Data 511, Analyzing NFL Seasons

Kristofer Hormoz

Theses are all the packages used in this project.

library(tinytex)

library(caret)

library(psych)

library(ggplot2)

library(gridExtra)

library(rattle)

library(plyr)

library(RANN)

library(rpart)

library(rpart.plot)

In this project I will be analyzing some stats from the 2020 NFL season. In doing so we will go over everything that we did in Data 511.

I will be getting my data from https://www.pro-football-reference.com/

Section 1: Data Preparation Phase

This is the 2020 nfl season data set

```
nfl_2020_season <- read.csv("nfl_2020_season.csv")
colnames(nfl_2020_season) <- nfl_2020_season[1,] # had to fix the column names
nfl_2020_season <- nfl_2020_season[-c(1),] # deleted a row

summary(nfl_2020_season[,c(1:5)]) # just looking at first 5 variables
```

```
##       Rk                  Tm                  G                   PF           
##  Length:32          Length:32          Length:32          Length:32         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##      Yds           
##  Length:32         
##  Class :character  
##  Mode  :character  
```

```
# all of the variables are char when they should be numeric, we need to fix this
```

We will now fix some of the variables. Note that we will keep "Tm" which stands for team as a char variable. Variables 16 to 28 will be removed.

These will be the variables we will be analyzing in our project.

The variables used are;

1) RK: symbolizes team rank
2) Tm: symbolizes the team
3) PF: symbolizes points fought for (i.e total points gained in a season)
4) Yds: symbolizes yards gained by passing
5) Ply: symbolizes offensive plays (pass attempts + Rush attempts + Times Sacked)
6) Y/p: symbolizes yards per play
7) TO: symbolizes Turn overs lost
8) FL: Symbolizes fumbles lost
9) 1std: Symbolizes first downs gained
10) cmp: Symbolizes completions
11) Att: symbolizes attempts made by passing
12) Yds.1: Symbolizes Total yards gained by rushing
13) TD: symbolizes touchdowns
14) Int: Symbolizes interceptions thrown
15) Yds.1: Symbolizes total yards from passing
16) Yds.2: Symbolizes total yards made by rushing
17) year: symbolizes year

```
# now we will fix up pur data set and only keep variables we are interested in

nfl_2020_season <- nfl_2020_season[,c(1:21)]

nfl_2020_season$year <- 2020 # adding year 2020
nfl_2020_season$G <- NULL
str(nfl_2020_season) # str gives the structure of each variable.
```

```
## 'data.frame':    32 obs. of  21 variables:
##  $ Rk    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Tm    : chr  "Green Bay Packers" "Buffalo Bills" "Tampa Bay Buccaneers" "Tennessee Titans" ...
##  $ PF    : num  509 501 492 491 482 473 468 459 451 434 ...
##  $ Yds   : num  6224 6343 6145 6343 6023 ...
##  $ Ply   : num  990 1034 1017 1031 1045 ...
##  $ Y/P   : num  6.3 6.1 6 6.2 5.8 6.3 5.9 5.8 5.9 5.9 ...
##  $ TO    : num  11 22 17 12 17 16 18 18 15 26 ...
##  $ FL    : num  6 11 5 5 9 9 7 5 4 16 ...
##  $ 1stD  : num  358 397 364 381 367 397 327 356 364 359 ...
##  $ Cmp   : num  372 410 410 316 370 420 257 388 371 369 ...
##  $ Att   : num  526 596 626 485 522 630 406 563 552 551 ...
##  $ Yds.1 : num  4106 4620 4626 3653 3758 ...
##  $ TD    : num  48 40 42 33 28 40 27 40 24 28 ...
##  $ Int   : num  5 11 12 7 8 7 11 13 11 10 ...
##  $ NY/A  : num  7.5 7.4 7.1 7.2 6.8 7.4 6.3 6.5 7.1 7.3 ...
##  $ 1stD.1: num  216 240 238 203 199 255 142 216 201 206 ...
##  $ Att.1 : num  443 411 369 521 494 403 555 411 459 457 ...
##  $ Yds.2 : num  2118 1723 1519 2690 2265 ...
##  $ TD.1  : num  16 16 16 26 30 13 24 15 20 20 ...
##  $ Y/A   : num  4.8 4.2 4.1 5.2 4.6 4.5 5.5 4.8 4.3 4.2 ...
##  $ year  : num  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
```

