deliverable2

Craig Le

12/3/2020

Dataset

My second dataset was compiled from the official NBA website, and it includes more than just home and away final scores. It includes each individual team's box scores stats for every game of the 2018-2019 season. It is 2460 rows, and each row is not the box stats from one single game. To clarify one single game separates each team's box stats into individual rows. For example, April 10, 2019 the Warriors (GSW) played against the Grizzlies (MEM); the first row is the Grizzlies stat line, and the second row is the Warriors stat line. Both of these rows are stats from the same game just separated by team. From my first dataset I want to take the attendance column and use those values for a prediction model.

```
#url <- c("https://www.nba.com/stats/teams/boxscores/?Season=2018-19&SeasonType=Regular%20Season")
#NBA_box_stats <- read_html(url) %>% html_nodes("main") %>%
# html_nodes("[class = 'nba-stat-table' ]") %>% #html_text()

#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
```

```
## # A tibble: 2,460 x 24
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
```

```
##  1 MEM   MEM vs. G~ 04/10/2019  W        240   132    48    98       49
##  2 GSW   GSW @ MEM  04/10/2019  L        240   117    46    92       50
##  3 CHA   CHA vs. O~ 04/10/2019  L        240   114    41    78       52.6
##  4 ORL   ORL @ CHA  04/10/2019  W        240   122    48    88       54.5
##  5 MIN   MIN @ DEN  04/10/2019  L        240    95    39    91       42.9
##  6 DEN   DEN vs. M~ 04/10/2019  W        240    99    39    87       44.8
##  7 MIL   MIL vs. O~ 04/10/2019  L        240   116    43   100       43
##  8 OKC   OKC @ MIL  04/10/2019  W        240   127    48    99       48.5
##  9 IND   IND @ ATL  04/10/2019  W        240   135    45    98       45.9
## 10 ATL   ATL vs. I~ 04/10/2019  L        240   134    43   103       41.7
## # ... with 2,450 more rows, and 15 more variables: THREE_PTM <dbl>,
## #   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>
```

```
summary(NBA_box_stats)
```

```
##      Team              Match Up          Game Date           W_or_L
##  Length:2460        Length:2460        Length:2460        Length:2460
##  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.   : 64.00
##  1st Qu.:240.0   1st Qu.:103.0   1st Qu.:38.00   1st Qu.: 85.00
##  Median :240.0   Median :111.0   Median :41.00   Median : 89.00
##  Mean   :241.6   Mean   :111.2   Mean   :41.08   Mean   : 89.21
##  3rd Qu.:240.0   3rd Qu.:120.0   3rd Qu.:44.00   3rd Qu.: 94.00
##  Max.   :340.0   Max.   :168.0   Max.   :61.00   Max.   :123.00
##    FG_PERCENT      THREE_PTM       THREE_PTA     THREE_PTPERCENT
##  Min.   :27.80   Min.   : 2.00   Min.   :12.00   Min.   :11.50
##  1st Qu.:42.60   1st Qu.: 9.00   1st Qu.:27.00   1st Qu.:29.60
##  Median :46.00   Median :11.00   Median :32.00   Median :35.30
##  Mean   :46.14   Mean   :11.36   Mean   :32.01   Mean   :35.52
##  3rd Qu.:49.50   3rd Qu.:14.00   3rd Qu.:37.00   3rd Qu.:40.92
##  Max.   :64.90   Max.   :27.00   Max.   :70.00   Max.   :84.20
##       FTM             FTA          FT_PERCENT          OREB
##  Min.   : 2.00   Min.   : 4.00   Min.   : 26.30   Min.   : 1.00
##  1st Qu.:13.00   1st Qu.:18.00   1st Qu.: 70.00   1st Qu.: 8.00
##  Median :17.00   Median :23.00   Median : 77.10   Median :10.00
##  Mean   :17.68   Mean   :23.07   Mean   : 76.71   Mean   :10.35
##  3rd Qu.:22.00   3rd Qu.:28.00   3rd Qu.: 84.00   3rd Qu.:13.00
##  Max.   :44.00   Max.   :54.00   Max.   :100.00   Max.   :26.00
##       DREB            REB             AST             STL
##  Min.   :18.00   Min.   :22.00   Min.   :10.00   Min.   : 0.000
##  1st Qu.:31.00   1st Qu.:41.00   1st Qu.:21.00   1st Qu.: 6.000
##  Median :35.00   Median :45.00   Median :24.00   Median : 7.000
##  Mean   :34.82   Mean   :45.17   Mean   :24.59   Mean   : 7.634
##  3rd Qu.:38.00   3rd Qu.:50.00   3rd Qu.:28.00   3rd Qu.: 9.000
##  Max.   :55.00   Max.   :71.00   Max.   :42.00   Max.   :20.000
##       BLK              TOV              PF         PLUS/MINUS
##  Min.   : 0.000   Min.   : 3.00   Min.   : 9.0   Min.   :-56
```

```
##  1st Qu.: 3.000   1st Qu.:11.00   1st Qu.:18.0   1st Qu.: -9
##  Median : 5.000   Median :14.00   Median :21.0   Median :  0
##  Mean   : 4.953   Mean   :14.08   Mean   :20.9   Mean   :  0
##  3rd Qu.: 6.000   3rd Qu.:17.00   3rd Qu.:24.0   3rd Qu.:  9
##  Max.   :19.000   Max.   :29.00   Max.   :38.0   Max.   : 56
```

It is really interesting looking at the summary of this table. From the summary I can really get a sense of how much basketball is sport with a lot of high and lot of low points. For example, in the summary of the 3 point field goal attempts the max value is 70 and the minimum value is 12. I looked through my data and found the Houston Rockets were responsible for the 70 attempts, and the Los Angeles Clippers had the lowest 3 point attempts. I also found that Houston was responsible for the top three most 3 pointers attempted that season. Almost every column features a large gap between the min and max values. Another example is the point totals. The minimum points was 68 and the max was 168. The max points came from a quadruple overtime game between the Chicago Bulls and Atlanta Hawks, so this max point value is definitely an outlier data point.

