Parameters associated with Injuries in the NFL

Brody Johnson

2024-06-21

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

# 1. Summary/Abstract

The goal of this project will be to find variables related to injuries observed in the NFL. That is, find the common play types, game situation (time remaining, down, yards to go, etc.), and injured player metrics (position, height, body weight, etc.) that has a correlation to causing the resulting injury.

# 2. Introduction

## 2.1 General Background Information

For years, the NFL has been making changes in the game in an attempt to prevent player injuries. These changes range from rule modifications to investment in advancements to the players pads and helmets. Hopefully, the findings from this project can help inform players, coaches, and who knows, the NFL about the metrics for a higher risk in injury.

## 2.2 Description of data and data source

Three data sets have been identified for this project. The primary data set (“Injury Player Data” obtained from a Github repository contains 1,586 observations) consists of every play ran in the NFL containing an injury during the 2019-2020 season. Some of the things contained in this data include the teams playing, the week in the season, play description, injury area, and injured player metrics. Two other data sets were obtained from Kaggle. The first is a “cumulative play data” containing information about every play ran in the 2019-2020 NFL season (42,186 observations). Information such as the teams playing, play description, and time in the game when it was ran. The last data set is the “player demographic data” containing information about all NFL players in the 2016-2019 seasons (11,145 observations). This data set has player metrics such as height, weight, date of birth, etc.

## 2.3 Questions/Hypotheses to be addressed

My hope is to answer the following question: What metrics, if any, are available that can help predict whether an injury may occur? In turn, this has the potential to answer more questions. For example: 1. For coaches, are there specific play types that lead to more injuries? 2. For players, is there a goal weight for certain positions that can reduce the likelihood of injury?

\*\*\*\*\*To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020).\*\*\*\*\*\*

# 3. Methods

The primary idea behind the analysis for this project is to look at various parameters when an injury occurred and compare them to the rest of the NFL. If the parameters are related to the play itself, the paramter in the injruy data set will be compared to cumulative play data set. Likewise, if the paramter is related to the injured player themselves, this paramter will be copmared to the cumulative player data set. If there is a statistical difference between paramters, we can likely draw some conclusion about a parameter causing injuries.

With this being said, it is imperative to ensure parameters from both data sets are in the same format in order to compare them. Other data cleaning items will include vetting the data sets for unexpected/missing values and converting variables to another variable that easier to use. For example, the play descriptions in the injury data set and cumulative play data set are very extensive in what happened during the play. Therefore, this will be boiled down to be either a run play, short pass, etc.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.3 Data import and cleaning

In both the “Injury Player Data” set and “cumulative play data” set, the play descriptions will need to be boiled down to simplified play types (pass, run, fumble, etc.). Next for the “Injury Player Data” set, the “injury\_area” parameter will be summarized as upper and lower body injuries. This is due to the overwhelming number of injuries being classified as one or the other already so this will classify the remaining injuries as such. For the injury data set and the “player demographic data”, player ages will need to be calculated from their birth-dates.

As for the “cumulative player data” set, empty columns will be removed and 0’s in the down column will be converted to “NA” values which indicate kickoffs & PATs. Lastly, a lot of unnecessary plays are included in the “cumulative play data” set such as end of quarter/games, 2-minute warnings, and timeouts so these observations can be filtered out.

In the “player demographic data” set, observations for players height were recorded in both inches and feet-inches so this was cleaned to only contain data in inches. Lastly, the players position was converted to a factor for analysis.

#Loading the processed Data Sets  
#Raw data files were cleaned in the processingcode.R (../../code/processing-code/) then saved as .rds file in the processed-data folder  
  
# Injury Player Data  
injuries=readRDS("../../data/processed-data/injuriesprocesseddata.rds")   
  
# Cumulative Play data  
pbp=readRDS("../../data/processed-data/pbpprocesseddata.rds")   
  
#Player Demographic data  
players=readRDS("../../data/processed-data/playersprocesseddata.rds")

## 3.4 Statistical analysis

First, I will select variables that show a potential difference between the injury data set and cumulative data sets through exploratory analysis. From these variables, I will run inferential tests for the numeric variables and a Chi-square test for the categorical variables.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Structures of each data set

#Structure of Injury Data Set  
glimpse(injuries)

