 0.176
## 10 Kawhi Leonard         0.174
```

As we can see from the histogram, the player impact factor, with a minimum of .112 and a maximum of .23, is quite right-skewed. The median player impact factor is .131, and the mean is .139, evidence of the right skew. The mode is around the median. The maximum, .23, is a significant outlier, and is that of Russell Westbrook, the same player who had by far the highest usage percentage; clearly, his data will need to be examined more closely later to see if it ultimately affects our model.

Next, we'll look at average points scored per game:




```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  15.3   1.5   9.5   14.6  21.4  31.6
```

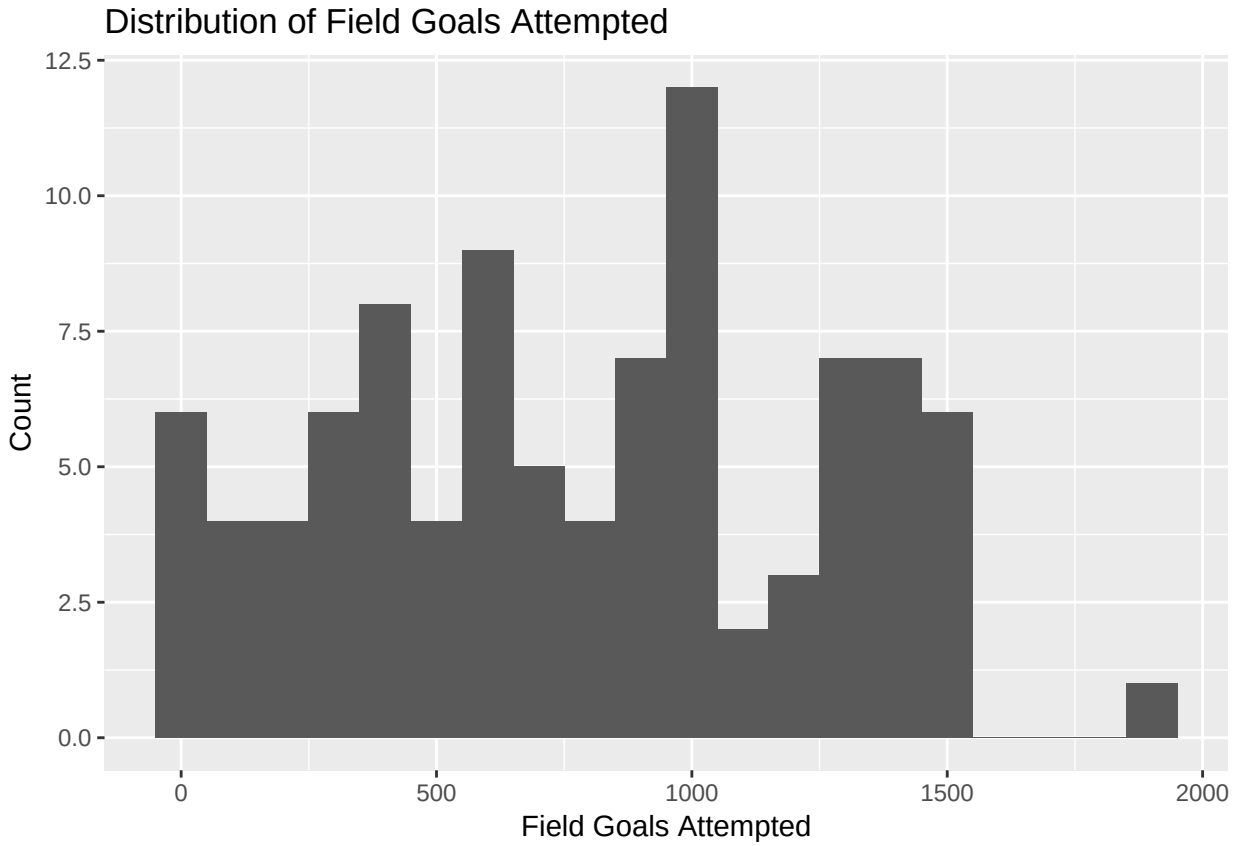
As we can see from the histogram, the distribution of average points scored is slightly normal, with some obvious departures from normality. The median number of points scored per game is 14.6, and the mean is 15.28232. The maximum is 31.6, but this does not seem to be an obvious outlier.

Next, we'll examine summary statistics about players' Twitter activity in 2015-2016:

```
## # A tibble: 2 x 2
##   ACTIVE_TWITTER_LAST_YEAR     n
##   <fct>                    <int>
## 1 0                            2
## 2 1                           93
```

Out of the 95 players in our modified dataset, 2 were not active on Twitter the year before the data was collected and and 93 were.

Next, we'll examine the distribution of field goals attempted:

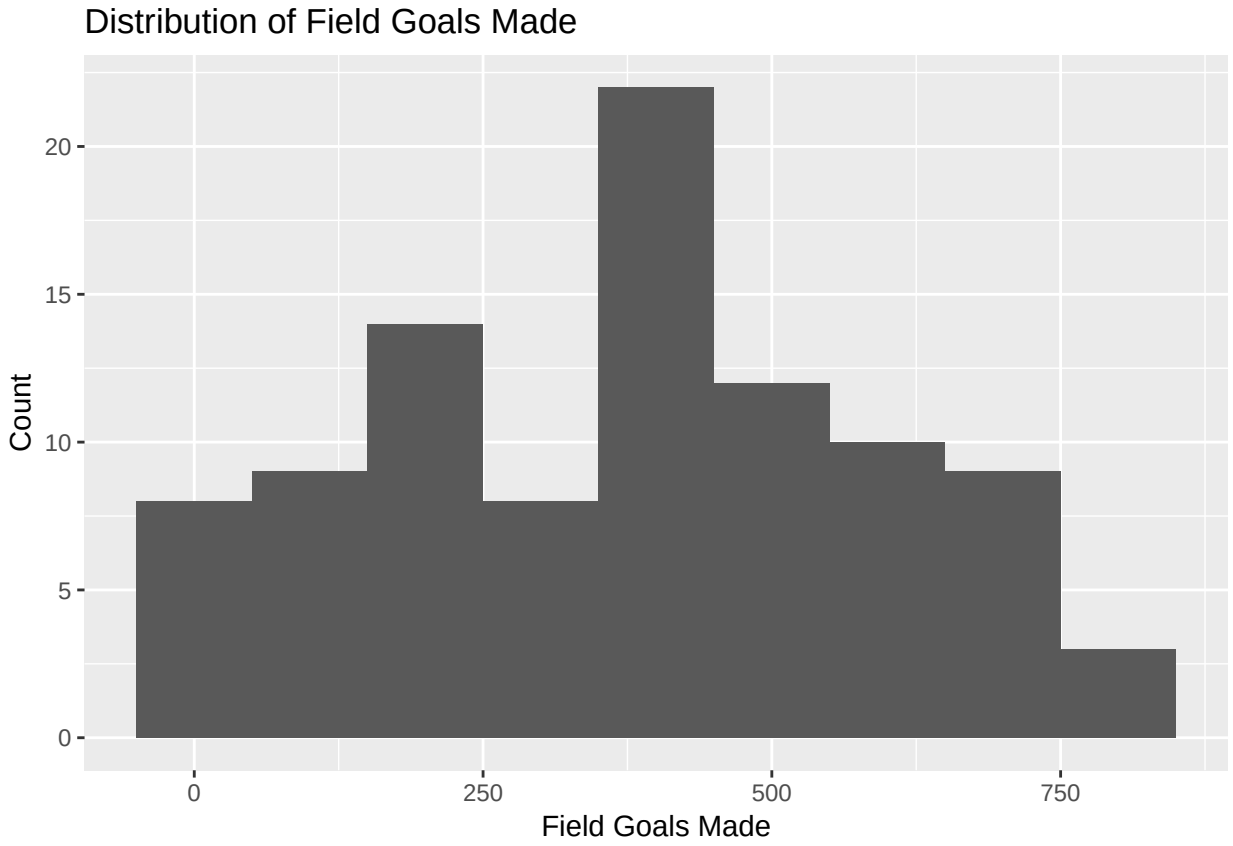



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  785.     1  412.    785  1140  1941
```

```
## # A tibble: 1 x 2
##   PLAYER_NAME         FGA
##   <chr>             <dbl>
## 1 Russell Westbrook  1941
```

The distribution of field goals attempted is non-normal but does not have a particular skew in either direction. Since this is a cumulative measure, it likely largely depends on a player's position and the amount of playing time they have had in NBA games. The outlier, at 1941 field goals attempted, is Russell Westbrook. This variable varies between 1 and 1941 field goals attempted, with a median of 785 and a mean of 785.27.

Now, we'll look at the distribution of field goals made:




```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  380.     1  216.    393  544.   824
```

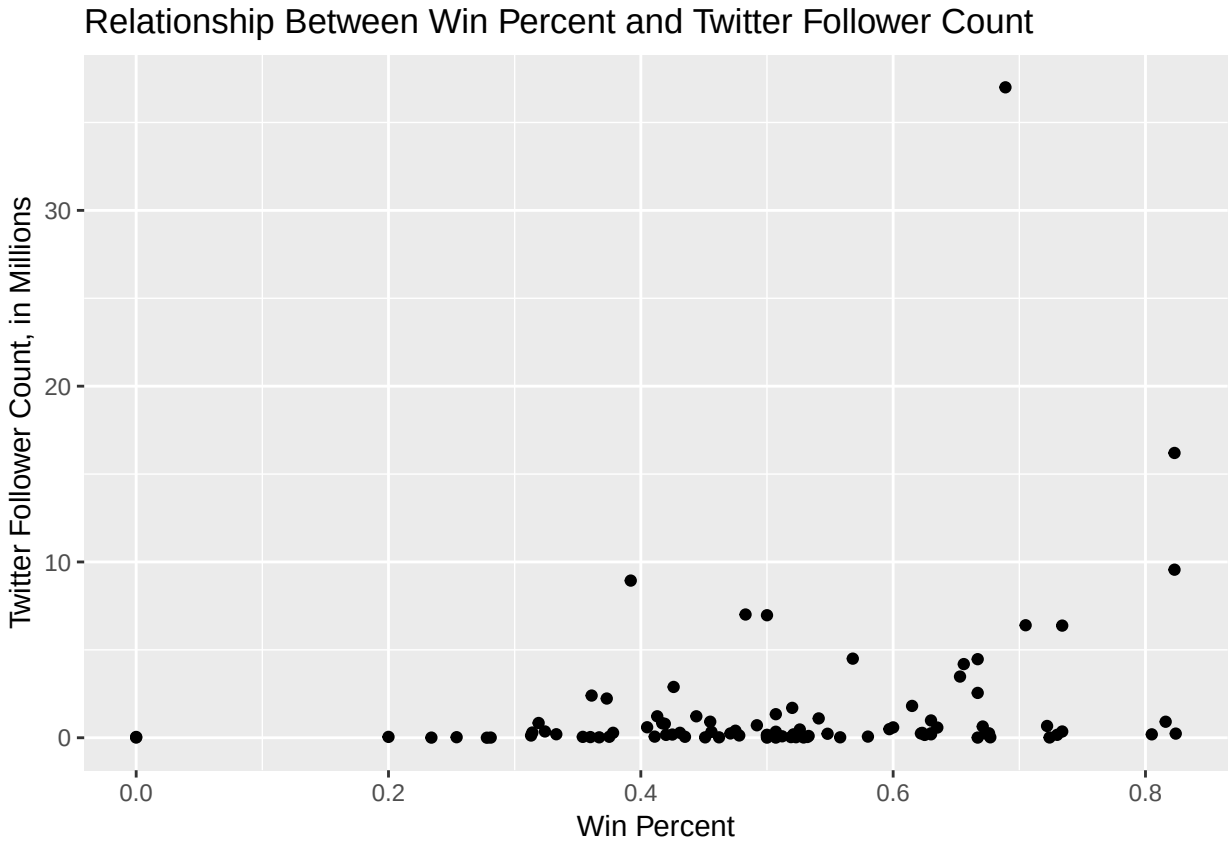
```
## # A tibble: 1 x 2
##   PLAYER_NAME         FGM
##   <chr>             <dbl>
## 1 Russell Westbrook   824
```