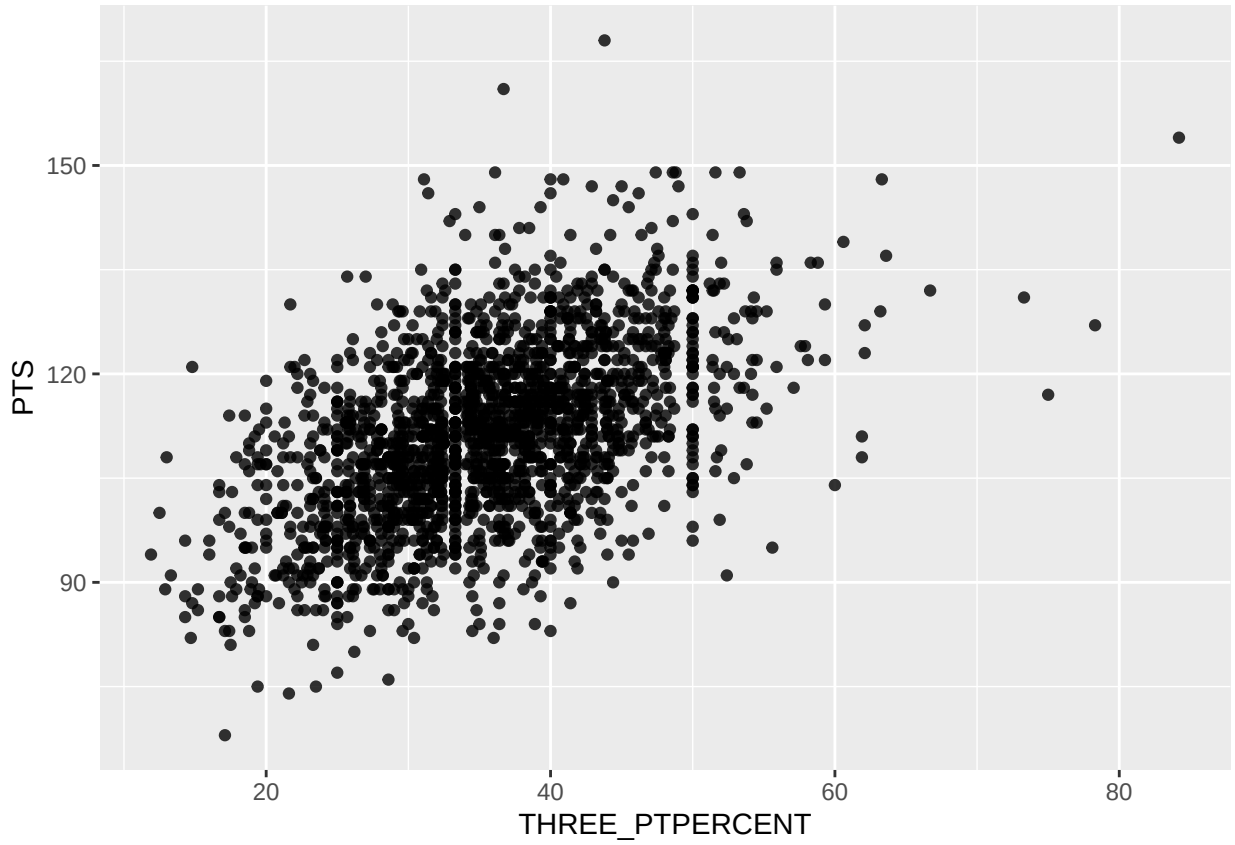
Model Planning and Building

Before I try to incorporate home court vs away factors, I want to do a general points prediction of all teams without home or away factors. I am using rebounds and assists as the variables for predicting points scored. Generally in basketball getting more rebounds means that a team is able to posses the ball more teams which leads to more opportunities to score, so based on that thought rebounds should be a solid variable to use in my model. In basketball an assist is counted when there is a pass made that directly leads to a basket getting scored, so numerically the more assists a team has than the more points they have. Also getting more assists in a game usually indicates that a team is executing their sets very well and also working really well as a team, as a result these un-trackable factors usually lead to more points being scored.

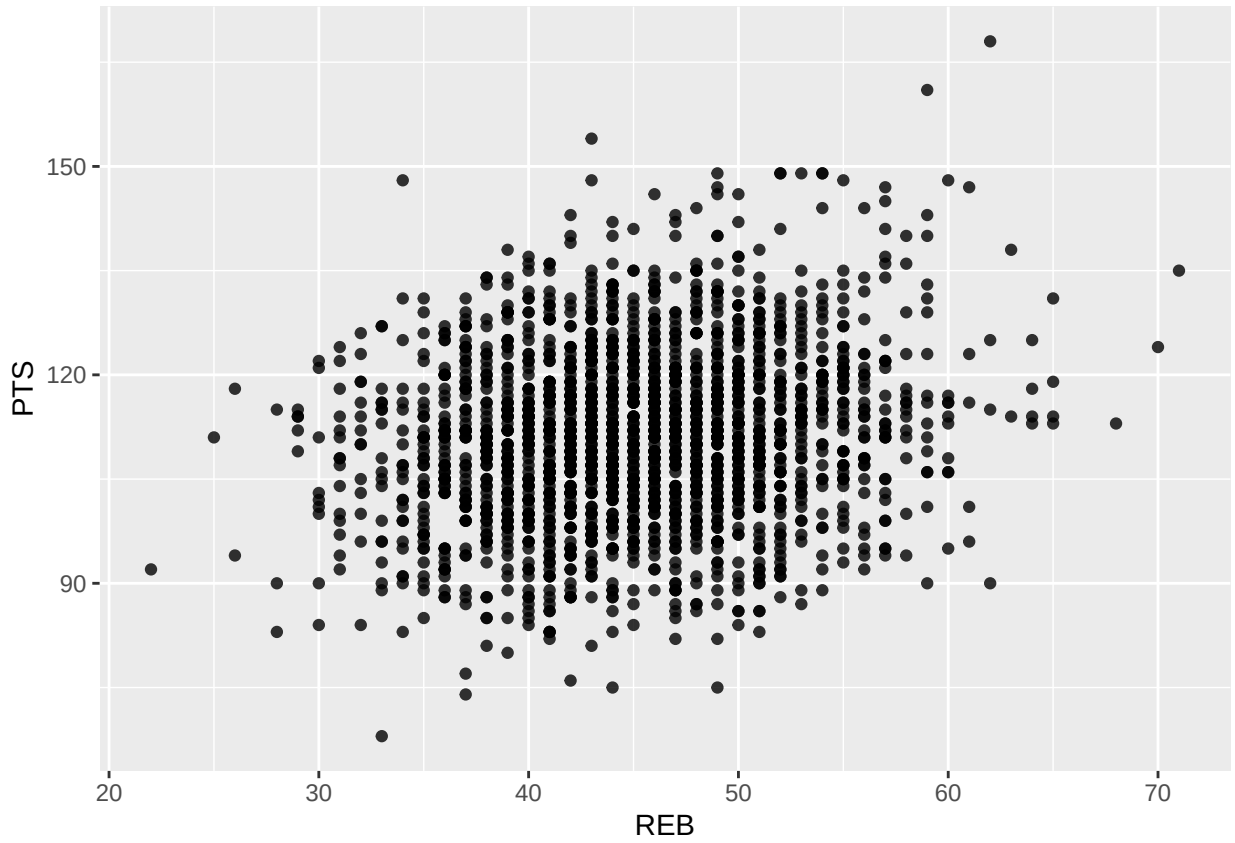
```
#Partitioning my test set for my model
#Data split 60 training 20 validation and 20 for testing
leftover_rows <- as.vector(createDataPartition(NBA_box_stats$PTS, p = 0.8, list = FALSE))
test_set <- NBA_box_stats[-leftover_rows, ]
leftover <- NBA_box_stats[leftover_rows, ]
leftover
```

```
## # A tibble: 1,969 x 24
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 MEM   MEM vs. G~ 04/10/2019  W        240   132    48    98       49
##  2 CHA   CHA vs. O~ 04/10/2019  L        240   114    41    78       52.6
##  3 MIN   MIN @ DEN  04/10/2019  L        240    95    39    91       42.9
##  4 DEN   DEN vs. M~ 04/10/2019  W        240    99    39    87       44.8
##  5 OKC   OKC @ MIL  04/10/2019  W        240   127    48    99       48.5
##  6 IND   IND @ ATL  04/10/2019  W        240   135    45    98       45.9
##  7 POR   POR vs. S~ 04/10/2019  W        240   136    53    91       58.2
##  8 UTA   UTA @ LAC  04/10/2019  L        265   137    47   106       44.3
##  9 BKN   BKN vs. M~ 04/10/2019  W        240   113    43   114       37.7
## 10 DET   DET @ NYK  04/10/2019  W        240   115    41    85       48.2
## # ... with 1,959 more rows, and 15 more variables: THREE_PTM <dbl>,
## #   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>
```

