Final Project

Justin Wininger, Otis Murray, Henry Katz

2023-11-29

Introduction

This project explores the relationships between Total Yards Per Game (YPG) and the Points Per Game (PPG) that a team scores in both the National Football League (NFL), as well as the Division 1 Football Bowl Subdivision (FBS) of College football. There are more college football teams than just the FBS, but for simplicity's sake, this project will refer to the Division 1 Football Bowl Subdivision teams as simply "college" teams. These results are then compared to answer the question of whether PPG and YPG have a differing relationship between the NFL and FBS. This project first examines the individual relationships between the NFL teams' YPG and their PPG, as well as examining how passing yards per game (PYPG) and rushing yards per game (RYPG) affect those results. This process is then repeated for the FBS, and then the resulting data is used to conclude how similar these relationships are in the FBS versus the more competitive NFL. This project heavily relies on simple linear regression (SLR) and multiple linear regression (MLR) in order to display the differing trends between PPG and the YPG.

NFL Data

First, we set off to explore the NFL data set that we scraped from pro-football-reference.com. This data set included a table with all 32 NFL teams, as well as statistics such as total yards, total passing yards, total rushing yards, etc. We then proceeded to tidy this table into one that only involved the Team, PPG, YPG, PYPG, and RYPG. Using this table we then created scatter plots for the three relationships: YPG v.s. PPG, PYPG v.s. PPG, and RYPG v.s. PPG. Using these scatter plots we will analyze using SLR and MLR.

NFL Visualizations

Below is a table that attempts to show a relationship between PPG and YPG in the NFL.

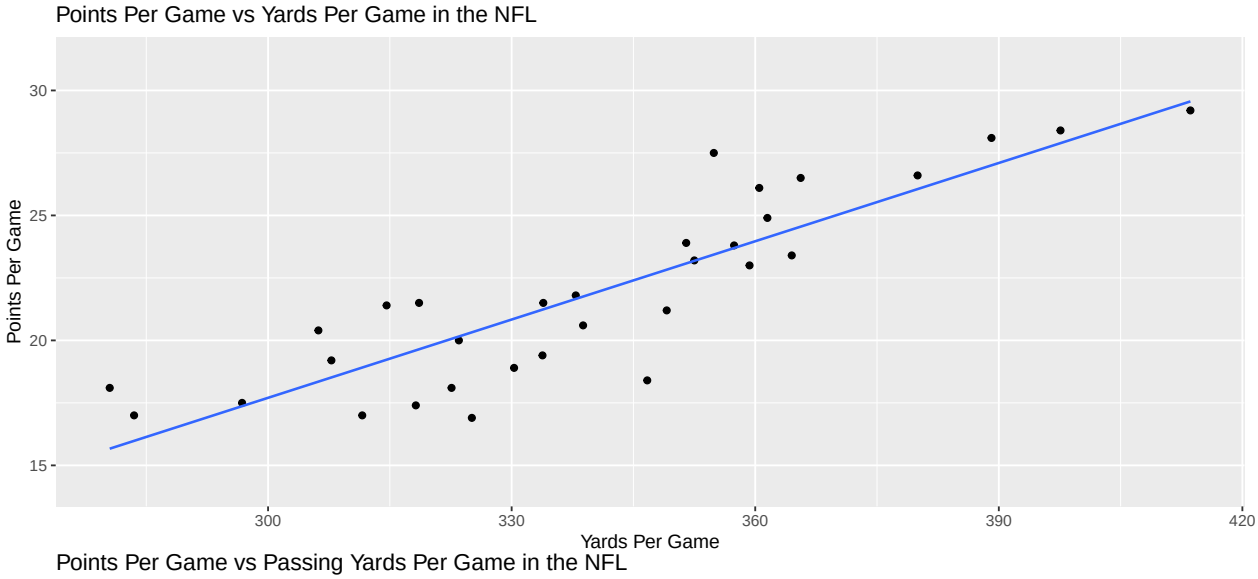
In this table, the top 14 NFL teams are sorted by their rank (out of 32) in PPG in ascending order. Their rank in YPG is also shown. On the far left is each team's ranking alphabetically. This is done to establish a baseline. This table attempts to show that, in general, the teams ranked highly in PPG are also ranked highly in YPG. For example, the top three teams in PPG, the Kansas City Chiefs, Buffalo Bills, and Philadelphia Eagles respectively, are also the top three teams in YPG, in the same order. The fourth ranked team in PPG, the Dallas Cowboys is only 11th in YPG, so the pattern is not true for all 32 teams. However, of the top 10 teams in PPG, only two are outside the top 10 in YPG, the aforementioned Cowboys and the Seattle Seahawks. Those two teams are ranked 11th and 13th in YPG respectively. The fact that the top 10 teams in PPG are all top 13 in YPG is an excellent indicator that there may be a relationship between PPG and YPG. We will conduct further analysis to see whether there really is a relationship between PPG and YPG or not.

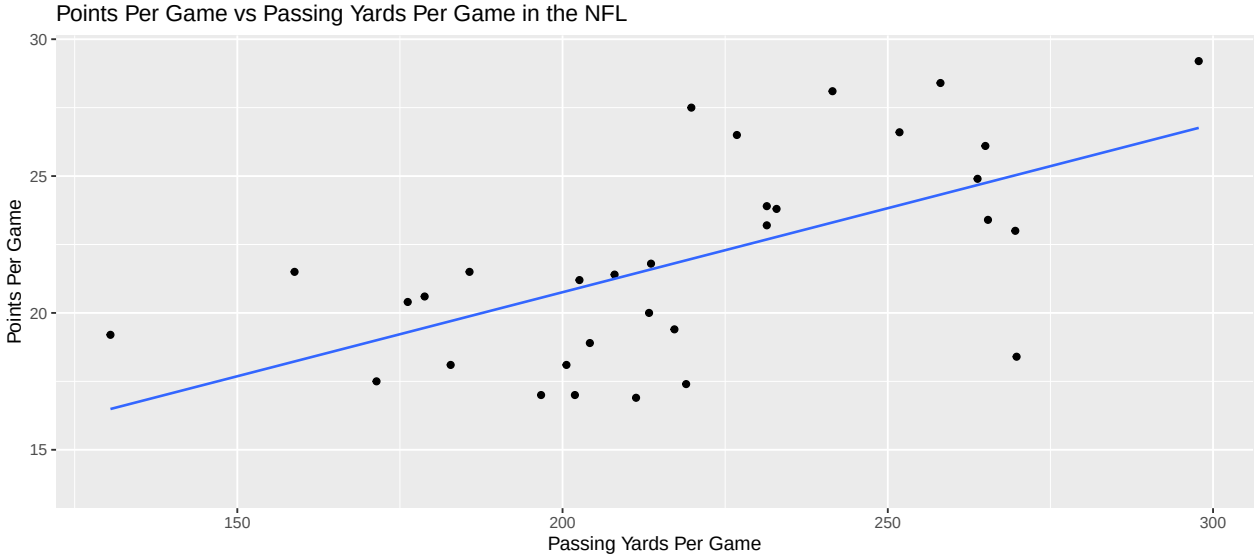
Table 1: Every NFL team's rank in PPG and YPG

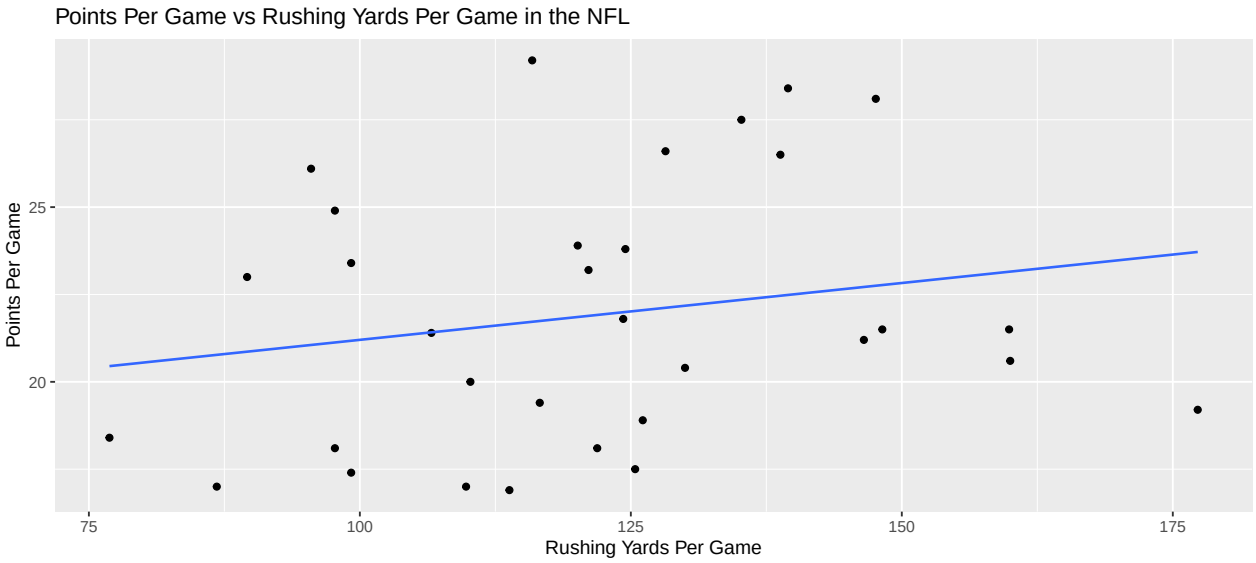
	Team	YPG Rank	PPG Rank
10	Chiefs	1	1
4	Bills	2	2
15	Eagles	3	3
13	Cowboys	11	4
20	Lions	4	5
1	49ers	5	6
3	Bengals	8	7
32	Vikings	7	8
28	Seahawks	13	9
18	Jaguars	10	10
14	Dolphins	6	11
24	Raiders	12	12
9	Chargers	9	13
21	Packers	17	14

