Parameters associated with Injuries in the NFL

Brody Johnson

2024-07-28

# 1. Summary/Abstract

The goal of this project will be to conduct a descriptive analysis on 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

Over the past three decades, the NFL has come under increased scrutiny over the dangers of American football (PBS Frontline (n.d.)). What initially began as concerns over the persistent brutality central to the sport, shifted into the more insidious issue of concussions and their role in the post-mortem diagnosis of Chronic Traumatic Encephalitic (CTE) (Mayo Clinic (n.d.)). The NFL has been making changes in the game in an attempt to prevent player injuries 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” 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. The next 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?

# 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 parameter in the “injury player data” set will be compared to “cumulative play data” set. Likewise, if the parameter is related to the injured player themselves, this parameter will be compared to the “cumulative player data” set.

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 is easier to use. For example, the play descriptions in the “injury player 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 Data aquisition

The “Injury Player Data” set was found from a Github repository. The repository belongs to a group of students who wanted to do a similar analysis looking at NFL injuries and various factors that could affect them.

https://github.com/sammieerne/NFL-Injury-Analysis/tree/main/Data

The “cumulative play data” set comes from a website called NFLsavant.com. This website is dedicated to providing NFL statistics to the public.

https://nflsavant.com/about.php

Lastly, the “player demographic data” set was also taken from a Github repository. Similar to the other repository, this one belongs to a different group who wanted to analyze injuries in the NFL.

https://github.com/ericcrouse/NFL-Injury-Exploration/tree/main/data/raw

## 3.2 Data import and cleaning

In both the “Injury Player Data” set and “cumulative play data” set, the play descriptions will need to be simplified to 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 player data” set and the “player demographic data”, player ages will need to be calculated from their birth-dates. Lastly, a new variable was created to tell whether the injured player was on the home or away team.

As for the “cumulative play 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")

## 3.3 Statistical analysis

I will depict specific variables that can potentially show a relationship to number of injuries. As applicable, the same variables will be depicted from the supporting data sets for comparison.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Structures of each data set

Below are glimpses into each of the three data sets showing the variables, variable types, and the first several observations as an exmaple of what the data looks like.

#Structure of Injury Data Set  
glimpse(injuries)

Rows: 1,586  
Columns: 52  
$ 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…  
$ InjPlayerHomeOrAway <chr> "Away", "Home", "Away", "Away", "Away", "Hom…

#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…

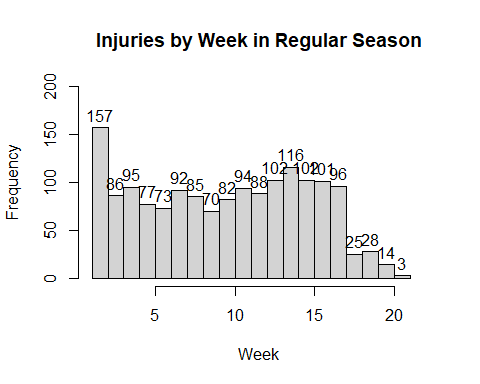
#Structure of All Players Data Set  
glimpse(players)

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 Parameters only in the *Injury* Data Set

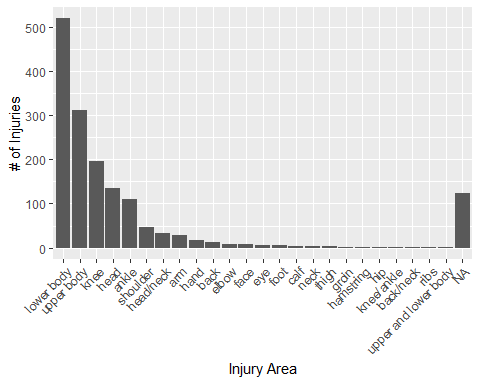
Looking at the number of injuries each week in the regular season (17 weeks), there’s a sharp drop after the first week, steadily climbs nearing the end of the season, and another drop at the end of the season. Intuitively, this makes sense in that players may not be physically ready for the intensity of the first 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=25,main="Injuries by Week in Regular Season", xlim=c(1,22),ylim=c(0,200),xlab="Week",labels =T)



