Project RD

John Gleason

5/4/2022

Intro

What:

For my project I examined a PGA Tour dataset that contained players stats from 2019 to 2015. I am using this data to shed light on what area of the game effective the player score the most. I want to see what area of the game, Tee shots, irons shots or putting, separate the elite golfers from the rest of the pack.

Why:

A new approach to the game of golf on the PGA tour has caught some momentum in the last few years. Historically, the thought was to score well on tour players must be extremely efficient on the putting green. In today game, there are a hand full of players that are beginning to focus more on driving distance than putting efficiency. I want to dig deeper into which approach is backed by the numbers.

How:

I have pulled 5 years of data from pgatour.com site for the top 200 players in the world. I modified the dataset to contain categories that fit the scope of the project and reduced the years of data from 5 to 1 create a smaller sample size. After pulling and modifying the data I will begin exploring the data. I will create a histogram for all the variables to show the normal distribution. Then I will find the correlation between the explanatory and response variables. Finally, I will create a regression model using Scoring Average as my response variable and Driving Distance, Greens In Regulation (GIR) and Average Putts Per Round as my explanatory variables. This model will hopefully show me what explanatory variable has the greatest and least effect on the response variable.

Body

Audience Knowledge:

The dataset that will be used to determine what area of the golf game effect the score the most will contain 2019 PGA player data. The dataset has column contain Players Name, Driving Distance, Putts Per Round, FedEx Cup Points, Greens In Regulation (GIR) and Scoring Average. Driving Distance measures the average length the golf ball goes off the tee. GIR is the percentage the player is able to hit the green in regulation, meaning if the hole is a par 4 they reach the green on their second shot. Average Putts Per Round is how many time the player hits the ball when it is on the green over a 18 hole period.

Specific Topic:

I have chosen this as my response variable to be Scoring Average as the Fed Ex Cup Point System can be skewed because different tournament award different amount of points. I am using Driving Distance, GIR and Putts Per Round as the explanatory variable and categorize them as Driving distance = Long Game, GIR = Mid Game, Putt per Round = Short Game

```
library(readxl)
library(tidyverse)
```

```
## Warning: replacing previous import ’lifecycle::last_warnings’ by
## ’rlang::last_warnings’ when loading ’pillar’
```

```
## Warning: replacing previous import ’lifecycle::last_warnings’ by
## ’rlang::last_warnings’ when loading ’hms’
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.6
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package ’tidyr’ was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(rmarkdown)
```

```
pgafulldata_1_copy <- read_excel("~/Desktop/pgafulldata (1) copy.xlsx")
```

```
head(pgafulldata_1_copy)
```

```
## # A tibble: 6 x 11
##   ‘PLAYER NAME‘    DrivingDistance ‘Approaches fr~‘ PuttsPerRound FedexCupPoints
##   <chr>                      <dbl> <chr>                    <dbl>          <dbl>
## 1 A.J. McInerney ~             NA   <NA>                     NA               NA
## 2 Aaron Wise (201~             NA   <NA>                     NA               NA
## 3 Aaron Wise (201~            303. "33’ 5\""                 29.2           1086
## 4 Aaron Wise (201~            303. "33’ 10\""                29.2            400
## 5 Abraham Ancer (~            276. "32’ 4\""                 29.1            147
## 6 Abraham Ancer (~            296. "33’ 0\""                 28.9            589
## # ... with 6 more variables: AverageDistancePuttsmade <chr>, GIR <dbl>,
## #   OfficialMoney <dbl>, DrivingAccuracyPercentage <dbl>, ScoringAverage <dbl>,
## #   FedexPointCat <dbl>
```

Step 1:Clean Data

The first step in my project is to clean my data set. To do this I must remove all the columns with null values, separate column 1 into 2 different columns (Player Name and Year), remove the columns that I do not need for this project and finally filtering on just 2019 data.

```
pga_data_2 <- pgafulldata_1_copy %>% na.omit()
```

```
pga_data_2$Year.Date <- str_sub(pga_data_2$'PLAYER NAME', -6)
```

```
pga_data_2$PLAYER.NAME <- str_sub(pga_data_2$'PLAYER NAME',0 ,-7)
```

```
PGA <- pga_data_2[-c(1,3,6,8,9)]
```

```
PGA2019 <- PGA%>% top_n(177, Year.Date)
```

```
head(PGA2019)
```

```
## # A tibble: 6 x 8
##   DrivingDistance PuttsPerRound FedexCupPoints   GIR ScoringAverage
##             <dbl>         <dbl>          <dbl> <dbl>          <dbl>
## 1            303.          29.2            400  80.1           70.7
## 2            293.          28.8            622  74.6           70.6
## 3            291.          28.7            818  77.2           70.5
## 4            292           29.2            719  75.7           71.5
## 5            301.          29.0            524  78.9           70.8
## 6            299.          29.0           1124  81.2           69.7
## # ... with 3 more variables: FedexPointCat <dbl>, Year.Date <chr>,
## #   PLAYER.NAME <chr>
```

