deliverable3

Craig Le

12/14/2020

Reviewing and Revising

From deliverable 1 to deliverable 2, my project has gone pretty smoothly. My datasets were not too difficult to find, and I have only had to do minimal tidying to both of them. The main bit of tidying that I had to do was changing column names in order to make them suitable for performing code operations on them. Also for my second dataset I decided to keep all rows separate in order to do an explicit separation between the home and away team stats. I created two extra tibbles that hold every single away team and home team. I then added a column that marked them as either home (H) or away (A), so that when I merged the two tables back together I knew for sure that each home and away team was properly marked.

```
#Renaming some columns that have unwanted symbols
NBA_box_stats <- as_tibble(read_csv("2018-19_detailed_box - Sheet1.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Team = col_character(),
##   `Match Up` = col_character(),
##   `Game Date` = col_character(),
##   `W/L` = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
NBA_box_stats <- NBA_box_stats %>% rename("PLUS/MINUS" = "#ERROR!")
NBA_box_stats <- NBA_box_stats %>% rename("THREE_PTM" = "3:00 PM")
NBA_box_stats <- NBA_box_stats %>% rename("THREE_PTA" = "3PA")
NBA_box_stats <- NBA_box_stats %>% rename("THREE_PTPERCENT" = "3P%")
NBA_box_stats <- NBA_box_stats %>% rename("FG_PERCENT" = "FG%")
NBA_box_stats <- NBA_box_stats %>% rename("FT_PERCENT" = "FT%")
NBA_box_stats <- NBA_box_stats %>% rename("W_or_L" = "W/L")
NBA_box_stats <- NBA_box_stats %>% rename("Match_Up" = "Match Up")
NBA_box_stats <- NBA_box_stats %>% rename("Game_Date" = "Game Date")
#NBA_box_stats
#summary(NBA_box_stats)

Home_games <- NBA_box_stats %>% filter(str_detect(Match_Up, 'vs.'))
Away_games <- NBA_box_stats %>% filter(str_detect(Match_Up, '@'))

Home_games$Status <- paste("H")
```

```r
Away_games$Status <- paste("A")

NBA_box_stats <- as_tibble(merge(Home_games, Away_games, all = TRUE))

Home_games
```

```
## # A tibble: 1,230 x 25
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 MEM   MEM vs.~ 04/10/20~ W        240   132    48    98       49          21
##  2 CHA   CHA vs.~ 04/10/20~ L        240   114    41    78       52.6         8
##  3 DEN   DEN vs.~ 04/10/20~ W        240    99    39    87       44.8        10
##  4 MIL   MIL vs.~ 04/10/20~ L        240   116    43   100       43          15
##  5 ATL   ATL vs.~ 04/10/20~ L        240   134    43   103       41.7        17
##  6 POR   POR vs.~ 04/10/20~ W        240   136    53    91       58.2        14
##  7 LAC   LAC vs.~ 04/10/20~ W        265   143    54   106       50.9        12
##  8 BKN   BKN vs.~ 04/10/20~ W        240   113    43   114       37.7        21
##  9 NYK   NYK vs.~ 04/10/20~ L        240    89    31    77       40.3         9
## 10 SAS   SAS vs.~ 04/10/20~ W        240   105    41    88       46.6         8
## # ... with 1,220 more rows, and 15 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>
```

```r
summary(Home_games)
```

```
##      Team             Match_Up          Game_Date           W_or_L
##  Length:1230        Length:1230        Length:1230        Length:1230
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       MIN             PTS             FGM             FGA
##  Min.   :240.0   Min.   : 77.0   Min.   :26.00   Min.   : 64.00
##  1st Qu.:240.0   1st Qu.:104.0   1st Qu.:38.00   1st Qu.: 85.00
##  Median :240.0   Median :112.0   Median :41.00   Median : 89.00
##  Mean   :241.6   Mean   :112.6   Mean   :41.55   Mean   : 89.23
##  3rd Qu.:240.0   3rd Qu.:121.0   3rd Qu.:45.00   3rd Qu.: 93.00
##  Max.   :340.0   Max.   :161.0   Max.   :61.00   Max.   :123.00
##    FG_PERCENT      THREE_PTM       THREE_PTA     THREE_PTPERCENT
##  Min.   :30.60   Min.   : 2.00   Min.   :14.00   Min.   :11.90
##  1st Qu.:43.00   1st Qu.: 9.00   1st Qu.:27.00   1st Qu.:30.00
##  Median :46.50   Median :11.00   Median :32.00   Median :35.70
##  Mean   :46.67   Mean   :11.52   Mean   :32.11   Mean   :35.92
##  3rd Qu.:50.00   3rd Qu.:14.00   3rd Qu.:37.00   3rd Qu.:41.65
##  Max.   :62.40   Max.   :27.00   Max.   :70.00   Max.   :84.20
##       FTM             FTA          FT_PERCENT          OREB
##  Min.   : 2.00   Min.   : 5.00   Min.   : 37.50   Min.   : 2.00
##  1st Qu.:13.00   1st Qu.:18.00   1st Qu.: 70.40   1st Qu.: 8.00
##  Median :18.00   Median :23.00   Median : 76.90   Median :10.00
##  Mean   :17.95   Mean   :23.40   Mean   : 76.68   Mean   :10.48
```

```
##  3rd Qu.:22.00   3rd Qu.:28.75   3rd Qu.: 84.00   3rd Qu.:13.00
##  Max.   :44.00   Max.   :51.00   Max.   :100.00   Max.   :25.00
##       DREB            REB             AST             STL
##  Min.   :18.00   Min.   :22.00   Min.   :10.00   Min.   : 1.000
##  1st Qu.:32.00   1st Qu.:41.00   1st Qu.:21.00   1st Qu.: 5.000
##  Median :35.00   Median :46.00   Median :25.00   Median : 7.000
##  Mean   :35.34   Mean   :45.82   Mean   :25.14   Mean   : 7.606
##  3rd Qu.:39.00   3rd Qu.:50.00   3rd Qu.:28.75   3rd Qu.:10.000
##  Max.   :53.00   Max.   :70.00   Max.   :42.00   Max.   :20.000
##       BLK              TOV              PF          PLUS/MINUS
##  Min.   : 0.000   Min.   : 3.00   Min.   : 9.00   Min.   :-56.000
##  1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.00   1st Qu.: -7.000
##  Median : 5.000   Median :14.00   Median :21.00   Median :  4.000
##  Mean   : 5.073   Mean   :14.04   Mean   :20.71   Mean   :  2.724
##  3rd Qu.: 7.000   3rd Qu.:17.00   3rd Qu.:23.00   3rd Qu.: 12.000
##  Max.   :19.000   Max.   :27.00   Max.   :38.00   Max.   : 50.000
##     Status
##  Length:1230
##  Class :character
##  Mode  :character
##
##
##
```

```
Away_games
```

```
## # A tibble: 1,230 x 25
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 GSW   GSW @ M~ 04/10/20~ L        240   117    46    92       50          13
##  2 ORL   ORL @ C~ 04/10/20~ W        240   122    48    88       54.5        11
##  3 MIN   MIN @ D~ 04/10/20~ L        240    95    39    91       42.9        13
##  4 OKC   OKC @ M~ 04/10/20~ W        240   127    48    99       48.5        23
##  5 IND   IND @ A~ 04/10/20~ W        240   135    45    98       45.9        12
##  6 SAC   SAC @ P~ 04/10/20~ L        240   131    50    96       52.1        18
##  7 UTA   UTA @ L~ 04/10/20~ L        265   137    47   106       44.3        14
##  8 MIA   MIA @ B~ 04/10/20~ L        240    94    38    98       38.8         8
##  9 DET   DET @ N~ 04/10/20~ W        240   115    41    85       48.2        14
## 10 DAL   DAL @ S~ 04/10/20~ L        240    94    37    91       40.7        11
## # ... with 1,220 more rows, and 15 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>
```