The distribution of field goals made is somewhat normal, with a mean of 379.8 and a median of 392 showing a slight right skew. The minimum is 1 and the maximum is 824, by Russell Westbrook.

Bivariate

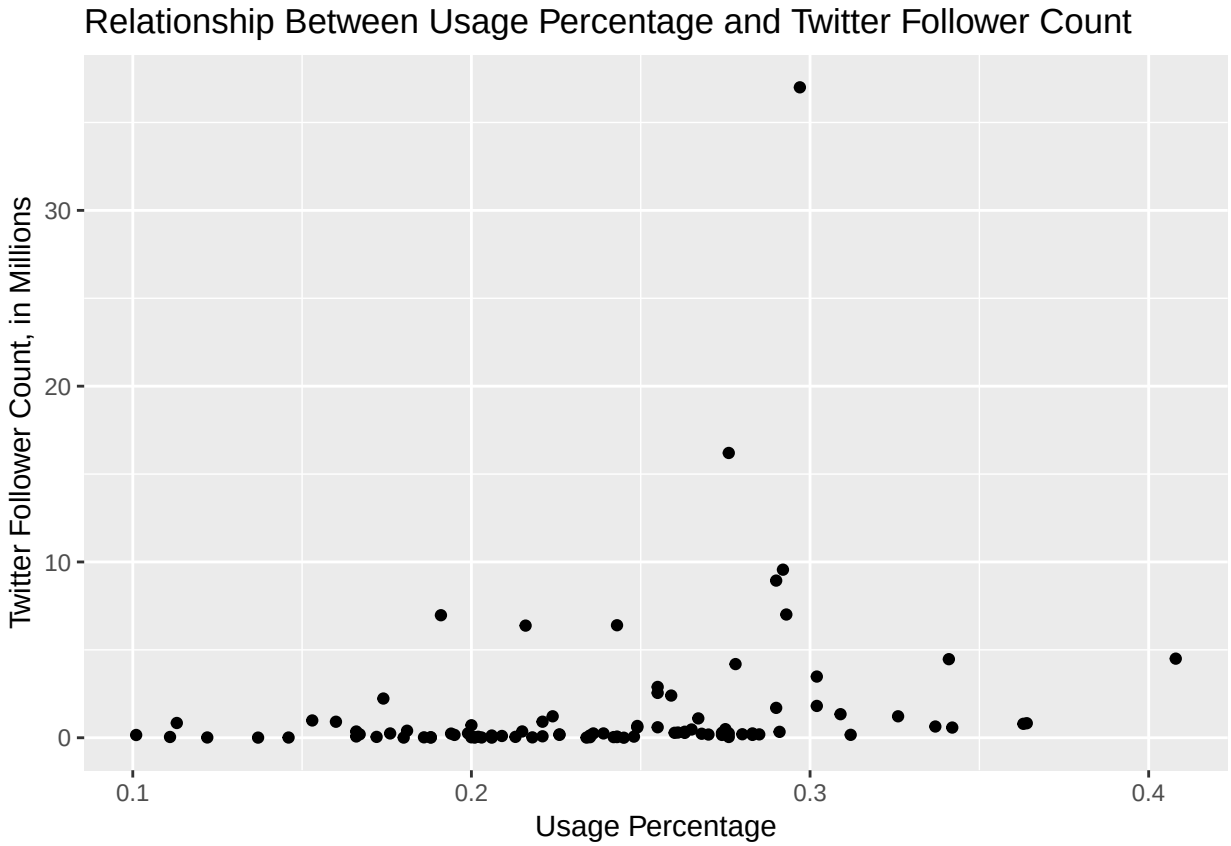
Next, we will do bivariate EDA, looking into the relationships of some of the predictor variables with the response variables. We won't do bivariate EDA on player name, Twitter handle, age, team abbreviation, or whether the players were active on Twitter last year, instead focusing on terms we believe may play a more nuanced / important role in predicting Twitter followers.

First, we'll look for a relationship between win percentage and the number of Twitter followers in millions:



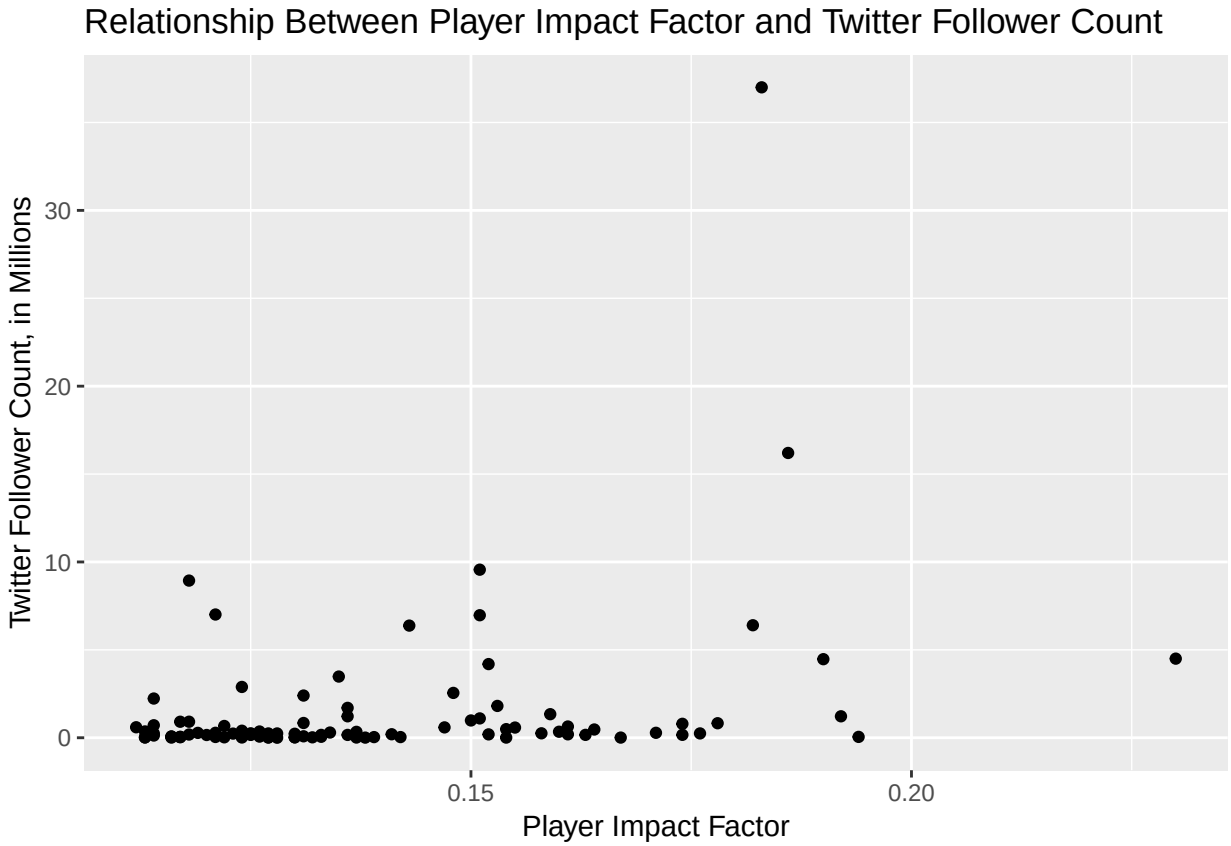
From the above plot, we can see that win percent and Twitter follower count may have a very weak positive orrelations. The players with significantly higher-than-average Twitter follower counts tend to have higher win percentages; however, this relationship is very weak.