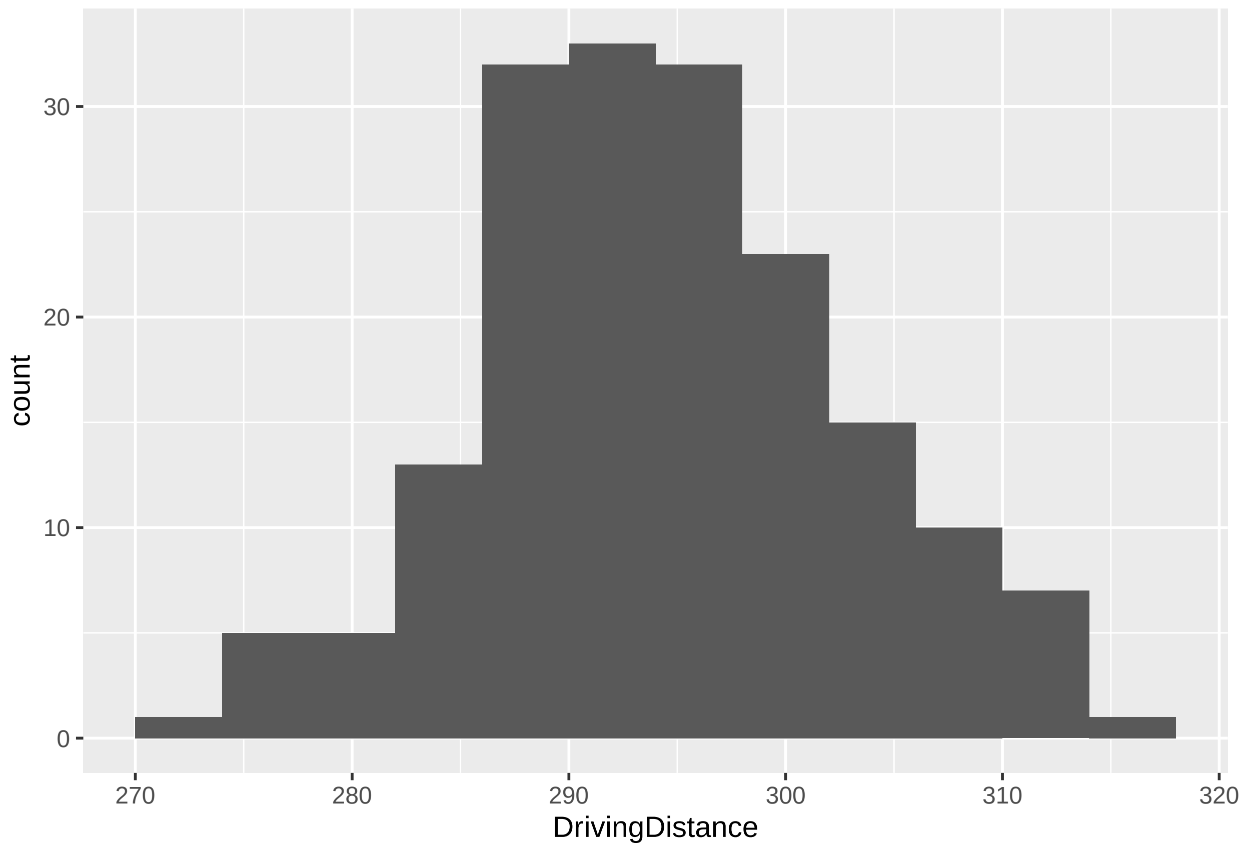
Step 1 Continued:

As you can see below I have cleaned me data set and I am now ready to create my model