Going further into that exploration, below are three scatter plots. The first shows the relationship between PPG and YPG in the NFL. The second shows the relationship between PPG and PYPG in the NFL. The third shows the relationship between PPG and RYPG in the NFL.

```
## Warning: Using the `size` aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the `linewidth` aesthetic instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```







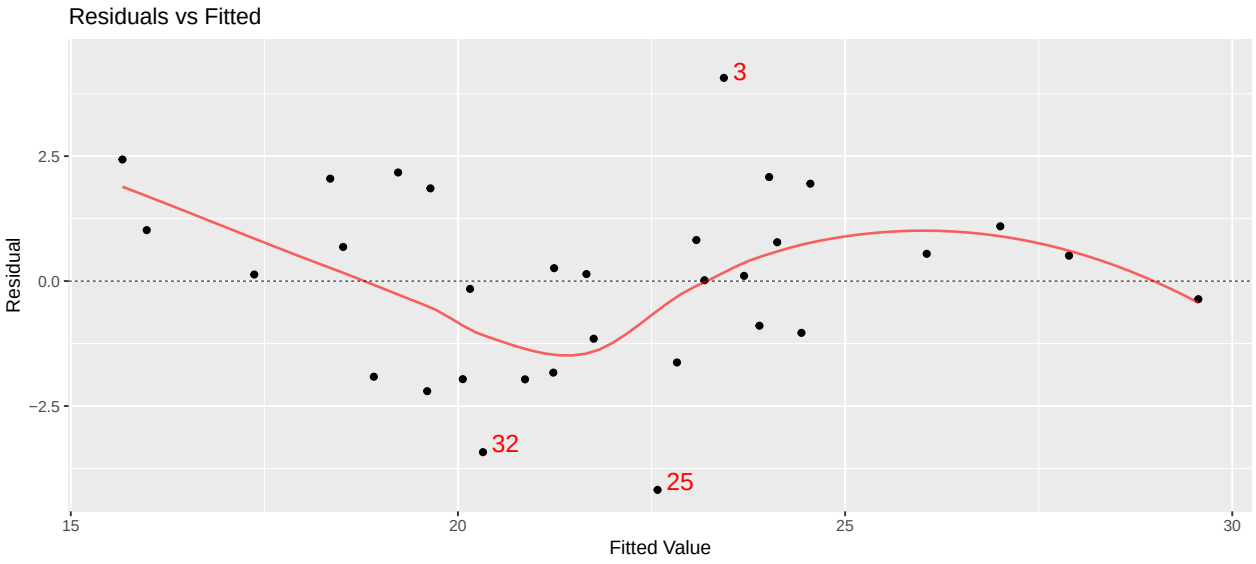
All three scatter plots show a positive, linear relationship. The first two, PPG vs YPG and PPG vs PYPG respectively, show a very clear and obvious positive linear relationship. The third plot, PPG vs RYPG, shows a less clear and obvious positive linear relationship. There appears to be some correlation, but not a lot. Regardless, the most important thing is that y increases as x increases. This is true for all three plots, whether x is YPG, PYPG, or RYPG.

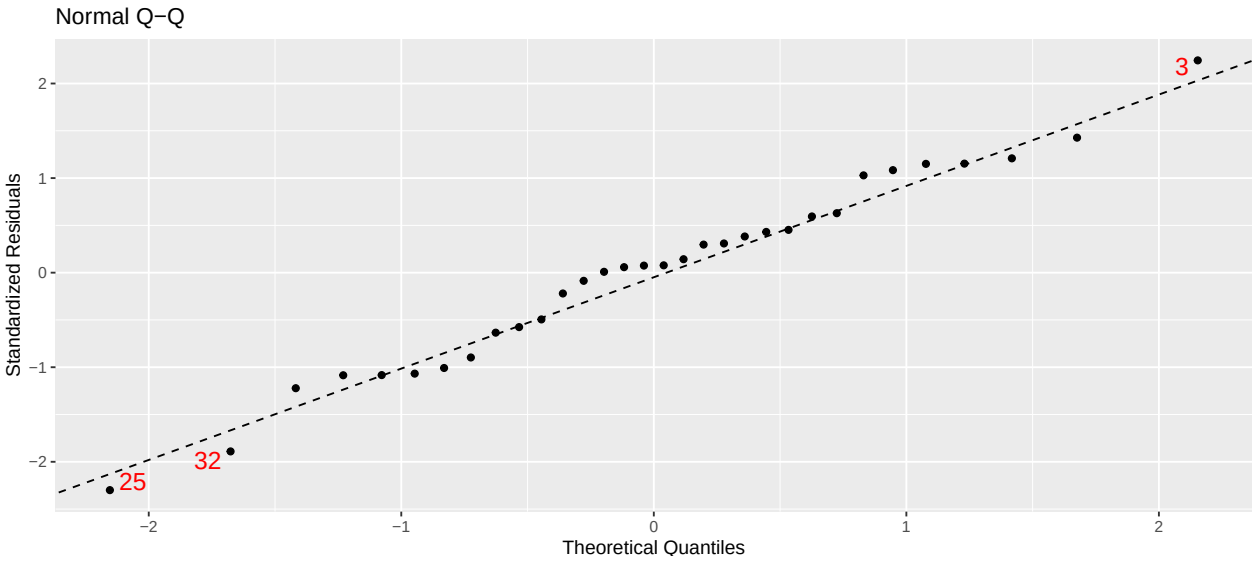
NFL Analysis

Simple Linear Regression

We decided to perform SLR to see if PPG and YPG are related. In order to infer, some conditions must be met. The data must be linear, have equal variance, and must be normal. Below are two plots that asses whether or not the conditions are met.

```
## `geom_smooth()` using formula = 'y ~ x'
```





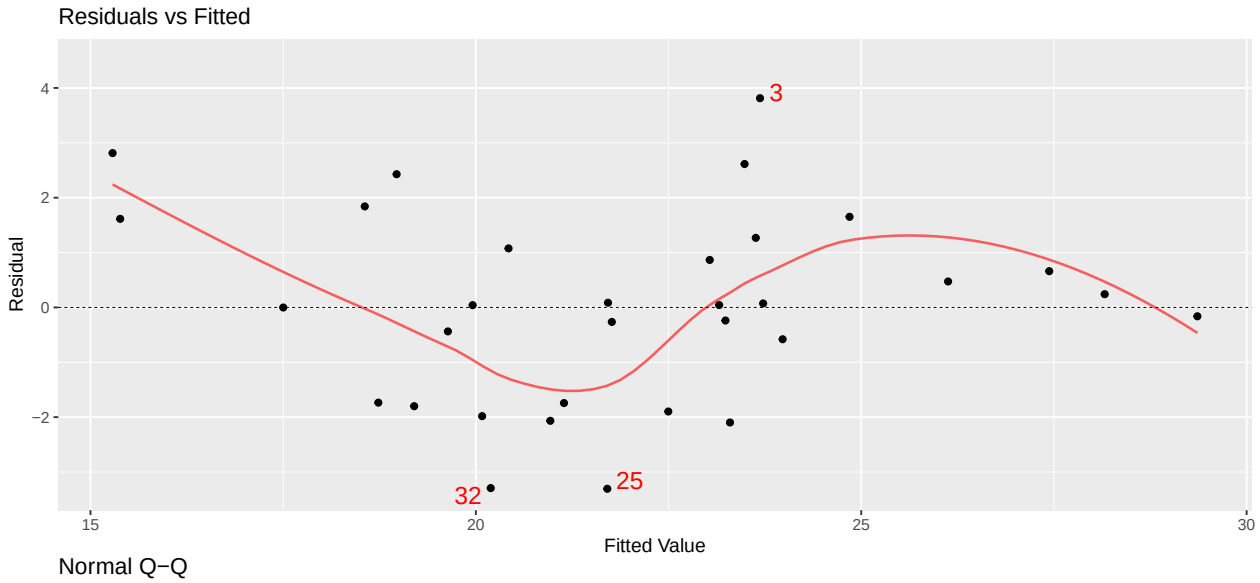
The first condition that must be met is linearity. The explanatory and response variables must have a linear relationship. This was proven by the first scatter plot of PPG and YPG. The linearity condition, therefore, is met. The second condition that must be met is equal variance. The first plot above is a scatter plot of residuals versus fitted values. The points in this plot must show an equal spread across the width. If they do, then the data have equal variance. Our residual vs fitted plot is fairly good. It is certainly not perfect, but nothing ever is in the real world. The plot shows enough spread and no obvious patterns, so the equal variance condition, therefore, is met. The third and final condition that must be met is normality. The data must be normally distributed. A Q-Q plot assesses this condition. If the Q-Q plot shows a strong linear relationship then the data are normal. Our QQ plot shows a relatively strong linear relationship. Like before, it is not perfect, but it is good enough. The normality condition, therefore, is met.

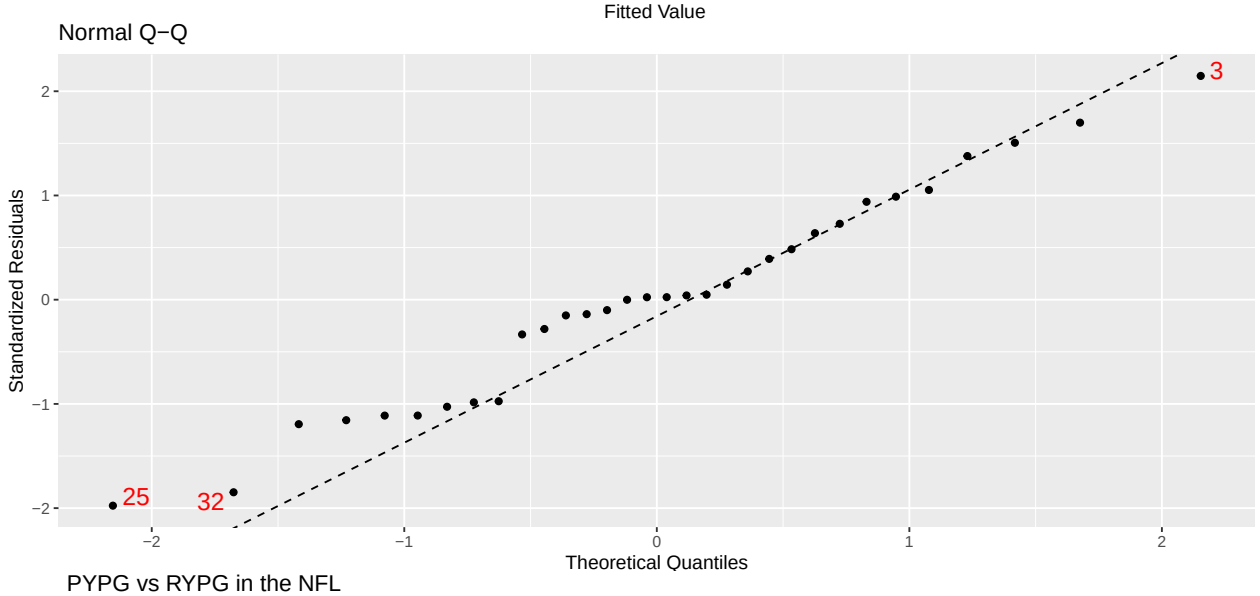
Because all the conditions for inference are met, we can now infer from the data. The PPG vs YPG model has a test statistic of 9.842055 which corresponds to a p-value of $6.620226Ö10^{-11}$.

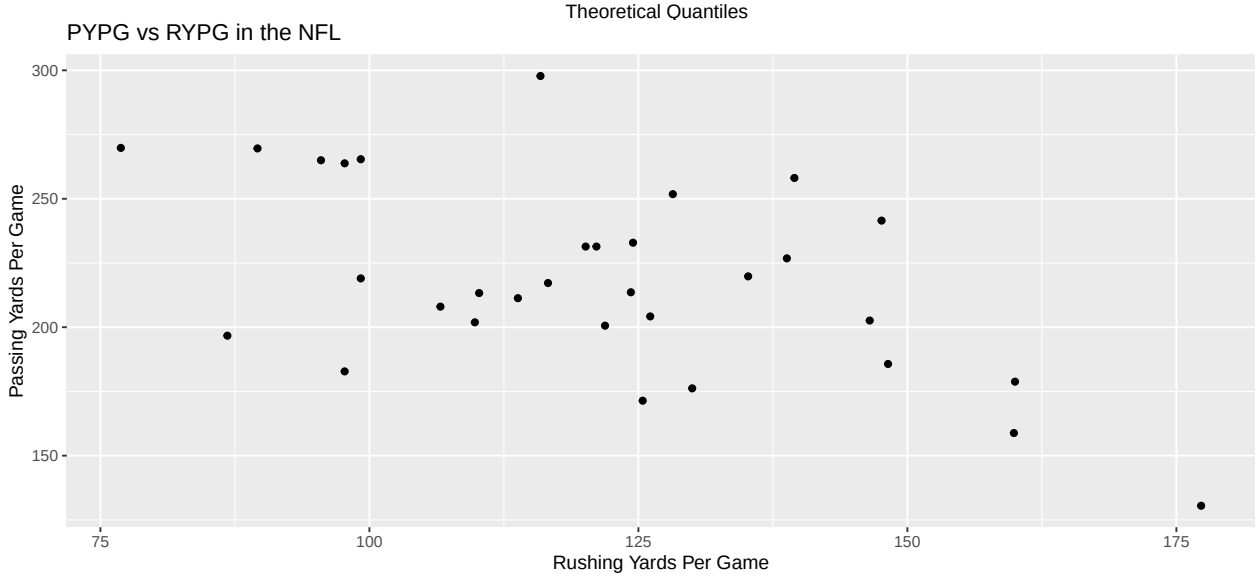
Multiple Linear Regression

We also decided to perform MLR to see if PPG is related to PYPG and RYPG. The conditions for inference for MLR are the same as SLR, with one additional condition. There must not be multicollinearity. Below are plots that asses the conditions for MLR.

