Data Analysis

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

2020-04-14

```
library(tidyverse)
library(broom)
library(stringr)
library(knitr)
library(tidyverse)
library(gridExtra)
```

```
nba_social_power <- read_csv("../data/nba.csv")

nba_2016_2017_100 <- nba_social_power

nba_social_power_mod <- nba_social_power %>%
  filter(TWITTER_HANDLE != "0") %>%
  select(PLAYER_NAME,
         TEAM_ABBREVIATION,
         AGE,
         W_PCT,
         OFF_RATING,
         DEF_RATING,
         NET_RATING,
         AST_RATIO,
         REB_PCT,
         USG_PCT,
         PIE,
         SALARY_MILLIONS,
         ACTIVE_TWITTER_LAST_YEAR,
         TWITTER_FOLLOWER_COUNT_MILLIONS,
         PTS)
```

Research Question and Objective

The research question explored in this analysis is as follows: Is there a relationship between measures of athletic success (like win percentage, offensive and defensive ratings, etc.) and internet "popularity," measured in number of twitter followers of NBA athletes?

The objective of this analysis is to predict Twitter follower counts of NBA players using measures of athletic success.

The Data

The data set we use in this analysis 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 stats and the number of twitter followers, we filtered for players who had an active twitter account, by filtering for values where `TWITTER_HANDLE` is not n/a. After filtering, we have 95 observations.

The response variable is `TWITTER_FOLLOWER_COUNT_MILLIONS`, which measures players' Twitter follower count at the time the data was collected.

Exploratory Data Analysis

Univariate

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

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

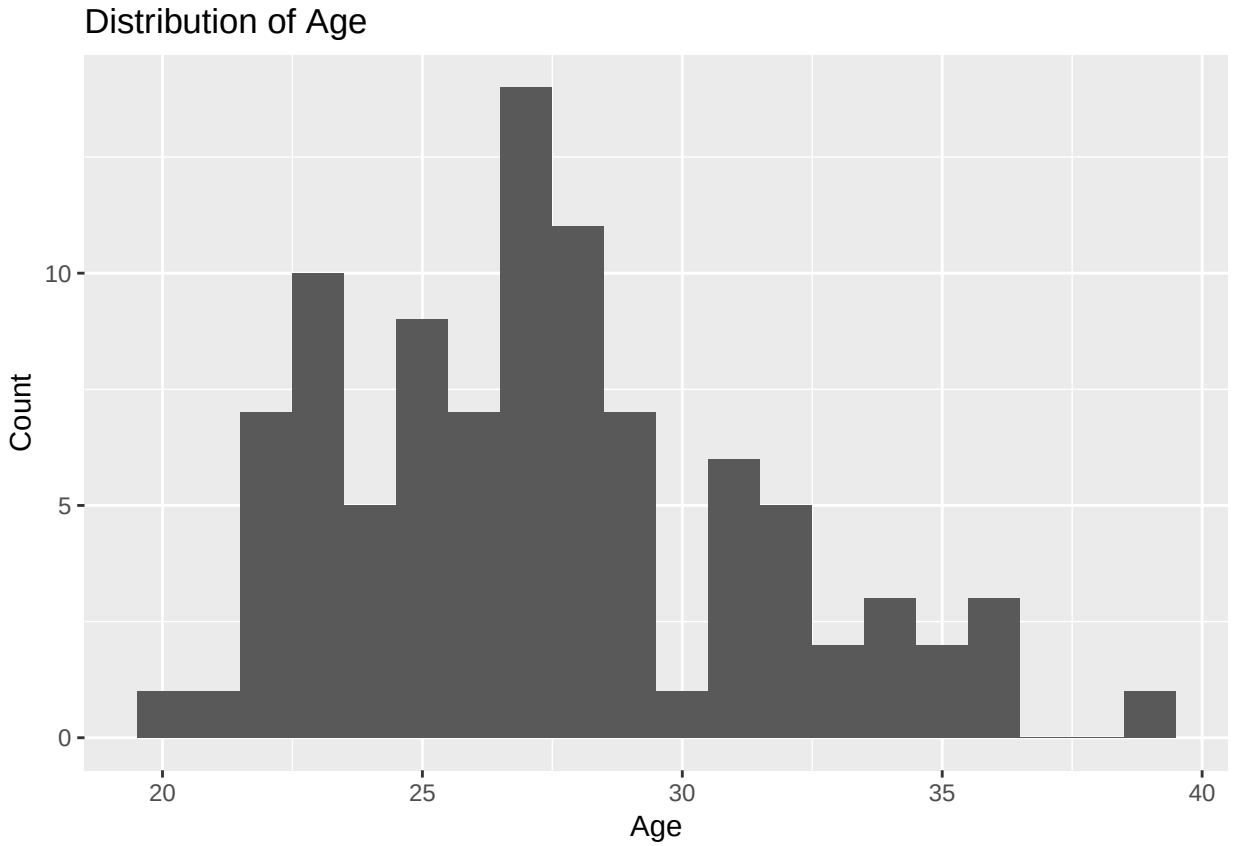
```
nba_social_power_mod %>%
  count(TEAM_ABBREVIATION) %>%
  arrange(n)
```

```
## # 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, I'll explore the distribution of the `AGE` variable in the dataset:

```
ggplot(data = nba_social_power_mod, aes(x= AGE)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Age", y = "Count", title = "Distribution of Age")
```




```
nba_social_power_mod %>%
  summarise(mean = mean(AGE), min= min(AGE), Q1 = quantile(AGE, .25), median = median(AGE),
            Q3 = quantile(AGE, .75), max = max(AGE))
```

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

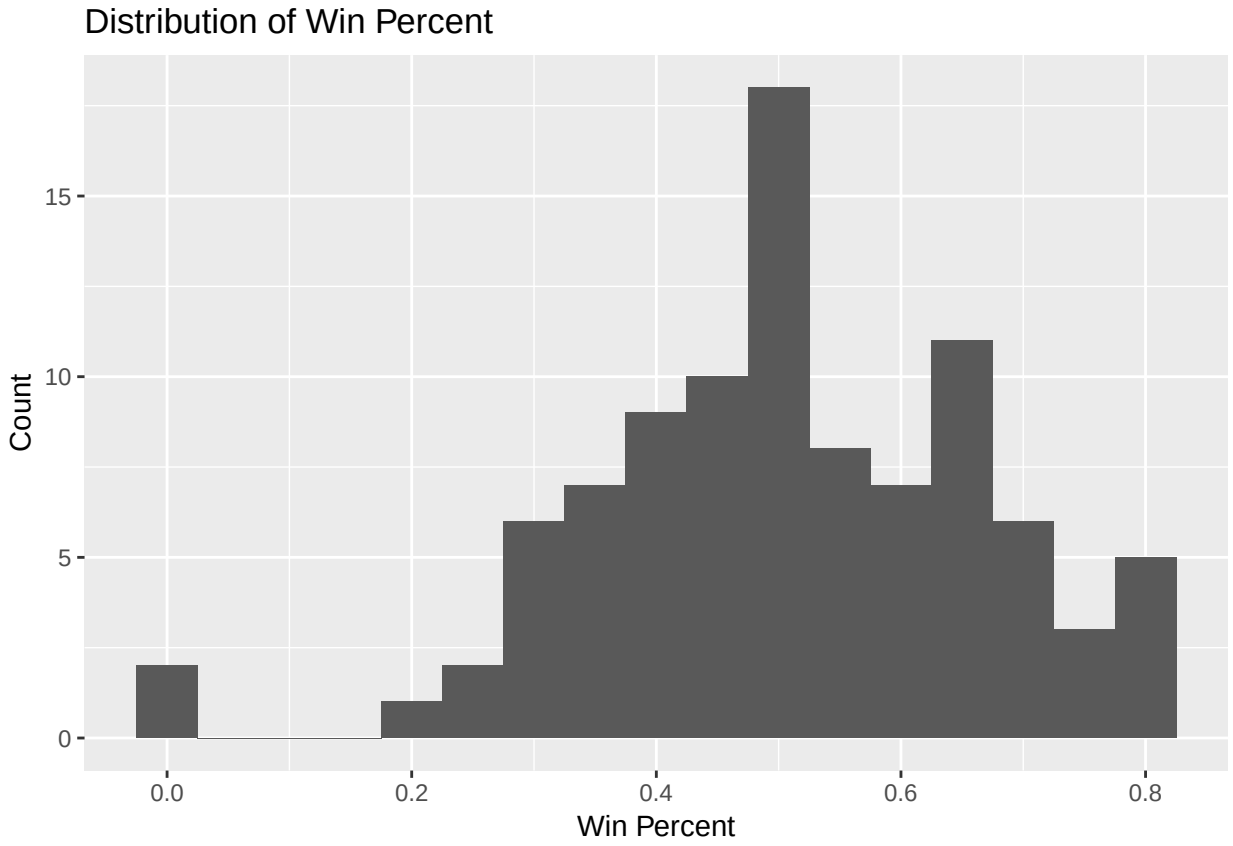
```
nba_social_power_mod %>%
  arrange(desc(AGE)) %>%
  head(1)
```

```
## # A tibble: 1 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 Dirk Nowit~ DAL                 39 0.426       105.       106.       -1.7
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <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, I'll examine the distribution of win percent, `W_PCT`:

```
ggplot(data = nba_social_power_mod, aes(x= W_PCT)) +
  geom_histogram(binwidth = .05) +
  labs(x = "Win Percent", y = "Count", title = "Distribution of Win Percent")
```



