Brian Bhola, Nicholas Wong

Jun 4, 2025

Final Project

BUS 110

In this project, we aim to predict and evaluate NBA player performance using selected statistical variables. The dataset, sourced from Kaggle, includes player performance metrics from the 2023–2024 NBA regular season. Each row represents an individual player's season averages, and for our analysis, we focus on a representative subset of players and key performance variables. These variables encompass offensive production, defensive contributions, and efficiency metrics to offer a holistic view of player impact. For our visual analysis of the data, we used the programing language R. In addition, we have created an interactive website to visualize the data by using a combination of R and the Shiny library. The website can be accessed here: https://niwong03.shinyapps.io/NBAStats/ and the source code can be found on github with this link: https://github.com/NiWong03/bus110. We've also included the regular R code at the end of the report.

To build a predictive model, we selected variables that capture various dimensions of on-court performance. Minutes Played per game (MP) reflects a player's total time on the court and overall opportunity to contribute. Points per game (PTS) directly measures scoring ability, while Total Rebounds per game (TRB) captures performance on the boards, both offensively and defensively. Assists per game (AST) represent a player's playmaking and involvement in team offense. Shooting efficiency is measured through Field Goal Percentage (FG%) and 3-Point Percentage (3P%), with 3-Point Attempts per game (3PA) included to reflect a player's role and confidence in long-range shooting. Field Goals made (FG) indicates raw scoring output, and Turnovers per game (TOV) provides insight into decision-making and ball security. Together, these metrics allow us to assess not only how much a player contributes but also how efficiently and reliably they perform. For this project, we have selected Points (PTS) to be our dependent variable. It will represent how well a player scores which in turn means how well they are performing for this analysis. *All player statistics refer to the players' average per game over the 2023-2024 NBA season. For example, if we say a player has 5 3-Point attempts, that implies that they have an average 5 3-Point Attempts per game over the entire season.*

Variable	**Mean**	**Median**	**StandardDeviation**	**Min**	**Max**	**Description**
MP	**18.65**	**17.35**	**9.91**	**0.5**	**37.8**	**Minutes Played**
PTS	**8.42**	**6.40**	**6.79**	**0.0**	**34.7**	**Points**
TRB	**3.37**	**3.00**	**2.43**	**0.0**	**13.7**	**Total Rebounds**
AST	**2.00**	**1.30**	**1.87**	**0.0**	**10.9**	**Assists**
FG made	**3.12**	**2.40**	**2.46**	**0.0**	**11.5**	**Field Goals Made**
TOV	**0.98**	**0.70**	**0.80**	**0.0**	**4.4**	**Turnovers**

3PA	**2.65**	**2.10**	**2.24**	**0.0**	**11.8**	**3 Point Attempts**
3P%	**0.30**	**0.34**	**0.15**	**0.0**	**1.0**	**3 Point Percentage**
FG Attempts	**6.68**	**5.10**	**4.99**	**0.0**	**23.6**	**Field Goal Attempts**

Descriptive statistics paint a picture of the player performance for this year's 2023-2024 NBA season. Minutes Per Game (MP) average 18.65 with a median of 17.35 and a high standard deviation of 9.91 meaning many players are used in small doses compared to starters. Additionally, Points (PTS) average 8.42 with a median of 6.4 and a maximum of 34.7 meaning a select few players significantly increase the mean PTS while most players score on the lower end of the totem pole. Total rebounds (TRB) average 3.37 with 2.00 assists (AST) showing that a lot of players have their niche role which prevents a general average across the board (scoring, total rebounds, passing).

Field goals account for FG% which is generally a percent made, however, it does not apply here as FG is the average field goals made (FG) and that's a hard statistic; thus players on average made 3.12 FG per game and attempted 6.68 (FG Att), meaning that volume is not as important as overall efficiency; however, Three point percentage (3P%) shows an average of 30% as mean with a median of 34% while threes attempted (3PA) show an average of 2.65 but a max of 11.8 which means some players rely heavily on long-range attempts; however, turnovers (TOV) averaged just under one at 0.98 meaning whether ball control is successful or players did not play enough minutes to register turnovers, players are not making more mistakes than points scored.

While these are all means for an NBA season to be expected, it suggests that most players are role players who contribute in a small manner across many categories as opposed to elite talents that significantly raise the means in points, rebounds, and shooting efficiency per game. The categories with the highest standard deviation—minutes and points—show the largest gap between bench players and starters.