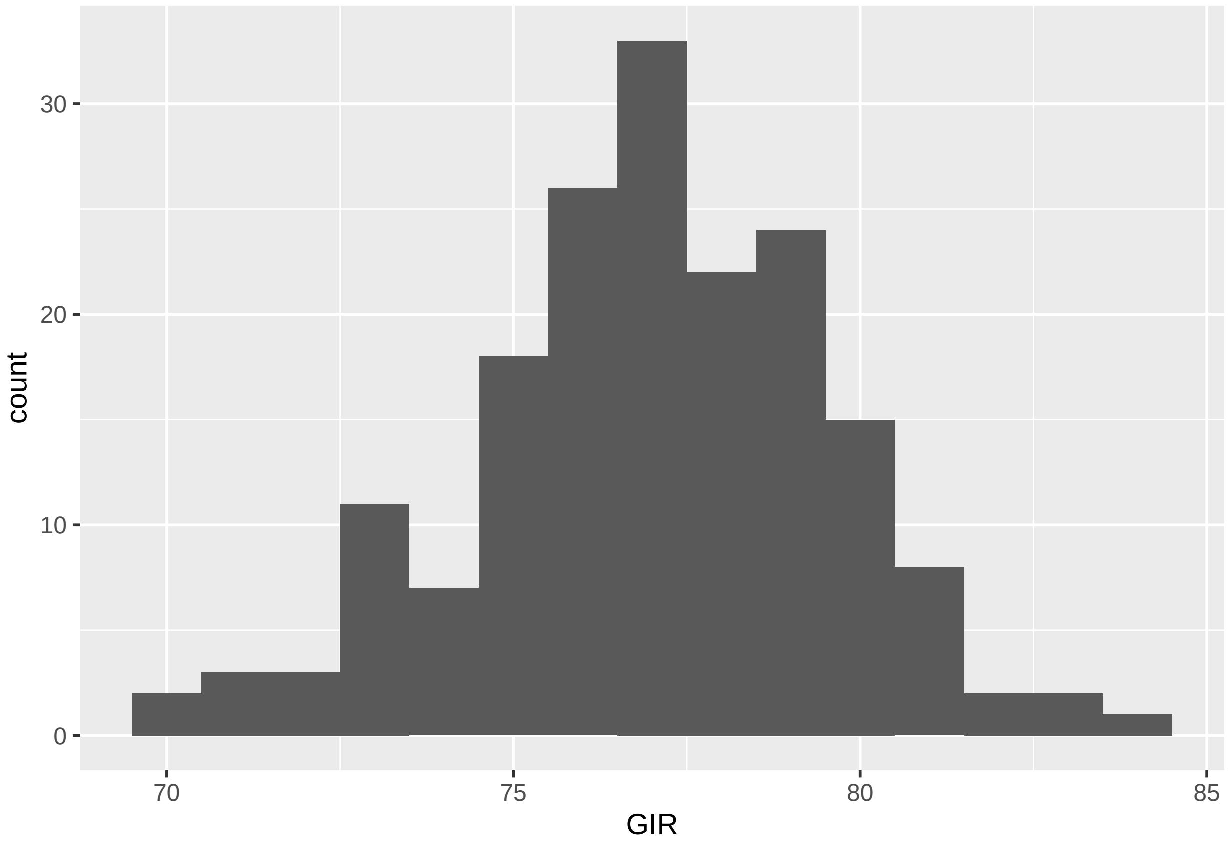
Step 2: Normal Distrubtion

Next, I will be testing for normal distribution between my explanatory variables. By creating a histogram for each explanatory variable it will give me an idea of the data distribution that I will be dealing with

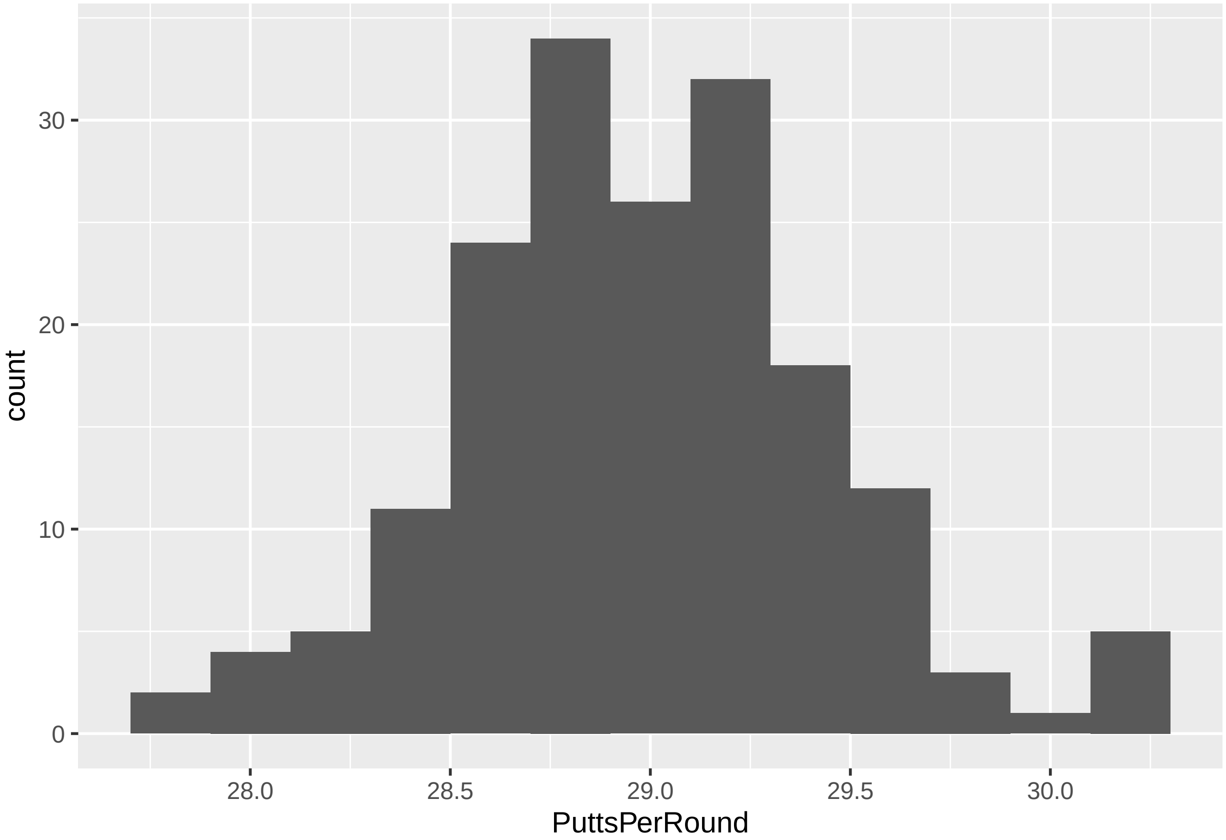
```
ggplot(PGA2019) + geom_histogram(aes(x=DrivingDistance), binwidth = 4)
```



```
ggplot(PGA2019) + geom_histogram(aes(x=GIR), binwidth = 1)
```



```
ggplot(PGA2019) + geom_histogram(aes(x=PuttsPerRound), binwidth = .2)
```