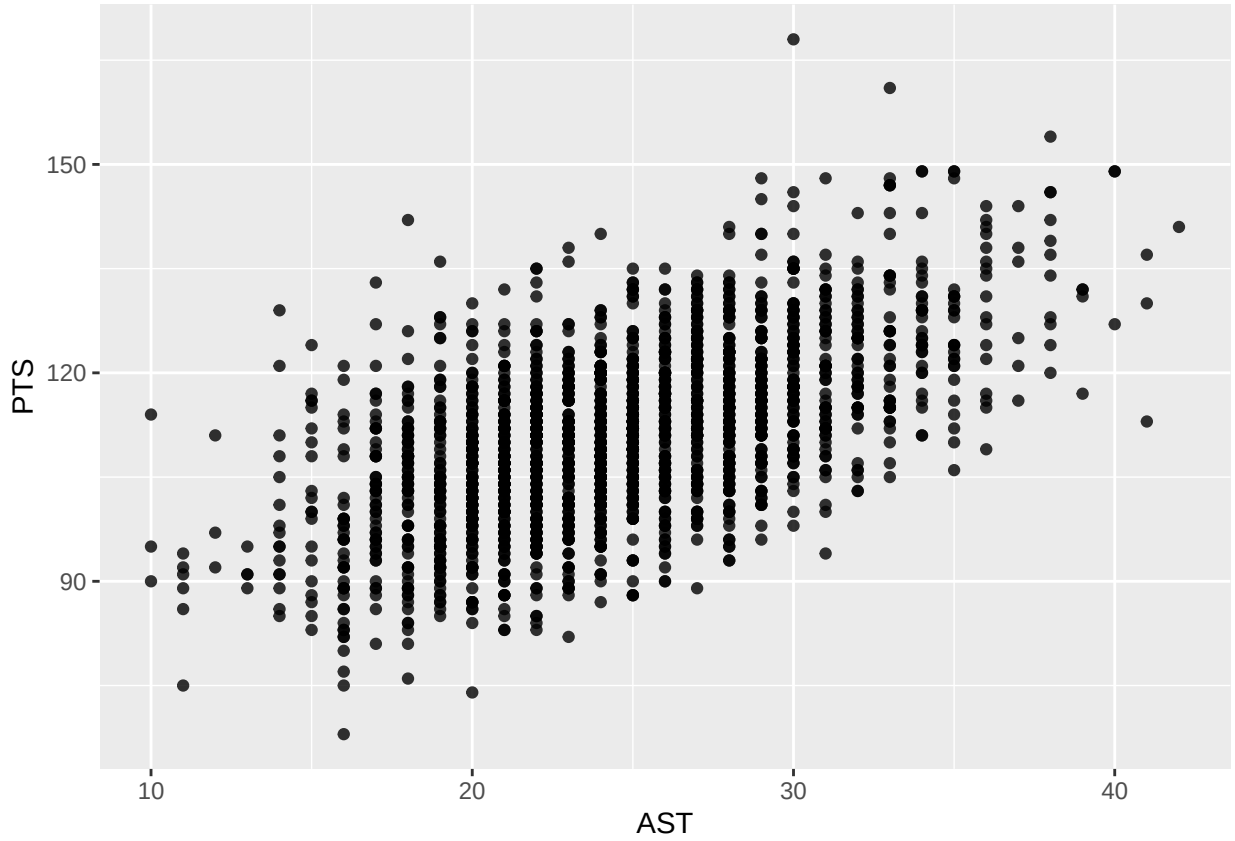
```r
#summary(leftover)

#Initial exploratory graphs to check my initial thinking
#and also to explore other options

ggplot(data = leftover) +
  geom_point(mapping = aes(x = THREE_PTPERCENT, y = PTS), alpha = .8)
```



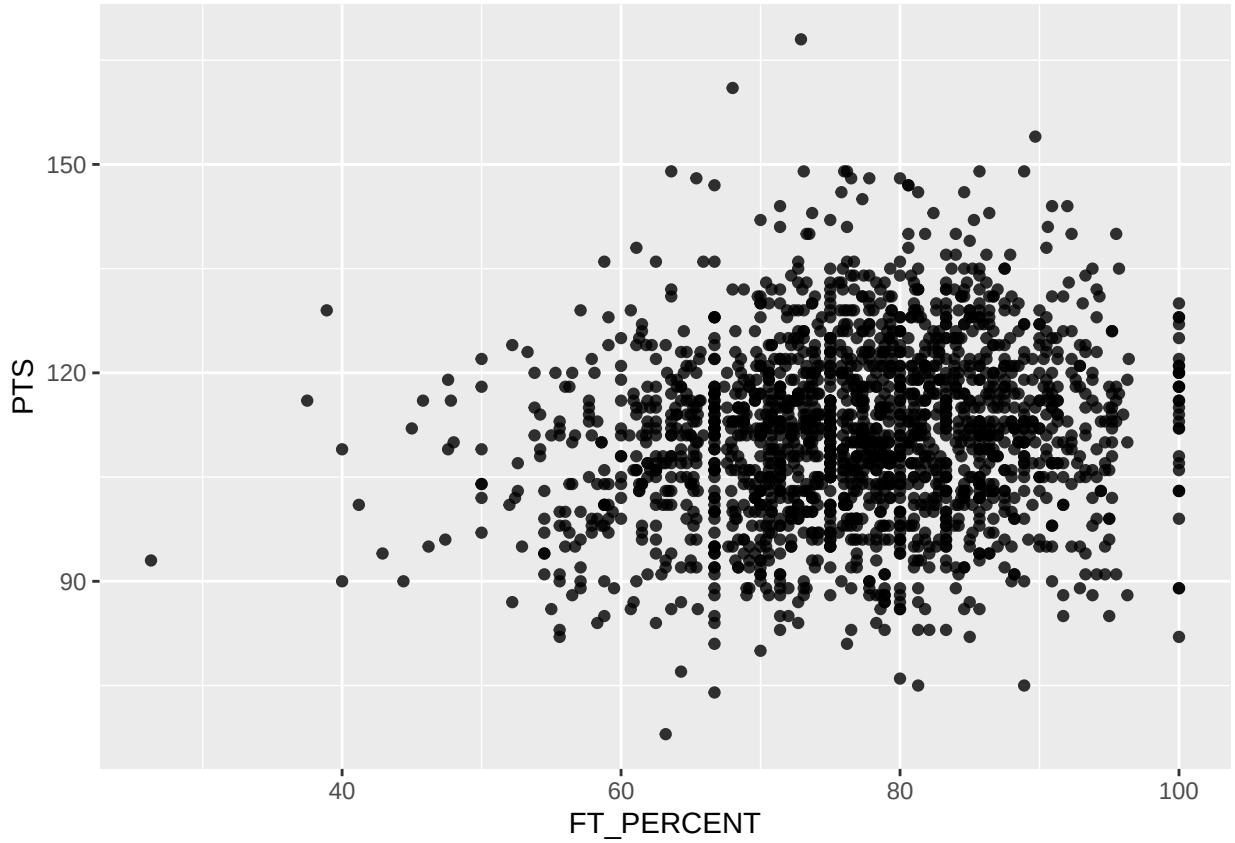

```r
ggplot(data = leftover) +
  geom_point(mapping = aes(x = REB, y = PTS), alpha = .8)
```



