Final Writeup

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

2020-04-25

Section 1: Introduction

use patchwork for plots remember no variable names in narrative (use actual terms) added FGM and FGA (field goals made/attempted) removed NETRATING

Complete EDA:

The plots that were included in the EDA in the main body of our analysis are NOT included again here, only the ones that were created but we did not decide to include above.

Univariate

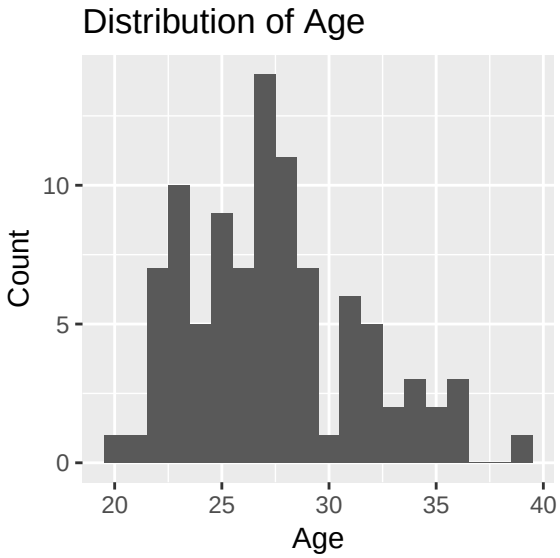
First, we will do univariate EDA on the dataset. Player name will be used to refer to observations in our dataset, but since each player name is distinct we do not need to do EDA on the `PLAYER_NAME` variable.

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 the `AGE` variable in the dataset:

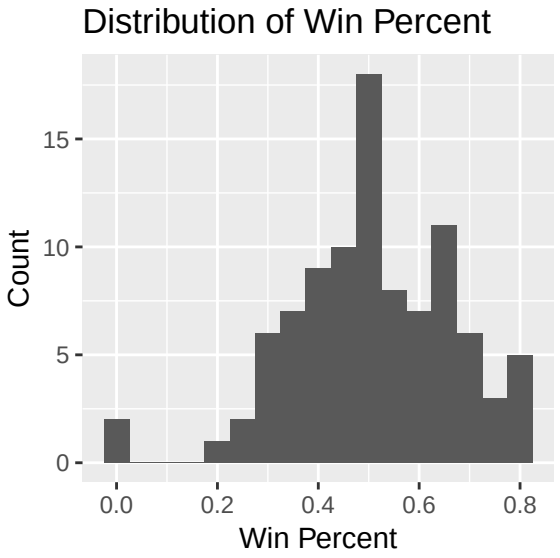



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  27.4    20  24.5     27    29    39
```

```
## # A tibble: 1 x 16
##   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 9 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>
```

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.

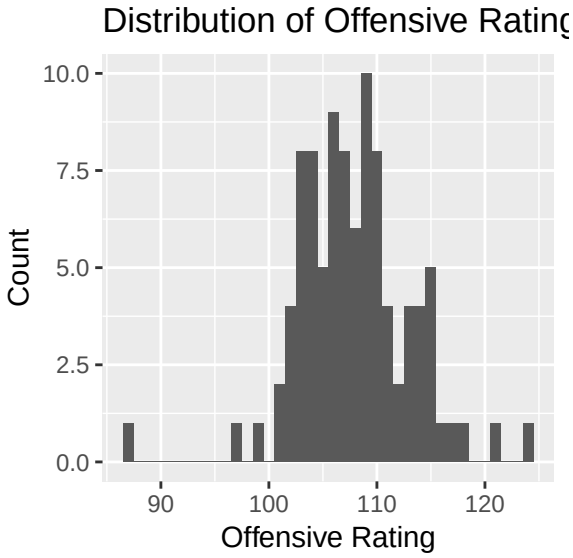
Now, we'll examine the distribution of win percent, `W_PCT`:



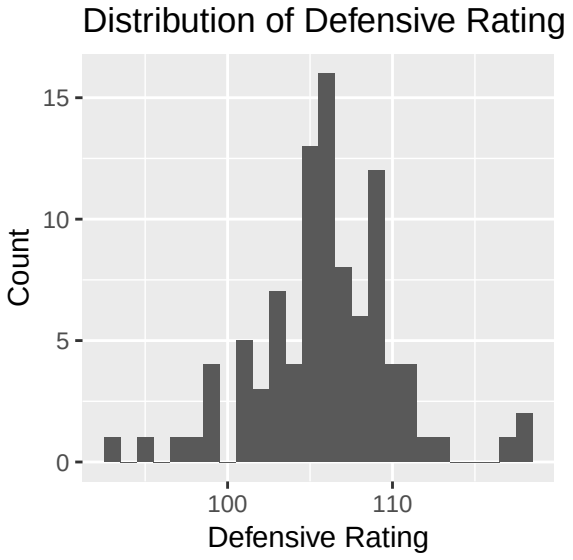
```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.511     0 0.418  0.507  0.63 0.824
```

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, `OFF_RATING` and defensive rating `DEF_RATING`:




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

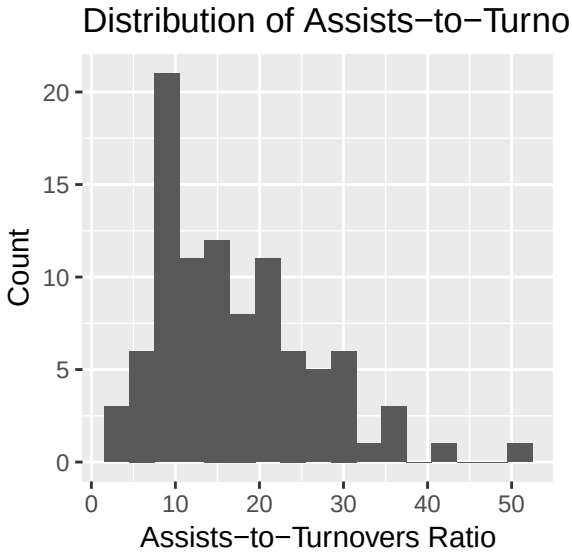



```
## # A tibble: 1 x 3
##     min median   max
```

```
##   <dbl>  <dbl> <dbl>
## 1    93    106  118.
```

Defensive rating, offensive rating, and net rating do not stray far from 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. The distribution of net rating has multiple nearly equal modes around -2 to -3 and around 1 and 3. The median net rating is 1.5, and the net ratings in the dataset vary from -17.2 to 18.7.

Next, we'll look at the distribution of the assist-to-turnovers ratio, `AST_RATIO`:

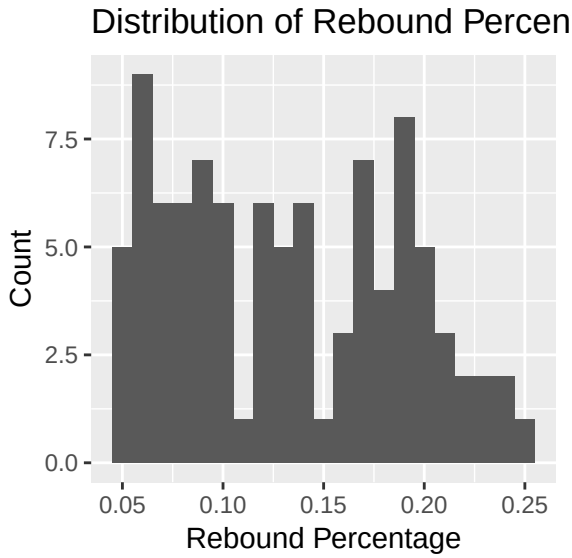



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  17.1     4  9.75     15  22.2  51.5
```

```
## # A tibble: 2 x 16
##   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 9 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>
```

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.