```
## `geom_smooth()` using formula = 'y ~ x'
```







In MLR, the linearity condition is met if all predictors have a linear relationship. This linear relationship is

seen in the second and third scatter plots from the visualization section. Those scatter plots prove that all predictors have a linear relationship, even if some are not that good. The linearity condition, therefore, is met. Unlike linearity, equal variance and normality are the same as before. The residual versus fitted plot shows a decent spread across the width, so the equal variance condition, therefore, is met. The Q-Q plot shows a weak linear relationship, but there still is one. We cannot say for sure if normality is met, but we will still proceed with caution as if it is met. Multicollinearity occurs if two independent variables have a high correlation. The two independent variables are the two explanatory variables, PYPG and RYPG. The third plot, a scatter plot that shows the relationship between PYPG and RYPG, should be a good indicator of if multicollinearity is present or not. If this scatter plot showed a strong relationship, it would be evidence that multicollinearity is present. It does not, however. Additionally, variance inflation factor (VIF) is a direct measure of multicollinearity. Generally, a value of five or above means multicollinearity is present. The VIF value for both predictors is 1.421673, so multicollinearity is not present and the multicollinearity condition, therefore, is met.

All conditions for inference are met. The MLR model has two predictors, and will therefore have two test statistics and p-values. For PYPG, the test statistic is 9.814515 which corresponds to a p-value of $1.007533Ö10^{-10}$. For RYPG, the test statistic is 7.282251 which corresponds to a p-value of $5.086612Ö10^{-8}$.

Regression Conclusion

All p-values are essentially zero. For the SLR, this means that we have proven that PPG and YPG have a linear relationship. As YPG increases, so does PPG. For the MLR, this means that we can conclude that both predictors are statistically significant. Both RPYG and PYPG, therefore, significantly affect PPG.

College Data

Below we examine the College Football data set that we wrangled from sports-reference.com. Unlike the NFL data set, likely due to the large difference in number of teams, this data set had far less unnecessary statistics that had to be removed in order to create a tided table. The data was then examined for the same relationships as the NFL data set: YPG v.s. PPG, PYPG v.s. PPG, and RYPG v.s. PPG. Further, we then created three more scatter plots and repeated the process of analyzing them using SLR, and MLR.

College Visualizations

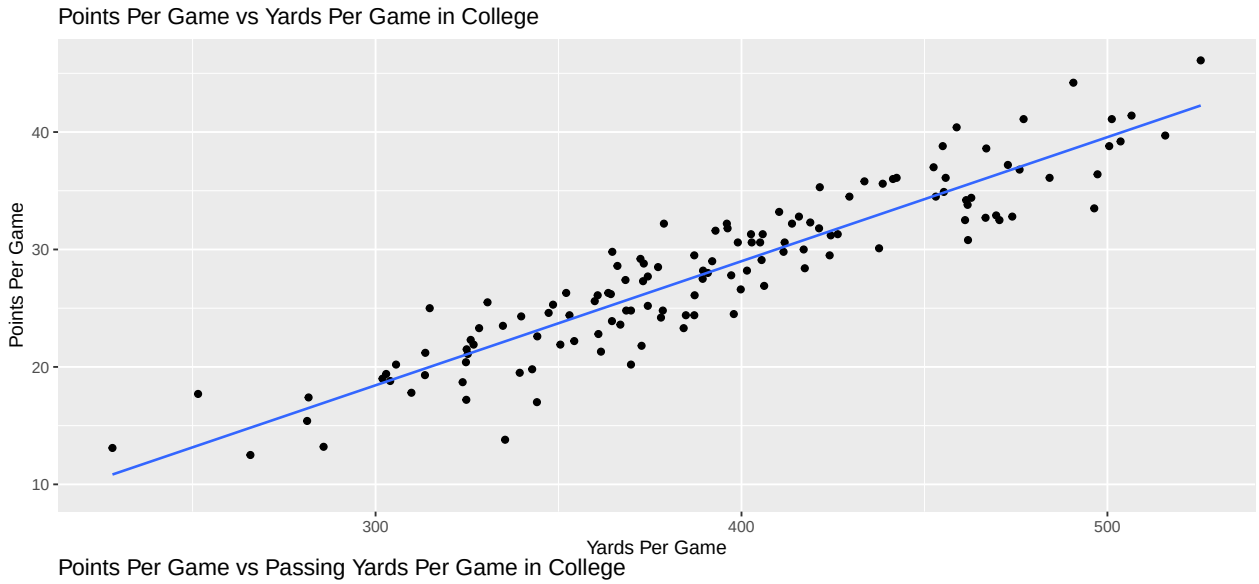
Below is a table that attempts to show a relationship between PPG and YPG in College.

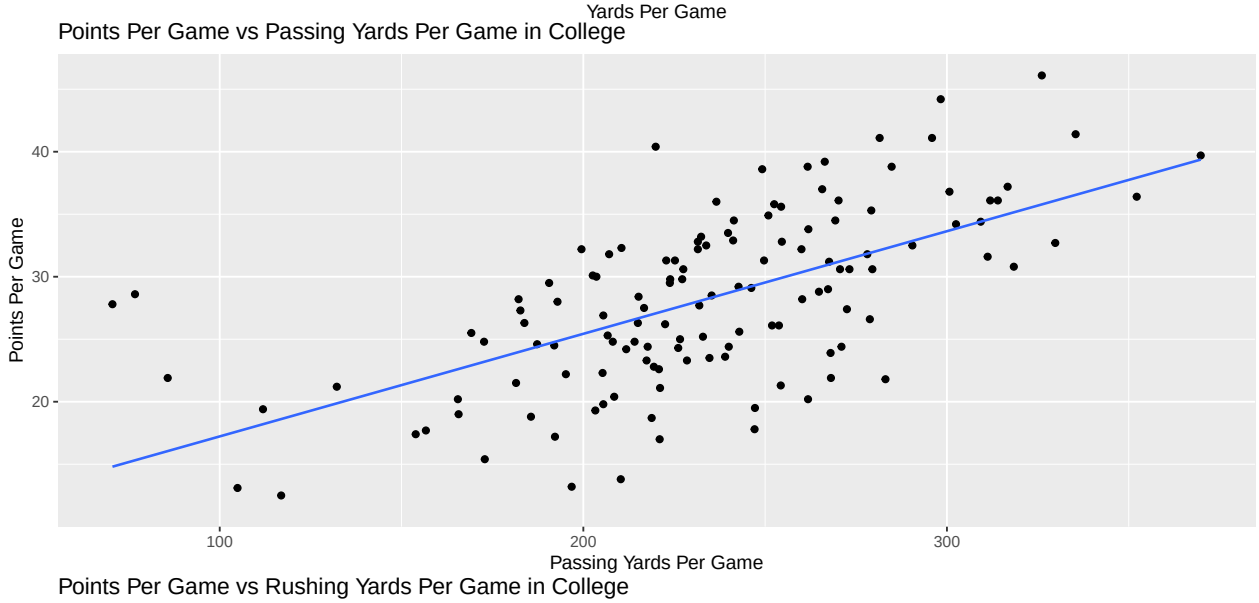
This table is the exact same as the NFL table in how it is formatted. It shows the top 14 teams' ranks in PPG (out of 131) in ascending order. It also shows the team's rank in YPG. While this table is not quite as perfect as the NFL table, it is still pretty good. The top ranked team in PPG was also the top ranked team in YPG. Additionally, all of the top 14 teams in PPG are in the top 29 in YPG. 29 divided by 131 is roughly 22%, so there appears to be a relationship between PPG and YPG in college as well as the NFL. However, we will conduct further analysis to determine if there truly is a relationship.

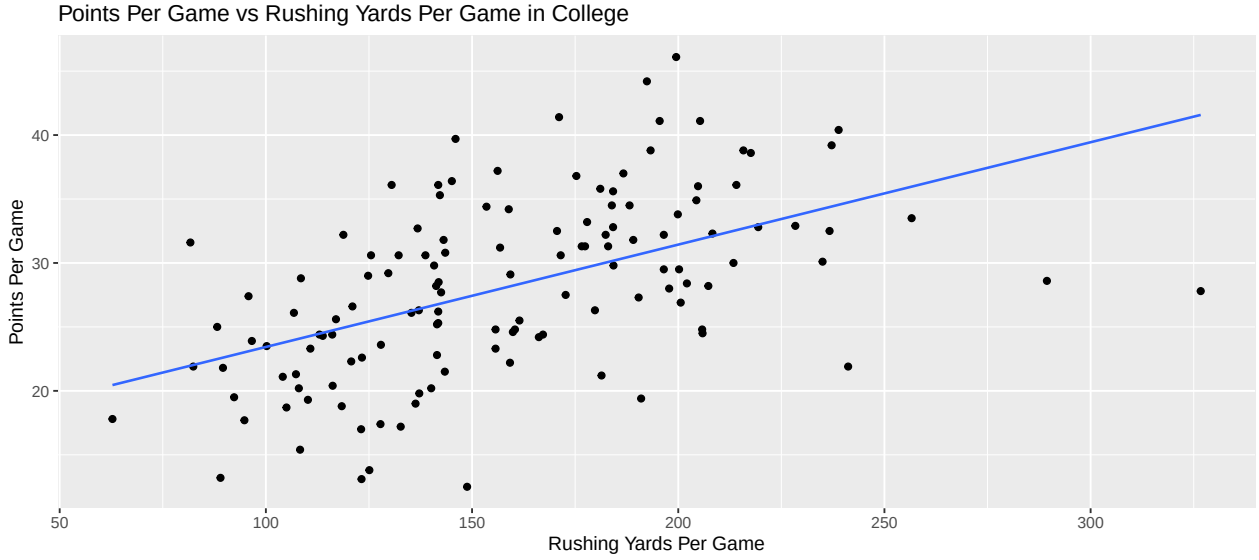
Delving deeper into our analysis, under the table are three scatter plots. They are the same plots in the same order as before, just with data from college teams as opposed to NFL teams. The first shows the relationship between PPG and YPG. The second shows the relationship between PPG and PYPG. The third shows the relationship between PPG and RYPG.

Table 2: Every college team's rank in PPG and YPG

	Team	YPG Rank	PPG Rank
103	Tennessee	1	1
80	Ohio State	9	2
116	USC	3	3
3	Alabama	11	4
35	Georgia	5	5
62	Michigan	24	6
125	Washington	2	7
114	UCLA	4	8
85	Oregon	6	9
101	TCU	27	10
117	Utah	17	11
94	SMU	14	12
45	James Madison	29	13
120	UTSA	12	14







Like before, all three plots show a positive linear relationship. All three appear to show a much stronger

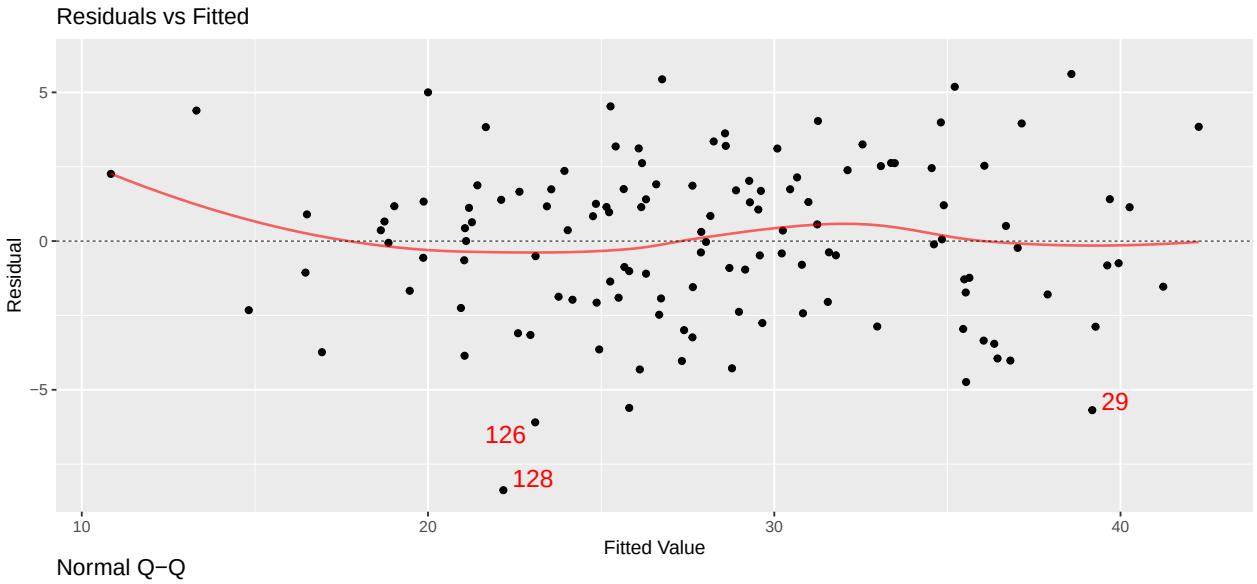
relationship than their NFL counterparts. But, like the NFL plots, PPG vs RYPG shows a much weaker relationship than the other two. Even though it is weaker, it is certainly not weak. The relationship between PPG and RYPG is relatively strong.

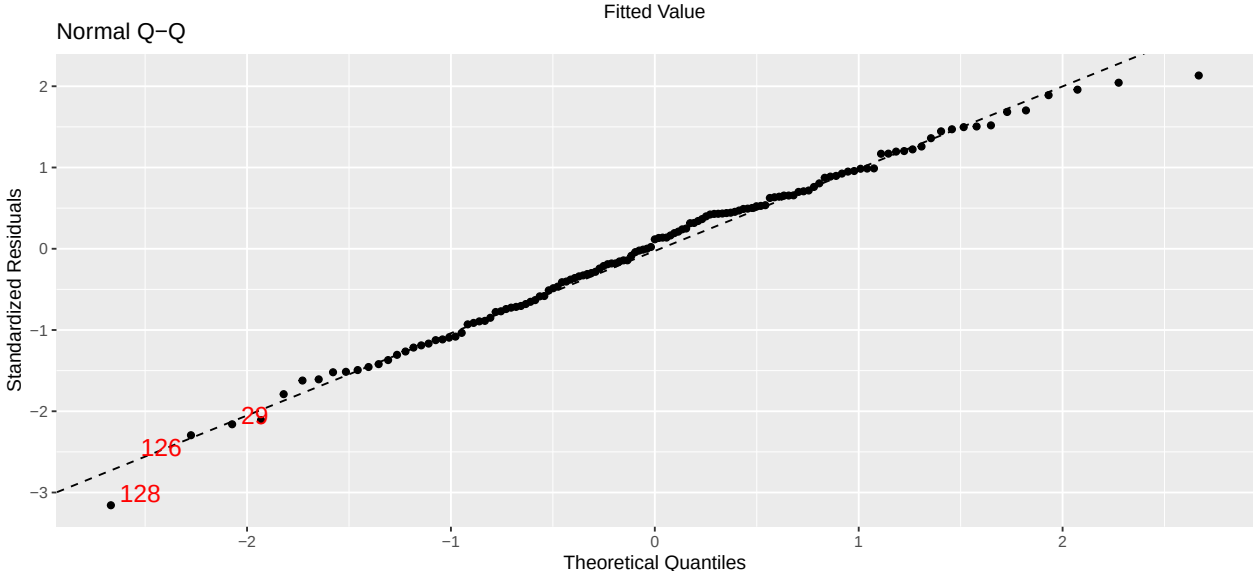
College Analysis

Simple Linear Regression

We wanted to perform SLR to see if PPG and YPG are related. Like before, the conditions for inference must be met. Below are two plots that attempt to assess these conditions.