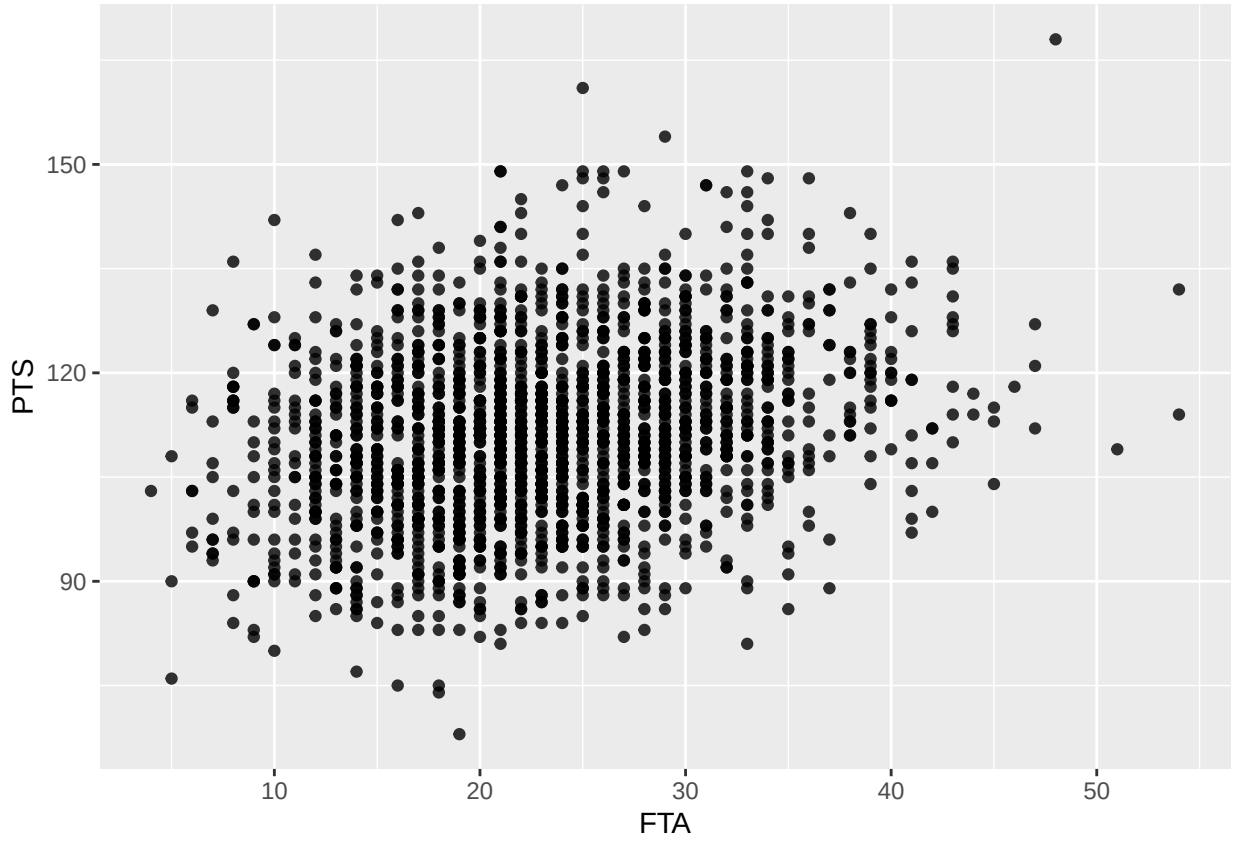

```
ggplot(data = leftover) +
  geom_point(mapping = aes(x = AST, y = PTS), alpha = .8)
```




```
ggplot(data = leftover) +
  geom_point(mapping = aes(x = FT_PERCENT, y = PTS), alpha = .8)
```




```
ggplot(data = leftover) +
  geom_point(mapping = aes(x = FTA, y = PTS), alpha = .8)
```




```
# The assists vs points graph has an expected increasing trend. Surprisingly the
# rebounds vs points did not have as strong of a relationship as I thought, but
# there is still an increasing trend. Free throw attempts and percentage were
# similar to the rebound graph. The three point percentage graph showed a pretty strong
# relationship to points, so maybe it could used as a variable in the future.

#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 24
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 MEM   MEM vs. G~ 04/10/2019  W        240   132    48    98       49
##  2 MIN   MIN @ DEN  04/10/2019  L        240    95    39    91       42.9
##  3 DEN   DEN vs. M~ 04/10/2019  W        240    99    39    87       44.8
##  4 IND   IND @ ATL  04/10/2019  W        240   135    45    98       45.9
##  5 POR   POR vs. S~ 04/10/2019  W        240   136    53    91       58.2
##  6 UTA   UTA @ LAC  04/10/2019  L        265   137    47   106       44.3
##  7 BKN   BKN vs. M~ 04/10/2019  W        240   113    43   114       37.7
##  8 DET   DET @ NYK  04/10/2019  W        240   115    41    85       48.2
##  9 NYK   NYK vs. D~ 04/10/2019  L        240    89    31    77       40.3
## 10 CHI   CHI @ PHI  04/10/2019  L        240   109    45    95       47.4
## # ... with 1,468 more rows, and 15 more variables: THREE_PTM <dbl>,
```

```
## #   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>
```