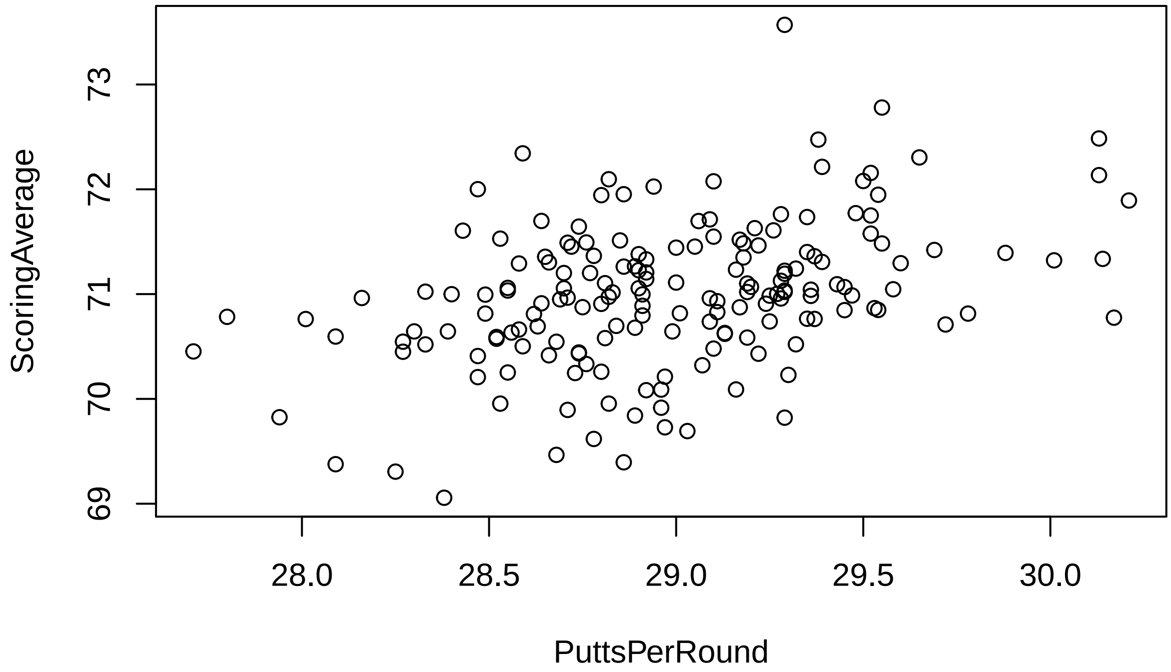
Step 2 continued

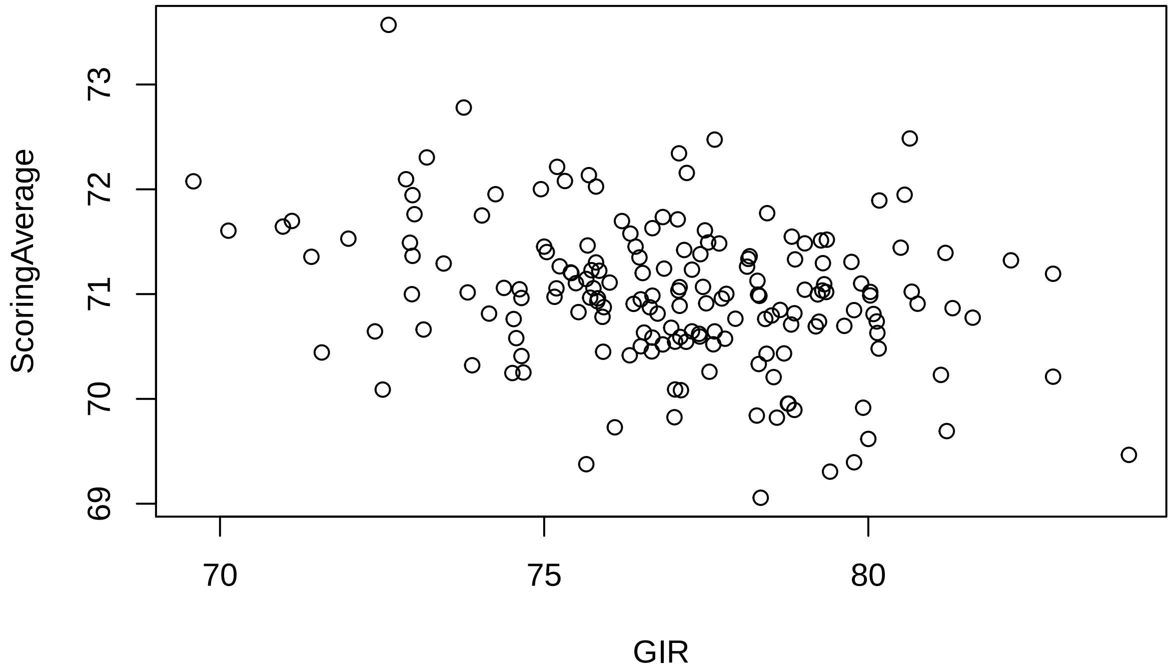
As see above, all three of my explanatory variables Driving Distance, GIR and Putting Per Round Average have a normal distribution.