The next lines of code, we are downloading more data and preparing it. Seasons 2016 to 2019 will be added.

```
nfl_2019_season <- read.csv("nfl_2019_season.csv")

colnames(nfl_2019_season) <- nfl_2019_season[1,] # had to fix the column names
nfl_2019_season <- nfl_2019_season[-c(1),] # deleted a row

nfl_2018_season <- read.csv("nfl_2018_season.csv")

colnames(nfl_2018_season) <- nfl_2018_season[1,] # had to fix the column names
nfl_2018_season <- nfl_2018_season[-c(1),] # deleted a row

nfl_2017_season <- read.csv("nfl_2017_season.csv")

colnames(nfl_2017_season) <- nfl_2017_season[1,] # had to fix the column names
nfl_2017_season <- nfl_2017_season[-c(1),] # deleted a row

nfl_2016_season <- read.csv("nfl_2016_season.csv")

colnames(nfl_2016_season) <- nfl_2016_season[1,] # had to fix the column names
nfl_2016_season <- nfl_2016_season[-c(1),] # deleted a row =
```

Combining the Data into One Data Frame

Here we are merging the data sets by the rows into one data frame. This will Give a new data frame "nfl" with 160 observations with 15 variables.

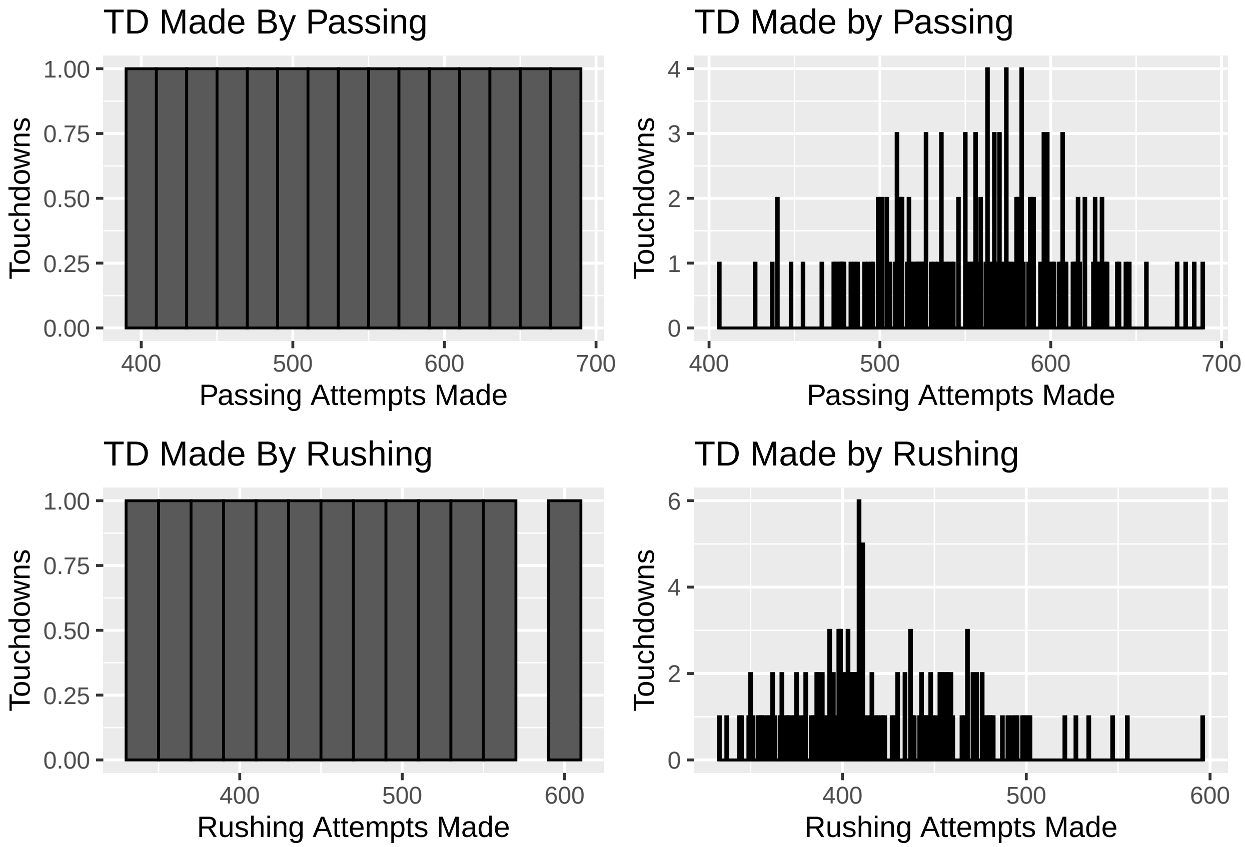
```
nfl <- rbind(nfl_2020_season[,c(21,1:20)],nfl_2019_season[,c(21,1:20)],
             nfl_2018_season[,c(21,1:20)],nfl_2017_season[,c(21,1:20)],
             nfl_2016_season[,c(21,1:20)])

nfl$pass_Att <- nfl$Att
nfl$rush_Att <- nfl$Att.1
nfl$pass_TD <- nfl$TD
nfl$rush_TD <- nfl$TD.1

nfl$Att <- NULL
nfl$Att.1 <- NULL
nfl$TD <- NULL
nfl$TD.1 <- NULL
```