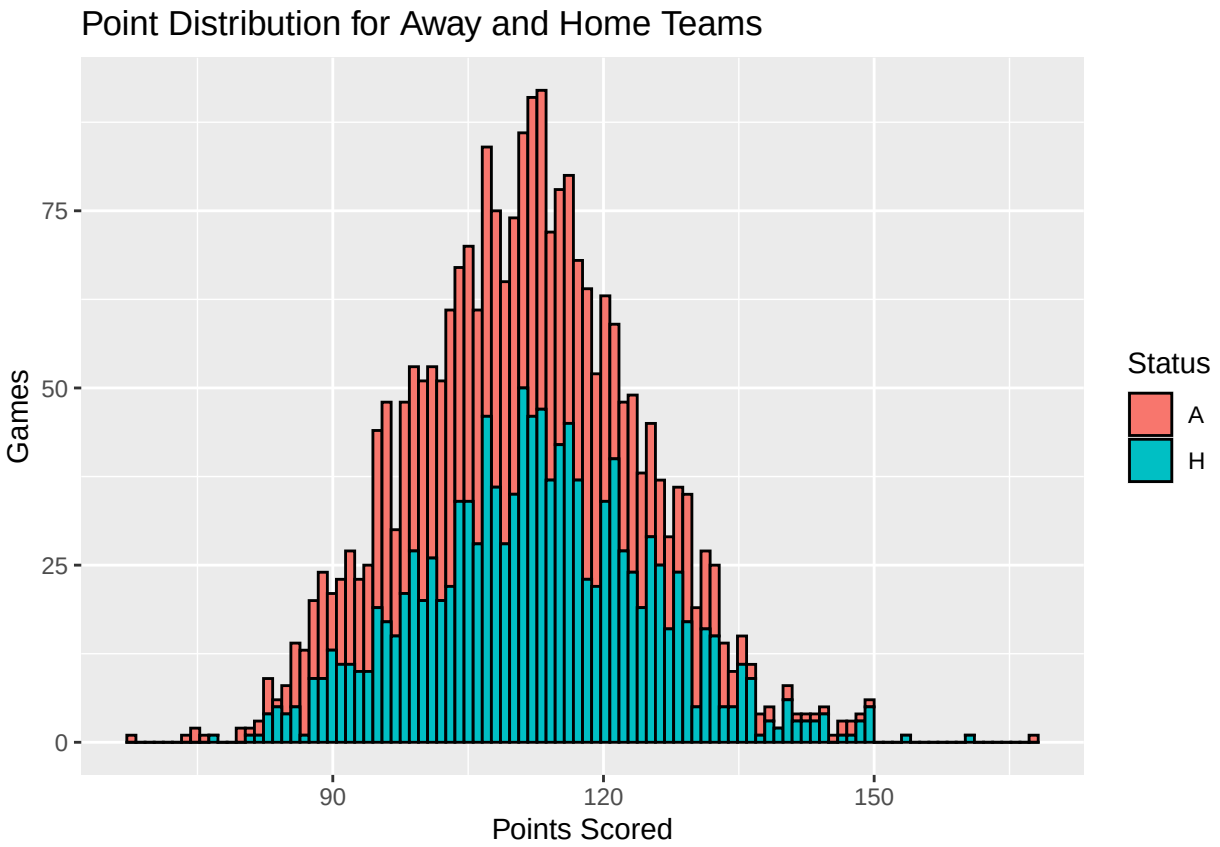
```
summary(Away_games)
```

```
##      Team             Match_Up          Game_Date            W_or_L
##  Length:1230        Length:1230        Length:1230        Length:1230
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##       MIN             PTS             FGM             FGA
##  Min.   :240.0   Min.   : 68.0   Min.   :25.00   Min.   : 65.00
##  1st Qu.:240.0   1st Qu.:101.0   1st Qu.:37.00   1st Qu.: 85.00
##  Median :240.0   Median :110.0   Median :41.00   Median : 89.00
##  Mean   :241.6   Mean   :109.8   Mean   :40.61   Mean   : 89.19
##  3rd Qu.:240.0   3rd Qu.:118.0   3rd Qu.:44.00   3rd Qu.: 94.00
##  Max.   :340.0   Max.   :168.0   Max.   :58.00   Max.   :119.00
##    FG_PERCENT      THREE_PTM       THREE_PTA    THREE_PTPERCENT      FTM
##  Min.   :27.80   Min.   : 2.00   Min.   :12.0   Min.   :11.50   Min.   : 3.00
##  1st Qu.:42.10   1st Qu.: 9.00   1st Qu.:27.0   1st Qu.:29.40   1st Qu.:13.00
##  Median :45.60   Median :11.00   Median :32.0   Median :35.00   Median :17.00
##  Mean   :45.62   Mean   :11.21   Mean   :31.9   Mean   :35.13   Mean   :17.41
##  3rd Qu.:48.90   3rd Qu.:14.00   3rd Qu.:37.0   3rd Qu.:40.50   3rd Qu.:21.00
##  Max.   :64.90   Max.   :26.00   Max.   :61.0   Max.   :78.30   Max.   :41.00
##       FTA          FT_PERCENT          OREB            DREB
##  Min.   : 4.00   Min.   : 26.30   Min.   : 1.00   Min.   :19.0
##  1st Qu.:18.00   1st Qu.: 70.00   1st Qu.: 7.00   1st Qu.:30.0
##  Median :22.00   Median : 77.30   Median :10.00   Median :34.0
##  Mean   :22.75   Mean   : 76.74   Mean   :10.22   Mean   :34.3
##  3rd Qu.:28.00   3rd Qu.: 84.00   3rd Qu.:12.00   3rd Qu.:38.0
##  Max.   :54.00   Max.   :100.00   Max.   :26.00   Max.   :55.0
##       REB             AST             STL             BLK
##  Min.   :24.00   Min.   :10.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:40.00   1st Qu.:20.00   1st Qu.: 6.000   1st Qu.: 3.000
##  Median :44.00   Median :24.00   Median : 8.000   Median : 5.000
##  Mean   :44.51   Mean   :24.03   Mean   : 7.662   Mean   : 4.833
##  3rd Qu.:49.00   3rd Qu.:27.00   3rd Qu.: 9.000   3rd Qu.: 6.000
##  Max.   :71.00   Max.   :41.00   Max.   :19.000   Max.   :14.000
##       TOV              PF         PLUS/MINUS        Status
##  Min.   : 3.00   Min.   : 9.0   Min.   :-50.000   Length:1230
##  1st Qu.:11.00   1st Qu.:18.0   1st Qu.:-12.000   Class :character
##  Median :14.00   Median :21.0   Median : -4.000   Mode  :character
##  Mean   :14.12   Mean   :21.1   Mean   : -2.724
##  3rd Qu.:17.00   3rd Qu.:24.0   3rd Qu.:  7.000
##  Max.   :29.00   Max.   :34.0   Max.   : 56.000
```

Looking at the summaries of the away and home data there are some surprising observations that can be made. For example, I really expected the field goal percent to have a bigger discrepancy between home and away teams, but the averages between the two are actually quite similar. The home team had about a 1 percent advantage at 46.67 percent compared to away teams at 45.62. Also looking at three point percentage both values are around the 35 percent mark. This solidifies further that the league average is about 35 percent. Also another tentative take away from these two values is that it seems like no matter the location shooting percentages are gonna be really similar.

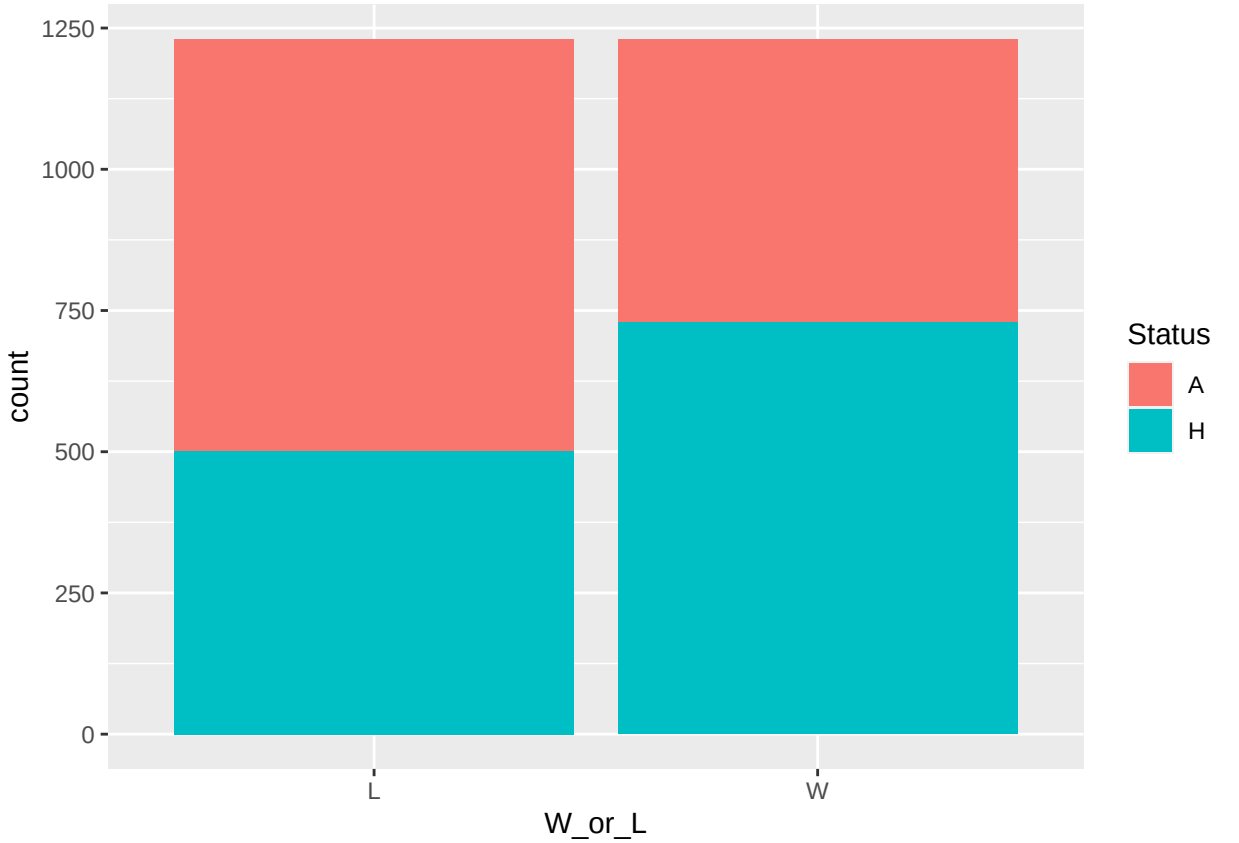
```
ggplot(data = NBA_box_stats) +
  geom_histogram(mapping = aes(x = PTS, fill = Status), bins = 100, color = 'black') +
  labs(title = "Point Distribution for Away and Home Teams", x = "Points Scored", y = "Games")
```