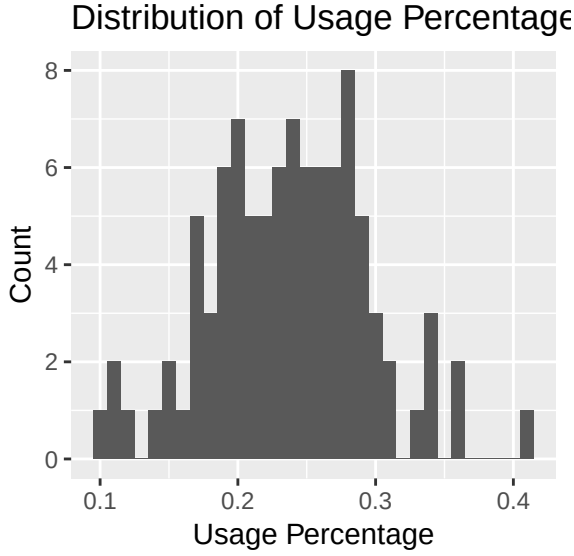
Next, we'll examine the variation of `REB_PCT`, the percent of rebounds a player makes:




```
## # 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 .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 `USG_PCT`, usage percentage, which is an estimate of how often a player makes team plays:

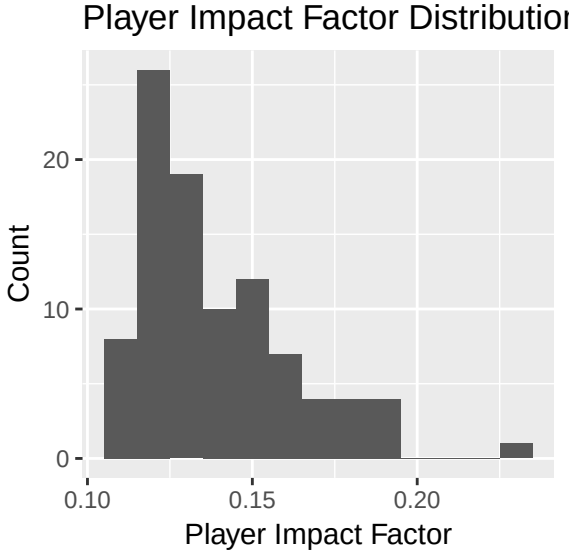



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.238 0.101   0.2  0.242 0.276 0.408
```

```
## # A tibble: 1 x 16
##   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 9 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>
```

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 `PIE`, player impact factor, a statistic roughly measuring a player's impact on the games that they play that's used by nba.com:



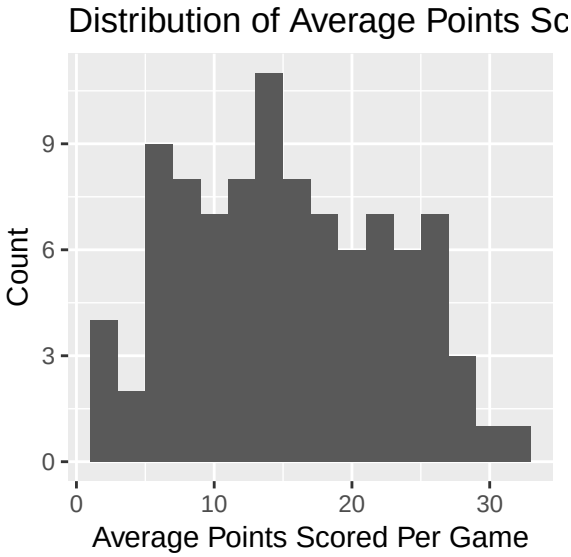

```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.139 0.112 0.122  0.131 0.152  0.23
```

```
## # 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 the variable `ACTIVE_TWITTER_LAST_YEAR`, which tells us whether or not each player posted on Twitter the year before the data was collected:

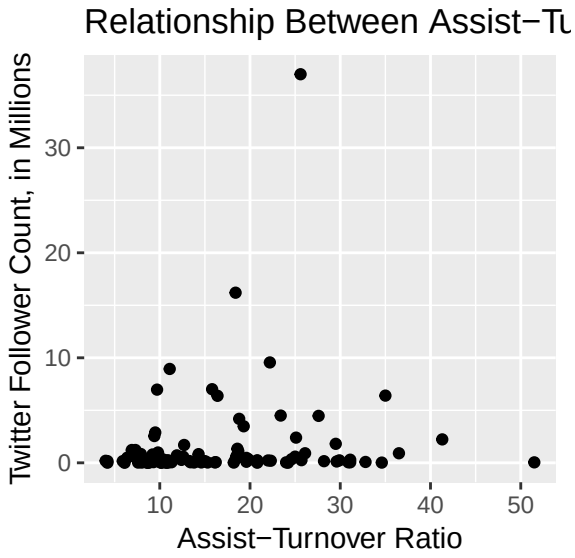
```
## # A tibble: 2 x 2
##   ACTIVE_TWITTER_LAST_YEAR     n
##   <fct>                    <int>
## 1 0                            2
## 2 1                           93
```

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.

Bivariate

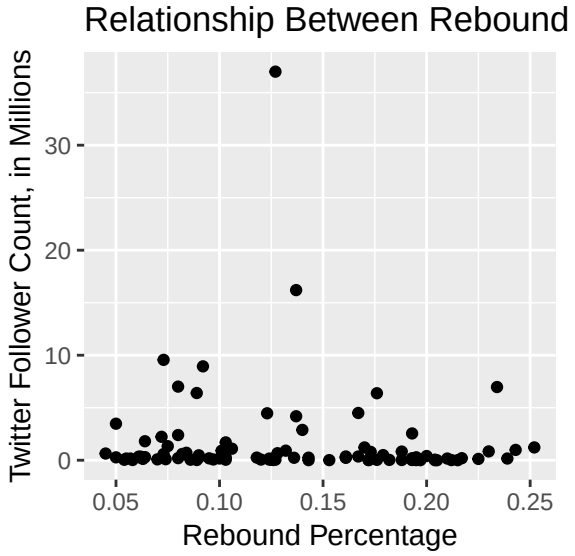
Next, we will do bivariate EDA, looking into the relationships of some of the predictor variables with the response variables. We won't do bivariate EDA on player name, Twitter handle, age, team abbreviation, or whether the players were active on Twitter last year, instead focusing on terms we believe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

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.

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 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.

Final Model and Model Fit Statistics

Thus, our final model is:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.270	1.495	-0.849	0.398	-4.242	1.702
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_lyear1	2.379	1.508	1.578	0.118	-0.618	5.376
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
```