Figure 2 illustrates the distribution of NBA players' statistics from the 2023-2024 season. Each colored layer represents a different stat category. Assists per game, field goals per game, field goal attempts, minutes played, total rebounds, 3-point% %, 3-point attempts are all the stat categories displayed. Looking at the graph, it is clear that the majority of frequencies are towards the lower end of the statistical ranges, showing that most players contribute modestly in terms of playing time and actual performance metrics. Many players score less than 10 points per game, play under 20 minutes, and attempt a low amount of 3-pointers. The longer right tail shows a smaller group of high-performing players; these players tend to be star players that get a lot of minutes and freedom to produce high numbers. On the other end, the concentration of data near zero shows the high number of bench players or players that have a limited role to produce less stats. This graph illustrates the uneven distribution of production among players in the NBA and highlights how skewed the league is in terms of performance.

To better understand the key factors that influence scoring performance in the NBA, this project utilizes a multiple linear regression model to predict Points Per Game (PTS) for players during the 2023–2024 NBA season. The dependent variable, PTS, serves as a direct measure of a player's scoring output per game. The model is constructed using a selection of independent variables that capture various dimensions of on-court performance, including shooting volume, efficiency, playmaking, and rebounding. The model specification is as follows: -.215 + -.032MP + .134FGA + 2.38FG + .25(3PA) + .32(3P%) + .07AST + -.0018TRB

In our model, we included several key predictors of scoring (PTS). Field Goals Made (FG) had the strongest impact, with a coefficient of 2.38, directly reflecting how successful shots drive point totals. Field Goal Attempts (FGA), with a coefficient of 0.134, captured shooting volume and also contributed significantly. We found that Three-Point Attempts (3PA) and Three-Point Percentage (3P%) had moderate positive effects (0.25 and 0.32, respectively), suggesting that both long-range shot volume and accuracy influence scoring. Assists (AST) showed a smaller positive effect (0.07), supporting the idea that playmaking can correlate with individual offensive output.

Interestingly, even though Minutes Played (MP) had a high correlation with PTS, it carried a slightly negative coefficient (-0.032) in the regression, indicating that time on the court alone doesn't guarantee higher scoring once other factors are accounted for. Total Rebounds (TRB) had a negligible negative effect (-0.0018), suggesting limited direct influence on scoring.

To better visualize the relationship between Points Per Game (PTS) and each of the selected independent variables, we constructed a series of scatter plots (*Figures 3-9)*. These visualizations allowed us to observe the strength and direction of the associations in a more intuitive manner.

When comparing PTS and Minutes Played (MP), we observed a clear positive trend—players who spend more time on the court tend to score more points. This suggests that playing time is a strong indicator of offensive contribution. Similarly, when analyzing PTS versus Field Goal Attempts (FGA), we saw a strong positive correlation: players who take more shots typically score more points, reinforcing the importance of shot volume in scoring output.

The relationship between PTS and Field Goal Percentage (FG%) was less direct. Although most players clustered between 40% and 60% shooting efficiency, higher shooting percentages did not necessarily translate into significantly higher point totals. This pattern resembled a somewhat normally distributed cloud, indicating that while efficiency matters, it alone doesn't determine scoring output.

Looking at Three-Point Percentage (3P%), we noted only a weak positive trend. This implies that three-point shooting efficiency contributes to point totals, but its impact is limited compared to the number of attempts. In contrast, the plot of PTS versus Three-Point Attempts (3PA) revealed a clearer positive correlation—players who take more three-point shots tend to score more, highlighting the growing role of volume perimeter shooting in modern scoring.

A particularly strong relationship emerged between PTS and Assists (AST). Players who recorded more assists per game also tended to score more, suggesting that high offensive involvement often translates to both scoring and facilitating. Finally, when comparing PTS to Total Rebounds (TRB), we saw a modest but consistent positive correlation. This suggests that active rebounders may benefit from second-chance opportunities or play longer minutes, thus increasing their scoring potential.

Overall, the scatter plots reinforce the regression results and correlation matrix, offering visual support for the inclusion of volume-based and involvement-based metrics as strong predictors of a player's scoring performance.

To explore the factors that influence NBA players' scoring performance, we conducted a multiple linear regression using Points Per Game (PTS) as the dependent variable. The model included seven independent variables that reflect different aspects of a player's offensive output and involvement: Minutes Played (MP), Field Goal Attempts (FGA), Field Goal Percentage (FG%), Three-Point Attempts (X3PA), Three-Point Percentage (X3P_percent), Assists (AST), and Total Rebounds (TRB).