This histogram is an improved display of the point distributions for home and away teams. I now have them graphed on the same axis. Yet again it displays that home and away teams are relatively similar in their points scored. However, it also displays that the home team has many more games where they are scoring very high point totals. Also the away team has higher count of lower scoring games than the home team.

- Average Home Points Scored = 112.6
- Average Away Points Scored = 109.8

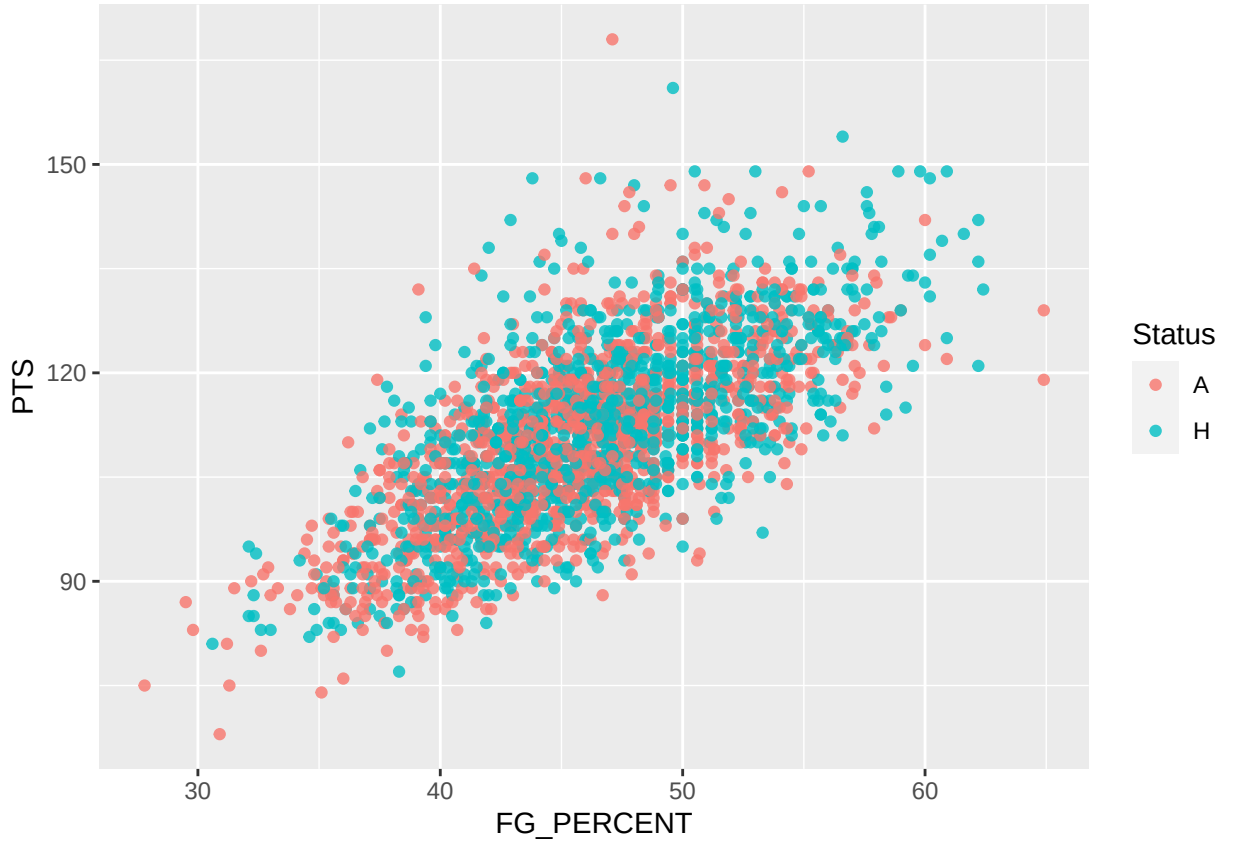
```
ggplot(data = NBA_box_stats) +
  geom_bar(mapping = aes(x = W_or_L, fill = Status))
```



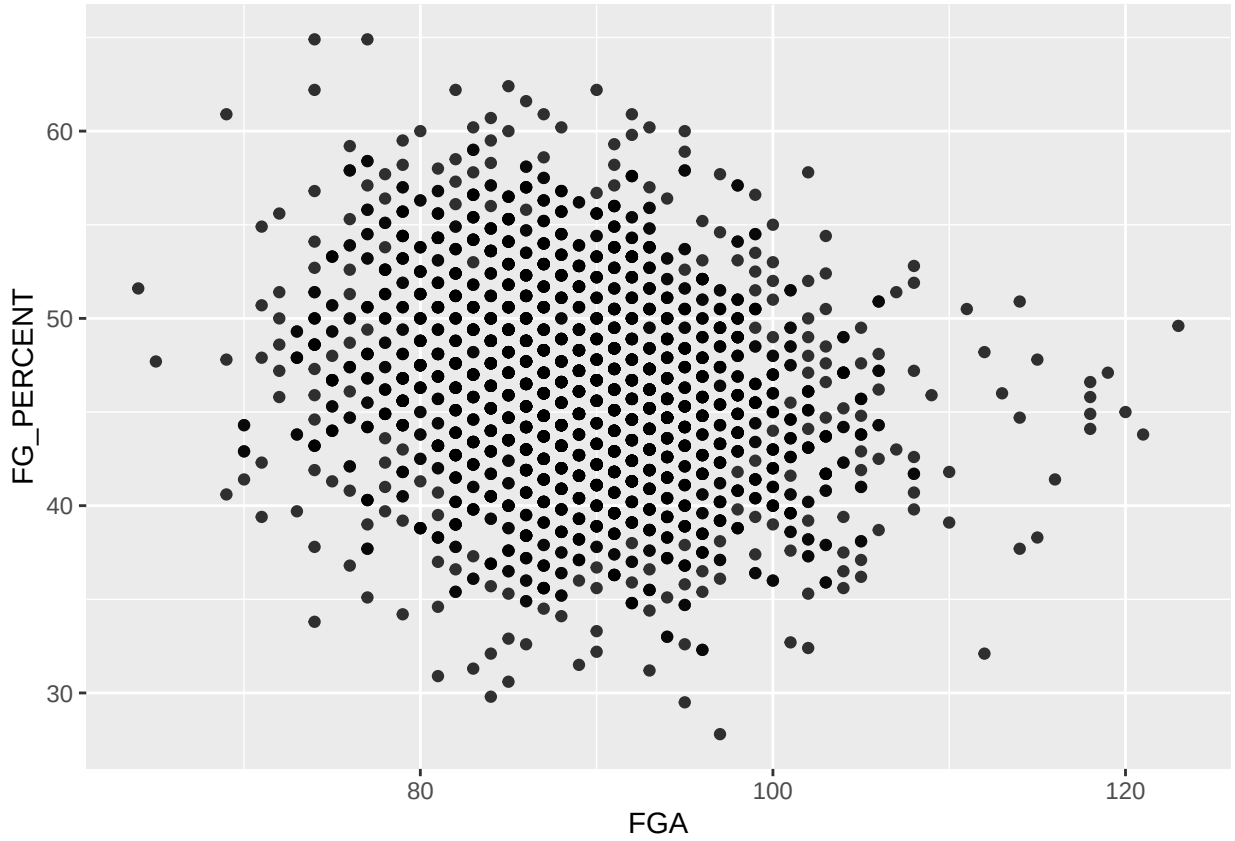
This graph is a simple bar chart displaying that home teams won more games than away teams. These results are not completely based in statistics; there are other untrackable factors that will play a part in the performance of certain team. For example, the amount of travel, back to back games, many games within a short stretch of time, and whether a certain team's roster is at full strength.