```
## `geom_smooth()` using formula = 'y ~ x'
```





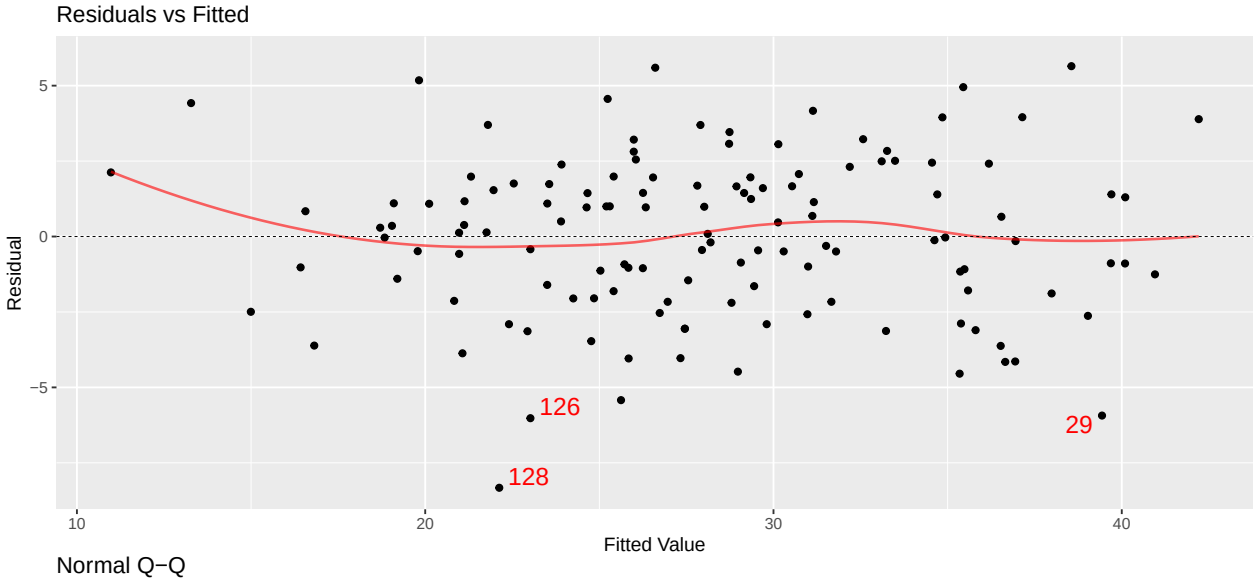
The conditions for inference, linearity, equal variance, and normality are the same for all instances of SLR. The data are linear because the scatter plot of PPG vs YPG shows a strong and obvious linear relationship. The linearity condition, therefore, is met. The residuals versus fitted plot shows a nice even spread. The equal variance condition, therefore, is met. Finally, the Q-Q plot shows a strong linear relationship, so the normality condition, therefore, is met.

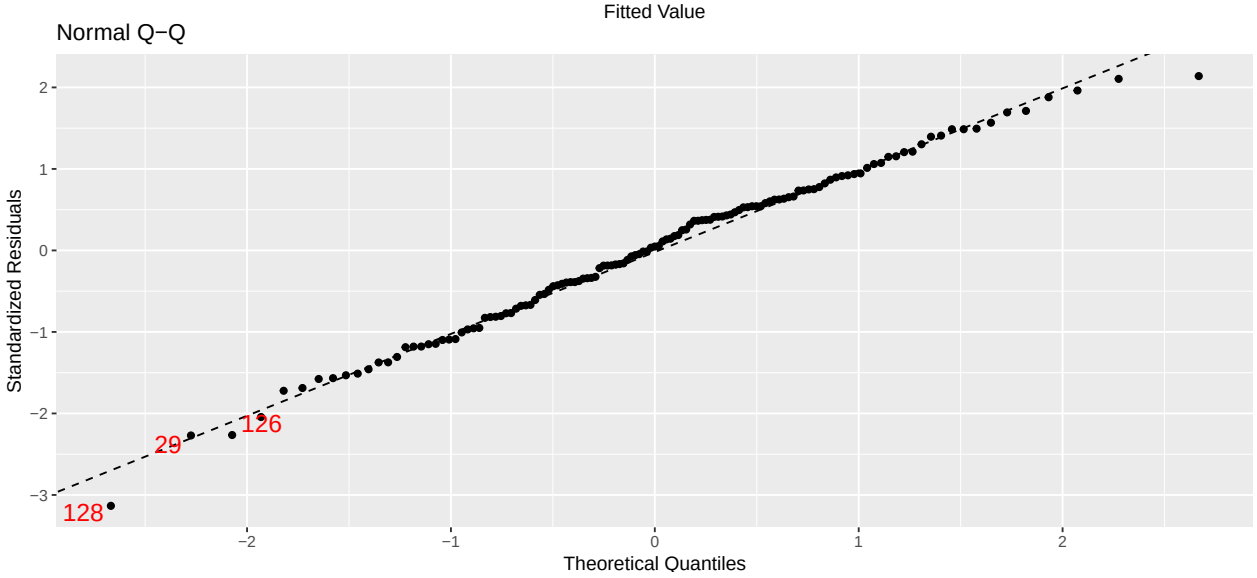
All conditions for inference are met. This model has a test statistic of 27.409820 which corresponds to a p-value of 1.213549Ö10^{-55}.

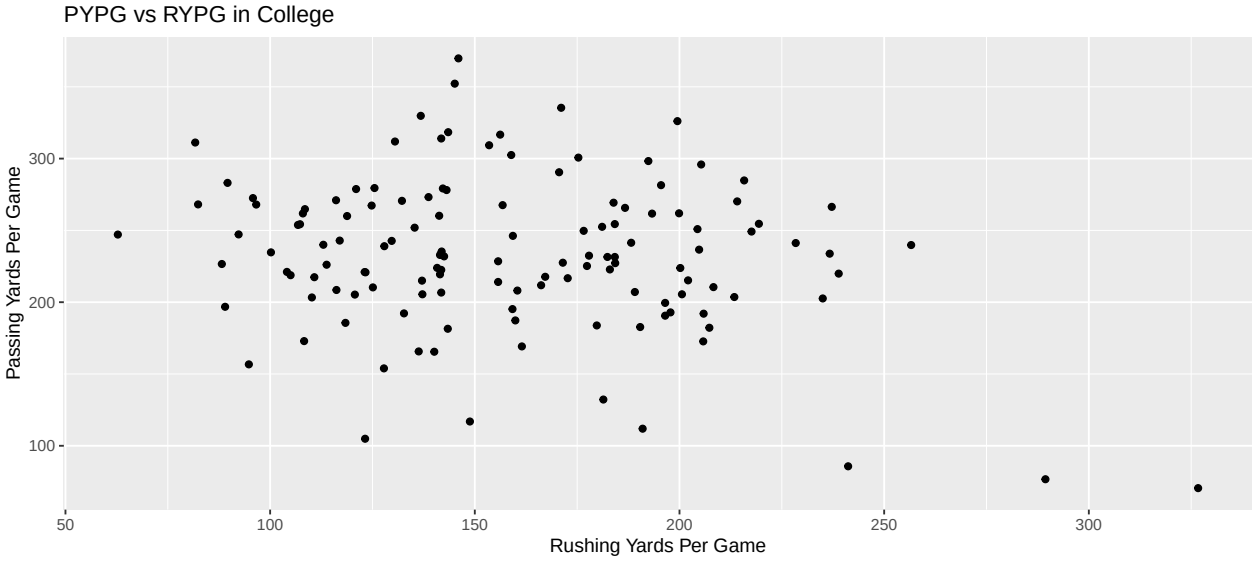
Multiple Linear Regression

Like the NFL analysis, we have also decided to perform MLR to see if PPG is related to PYPG and RYPG. The conditions that must be met are linearity, equal variance, normality, and multicollinearity. Below are plots that assess whether or not those conditions are met.