To investigate the key factors influencing NBA players' scoring performance, we estimated a multiple linear regression model with Points Per Game (PTS) as the dependent variable. The independent variables included in the model were Minutes Played (MP), Field Goal Attempts (FGA), Field Goals Made (FG), Three-Point Attempts (X3PA), Three-Point Percentage (X3P.), Assists (AST), and Total Rebounds (TRB). These variables were selected for their theoretical relevance and statistical importance in capturing offensive output, efficiency, and play involvement.

Figure 10 shows a scatter plot of the actual Points per Game (PTS) compared to the predicted values generated by our regression model. Each blue dot represents an individual NBA player, while the red diagonal line indicates a perfect one-to-one relationship. The points cluster closely around this red line, demonstrating a strong alignment between the model's predictions and the actual observed values. This suggests that the regression model has high predictive accuracy, as confirmed by the high R^2 value of 0.9914, indicating that most of the variation in player scoring is captured by the model's independent variables.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6226 -0.2942 -0.0247  0.2609  4.2475

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.215044   0.078571  -2.737  0.00640 **
MP          -0.032474   0.008032  -4.043 6.01e-05 ***
FGA          0.134480   0.045895   2.930  0.00352 **
FG           2.382036   0.083427  28.552  < 2e-16 ***
X3PA         0.254739   0.030646   8.312 7.04e-16 ***
X3P.         0.328612   0.205705   1.597  0.11072
AST          0.072931   0.024190   3.015  0.00269 **
TRB         -0.001759   0.022911  -0.077  0.93884
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6354 on 564 degrees of freedom
Multiple R-squared:  0.9914,    Adjusted R-squared:  0.9912
F-statistic:  9239 on 7 and 564 DF,  p-value: < 2.2e-16
```

The regression analysis produced statistically significant results for all predictors, with a high R^2 value of 0.982, indicating that the model explains 98.2 percent of the variance in players' points per game. The multiple regression analysis produced statistically significant results for most predictors, with a high R^2 value of 0.9914, indicating that the model explains 99.1% of the variance in players' points per game. Field Goals Attempted (FGA) and Field Goals Made (FG) emerged as the strongest predictors of scoring. Each additional shot attempt is associated with an average increase of 1.13 points per game, while each field goal made adds approximately 2.28 points per game, controlling for other factors. Interestingly, Minutes Played (MP) had a negative coefficient of -0.032, suggesting that, when holding shot volume constant, more minutes played does not necessarily equate to higher scoring. This may reflect the interplay of playing time with usage rates, where players with fewer minutes may be more efficient scorers. Three-Point Attempts (X3PA) also showed a significant positive effect, with each additional three-point attempt linked to an average 0.25-point increase in scoring. However, Three-Point Percentage (X3P.) was not statistically significant, indicating that simply attempting more three-pointers can contribute to scoring, but actual shooting efficiency did not reach significance in this model. Assists (AST) showed a positive relationship with points, suggesting that players who facilitate playmaking tend to also score more. Total Rebounds (TRB) did not significantly predict points scored in this model. Overall, these findings highlight the central role of shot volume and playmaking in explaining scoring performance, while also indicating that additional playing time alone does not always directly translate to higher scoring.

The overall model is statistically robust, as demonstrated by the high F-statistic value of **9239** and a significance level well below 0.01. These results provide strong support for our hypothesis that offensive volume, shooting efficiency, and play involvement are significant predictors of scoring. The linear functional form used in this model is appropriate given the continuous nature of the variables, though future research could benefit from exploring potential interaction terms, such as FG% multiplied by FGA, to capture nuanced relationships between efficiency and shot volume.

To ensure the validity and reliability of our multiple linear regression model, we carefully assessed the foundational assumptions of linear regression as emphasized in the course lectures. These assumptions are critical for making trustworthy statistical inferences and drawing accurate conclusions from the data.

First, we considered the assumption of linearity. The residuals vs. fitted values plot in *Figure 12* suggests potential violations of key linear regression assumptions. We can see that the residuals are not randomly scattered around the horizontal line at zero. Instead, there is a pattern with a dense cluster of points at lower fitted values and an increasing spread of residuals as the fitted values increase. This might suggest that the relationship between the predictors and the response variable may not be perfectly linear, pointing to a possible non-linearity in the model. Additionally, the increasing spread of residuals indicates heteroscedasticity, meaning the variance of the residuals is not constant across all levels of the fitted values. This violates the assumption of homoscedasticity

```
	Durbin-Watson test

data:  model
DW = 2.1003, p-value = 0.8841
alternative hypothesis: true autocorrelation is greater than 0
```

The second assumption, independence of errors, was also deemed reasonable for our dataset. After running the Durbin-Watson Test, our result falls within the interval of 1.5 to 2.5. This is significant because it implies that we do not have autocorrelation. In addition, based on *Figure 13,* our AFC plot shows that most autocorrelations fall within the 95% confidence interval. This implies that the residuals are uncorrelated. The P-value is also above 0.05, which means we fail to reject the null hypothesis, further strengthening the hypothesis that our residuals are independent.