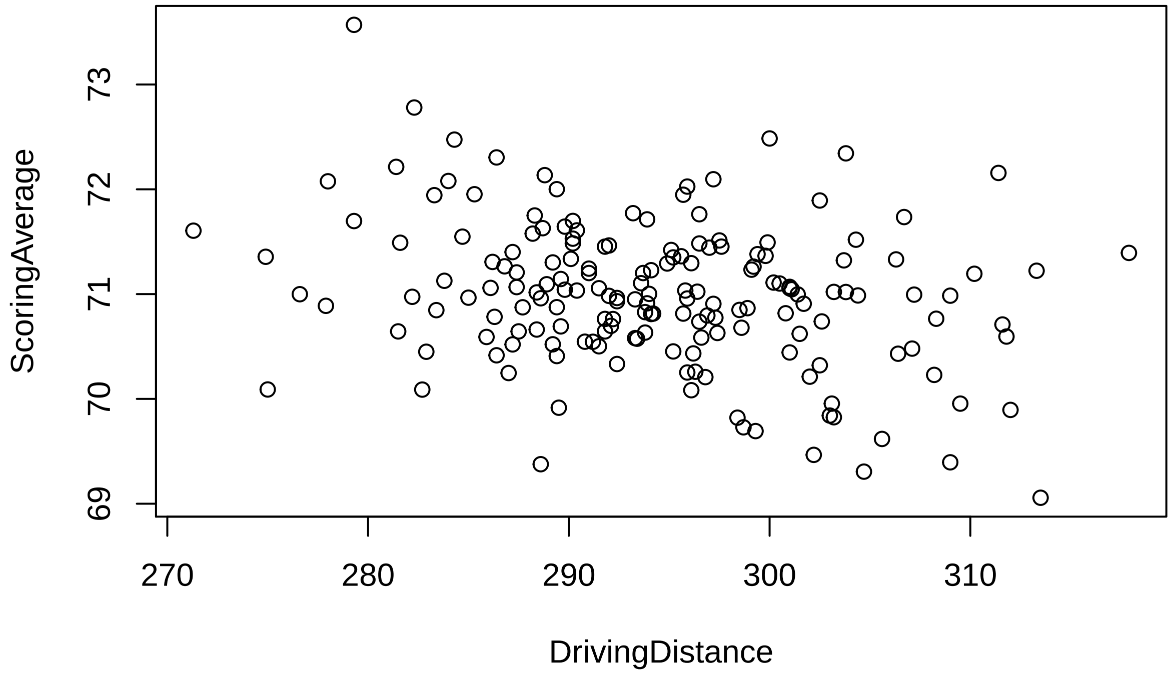
Step 3: Correlation

Before creating my regression model, I must look at the correlation between each explanatory variable compared to the response variable. Below I have created a scatter play that will show me the strength of the correlation be each variable.

```
plot(ScoringAverage ~ PuttsPerRound + GIR + DrivingDistance, data = PGA2019)
```


ScoringAverage
69
70
71
72
73
28.0
28.5
29.0
29.5
30.0
PuttsPerRound


ScoringAverage
69
70
71
72
73
70
75
80
GIR



```
cor(PGA2019$ScoringAverage, PGA2019$GIR)
```

```
## [1] -0.3067881
```

```
cor(PGA2019$ScoringAverage, PGA2019$DrivingDistance)
```

```
## [1] -0.2909246
```

```
cor(PGA2019$ScoringAverage, PGA2019$PuttsPerRound)
```

```
## [1] 0.4030289
```

Step 3 continued

As seen above the rank of strength of correlation is Putts Per Round (40%), GIR (30%) and finally Driving Distance (29%). This rank will feed into the order I put the variables into my regression model.