Next, we'll examine whether there is a relationship between usage percentage and Twitter follower count:



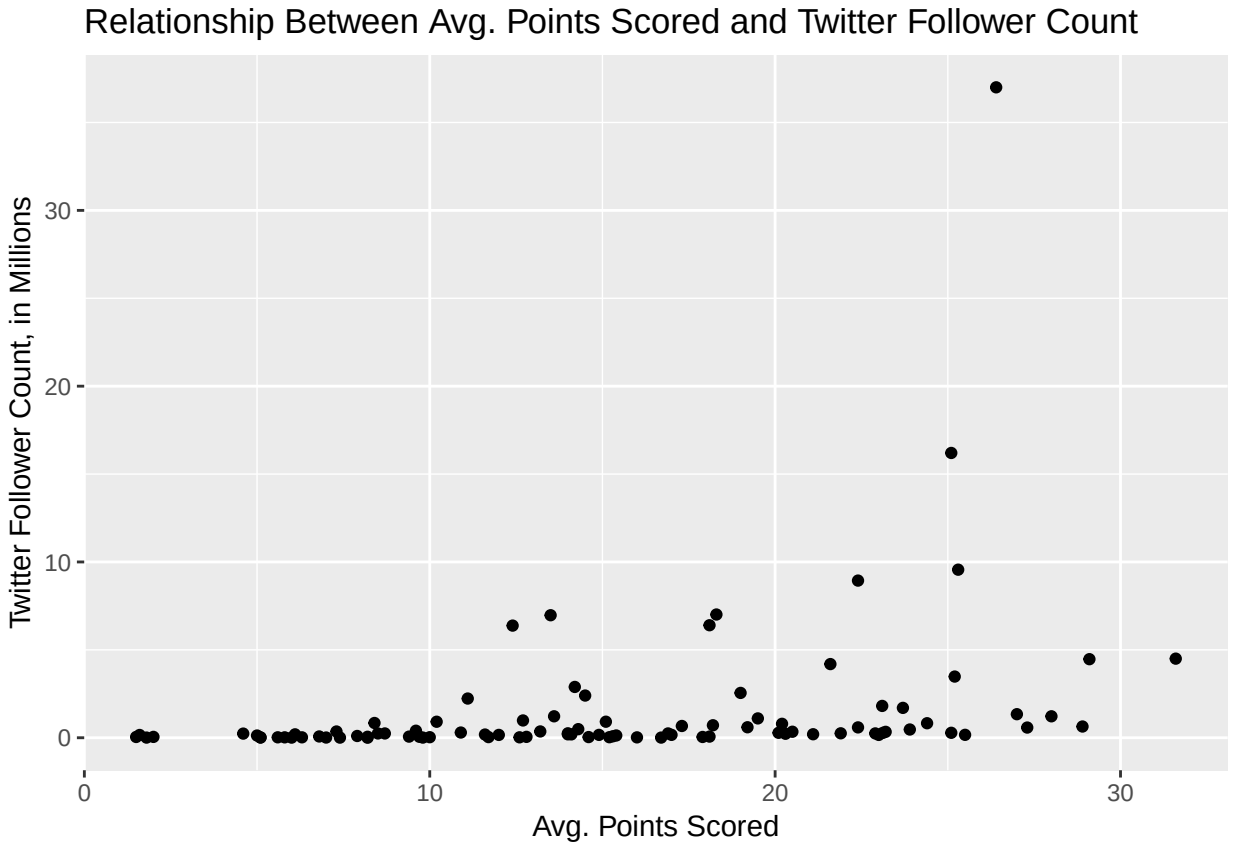
There appears to be a very weak positive correlation between usage percentage and Twitter follower count; players with a high usage percentage tend to have more Twitter followers, on average, than those with a lower usage percentage.

Next, we'll examine whether there is a relationship between player impact factor (PIE) and Twitter follower count:



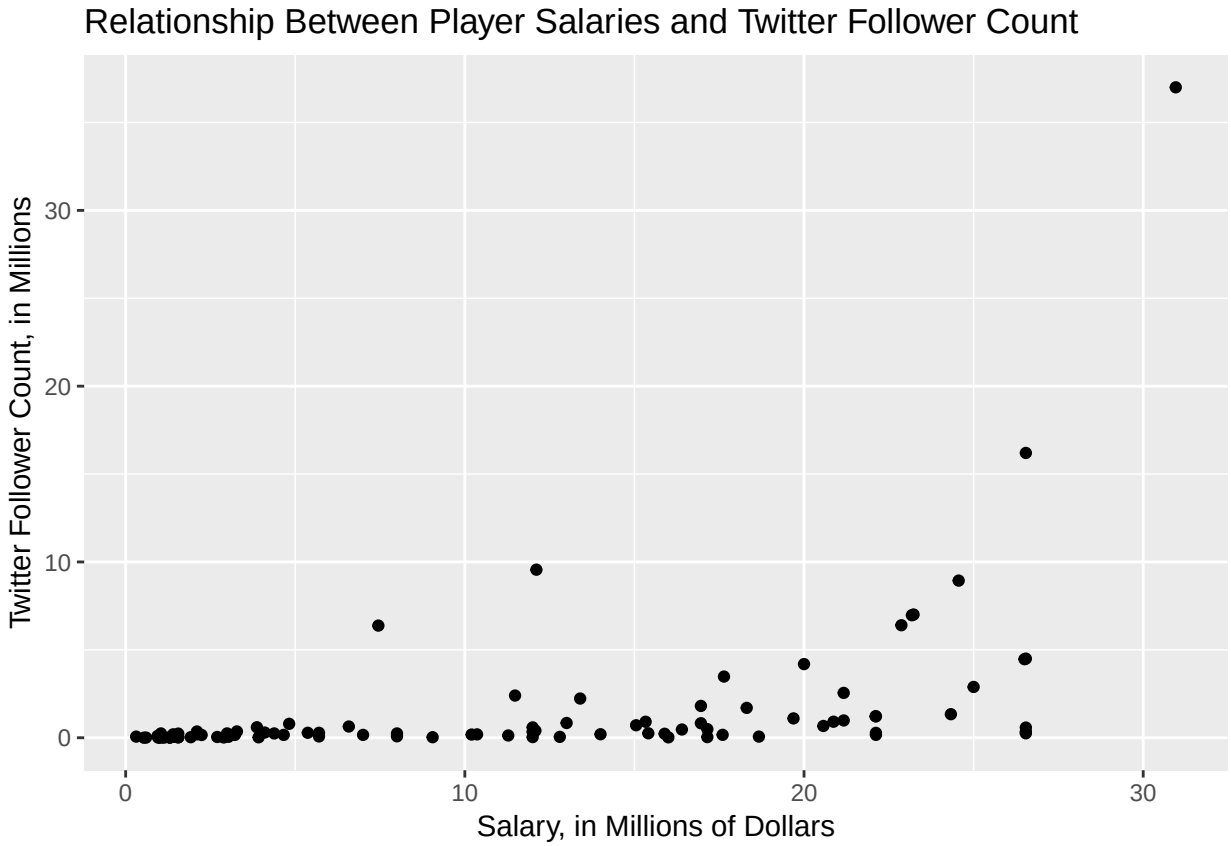
There appears to be a somewhat positive correlation between player impact factor and Twitter follower count; players with a high player impact factor tend to have more Twitter followers, on average, than those with a lower player impact factor.

Now, we'll look for a relationship between average points scored per game and Twitter follower count:



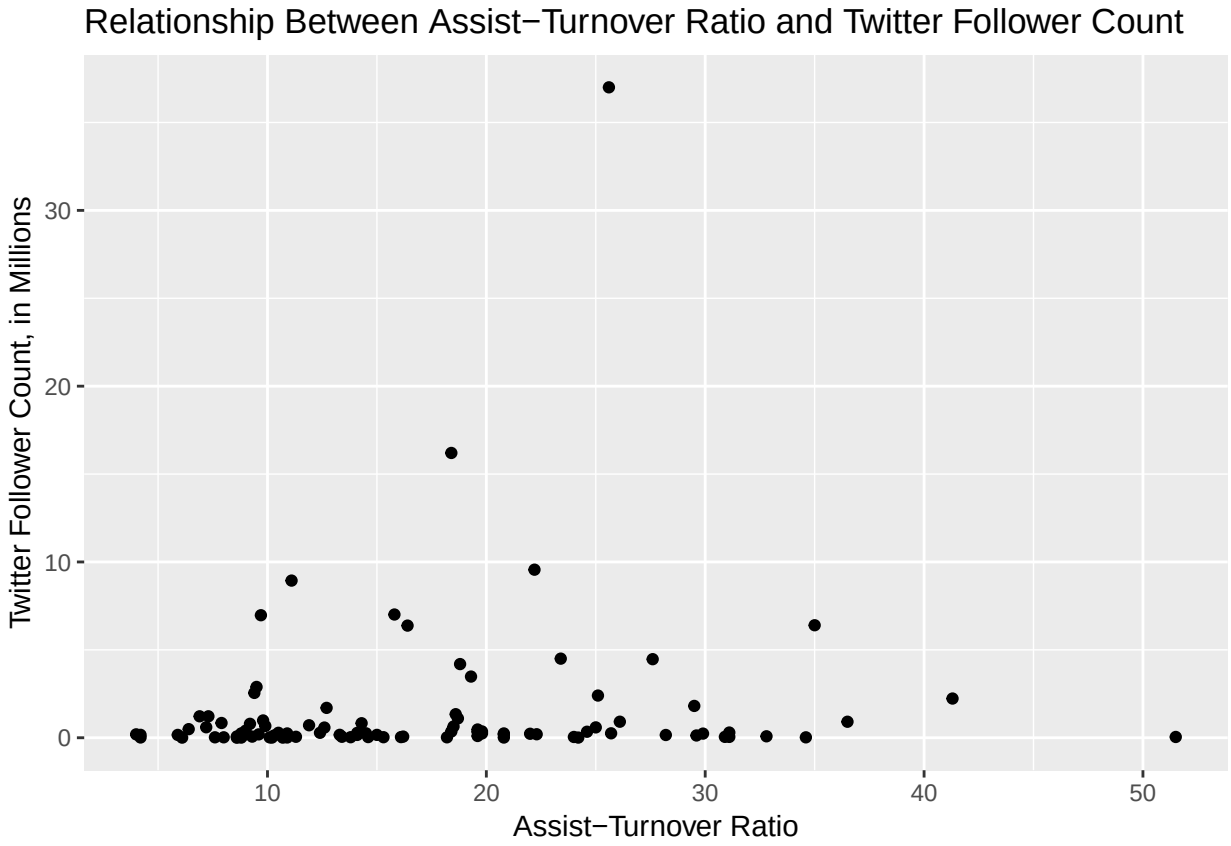
There appears to be a weak positive correlation between average points scored and Twitter follower count; players with higher avg. points scored tend to have more Twitter followers than those with lower avg. points scored.