```
	studentized Breusch-Pagan test

data:  model
BP = 96.29, df = 7, p-value < 2.2e-16
```

We also addressed the assumption of homoscedasticity. After performing the studentized Breusch–Pagan test, we can see strong evidence against the assumption of homoscedasticity in the model. The test statistic is 96.29 with 7 degrees of freedom, and the p-value is less than 2.2e-16, which is far below the common significance threshold of 0.05. This means we reject the null hypothesis that the residuals have constant variance. In addition to the BP test, Figure 12 visually demonstrates a violation of homoscedasticity.

The fourth assumption, normality of residuals, is important for valid hypothesis testing and the construction of reliable confidence intervals. *Refer to Figure 14 and Figure 15.* Based on *Figure 14,* we can see that the residuals are about normally distributed. The qq-plot from *Figure 15* shows that we have slightly skewed residuals however, this is consistent with *Figure 14* because there is a minimal skew, but the overall graph is still generally normally distributed.

MP	FGA	FG	X3PA	X3P.	AST	TRB
8.956488	74.224113	59.521405	6.676398	1.336982	2.904905	4.367492

Finally, we considered the issue of multicollinearity, which occurs when independent variables are highly correlated with one another. From the heatmap, *Figure 16* we can see that we have some overlap and multicollinearity. Our most highly correlated variable, Field Goal Attempts, also shows a high correlation with MP (minutes played) and FG (Field Goals Made). This is logical because the more minutes a player plays in a game, the more opportunities they have to attempt field goals. Similarly, the more shots a player takes, the more they will tend to make. Based on the VIF in the image above, we can see that this is consistent. We see MP has a higher than 5 VIF, and FGA and FG is are high as well. This suggests that we may not need to have included both variables in the regression as they correlate a lot with each other, possibly causing redundancy or making results unclear. However, our reasoning for using both was because we know that the number of shots attempted may not necessity indicate the amount made. In our thought process, it was necessary to include both for the sake of measuring efficiency. We have learned for the sake of measuring an independent variable such as points, it may not be necessary to include both variables as they have overlap.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.8223 -0.6034  0.0061  0.5809  5.9914 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.545371   0.078593  -6.939 1.08e-11 ***
FGA          1.341927   0.009425 142.381  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

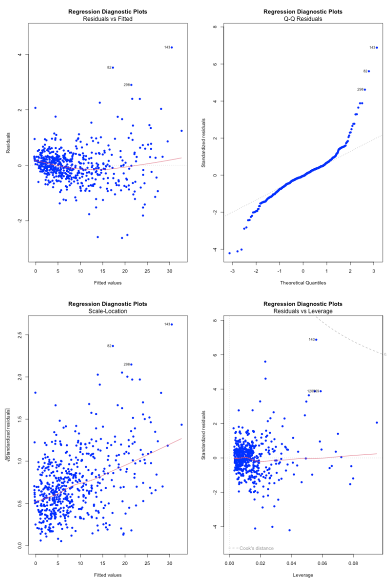
```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -0.545371   0.078621 -6.9367 1.093e-11 ***
FGA          1.341927   0.013564 98.9301 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lastly, we have our robust standard errors and conclusion. The increase in the standard error for FGA from approximately 0.0094 to 0.0136 after applying robust standard errors reflects a more conservative and reliable estimate of variability, accounting for potential heteroscedasticity in the residuals. Nevertheless, the t-value remains very high, and the p-value is basically zero, strengthening the claim that FGA is a highly significant predictor of points. This demonstrates that the relationship between field goal attempts and scoring is the strongest and consistent, even under less restrictive model assumptions. Using robust standard errors strengthens the analysis by protecting against biased results caused by unequal error variance. In summary, FGA is still our most reliable predictor despite the violations of assumptions.