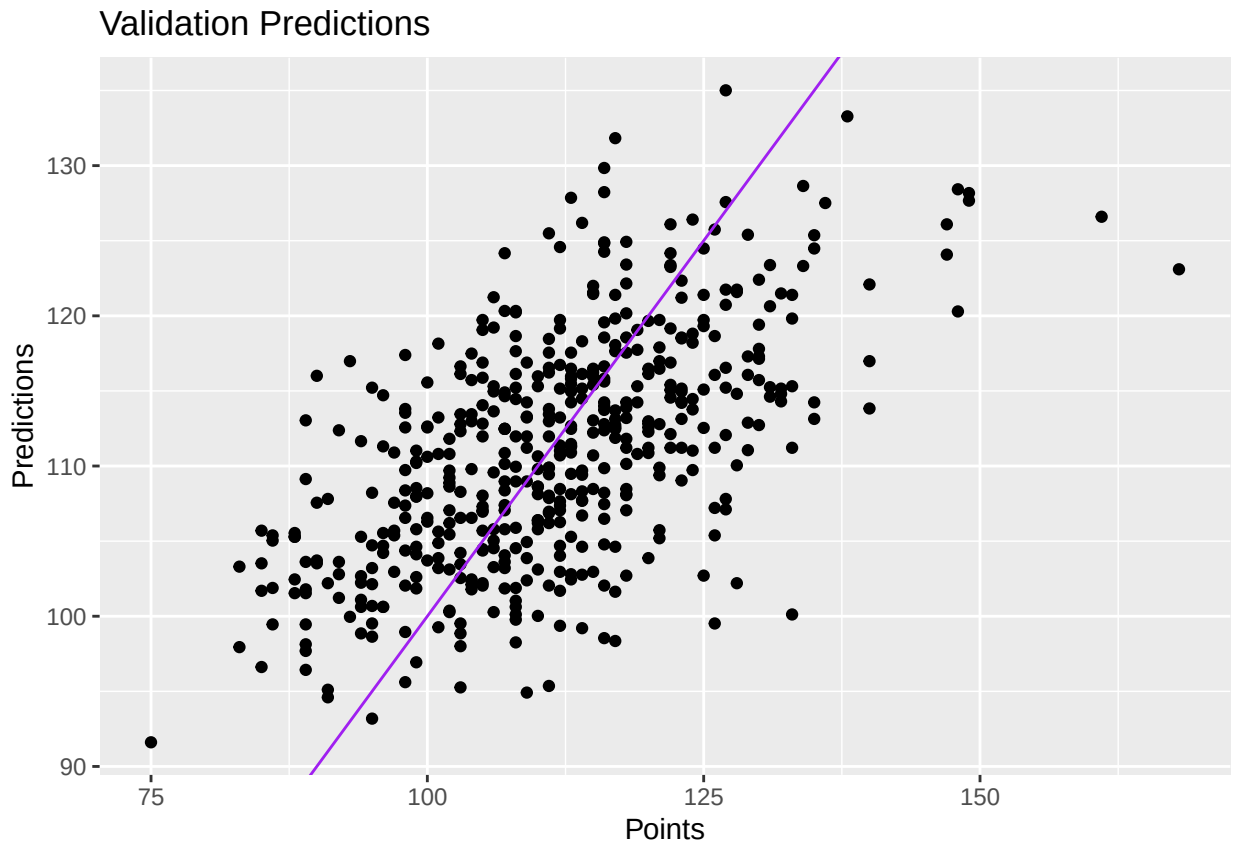
```
validate_rows
```

```
## # A tibble: 491 x 24
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 CHA   CHA vs. O~ 04/10/2019  L        240   114    41    78       52.6
##  2 OKC   OKC @ MIL  04/10/2019  W        240   127    48    99       48.5
##  3 PHI   PHI vs. C~ 04/10/2019  W        240   125    52    93       55.9
##  4 BOS   BOS @ WAS  04/09/2019  W        240   116    45    99       45.5
##  5 TOR   TOR @ MIN  04/09/2019  W        240   120    46    88       52.3
##  6 MIN   MIN vs. T~ 04/09/2019  L        240   100    38    91       41.8
##  7 POR   POR @ LAL  04/09/2019  W        240   104    37    90       41.1
##  8 OKC   OKC @ MIN  04/07/2019  W        240   132    48    92       52.2
##  9 ATL   ATL @ MIL  04/07/2019  L        240   107    40   100       40
## 10 BOS   BOS vs. O~ 04/07/2019  L        240   108    42    89       47.2
## # ... with 481 more rows, and 15 more variables: THREE_PTM <dbl>,
## #   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>
```

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

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

```
## # A tibble: 491 x 25
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 CHA   CHA vs. O~ 04/10/2019  L        240   114    41    78       52.6
##  2 OKC   OKC @ MIL  04/10/2019  W        240   127    48    99       48.5
##  3 PHI   PHI vs. C~ 04/10/2019  W        240   125    52    93       55.9
##  4 BOS   BOS @ WAS  04/09/2019  W        240   116    45    99       45.5
##  5 TOR   TOR @ MIN  04/09/2019  W        240   120    46    88       52.3
##  6 MIN   MIN vs. T~ 04/09/2019  L        240   100    38    91       41.8
##  7 POR   POR @ LAL  04/09/2019  W        240   104    37    90       41.1
##  8 OKC   OKC @ MIN  04/07/2019  W        240   132    48    92       52.2
##  9 ATL   ATL @ MIL  04/07/2019  L        240   107    40   100       40
## 10 BOS   BOS vs. O~ 04/07/2019  L        240   108    42    89       47.2
## # ... with 481 more rows, and 16 more variables: THREE_PTM <dbl>,
## #   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>, pred <dbl>
```

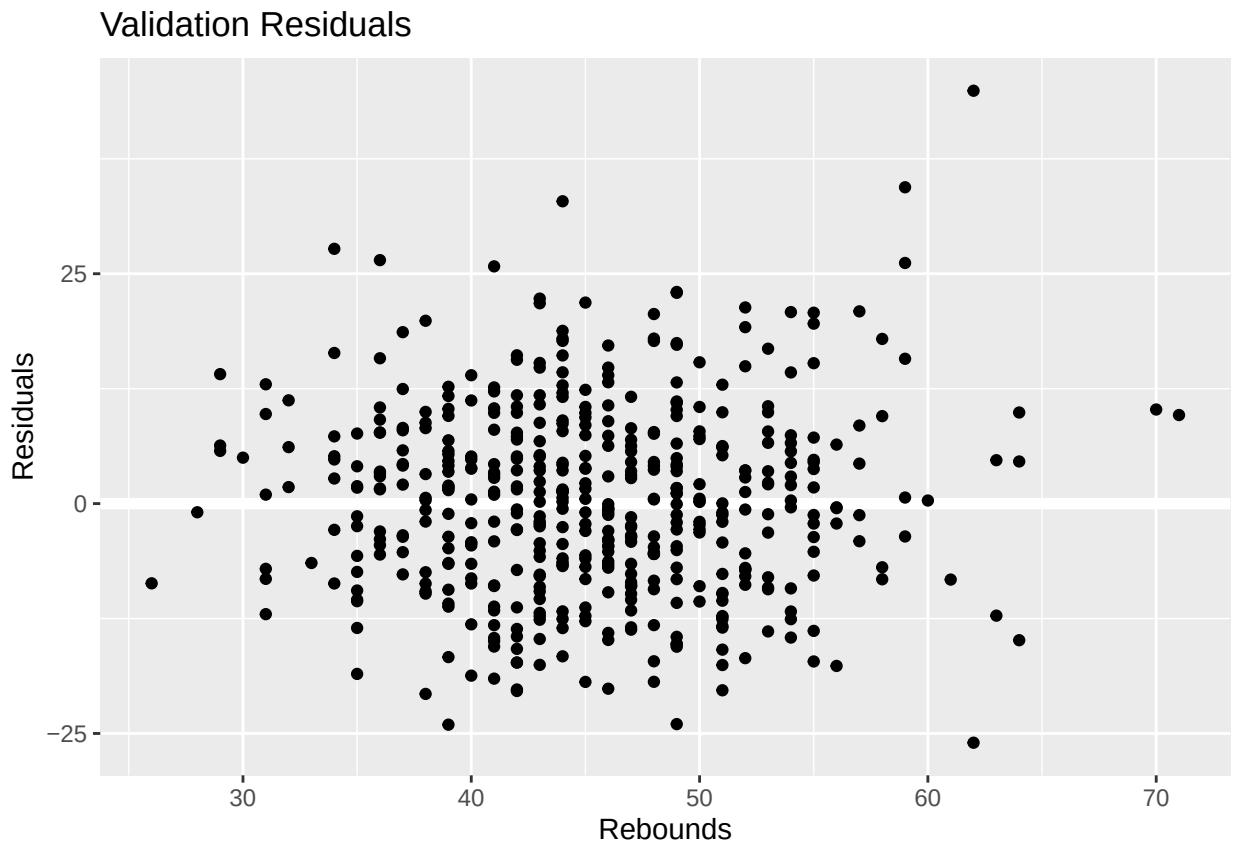
```
ggplot(data = predictions, mapping = aes(x = PTS, y = pred)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "purple") +
  labs(y = "Predictions", x = "Points", title = "Validation Predictions")
```