Model Refining

```
ggplot(data = NBA_box_stats) +
  geom_point(mapping = aes(x = FG_PERCENT, y = PTS, color = Status), alpha = .8)
```




```
ggplot(data = NBA_box_stats) +
  geom_point(mapping = aes(x = FGA, y = FG_PERCENT), alpha = .8)
```



The first graph displays field goal percentage vs points scored; visually there is definitely an upwards trend between the two variables. The second graph is a plot between field goal attempts and field goal percentage; there is not a very clear correlation between the two, but it is important to note that after the 100 fga mark there are not many values that are above 50 percent in terms of the shooting percentage. Also the really high field goal percentage values occured mainly when a team took less than about 100 fgas.

To try and improve the performance of my previous model I decided to try and also incorporate field goal percentage.

```
#Partitioning my test set for my model
#Data split 60 training 20 validation and 20 for testing
leftover_rows <- as.vector(createDataPartition(NBA_box_stats$PTS, p = 0.8, list = FALSE))
test_set <- NBA_box_stats[-leftover_rows, ]
leftover <- NBA_box_stats[leftover_rows, ]
leftover
```

```
## # A tibble: 1,969 x 25
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ B~ 01/09/20~ L        240   100    36    99       36.4         6
##  2 ATL   ATL @ B~ 12/16/20~ L        240   127    47    85       55.3        14
##  3 ATL   ATL @ C~ 11/06/20~ L        240   102    41    85       48.2         6
##  4 ATL   ATL @ C~ 03/03/20~ W        240   123    42    89       47.2        21
##  5 ATL   ATL @ C~ 10/21/20~ W        240   133    48    98       49          22
##  6 ATL   ATL @ C~ 10/30/20~ L        240   114    44    82       53.7        15
##  7 ATL   ATL @ D~ 12/12/20~ L        240   107    43    93       46.2        11
##  8 ATL   ATL @ D~ 11/15/20~ L        240    93    32    92       34.8         9
```

```
##  9 ATL   ATL @ H~ 02/25/20~ L        240   111    39    80       48.8        17
## 10 ATL   ATL @ I~ 11/17/20~ L        240    89    31    87       35.6         8
## # ... with 1,959 more rows, and 15 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>
```

```
#Creating the training and validation sets
training_rows <- as.vector(createDataPartition(leftover$PTS, p = 0.75, list = FALSE))
validate_rows <- leftover[-training_rows, ]
training <- leftover[training_rows, ]
training
```

```
## # A tibble: 1,478 x 25
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ B~ 01/09/20~ L        240   100    36    99       36.4         6
##  2 ATL   ATL @ B~ 12/16/20~ L        240   127    47    85       55.3        14
##  3 ATL   ATL @ C~ 11/06/20~ L        240   102    41    85       48.2         6
##  4 ATL   ATL @ C~ 10/21/20~ W        240   133    48    98       49          22
##  5 ATL   ATL @ D~ 12/12/20~ L        240   107    43    93       46.2        11
##  6 ATL   ATL @ D~ 11/15/20~ L        240    93    32    92       34.8         9
##  7 ATL   ATL @ H~ 02/25/20~ L        240   111    39    80       48.8        17
##  8 ATL   ATL @ M~ 03/04/20~ L        240   113    38    91       41.8        17
##  9 ATL   ATL @ M~ 11/27/20~ W        240   115    38    80       47.5        12
## 10 ATL   ATL @ M~ 01/04/20~ L        240   112    37    85       43.5        14
## # ... with 1,468 more rows, and 15 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>
```

```
validate_rows
```

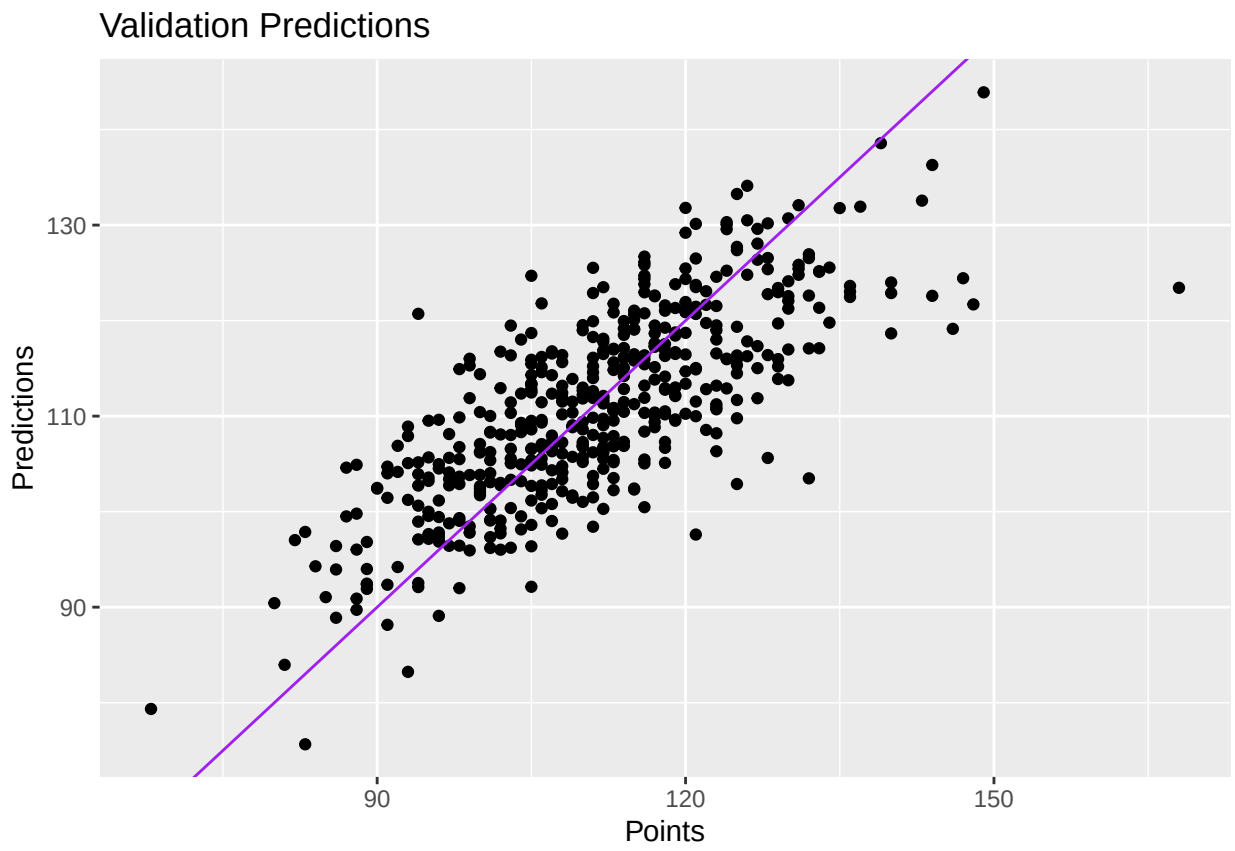
```
## # A tibble: 491 x 25
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ C~ 03/03/20~ W        240   123    42    89       47.2        21
##  2 ATL   ATL @ C~ 10/30/20~ L        240   114    44    82       53.7        15
##  3 ATL   ATL @ I~ 11/17/20~ L        240    89    31    87       35.6         8
##  4 ATL   ATL @ N~ 10/17/20~ L        240   107    41    90       45.6        10
##  5 ATL   ATL @ O~ 04/05/20~ L        240   113    41    94       43.6        10
##  6 ATL   ATL @ P~ 02/02/20~ W        240   118    43    84       51.2        14
##  7 ATL   ATL @ P~ 01/26/20~ L        240   111    42    92       45.7        12
##  8 ATL   ATL vs.~ 11/25/20~ W        240   124    50    94       53.2        16
##  9 ATL   ATL vs.~ 10/24/20~ W        240   111    37    91       40.7        15
## 10 ATL   ATL vs.~ 03/13/20~ W        240   132    52   103       50.5        17
## # ... with 481 more rows, and 15 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>
```

```r
#Training my model on the training set
model <- lm(PTS ~ FG_PERCENT + REB + AST, data = training)
```

```r
predictions <- add_predictions(validate_rows, model)
predictions
```

```
## # A tibble: 491 x 26
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ C~ 03/03/20~ W        240   123    42    89       47.2        21
##  2 ATL   ATL @ C~ 10/30/20~ L        240   114    44    82       53.7        15
##  3 ATL   ATL @ I~ 11/17/20~ L        240    89    31    87       35.6         8
##  4 ATL   ATL @ N~ 10/17/20~ L        240   107    41    90       45.6        10
##  5 ATL   ATL @ O~ 04/05/20~ L        240   113    41    94       43.6        10
##  6 ATL   ATL @ P~ 02/02/20~ W        240   118    43    84       51.2        14
##  7 ATL   ATL @ P~ 01/26/20~ L        240   111    42    92       45.7        12
##  8 ATL   ATL vs.~ 11/25/20~ W        240   124    50    94       53.2        16
##  9 ATL   ATL vs.~ 10/24/20~ W        240   111    37    91       40.7        15
## 10 ATL   ATL vs.~ 03/13/20~ W        240   132    52   103       50.5        17
## # ... with 481 more rows, and 16 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>, pred <dbl>
```