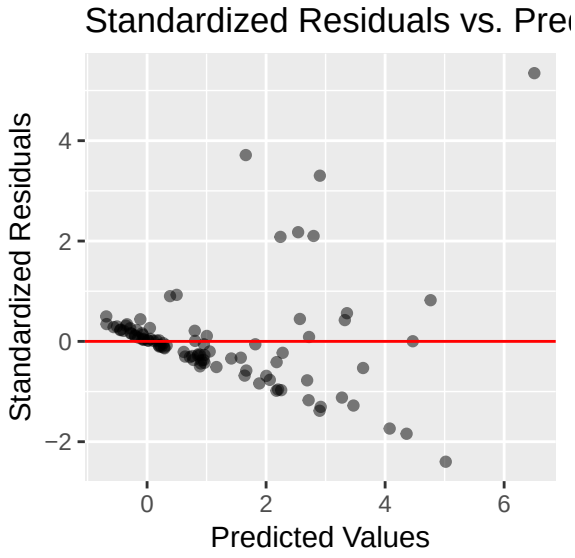
```
## [1] 0.3178268
```

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 variation in the number of Twitter followers (in millions).

Discussion of Assumptions

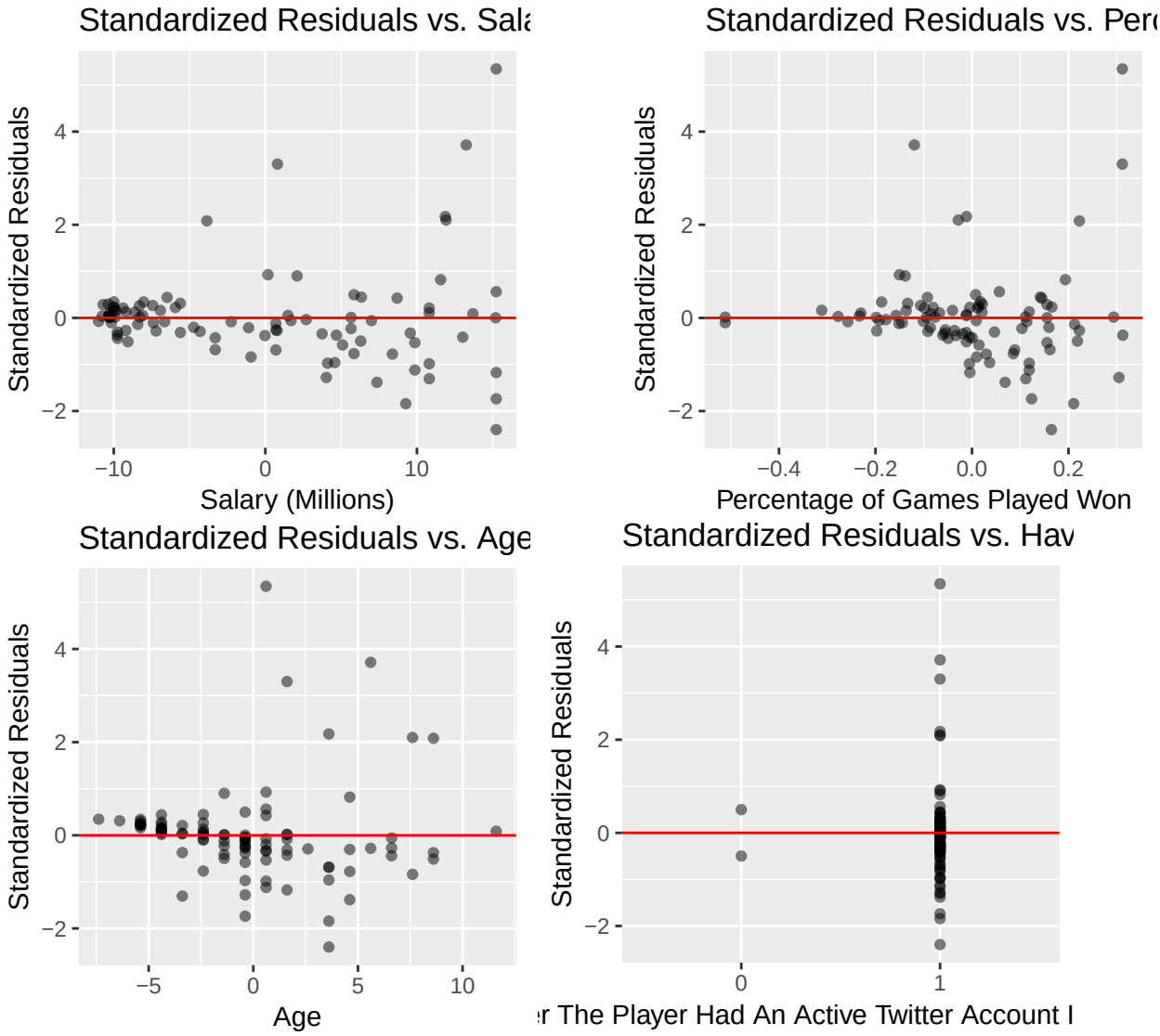
Next, we checked the linearity, constant variance, normality, and independence assumptions for multiple linear regression.

First, we checked whether the response, the number of Twitter followers (in millions), had a linear relationship with the predictor variables in the model by plotting the standardized residuals vs. predicted values:



The height of the cloud of points varies as you move from left to right. Points are clustered at the very left, but they are sparser as you move along the graph. Therefore, constant variance is not satisfied.

There does not seem to be a systematic pattern in the plot of the standardized residuals vs. predicted values. Hence, this plot presents no issues with the linearity assumption.

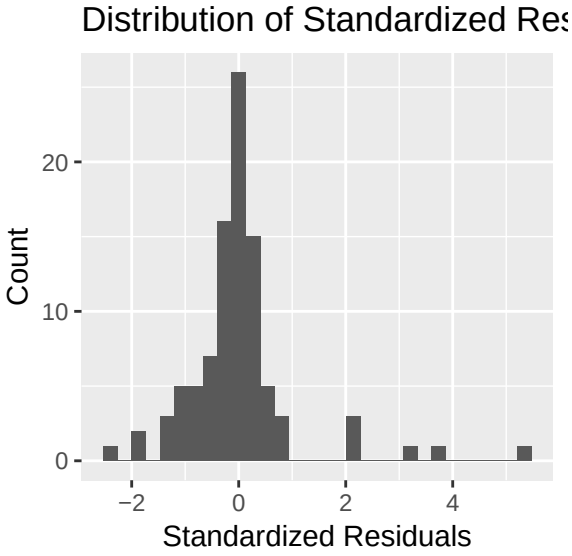


There looks to be a pattern in the plot because towards the end it curves upward drastically. Therefore, this pattern suggests that interactions or higher-order terms (like quadratic terms) are required. Thus, linearity is not satisfied.