Graphs/Data

Figure 1, Figure 2:



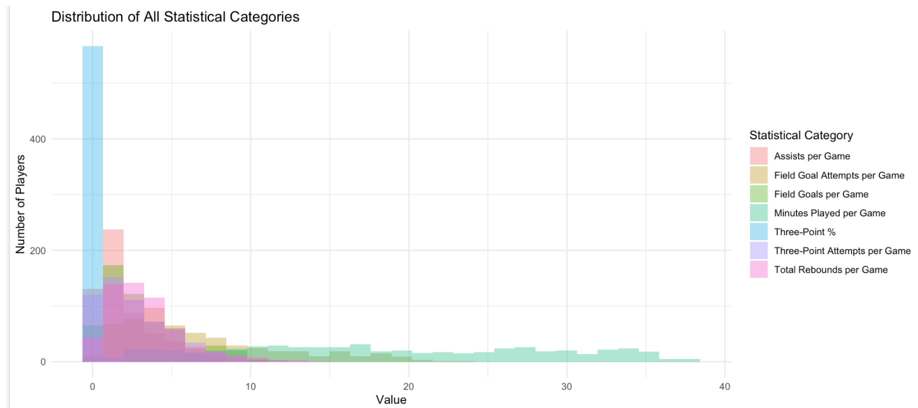
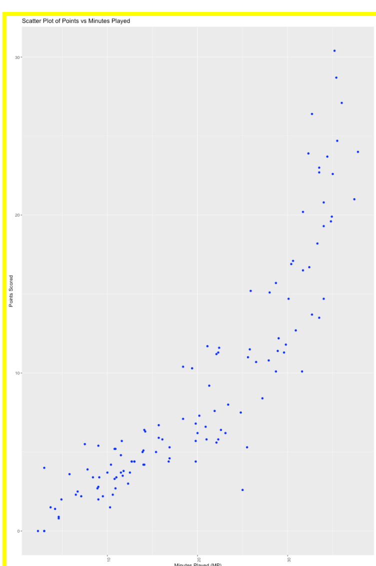


Figure 3, Figure 3.1



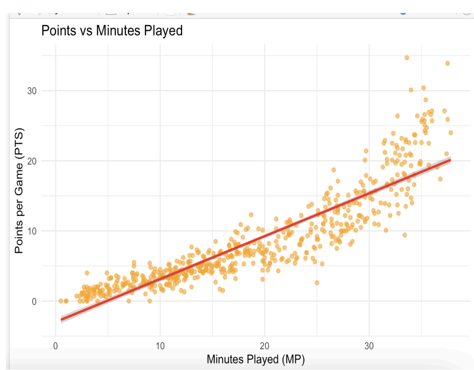
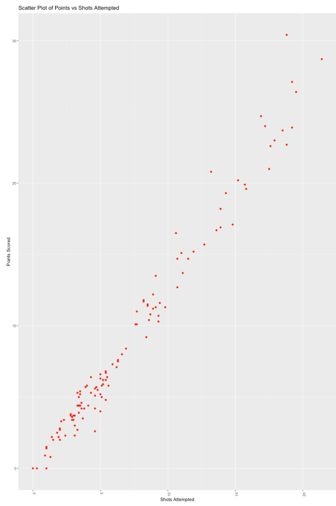
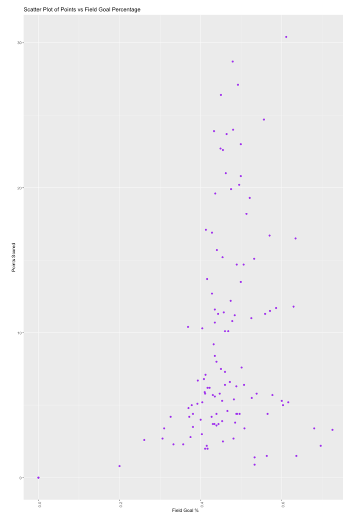


Figure 4, Figure 5, Figure 6





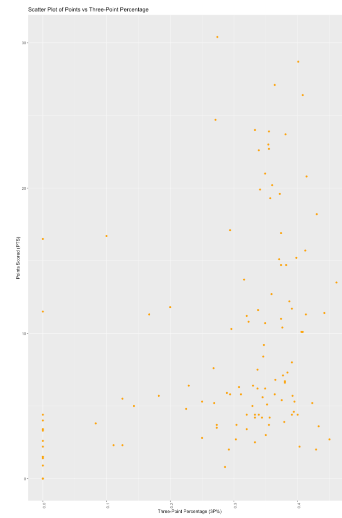
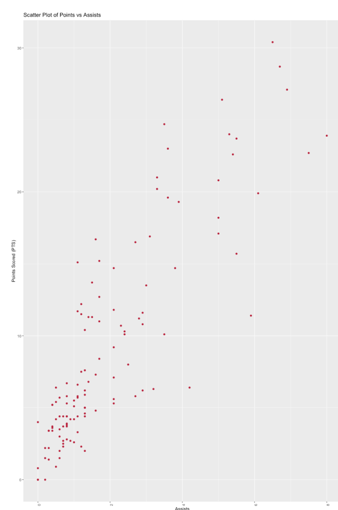