Exploratory Data Analysis Phase

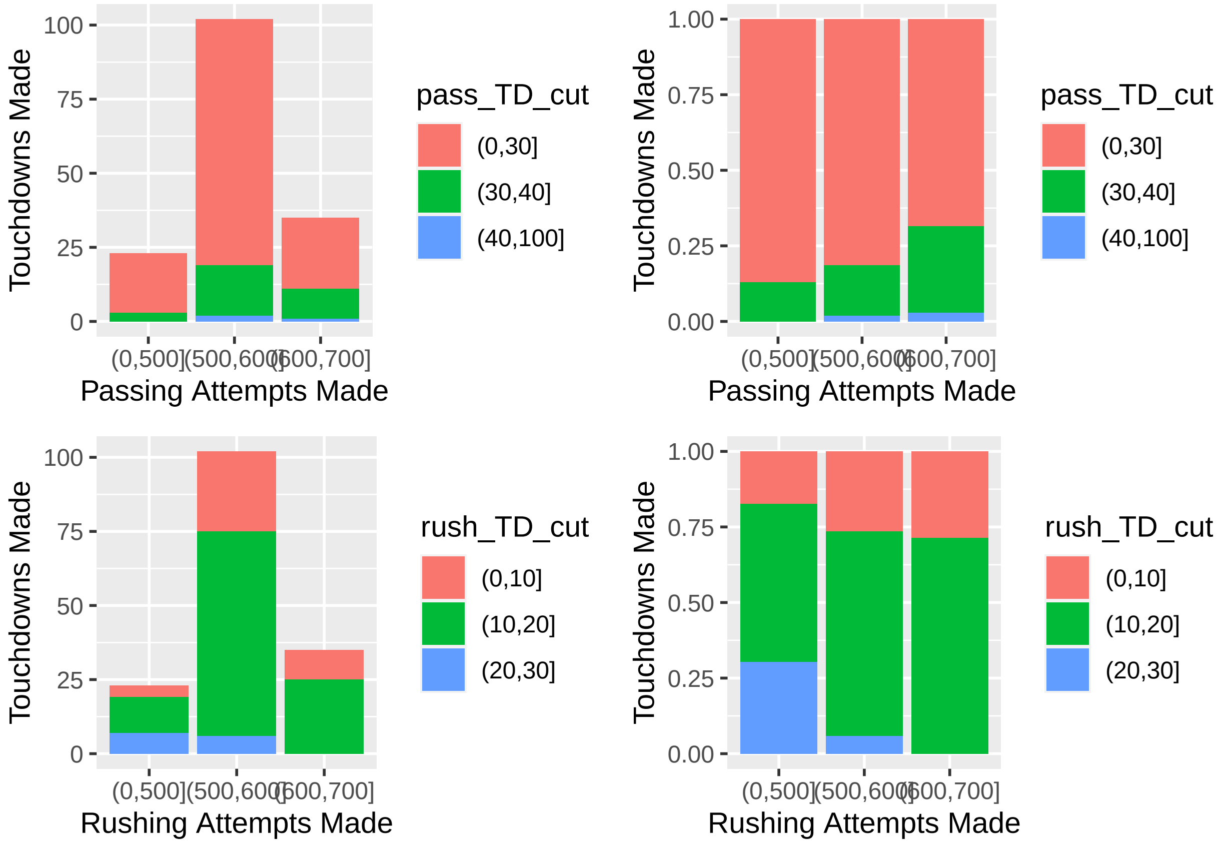
step (2) using decision trees to find optimal bins

```r
# Normalized histogram w churn overlay.

grid.arrange(
ggplot(nfl, aes(pass_Att)) +
  geom_histogram(aes(fill = pass_TD),
                 color = "black", binwidth= 20, position = "fill") +
  xlab("Passing Attempts Made") +
  ylab("Touchdowns") +
  ggtitle("TD Made By Passing"),



# Non-normalized histogram w overlay.
ggplot(nfl, aes(pass_Att)) +
  geom_histogram(aes(fill = pass_TD),
                 color = "black", binwidth = 1, position = "stack") +
  xlab("Passing Attempts Made") +
  ylab("Touchdowns") +
  ggtitle("TD Made by Passing"),


ggplot(nfl, aes(rush_Att)) +
  geom_histogram(aes(fill = rush_TD),
                 color = "black", binwidth= 20, position = "fill") +
  xlab("Rushing Attempts Made") +
  ylab("Touchdowns") +
  ggtitle("TD Made By Rushing"),



# Non-normalized histogram w overlay.
ggplot(nfl, aes(rush_Att)) +
  geom_histogram(aes(fill = rush_TD),
                 color = "black", binwidth = 1, position = "stack") +
  xlab("Rushing Attempts Made") +
  ylab("Touchdowns") +
  ggtitle("TD Made by Rushing")

)
```