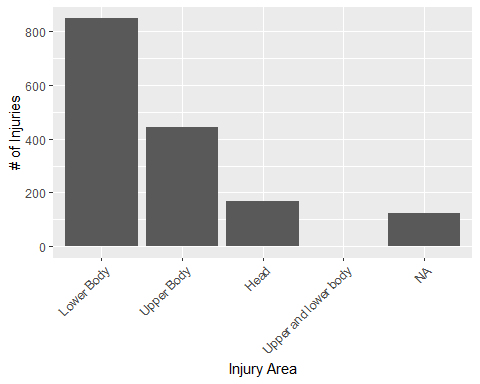
Below depicts the number of injuries for each reported injury area. The top results were simply reported as upper or lower body.

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))



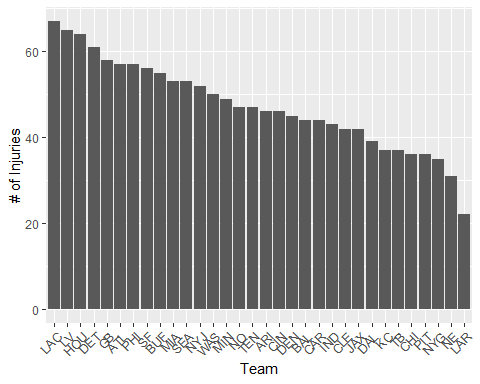
Below is a plot of the number of injuries by injury area but the injury areas were summarized to either be upper/lower body or head injuries. I kept head injuries separate due to the interest around concussions in football and see how they stack up against all other injuries.

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))



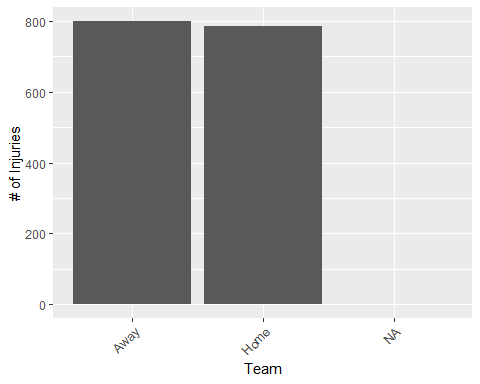
Out of curiosity, I plotted below the frequency of injuries for each team during the regular season. I was not expecting much but it is interesting to see about a triple in number of injuries from the lowest team (LA Rams) and the highest (LA Chargers). Although this graph filters out the post-season games, the four teams that made division championship games are relatively spread out across this graph and consist of the Chiefs (KC), Titans (TEN), 49ers (SF), and Packers (GB). Point being the most competitive teams are relatively unaffected by number of injuries.

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))



Below depicts the number of injuries for the players that belonged to the home or away team.

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



At initial glance, the bar plot above makes it appear the number of injuries are roughly equal between the home and away teams. Below, I created a table to see the number of observations along with their respective proportions. Sure enough, there does not appear to be a large difference in the number of injuries on a team whether they are playing at home or away.

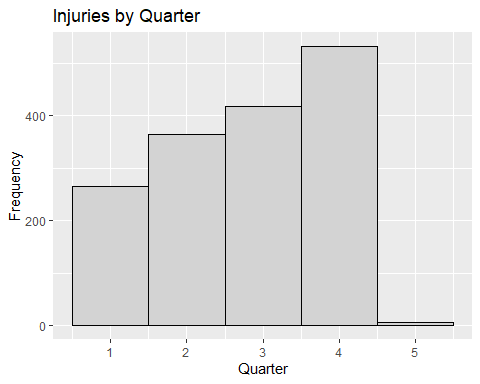
counts = table(injuries$InjPlayerHomeOrAway) #Counts of each factor   
proportions = prop.table(counts) #Proportions of each factor  
result = cbind(Counts = counts, Percentage = round(proportions\*100,1)) #Combine results  
print(result)