Rows: 1,586  
Columns: 51  
$ player\_id <chr> "2019\_ARI\_1\_Murray", "2019\_ARI\_13\_Kirk", "20…  
$ game\_id <chr> "2019\_16\_ARI\_SEA", "2019\_04\_SEA\_ARI", "2019\_…  
$ home\_team <chr> "SEA", "ARI", "TB", "TB", "NO", "ARI", "SEA"…  
$ away\_team <chr> "ARI", "SEA", "ARI", "ARI", "ARI", "LAR", "A…  
$ season\_type <chr> "REG", "REG", "REG", "REG", "REG", "REG", "R…  
$ week.x <int> 16, 4, 10, 10, 8, 13, 16, 16, 7, 17, 8, 9, 1…  
$ posteam <chr> "ARI", "ARI", "TB", "TB", "ARI", "LA", "ARI"…  
$ posteam\_type <chr> "away", "home", "home", "home", "away", "awa…  
$ defteam <chr> "SEA", "SEA", "ARI", "ARI", "NO", "ARI", "SE…  
$ side\_of\_field <chr> "SEA", "SEA", "TB", "TB", "ARI", "LA", "ARI"…  
$ yardline\_100 <int> 31, 15, 72, 84, 67, 91, 81, 10, 67, 38, 35, …  
$ game\_date <date> 2019-12-22, 2019-09-29, 2019-11-10, 2019-11…  
$ quarter\_seconds\_remaining <int> 735, 26, 453, 437, 269, 653, 890, 40, 155, 1…  
$ half\_seconds\_remaining <int> 1635, 26, 1353, 437, 1169, 1553, 890, 40, 15…  
$ game\_seconds\_remaining <int> 1635, 26, 1353, 437, 1169, 1553, 890, 1840, …  
$ game\_half <chr> "Half2", "Half2", "Half2", "Half2", "Half2",…  
$ qtr <int> 3, 4, 3, 4, 3, 3, 4, 2, 4, 3, 1, 1, 1, 2, 3,…  
$ down <int> 3, 2, 2, 1, 3, 3, 1, 1, 4, 3, 1, 1, 3, 1, 2,…  
$ time <dbl> 0.510416667, 0.018055556, 0.314583333, 0.303…  
$ yrdln <chr> "SEA 31", "SEA 15", "TB 28", "TB 16", "ARI 3…  
$ ydstogo <int> 12, 3, 7, 10, 2, 9, 10, 10, 15, 4, 10, 10, 2…  
$ desc <chr> "(12:15) (Shotgun) 1-K.Murray scrambles left…  
$ injured\_first\_name <chr> "K", "C", "P", "P", "C", "J", "K", "K", "H",…  
$ injured\_last\_name <chr> "Murray", "Kirk", "Peterson", "Peterson", "E…  
$ injured\_team <chr> "ARI", "ARI", "ARI", "ARI", "ARI", "ARI", "A…  
$ injured\_player\_num <int> 1, 13, 21, 21, 29, 34, 41, 41, 43, 54, 54, 7…  
$ Contact..non.contact <chr> "non contact", "contact", "non contact", "no…  
$ injury.area <chr> "lower body", "knee", "lower body", "lower b…  
$ player.role <chr> "ball carrier", "pass catcher", "AFP", "pas …  
$ season.x <int> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 20…  
$ season.y <int> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 20…  
$ team <chr> "ARI", "ARI", "ARI", "ARI", "ARI", "ARI", "A…  
$ position <fct> QB, WR, DB, DB, RB, DB, RB, RB, LB, LB, LB, …  
$ depth\_chart\_position <chr> "QB", "WR", "CB", "CB", "RB", "DB", "RB", "R…  
$ jersey\_number <int> 1, 13, 21, 21, 29, 34, 41, 41, 43, 54, 54, 7…  
$ status <chr> "Active", "Active", "Active", "Active", "Act…  
$ full\_name <chr> "Kyler Murray", "Christian Kirk", "Patrick P…  
$ first\_name <chr> "Kyler", "Christian", "Patrick", "Patrick", …  
$ last\_name <chr> "Murray", "Kirk", "Peterson", "Peterson", "E…  
$ birth\_date <date> 1997-08-07, 1996-11-18, 1990-07-11, 1990-07…  
$ height <int> 70, 71, 73, 73, 69, 71, 73, 73, 73, 76, 76, …  
$ weight <int> 207, 200, 203, 203, 210, 190, 211, 211, 235,…  
$ college <chr> "Oklahoma", "Texas A&amp;M", "Louisiana Stat…  
$ years\_exp <int> 0, 1, 8, 8, 1, 0, 3, 3, 2, 5, 5, 4, 8, 5, 10…  
$ game\_type <chr> "REG", "REG", "REG", "REG", "REG", "REG", "R…  
$ football\_name <chr> "Kyler", "Christian", "Patrick", "Patrick", …  
$ smart\_id <chr> "32004d55-5267-0413-8d36-a5c3fd781aa0", "320…  
$ entry\_year <int> 2019, 2018, 2011, 2011, 2018, 2019, 2016, 20…  
$ play\_type <chr> "Scramble", "Short Pass", "Short Pass", "Dee…  
$ injury.area.new <chr> "Lower Body", "Lower Body", "Lower Body", "L…  
$ age <dbl> 22.5, 22.9, 29.4, 29.4, 23.6, 21.4, 26.0, 26…