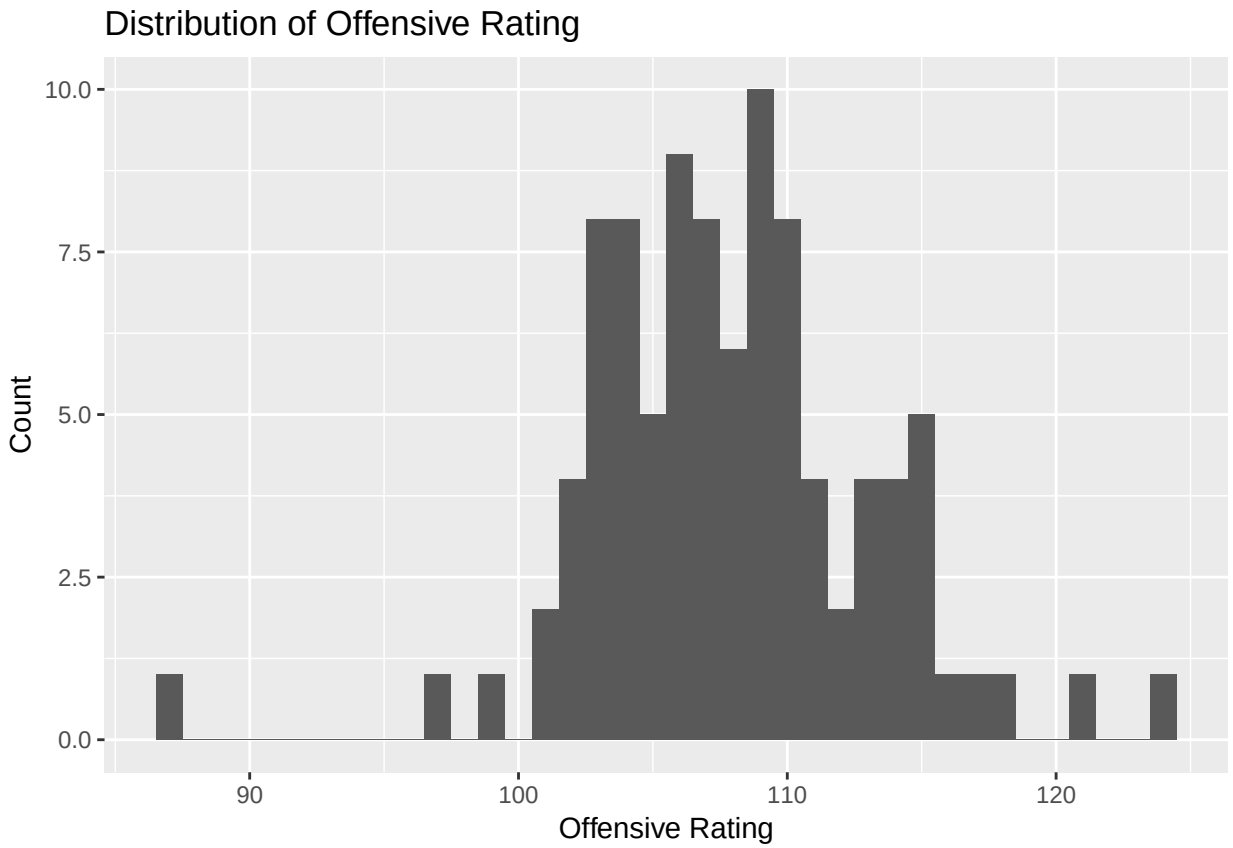

```
nba_social_power_mod %>%
  summarise(mean = mean(W_PCT), min= min(W_PCT), Q1 = quantile(W_PCT, .25),
            median = median(W_PCT), Q3 = quantile(W_PCT, .75), max = max(W_PCT))
```

```
## # 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, I'll look at the distributions for offensive rating, `OFF_RATING` and defensive rating `DEF_RATING`, as well as the distribution for `NET_RATING`, which is the average of the offensive and defensive rating:

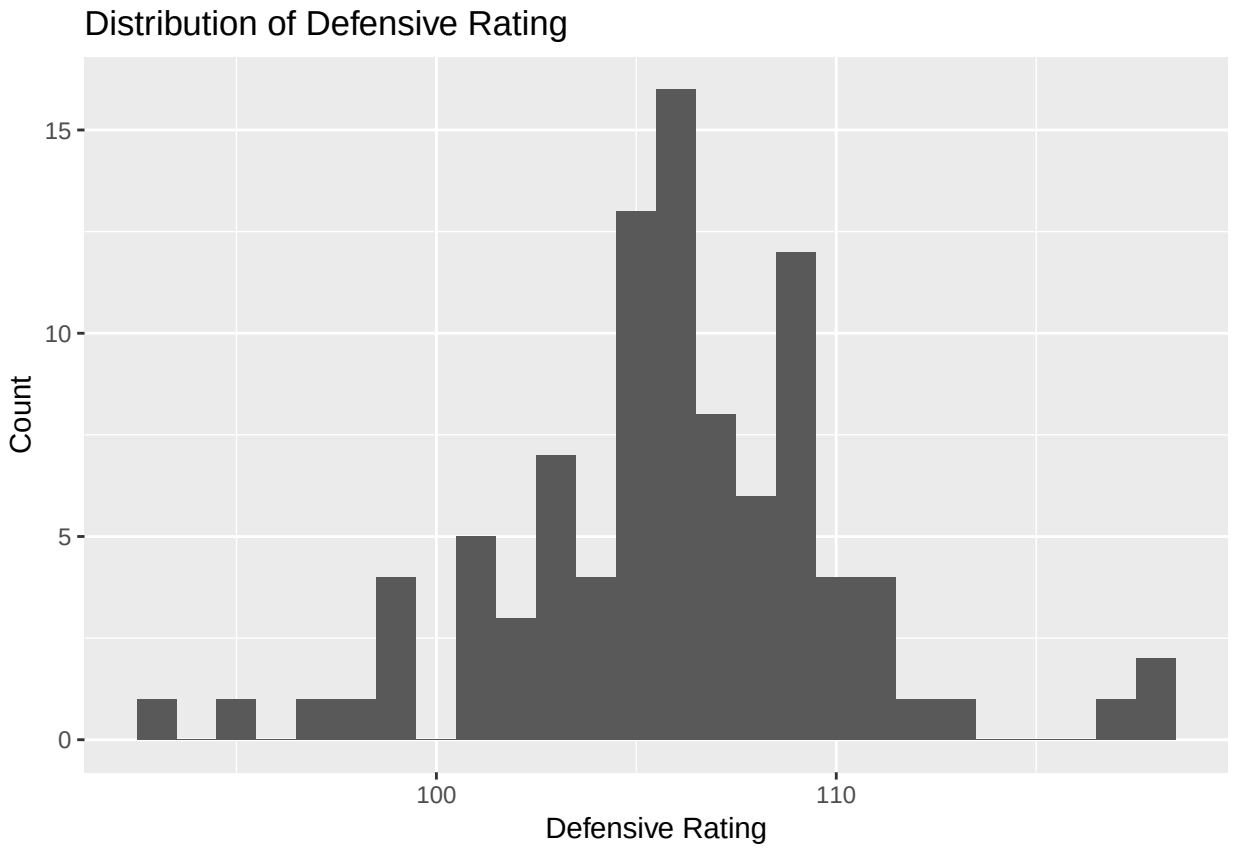
```
ggplot(data = nba_social_power_mod, aes(x= OFF_RATING)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Offensive Rating", y = "Count", title = "Distribution of Offensive Rating")
```




```
nba_social_power_mod %>%
  summarise(min= min(OFF_RATING), median = median(OFF_RATING), max = max(OFF_RATING))
```

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

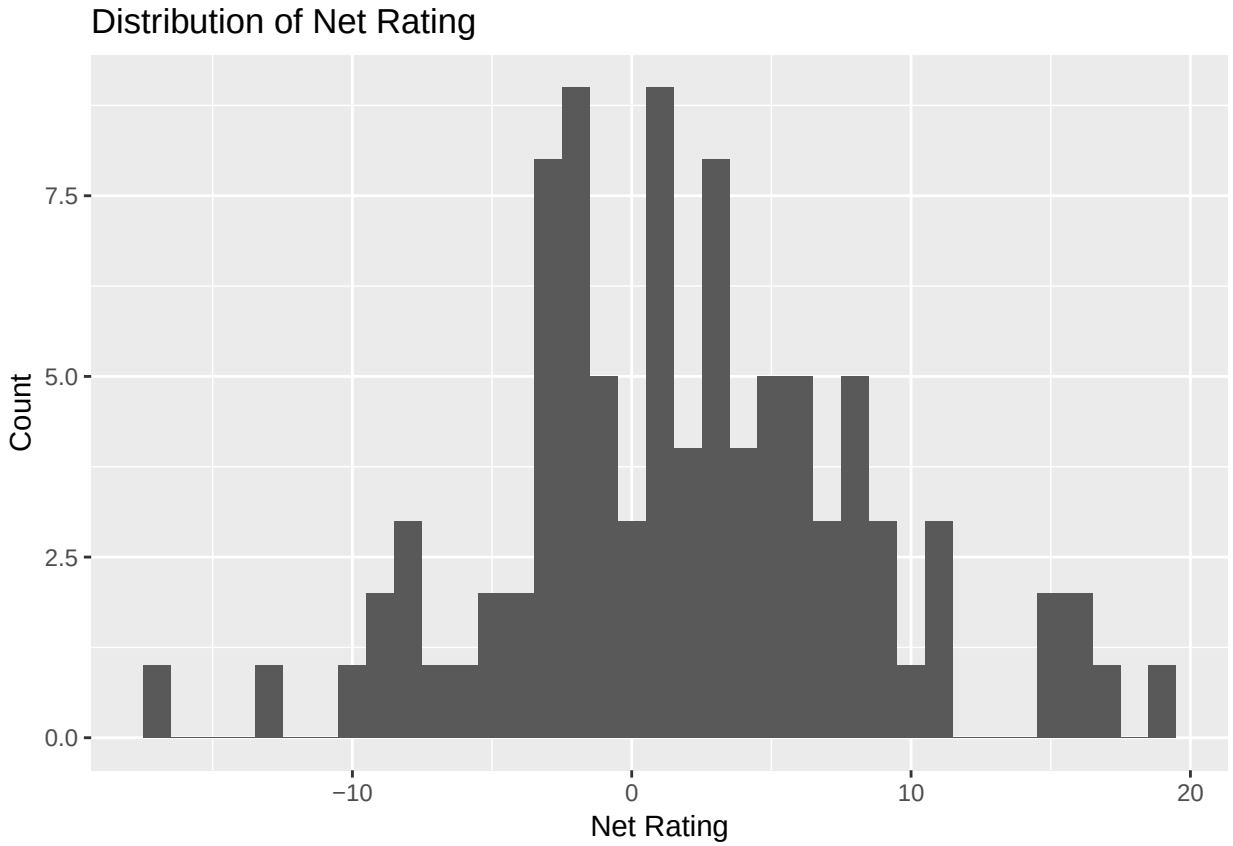
```
ggplot(data = nba_social_power_mod, aes(x= DEF_RATING)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Defensive Rating", y = "Count", title = "Distribution of Defensive Rating")
```




```
nba_social_power_mod %>%
  summarise(min= min(DEF_RATING), median = median(DEF_RATING), max = max(DEF_RATING))
```

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