```
## `geom_smooth()` using formula = 'y ~ x'
```







The linearity condition is met because, as seen in the second and third scatter plot, all predictors have a relatively strong linear relationship. The residuals versus fitted plot shows a nice spread, so the equal variance condition is met. The Q-Q plot shows a very strong linear relationship, so the normality condition is met. Finally, because the scatter plot of PYPG vs RYPG does how show a relationship, and the individual VIF values are only 1.057908, multicollinearity is not present, so the multicollinearity condition, therefore, is met.

All conditions for inference are met. The MLR model has two predictors, and will therefore have two test statistics and p-values. For PYPG, the test statistic is 22.620628 which corresponds to a p-value of $1.494167Ö10^{-46}$. For RYPG, the test statistic is 20.271545 which corresponds to a p-value of $8.880033Ö10^{-42}$.

Regression Conclusion

We reach the same conclusion as the NFL data. All p-values are essentially zero. For the SLR, this means that we have proven that PPG and YPG have a linear relationship. As YPG increases, so does PPG. For the MLR, this means that we can conclude that both predictors are statistically significant. Both RPYG and PYPG, therefore, significantly affect PPG.

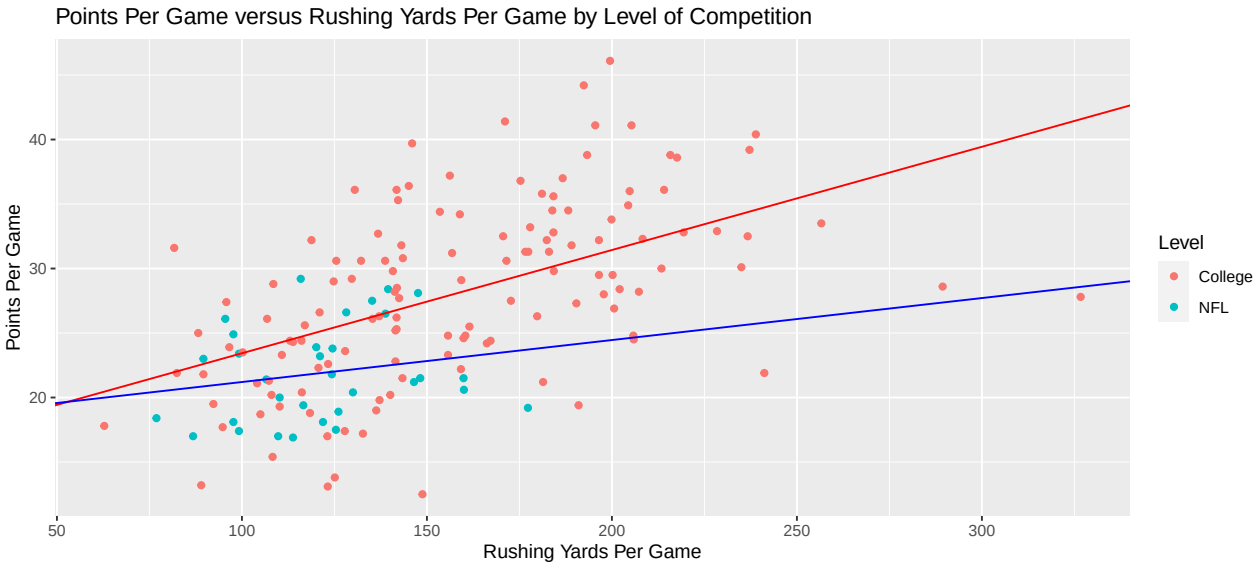
Comparison

Using these conclusions we can see that when comparing the YPG v.s. PPG relationship between the NFL and FBS, they both display similar results. Despite, the two leagues experiencing differing levels of competition, all 6 relationships over the two leagues were statistically significant. Below we are going to compare these relationships using comparative visualizations, specifically using scatter plots with least-square regression lines. We are then going to further compare these relationships using MLR. We're comparing all 6 of these relationships because we want to make sure that we check not only if the relationship between YPG and PPG differ between NFL and College, but if they differ between PYPG/RYPG and PPG as well. This extra layer of comparison just adds to the initial question about whether YPG affects PPG, as it helps us narrow down how the passing yards or rushing yards influence this relationship.

Comparative Visualizations

To see how the relationship between PPG and YPG, PYPG, and RYPG is different for the NFL versus College, scatter plots with the two on the same axis separated by level of competition would be helpful.

Below are three scatter plots, each containing points from both levels of competition. These plots do contain the least-squares regression line for each level. This was done to make it easier to see how the relationships differ at each level. The first plot shows the relationship between PPG and YPG by level of competition. The second shows the relationship between YPG and PYPG by level of competition. The third shows the relationship between YPG and RYPG by level of competition.



These three plots are very intriguing. The first plot shows that the relationship between PPG and YPG is very similar in both the NFL and College. Because the slopes of the lines of best fit are almost identical, the relationships are also almost identical. The other two plots show that the relationships between PPG and PYPG and PPG and RYPG differ for the two levels of competition. Whether these differences are statistically significant will be further explored.

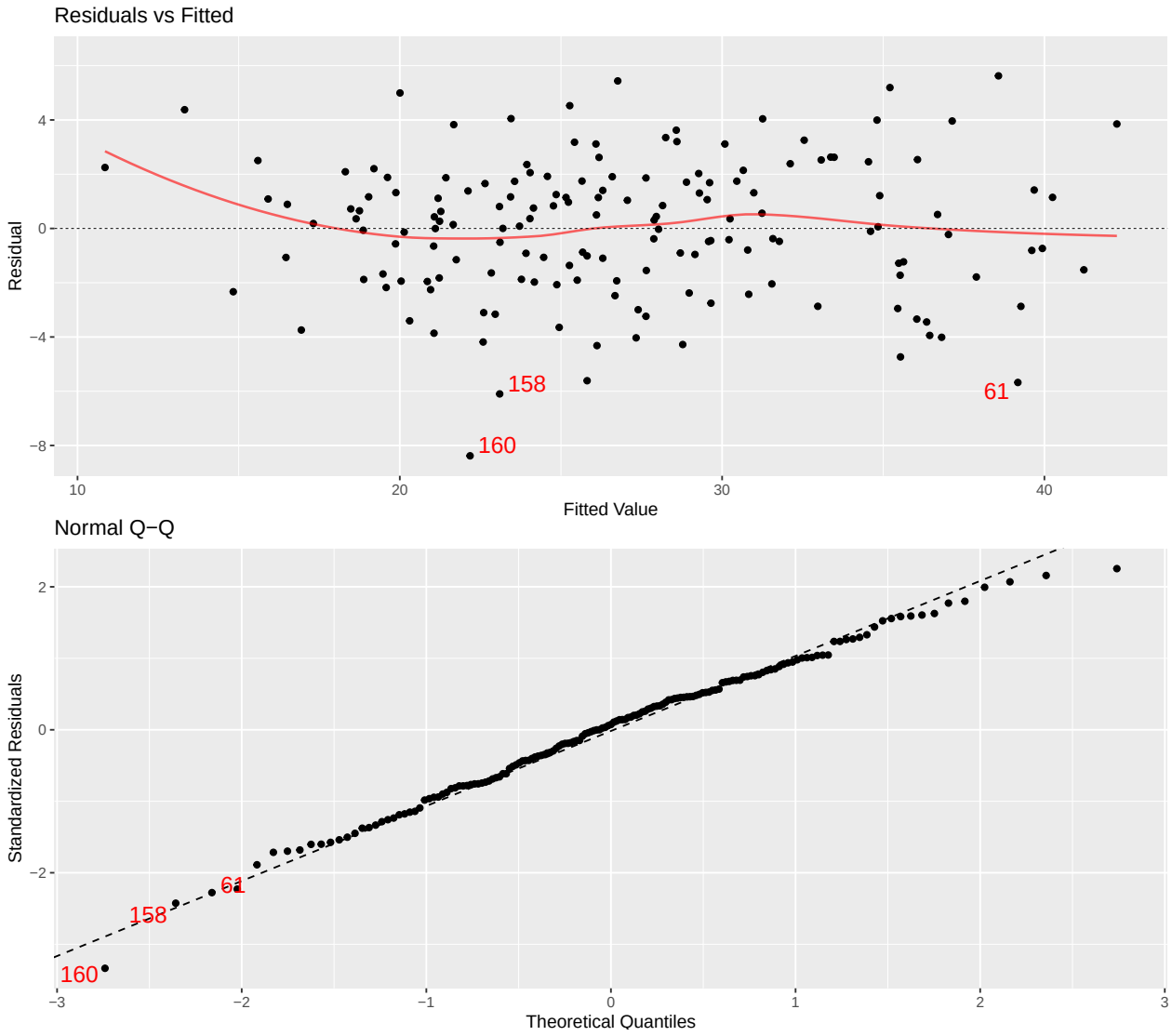
Comparative Analysis

To analyze whether the level of competition affects any of the three relationships we've been observing so far, we will perform more MLR.

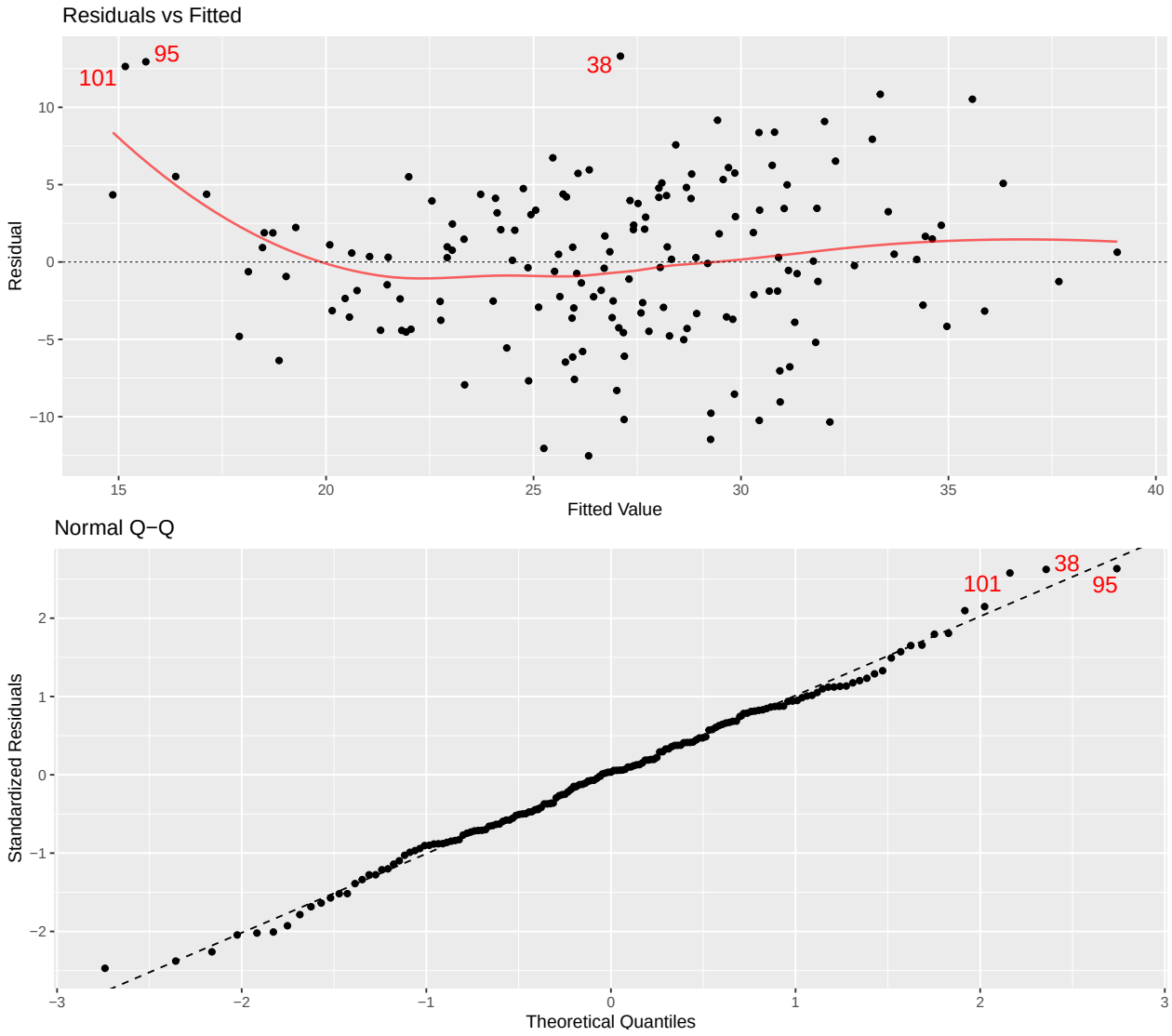
Conditions for Inference

To infer from the results of MLR, all conditions for inference must be met. The conditions are the same as before. Multicollinearity cannot be an issue this time, though, because there is only one quantitative predictor. So we can safely say the condition is met. Below are six plots, three residual versus fitted, and three Q-Q plots. The first residual versus fitted and Q-Q plot are about the the relationship between YPG and YPG by level of competition. The second residual versus fitted and Q-Q plot are about the the relationship between YPG and PYPG by level of competition. The third residual versus fitted and Q-Q plot are about the the relationship between YPG and RYPG by level of competition.

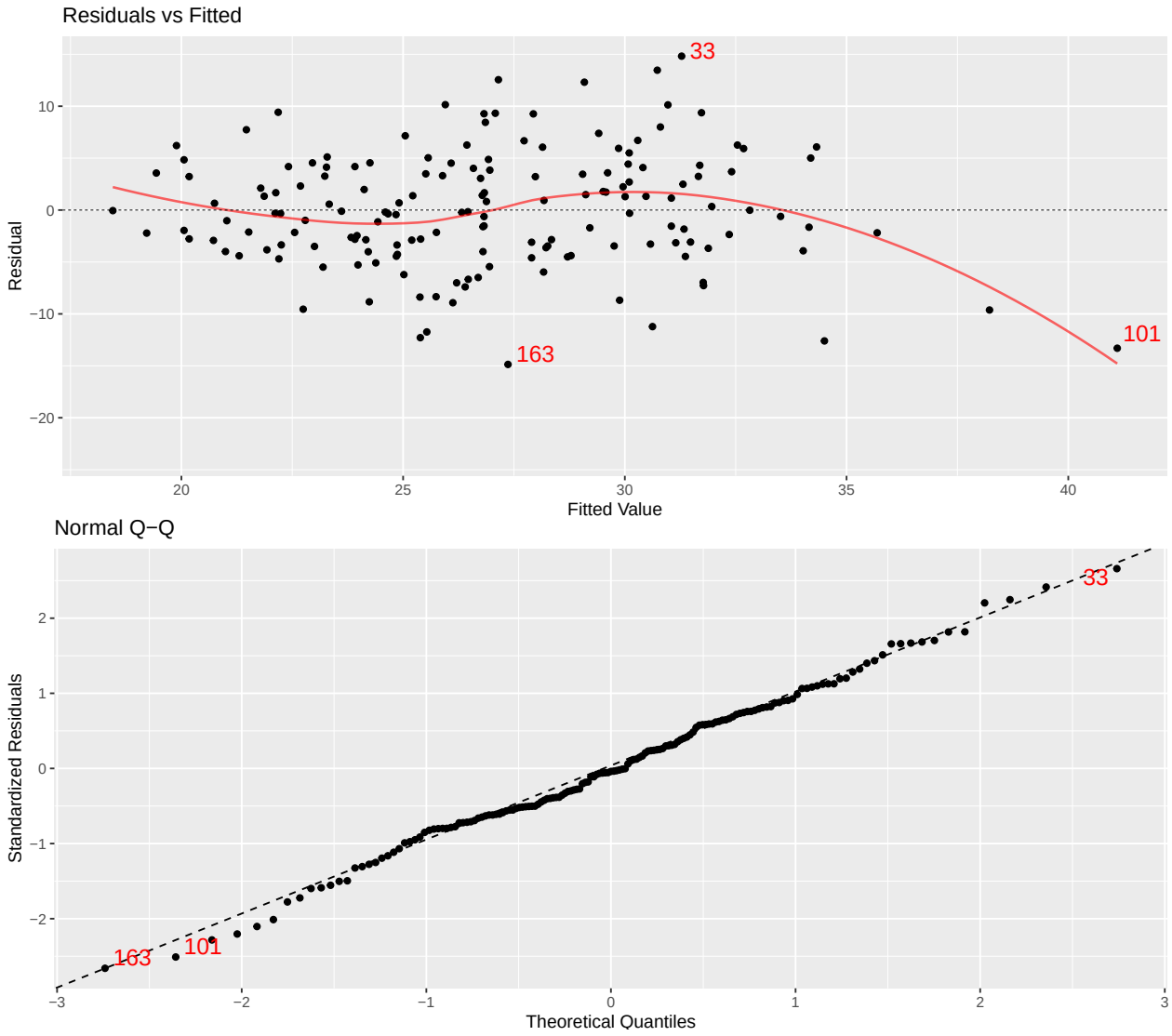
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