Counts Percentage  
Away 800 50.5  
Home 785 49.5

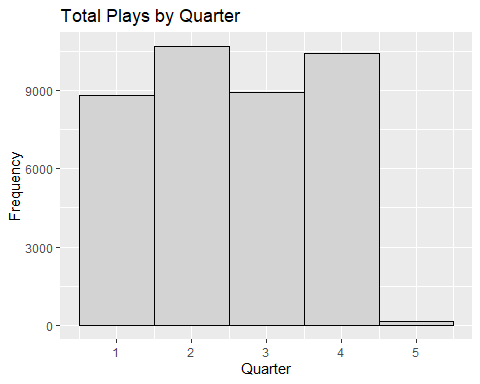
### 4.1.3 Parameters in the *Injury* Data Set and *Play-by-play* Data Set

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.

ggplot(injuries, aes(x = qtr)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Injuries by Quarter", x = "Quarter", y = "Frequency")



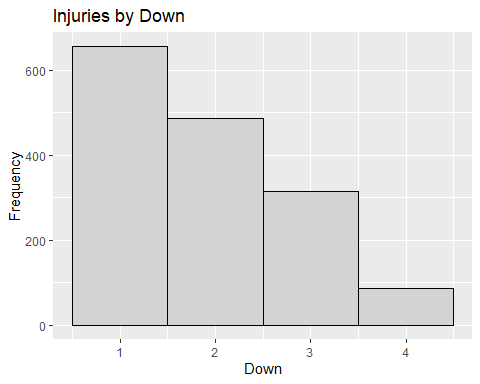
ggplot(pbp, aes(x = Quarter)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Total Plays by Quarter", x = "Quarter", y = "Frequency")



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

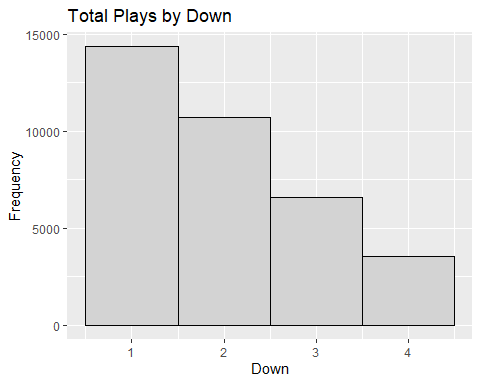
ggplot(injuries, aes(x = down)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Injuries by Down", x = "Down", y = "Frequency")

Warning: Removed 42 rows containing non-finite outside the scale range  
(`stat\_bin()`).



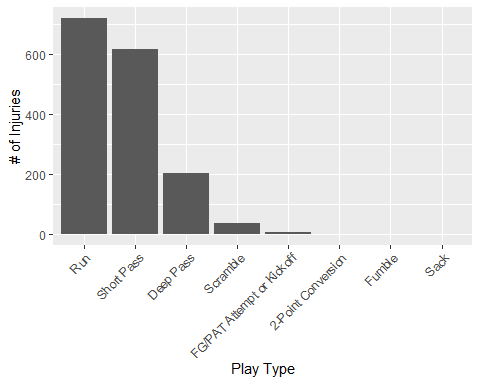
ggplot(pbp, aes(x = Down)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Total Plays by Down", x = "Down", y = "Frequency")

Warning: Removed 3687 rows containing non-finite outside the scale range  
(`stat\_bin()`).