```
ggplot(data = nba_social_power_mod, aes(x= NET_RATING)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Net Rating", y = "Count", title = "Distribution of Net Rating")
```



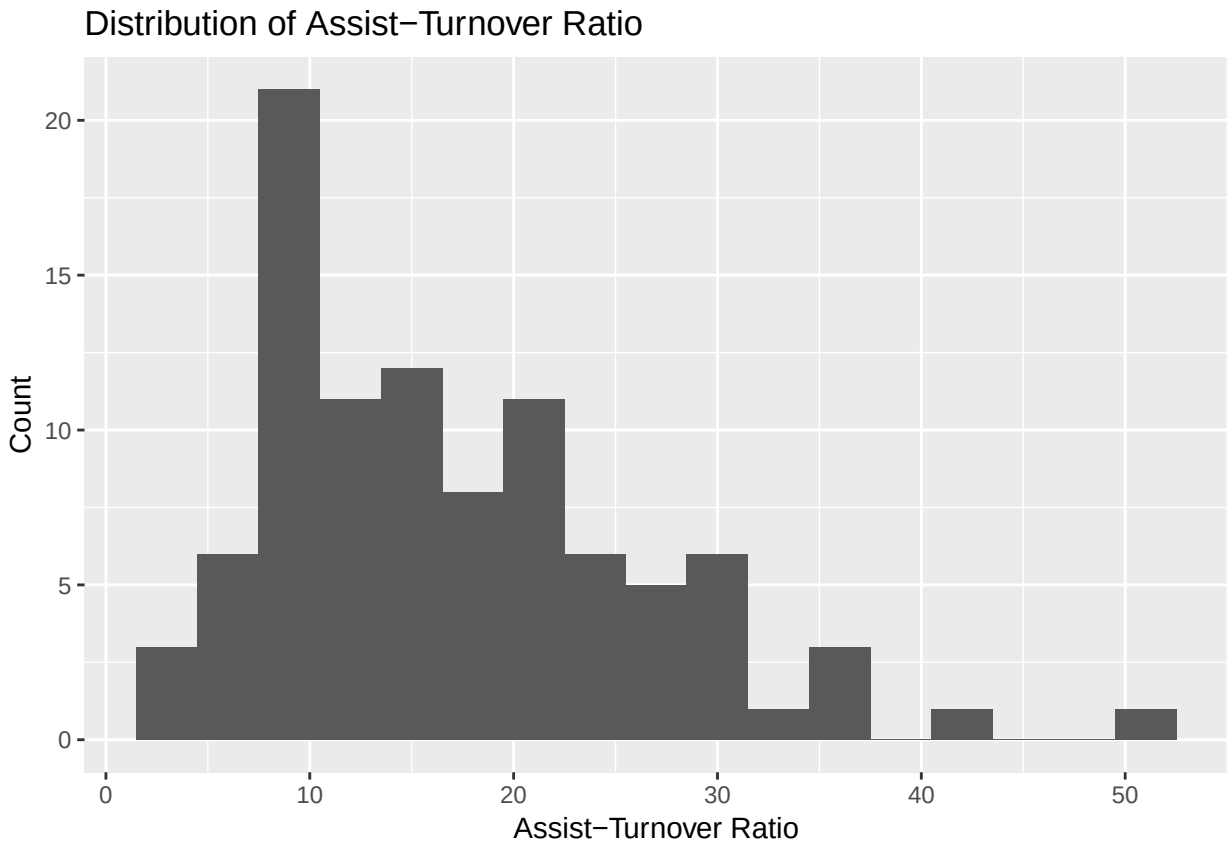

```
nba_social_power_mod %>%
  summarise(min= min(NET_RATING), median = median(NET_RATING), max = max(NET_RATING))
```

```
## # A tibble: 1 x 3
##     min median   max
##   <dbl>  <dbl> <dbl>
## 1 -17.2    1.5  18.7
```

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

```
ggplot(data = nba_social_power_mod, aes(x= AST_RATIO)) +
  geom_histogram(binwidth = 3) +
  labs(x = "Assist-Turnover Ratio", y = "Count", title = "Distribution of Assist-Turnover Ratio")
```




```
nba_social_power_mod %>%
  summarise(mean = mean(AST_RATIO), min= min(AST_RATIO), Q1 = quantile(AST_RATIO, .25), median =   media
```

```
## # 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
```

```
nba_social_power_mod %>%
  arrange(desc(AST_RATIO)) %>%
  head(2)
```

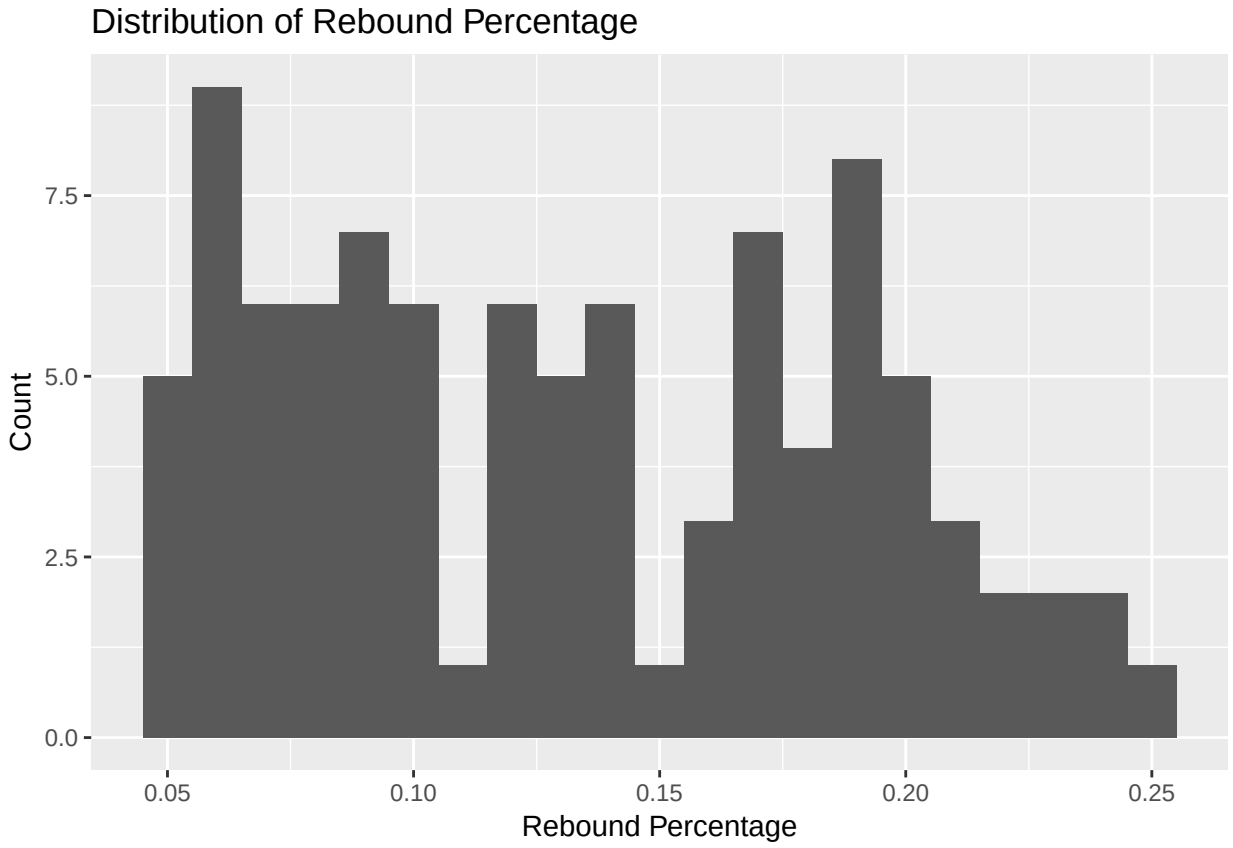
```
## # A tibble: 2 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 Jarnell St~ DEN                 23 0           115.       118.       -3.1
## 2 Ricky Rubio MIN                 26 0.373       109.       110.       -1
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <dbl>
```

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

```
ggplot(data = nba_social_power_mod, aes(x= REB_PCT)) +
  geom_histogram(binwidth = .01) +
  labs(x = "Rebound Percentage", y = "Count", title = "Distribution of Rebound Percentage")
```



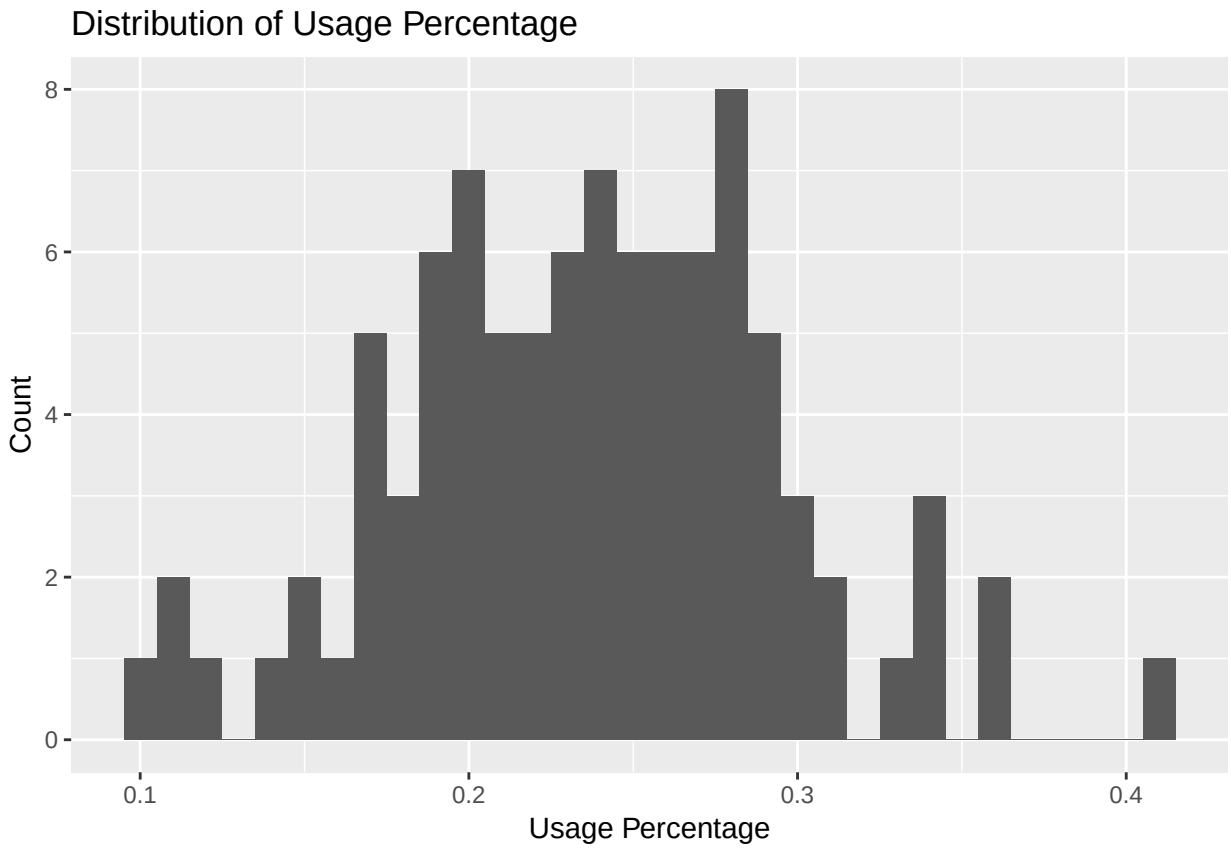

```
nba_social_power_mod %>%
  summarise(mean = mean(REB_PCT), min= min(REB_PCT), Q1 = quantile(REB_PCT, .25), median =   median(REB
```

```
## # 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, I'll look at `USG_PCT`, usage percentage, which is an estimate of how often a player makes team plays:

```
ggplot(data = nba_social_power_mod, aes(x= USG_PCT)) +
  geom_histogram(binwidth = .01) +
  labs(x = "Usage Percentage", y = "Count", title = "Distribution of Usage Percentage")
```




```
nba_social_power_mod %>%
  summarise(mean = mean(USG_PCT), min= min(USG_PCT), Q1 = quantile(USG_PCT, .25), median =   median(USG
```

```
## # 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
```