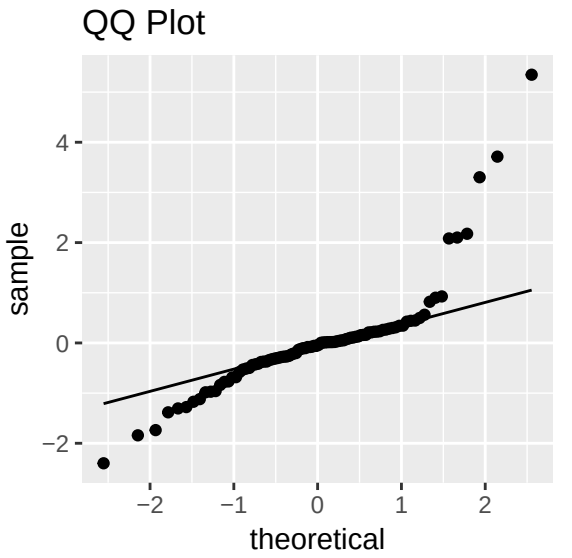
There looks to be a pattern in the plot because towards the end it also curves upward drastically. Therefore, this pattern suggests that interactions or higher-order terms (like quadratic terms) are required. Linearity is not satisfied.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

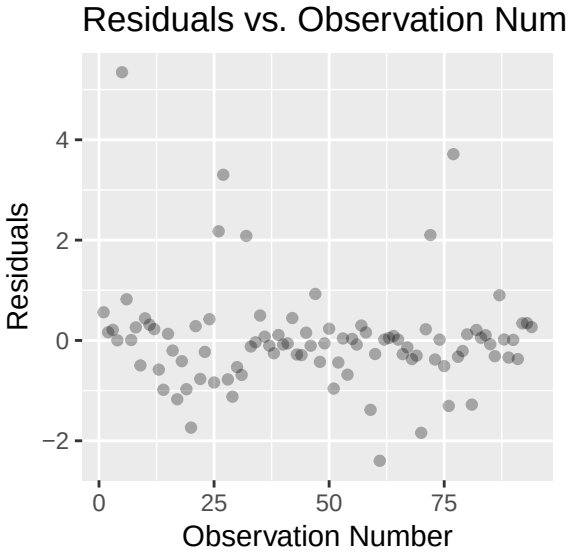


Normality is violated because the histogram of residuals is heavily right skewed because the height of the graphs decreases dramatically from right to left.



However, normality is not satisfied because some of the points don't follow the diagonal line.

Lastly, we will check for independence. Data was not taken over time, so we know there is no temporal correlation. There is also no spatial correlation because data was not taken in space. We can check to see if there is some structure/order to the dataset according to observation number. There is no purposeful order to how the dataset was collected according to Kaggle, but we will check just in case:



Looking at the graph, there is no distinguishable pattern in the graph. The number of Twitter followers of one player will not affect the number of Twitter followers of another player. Therefore, independence is satisfied.

Therefore, because constant variance and normality and linearity are not satisfied, but independence is, we will make some adjustments to our model. We can adjust for linearity by adding squared terms for salary and percentage of games played won because those standardized residuals vs. predictor variable plots violated linearity. For constant variance, we can log-transform the response variable because our standardized residuals vs. predicted values violated constant variance. Lastly, for normality regression there were outliers observed in the histograms, so we can try to remove all 7 outlier players from our model as mentioned previously above because Kevin Durant, LeBron James, Josh Huestis, JaVale McGee, David West, Jarnell Stokes, and Carmelo Anthony are considered high leverage players (and hence potential influential points).

After making these adjustments our model looks like this:

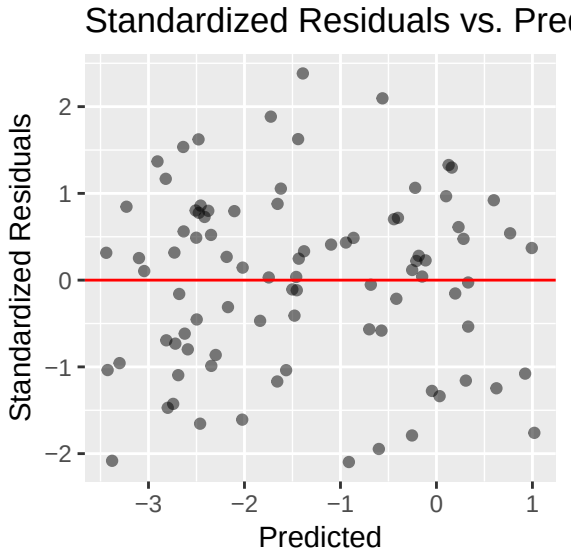
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.169	0.279	-4.188	0.000	-1.725	-0.614
salary_millionsCent	0.148	0.020	7.532	0.000	0.109	0.187
I(salary_millionsCent * salary_millionsCent)	-0.004	0.003	-1.592	0.115	-0.010	0.001
w_pctCent	2.081	1.218	1.708	0.091	-0.342	4.505
I(w_pctCent * w_pctCent)	6.565	7.571	0.867	0.388	-8.496	21.625
salary_millionsCent:w_pctCent	-0.109	0.173	-0.629	0.531	-0.454	0.236

However, because the interaction effect and `w_pctCent` squared terms and the `salary_millionsCent` have p-values that are very large and > 0.05, we will remove them from our model to produce this model. We will attempt to see if the log transformation of our response variable is enough to support our assumptions.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.398	0.149	-9.370	0.000	-1.694	-1.101
salary_millionsCent	0.133	0.018	7.250	0.000	0.097	0.170
w_pctCent	2.336	1.152	2.029	0.046	0.047	4.626

We will check for all assumptions again:

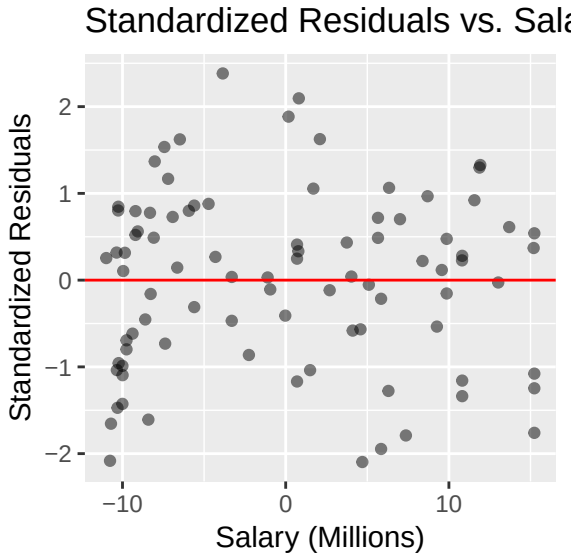
First, we will check for linearity and whether the response variable has a linear relationship with the predictor variables in the model. We will check the plot of standardized residuals vs. predicted values:



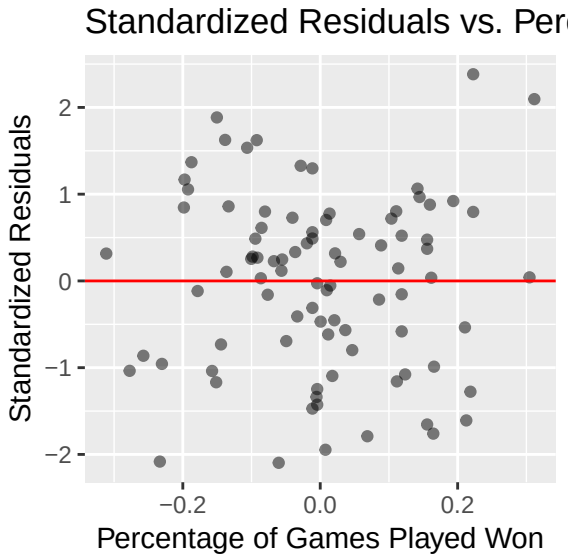
When observing for constant variance, the height of the cloud of points seems to be constant as you move from left to right. Therefore, constant variance is satisfied.