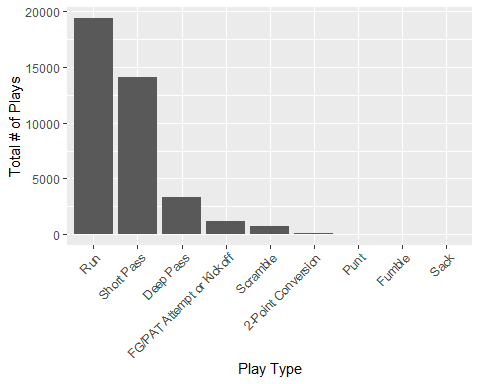


Below are two histograms depicting first the number of injuries and then second the total number of plays by different play types. After a glance, there does not appear to be a significant difference in number of injuries and total number of plays ran by play type.

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))



Above, we can tell the order of play types are the same between injuries and total number of plays (e.g. there are more run plays causing injuries and in total relative to deep passes). Next, I wanted to compare the proportions of these play types. Below is a table of percentages of injuries and total plays by play type. From this, we can tell the percentages are relatively close to one another. It appears run plays have injuries about 4.5% less often compared to how often they’re ran. Similarly, pass plays have injuries occur 2.5-4% more often compared to how often they are ran.

#Injury Data set  
inj\_counts = table(injuries$play\_type) #Counts of each play type  
inj\_proportions = prop.table(inj\_counts) #Proportions of each play type   
  
#PBP data set  
total\_play\_counts = table(pbp$play\_type) #Counts of each play type  
total\_play\_counts = total\_play\_counts[-5]  
total\_play\_proportions = prop.table(total\_play\_counts) #Proportions of each play type   
  
#Combining results  
result = data.frame(round(inj\_proportions\*100,1), round(total\_play\_proportions\*100,1))  
result = result[,-3]   
colnames(result) = c("Play Type", "Injuries Percentages", "Total Plays Percentages") #Renaming columns  
result=result[order(result$`Injuries Percentages`, decreasing = TRUE),]  
print(result)

Play Type Injuries Percentages Total Plays Percentages  
5 Run 45.5 49.9  
8 Short Pass 38.8 36.2  
2 Deep Pass 12.8 8.5  
7 Scramble 2.4 1.9  
3 FG/PAT Attempt or Kickoff 0.4 3.0  
1 2-Point Conversion 0.1 0.3  
4 Fumble 0.1 0.1  
6 Sack 0.1 0.0

Now that the play types are known, I am interested if there is a difference in plays ran between the teams with the lowest and highest number of injuries (Rams and Chargers, respectively). For one team having triple the number of injuries, the percentage of play types are closer than I would have expected. The biggest difference is between the number of deep passes. According to the table above, deep passes are one of the more dangerous play types relative to how often they are ran. Still with only a ~3% difference in deep pass plays between the two teams, there must be more factors contributing to the Chargers having so many more injuries.

#Rams   
ram\_df = pbp[pbp$OffenseTeam == 'LA',] #Filter data set  
rams\_inj\_counts = table(ram\_df$play\_type) #Counts of each play type   
rams\_inj\_proportions = prop.table(rams\_inj\_counts) #Proportions of each play type   
rams\_inj\_proportions = rams\_inj\_proportions[-5]  
  
#Chargers  
cha\_df = pbp[pbp$OffenseTeam == 'LAC',] #Filter data set  
cha\_inj\_counts = table(cha\_df$play\_type) #Counts of each play type   
cha\_inj\_proportions = prop.table(cha\_inj\_counts) #Proportions of each play type   
  
#Combining results  
result = data.frame(round(rams\_inj\_proportions\*100,1), round(cha\_inj\_proportions\*100,1))  
print(result)

Var1 Freq Var1.1 Freq.1  
1 2-Point Conversion 0.4 2-Point Conversion 0.1  
2 Deep Pass 8.0 Deep Pass 10.9  
3 FG/PAT Attempt or Kickoff 3.0 FG/PAT Attempt or Kickoff 2.9  
4 Fumble 0.2 Fumble 0.1  
5 Run 47.1 Run 46.9  
6 Scramble 0.4 Scramble 0.4  
7 Short Pass 40.8 Short Pass 38.7

result = result[,-3]   
colnames(result) = c("Play Type", "Rams Percentages", "Chargers Percentages") #Renaming columns  
result=result[order(result$`Rams Percentages`, decreasing = TRUE),]  
print(result)