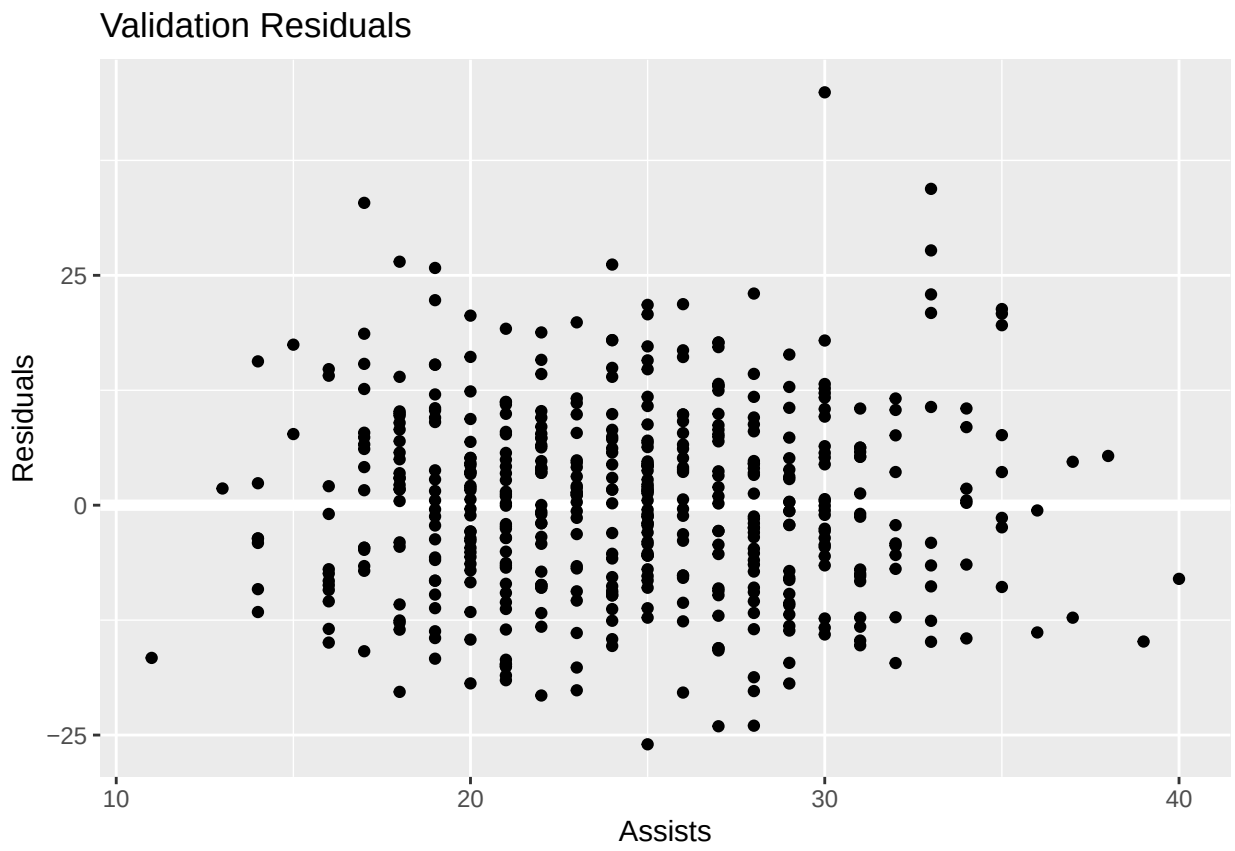

```
# At a glance my validation set predictions look decent. There are many points
# that are near the perfect prediction line, and there are points that fall directly on
# that line. Also there is a decently clear trend that follows the prediction line.
# However, there are also many clear outlier data values that can be seen on the graph.

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

```
## # A tibble: 491 x 25
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 CHA   CHA vs. O~ 04/10/2019  L        240   114    41    78       52.6
##  2 OKC   OKC @ MIL  04/10/2019  W        240   127    48    99       48.5
##  3 PHI   PHI vs. C~ 04/10/2019  W        240   125    52    93       55.9
##  4 BOS   BOS @ WAS  04/09/2019  W        240   116    45    99       45.5
##  5 TOR   TOR @ MIN  04/09/2019  W        240   120    46    88       52.3
##  6 MIN   MIN vs. T~ 04/09/2019  L        240   100    38    91       41.8
##  7 POR   POR @ LAL  04/09/2019  W        240   104    37    90       41.1
##  8 OKC   OKC @ MIN  04/07/2019  W        240   132    48    92       52.2
##  9 ATL   ATL @ MIL  04/07/2019  L        240   107    40   100       40
## 10 BOS   BOS vs. O~ 04/07/2019  L        240   108    42    89       47.2
## # ... with 481 more rows, and 16 more variables: THREE_PTM <dbl>,
## #   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>, 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")
```




```
# Both of my residual graphs seem pretty solid. There does not seem to be any noticeable
# trends for both the assists and rebounds. This indicates that my model did a solid job at
# removing patterns that might have existed.

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

```
## [1] 0.3624039
```

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

```
## [1] 8.266143
```

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

```
## [1] 10.37754
```

```
model
```

```
##
## Call:
## lm(formula = PTS ~ REB + AST, data = training)
##
```

```
## Coefficients:
## (Intercept)          REB          AST
##     64.9079       0.2522       1.4184
```