Step 4: Regression

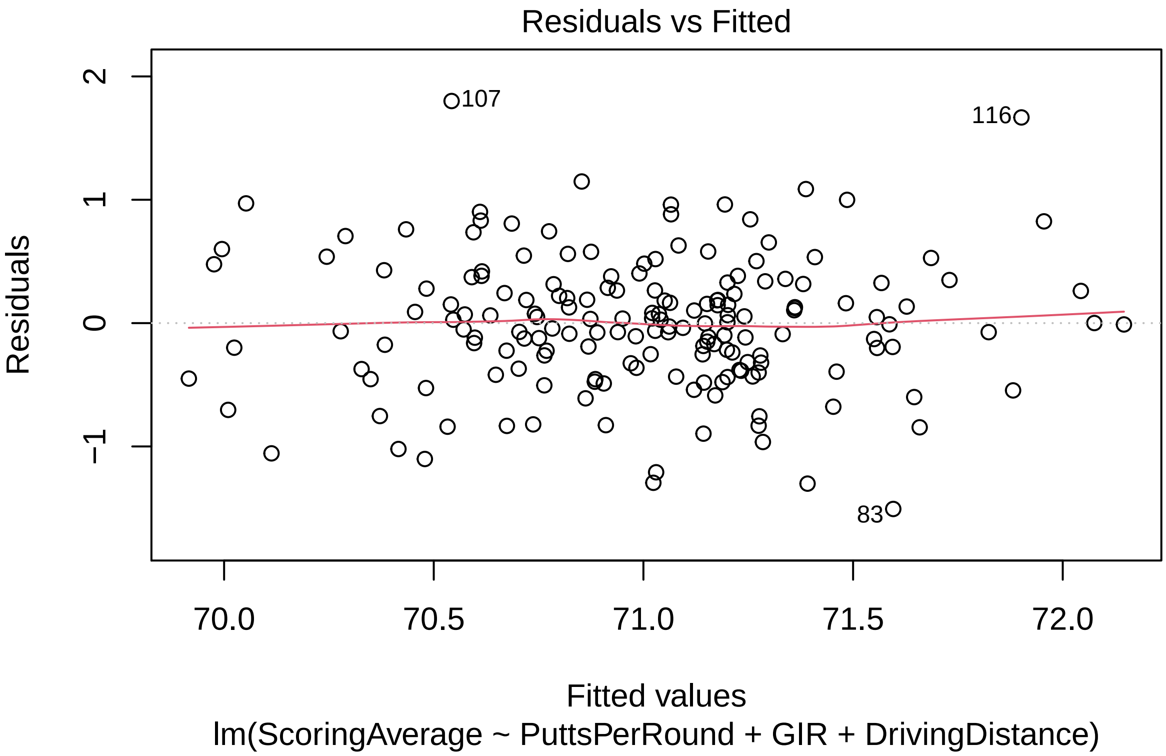
I have created a regression model based on the correlation of the explanatory variable compared to my response variable. Variable order in the model is Putts per round + GIR + Driving distance to find out the scoring average of the player.

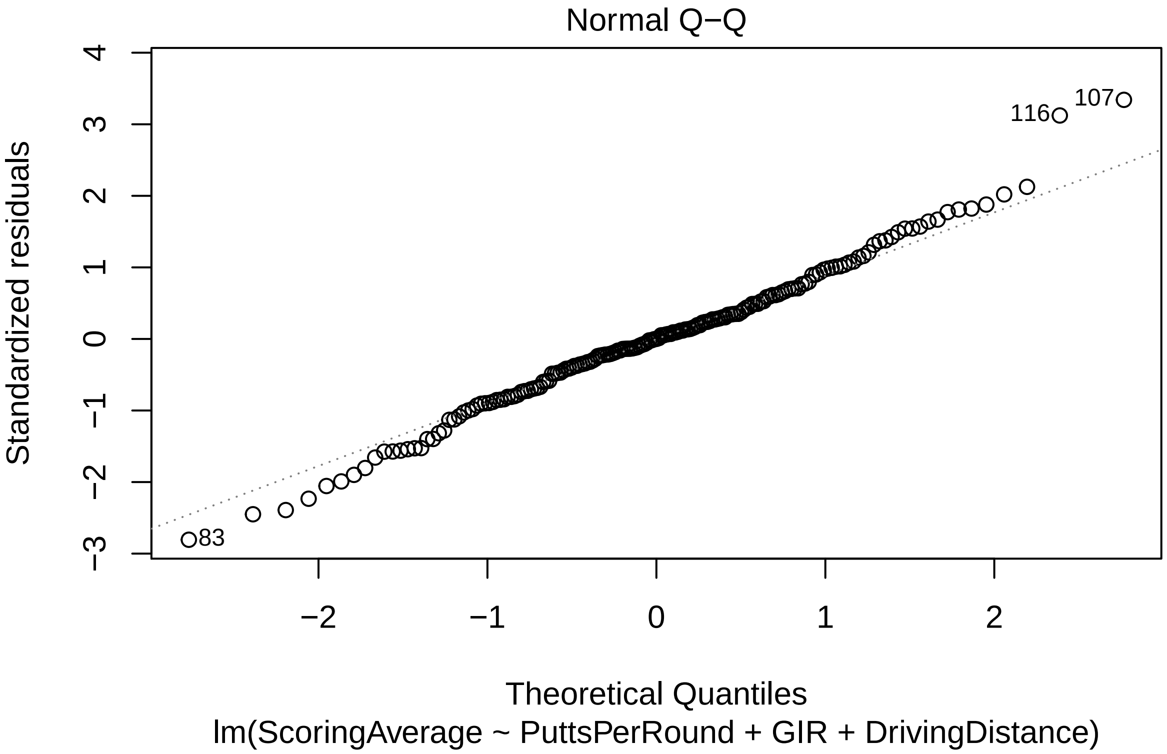
```
mode19<- lm(ScoringAverage ~PuttsPerRound + GIR+ DrivingDistance, data = PGA2019)
```