#Structure of Play-by-Play Data  
glimpse(pbp)

Rows: 38,966  
Columns: 42  
$ GameId <int> 2019122201, 2019122201, 2019122201, 201…  
$ GameDate <date> 2019-12-22, 2019-12-22, 2019-12-22, 20…  
$ Quarter <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 2, …  
$ Minute <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 1, 8, …  
$ Second <int> 40, 47, 51, 57, 3, 15, 20, 25, 43, 4, 8…  
$ OffenseTeam <chr> "LAC", "LAC", "LAC", "LAC", "LAC", "LAC…  
$ DefenseTeam <chr> "LV", "LV", "LV", "LV", "LV", "LV", "LV…  
$ Down <int> 3, 2, 1, 2, 1, 3, 2, 1, 2, NA, 1, 1, 1,…  
$ ToGo <int> 9, 9, 9, 2, 10, 6, 6, 10, 4, 0, 1, 10, …  
$ YardLine <int> 91, 91, 91, 86, 78, 68, 68, 64, 55, 85,…  
$ SeriesFirstDown <int> 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, …  
$ NextScore <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ Description <chr> "(1:40) (SHOTGUN) 17-P.RIVERS PASS INCO…  
$ TeamWin <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ SeasonYear <int> 2019, 2019, 2019, 2019, 2019, 2019, 201…  
$ Yards <int> 0, 0, 0, 0, 8, 10, 0, 4, 9, 0, 1, 18, 0…  
$ Formation <chr> "SHOTGUN", "SHOTGUN", "SHOTGUN", "SHOTG…  
$ PlayType <chr> "PASS", "PASS", "PASS", "PASS", "PASS",…  
$ IsRush <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, …  
$ IsPass <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, …  
$ IsIncomplete <int> 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, …  
$ IsTouchdown <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, …  
$ PassType <chr> "SHORT LEFT", "SHORT LEFT", "SHORT RIGH…  
$ IsSack <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsChallenge <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsChallengeReversed <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ Challenger <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ IsMeasurement <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsInterception <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsFumble <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsPenalty <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsTwoPointConversion <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ IsTwoPointConversionSuccessful <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ RushDirection <chr> "", "", "", "", "", "", "", "", "", "",…  
$ YardLineFixed <int> 9, 9, 9, 14, 22, 32, 32, 36, 45, 15, 1,…  
$ YardLineDirection <chr> "OPP", "OPP", "OPP", "OPP", "OPP", "OPP…  
$ IsPenaltyAccepted <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ PenaltyTeam <chr> "", "", "", "OAK", "", "", "", "", "", …  
$ IsNoPlay <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ PenaltyType <chr> "", "", "", "DEFENSIVE OFFSIDE", "", ""…  
$ PenaltyYards <int> 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ play\_type <chr> "Short Pass", "Short Pass", "Short Pass…