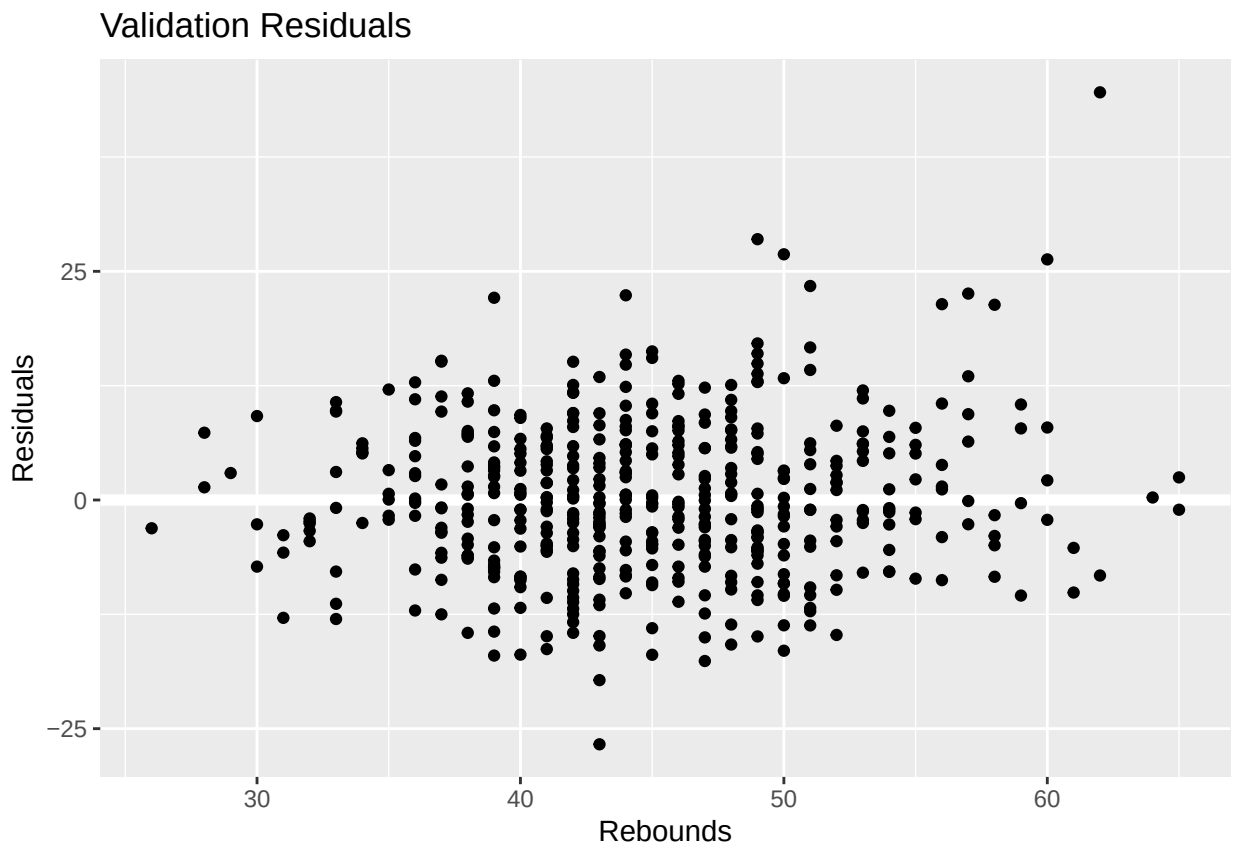
```r
ggplot(data = predictions, mapping = aes(x = PTS, y = pred)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "purple") +
  labs(y = "Predictions", x = "Points", title = "Validation Predictions")
```



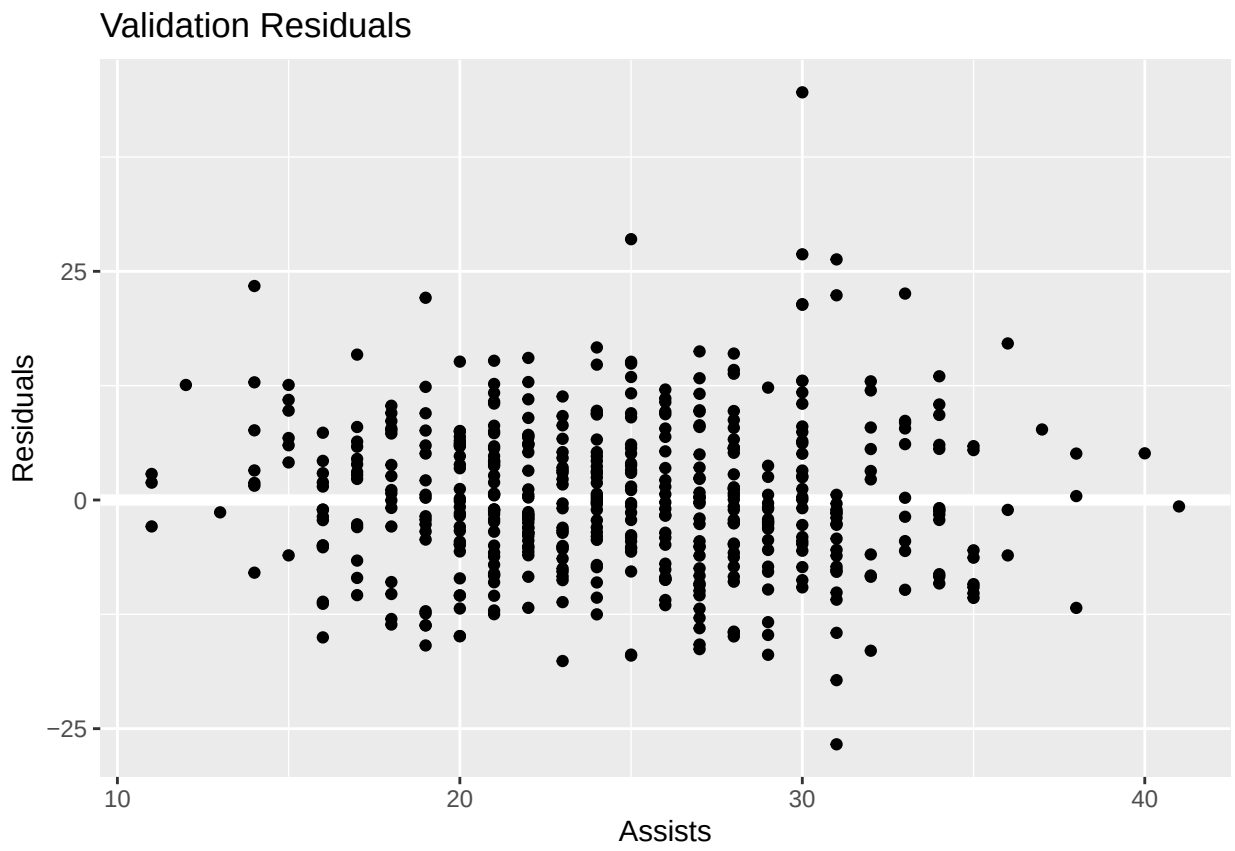

```
validate_resid <- add_residuals(validate_rows, model)
validate_resid
```

```
## # A tibble: 491 x 26
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ C~ 03/03/20~ W        240   123    42    89       47.2        21
##  2 ATL   ATL @ C~ 10/30/20~ L        240   114    44    82       53.7        15
##  3 ATL   ATL @ I~ 11/17/20~ L        240    89    31    87       35.6         8
##  4 ATL   ATL @ N~ 10/17/20~ L        240   107    41    90       45.6        10
##  5 ATL   ATL @ O~ 04/05/20~ L        240   113    41    94       43.6        10
##  6 ATL   ATL @ P~ 02/02/20~ W        240   118    43    84       51.2        14
##  7 ATL   ATL @ P~ 01/26/20~ L        240   111    42    92       45.7        12
##  8 ATL   ATL vs.~ 11/25/20~ W        240   124    50    94       53.2        16
##  9 ATL   ATL vs.~ 10/24/20~ W        240   111    37    91       40.7        15
## 10 ATL   ATL vs.~ 03/13/20~ W        240   132    52   103       50.5        17
## # ... with 481 more rows, and 16 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>, resid <dbl>
```

