Curating the Dataset

December 24, 2017

1 Direction 1

1.1 A) Curate NFL Data

The purpose of this notebook is to curate a data set that determines the outcome of a game and relevant statistics that I believe will predict the outcome of the game. There are 1791 unique games from 2009 to 2015 that are within the main data set. The following variables were created: * Average Yard Gained by Passing * Average Yard Gained by Rushing * Total Yards Allowed by Defense Passing * TotalYards Allowed by Defense Rushing * Turnovers by Offense * Turnovers caused by Defense * Scores by Offense * Scores by Denfense

A function was created to determine these statistics for each of the games. Though there are many other statistics that could have been created, this is a first attempt to understand the dynamics of a good team.

In [25]:
```
## Load Appropriate Packages and Data
library(dplyr)
load("nfldata.rda")
```

In [2]:
```
nfl_data <- tbl_df(pbp_data)
```

```
##-----------------------
# Purpose: To determine the outcome of each game based on
# the away and home teams performance
# Input: Data from 1 game
# Output: Average Pass and Rush Yards, Total Defense Pass
# and Rush Yards allowed, Turnovers by offense
# Score, and Outcome
##-----------------------
outcome_of_game <- function(tmp_data){
    # Determine Teams
    Teams = tmp_data %>% select(HomeTeam,AwayTeam)
    HomeTEAM = Teams[1,1]
    AwayTEAM = Teams[1,2]

    ## HomeTeam Offense and AwayTeam Defense
```

```r
# Determine average rush yards by home team and Total
# rush yards allowed by awawy team
Rush_Yards_HT <- tmp_data %>%
    filter(posteam == as.character(HomeTEAM) &
           PlayType == "Run") %>%
    summarise(avgRushYard_HT = mean(Yards.Gained),
              totalRushYards_AT = sum(Yards.Gained))
# Determine average pass yards by home team and Total
# pass yards allowed by awawy team
Pass_Yards_HT <- tmp_data %>%
    filter(posteam == as.character(HomeTEAM) &
           PlayType == "Pass") %>%
    summarise(avgPassYard_HT = mean(Yards.Gained),
              totalPassYards_AT = sum(Yards.Gained))


## AwayTeam Offense and HomeTeam Defense
# Determine average rush yards by home team and
# Total rush yards allowed by awawy team
Rush_Yards_AT <- tmp_data %>%
    filter(posteam == as.character(AwayTEAM) &
           PlayType == "Run") %>%
    summarise(avgRushYard_AT = mean(Yards.Gained),
              totalRushYards_HT = sum(Yards.Gained))
# Determine average pass yards by home team and
# Total pass yards allowed by awawy team
Pass_Yards_AT <- tmp_data %>%
    filter(posteam == as.character(AwayTEAM) &
           PlayType == "Pass") %>%
    summarise(avgPassYard_AT = mean(Yards.Gained),
              totalPassYards_HT = sum(Yards.Gained))


## HomeTeam Turnovers by offense and AwayTeam by Defense
## Note 1: I am not concern which team recovers
## Note 2: I only want to see what happens
## during non special teams plays
TurnOvers_HT <- tmp_data %>%
    filter(posteam == as.character(HomeTEAM) &
       (InterceptionThrown == 1| Fumble == 1) &
       (PlayType == "Run" | PlayType == "Pass") ) %>%
    summarise(TurnOver_HT = n())

## HomeTeam Turnovers by offense and AwayTeam by Defense
## Note 1: I am not concern which team recovers
## Note 2: I only want to see what happens
## during non special teams plays
TurnOvers_AT <- tmp_data %>%
```

```
    filter(posteam == as.character(AwayTEAM) &
       (InterceptionThrown == 1| Fumble == 1) &
       (PlayType == "Run" | PlayType == "Pass") ) %>%
    summarise(TurnOver_AT = n())


## Game Conclusion
last_row <- tmp_data[nrow(tmp_data),]


## Determine score of the home and away teams
if(last_row$posteam == as.character(HomeTEAM)){
    HomeTeam_Score <- as.numeric(last_row$PosTeamScore)
    AwayTeam_Score <- as.numeric(last_row$DefTeamScore)
} else {
    HomeTeam_Score <- as.numeric(last_row$DefTeamScore)
    AwayTeam_Score <- as.numeric(last_row$PosTeamScore)
}

## Determine outcome of the home and away teams
if( is.na(HomeTeam_Score) == T | is.na(AwayTeam_Score) == T){
    HomeTeam_Outcome <- NA
    AwayTeam_Outcome <- NA
} else if(HomeTeam_Score > AwayTeam_Score) {
    HomeTeam_Outcome <- 1
    AwayTeam_Outcome <- 0
} else if(HomeTeam_Score < AwayTeam_Score) {
    HomeTeam_Outcome <- 0
    AwayTeam_Outcome <- 1
} else{
    HomeTeam_Outcome <- NA
    AwayTeam_Outcome <- NA
}


## Put information together by home and away stats
row1 <- cbind(HomeTEAM,Pass_Yards_HT[1],Rush_Yards_HT[1],
      Pass_Yards_AT[2],Rush_Yards_AT[2],
      TurnOvers_HT,TurnOvers_AT,
      HomeTeam_Score,HomeTeam_Outcome,"H")
row2 <- cbind(AwayTEAM,Pass_Yards_AT[1],Rush_Yards_AT[1],
      Pass_Yards_HT[2],Rush_Yards_HT[2],
      TurnOvers_AT,TurnOvers_HT,
      AwayTeam_Score,AwayTeam_Outcome,"A")

colnames(row1) <- NULL
colnames(row2) <- NULL
return(list(home = row1,away = row2,
```

```
                    pos_score = last_row$PosTeamScore,
                    def_score = last_row$DefTeamScore ))
        }
```

In [ ]:
```
        ## Determine all game ids
        GID <- nfl_data %>% select(GameID) %>% unique()
        homeTeamData <- vector()
        awayTeamData <- vector()

        ## Go through each game and calculate the statistics
        ## and combine the stats to a data.frame
        for(i in 1:nrow(GID)){
            tmp1 <- nfl_data %>% filter(GameID == as.character(GID[i,1]) )
            tmp2 <- tmp1 %>% outcome_of_game()
            homeTeamData <- cbind(homeTeamData,t(as.matrix(tmp2$home)))
            awayTeamData <- cbind(awayTeamData,t(as.matrix(tmp2$away)))
            ##cat("Game ID:", as.character(GID[i,1]),"  ",i,"\n")
            ##cat("Pos Score",tmp2$pos_score, " Def Score",tmp2$def_score, "\n")
        }
```

In [24]:
```
        ## Combine
        TeamData <- rbind(t(homeTeamData),t(awayTeamData))
        ## Create Column Names
        colnames(TeamData) <- c("Team","avgPassYards","avgRushYards",
                                "defTotalPassYards","defTotalRushYards",
                                "offTurnOvers","defTurnOvers",
                                "Score","Outcome","HomeORAway")
        ## Makes Data into data frame
        TeamData <- tbl_df(data.frame(TeamData))

        write.csv(TeamData,"nfl_direction1.csv",row.names = F)
```

A side note, through vigorous searching, it was found that if a team went to overtime, the conclusion of the game was marked as a tie. This implies these games will be ignored in the analysis.