All six plots were lumped together instead of being analyzed individually because they all have the same result. All three residual versus fitted plots show an even spread, so all three models meet the equal variance condition. Similarly, all three Q-Q plots show a very strong linear relationship, so the normality condition is met for all three models. Linearity is also met for all three, because of the scatter plots shown earlier. All predictors have a linear relationship, so the linearity condition, therefore, is met for all three models. All conditions for all models are met.

Points per Game versus Yards per Game

The PPG vs YPG model has two predictors, one for YPG and one for level of competition. The YPG predictor has a test statistic of 29.837410 which corresponds to a p-value of $2.884682Ö10^{-67}$. For level of competition, the test statistic is -1.482403 which corresponds to a p-value of 0.1402001. The YPG term is significant at the 0.05 level, but the level term is not. The YPG term being significant means that YPG affects PPG, which we already knew. But, the level term being insignificant means that the relationships between PPG and YPG by level of competition are not significantly different. This means that there is not enough evidence to conclude that the relationship between YPG and PPG at the NFL level and the relationship between YPG and PPG at the college level are different.

Points per Game versus Passing Yards per Game

The PPG vs YPG model has two predictors, one for PYPG and one for level of competition. The PYPG predictor has a test statistic of 9.949669 which corresponds to a p-value of $1.861977Ö10^{-18}$. For level of competition, the test statistic is -5.039062 which corresponds to a p-value of $1.252203Ö10^{-6}$. The PYPG term is significant, which means that, like before, PYPG affects PPG, which we already knew. The level term being significant means that there is sufficient evidence to conclude that the relationships between PPG and PYPG by level of competition are different.

Points per Game versus Rushing Yards per Game

The PPG vs YPG model has two predictors, one for RYPG and one for level of competition. The RYPG predictor has a test statistic of 7.313959 which corresponds to a p-value of $1.177076Ö10^{-11}$. For level of competition, the test statistic is -2.860394 which corresponds to a p-value of $4.796173Ö10^{-3}$. The RYPG term is significant, which means that RYPG affects PPG, but we already knew that. The level term is significant, which means that like before, here is sufficient evidence to conclude that the relationships between PPG and RYPG by level of competition are different.

Conclusion

This project has explored the relationship between PPG and YPG, PPG and PYPG, and PPG and RYPG at the NFL level of competition and the collegiate level of competition. Then, it compared those relationships to see if the NFL relationship differed from the collegiate relationship. To accomplish this, we first needed data to an analyze. We scraped data from pro-football-reference.com and sports-reference.com. We then had to wrangle the data. Then, we performed various exploratory data analysis, the majority of which was creating visualizations to see if those relationships were worth exploring. We determined that they were, and then began to analyze. We first performed SLR twice on the relationship between PPG and YPG at each level of competition. After checking the conditions for inference, we determined that for both levels, YPG significantly affected PPG. Then we performed MLR twice on the relationship between PPG, and PYPG and RYPG at each level of competition. After checking the conditions for inference, we determined that for both levels, both PYPG and RYPG significantly affected PPG. Finally, we performed more MLR to determine if the relationships we observed at the different levels of competition were significant or not. We looked at the relationship between PPG and YPG and wanted to see if the relationship at the NFL level was different than the relationship at the collegiate level. After checking the conditions, we determined that the relationships were not significantly different. Finally, we performed two more MLRs. One examined the relationship between PPG and PYPG, and the other examined the relationship between PPG and RYPG. Both sought to see if the relationship at the NFL level was different than the relationship at the collegiate level. They both found that the relationships were indeed statistically significantly different.

Code Appendix