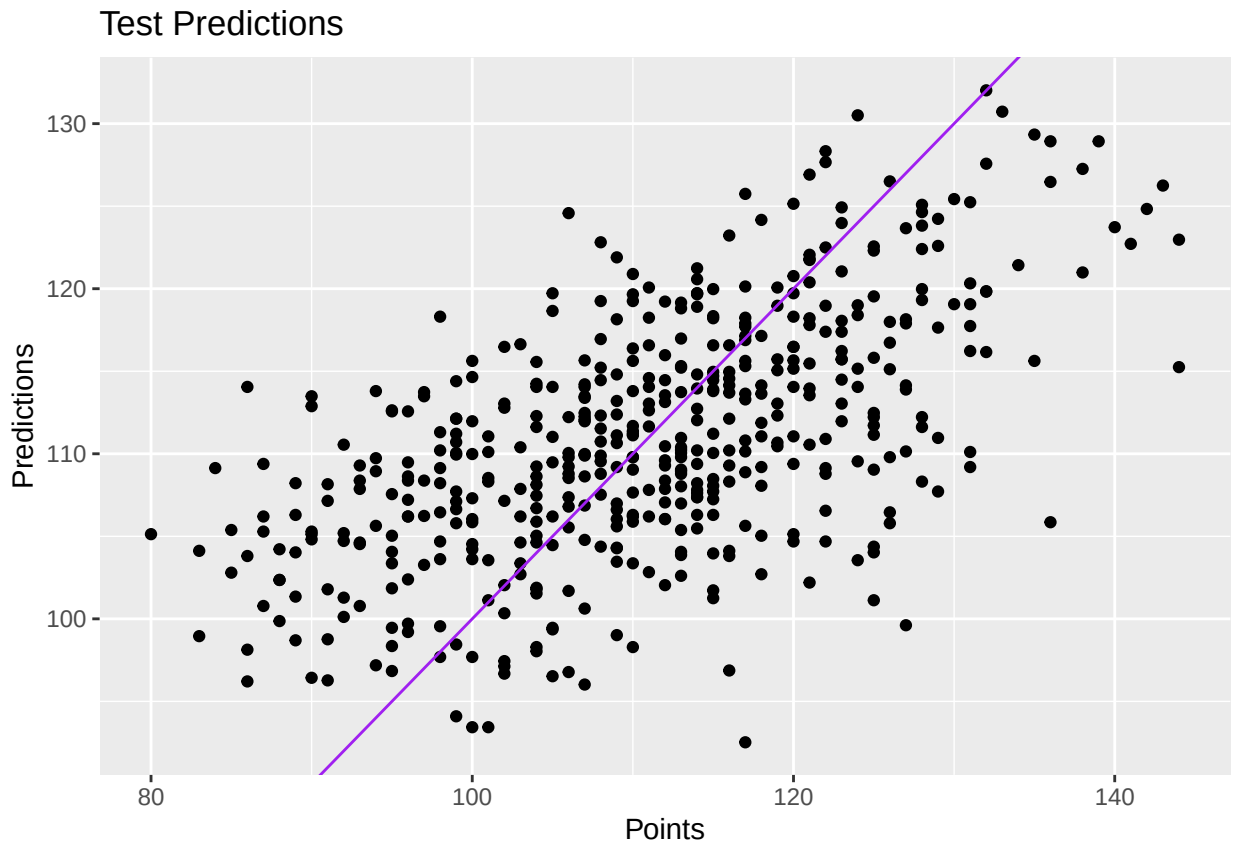
```
summary(model)
```

```
##
## Call:
## lm(formula = PTS ~ REB + AST, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.608  -7.123  -0.290   6.711  39.707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.90786    2.14620  30.243  < 2e-16 ***
## REB          0.25221    0.04020   6.275  4.6e-10 ***
## AST          1.41842    0.05134  27.630  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 1475 degrees of freedom
## Multiple R-squared:  0.3615, Adjusted R-squared:  0.3606
## F-statistic: 417.5 on 2 and 1475 DF,  p-value: < 2.2e-16
```

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

```
## # A tibble: 491 x 25
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 GSW   GSW @ MEM  04/10/2019  L        240   117    46    92       50
##  2 ORL   ORL @ CHA  04/10/2019  W        240   122    48    88       54.5
##  3 MIL   MIL vs. O~ 04/10/2019  L        240   116    43   100       43
##  4 ATL   ATL vs. I~ 04/10/2019  L        240   134    43   103       41.7
##  5 SAC   SAC @ POR  04/10/2019  L        240   131    50    96       52.1
##  6 LAC   LAC vs. U~ 04/10/2019  W        265   143    54   106       50.9
##  7 MIA   MIA @ BKN  04/10/2019  L        240    94    38    98       38.8
##  8 SAS   SAS vs. D~ 04/10/2019  W        240   105    41    88       46.6
##  9 DAL   DAL @ SAS  04/10/2019  L        240    94    37    91       40.7
## 10 NOP   NOP vs. G~ 04/09/2019  L        240   103    44    99       44.4
## # ... with 481 more rows, and 16 more variables: THREE_PTM <dbl>,
## #   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>, pred <dbl>
```

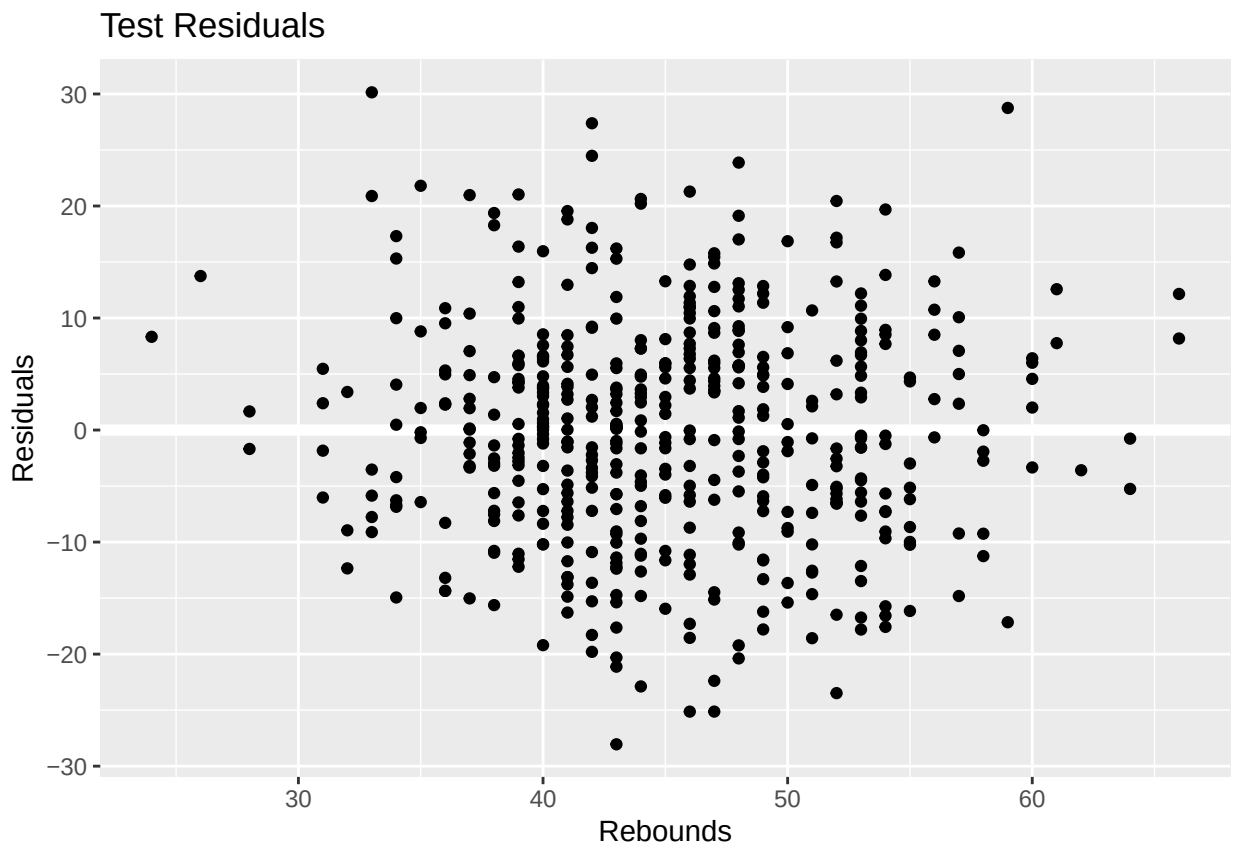
```
ggplot(data = predictions, mapping = aes(x = PTS, y = pred)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "purple") +
  labs(y = "Predictions", x = "Points", title = "Test Predictions")
```