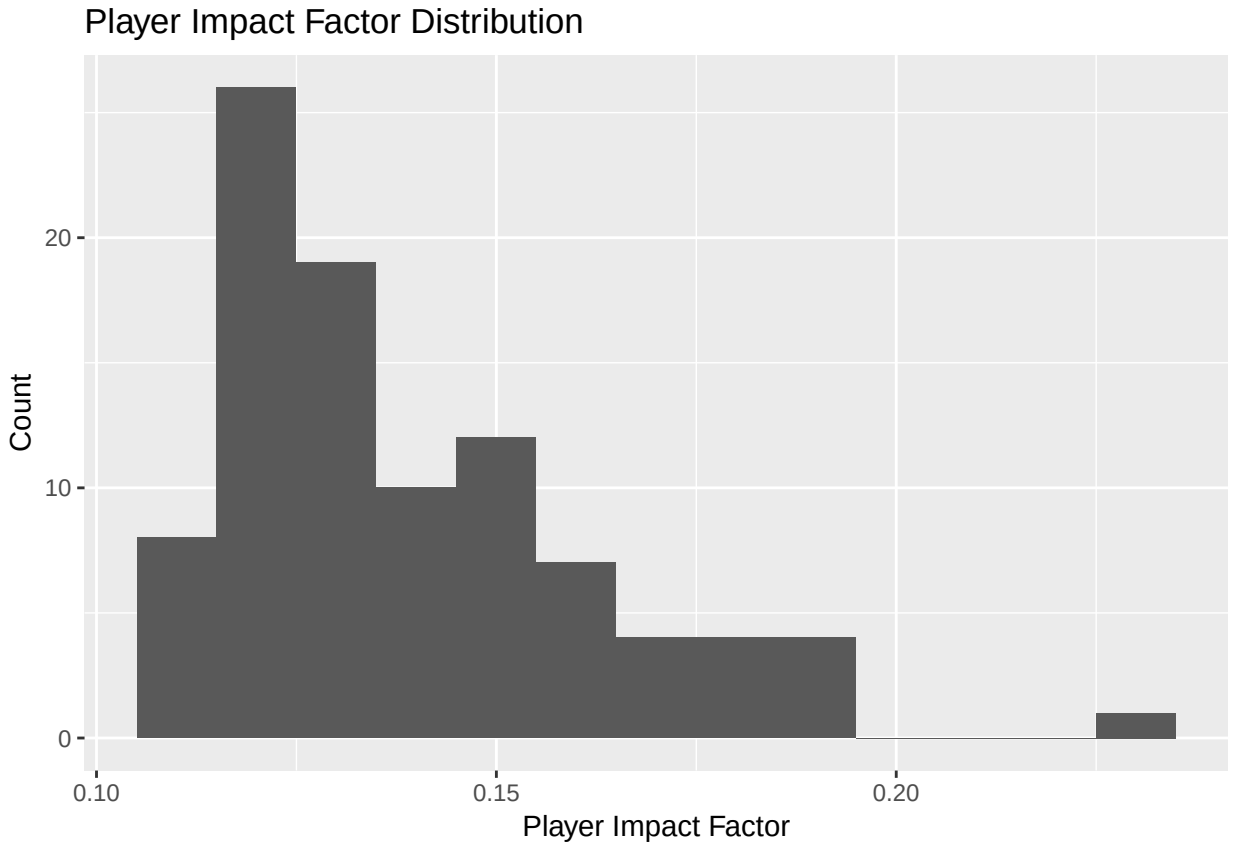
```
nba_social_power_mod %>%
  arrange(desc(USG_PCT)) %>%
  head(1)
```

```
## # A tibble: 1 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 Russell We~ OKC                 28 0.568       108.       105.        3.3
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <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, I'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:

```
ggplot(data = nba_social_power_mod, aes(x= PIE)) +
  geom_histogram(binwidth = .01) +
  labs(x = "Player Impact Factor", y = "Count", title = "Player Impact Factor Distribution")
```




```
nba_social_power_mod %>%
  summarise(mean = mean(PIE), min= min(PIE), Q1 = quantile(PIE, .25), median =   median(PIE),  Q3 = qua
```

```
## # 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
```

```
nba_social_power_mod %>%
  arrange(desc(PIE)) %>%
  head(1)
```

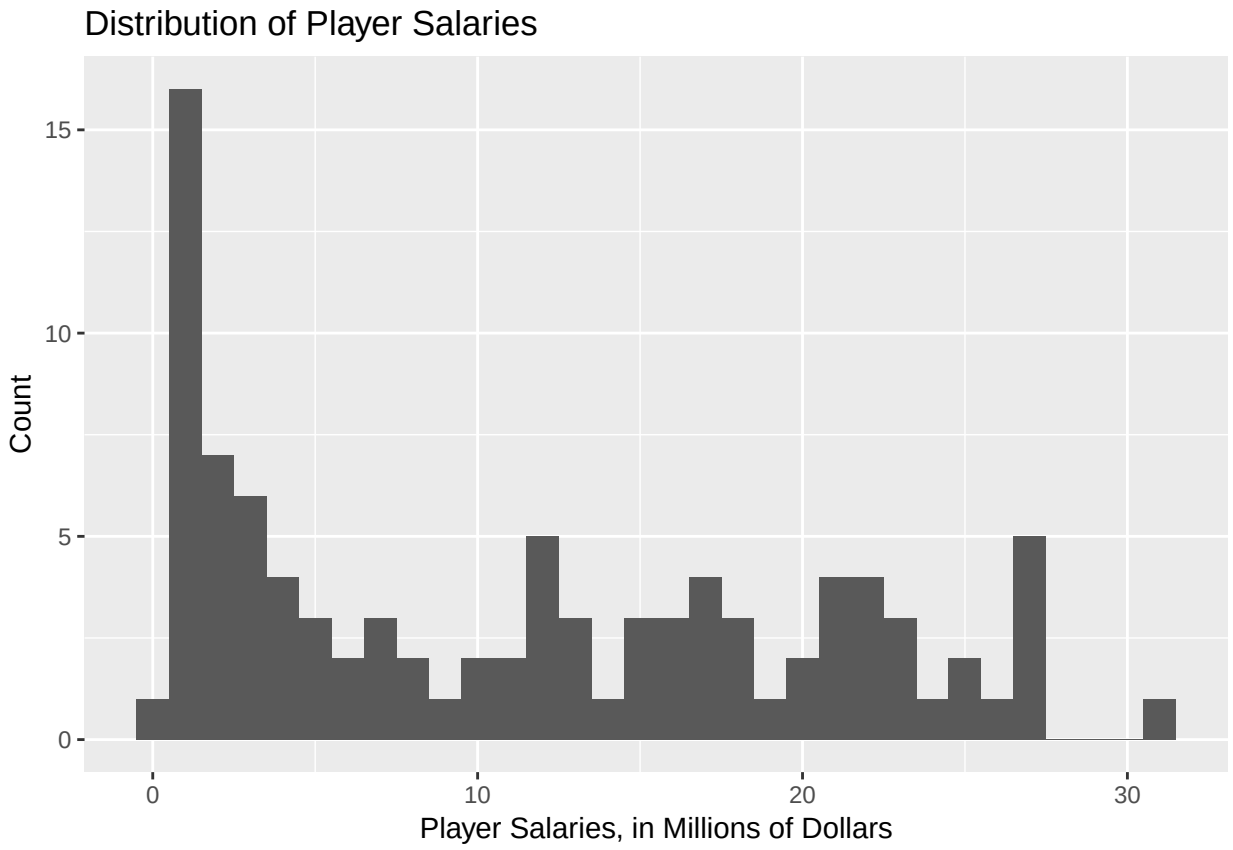
```
## # A tibble: 1 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 Russell We~ OKC                 28 0.568       108.       105.        3.3
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <dbl>
```

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

```
ggplot(data = nba_social_power_mod, aes(x= SALARY_MILLIONS)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Player Salaries, in Millions of Dollars", y = "Count", title = "Distribution of Player Sala
```




```
nba_social_power_mod %>%
  summarise(mean = mean(SALARY_MILLIONS), min= min(SALARY_MILLIONS), Q1 = quantile(SALARY_MILLIONS, .25)
```

```
## # 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
```

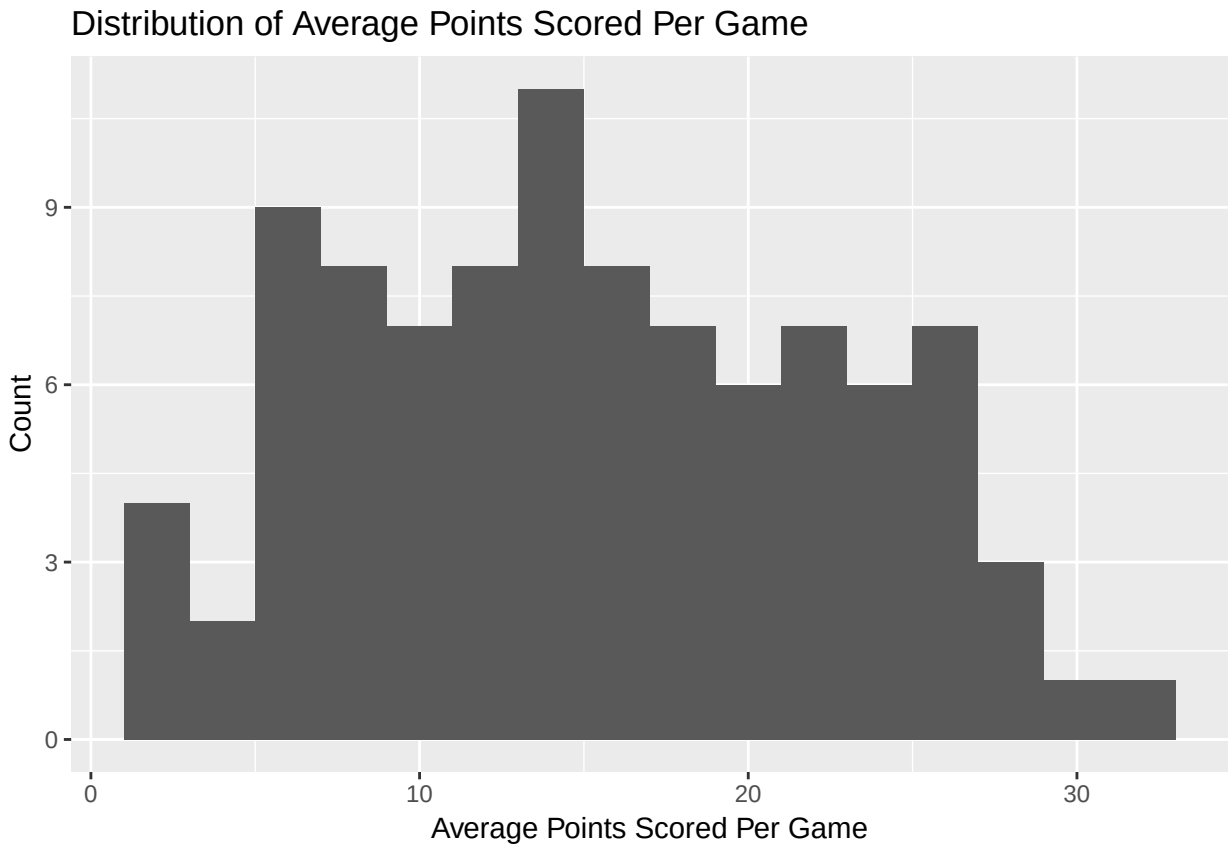
```
nba_social_power_mod %>%
  arrange(desc(SALARY_MILLIONS)) %>%
  head(1)
```