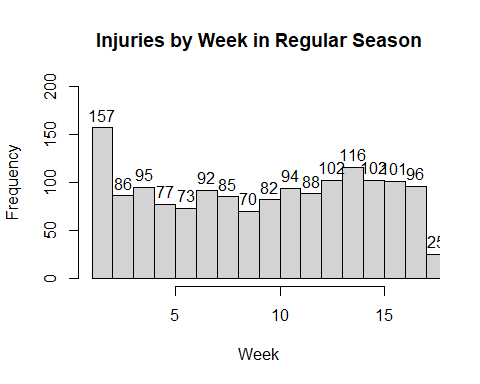
#Structure of All Players Data Set  
glimpse(players)

Rows: 1,303  
Columns: 8  
$ nflId <int> 2539334, 2539653, 2543850, 2555162, 2555255, 2555543, 2556…  
$ height <dbl> 72, 70, 69, 73, 75, 73, 70, 72, 78, 77, 72, 72, 68, 75, 66…  
$ weight <int> 190, 186, 186, 227, 232, 216, 211, 200, 243, 250, 223, 198…  
$ birthDate <date> 1990-09-10, 1988-11-01, 1991-12-18, 1994-11-04, 1993-07-0…  
$ collegeName <chr> "Washington", "Southeastern Louisiana", "Purdue", "Louisia…  
$ position <fct> DB, DB, DB, LB, LB, DB, DB, WR, QB, TE, RB, WR, WR, DL, RB…  
$ displayName <chr> "Desmond Trufant", "Robert Alford", "Ricardo Allen", "Deio…  
$ age <dbl> 29.1, 30.9, 27.8, 24.9, 26.3, 24.2, 27.0, NA, 30.7, 28.9, …

### 4.1.2 Exploratory Analysis of parameters only in the *Injury* Data Set

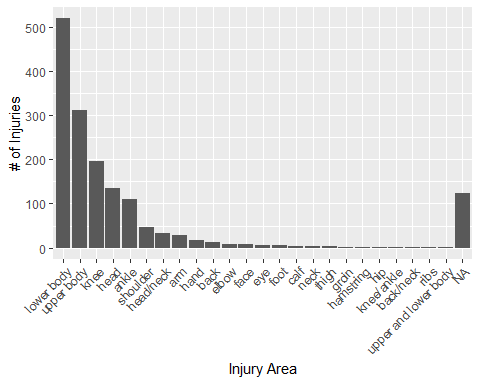
Looking at the number of injuries each week in the regular season, there’s a sharp drop after the first week, steadily climbs nearing the end of the season, and another drop at the end of the season. Intuitively, this makes sense in that players may not be physically readily for the intensity of a regular-season game despite training camps and preseason games. After that, the slight rise in the second half of the season could be due to exhaustion through a long season and/or teams fighting more for playoff spots. Lastly, the steep drop in the last week is likely due to teams knowing whether or not they have made playoffs and therefore, do not want to incur an injury to go into either the post-season or off-season.

hist(injuries$week.x, breaks=21,main="Injuries by Week in Regular Season", xlim=c(1,17),ylim=c(0,200),xlab="Week",labels =T)



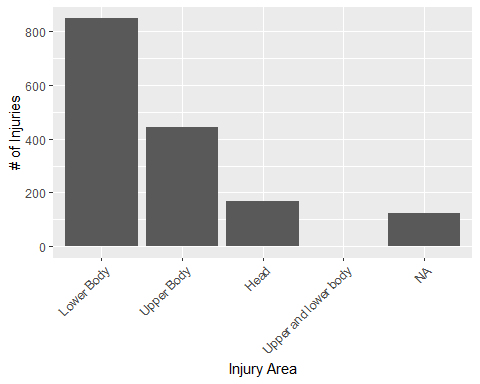
Below depicts the number of injuries for each reported injury area. The top results were simply reported as upper or lower body.

injarea\_freq=injuries %>% group\_by(injury.area) %>% summarize(count=n())   
ggplot(data = injarea\_freq, aes(x = reorder(injury.area, -count), y = count))+  
 geom\_col() +  
 labs(x = "Injury Area", y="# of Injuries") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