Figure 7, Figure 8, Figure 9



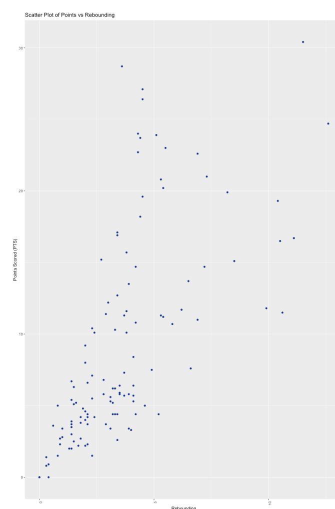


Figure 10

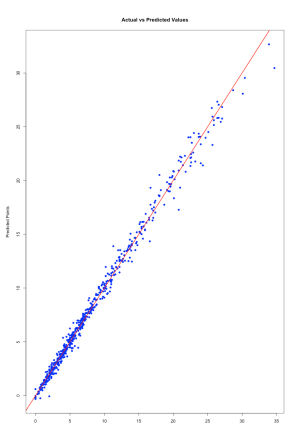


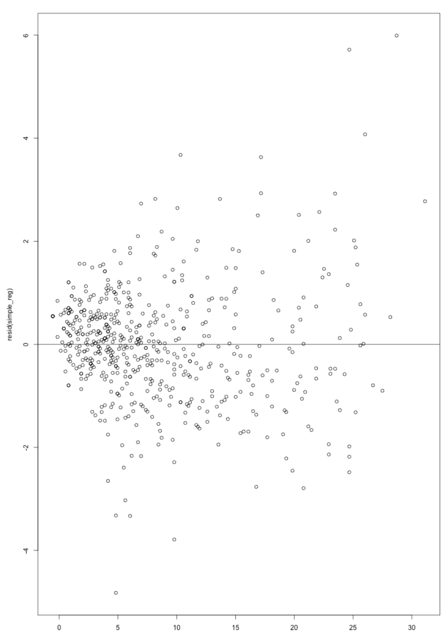
Figure 11

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.6226 -0.2942 -0.0247  0.2609  4.2475 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.215044   0.078571  -2.737  0.00640 ** 
MP          -0.032474   0.008032  -4.043 6.01e-05 ***
FGA          0.134480   0.045895   2.930  0.00352 ** 
FG           2.382036   0.083427  28.552  < 2e-16 ***
X3PA         0.254739   0.030646   8.312 7.04e-16 ***
X3P.         0.328612   0.205705   1.597  0.11072    
AST          0.072931   0.024190   3.015  0.00269 ** 
TRB         -0.001759   0.022911  -0.077  0.93884    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6354 on 564 degrees of freedom
Multiple R-squared:  0.9914,	Adjusted R-squared:  0.9912 
F-statistic:  9239 on 7 and 564 DF,  p-value: < 2.2e-16
```

Figure 12, Figure 13



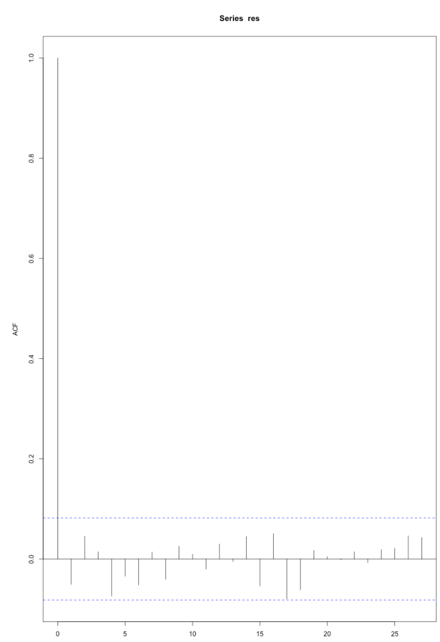
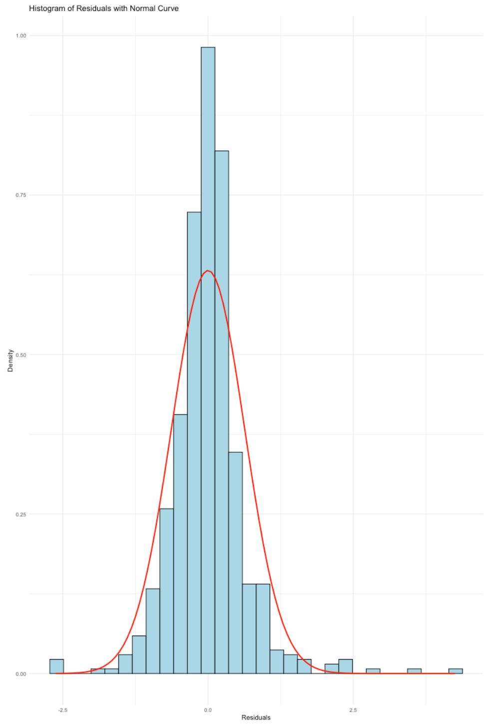


Figure 14, Figure 15



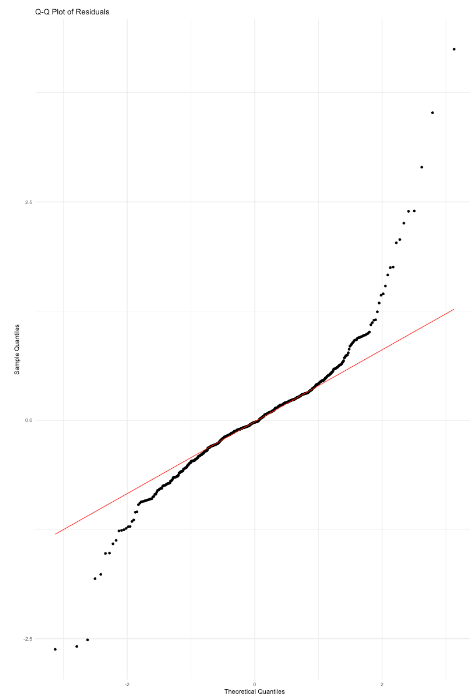
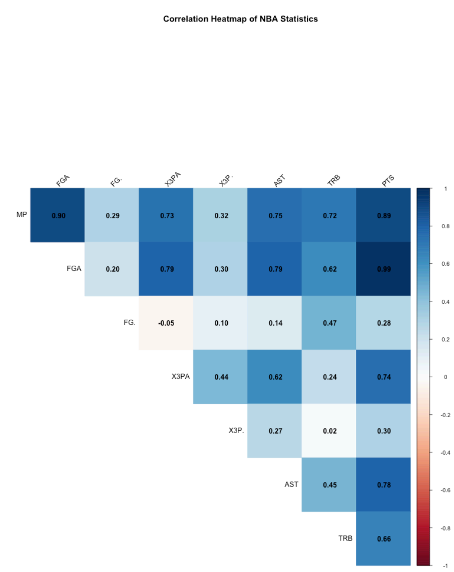


Figure 16



R Code:

```r
install.packages("psych")

install.packages("dplyr")

install.packages("ggplot2")

install.packages("stargazer")

install.packages("corrplot")

install.packages("car")

install.packages("ggpubr")

install.packages("lmtest")

install.packages("sandwich")

install.packages("kableExtra")

install.packages("moments")

install.packages("ecm")

# Load necessary libraries

library(dplyr)