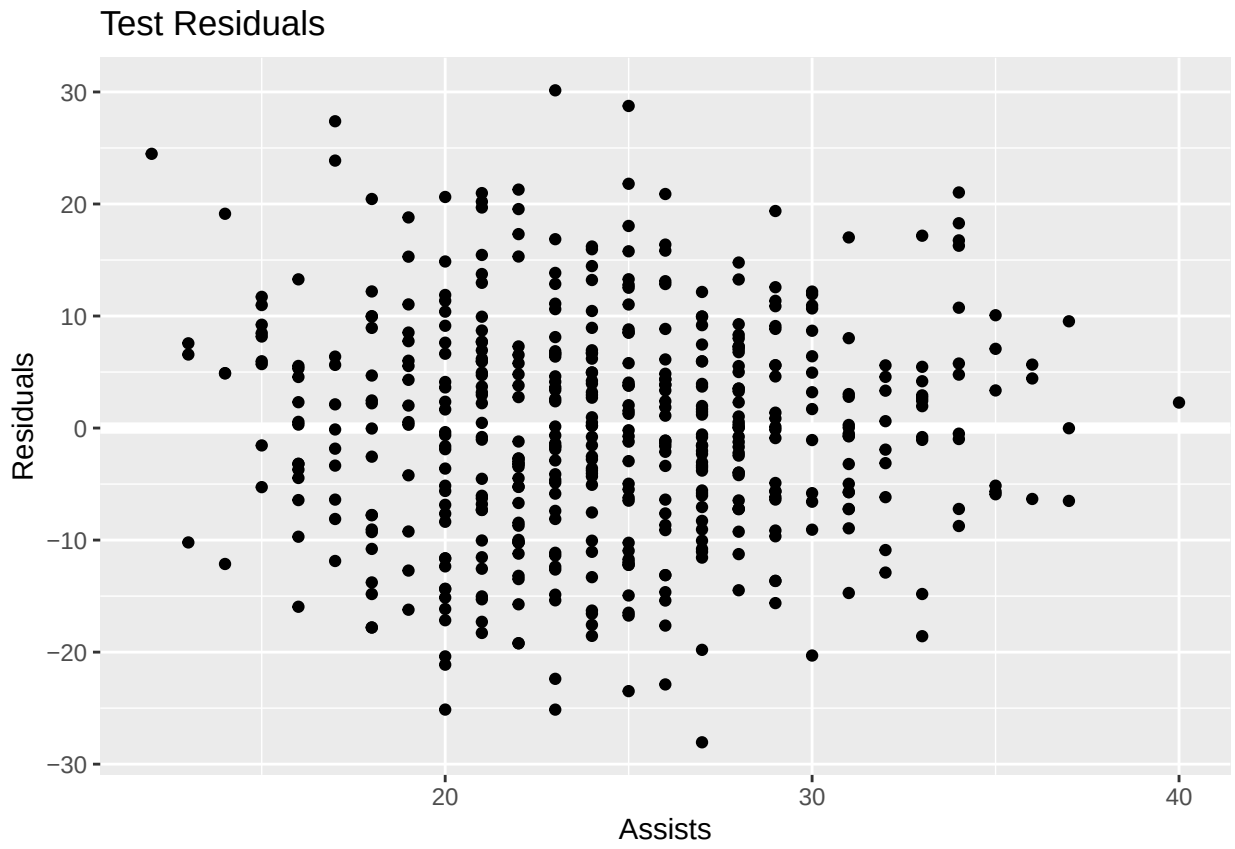

```
test_resids <- add_residuals(test_set, model)
test_resids
```

```
## # A tibble: 491 x 25
##    Team  `Match Up` `Game Date` W_or_L   MIN   PTS   FGM   FGA FG_PERCENT
##    <chr> <chr>      <chr>       <chr>  <dbl> <dbl> <dbl> <dbl>      <dbl>
##  1 GSW   GSW @ MEM  04/10/2019  L        240   117    46    92       50
##  2 ORL   ORL @ CHA  04/10/2019  W        240   122    48    88       54.5
##  3 MIL   MIL vs. O~ 04/10/2019  L        240   116    43   100       43
##  4 ATL   ATL vs. I~ 04/10/2019  L        240   134    43   103       41.7
##  5 SAC   SAC @ POR  04/10/2019  L        240   131    50    96       52.1
##  6 LAC   LAC vs. U~ 04/10/2019  W        265   143    54   106       50.9
##  7 MIA   MIA @ BKN  04/10/2019  L        240    94    38    98       38.8
##  8 SAS   SAS vs. D~ 04/10/2019  W        240   105    41    88       46.6
##  9 DAL   DAL @ SAS  04/10/2019  L        240    94    37    91       40.7
## 10 NOP   NOP vs. G~ 04/09/2019  L        240   103    44    99       44.4
## # ... with 481 more rows, and 16 more variables: THREE_PTM <dbl>,
## #   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>, 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")
```




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

```
## [1] 0.3539825
```

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

```
## [1] 7.950831
```

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

```
## [1] 9.936166
```

Observations

In general my model does a decent job at predicting points scored. When comparing my training and testing results, there is not anything to be too concerned about. The prediction results of each are very similar and both visualtions in the same way display that my model could definitely be improved. Between the residual graphs of training and testing, there are not any major discrepancies that need to be addressed.

My R^2, MAE, RMSE values are also similar between testing and training.

For training:

- R^2 = 0.3624039
- MAE = 8.266143
- RMSE = 10.37754

For testing:

- R^2 = 0.3539825
- MAE = 7.950831
- RMSE = 9.936166

There does not seem to be any evidence of overfitting or extremely better performance with the test set. In testing my model actually showed slightly worst performance than in training.

Potential Social and Ethical Implications

In recent years basketball has become more reliant on analytics, but there are still many unmeasurable factors during a basketball game. A simple model like mine doesn't really have enough substance to fully confirm that assists and rebounds lead to more points. It does a decent job at getting a general understanding of how these variables can affect the points scored. Ethically this model could be used to present misleading information about the effects of rebounding and total assists. For example, I could choose a prediction value that happens to be a perfect prediction and use that as confirmation bias for the general belief that "Better rebounding = more points scored". But in reality that statement is not completely true.

Fine tuning my goal

My initial model gave a good general idea about more in depth relationships between my variables. I would still like to be able to analyze home court advantage, but I need to make adjustments and changes to my datasets in order to make that happen.