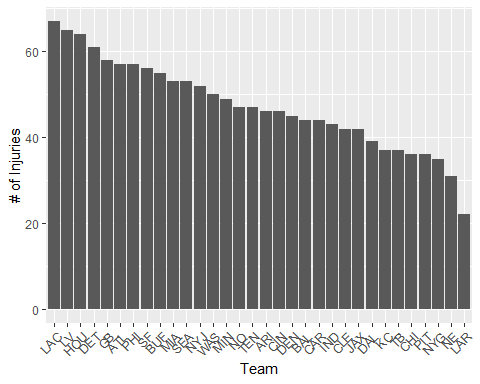
Below is a plot of the number of injuries by injury area but the injury areas were summarized to either be upper/lower body or head injuries. I kept head injuries separate due to the interest around concussions in football and see how they stack up against all other injuries.

injareanew\_freq=injuries %>% group\_by(injury.area.new) %>% summarize(count=n())   
ggplot(data = injareanew\_freq, aes(x = reorder(injury.area.new, -count), y = count))+  
 geom\_col() +  
 labs(x = "Injury Area", y="# of Injuries") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



Out of curiosity, I plotted below the frequency of injuries for each team during the regular season. I was not expecting much but it is interesting to see about a triple in number of injuries from the lowest team (LA Rams) and the highest (LA Chargers).

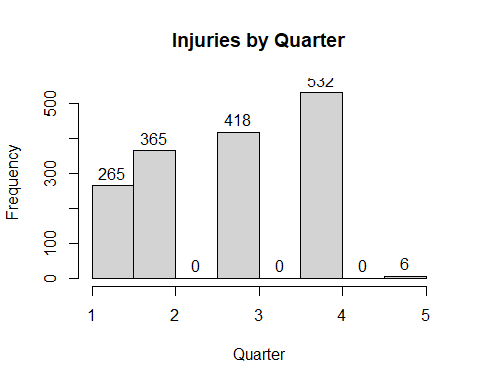
team\_freq=injuries %>% filter(week.x<18) %>% group\_by(team) %>% summarize(count=n()) #filtered to only contain regular season games (weeks 1-17)  
ggplot(data = team\_freq, aes(x = reorder(team, -count), y = count))+  
 geom\_col() +  
 labs(x = "Team", y="# of Injuries") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



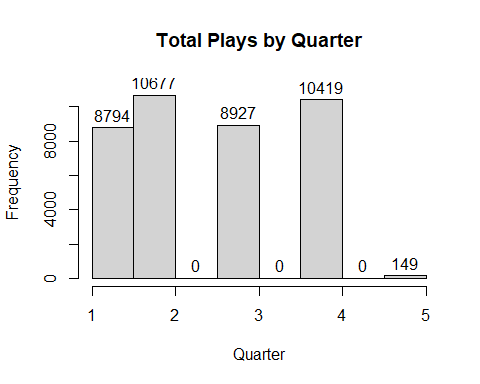
### 4.1.3 Exploratory Analysis of parameters in the *Injury* Data Set and *Play-by-play* Data Set

The number of plays seem evenly distributed among the halves (slight increase in 2nd and 4th quarters likely due to running last minute plays). However, the number of injuries consistently increase throughout the game which also makes sense as players fatigue.

hist(injuries$qtr,main="Injuries by Quarter", ylim=c(0,550), xlim=c(1,5),xlab="Quarter",labels =T)

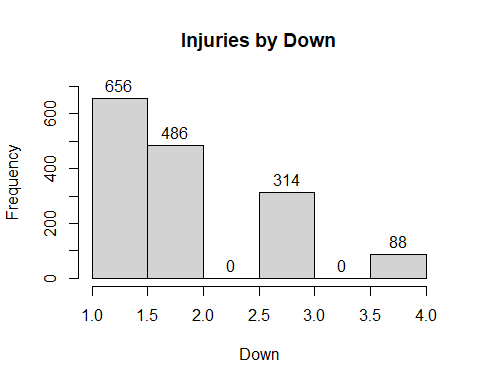


hist(pbp$Quarter, breaks=6,main="Total Plays by Quarter", ylim=c(0,11200), xlab="Quarter",labels =T)

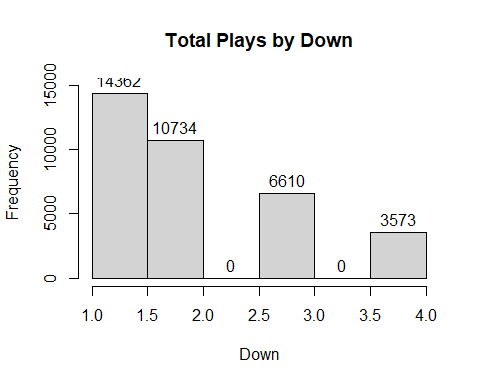


There appears to be a very similar trend among the number of plays per down and the number of injuries ocurring during each down. That is, they both decrease at similar rates.