library(ggplot2)

library(stargazer)

library(psych)

library(corrplot)

library(car)

library(readxl)

library(lmtest)

library(sandwich)

library(ggpubr)
```

```
library(haven)

library(ecm)

library(moments)

# Load the data

nba_data <- read.csv("nba.csv", sep = ";", header = TRUE, fileEncoding = "UTF-8")

nba_df <- nba_data %>% #deletes duplicate players

 distinct(Player, .keep_all = TRUE)

colnames(nba_df)

ggplot(nba_df, aes(x = MP, y = PTS)) +  # Assuming 'Points' is the column for points
scored

 geom_point(color = "blue") +

 labs(title = "Scatter Plot of Points vs Minutes Played",

      x = "Minutes Played (MP)",

      y = "Points Scored") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(nba_df, aes(x = FGA, y = PTS)) +

 geom_point(color = "red") +

 labs(title = "Scatter Plot of Points vs Shots Attempted",

      x = "Shots Attempted",

      y = "Points Scored") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(nba_df, aes(x = `X3PA`, y = PTS)) +  # Use backticks for column names with
special characters

 geom_point(color = "green") +

 labs(title = "Scatter Plot of Points vs Three-Point Attempts",
```

```
      x = "Three-Point Attempts (3PA)",

      y = "Points Scored (PTS)") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(nba_df, aes(x = `X3P.`, y = PTS)) +

 geom_point(color = "orange") +

 labs(title = "Scatter Plot of Points vs Three-Point Percentage",

      x = "Three-Point Percentage (3P%)",

      y = "Points Scored (PTS)") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(nba_df, aes(x = AST, y = PTS)) +