```
## # A tibble: 1 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 LeBron Jam~ CLE                 32 0.689       115.       107.        7.7
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <dbl>
```

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 makes the most by far, at 30.96 million dollars a year, is LeBron James.

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

```
ggplot(data = nba_social_power_mod, aes(x= PTS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "Average Points Scored Per Game", y = "Count", title = "Distribution of Average Points Scored
```




```
nba_social_power_mod %>%
  summarise(mean = mean(PTS), min= min(PTS), Q1 = quantile(PTS, .25), median =   median(PTS),  Q3 = quan
```

```
## # 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, I'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:

```
nba_social_power_mod %>%
  count(ACTIVE_TWITTER_LAST_YEAR)
```

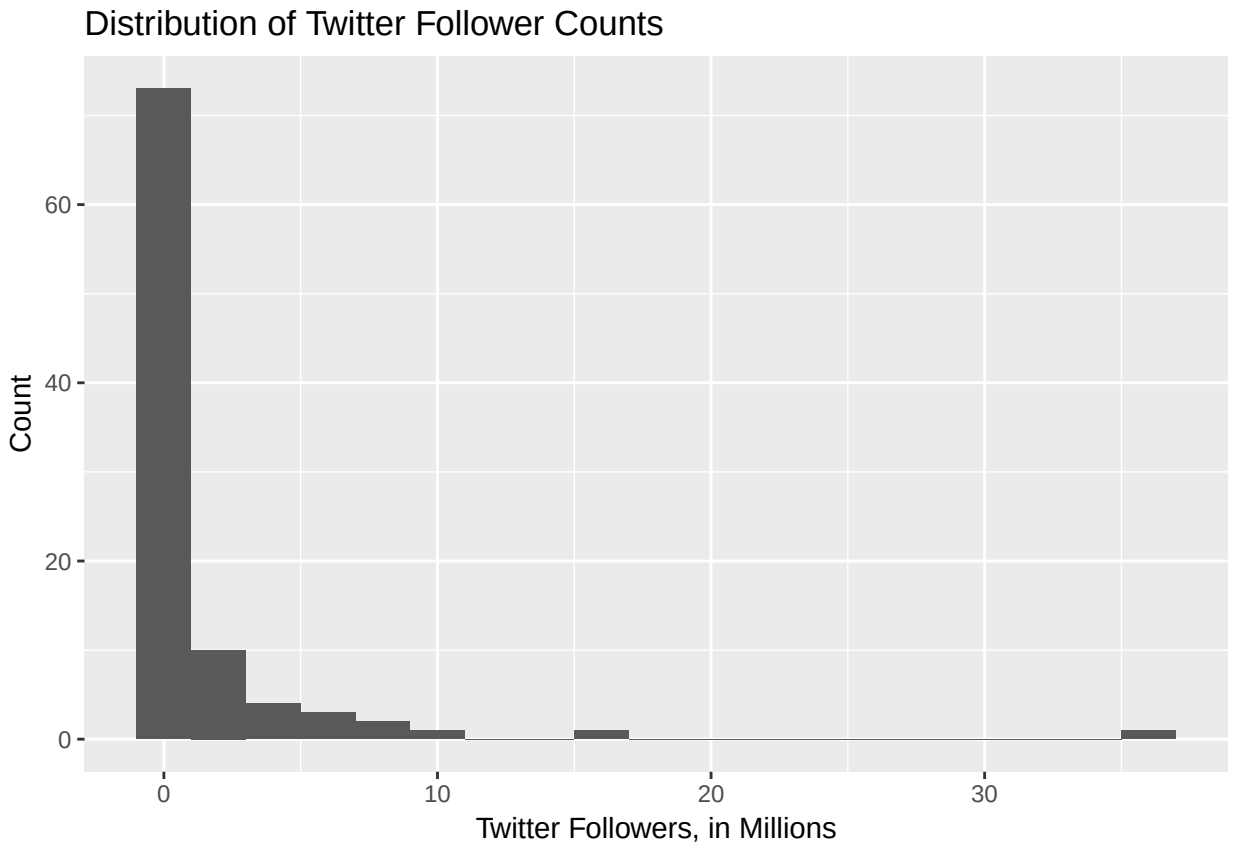
```
## # A tibble: 2 x 2
##   ACTIVE_TWITTER_LAST_YEAR     n
```

```
##                       <int> <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.

Finally, I'll look into the response variable, `TWITTER_FOLLOWER_COUNT_MILLIONS`:

```
ggplot(data = nba_social_power_mod, aes(x= TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "Twitter Followers, in Millions", y = "Count", title = "Distribution of Twitter Follower Cou
```




```
nba_social_power_mod %>%
  summarise(mean = mean(TWITTER_FOLLOWER_COUNT_MILLIONS), min= min(TWITTER_FOLLOWER_COUNT_MILLIONS), Q1
```

```
## # 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
```

```
nba_social_power_mod %>%
  arrange(desc(TWITTER_FOLLOWER_COUNT_MILLIONS)) %>%
  head(2)
```

```
## # A tibble: 2 x 15
##   PLAYER_NAME TEAM_ABBREVIATI~   AGE W_PCT OFF_RATING DEF_RATING NET_RATING
##   <chr>       <chr>            <int> <dbl>      <dbl>      <dbl>      <dbl>
## 1 LeBron Jam~ CLE                 32 0.689       115.       107.        7.7
## 2 Kevin Dura~ GSW                 28 0.823       117.       101.       16
```

```
## # ... with 8 more variables: AST_RATIO <dbl>, REB_PCT <dbl>,
## #   USG_PCT <dbl>, PIE <dbl>, SALARY_MILLIONS <dbl>,
## #   ACTIVE_TWITTER_LAST_YEAR <int>, TWITTER_FOLLOWER_COUNT_MILLIONS <dbl>,
## #   PTS <dbl>
```

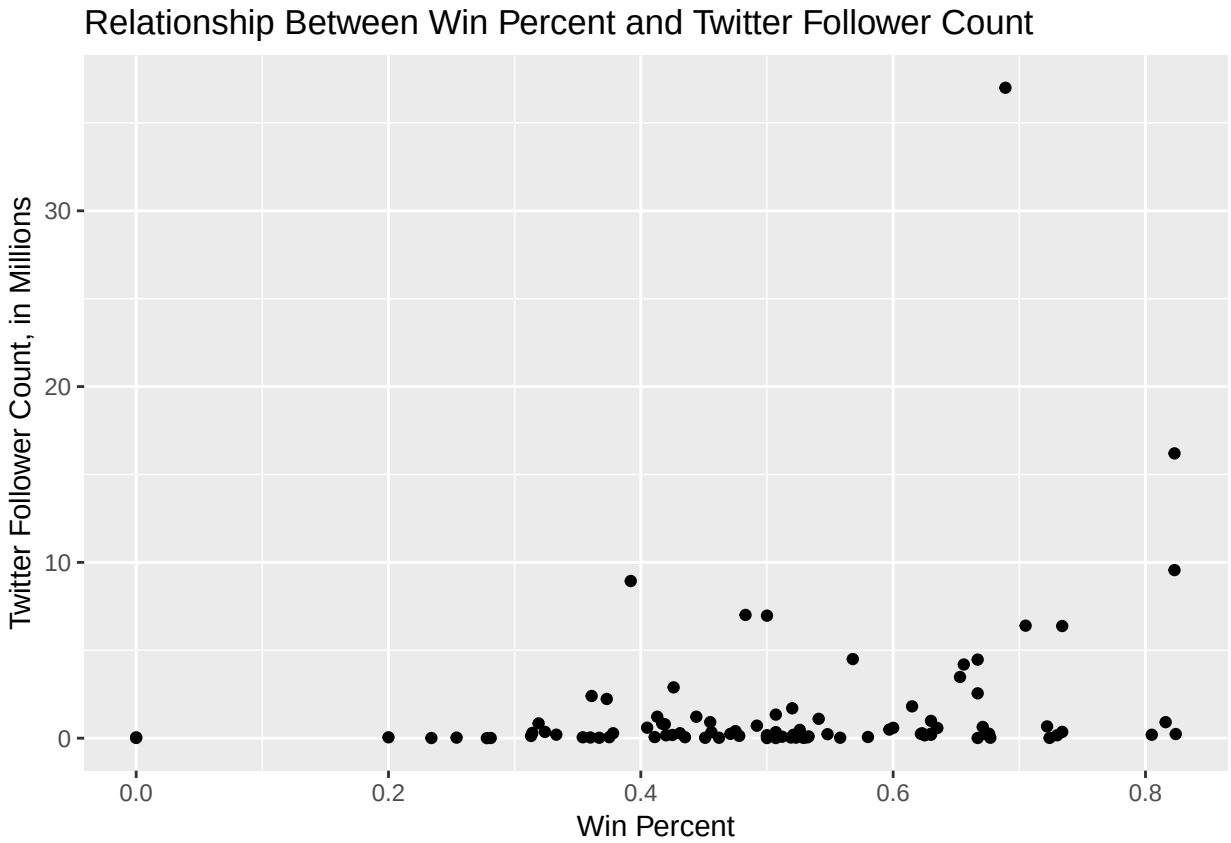
From the histogram, we can see that the distribution of twitter follower counts is extremely right-skewed. The number of twitter followers ranges from .002 million to 37 million, with a mean of 1.6 million and a median of.246 million. There are two obvious outliers: Kevin Durant, with 16.2 million followers,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

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, I'll look for a relationship between win percent and twitter followers in millions:

```
ggplot(nba_social_power_mod, aes(x = W_PCT, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Win Percent and Twitter Follower Count",
       x = "Win Percent", y = "Twitter Follower Count, in Millions")
```

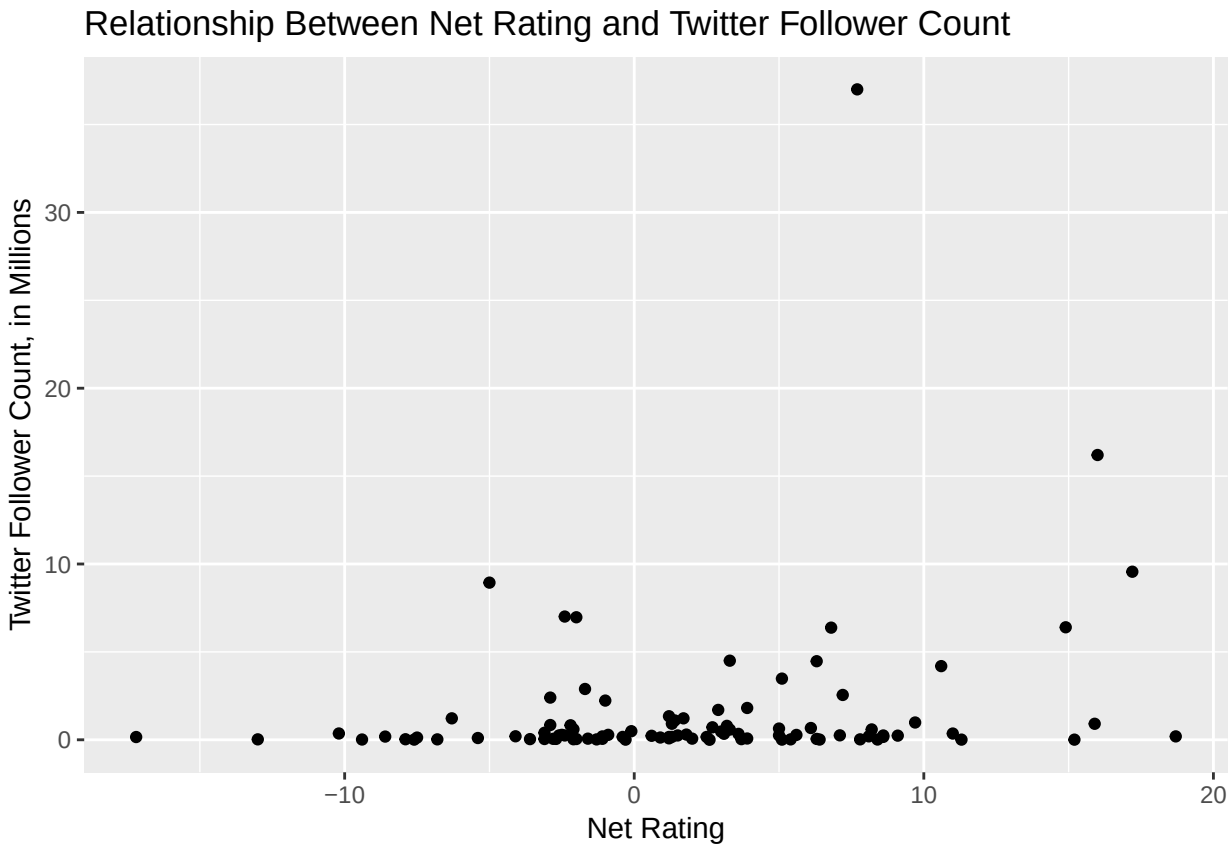


From the above plot, we can see that win percent and twitter follower count may have a very weak positive orrelations. The players with significantly higher-than-average Twitter follower counts tend to have higher win percentages; however, this relationship is very weak.

Next, I'll look at the relationship between net rating, which combines offensive and defensive rating, and

Twitter follower count:

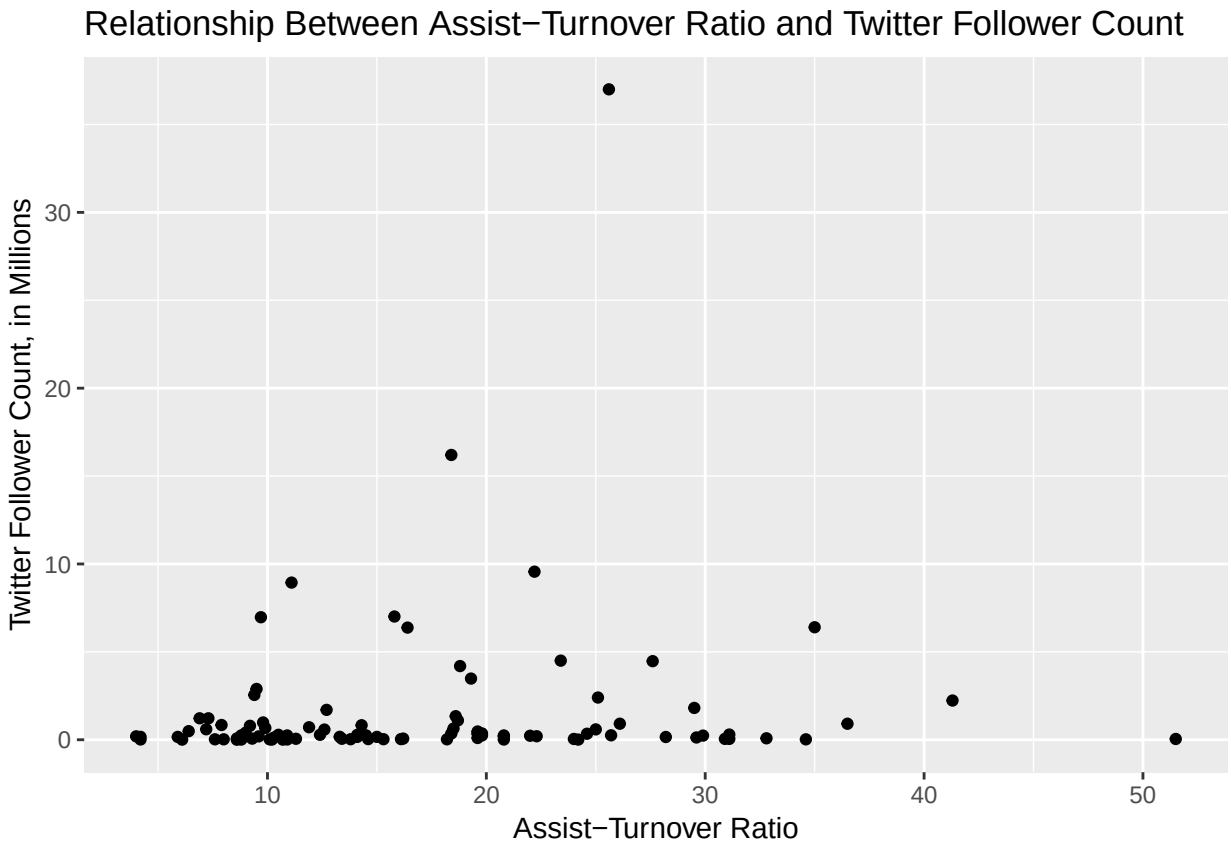
```
ggplot(nba_social_power_mod, aes(x = NET_RATING, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Net Rating and Twitter Follower Count",
       x = "Net Rating", y = "Twitter Follower Count, in Millions")
```



Similarly to win percent, Twitter follower count and net rating seem to have a very weak positive relationship. The players with the highest net ratings more often have higher-than-average Twitter follower counts, and more players with positive net ratings have high Twitter follower counts than those with negative net ratings.

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

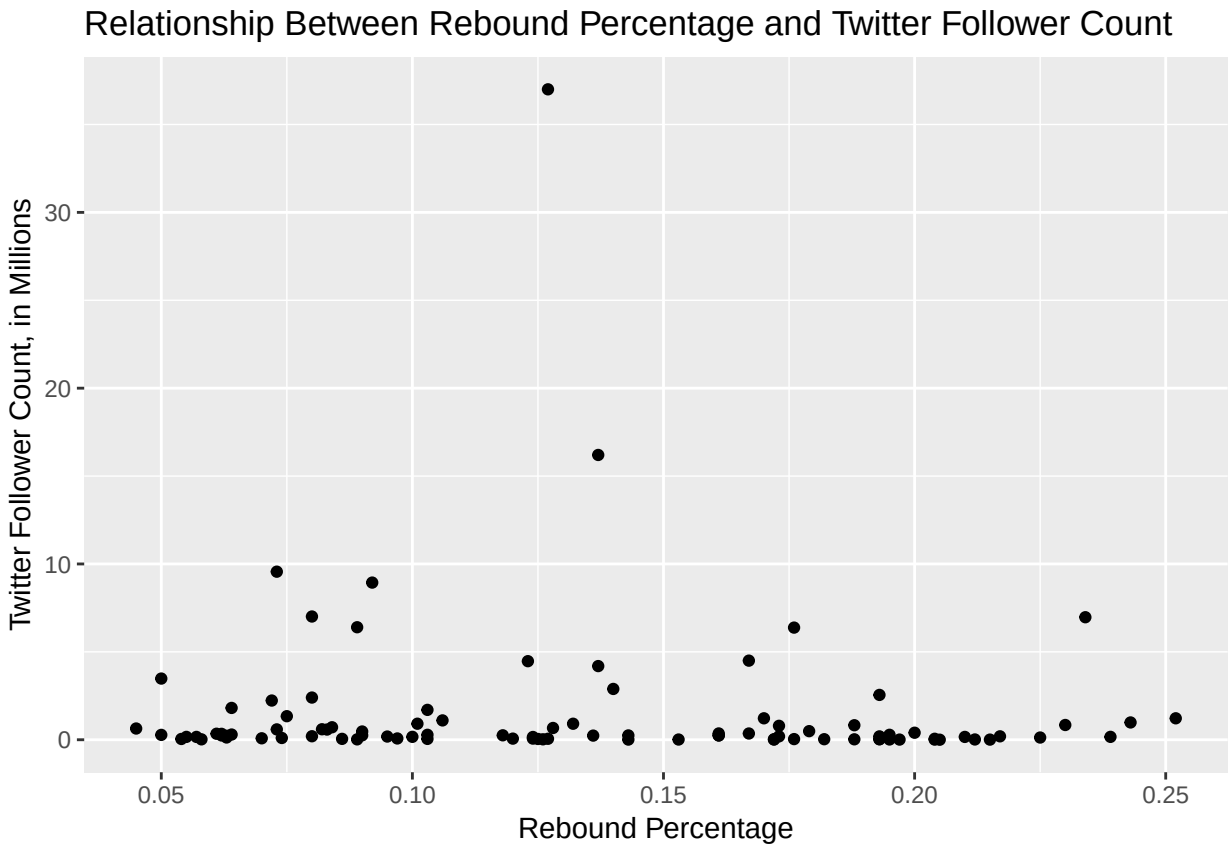
```
ggplot(nba_social_power_mod, aes(x = AST_RATIO, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Assist-Turnover Ratio and Twitter Follower Count",
       x = "Assist-Turnover Ratio", y = "Twitter Follower Count, in Millions")
```



There is no evident relationship between assist-turnover ratio and Twitter follower count.

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

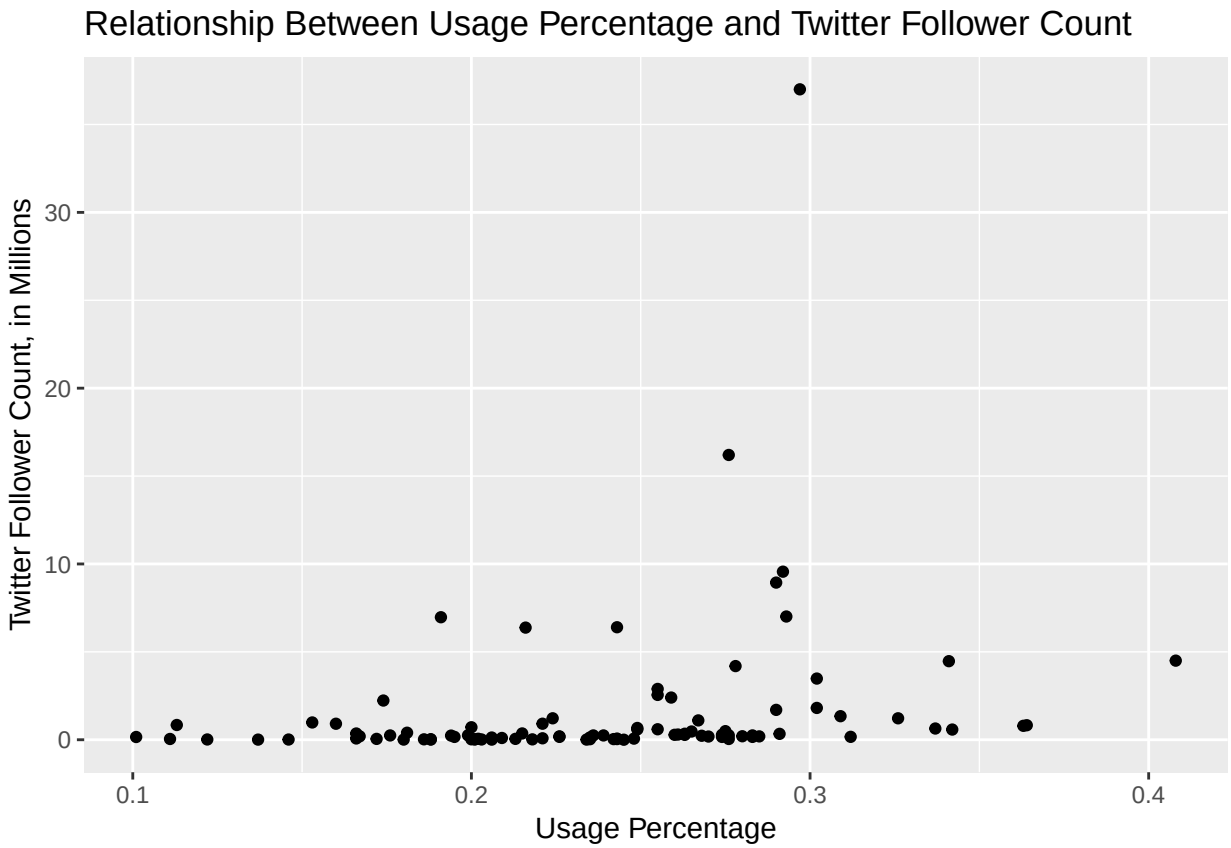
```
ggplot(nba_social_power_mod, aes(x = REB_PCT, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Rebound Percentage and Twitter Follower Count",
       x = "Rebound Percentage", y = "Twitter Follower Count, in Millions")
```



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

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

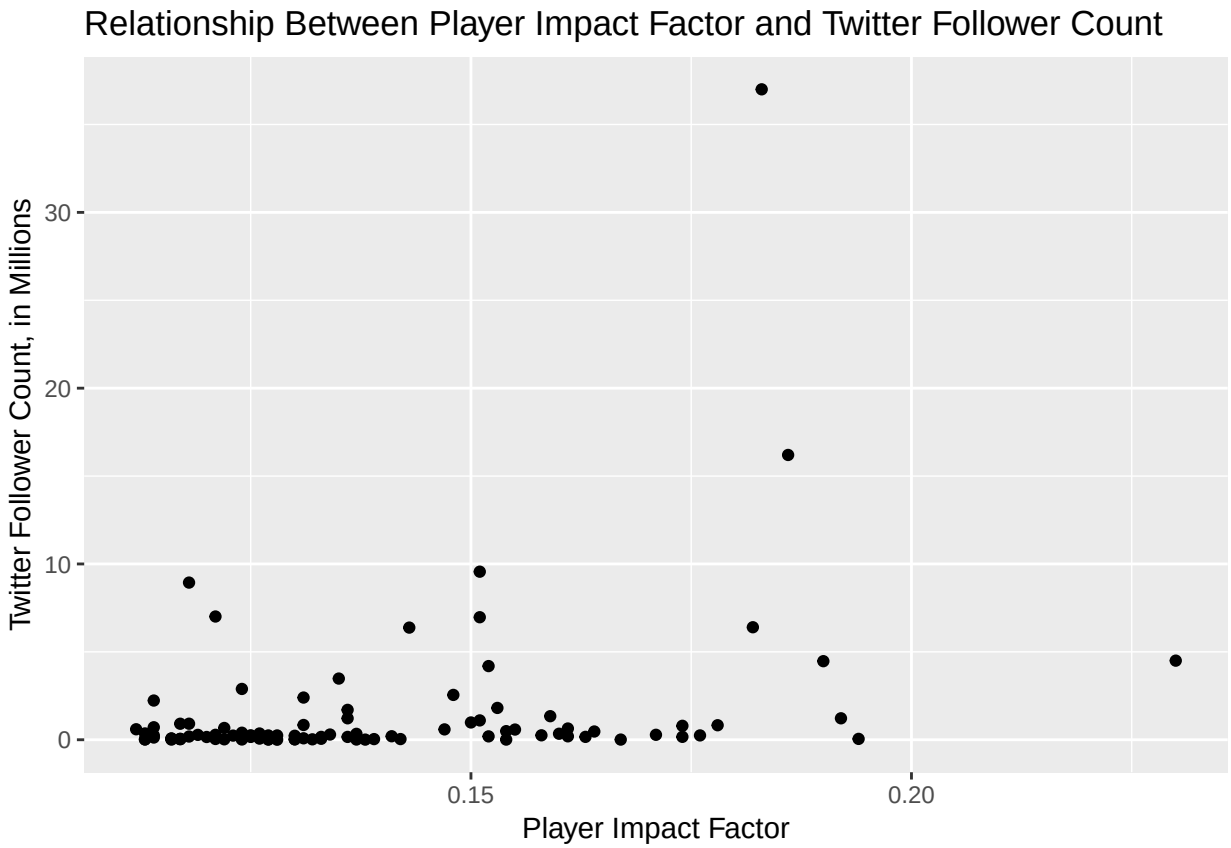
```
ggplot(nba_social_power_mod, aes(x = USG_PCT, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Usage Percentage and Twitter Follower Count",
       x = "Usage Percentage", y = "Twitter Follower Count, in Millions")
```



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, I'll examine whether there is a relationship between player impact factor and Twitter follower count:

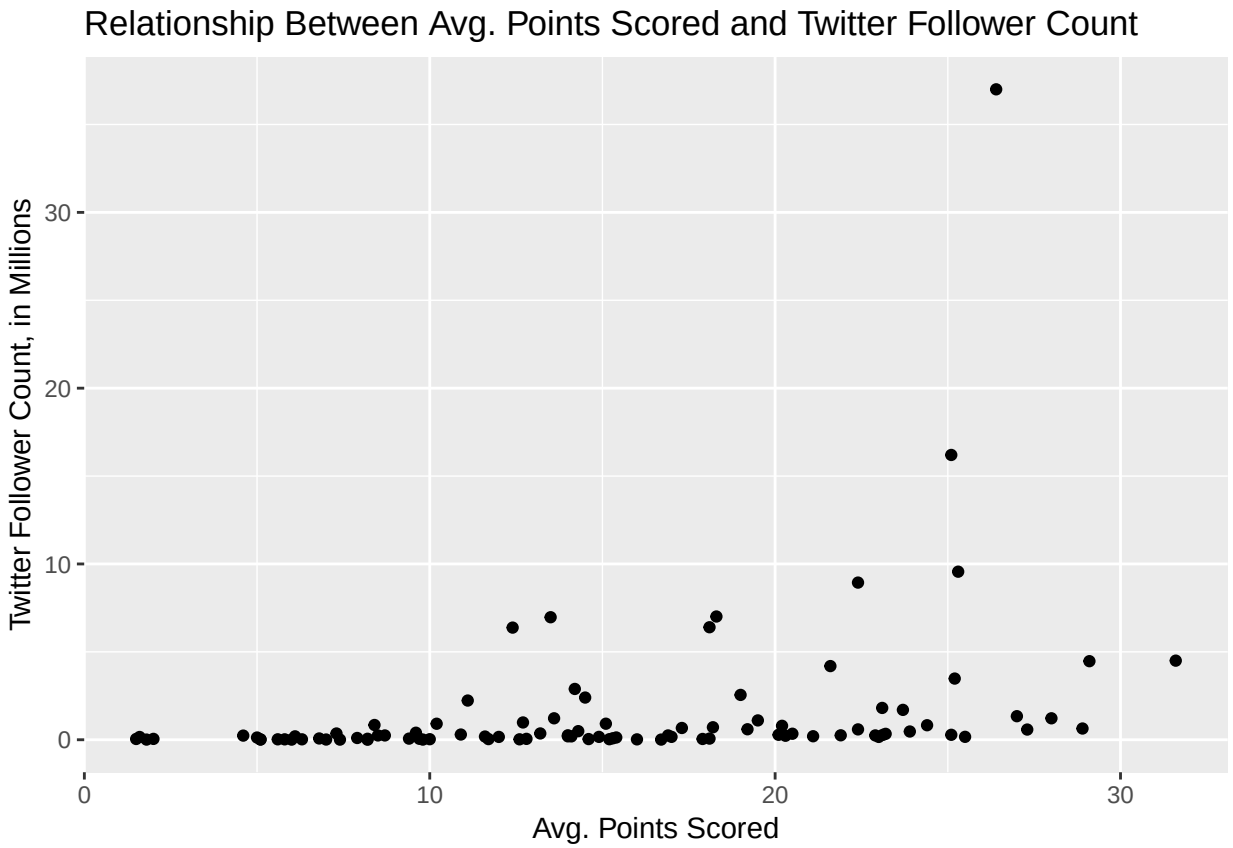
```
ggplot(nba_social_power_mod, aes(x = PIE, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Player Impact Factor and Twitter Follower Count",
       x = "Player Impact Factor", y = "Twitter Follower Count, in Millions")
```



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

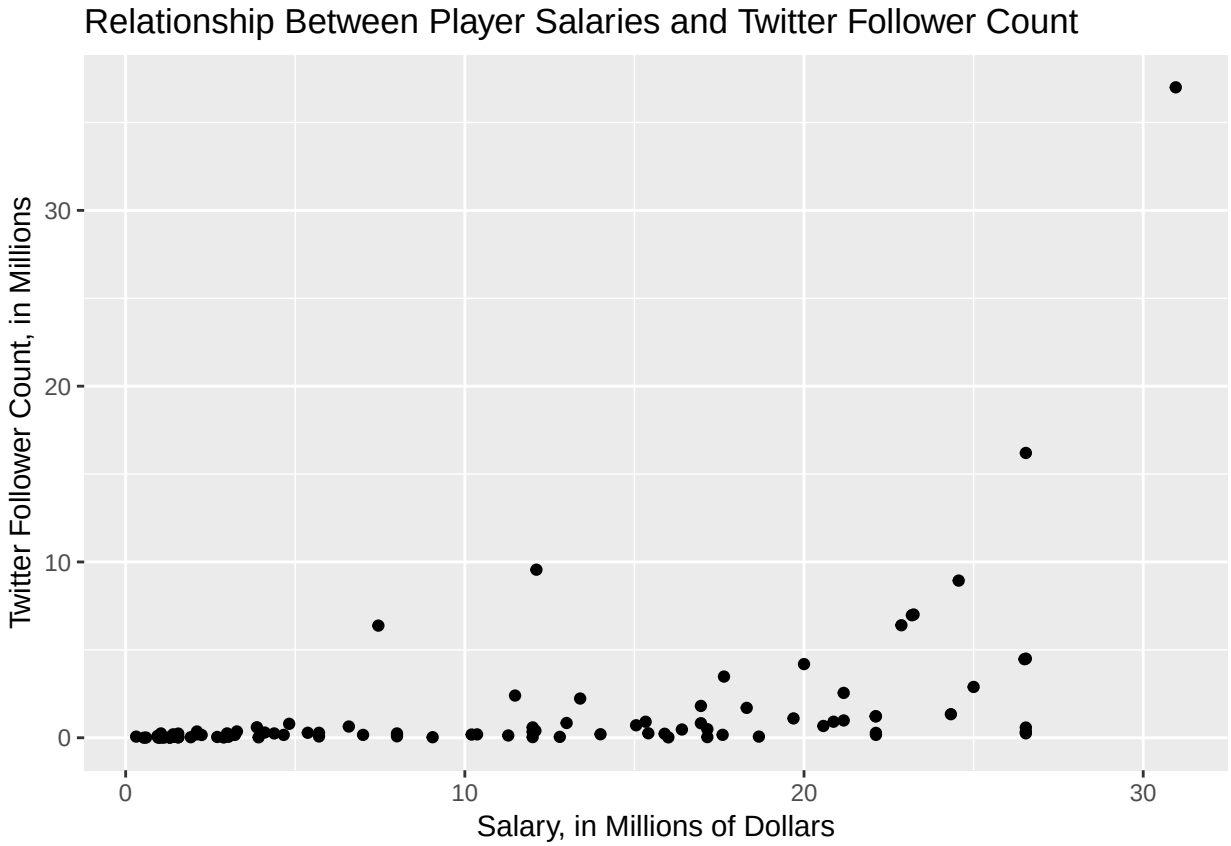
```
ggplot(nba_social_power_mod, aes(x = PTS, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Avg. Points Scored and Twitter Follower Count",
       x = "Avg. Points Scored", y = "Twitter Follower Count, in Millions")
```



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.

Finally, I'll look for a relationship between player salaries and points scored:

```
ggplot(nba_social_power_mod, aes(x = SALARY_MILLIONS, y = TWITTER_FOLLOWER_COUNT_MILLIONS)) +
  geom_point() +
  labs(title = "Relationship Between Player Salaries and Twitter Follower Count",
       x = "Salary, in Millions of Dollars", y = "Twitter Follower Count, in Millions")
```



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.

In sum, there appear to be weak to moderate positive correlations between Twitter follower count and win percentage, net rating, usage percentage, player impact factor, avg. points per game, and player salaries; there is no obvious relationship between Twitter follower count and assist-to-turnover ratio or rebound percentage.

patchwork: multinomial logistic lecture code

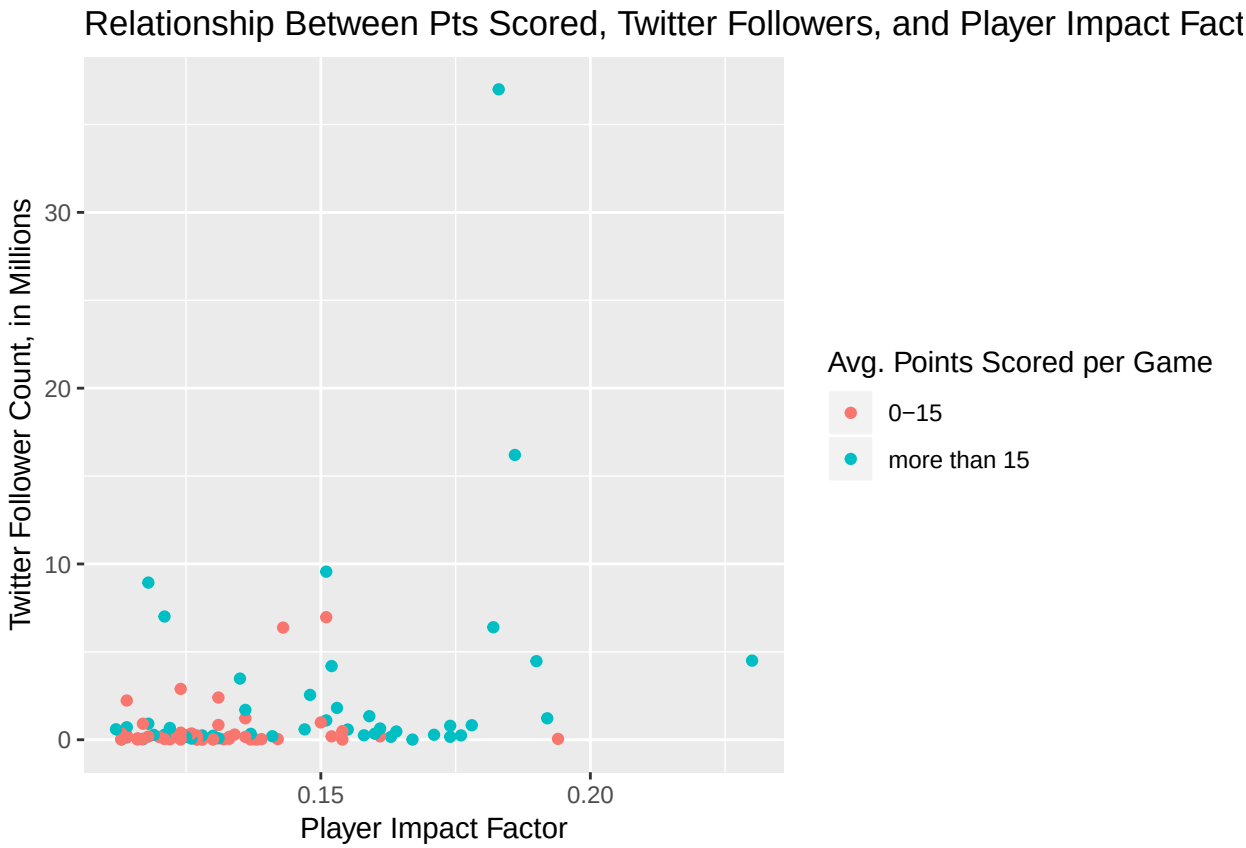
Multivariate Data Analysis

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

First, I'll look to see if points scored affects the way player impact factor relates to Twitter follower count.