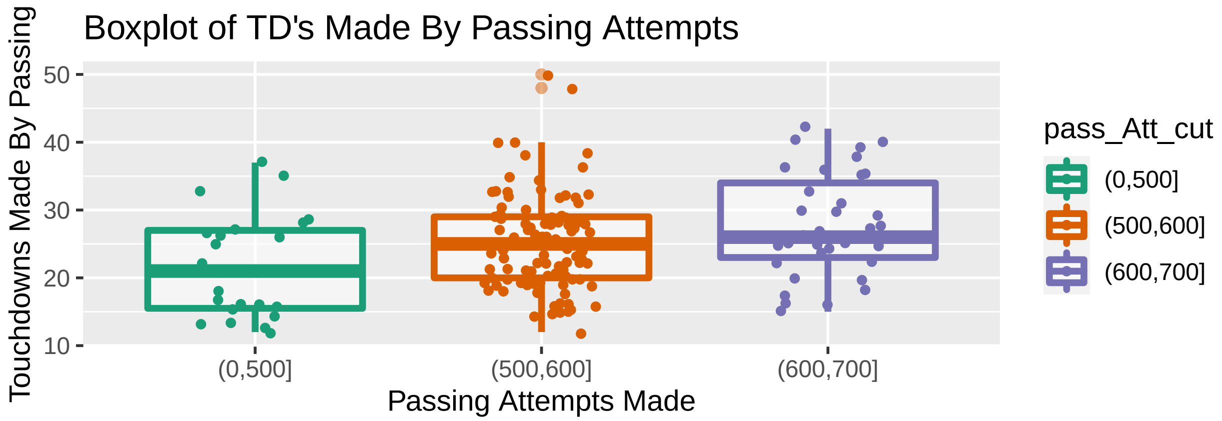
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