There is no obvious pattern and the shape of the graph seems to be linear. Hence, this plot presents no issues with the linearity assumption.

Next, we will observe the plot of residuals vs. predictors:



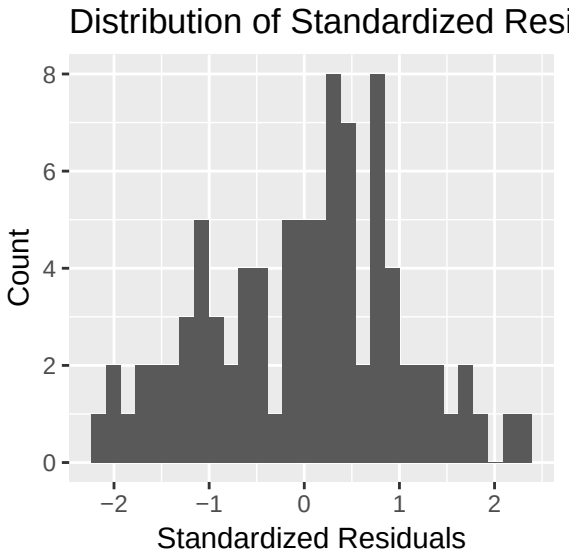
There is no distinguishable pattern in our plot because there are no discernible curves. Therefore, we satisfy linearity.



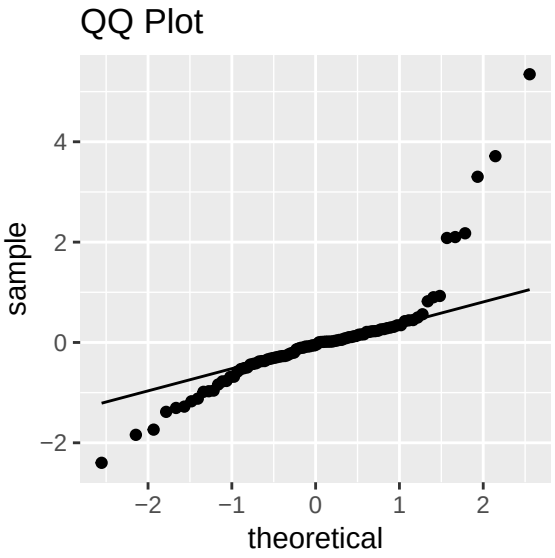
There looks to be no distinguishable pattern in the plot as well, so therefore linearity is supported because there have been previous violations of linearity.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

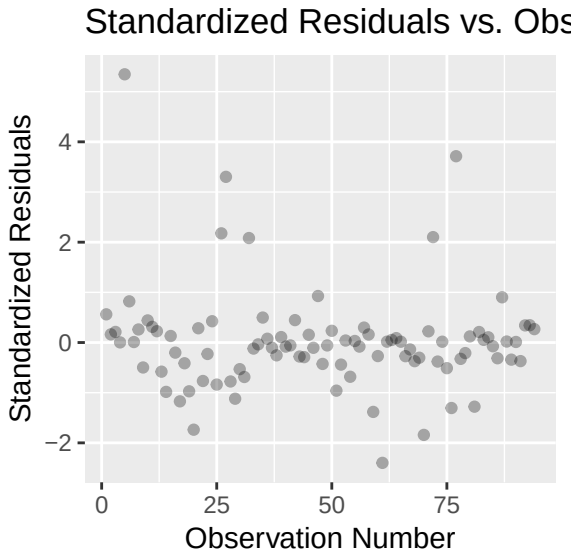


Normality is not supported because the histogram of residuals is not normal and is now slightly left skewed because the height increases as you move across the graph, then drastically decreases, but the distribution is unimodal.



However, normality is not satisfied because some of the points don't follow the diagonal line.

Lastly, we will check for independence. Data was not taken over time, so we know there is no temporal correlation. There is also no spatial correlation because data was not taken in space. We can check to see if there is some structure/order to the dataset according to observation number. There is no purposeful order the data was collected in according to Kaggle, but we will check just in case:



Looking at the graph, there is no distinguishable pattern in the graph. The number of Twitter followers of one player will not affect the number of Twitter followers of another player. Therefore, independence is satisfied.

Linearity, constant variance, and normality are all satisfied in this model. We have done everything we can to the best of our knowledge to fix normality, so we will continue on with this model for interpretations. Our final model output is:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.398	0.149	-9.370	0.000	-1.694	-1.101
salary_millionsCent	0.133	0.018	7.250	0.000	0.097	0.170
w_pctCent	2.336	1.152	2.029	0.046	0.047	4.626

`log(TWITTER_FOLLOWER_COUNT_MILLIONS)`-hat = -1.398 + 0.133 * `salary_millionsCent` + 2.336 * `w_pctCent`.

The p-values for each of these predictor variables are all < 0.05, so we know each of the predictor variables is a significant predictor for our response variable.

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: 2 x 2
##   names                   x
##   <chr>               <dbl>
## 1 salary_millionsCent  1.09
## 2 w_pctCent            1.09
```

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

Section 3: Discussion

Interpretations

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.398	0.149	-9.370	0.000	-1.694	-1.101
salary_millionsCent	0.133	0.018	7.250	0.000	0.097	0.170
w_pctCent	2.336	1.152	2.029	0.046	0.047	4.626