```
#tidyverse style guide
#load necessary packages
knitr::opts_chunk$set(echo = FALSE, dpi=300)
library("groundhog")
pkgs <- c('Stat2Data', 'tidyverse', 'mosaic', 'ggformula', 'Lock5Data', 'tinytex', 'car', 'leaps', 'HH'
groundhog.library(pkgs, '2023-11-29')
#create NFL table
nflTable <- read_html(
  x = "http://pfref.com/pi/share/d7riA"
) %>%
  html_elements(css = "table") %>%
  html_table()
nflOffense <- nflTable[[1]]
#delete unnecessary data
nflOffense <- subset(nflOffense, select=-c(1, 6:12, 14:18, 20:28))
nflOffense <- nflOffense[-c(1, 22:23, 36:38), ]

#rename columns
nflOffense <- nflOffense%>%rename(
  `Team`=1, `Games`=2, `Total Points`=3, `Total Yards`=4, `Total Passing Yards`=5, `Total Rushing Yards`

#squish team names
nflOffense$Team <- sub(".*\\s", "", nflOffense$Team)

#make every cell a number versus a character
nflOffense$Games <- as.numeric(nflOffense$Games)
nflOffense$`Total Points` <- as.numeric(nflOffense$`Total Points`)
nflOffense$`Total Yards` <- as.numeric(nflOffense$`Total Yards`)
nflOffense$`Total Passing Yards` <- as.numeric(nflOffense$`Total Passing Yards`)
nflOffense$`Total Rushing Yards` <- as.numeric(nflOffense$`Total Rushing Yards`)

#make per game stats
nflOffense$`Points Per Game` <- nflOffense$`Total Points` / nflOffense$Games
nflOffense$`Yards Per Game` <- nflOffense$`Total Yards` / nflOffense$Games
nflOffense$`Passing Yards Per Game` <- nflOffense$`Total Passing Yards` / nflOffense$Games
nflOffense$`Rushing Yards Per Game` <- nflOffense$`Total Rushing Yards` / nflOffense$Games

#remove unnecessary data
nflOffense <- subset(nflOffense, select=-c(2:6))

#round down to tenths place
nflOffense$`Points Per Game` <- round(nflOffense$`Points Per Game`, digits = 1)
nflOffense$`Yards Per Game` <- round(nflOffense$`Yards Per Game`, digits = 1)
nflOffense$`Passing Yards Per Game` <- round(nflOffense$`Passing Yards Per Game`, digits = 1)
nflOffense$`Rushing Yards Per Game` <- round(nflOffense$`Rushing Yards Per Game`, digits = 1)

#rearrange columns
nflOffense <- nflOffense[, c(1, 2, 4, 5, 3)]
#create new dataframe with just points per game
nflOffensePPG <- subset(nflOffense, select=-c(3:5))
#order the rows by PPG
```

```r
nfl0ffensePPG <- nfl0ffensePPG[order(nfl0ffensePPG$`Points Per Game`, decreasing = TRUE),]

#create new dataframe with just yards per game
nfl0ffenseYPG <- subset(nfl0ffense, select=-c(2:4))
#order the rows by YPG
nfl0ffenseYPG <- nfl0ffenseYPG[order(nfl0ffenseYPG$`Yards Per Game`, decreasing = TRUE),]

#rank NFL teams by PPG
nfl0ffensePPG <- nfl0ffensePPG%>%
  mutate(`PPG Rank`= c(1:32))
nfl0ffensePPG <- subset(nfl0ffensePPG, select=-c(2))

#rank NFL teams by YPG
nfl0ffenseYPG <- nfl0ffenseYPG%>%
  mutate(`YPG Rank`= c(1:32))
nfl0ffenseYPG <- subset(nfl0ffenseYPG, select=-c(2))

#merge ranked data frames
nflRanks <- merge(nfl0ffenseYPG, nfl0ffensePPG, by = "Team", all = TRUE)
nflRanks <- nflRanks[order(nflRanks$`PPG`, decreasing = FALSE),]

#only include the top 14 teams by PPG
nflRanks14 <- head(nflRanks, n=14)
#create a nice-looking table
nflRanks14%>%
  kable(
    caption="Every NFL team's rank in PPG and YPG",
    booktabs=TRUE,
    align=c("l", rep("c",6))
  )%>%
  kableExtra::kable_styling(
    bootstrap_options=c("striped", "condensed"),
    font_size=16
  )
#create scatter plot to show PPG vs YPG
gf_point(`Points Per Game`~`Yards Per Game`,
         data=nfl0ffense,
         linewidth=1,
         title= 'Points Per Game vs Yards Per Game in the NFL')%>%
  gf_lm()