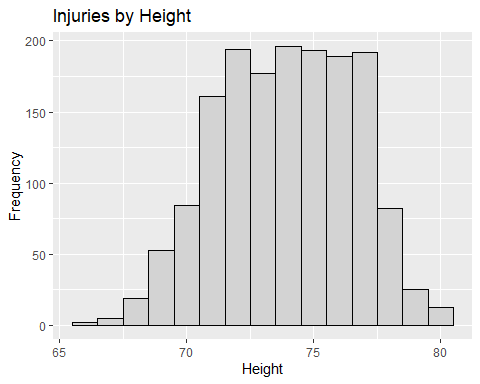
Play Type Rams Percentages Chargers Percentages  
5 Run 47.1 46.9  
7 Short Pass 40.8 38.7  
2 Deep Pass 8.0 10.9  
3 FG/PAT Attempt or Kickoff 3.0 2.9  
1 2-Point Conversion 0.4 0.1  
6 Scramble 0.4 0.4  
4 Fumble 0.2 0.1

### 4.1.4 Parameters in the *Injury* Data Set and *Cumulative Player* Data Set

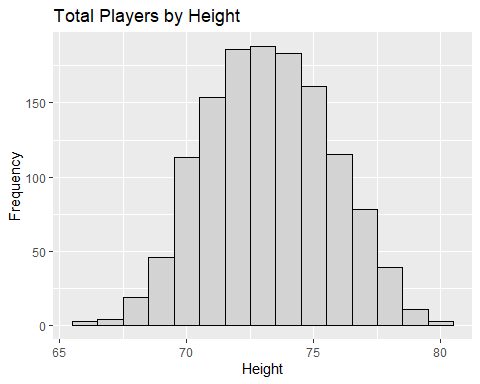
Below are the distributions of players height first of the injured players and then from the cumulative player data set. Both histograms appear to be normally distributed. One interesting difference is the injured players appear to have more players just above the center of the histogram. That is, it would suggest players just above the mean height (specifically with a height of 76-77 inches) have a higher inclination for injury.

ggplot(injuries, aes(x = height)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Injuries by Height", x = "Height", y = "Frequency")

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_bin()`).

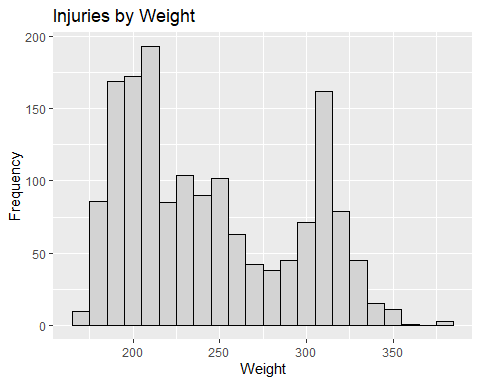


ggplot(players, aes(x = height)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Total Players by Height", x = "Height", y = "Frequency")

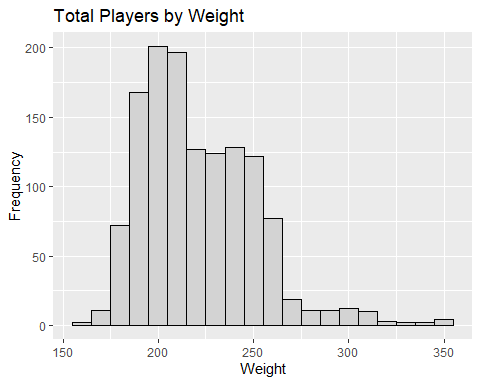


Below are the histograms depicting the players height. They are both skewed to the right with one interesting finding of a peak of injured players in the ~310lb range.