hist(injuries$down, breaks=5,main="Injuries by Down", ylim=c(0,700),xlim=c(1,4),xlab="Down",labels =T)

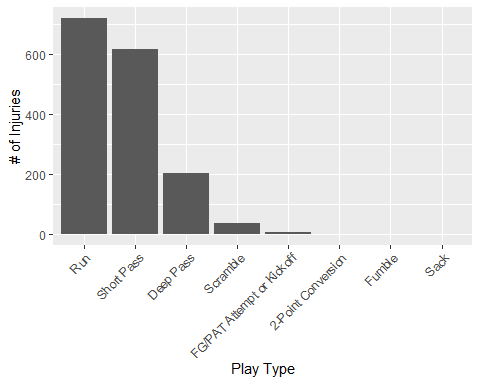


hist(pbp$Down, breaks=6,main="Total Plays by Down", ylim=c(0,14900),xlim=c(1,4), xlab="Down",labels =T)

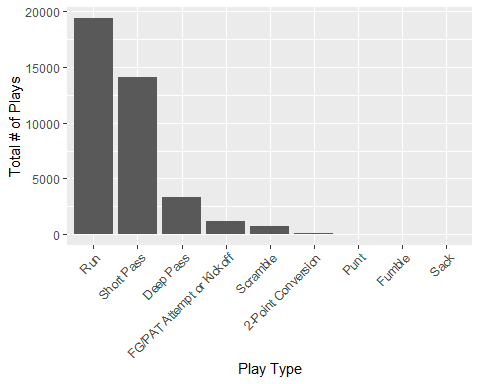


Below are two histograms depicting first the number of injuries and then second the total # of plays by different play types. After a glance, there does not appear to be a significant difference in number of injuries and total # of plays ran by play type.

play\_type\_freq=injuries %>% group\_by(play\_type) %>% summarize(count=n())   
ggplot(data = play\_type\_freq, aes(x = reorder(play\_type, -count), y = count))+  
 geom\_col() +  
 labs(x = "Play Type", y="# of Injuries") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



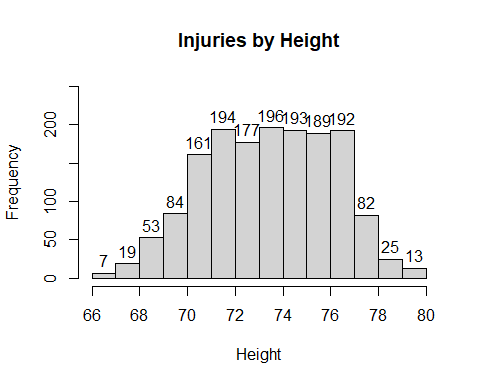
play\_type\_freq2=pbp %>% group\_by(play\_type) %>% summarize(count=n())   
ggplot(data = play\_type\_freq2, aes(x = reorder(play\_type, -count), y = count))+  
 geom\_col() +  
 labs(x = "Play Type", y="Total # of Plays") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



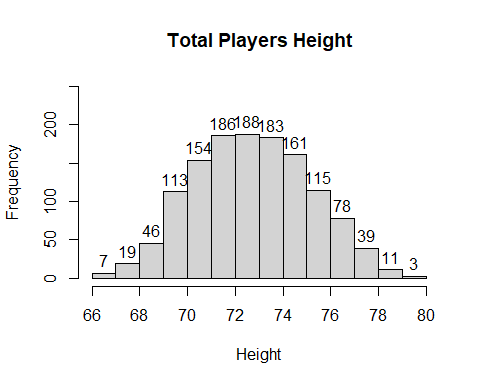
### 4.1.4 Exploratory Analysis of parameters in the *Injury* Data Set and *Cumulative Player* Data Set

Below are the distributions of players height first of the injured players and then from the cumulative player data set. Both histograms appear to be normally distributed. One interesting difference is the injured players appear to have more players just below and above the center of the histogram. That is, it would almost suggest players specifically under and above the mean height have a higher inclination for injury.