#create scatter plot to show PPG vs PYPG
gf_point(`Points Per Game`~`Passing Yards Per Game`,
         data=nfl0ffense,
         linewidth=1,
         title= 'Points Per Game vs Passing Yards Per Game in the NFL')%>%
  gf_lm()

##create scatter plot to show PPG vs RYPG
gf_point(`Points Per Game`~`Rushing Yards Per Game`,
         data=nfl0ffense,
         linewidth=1,
         title= 'Points Per Game vs Rushing Yards Per Game in the NFL')%>%
```

```
  gf_lm()
#create an SLR model
nflTotalYardsSLRModel <- lm(`Points Per Game`~`Yards Per Game`, data=nflOffense)

#create plots to examine conditions for inference
mplot(nflTotalYardsSLRModel, which=1)
mplot(nflTotalYardsSLRModel, which=2)

#create a matrix for output so it doesn't output into the file
nflTotalYardsSLRSummary <- summary(nflTotalYardsSLRModel)
nflTotalYardsSLROutput <- nflTotalYardsSLRSummary[["coefficients"]]
#create an MLR model
nflMLRModel <- lm(`Points Per Game`~`Passing Yards Per Game`+`Rushing Yards Per Game`, data=nflOffense)

#create plots to examine conditions for inference
mplot(nflMLRModel, which=1)
mplot(nflMLRModel, which=2)
gf_point(`Passing Yards Per Game`~`Rushing Yards Per Game`, data=nflOffense, title='PYPG vs RYPG in the
nflVif <- vif(nflMLRModel)

#create a matrix for output so it doesn't output into the file
nflMLRSummary<- summary(nflMLRModel)
nflMLROutput <- nflMLRSummary[["coefficients"]]
#scrape data from website
collegeTable <- read_html(
  x = "https://www.sports-reference.com/cfb/years/2022-team-offense.html"
) %>%
  html_elements(css = "table") %>%
  html_table()
collegeOffense <- collegeTable[[1]]

#remove unnecessary columns
collegeOffense <- subset(collegeOffense, select=-c(1, 3, 5:7, 9:10, 12:14, 16:25))

#rename column titles to match with NFL titles
collegeOffense <- collegeOffense%>%rename(
   `Team`=1, `Points Per Game`=2, `Passing Yards Per Game`=3, `Rushing Yards Per Game`=4, `Yards Per Ga
 )

#remove unnecessary rows
collegeOffense <- collegeOffense[-c(1, 22:23, 44:45, 66:67, 88:89, 110:111, 132:133),]

#make every number numeric, because some were character
collegeOffense$`Points Per Game` <- as.numeric(collegeOffense$`Points Per Game`)
collegeOffense$`Passing Yards Per Game` <- as.numeric(collegeOffense$`Passing Yards Per Game`)
collegeOffense$`Rushing Yards Per Game` <- as.numeric(collegeOffense$`Rushing Yards Per Game`)
collegeOffense$`Yards Per Game` <- as.numeric(collegeOffense$`Yards Per Game`)
#create new dataframe with just points per game
collegeOffensePPG <- subset(collegeOffense, select=-c(3:5))
#order the rows by PPG
collegeOffensePPG <- collegeOffensePPG[order(collegeOffensePPG$`Points Per Game`, decreasing = TRUE),]

#create new dataframe with just yards per game
```

```r
collegeOffenseYPG <- subset(collegeOffense, select=-c(2:4))
#order the rows by YPG
collegeOffenseYPG <- collegeOffenseYPG[order(collegeOffenseYPG$`Yards Per Game`, decreasing = TRUE),]

#rank every team by PPG
collegeOffensePPG <- collegeOffensePPG%>%
  mutate(`PPG Rank`= c(1:131))
collegeOffensePPG <- subset(collegeOffensePPG, select=-c(2))

#rank every team by YPG
collegeOffenseYPG <- collegeOffenseYPG%>%
  mutate(`YPG Rank`= c(1:131))
collegeOffenseYPG <- subset(collegeOffenseYPG, select=-c(2))

#merge ranked data frames
collegeRanks <- merge(collegeOffenseYPG, collegeOffensePPG, by = "Team", all = TRUE)
collegeRanks <- collegeRanks[order(collegeRanks$`PPG`, decreasing = FALSE),]
collegeRanks14 <- head(collegeRanks, n=14)
#create a nice-looking table
collegeRanks14%>%
  kable(
    caption="Every college team's rank in PPG and YPG",
    booktabs=TRUE,
    align=c("l", rep("c",6))
  )%>%
  kableExtra::kable_styling(
    bootstrap_options=c("striped", "condensed"),
    font_size=16,
  )
#create scatter plot to show PPG vs YPG
gf_point(`Points Per Game`~`Yards Per Game`,
         data=collegeOffense,
         linewidth=1,
         title= 'Points Per Game vs Yards Per Game in College')%>%
  gf_lm()

#create scatter plot to show PPG vs PYPG
gf_point(`Points Per Game`~`Passing Yards Per Game`,
         data=collegeOffense,
         linewidth=1,
         title= 'Points Per Game vs Passing Yards Per Game in College')%>%
  gf_lm()

#create scatter plot to show PPG vs RYPG
gf_point(`Points Per Game`~`Rushing Yards Per Game`,
         data=collegeOffense,
         linewidth=1,
         title= 'Points Per Game vs Rushing Yards Per Game in College')%>%
  gf_lm()
#create an SLR model
collegeTotalYardsSLRModel <- lm(`Points Per Game`~`Yards Per Game`, data=collegeOffense)

#create plots to assess conditions
```

```r
mplot(collegeTotalYardsSLRModel, which=1)
mplot(collegeTotalYardsSLRModel, which=2)

#create matrix so output isn't in file
collegeTotalYardsSLRSummary <- summary(collegeTotalYardsSLRModel)
collegeTotalYardsSLROutput <- collegeTotalYardsSLRSummary[["coefficients"]]
#create an SLR model
collegeMLRModel <- lm(`Points Per Game`~`Passing Yards Per Game`+`Rushing Yards Per Game`, data=college

#create plots to assess conditions
mplot(collegeMLRModel, which=1)
mplot(collegeMLRModel, which=2)
gf_point(`Passing Yards Per Game`~`Rushing Yards Per Game`, data=collegeOffense, title='PYPG vs RYPG in
collegeVif <- vif(collegeMLRModel)

#create matrix so output isn't in file
collegeMLRSummary<- summary(collegeMLRModel)
collegeMLROutput <- collegeMLRSummary[["coefficients"]]
#add level of competition to data frames
collegeOffense <- collegeOffense%>%
  mutate(`Level`= "College")
nflOffense <- nflOffense%>%
  mutate(`Level`= "NFL")

#merge data frames
offense <- rbind(nflOffense,collegeOffense)

#create new variable that is categorical but still a number for future MLR
offense$IndLevel <- as.numeric(offense$Level=="NFL")

#make visualization of PPG vs YPG by level and include line of best fit
offenseModel <- lm(`Points Per Game` ~ `Yards Per Game`+IndLevel+`Yards Per Game`*IndLevel, data = offe
b0 <- summary(offenseModel)$coeff[1,1]
b1 <- summary(offenseModel)$coeff[2,1]
b2 <- summary(offenseModel)$coeff[3,1]
b3 <- summary(offenseModel)$coeff[4,1]
gf_point(`Points Per Game` ~ `Yards Per Game`,
         color=~Level,
         data = offense,
         title= "Points Per Game versus Yards Per Game by Level of Competition") %>%
gf_abline(intercept = b0, slope = b1, color = "red") %>%
gf_abline(intercept = b0+b2, slope = b1+b3, color = "blue")

#make visualization of PPG vs PYPG by level and include line of best fit
passingOffenseModel <- lm(`Points Per Game` ~ `Passing Yards Per Game`+IndLevel+`Passing Yards Per Game
pb0 <- summary(passingOffenseModel)$coeff[1,1]
pb1 <- summary(passingOffenseModel)$coeff[2,1]
pb2 <- summary(passingOffenseModel)$coeff[3,1]
pb3 <- summary(passingOffenseModel)$coeff[4,1]
gf_point(`Points Per Game` ~ `Passing Yards Per Game`,
         color=~Level,
         data = offense,
         title= "Points Per Game versus Passing Yards Per Game by Level of Competition") %>%
```

```
gf_abline(intercept = pb0, slope = pb1, color = "red") %>%
gf_abline(intercept = pb0+pb2, slope = pb1+pb3, color = "blue")

#make visualization of PPG vs RYPG by level and include line of best fit
rushingOffenseModel <- lm(`Points Per Game` ~ `Rushing Yards Per Game`+IndLevel+`Rushing Yards Per Game
rb0 <- summary(rushingOffenseModel)$coeff[1,1]
rb1 <- summary(rushingOffenseModel)$coeff[2,1]
rb2 <- summary(rushingOffenseModel)$coeff[3,1]
rb3 <- summary(rushingOffenseModel)$coeff[4,1]
gf_point(`Points Per Game` ~ `Rushing Yards Per Game`,
         color=~Level,
         data = offense,
         title= "Points Per Game versus Rushing Yards Per Game by Level of Competition") %>%
gf_abline(intercept = rb0, slope = rb1, color = "red") %>%
gf_abline(intercept = rb0+rb2, slope = rb1+rb3, color = "blue")
#create MLR models
totalYardsMLRModel <- lm(`Points Per Game` ~ `Yards Per Game` + IndLevel , data=offense)
passingYardsMLRModel <- lm(`Points Per Game` ~ `Passing Yards Per Game` + IndLevel , data=offense)
rushingYardsMLRModel <- lm(`Points Per Game` ~ `Rushing Yards Per Game` + IndLevel , data=offense)

#make plots to assess conditions
mplot(totalYardsMLRModel, which=1)
mplot(totalYardsMLRModel, which=2)
mplot(passingYardsMLRModel, which=1)
mplot(passingYardsMLRModel, which=2)
mplot(rushingYardsMLRModel, which=1)
mplot(rushingYardsMLRModel, which=2)

#make matrix so output isn't in file
totalYardsMLRSummary<- summary(totalYardsMLRModel)
totalYardsMLROutput <- totalYardsMLRSummary[["coefficients"]]
passingYardsMLRSummary<- summary(passingYardsMLRModel)
passingYardsMLROutput <- passingYardsMLRSummary[["coefficients"]]
rushingYardsMLRSummary<- summary(rushingYardsMLRModel)
rushingYardsMLROutput <- rushingYardsMLRSummary[["coefficients"]]
```