ggplot(injuries, aes(x = weight)) +  
 geom\_histogram(binwidth = 10, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Injuries by Weight", x = "Weight", y = "Frequency")



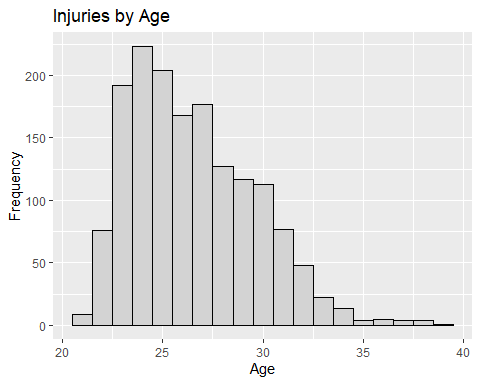
ggplot(players, aes(x = weight)) +  
 geom\_histogram(binwidth = 10, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Total Players by Weight", x = "Weight", y = "Frequency")



Both distributions of players age are below and have a similar distribution.

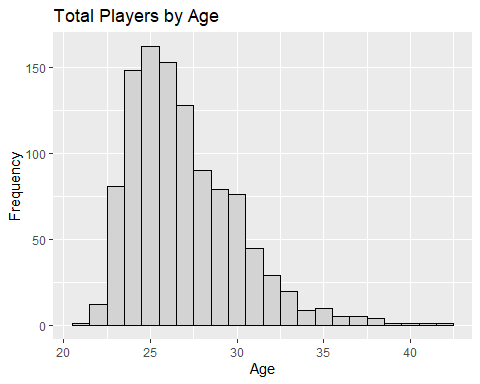
ggplot(injuries, aes(x = age)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Injuries by Age", x = "Age", y = "Frequency")

Warning: Removed 1 row containing non-finite outside the scale range  
(`stat\_bin()`).



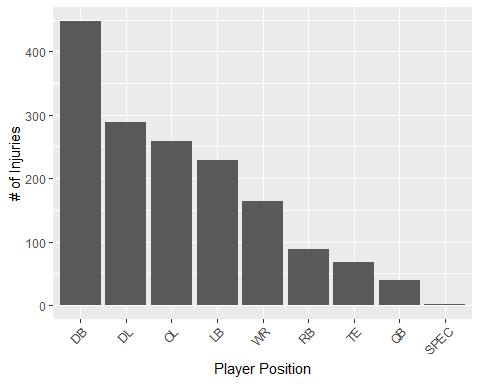
ggplot(players, aes(x = age)) +  
 geom\_histogram(binwidth = 1, fill = "lightgray", color = "black", alpha = 1) +  
 labs(title = "Total Players by Age", x = "Age", y = "Frequency")

Warning: Removed 242 rows containing non-finite outside the scale range  
(`stat\_bin()`).

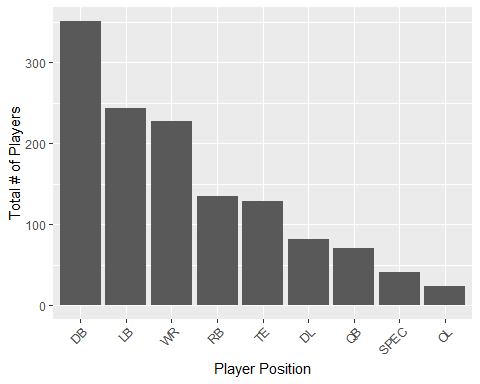


The primary finding from the player positions is that it appears the linemen (defensive and offensive) are injured more often relative to how many there are in the league.

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))



To make findings about the positions more clear, a table of percentages for each position can be found below for injured players and all of players. From the graphs, we could easily tell the defensive linemen and offensive linemen have higher percentages of being injured relative to their proportions in the league. The table below suggests linebackers, wide receivers, and running backs are injured 4-7% less relative to their respective proportions in the league.