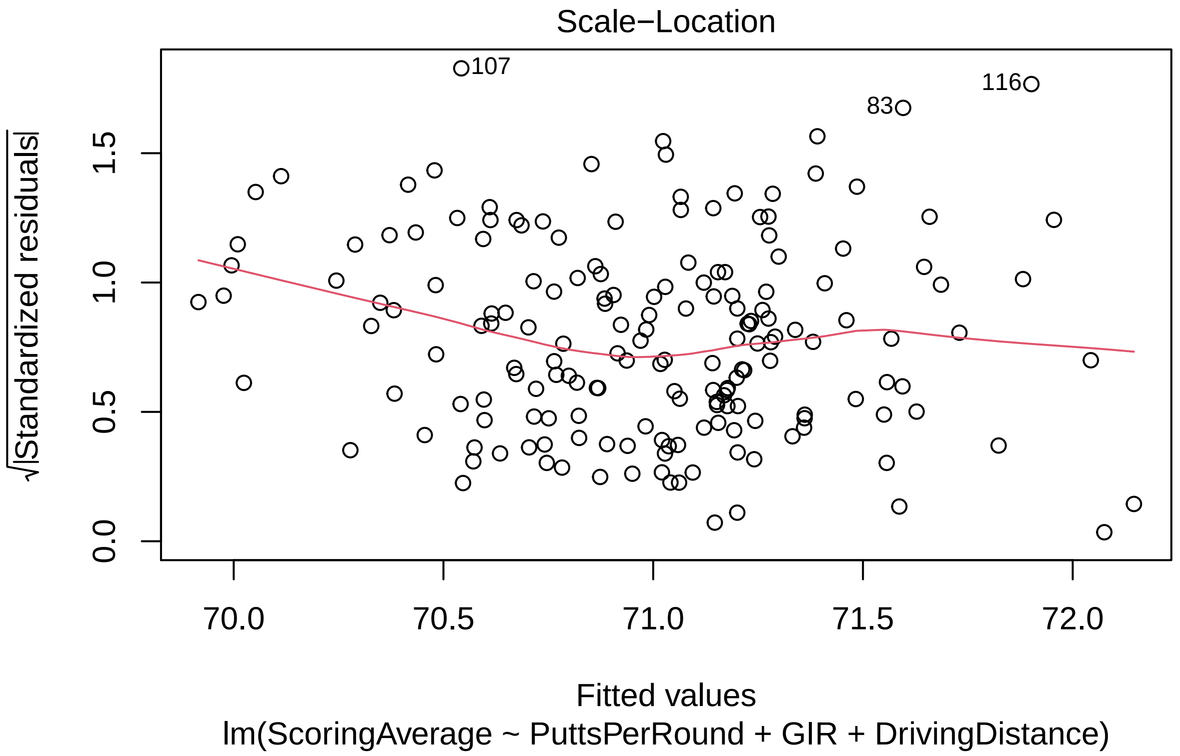
```
summary(mode19)
```

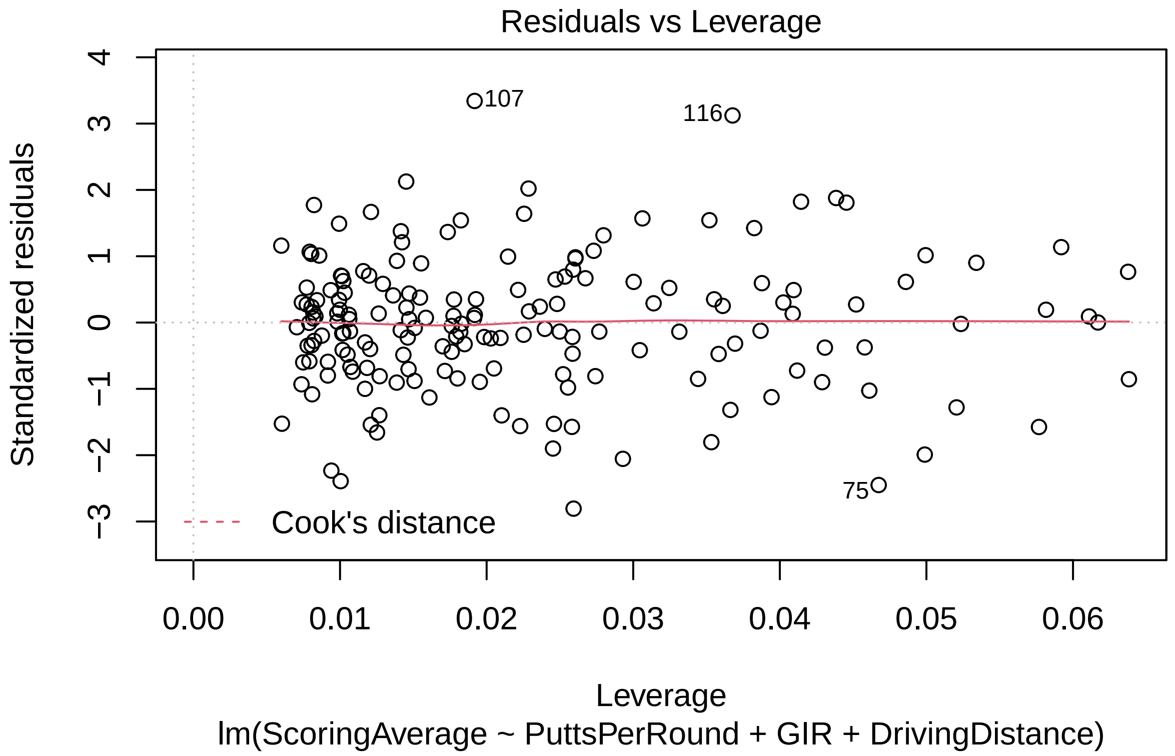
```
##
## Call:
## lm(formula = ScoringAverage ~ PuttsPerRound + GIR + DrivingDistance,
##     data = PGA2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50681 -0.32582  0.00064  0.31689  1.80044
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     59.047202   2.823292  20.914  < 2e-16 ***
## PuttsPerRound    0.820600   0.094443   8.689 2.70e-15 ***
## GIR             -0.103939   0.019302  -5.385 2.34e-07 ***
## DrivingDistance -0.013015   0.005631  -2.312    0.022 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5441 on 173 degrees of freedom
## Multiple R-squared:  0.386,  Adjusted R-squared:  0.3753
## F-statistic: 36.25 on 3 and 173 DF,  p-value: < 2.2e-16
```

```
plot(mode19)
```