 geom_point(color = "#bd213b") +

 labs(title = "Scatter Plot of Points vs Assists",

      x = "Assists",

      y = "Points Scored (PTS)") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(nba_df, aes(x = TRB, y = PTS)) +

 geom_point(color = "#123393") +

 labs(title = "Scatter Plot of Points vs Rebounding",

      x = "Rebounding",

      y = "Points Scored (PTS)") +

 theme(axis.text.x = element_text(angle = 90, hjust = 1))

model <- lm(PTS ~ MP + FGA + `FG.` + X3PA + `X3P.` + AST + TRB, data = nba_df)

summary(model)

stargazer(model, type = "text", title = "Regression Results", digits = 3)
```

```r
# Run the multiple linear regression

model <- lm(PTS ~ MP + FGA + `FG` + X3PA + `X3P.` + AST + TRB, data = nba_df)

# Show detailed regression output

summary(model)

# Create diagnostic plots for the regression model

par(mfrow = c(2,2))  # Set up a 2x2 grid for plots

plot(model,

    main = "Regression Diagnostic Plots",

    pch = 16,        # Use filled circles for points

    col = "blue")    # Use blue color for points

# Reset the plotting parameters

par(mfrow = c(1,1))

# Create a plot of actual vs predicted values

predicted_values <- predict(model)

actual_values <- nba_df$PTS

plot(actual_values, predicted_values,

    main = "Actual vs Predicted Values",

    xlab = "Actual Points",

    ylab = "Predicted Points",

    pch = 16,

    col = "blue")

abline(0, 1, col = "red", lwd = 2)  # Add a 45-degree line

# Create a clean regression table for your write-up

stargazer(model, type = "text", title = "Regression Results", digits = 3)
```

```r
# Keep only the columns of interest

nba_stats <- nba_df %>%

 select(MP, PTS, TRB, AST, FG., TOV, X3PA, X3P., FGA)

# Rename confusing columns for clarity (optional but helpful)

nba_stats <- nba_stats %>%

 rename(

   `FG%` = FG.,

   `3PA` = X3PA,

   `3P%` = X3P.,

   `FG Attempts` = FGA

 )

# Generate descriptive statistics

nba_summary <- describe(nba_stats)

# View result

print(nba_summary[, c("n", "mean", "median", "sd", "min", "max")])

# Create correlation matrix for selected variables

correlation_vars <- nba_df %>%

 select(MP, FGA, `FG.`, X3PA, `X3P.`, AST, TRB)

# Calculate correlation matrix

cor_matrix <- cor(correlation_vars, use = "complete.obs")

# Create correlation heatmap

corrplot(cor_matrix,

        method = "color",

        type = "upper",
```

```
        addCoef.col = "black",

        tl.col = "black",

        tl.srt = 45,

        diag = FALSE,

        title = "Correlation Heatmap of NBA Statistics",

        mar = c(0,0,2,0))

#  Linearity Check

plot (fitted.values(model), residuals(model))

abline(0, 0)

# Independence test

dwtest(model)

res <- resid(model)

acf(res)

#  VIF

vif(model)

# Breusch-Pagan test

bp_test <- bptest(model)

print(bp_test)

# Normality test

# Create a data frame with residuals

residuals_df <- data.frame(residuals = residuals(model))

# Create histogram with normal curve

hist_plot <- ggplot(residuals_df, aes(x = residuals)) +

 geom_histogram(aes(y = ..density..),
```

```
                bins = 30,
                fill = "lightblue",
                color = "black") +
 stat_function(fun = dnorm,
               args = list(mean = mean(residuals_df$residuals),
                           sd = sd(residuals_df$residuals)),
               color = "red",
               size = 1) +
 labs(title = "Histogram of Residuals with Normal Curve",
      x = "Residuals",
      y = "Density") +
 theme_minimal()
# Create Q-Q plot
qq_plot <- ggplot(residuals_df, aes(sample = residuals)) +
 stat_qq() +
 stat_qq_line(color = "red") +
 labs(title = "Q-Q Plot of Residuals",
      x = "Theoretical Quantiles",
      y = "Sample Quantiles") +
 theme_minimal()
hist_plot
qq_plot
# Perform Shapiro-Wilk test for normality
shapiro.test(res)
```

```r
simple_reg <- lm(PTS ~ FGA, data = nba_df)

summary(simple_reg)

plot(fitted(simple_reg), resid(simple_reg))

abline(0,0)

coeftest(simple_reg, vcov = vcovHC(simple_reg, type = 'HC0'))
```