```
## # A tibble: 1 x 2
##   mean_salary mean_winpct
##         <dbl>       <dbl>
## 1        11.3       0.511
```

Again, the final model is displayed above. Its equation is: `log(TWITTER_FOLLOWER_COUNT_MILLIONS)`-hat = -1.398 + 0.133 * `salary_millionsCent` + 2.336 * `w_pctCent`.

The intercept of the model is -1.398, which means that a player with a salary of 11.3 million dollars (average salary) and win percentage of 0.511 (average win percentage) are expected to have exp(-1.398) million, or approximately 247,090.7, Twitter followers.

The coefficient of `salary_millionsCent` is 0.133, which means that, for every 1 million increase in player salary, a player's Twitter followers are expected to multiply by a factor of exp(0.133) = 1.142, holding all else constant.

The coefficient of `w_pctCent` is 2.336. This coefficient is misleading, as an increase of 1 in win percent is only possible if a team has 0 wins. So, this coefficient is better explained by every 0.1 increase in team win percentage, a player's Twitter followers are expected to multiply by a factor of exp(0.2336) = 1.263, holding all else constant.

Interpreting coefficients is not particularly important, however, as our objective is to predict Twitter followers.

Predictions

We have carefully selected 3 players from the 2016-2017 season who are not included in the dataset, and will test the 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 percent data was calculated from team data (stats.nba.com). The number of Twitter followers was collected from various articles written in 2017.

Derrick Rose:

`salary_millionsCent` = 21.3 - 11.3 = 10 `w_pctCent` = 0.406 - 0.511 = -0.105

-1.398 + 0.133 * (10) + 2.336 * (-0.105) = -0.313

Derrick Rose's Twitter followers are expected to be exp(-0.313) = 731,249.9 followers. He actually had 2.49 million Twitter followers, so this is an extreme underprediction.

Wesley Matthews:

`salary_millionsCent` = 17.1 - 11.3 = 5.8 `w_pctCent` = 0.397 - 0.511 = -0.114

-1.398 + 0.133 * (5.8) + 2.336 * (-0.114) = -0.893

Wesley Matthews's Twitter followers are expected to be exp(-0.893) = 409,425.6 Twitter followers. He actually had 241,000 Twitter followers, so this is another overprediction. This might be attributed to his high salary, however, so the model overpredicts for that reason.

Boris Diaw:

`salary_millionsCent` = 7 - 11.3 = -4.3 `w_pctCent` = 0.616 - 0.511 = 0.105

-1.398 + 0.133 * (-4.3) + 2.336 * (0.105) = -1.725

Boris Diaw's Twitter followers are expected to be exp(-1.725) = 178,173.1 followers. He actually had 462,000 Twitter followers (gross underestimate). This time, his salary contributed to the underprediction.

The model has poor preditive capacity. The most likely reason this is the case is due to the large number of players on a basketball team (15) and the few stars in the NBA who have large amounts of Twitter followers (in the millions). We would like some feedback about how to improve this by possibly adding more significant terms.

It is important to note that the dates at which these Twitter follow 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 these articles came from during the season as well.

Section 4: Limitations

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. Because we removed LeBron, we will use standardized residuals when checking assumptions.

All in all, we completed the analysis with 94 observations, which is a sufficiently large sample population to build a model which can be generalized to a larger population of NBA athletes.

Section 5: Conclusion

Section 6: Additional Work

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)	-3.270	2.620	-1.248	0.216	-8.486	1.946
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
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_lyear1	4.384	2.629	1.668	0.099	-0.848	9.617
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)	-3.068	2.510	-1.222	0.225	-8.059	1.923

term	estimate	std.error	statistic	p.value	conf.low	conf.high
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_lyear1	4.378	2.532	1.729	0.087	-0.657	9.413
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
##                    -2.69513922                     0.30459234
##                  ast_ratioCent                     usg_pctCent
##                     0.08454854                    26.69292449
##            salary_millionsCent                         fgmCent
##                     0.15051918                     0.04677193
##                        fgaCent          active_twitter_lyear1
##                    -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_lyear            <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)	-1.603	1.509	-1.062	0.291	-4.604	1.398
ageCent	0.137	0.067	2.050	0.043	0.004	0.270

term	estimate	std.error	statistic	p.value	conf.low	conf.high
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_lyear1	2.743	1.524	1.800	0.075	-0.287	5.772
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
```

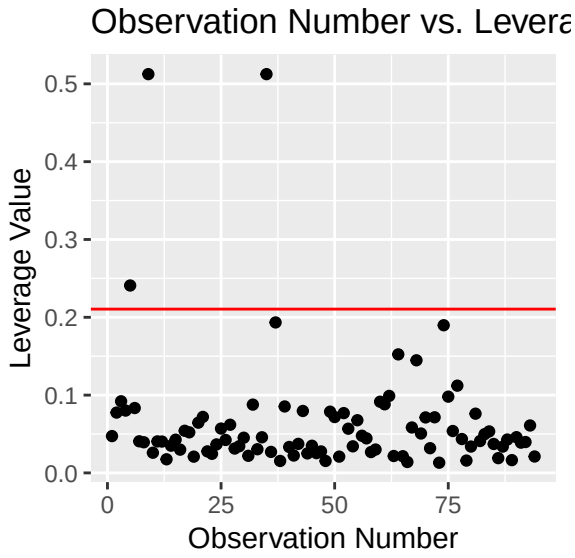
```
## # A tibble: 5 x 2
##   PLAYER_NAME          FGM
##   <chr>              <dbl>
## 1 Russell Westbrook    824
## 2 Karl-Anthony Towns   802
## 3 Anthony Davis        770
## 4 LeBron James         736
## 5 DeMar DeRozan        721
```

```
## # 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:

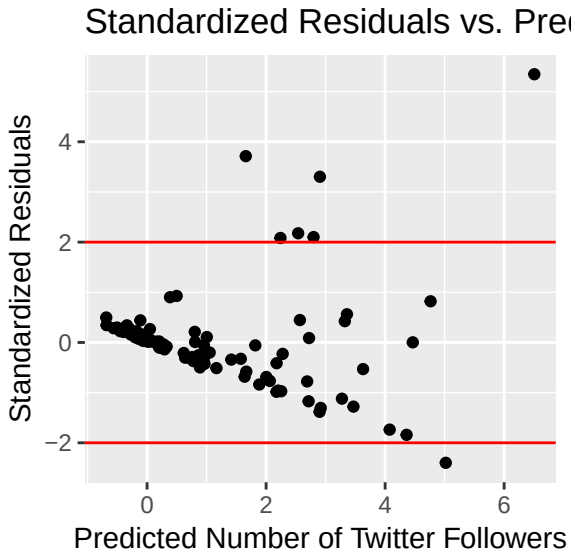



```
## # A tibble: 3 x 2
##   obs_num  .hat
##     <int> <dbl>
## 1       5 0.241
## 2       9 0.512
## 3      35 0.512
```

```
## # 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:

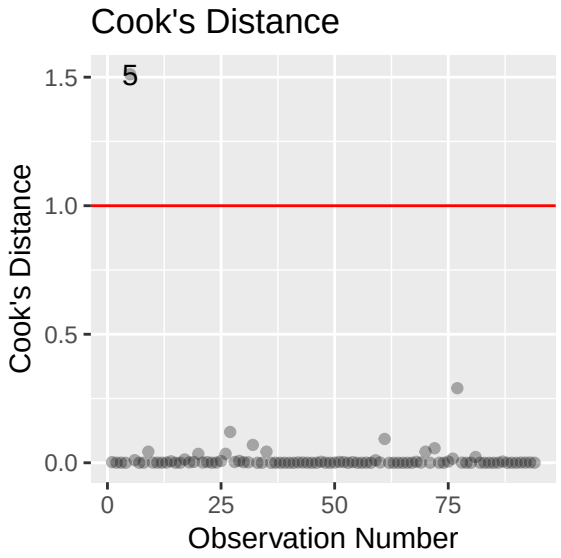



```
## # A tibble: 7 x 2
##   obs_num .std.resid
##     <int>      <dbl>
## 1       5       5.35
## 2      26       2.18
## 3      27       3.30
## 4      32       2.08
## 5      61      -2.40
## 6      72       2.10
## 7      77       3.71
```

```
## # 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)	-1.270	1.495	-0.849	0.398	-4.242	1.702
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_lyear1	2.379	1.508	1.578	0.118	-0.618	5.376
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.

Complete EDA:

The plots that were included in the EDA in the main body of our analysis are NOT included again here, only the ones that were created but we did not decide to include above.

Univariate

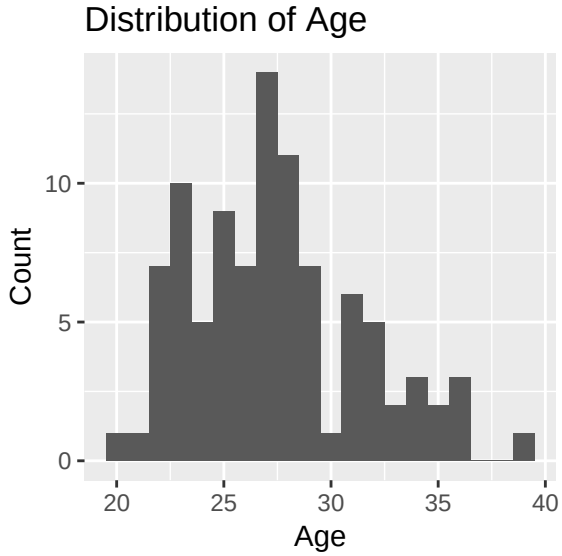
First, we will do univariate EDA on the dataset. Player name will be used to refer to observations in our dataset, but since each player name is distinct we do not need to do EDA on the `PLAYER_NAME` variable.

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 the `AGE` variable in the dataset:

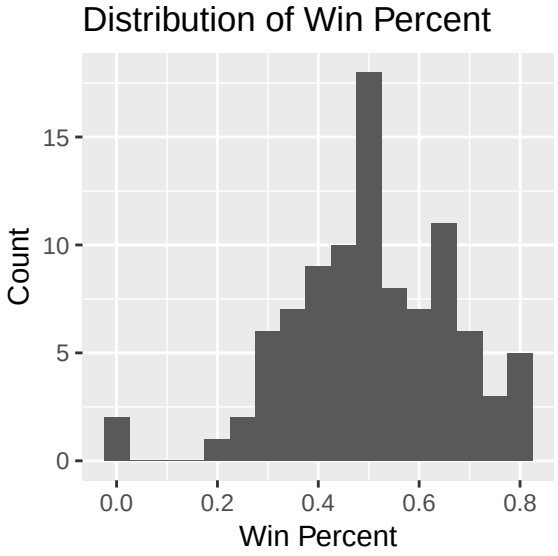



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  27.4    20  24.5     27    29    39
```

```
## # 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 <fct>,
## #   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.

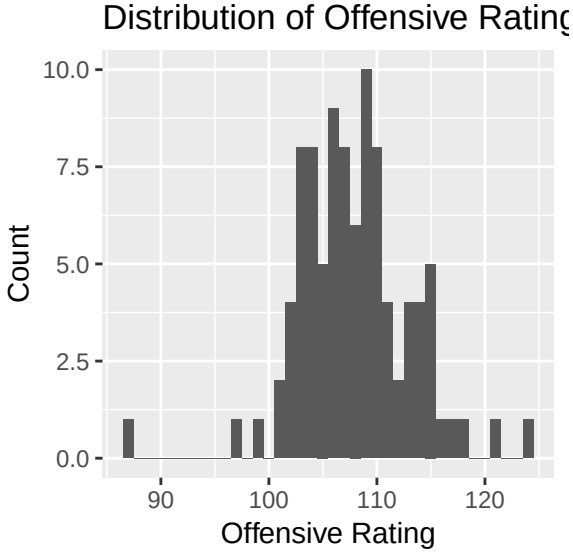
Now, we'll examine the distribution of win percent, `W_PCT`:



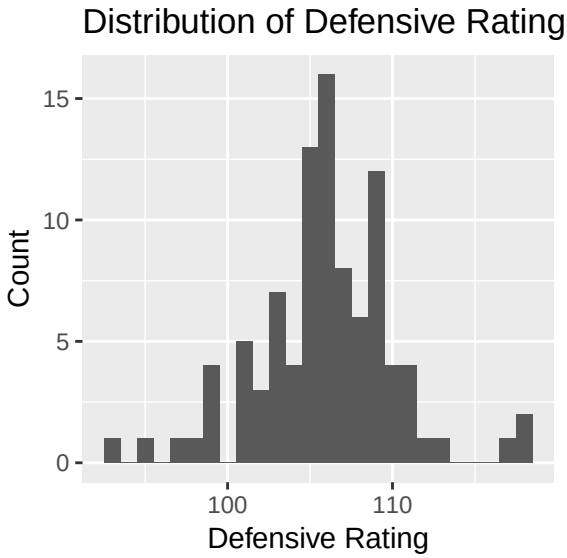

```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.511     0 0.418  0.507  0.63 0.824
```

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, `OFF_RATING` and defensive rating `DEF_RATING`:



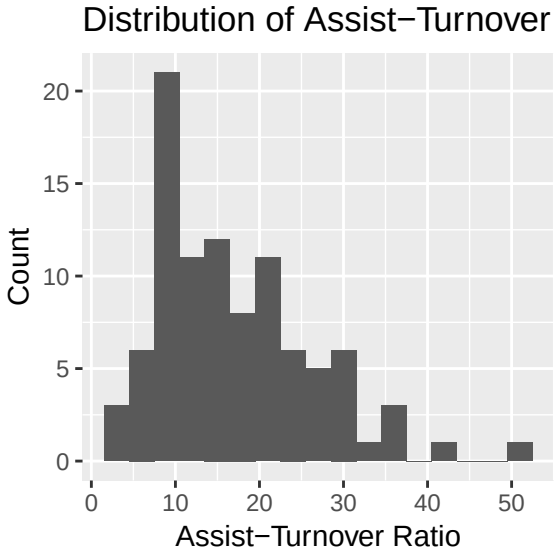

```
## # A tibble: 1 x 3
##     min median   max
##   <dbl>  <dbl> <dbl>
## 1  86.8   108.  124.
```




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

Defensive rating, offensive rating, and net rating do not stray far from 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. The distribution of net rating has multiple nearly equal modes around -2 to -3 and around 1 and 3. The median net rating is 1.5, and the net ratings in the dataset vary from -17.2 to 18.7.

Next, we'll look at the distribution of the assist-to-turnovers ratio, `AST_RATIO`:



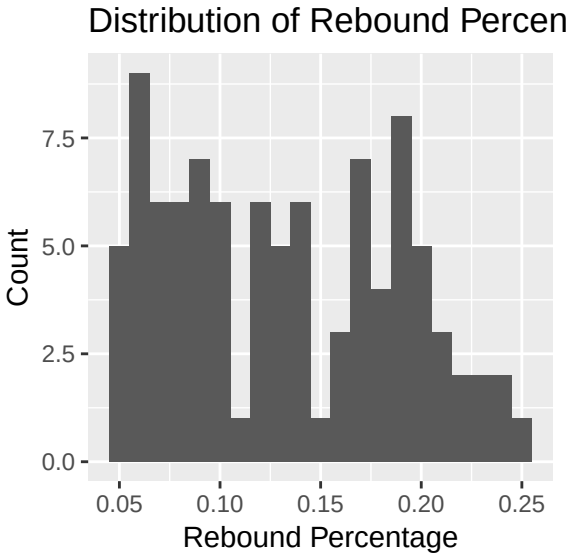

```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1  17.1     4  9.75     15  22.2  51.5
```

```
## # 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 <fct>,
## #   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 assist-turnover 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 assist-turnover 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 assist-turnover ratios, 51.5 and 41.3, are Jarnell Stokes and Ricky Rubio, respectively.