Next, we'll look for a relationship between player salaries and points scored:



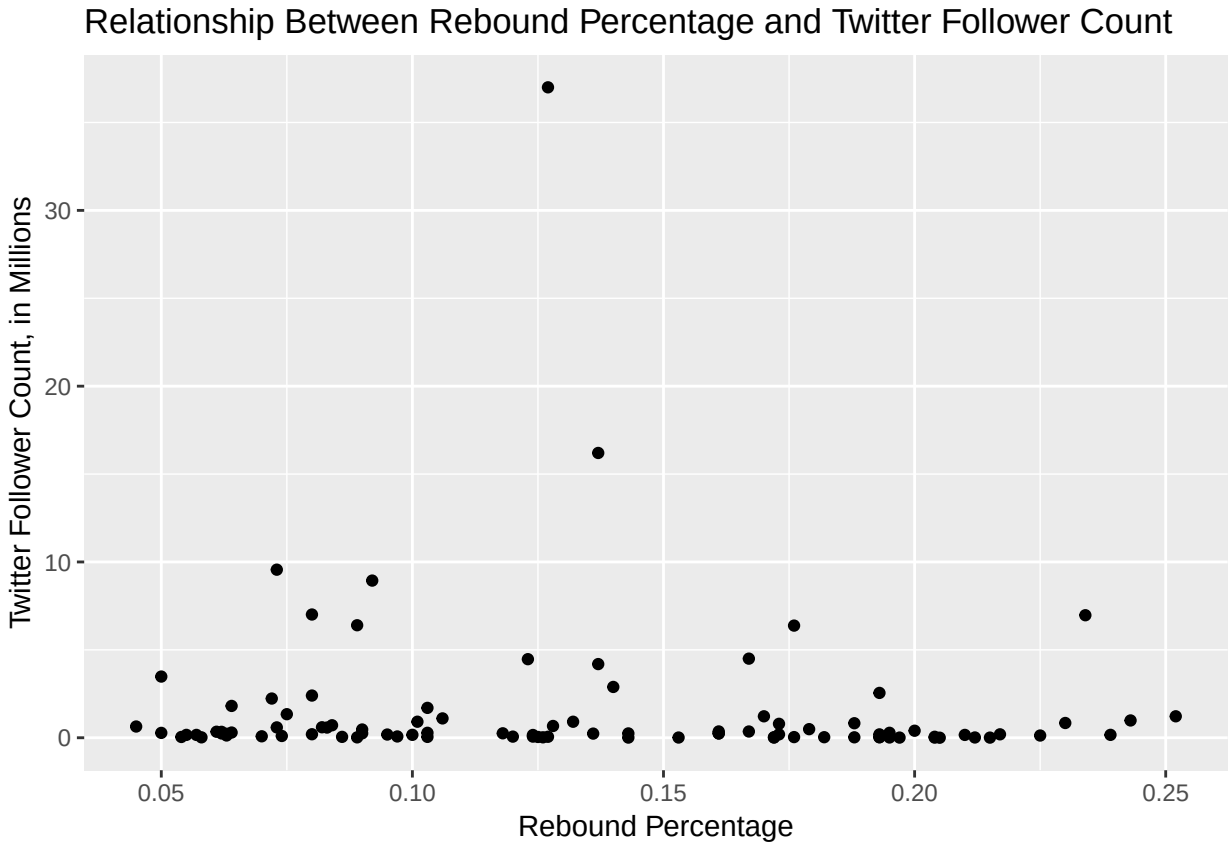
There appears to be a positive correlation between salary and Twitter follower count; the players with higher salaries tend to have higher Twitter follower counts.

Here, we'll look at the relationship between assists-to-turnovers ratio and Twitter follower count:



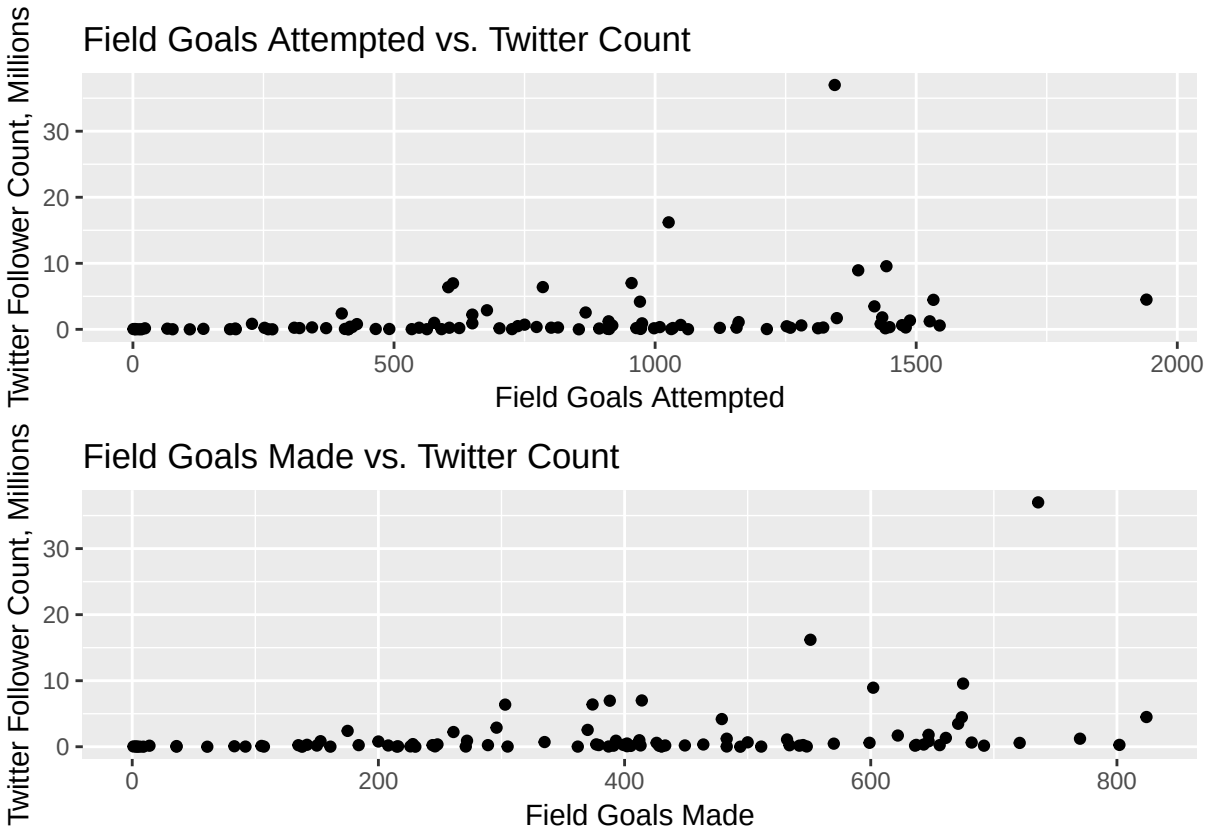
There is no evident relationship between assists-to-turnovers ratio and Twitter follower count.

Next, we'll examine whether there is a relationship between rebound percentage and Twitter follower count:



There is no evident relationship between rebound percentage and Twitter follower count.

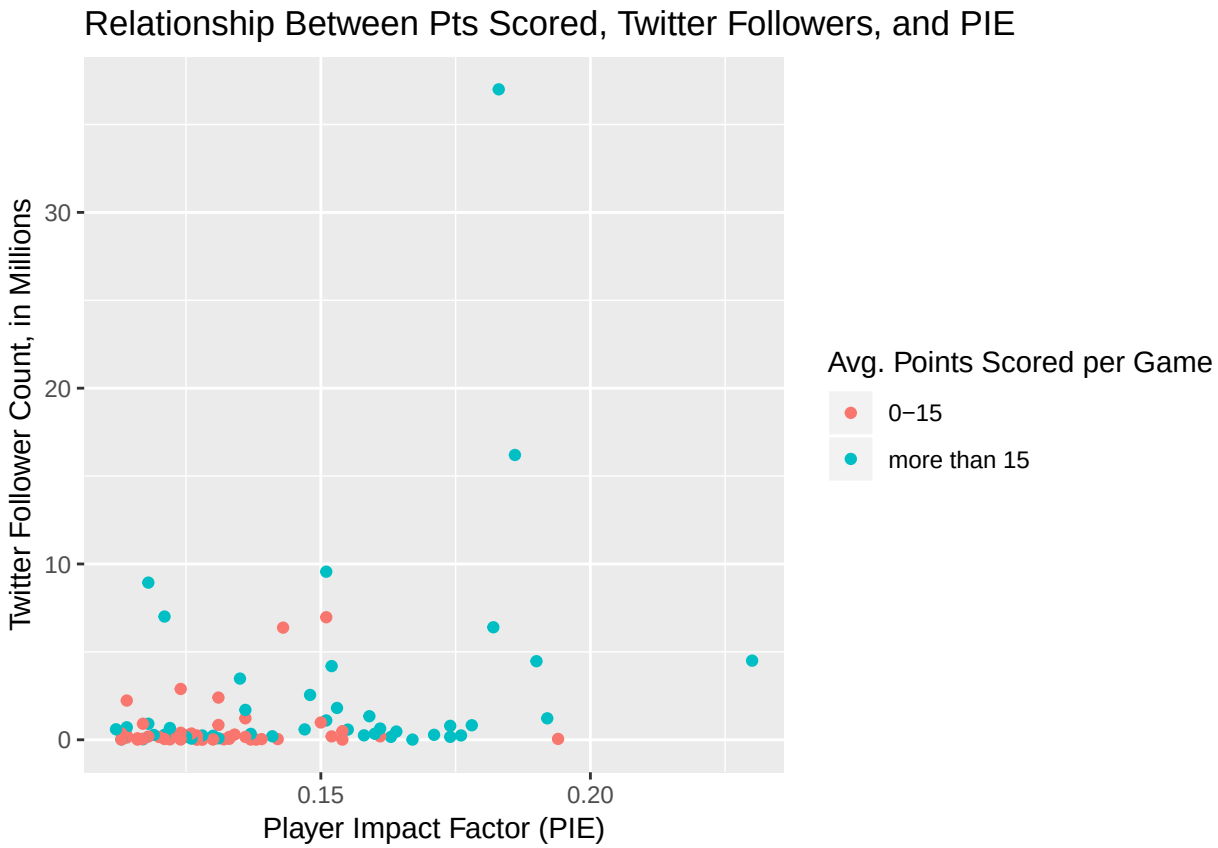
Now, we'll look for a relationship between field goals attempted and Twitter follower count as well as between field goals made and Twitter follower count:



Multivariate Data Analysis

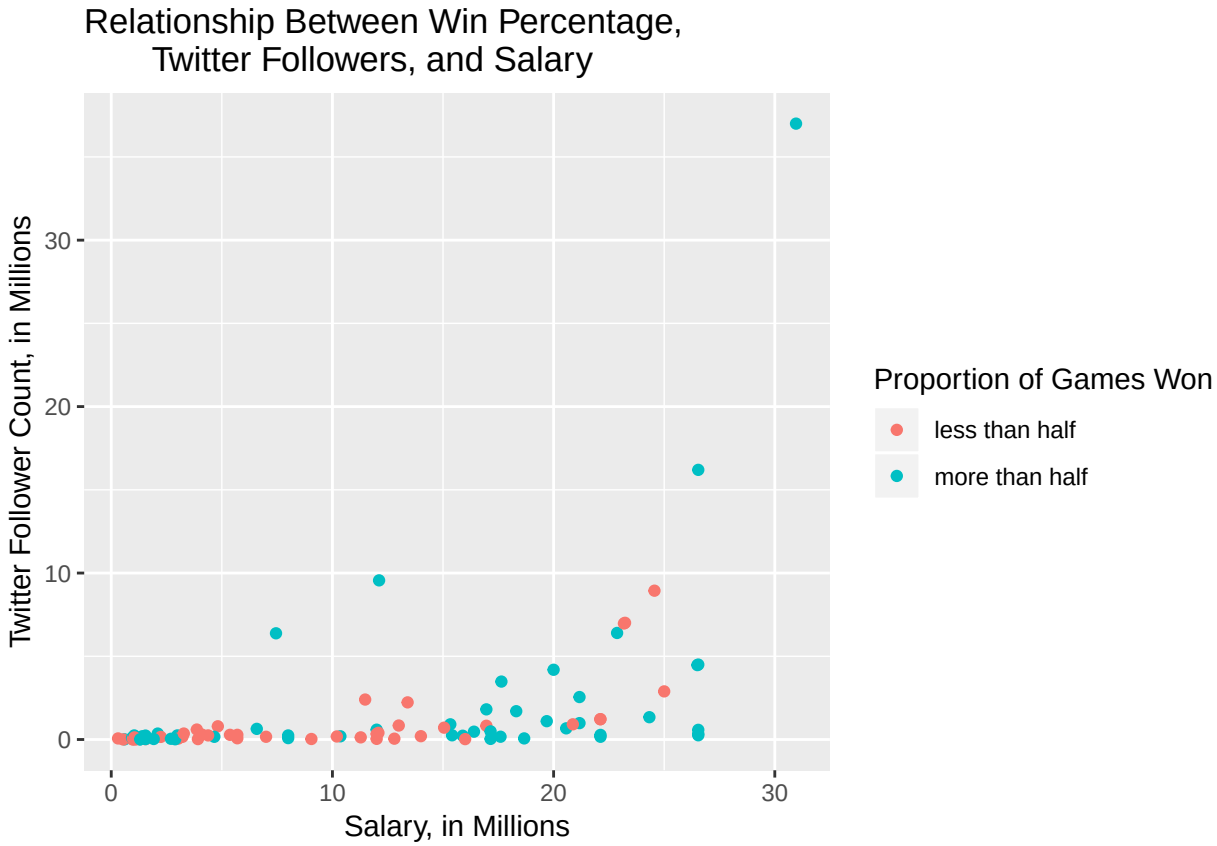
Now, we'll do some multivariate analysis. In this section, we are looking for predictor variables that may affect the way other predictor variables relate to the response variable.

First, we'll look to see if points scored affects the way player impact factor (PIE) relates to Twitter follower count:



As we can see from the above color-coded scatterplot, many of the players with the most points scored have higher player impact factors, and player impact factor values have a weak positive correlation with the Twitter follower count. This could be an opportunity for an interaction term.

Next, we'll try to determine whether win percentage affects the way salary relates to the Twitter follower count:



As we can see from the scatterplot, players with higher win percentages tend to be paid more, and high salary has a weak positive correlation with the Twitter follower count. This could also be an opportunity for an interaction term.

Initial Model

First, we fit a model with thirteen main effect terms – mean-centered age, mean-centered assists-to-turnovers ratio, mean-centered player offensive rating, mean-centered player defensive rating, mean-centered player impact factor (PIE), mean-centered rebound percentage, mean-centered usage percentage, mean-centered salary, mean-centered win percentage, mean-centered points scored, mean-centered field goals made, mean-centered field goals attempted, and whether the player has an active Twitter account in 2015-2016 – and also considered interactions between mean-centered salary and mean-centered win percentage and mean-centered player impact factor (PIE) and mean-centered points scored:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.114	0.431	2.584	0.012	0.256	1.973
ageCent	0.270	0.117	2.314	0.023	0.038	0.502
ast_ratioCent	0.066	0.062	1.054	0.295	-0.058	0.190
off_ratingCent	0.041	0.100	0.407	0.685	-0.159	0.241
def_ratingCent	0.051	0.116	0.441	0.660	-0.179	0.281
PIECent	-12.266	27.217	-0.451	0.653	-66.440	41.908
reb_pctCent	-3.412	12.213	-0.279	0.781	-27.723	20.898
usg_pctCent	22.823	14.896	1.532	0.129	-6.826	52.472
salary_millionsCent	0.146	0.068	2.143	0.035	0.010	0.282
w_pctCent	4.023	3.770	1.067	0.289	-3.482	11.527
ptsCent	0.027	0.201	0.136	0.892	-0.372	0.427

term	estimate	std.error	statistic	p.value	conf.low	conf.high
fgmCent	0.044	0.013	3.253	0.002	0.017	0.071
fgaCent	-0.022	0.007	-3.412	0.001	-0.035	-0.009
active_twitter_lyear0	-4.384	2.629	-1.668	0.099	-9.617	0.848
salary_millionsCent:w_pctCent	0.817	0.338	2.419	0.018	0.145	1.490
PIECent:ptsCent	2.552	2.206	1.157	0.251	-1.839	6.943

Backward Selection (Iteration 1)

We will now perform the first iteration of backward selection using AIC as the selection criterion:

```
## Start:  AIC=256.02
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + off_ratingCent +
##     def_ratingCent + PIECent + reb_pctCent + usg_pctCent + salary_millionsCent +
##     w_pctCent + ptsCent + fgmCent + fgaCent + active_twitter_lyear +
##     salary_millionsCent * w_pctCent + PIECent * ptsCent
##
##                                 Df Sum of Sq    RSS    AIC
## - reb_pctCent                    1     0.992 1005.2 254.12
## - off_ratingCent                 1     2.101 1006.4 254.22
## - def_ratingCent                 1     2.473 1006.7 254.26
## - ast_ratioCent                  1    14.123 1018.4 255.35
## - PIECent:ptsCent                1    17.012 1021.3 255.62
## <none>                                       1004.3 256.02
## - usg_pctCent                    1    29.844 1034.1 256.80
## - active_twitter_lyear           1    35.367 1039.6 257.31
## - ageCent                        1    68.091 1072.3 260.25
## - salary_millionsCent:w_pctCent  1    74.371 1078.6 260.81
## - fgmCent                        1   134.485 1138.8 265.96
## - fgaCent                        1   148.018 1152.3 267.08
##
## Step:  AIC=254.12
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + off_ratingCent +
##     def_ratingCent + PIECent + usg_pctCent + salary_millionsCent +
##     w_pctCent + ptsCent + fgmCent + fgaCent + active_twitter_lyear +
##     salary_millionsCent:w_pctCent + PIECent:ptsCent
##
##                                 Df Sum of Sq    RSS    AIC
## - off_ratingCent                 1     2.021 1007.3 252.31
## - def_ratingCent                 1     2.615 1007.9 252.36
## - PIECent:ptsCent                1    16.272 1021.5 253.64
## <none>                                       1005.2 254.12
## - ast_ratioCent                  1    28.691 1033.9 254.79
## - usg_pctCent                    1    31.709 1037.0 255.07
## - active_twitter_lyear           1    34.674 1039.9 255.34
## - ageCent                        1    71.922 1077.2 258.68
## - salary_millionsCent:w_pctCent  1    75.264 1080.5 258.98
## - fgmCent                        1   141.844 1147.1 264.66
## - fgaCent                        1   151.488 1156.7 265.45
##
## Step:  AIC=252.31
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + def_ratingCent +
##     PIECent + usg_pctCent + salary_millionsCent + w_pctCent +
```

```
##     ptsCent + fgmCent + fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent +
##     PIECent:ptsCent
##
##                                 Df Sum of Sq    RSS    AIC
## - def_ratingCent                 1     4.664 1011.9 250.75
## - PIECent:ptsCent                1    14.257 1021.5 251.64
## <none>                                       1007.3 252.31
## - usg_pctCent                    1    30.100 1037.4 253.10
## - ast_ratioCent                  1    31.424 1038.7 253.23
## - active_twitter_lyear           1    33.980 1041.3 253.46
## - ageCent                        1    70.353 1077.6 256.72
## - salary_millionsCent:w_pctCent  1    92.123 1099.4 258.62
## - fgmCent                        1   140.414 1147.7 262.70
## - fgaCent                        1   150.523 1157.8 263.54
##
## Step:  AIC=250.75
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + PIECent +
##     usg_pctCent + salary_millionsCent + w_pctCent + ptsCent +
##     fgmCent + fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent +
##     PIECent:ptsCent
##
##                                 Df Sum of Sq    RSS    AIC
## - PIECent:ptsCent                1    13.909 1025.8 250.04
## <none>                                       1011.9 250.75
## - ast_ratioCent                  1    33.048 1045.0 251.80
## - active_twitter_lyear           1    34.460 1046.4 251.93
## - usg_pctCent                    1    35.228 1047.2 252.00
## - ageCent                        1    66.760 1078.7 254.81
## - salary_millionsCent:w_pctCent  1    99.502 1111.4 257.66
## - fgmCent                        1   151.084 1163.0 261.97
## - fgaCent                        1   154.094 1166.0 262.21
##
## Step:  AIC=250.04
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + PIECent +
##     usg_pctCent + salary_millionsCent + w_pctCent + ptsCent +
##     fgmCent + fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## - ptsCent                        1     0.455 1026.3 248.09
## - PIECent                        1     0.930 1026.8 248.13
## <none>                                       1025.8 250.04
## - ast_ratioCent                  1    33.045 1058.9 251.06
## - active_twitter_lyear           1    34.602 1060.5 251.19
## - usg_pctCent                    1    39.131 1065.0 251.60
## - ageCent                        1    69.400 1095.2 254.26
## - salary_millionsCent:w_pctCent  1   105.759 1131.6 257.36
## - fgmCent                        1   145.528 1171.4 260.65
## - fgaCent                        1   148.123 1174.0 260.86
##
## Step:  AIC=248.08
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + PIECent +
##     usg_pctCent + salary_millionsCent + w_pctCent + fgmCent +
##     fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
```

```
##                                 Df Sum of Sq    RSS    AIC
## - PIECent                        1     0.803 1027.1 246.16
## <none>                                       1026.3 248.09
## - ast_ratioCent                  1    33.385 1059.7 249.13
## - active_twitter_lyear           1    36.163 1062.5 249.38
## - usg_pctCent                    1    55.786 1082.1 251.11
## - ageCent                        1    68.956 1095.3 252.26
## - salary_millionsCent:w_pctCent  1   111.612 1137.9 255.89
## - fgaCent                        1   148.408 1174.7 258.92
## - fgmCent                        1   152.494 1178.8 259.25
##
## Step:  AIC=246.16
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + usg_pctCent +
##     salary_millionsCent + w_pctCent + fgmCent + fgaCent + active_twitter_lyear +
##     salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## <none>                                       1027.1 246.16
## - ast_ratioCent                  1    33.008 1060.1 247.16
## - active_twitter_lyear           1    36.117 1063.2 247.44
## - usg_pctCent                    1    63.678 1090.8 249.87
## - ageCent                        1    70.344 1097.5 250.45
## - salary_millionsCent:w_pctCent  1   111.016 1138.1 253.91
## - fgaCent                        1   168.949 1196.1 258.63
## - fgmCent                        1   174.135 1201.2 259.04
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.310	0.381	3.436	0.001	0.552	2.068
ageCent	0.266	0.110	2.413	0.018	0.047	0.485
ast_ratioCent	0.078	0.047	1.653	0.102	-0.016	0.171
usg_pctCent	25.026	10.902	2.296	0.024	3.350	46.701
salary_millionsCent	0.138	0.061	2.259	0.026	0.017	0.259
w_pctCent	4.426	2.706	1.636	0.106	-0.954	9.807
fgmCent	0.042	0.011	3.796	0.000	0.020	0.063
fgaCent	-0.021	0.006	-3.739	0.000	-0.032	-0.010
active_twitter_lyear0	-4.378	2.532	-1.729	0.087	-9.413	0.657
salary_millionsCent:w_pctCent	0.915	0.302	3.031	0.003	0.315	1.515

Based on the output displayed above from the first iteration of backward selection (using AIC as the selection criterion), five main effect terms (mean-centered player offensive rating, mean-centered player defensive rating, mean-centered player impact factor, mean-centered rebound percentage, and mean-centered points scored) and the interaction between mean-centered player impact factor and mean-centered points scored were removed.

However, three terms in the selected model – mean-centered assists-to-turnovers ratio, mean-centered win percentage, and whether the player had an active Twitter account in 2015-2016 – have high p-values, 0.102, 0.106, and 0.087 respectively. Furthermore, the confidence intervals for these slope coefficients, [-0.016, 0.171], [-0.954, 9.807] and [-0.657, 9.413] respectively, include zero. Mean-centered win percentage will need to remain in the model to keep the statistically significant interaction between mean-centered salary and mean-centered win percentage. But, we can reasonably infer mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016 may be a particularly troublesome predictors in the model.

We will proceed with the second iteration of backward selection and will revisit the issue of these troublesome predictors after viewing the selected linear regression model based on adjusted R-squared.

Backward Selection (Iteration 2)

Next, we will perform the second iteration of backward selection using adjusted R-squared as the selection criterion:

```
##                    (Intercept)                         ageCent
##                     1.37224766                      0.30459234
##                  ast_ratioCent                     usg_pctCent
##                     0.08454854                     26.69292449
##            salary_millionsCent                         fgmCent
##                     0.15051918                      0.04677193
##                         fgaCent          active_twitter_lyear0
##                    -0.02298323                     -4.06738689
## salary_millionsCent:w_pctCent
##                     0.74874489
```

Based on the output displayed above from the second iteration of backward selection (using adjusted R-squared as the selection criterion), six main effect terms (mean-centered player offensive rating, mean-centered player defensive rating, mean-centered player impact factor, mean-centered rebound percentage, mean-centered win percentage, and mean-centered points scored) and and the interaction between mean-centered player impact factor and mean-centered points scored were removed.

Model Comparison: AIC vs. Adjusted R-squared

We noticed the model selected using AIC as the selection criterion includes an additional quantitative term (mean-centered win percentage) which was omitted in the model selected based on adjusted R-squared.

We decided to keep mean-centered win percentage in the model so that the statistically significant interaction between mean-centered salary and mean-centered win percentage could remain in the model as well.

Then, we revisited the issue of troublesome (seemingly insignificant) predictors. Unfortunately, both selected models included mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016; thus, we had to decide whether to keep the variables in the model, or to ignore the results from the two iterations of backward selection and remove them.

To answer this question, we compared the AIC and adjusted R-squared values for the model selected based on AIC as the selection criterion (first iteration) and the same model without mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016:

```
## [1] 517.7576
```

```
## [1] 0.3883516
```

```
## [1] 519.8771
```

```
## [1] 0.3626515
```

Based on the above output, the AIC of the model with all terms is roughly 517.76. Conversely, the AIC of the model without the statistically insignificant terms is roughly 519.88.

Moreover, the adjusted R-squared value for the model with all terms is roughly 0.388, whereas the adjusted R-squared value for the model without the statistically insignificant terms is about 0.363.

Therefore, the model with all terms maximizes adjusted R-squared and minimizes AIC. Since adjusted R-squared penalizes for unnecessary predictors, the fact that the model with all terms had a higher adjusted R-squared value means that, despite the high p-values and the presence of zero in the confidence intervals associated with mean-centered assists-to-turnovers ratio and whether the player had an active Twitter account in 2015-2016, we can conclude these predictors are valuable in predicting the response, the number of Twitter followers (in millions).

Impact of Prominent Players

Lastly, before discussing assumptions, we examined the impact of including versus excluding prominent athletes in our model. Since he is widely regarded as one of the best NBA players of all-time, we used LeBron James as a case study for preliminary analysis:

```
## # A tibble: 10 x 2
##    PLAYER_NAME       SALARY_MILLIONS
##    <chr>                       <dbl>
##  1 LeBron James                 31.0
##  2 Russell Westbrook            26.5
##  3 Kevin Durant                 26.5
##  4 Mike Conley                  26.5
##  5 DeMar DeRozan                26.5
##  6 Al Horford                   26.5
##  7 James Harden                 26.5
##  8 Dirk Nowitzki                25
##  9 Carmelo Anthony              24.6
## 10 Damian Lillard               24.3
```

```
## # A tibble: 10 x 2
##    PLAYER_NAME       TWITTER_FOLLOWER_COUNT_MILLIONS
##    <chr>                                       <dbl>
##  1 LeBron James                                37
##  2 Kevin Durant                                16.2
##  3 Stephen Curry                                9.56
##  4 Carmelo Anthony                              8.94
##  5 Dwyane Wade                                  7.01
##  6 Dwight Howard                                6.97
##  7 Chris Paul                                   6.4
##  8 Pau Gasol                                    6.38
##  9 Russell Westbrook                            4.5
## 10 James Harden                                 4.47
```

Based on the tables above, it is seems like LeBron James is an outlier, both in regard to his annual salary and Twitter follower count. So, we will remove LeBron from the dataset and see how the model changes:

```
## Observations: 94
## Variables: 29
## $ PLAYER_NAME                     <chr> "Russell Westbrook", "Demetrius Jac...
## $ TEAM_ABBREVIATION               <chr> "OKC", "BOS", "NOP", "HOU", "GSW", ...
## $ AGE                             <dbl> 28, 22, 24, 27, 28, 32, 26, 22, 26,...
## $ W_PCT                           <dbl> 0.568, 0.200, 0.413, 0.667, 0.823, ...
## $ OFF_RATING                      <dbl> 107.9, 124.2, 104.2, 113.6, 117.2, ...
## $ DEF_RATING                      <dbl> 104.6, 117.8, 102.5, 107.3, 101.3, ...
## $ AST_RATIO                       <dbl> 23.4, 31.1, 7.3, 27.6, 18.4, 35.0, ...
## $ REB_PCT                         <dbl> 0.167, 0.103, 0.170, 0.123, 0.137, ...
## $ USG_PCT                         <dbl> 0.408, 0.172, 0.326, 0.341, 0.276, ...
## $ PIE                             <dbl> 0.230, 0.194, 0.192, 0.190, 0.186, ...
## $ SALARY_MILLIONS                 <dbl> 26.54, 1.45, 22.12, 26.50, 26.54, 2...
## $ ACTIVE_TWITTER_LAST_YEAR        <fct> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,...
## $ TWITTER_FOLLOWER_COUNT_MILLIONS <dbl> 4.500, 0.049, 1.220, 4.470, 16.200,...
## $ PTS                             <dbl> 31.6, 2.0, 28.0, 29.1, 25.1, 18.1, ...
## $ FGM                             <dbl> 824, 3, 770, 674, 551, 374, 647, 65...
## $ FGA                             <dbl> 1941, 4, 1526, 1533, 1026, 785, 143...
## $ ageCent                         <dbl> 0.6105263, -5.3894737, -3.3894737, ...
```

```
## $ ast_ratioCent                  <dbl> 6.274737, 13.974737, -9.825263, 10....
## $ off_ratingCent                 <dbl> -0.009473684, 16.290526316, -3.7094...
## $ def_ratingCent                 <dbl> -1.39368421, 11.80631579, -3.493684...
## $ fgaCent                        <dbl> 1155.7263158, -781.2736842, 740.726...
## $ fgmCent                        <dbl> 444.157895, -376.842105, 390.157895...
## $ PIECent                        <dbl> 0.09073684, 0.05473684, 0.05273684,...
## $ reb_pctCent                    <dbl> 0.033778947, -0.030221053, 0.036778...
## $ usg_pctCent                    <dbl> 0.170, -0.066, 0.088, 0.103, 0.038,...
## $ salary_millionsCent            <dbl> 15.2351368, -9.8548632, 10.8151368,...
## $ w_pctCent                      <dbl> 0.056589474, -0.311410526, -0.09841...
## $ ptsCent                        <dbl> 16.3176842, -13.2823158, 12.7176842...
## $ active_twitter_1year           <fct> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,...
```

Based on the above output, we can see LeBron has been removed from the dataset (94 observations remaining).

Now, to assess the impact of his absence on the final model:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.140	0.229	4.978	0.000	0.684	1.595
ageCent	0.137	0.067	2.050	0.043	0.004	0.270
ast_ratioCent	0.018	0.029	0.631	0.530	-0.039	0.075
usg_pctCent	11.353	6.627	1.713	0.090	-1.826	24.532
salary_millionsCent	0.095	0.037	2.579	0.012	0.022	0.168
w_pctCent	3.880	1.623	2.391	0.019	0.653	7.107
fgmCent	0.010	0.007	1.445	0.152	-0.004	0.024
fgaCent	-0.005	0.004	-1.539	0.128	-0.013	0.002
active_twitter_1year0	-2.743	1.524	-1.800	0.075	-5.772	0.287
salary_millionsCent:w_pctCent	0.492	0.184	2.671	0.009	0.126	0.858