hist(injuries$height, breaks=10,main="Injuries by Height", ylim=c(0,250), xlab="Height",labels =T)

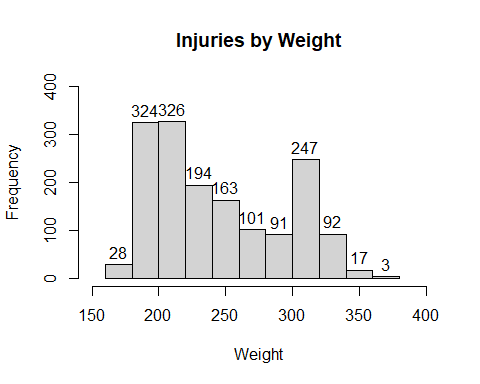


hist(players$height, breaks=10,main="Total Players Height", ylim=c(0,250), xlab="Height",labels =T)

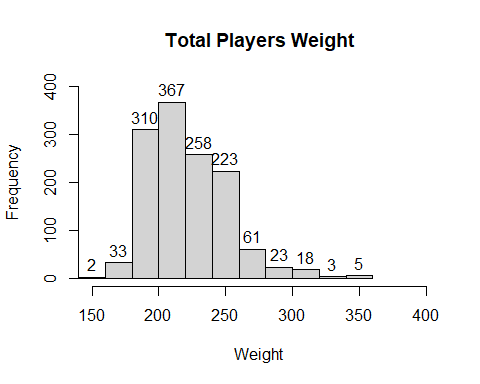


Below are the histograms depicting the players height. They are both skewed to the right with one interesting finding of a peak of injured players in the ~300lb range.

hist(injuries$weight, breaks=10,main="Injuries by Weight", xlim=c(150,400),ylim=c(0,400), xlab="Weight",labels =T)

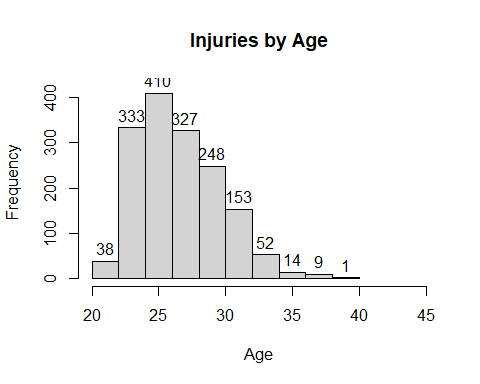


hist(players$weight, breaks=10,main="Total Players Weight", xlim=c(150,400),ylim=c(0,400), xlab="Weight",labels =T)

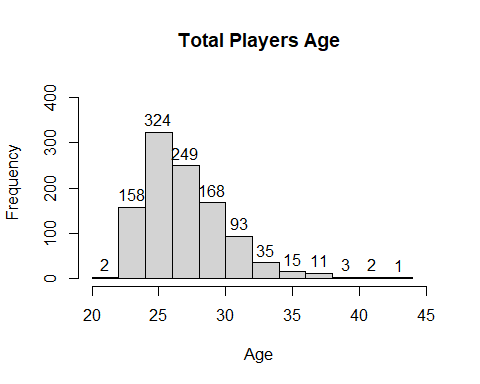


Both distributions of players age are below and have a similar skewness to the right.

hist(injuries$age, breaks=10,main="Injuries by Age", xlim=c(20,45),ylim=c(0,425), xlab="Age",labels =T)

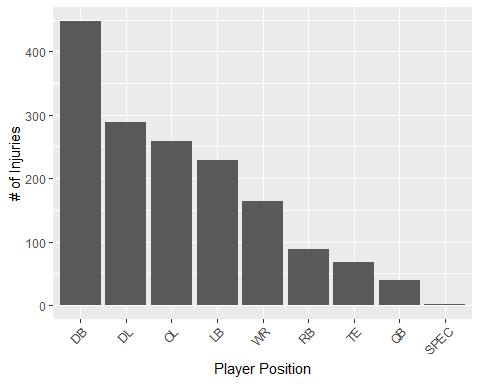


hist(players$age, breaks=10,main="Total Players Age", xlim=c(20,45),ylim=c(0,425), xlab="Age",labels =T)