```
nba_social_power_mod1 <- nba_social_power_mod %>%
  mutate(PTS_CAT = case_when(
    PTS <= 15 ~ "0-15" ,
    PTS > 15 ~ "more than 15"
  ))
```

```
ggplot(data = nba_social_power_mod1, aes(x = PIE, y = TWITTER_FOLLOWER_COUNT_MILLIONS, color = PTS_CAT))
        geom_point() +
  labs(x = "Player Impact Factor", y= "Twitter Follower Count, in Millions",
       color = "Avg. Points Scored per Game",
       title = "Relationship Between Pts Scored, Twitter Followers, and Player Impact Factor")
```

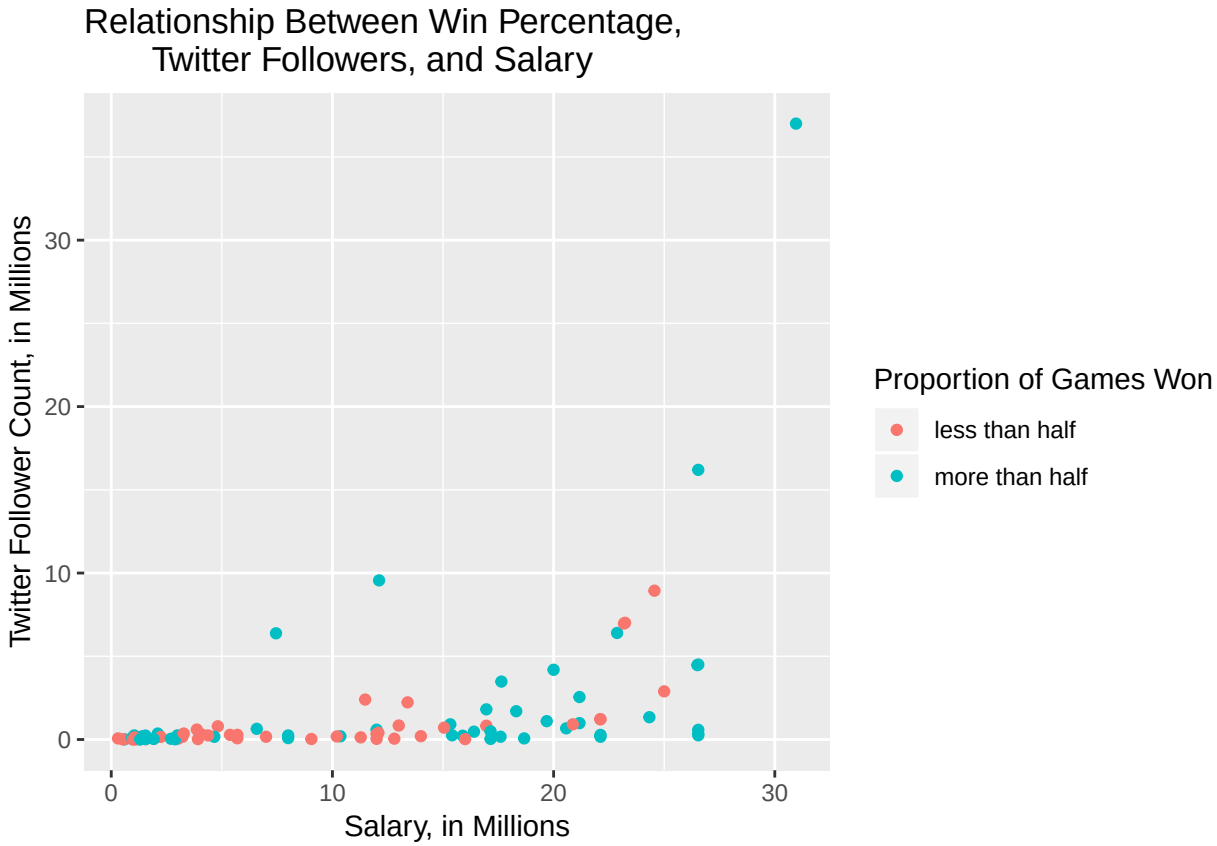


As we can see from the above color-coded boxplot, many of the players with the highest points scored numbers have much higher player impact factors, and player impact factor values have a weak positive correlation with Twitter follower count. This could be an opportunity for an interaction term.

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

```
nba_social_power_mod1 <- nba_social_power_mod %>%
  mutate(W_PCT_CAT = case_when(
    W_PCT <= .5 ~ "less than half" ,
    W_PCT > .5 ~ "more than half"
  ))
```

```
ggplot(data = nba_social_power_mod1, aes(x = SALARY_MILLIONS, y = TWITTER_FOLLOWER_COUNT_MILLIONS,color
         geom_point() +
  labs(x = "Salary, in Millions", y= "Twitter Follower Count, in Millions",
       color = "Proportion of Games Won",
       title = "Relationship Between Win Percentage,
       Twitter Followers, and Salary")
```



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 Twitter follower count. This could also be an opportunity for an interaction term.