#Injury Data set  
inj\_counts = table(injuries$position) #Counts of each position  
inj\_proportions = prop.table(inj\_counts) #Proportions of each position  
  
#Players Data set  
total\_play\_counts = table(players$position) #Counts of each position  
total\_play\_proportions = prop.table(total\_play\_counts) #Proportions of each position  
  
#Combine results  
result = data.frame(round(inj\_proportions\*100,1), round(total\_play\_proportions\*100,1))  
result = result[,-3]   
colnames(result) = c("Player Position", "Injuries Percentages", "Total Percentages") #Renaming columns  
result=result[order(result$`Injuries Percentages`, decreasing = TRUE),]  
print(result)

Player Position Injuries Percentages Total Percentages  
1 DB 28.2 26.9  
2 DL 18.2 6.3  
4 OL 16.3 1.8  
3 LB 14.4 18.6  
9 WR 10.3 17.5  
6 RB 5.6 10.4  
8 TE 4.3 9.8  
5 QB 2.5 5.4  
7 SPEC 0.1 3.1

# 5. Discussion

## 5.1 Summary and Interpretation

The purpose of this analysis was to find the variables within these data sets that have a correlation to causing injuries. First, what kind of injuries? Although we have limited information on the specifics, it should be noted lower body injuries occur more often than upper body and head injuries.

The parameters that appeared to have an effect on injuries would consist of the week in the season, quarter, play type, and player position. The finding from each parameter is summarized below. It is worth mentioning players height and weight also had an effect but would be highly correlated to the player position.  
1. Week during season - A spike of injuries at the beginning of the season followed by a steady increase throughout the rest of the season. Likely due to being unprepared for a regular season game and then more injuries throughout the season as fatigue sets in. 2. Quarter of a game - Injuries steadily increases throughout the course of a game likely due to fatigue. 3. Play type - There is a 2-5% difference in number of injuries compared to the total number ran. Runs have slight lower chance of injury while pass plays have a slight higher chance. 4. Player Position - By and large, linemen (offensive and defensive) have a greater chance of injury.

A notable mention would be number of injuries per team during the regular season. One can expect a difference due to coaching staff, team culture, ambient conditions of home field but the LA Chargers having three times the number of injuries as the LA Rams was unexpected.

## 5.2 Strengths and Limitations

Strengths of this analysis are that it was thorough as a descriptive analysis. As applicable, parameters from the injury data set were compared to the population of the NFL to see if there were any meaningful differences.

With this being said, this analysis consisted of comparing plots and percentages. The next step of this would be to conduct inferential analysis to solidify the findings in this initial analysis. Furthermore, this analysis would be interesting to conduct on multiple years worth of data. This analysis was of a single NFL season and therefore, it would be interesting to see a trend in this data across different seasons.

## 5.3 Conclusions

First, the increased number of injuries being in the lower body is a reminder to coaches and players to stay hydrated and stretch before games to avoid pulled muscles. Although the data does not call out pulled muscles being a major contributor, this is a good reminder to limit injuries if you can.

As for coaches, be aware of the variables discussed in this analysis that have a higher chance of injuries. Injuries occur more often during the first week of the season so perhaps adjustments can be made in the preseason to better prepare the players. The coaches with a higher proportion of injuries should consider collaborating with the coaches of lower numbers to see if there is something different they can do to prevent these for the players. Preseason rituals are likely proprietary but I would think it would be in the interest of the NFL to investigate this and come up with a best practices for the teams.

# 6. References

Mayo Clinic. (n.d.). *Chronic traumatic encephalopathy (CTE)*. <https://www.mayoclinic.org/diseases-conditions/chronic-traumatic-encephalopathy/symptoms-causes/syc-20370921>.

PBS Frontline. (n.d.). *Timeline: The NFL’s concussion crisis*. <https://www.pbs.org/wgbh/pages/frontline/sports/league-of-denial/timeline-the-nfls-concussion-crisis/>.