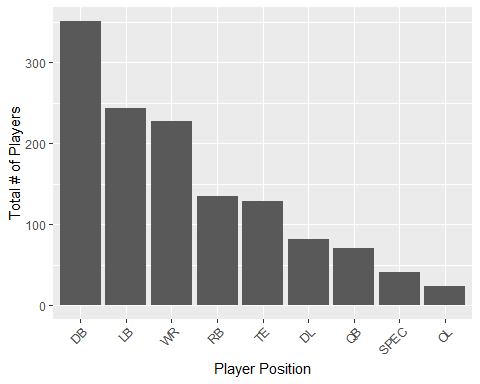


The primary finding from the player positions is that it appears the linemen (defensive and ofensive) are injured more often relative to how many there are in the league.

position\_freq=injuries %>% group\_by(position) %>% summarize(count=n())   
ggplot(data = position\_freq, aes(x = reorder(position, -count), y = count))+  
 geom\_col() +  
 labs(x = "Player Position", y="# of Injuries") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



position\_freq2=players %>% group\_by(position) %>% summarize(count=n())   
ggplot(data = position\_freq2, aes(x = reorder(position, -count), y = count))+  
 geom\_col() +  
 labs(x = "Player Position", y="Total # of Players") +   
 theme(axis.text.x = element\_text(angle=45, hjust = 1))



## 4.2 Basic statistical analysis

From the exploratory analysis, it appears there three numeric variables that may show a difference and one categorical. The three numeric variables are quarter, height, and weight while the single categorical variable is player position.

#Inferential test looking at Quarter  
shapiro.test(injuries$qtr) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: injuries$qtr  
W = 0.8555, p-value < 2.2e-16

shapiro.test(sample(pbp$Quarter, size=5000, replace=F)) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: sample(pbp$Quarter, size = 5000, replace = F)  
W = 0.86467, p-value < 2.2e-16

#Will do Wilcoxon rank-sum test b/c of non-normal data   
  
wilcox.test(injuries$qtr, pbp$Quarter, exact=F) #p-value<0.05. Reject Ho and therefore, one group tends to have larger values than the other

Wilcoxon rank sum test with continuity correction  
  
data: injuries$qtr and pbp$Quarter  
W = 34440728, p-value = 1.194e-15  
alternative hypothesis: true location shift is not equal to 0

#Inferential test looking at Height  
shapiro.test(injuries$height) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: injuries$height  
W = 0.97684, p-value = 2.618e-15

shapiro.test(players$height) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: players$height  
W = 0.98243, p-value = 1.75e-11

#Will do Wilcoxon rank-sum test b/c of non-normal data   
  
wilcox.test(injuries$height, players$height, exact=F) #p-value<0.05. Reject Ho and therefore, one group tends to have larger values than the other

Wilcoxon rank sum test with continuity correction  
  
data: injuries$height and players$height  
W = 1191966, p-value = 6.506e-13  
alternative hypothesis: true location shift is not equal to 0

#Inferential test looking at Weight  
shapiro.test(injuries$weight) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: injuries$weight  
W = 0.91678, p-value < 2.2e-16

shapiro.test(players$weight) #p-value<0.05. Reject Ho & don't assume normal.

Shapiro-Wilk normality test  
  
data: players$weight  
W = 0.94389, p-value < 2.2e-16

#Will do Wilcoxon rank-sum test b/c of non-normal data   
  
wilcox.test(injuries$weight, players$weight, exact=F) #p-value<0.05. Reject Ho and therefore, one group tends to have larger values than the other

Wilcoxon rank sum test with continuity correction  
  
data: injuries$weight and players$weight  
W = 1271297, p-value < 2.2e-16  
alternative hypothesis: true location shift is not equal to 0

#Chi-square test looking at player position  
factor\_levels = levels(injuries$position)  
chisq.test(sample(injuries$position, size=1303, replace=F), players$position) #p-value>0.05. Accept Ho and therefore, the two data sets have equal proportions among the player positions

Warning in chisq.test(sample(injuries$position, size = 1303, replace = F), :  
Chi-squared approximation may be incorrect

Pearson's Chi-squared test  
  
data: sample(injuries$position, size = 1303, replace = F) and players$position  
X-squared = 74.871, df = 64, p-value = 0.1662

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 1](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>