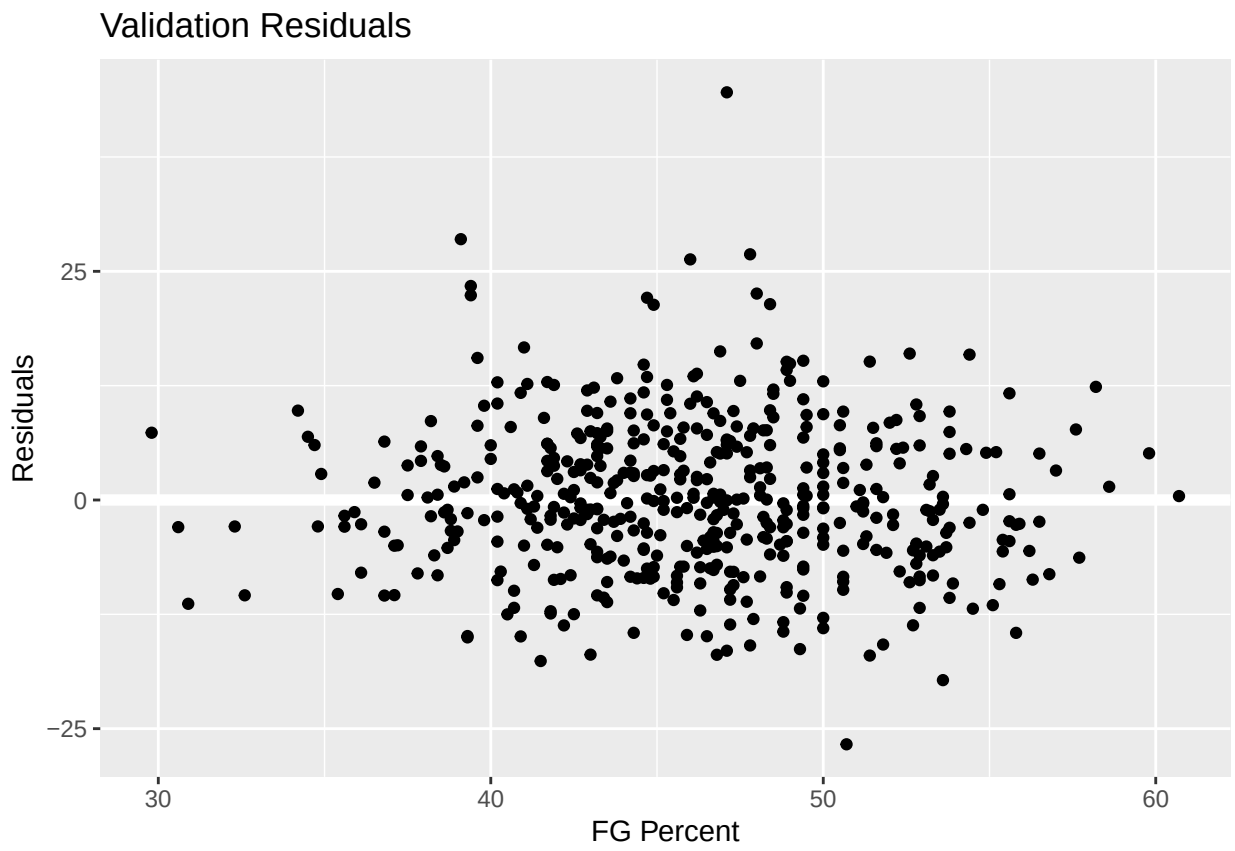
```
ggplot(validate_resid, aes(REB, resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "Rebounds", title = "Validation Residuals")
```




```
ggplot(validate_resid, aes(AST, resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "Assists", title = "Validation Residuals")
```




```
ggplot(validate_resid, aes(FG_PERCENT, resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "FG Percent", title = "Validation Residuals")
```




```
# Calculating goodness-of-fit measures for my model on the validation set
R2(predictions$pred, predictions$PTS)
```

```
## [1] 0.5825813
```

```
MAE(predictions$pred, predictions$PTS)
```

```
## [1] 6.524534
```

```
RMSE(predictions$pred, predictions$PTS)
```

```
## [1] 8.406944
```

Yet again my validation predictions look pretty solid. There are a good amount of points that are on the perfect prediction line, and there are not many extremely far off predictions; most of the predictions stay close to the line.

My residual plots for assists and rebounds are also decent. They are both pretty random without any noticeable patterns. However, my field goal percent residuals do not look as reliable. The plot still has a good amount of randomness, but there is definitely a lot of concentration between 50 and 40 percent.

```
model
```

```
##
## Call:
## lm(formula = PTS ~ FG_PERCENT + REB + AST, data = training)
##
## Coefficients:
## (Intercept)   FG_PERCENT          REB          AST
##     12.0086       1.3740       0.4380       0.6518
```

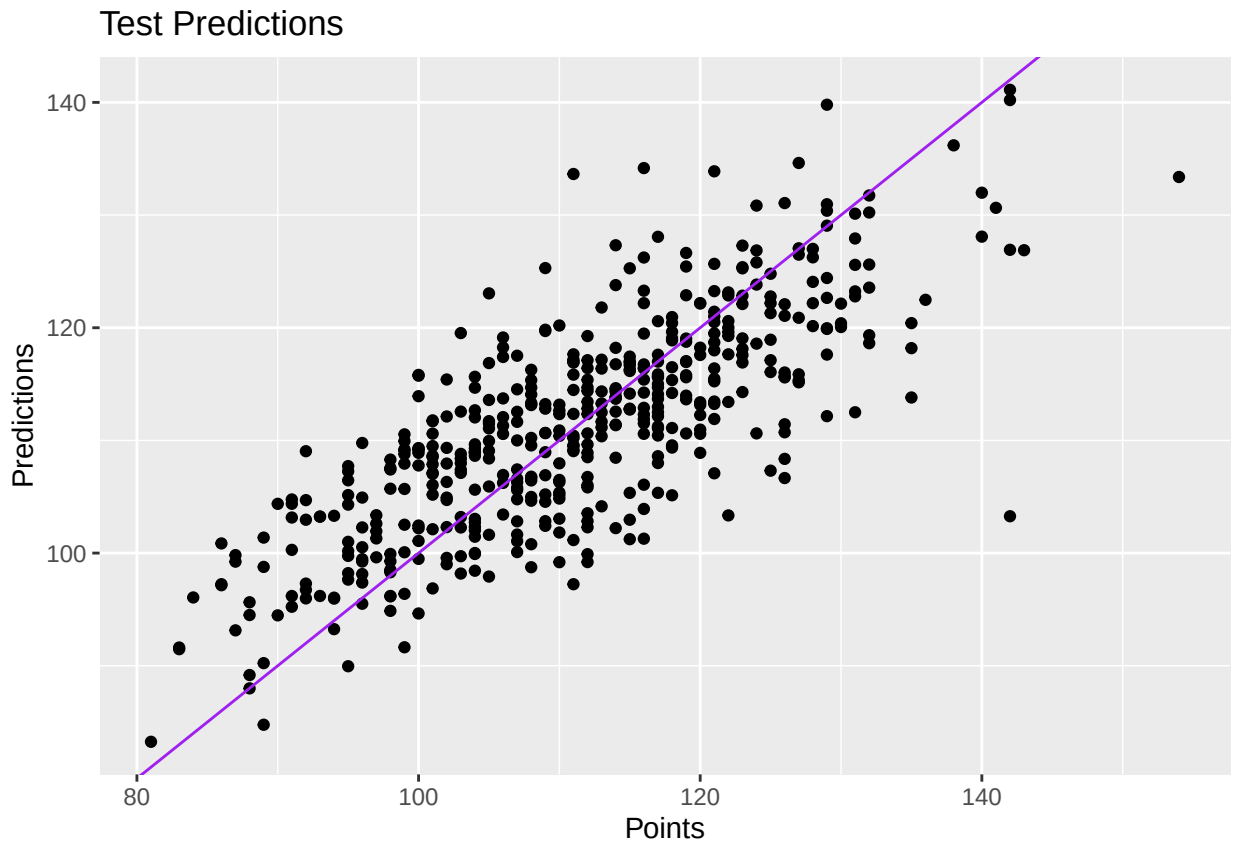
```
summary(model)
```

```
##
## Call:
## lm(formula = PTS ~ FG_PERCENT + REB + AST, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.537  -5.595  -0.659   5.143  36.211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.00862    2.49879   4.806  1.7e-06 ***
## FG_PERCENT   1.37401    0.04834  28.423  < 2e-16 ***
## REB          0.43795    0.03170  13.816  < 2e-16 ***
## AST          0.65181    0.05064  12.872  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.133 on 1474 degrees of freedom
## Multiple R-squared:  0.5915, Adjusted R-squared:  0.5907
## F-statistic: 711.4 on 3 and 1474 DF,  p-value: < 2.2e-16
```

```
predictions <- add_predictions(test_set, model)
predictions
```

```
## # A tibble: 491 x 26
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ B~ 03/16/20~ L        240   120    43   100       43          15
##  2 ATL   ATL @ B~ 12/14/20~ L        240   108    36    86       41.9        16
##  3 ATL   ATL @ C~ 11/28/20~ L        240    94    32    93       34.4        11
##  4 ATL   ATL @ C~ 01/23/20~ W        240   121    45    90       50          15
##  5 ATL   ATL @ D~ 12/23/20~ W        240    98    36    91       39.6         6
##  6 ATL   ATL @ G~ 11/13/20~ L        240   103    40    89       44.9        12
##  7 ATL   ATL @ I~ 12/31/20~ L        240   108    41    95       43.2        13
##  8 ATL   ATL @ L~ 01/28/20~ W        240   123    44    87       50.6        10
##  9 ATL   ATL @ L~ 11/11/20~ L        240   106    41    95       43.2        13
## 10 ATL   ATL @ M~ 10/19/20~ L        240   117    41    83       49.4        14
## # ... with 481 more rows, and 16 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>, pred <dbl>
```