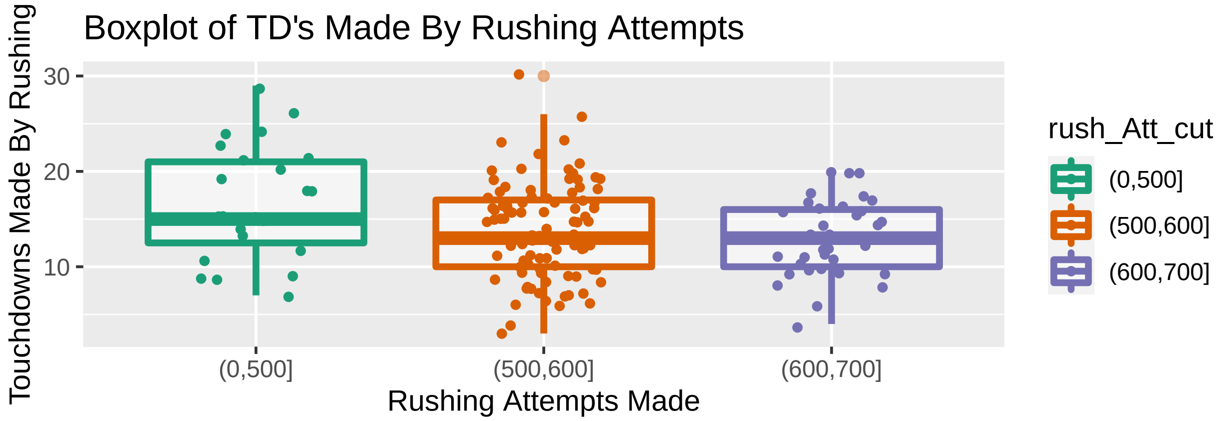


```
nfl$pass_TD_cut <- cut(nfl$pass_TD, breaks = c(0,30,40,100))
nfl$rush_TD_cut <- cut(nfl$rush_TD, breaks = c(0,10,20,30))

nfl$pass_Att_cut <- cut(nfl$pass_Att, breaks = c(0,500,600,700))
nfl$rush_Att_cut <- cut(nfl$rush_Att, breaks = c(0,400,475,600))

nfl$pass_Att_cut <- as.factor(nfl$pass_Att_cut)
nfl$rush_Att_cut <- as.factor(nfl$pass_Att_cut)
nfl$pass_TD_cut <- as.factor(nfl$pass_TD_cut)
nfl$rush_TD_cut <- as.factor(nfl$rush_TD_cut)

grid.arrange(
# non normalized
ggplot(nfl, aes(pass_Att_cut))+
  geom_bar(aes(fill = pass_TD_cut),position = "stack")+
  xlab("Passing Attempts Made")+
  ylab("Touchdowns Made"),

#normalized
ggplot(nfl, aes(pass_Att_cut))+
  geom_bar(aes(fill = pass_TD_cut) ,position = "fill")+
  xlab("Passing Attempts Made")+
  ylab("Touchdowns Made"),

ggplot(nfl, aes(rush_Att_cut))+
  geom_bar(aes(fill = rush_TD_cut),position = "stack")+
  xlab("Rushing Attempts Made")+
  ylab("Touchdowns Made"),

#normalized
ggplot(nfl, aes(rush_Att_cut))+
  geom_bar(aes(fill = rush_TD_cut) ,position = "fill")+
  xlab("Rushing Attempts Made")+
  ylab("Touchdowns Made")

)
```



```
grid.arrange(

  ggplot(data = nfl,
       aes(x = pass_Att_cut, y = pass_TD, color = pass_Att_cut)) +
  geom_boxplot(size = 1.2, alpha = .5) +
  xlab("Passing Attempts Made") +
  ylab("Touchdowns Made By Passing") +
  labs(title = "Boxplot of TD's Made By Passing Attempts") +
  scale_color_brewer(palette="Dark2") +
  geom_jitter(shape=16, position=position_jitter(0.2)),


ggplot(data = nfl,
       aes(x = rush_Att_cut, y = rush_TD, color = rush_Att_cut)) +
  geom_boxplot(size = 1.2, alpha = .5) +
  xlab("Rushing Attempts Made") +
  ylab("Touchdowns Made By Rushing") +
  labs(title = "Boxplot of TD's Made By Rushing Attempts") +
  scale_color_brewer(palette="Dark2") +
  geom_jitter(shape=16, position=position_jitter(0.2))

)
```





```
# then do a contingency table

t1 <- table(nfl$pass_TD_cut, nfl$pass_Att_cut)
round(prop.table(t1,2)*100,2)
```

```
##
##            (0,500] (500,600] (600,700]
##   (0,30]     86.96     81.37     68.57
##   (30,40]    13.04     16.67     28.57
##   (40,100]    0.00      1.96      2.86
```

```
t2 <- table(nfl$rush_TD_cut, nfl$rush_Att_cut)
round(prop.table(t2,2)*100,2)
```

```
##
```

```
##           (0,500] (500,600] (600,700]
##   (0,10]    17.39     26.47     28.57
##   (10,20]   52.17     67.65     71.43
##   (20,30]   30.43      5.88      0.00
```