Residuals vs Fitted
Residuals
2
1
0
−1
107
116
83
70.0
70.5
71.0
71.5
72.0
Fitted values
lm(ScoringAverage ~ PuttsPerRound + GIR + DrivingDistance)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Theoretical Quantiles
lm(ScoringAverage ~ PuttsPerRound + GIR + DrivingDistance)
107
116
83


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
70.0
70.5
71.0
71.5
72.0
107
83
116
Fitted values
lm(ScoringAverage ~ PuttsPerRound + GIR + DrivingDistance)



Step 4 continued

The regression model shown above, 38.6% of the variability in scoring average is explained by putts per round, GIR and driving distance. The equation the model produced in yhat = 59.05 + .821* Putts per Round + -.104*GIR + -.013*DrivingDistance

Challeneges and Issue

I faced multiple issues and challenges during this project. First, there is no attribute that has a strong correlation to Scoring Average so it was difficult to dig through the data to find a relationship. Plus, there was different response variables that I could have used like FedEx Cup Point and top ten finishes but I believe the scoring average was the best indicator on a players performance. Additionally, the data set I used had a ton of data point (over 1,000), I had to reduce the data points to produce a useful model. Reducing the data size could have be done a lot of different ways but I chose to solely look at 2019 data.

Contribution

I built out a regression model that can be used to predicted a PGA tour player average score based on average putts per round, GIR and driving distance. The data set I used was Link this process I found out that all three variables contribute to the player scoring average but to different degrees. Putts per round has the greatest effect on scoring average with GIR being second and driving distance third by a slim margin.

Topics From Class

1) R Markdown

I completed this whole project in R Markdown. I had a little trouble with the formatting of the PDF and how to display the equation in an easy to look at format. I believe I could still use some work with R Markdown formatting but I was able to complete the project from start to finish using R Markdown

2) GitHub

I had no exposure to Git before this as I had to create an account for this class/project. I was able to create a connection to my Rstudio and have the GIT document automatically update. I still have a lot to learn about the ins and outs of GIT but this was a good first step in that process.

3) Cleansing Data

I cleaned the data in Rstudio which we learned in class. I uploaded the data as in which was in a messy format. I delete the null values, split one column into two columns, reduce the data size. Cleaning the data gave me the starting point I needed to begin the rest of my analysis.

4) Graphs: histogram and scatter plot

I created a histogram to show me the normal distribution of each explanatory variable. This ensured that the data was I working with described the entire population. I also used a scatter plots that we learned in class to show the linear relationship between two variables (Explanatory to the Response)

5) Correlation and regression

From the scatter plots, I created a correlation function to give a numeric value to the relationship between each explanatory value to the response. I took this information and created an regression model based on the top correlated value to the least. Form their I found the equation to predict a scoring average based on the three explanatory variables

Conclusion

Before diving into the data my prior knowledge about the game of golf allowed me to identify three areas of the game that I believe effect the scoring out come of a player. I believed that each area of the game driving, green in regulation and putting average could explain the scoring average but I did not know to which extend. After completing the project I can come to the conclusion that the three areas of the game I looked into do effect the scoring average but not in the way I thought. I learned that putting average (40%) has the strongest correlation to scoring average with green in regulation (30%) second and driving distance (29%) last. The biggest surprise to me is when I created a regression model to predict scoring average based on all three explanatory variables the model could only explain 38% of the data points. This told me that there is more that goes into the scoring average then these three variable that I originally thought. If I were to have more time to build out this model I would take into account more explanatory variables and build my regression model from their. I simply did not have the time to do this as I was to far long into my project when I realized this and there is to many variable to take into account. This project advanced my skill in regression model building and how I walk through the process from start to finish. I was able to pull my own data, clean the data, look at the data and finally see the relationship between the variables. Overall, this project was a positive learning experience on how to work with data in the future.