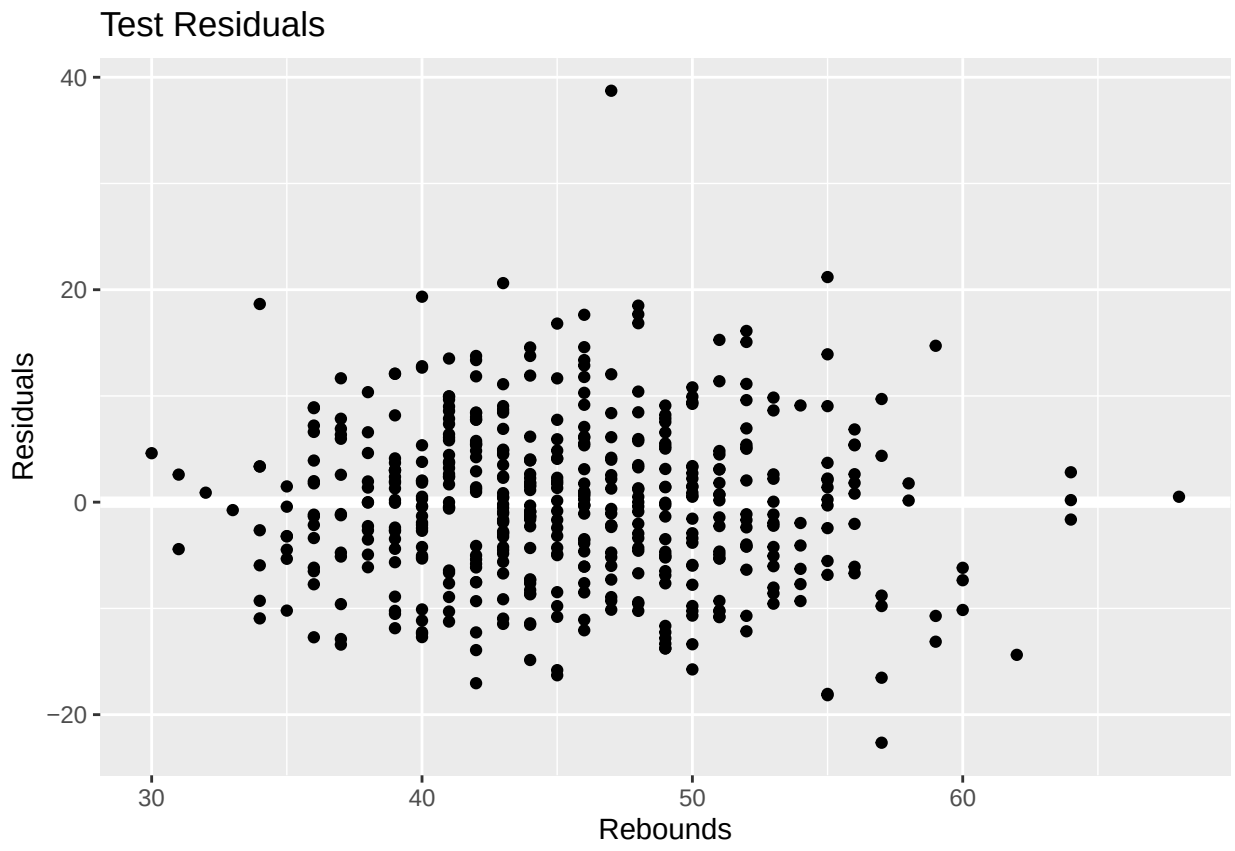
```
ggplot(data = predictions, mapping = aes(x = PTS, y = pred)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "purple") +
  labs(y = "Predictions", x = "Points", title = "Test Predictions")
```



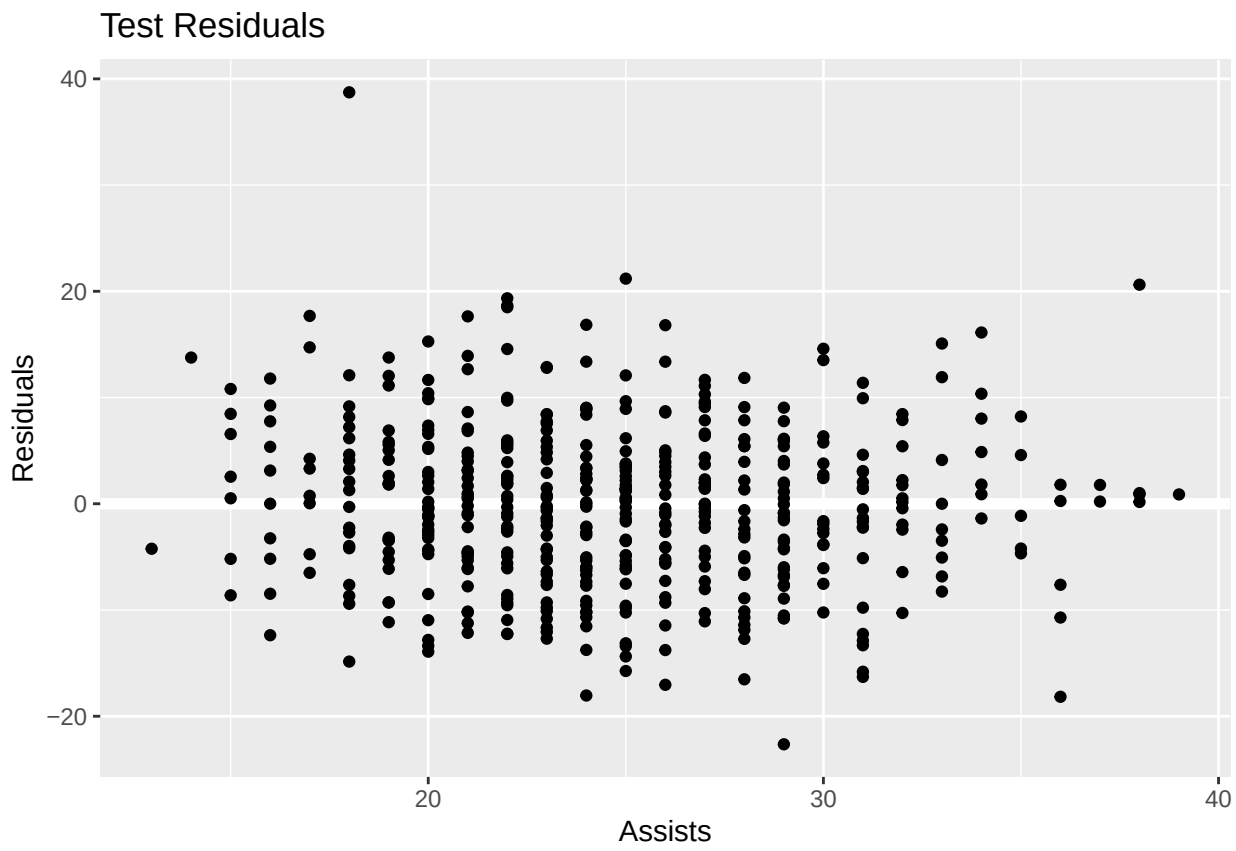

```
test_resids <- add_residuals(test_set, model)
test_resids
```

```
## # A tibble: 491 x 26
##    Team  Match_Up Game_Date W_or_L   MIN   PTS   FGM   FGA FG_PERCENT THREE_PTM
##    <chr> <chr>    <chr>     <chr>   <dbl> <dbl> <dbl> <dbl>      <dbl>     <dbl>
##  1 ATL   ATL @ B~ 03/16/20~ L         240   120    43   100       43          15
##  2 ATL   ATL @ B~ 12/14/20~ L         240   108    36    86       41.9        16
##  3 ATL   ATL @ C~ 11/28/20~ L         240    94    32    93       34.4        11
##  4 ATL   ATL @ C~ 01/23/20~ W         240   121    45    90       50          15
##  5 ATL   ATL @ D~ 12/23/20~ W         240    98    36    91       39.6         6
##  6 ATL   ATL @ G~ 11/13/20~ L         240   103    40    89       44.9        12
##  7 ATL   ATL @ I~ 12/31/20~ L         240   108    41    95       43.2        13
##  8 ATL   ATL @ L~ 01/28/20~ W         240   123    44    87       50.6        10
##  9 ATL   ATL @ L~ 11/11/20~ L         240   106    41    95       43.2        13
## 10 ATL   ATL @ M~ 10/19/20~ L         240   117    41    83       49.4        14
## # ... with 481 more rows, and 16 more variables: THREE_PTA <dbl>,
## #   THREE_PTPERCENT <dbl>, FTM <dbl>, FTA <dbl>, FT_PERCENT <dbl>, OREB <dbl>,
## #   DREB <dbl>, REB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>,
## #   PF <dbl>, `PLUS/MINUS` <dbl>, Status <chr>, resid <dbl>
```