```
## # A tibble: 15 x 2
##    PLAYER_NAME       USG_PCT
##    <chr>               <dbl>
##  1 Russell Westbrook   0.408
##  2 DeMarcus Cousins    0.364
##  3 Joel Embiid         0.363
##  4 DeMar DeRozan       0.342
##  5 James Harden        0.341
##  6 Isaiah Thomas       0.337
##  7 Anthony Davis       0.326
##  8 Kawhi Leonard       0.312
##  9 Damian Lillard      0.309
## 10 John Wall           0.302
## 11 Kyrie Irving        0.302
## 12 LeBron James        0.297
## 13 Dwyane Wade         0.293
## 14 Stephen Curry       0.292
## 15 Kemba Walker        0.291
```

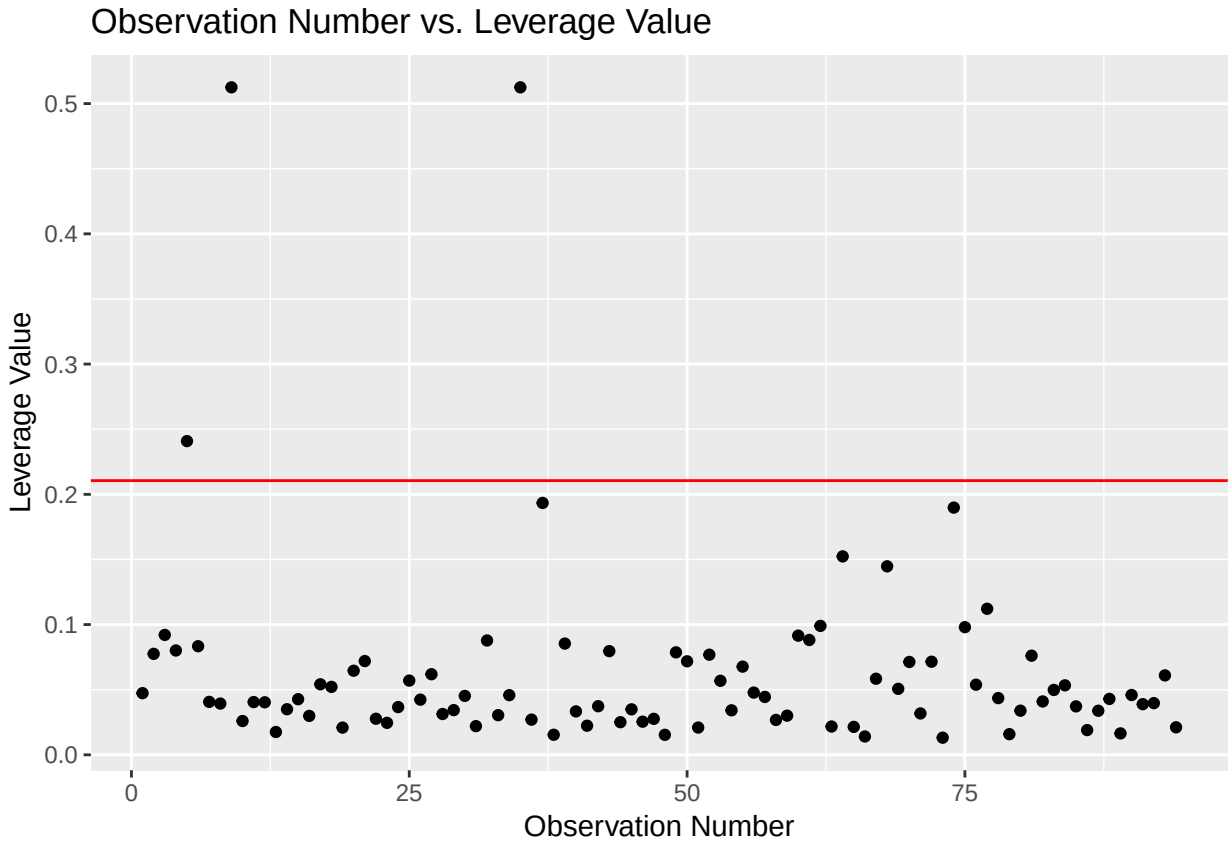
```
## # A tibble: 5 x 2
##   PLAYER_NAME          FGM
##   <chr>              <dbl>
## 1 Russell Westbrook    824
## 2 Karl-Anthony Towns   802
## 3 Anthony Davis        770
## 4 LeBron James         736
```

```
## 5 DeMar DeRozan        721
```

```
## # A tibble: 16 x 2
##    PLAYER_NAME          FGA
##    <chr>              <dbl>
##  1 Russell Westbrook   1941
##  2 DeMar DeRozan       1545
##  3 James Harden        1533
##  4 Anthony Davis       1526
##  5 Damian Lillard      1488
##  6 Karl-Anthony Towns  1480
##  7 Isaiah Thomas       1473
##  8 Kemba Walker        1449
##  9 Stephen Curry       1443
## 10 CJ McCollum         1441
## 11 John Wall           1435
## 12 DeMarcus Cousins    1432
## 13 Kyrie Irving        1420
## 14 Carmelo Anthony     1389
## 15 Paul George         1348
## 16 LeBron James        1344
```

Comparing the two models, we notice a relatively sizable discrepancy in the slope coefficient of mean-centered usage percentage. This makes sense since LeBron has the twelfth-highest usage percentage in the league (0.297), so eliminating him from the dataset dramatically affects the average usage percentage (as well as the spread, or standard deviation). We also see a discrepancy in the p-values for mean-centered field goals made and mean-centered field goals attempted. These predictors have become statistically insignificant without LeBron James. This also makes sense since LeBron made the fourth-most field goals (736) and attempted the sixteenth-most field goals (1344).

More generally, to determine whether prominent athletes are influential points, we will look at standardized residuals, leverage, and Cook's Distance:

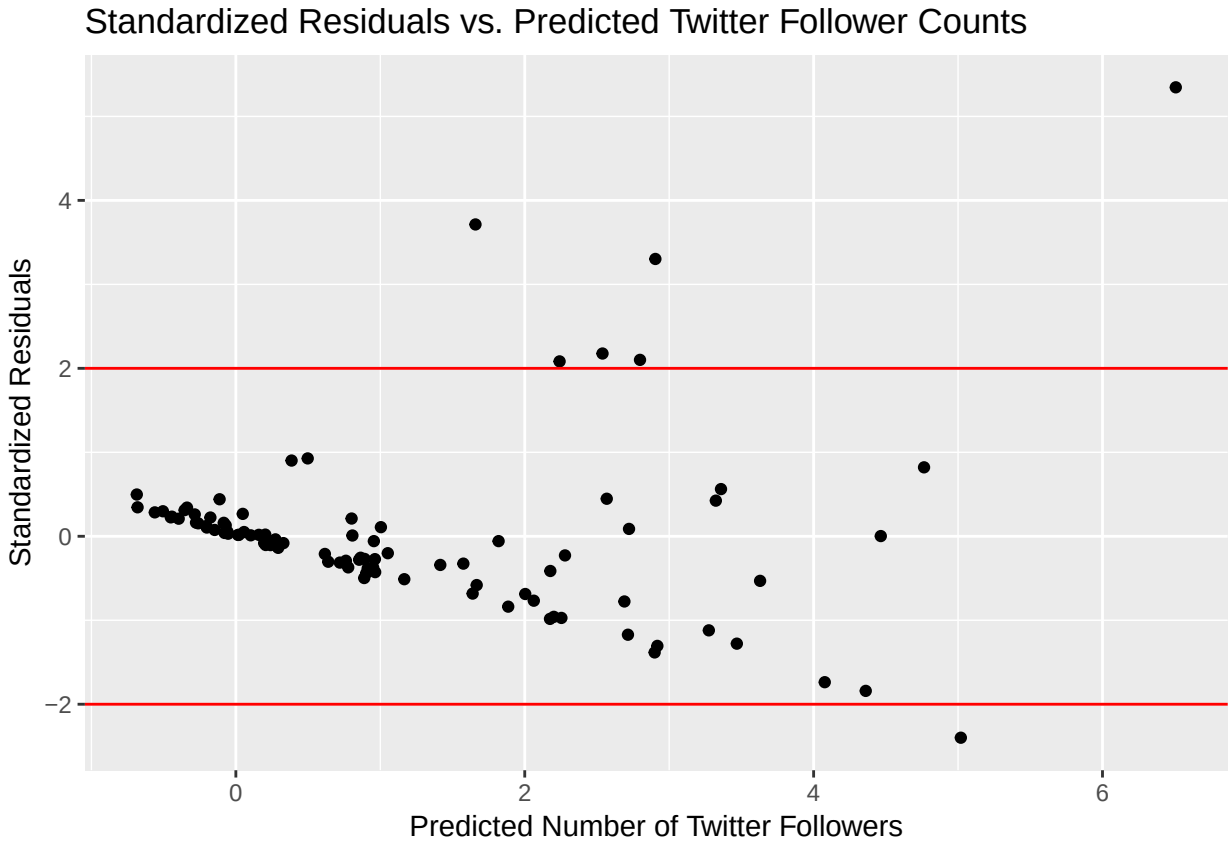



```
## # A tibble: 3 x 2
##   obs_num  .hat
##     <int> <dbl>
## 1       5 0.241
## 2       9 0.512
## 3      35 0.512
```