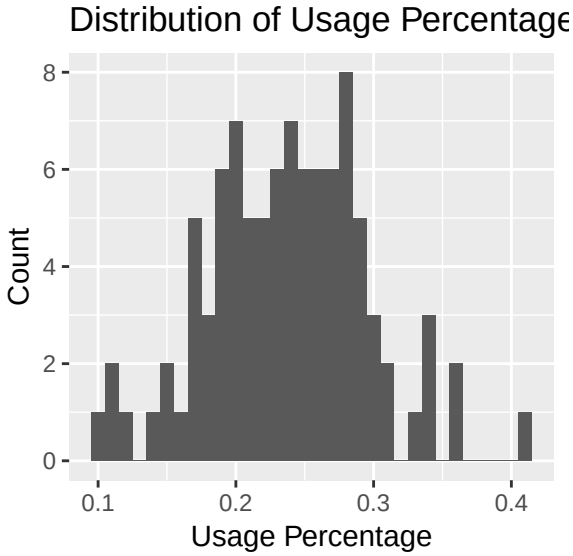
Next, we'll examine the variation of `REB_PCT`, the percent of rebounds a player makes:




```
## # 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 .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 `USG_PCT`, usage percentage, which is an estimate of how often a player makes team plays:

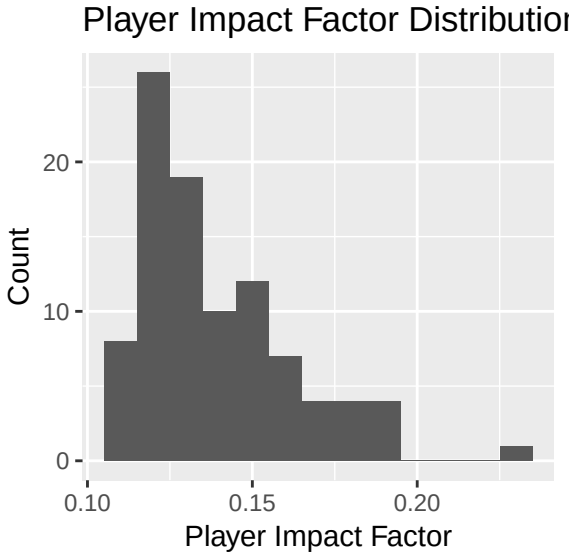



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.238 0.101   0.2  0.242 0.276 0.408
```

```
## # 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 <fct>,
## #   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 `PIE`, player impact factor, a statistic roughly measuring a player's impact on the games that they play that's used by nba.com:

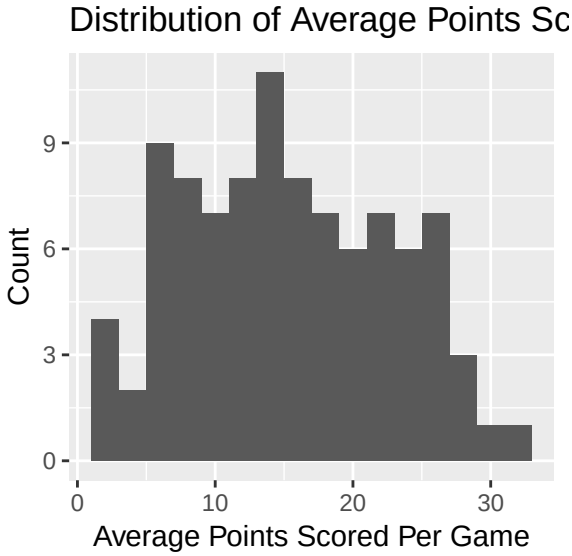



```
## # A tibble: 1 x 6
##    mean   min    Q1 median    Q3   max
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 0.139 0.112 0.122  0.131 0.152  0.23
```

```
## # 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 the variable `ACTIVE_TWITTER_LAST_YEAR`, which tells us whether or not each player posted on Twitter the year before the data was collected:

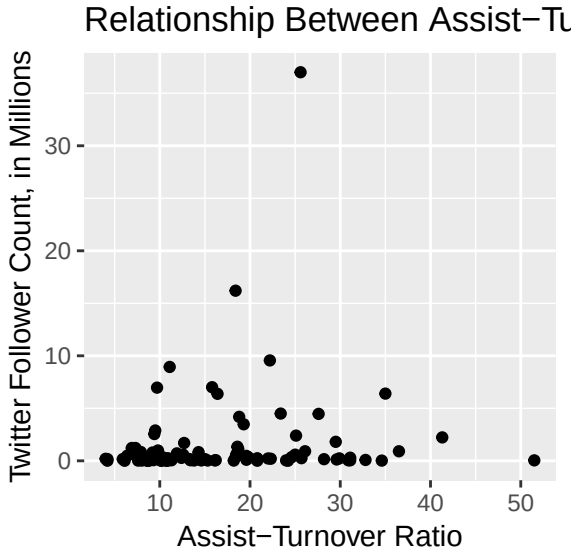
```
## # A tibble: 2 x 2
##   ACTIVE_TWITTER_LAST_YEAR     n
##   <fct>                    <int>
## 1 0                            2
## 2 1                           93
```

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.

Bivariate

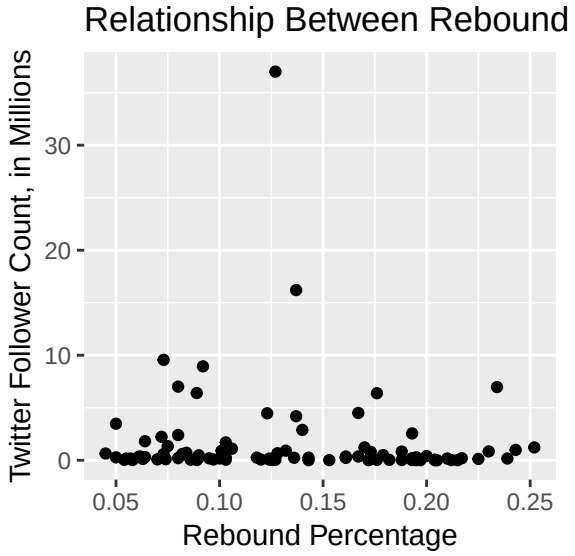
Next, we will do bivariate EDA, looking into the relationships of some of the predictor variables with the response variables. We won't do bivariate EDA on player name, Twitter handle, age, team abbreviation, or whether the players were active on Twitter last year, instead focusing on terms we believe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

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



There is no evident relationship between assist-turnover 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.

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)	-0.621	0.431	-1.441	0.157	-1.492	0.250
log(salary_millionsCent)	0.165	0.224	0.736	0.466	-0.288	0.618
w_pctCent	4.088	2.759	1.482	0.146	-1.489	9.665
log(salary_millionsCent):w_pctCent	-1.157	1.653	-0.700	0.488	-4.497	2.183

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.