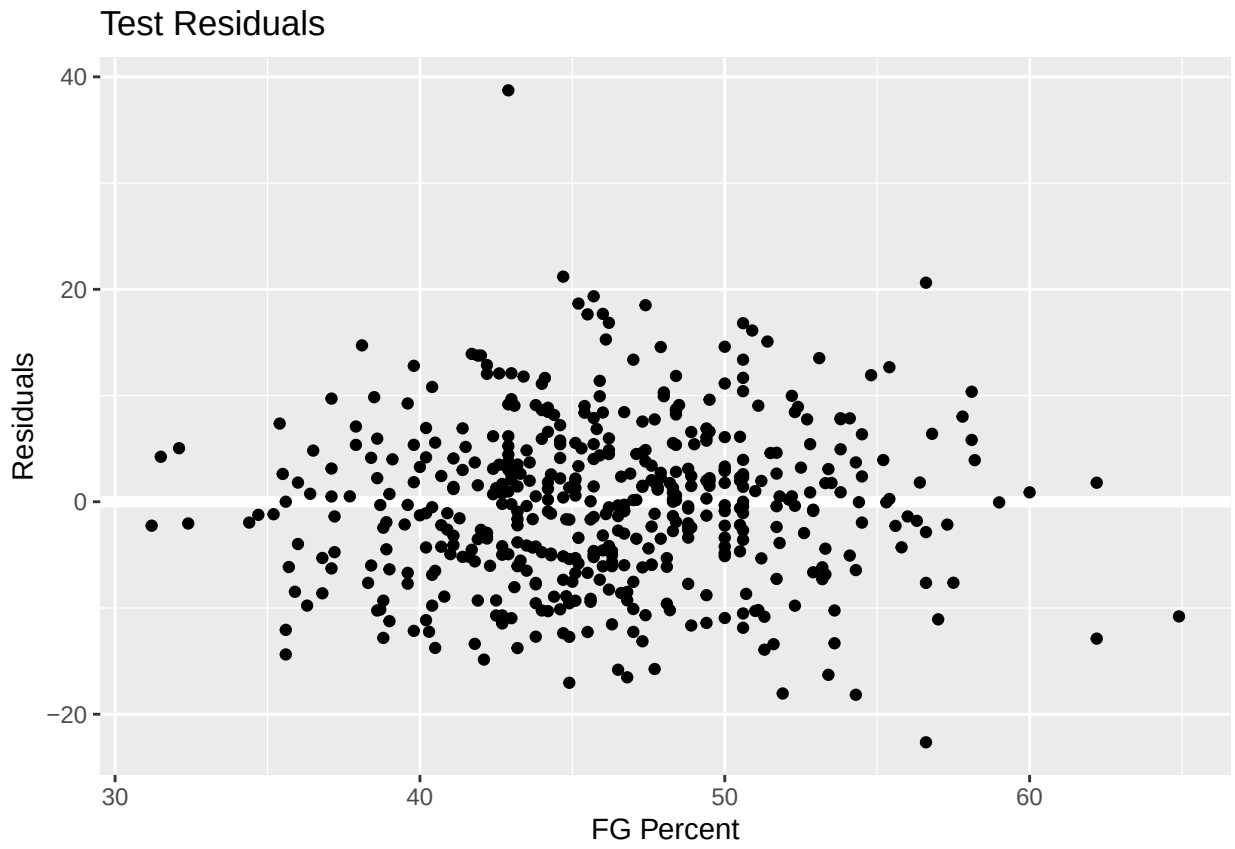
```
ggplot(data = test_resids, mapping = aes(x = REB, y = resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "Rebounds", title = "Test Residuals")
```




```
ggplot(data = test_resids, mapping = aes(x = AST, y = resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "Assists", title = "Test Residuals")
```




```
ggplot(data = test_resids, mapping = aes(x = FG_PERCENT, y = resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(y = "Residuals", x = "FG Percent", title = "Test Residuals")
```




```
# Calculating goodness-of-fit measures for my model on the test set
R2(predictions$pred, predictions$PTS)
```

```
## [1] 0.6046627
```

```
MAE(predictions$pred, predictions$PTS)
```

```
## [1] 5.944814
```

```
RMSE(predictions$pred, predictions$PTS)
```

```
## [1] 7.60912
```

Observations and Ethics discussion

Again my model visually seems like it does a decent job at predicting points scored. Analyzing my training and testing results there is not anything that looks concerning. Both my prediction results and test results display the same trends, and there is not any massive discrepancies between the visual plots.

My R^2, MAE, RMSE values are also very similar between testing and training. Between the previous iteration and this iteration my R^2 value is much nicer. In testing it went from about .35 to no .60. Also this time around my model in testing performed slightly better than in training. Based on this there does not seem to be any evidence of overfitting or extremely better performance with the test set.

For training:

- R^2 = 0.5825813
- MAE = 6.524534
- RMSE = 8.406944

For testing:

- R^2 = 0.6046627
- MAE = 5.944814
- RMSE = 7.60912

This model definitely performs better than my previous iteration that only used two variables to make predictions. However, there is still a considerable amount of uncertainty. One main concern is that field goal percent is not directly indicitive of points scored; for example, one team might have a 50 percent field percentage, but they could have taken considerably less shot attempts than another team that shot 40 percent from the floor. This is important because the team with less attempts and a higher percentage may have less field goals made which means less points. So in general this could also be misleading if someone is not paying attention.

Conclusions and discussion

Throughout the whole process of my project my main goals and data exploration took on a few changes. One main change was my exploration and use of the attendance figure. The attendance figure for each game proved to be not that useful to my project specifically. There is multiple reasons for this: a main reason is that each team's stadium has a maximum capacity, so more popular teams are consistently selling out their stadium which means that there is not much variance in their attendance numbers, and because a team is not scoring the same amount of points each game it makes it difficult to visualize any potential trends from that data. This is quite clearly visualized in my deliverable 1 plot of attendance vs home team points. It is easy to see a number of points that form practically horizontal lines throughout the graph. However, in a different context and project direction it could be very interesting. For example, in the future I could decide to use each team's game attendance figures in order to do analysis based on each team. In general I think my goal of analyzing away vs home team performance could use a lot of refining. Using the original base stats most likely is not detailed enough to come up with solid predictions; in order to have really solid analysis I would probably need to take a look at advance analytics which are much more detailed and provide a more in depth look to a team's performance. Overall, I think there is a lot of room for my project to grow. There is definitely much more data available that I could analyze in order to make a more solid in depth model.