```
## # A tibble: 6 x 2
##   obs_num PLAYER_NAME
##     <int> <chr>
## 1       5 Kevin Durant
## 2       6 LeBron James
## 3      10 Kawhi Leonard
## 4      11 Joel Embiid
## 5      36 Greg Monroe
## 6      75 Jarnell Stokes
```

Based on the threshold (`2 * (p + 1) / n`), Kevin Durant, LeBron James, Kawhi Leonard, Joel Embiid, Greg Monroe, and Jarnell Stokes are considered high leverage players (and hence potential influential points).

Now, to identify outliers within these candidates, we will look at the standardized residuals:

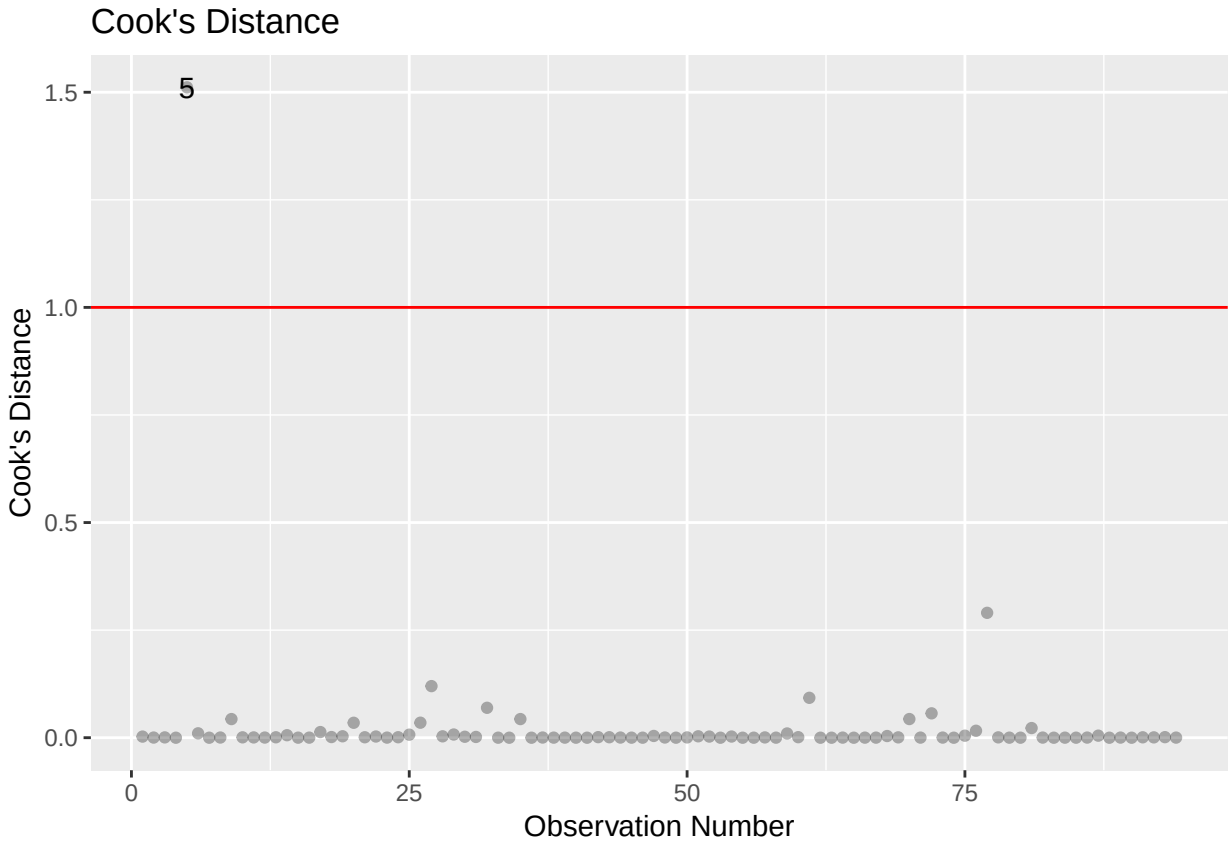



```
## # A tibble: 7 x 2
##   obs_num .std.resid
##     <int>      <dbl>
## 1       5       5.35
## 2      26       2.18
## 3      27       3.30
## 4      32       2.08
## 5      61      -2.40
## 6      72       2.10
## 7      77       3.71
```

```
## # A tibble: 3 x 2
##   obs_num PLAYER_NAME
##     <int> <chr>
## 1       6 LeBron James
## 2      30 DeAndre Jordan
## 3      78 Carmelo Anthony
```

The players with standardized residuals of magnitude greater than 2 are LeBron James, DeAndre Jordan, and Carmelo Anthony. Hence, LeBron, DeAndre, and Carmelo are outliers; however, it remains to be seen whether DeAndre and Carmelo impact the regression line (we already examined LeBron's effect).

To assess the impact of prominent athletes (identified by high leverage and/or high standardized residuals) on the regression line, we examine Cook's Distance:




```
## # A tibble: 1 x 2
##   obs_num PLAYER_NAME
##     <int> <chr>
## 1       6 LeBron James
```

It is clear from the plot of Cook's Distance vs. observation number that LeBron James is the only influential point.

Since our objective is to accurately predict the Twitter follower counts of NBA players, it is probably best to exclude LeBron to avoid overestimating for less prominent athletes. Hence, we will continue our analysis by using the multiple linear regression model without LeBron James.

Before finalizing this choice, we must consider the impact of LeBron's absence on the significance of certain predictors, namely mean-centered usage percentage, mean-centered field goals made, and mean-centered field goals attempted. We will perform backward selection on this new model (with adjusted slope coefficients due to the absence of LeBron James) with AIC as the selection criterion:

```
## Start:  AIC=147.43
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + ast_ratioCent + usg_pctCent +
##     salary_millionsCent + w_pctCent + fgmCent + fgaCent + active_twitter_lyear +
##     salary_millionsCent * w_pctCent
##
##                                Df Sum of Sq    RSS    AIC
## - ast_ratioCent                 1    1.7302 366.39 145.88
## <none>                                      364.65 147.43
## - fgmCent                       1    9.0591 373.71 147.74
## - fgaCent                       1   10.2794 374.93 148.04
## - usg_pctCent                   1   12.7393 377.39 148.66
## - active_twitter_lyear          1   14.0666 378.72 148.99
```

```
## - ageCent                          1   18.2468 382.90 150.02
## - salary_millionsCent:w_pctCent  1   30.9592 395.61 153.09
##
## Step:  AIC=145.88
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + usg_pctCent + salary_millionsCent +
##     w_pctCent + fgmCent + fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## - fgmCent                        1     7.330 373.72 145.74
## <none>                                       366.39 145.88
## - fgaCent                        1     8.609 374.99 146.06
## - usg_pctCent                    1    11.045 377.43 146.67
## - active_twitter_lyear           1    13.821 380.21 147.36
## - ageCent                        1    19.688 386.07 148.80
## - salary_millionsCent:w_pctCent  1    39.708 406.09 153.55
##
## Step:  AIC=145.74
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + usg_pctCent + salary_millionsCent +
##     w_pctCent + fgaCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## - fgaCent                        1     1.417 375.13 144.09
## - usg_pctCent                    1     6.659 380.37 145.40
## <none>                                       373.72 145.74
## - active_twitter_lyear           1    11.869 385.58 146.68
## - ageCent                        1    14.470 388.19 147.31
## - salary_millionsCent:w_pctCent  1    37.953 411.67 152.83
##
## Step:  AIC=144.09
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + usg_pctCent + salary_millionsCent +
##     w_pctCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## - usg_pctCent                    1     5.902 381.03 143.56
## <none>                                       375.13 144.09
## - active_twitter_lyear           1    11.398 386.53 144.91
## - ageCent                        1    18.264 393.40 146.56
## - salary_millionsCent:w_pctCent  1    41.350 416.48 151.92
##
## Step:  AIC=143.56
## TWITTER_FOLLOWER_COUNT_MILLIONS ~ ageCent + salary_millionsCent +
##     w_pctCent + active_twitter_lyear + salary_millionsCent:w_pctCent
##
##                                 Df Sum of Sq    RSS    AIC
## <none>                                       381.03 143.56
## - active_twitter_lyear           1    10.778 391.81 144.18
## - ageCent                        1    14.564 395.60 145.09
## - salary_millionsCent:w_pctCent  1    38.836 419.87 150.69
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.109	0.226	4.904	0.000	0.660	1.559
ageCent	0.109	0.059	1.834	0.070	-0.009	0.227
salary_millionsCent	0.097	0.027	3.552	0.001	0.043	0.152
w_pctCent	4.533	1.550	2.924	0.004	1.452	7.614

term	estimate	std.error	statistic	p.value	conf.low	conf.high
active_twitter_lyear0	-2.379	1.508	-1.578	0.118	-5.376	0.618
salary_millionsCent:w_pctCent	0.513	0.171	2.995	0.004	0.173	0.853

As we can see from the above output, mean-centered assists-to-turnovers ratio, mean-centered usage percentage, mean-centered field goals made, and mean-centered field goals attempted were removed. However, mean-centered age and whether the player had an active Twitter account in 2015-2016 – insignificant predictors (based on p-values and confidence intervals) – remain in the model. To determine whether these predictors should be removed, we compare the AIC and adjusted R-squared values for the final model selected based on AIC as the selection criterion (without LeBron) and the same model without mean-centered age and whether the player had an active Twitter account in 2015-2016:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.087	0.229	4.742	0.000	0.632	1.543
salary_millionsCent	0.109	0.027	4.035	0.000	0.055	0.162
w_pctCent	4.609	1.568	2.940	0.004	1.495	7.724
salary_millionsCent:w_pctCent	0.431	0.171	2.513	0.014	0.090	0.771

```
## [1] 412.3224
```

```
## [1] 0.3178268
```

```
## [1] 414.8326
```

```
## [1] 0.2851532
```

Based on the AIC and adjusted R-squared values displayed above, the model with all terms minimized AIC (412.32) and maximized adjusted R-squared (0.318). Therefore, we proceed with the model which includes mean-centered salary, mean-centered win percentage, mean-centered age, whether the player had an active Twitter account in 2015-2016, and the interaction between mean-centered salary and mean-centered win percentage.

As we can see from the dotplots above, both field goals made and field goals attempted appear to have weak positive relationships with Twitter follower counts.

In sum, there appear to be weak to moderate positive correlations between Twitter follower count and win percentage, usage percentage, player impact factor, avg. points per game, field goals made, field goals attempted, and player salaries; there is no obvious relationship between Twitter follower count and assists-to-turnovers ratio or rebound percentage.

There may have been some slight linearity issues with our current model. Though there is not a discernible pattern in our residuals plot for salary, there was slight curvature, so we were interested to see if we could fix this with a log-transformation:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.702	0.446	-1.576	0.123	-1.602	0.197
log(salary_millionsCent)	0.260	0.229	1.139	0.261	-0.201	0.721
w_pctCent	3.590	2.851	1.259	0.215	-2.163	9.344
log(salary_millionsCent):w_pctCent	-0.646	1.544	-0.418	0.678	-3.762	2.471

Our p-values are extremely high for all coefficients, which means that predicted variables are not significant predictors for the response variable. Therefore, we will not continue with this model, because our final model from before is significant and satisfies assumptions fairly well.