NBA Player Injuries Analysis

Lauren Cohen

STAT 4490 Fall 2022

Kennesaw State University

**Table of Contents**

1. Introduction

2. Data Collection and Munging

3. Descriptive Statistics

4. Hypotheses and Results

5. Conclusions

6. Potential Future Research

7. R Code

**Introduction**

Are there significant factors to show how NBA players get injured? From height, age, number of games played, and more variables, this research was to analyze NBA players injuries in a descriptive study. The reason for doing this research was as an athlete that played sports, seen some injuries and been injured, I wanted to see if there was a way to analyze injuries in athletes. The initial research was to look at sports such as track and field and basketball to see if they are a cause of back and hip injuries. Due to the limited data found, the project was shaped more to NBA players and the types of injuries they had and how many days they missed when getting injured.

**Data Collection and Munging**

The original dataset came from Hashtag Basketball NBA Injury Database. The main source of this website is collated data from over 4000 NBA injuries since 2010 and analyzed the recovery time needed for each of them. The data was from 2020-2022 and split into 10 separate data sets based on the most common injuries: NBA health and safety protocols, Illness, Sprained left ankle, Sprained right ankle, Sore left knee, Sore right knee, Concussion, Sore left ankle, Sore lower back, and left knee injury. In each of the data sets, the variables are player, team, date injured on, date returned on, and number of days missed. Since the data was split by injury, Microsoft Excel was used to combine all the datasets into one by stacking them on top of one another and adding the variable of type of injury. At the end there were 2182 observations and 6 variables.

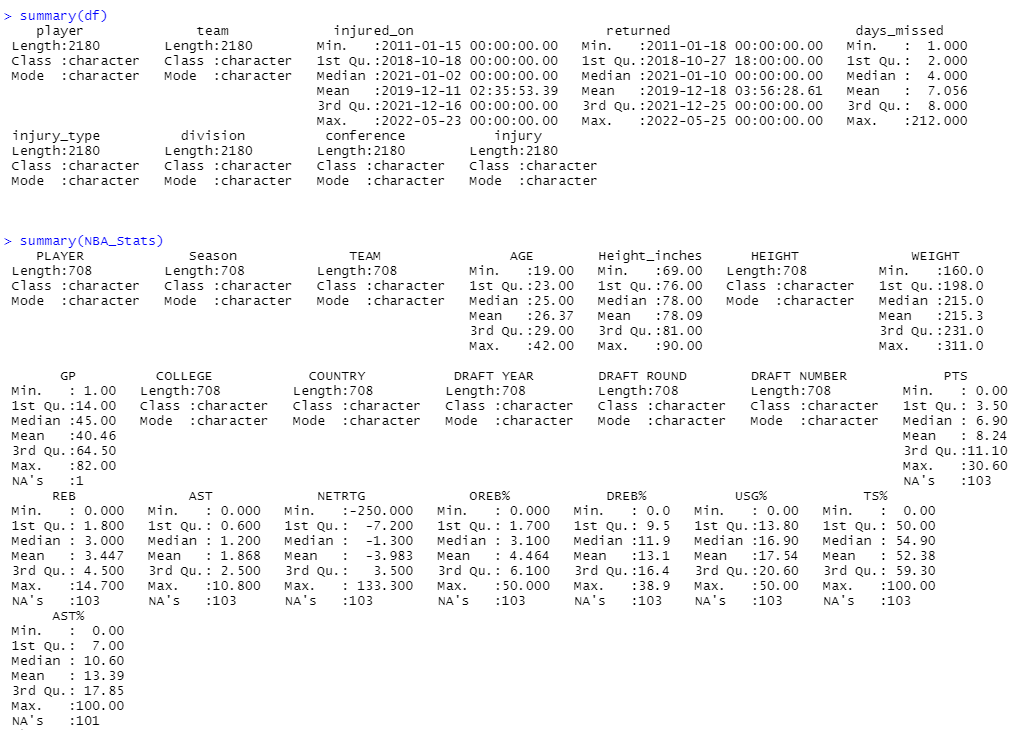
The second dataset that was collected was from NBA.com. To do further analysis of the players in the dataset, data was collected to find more variables such as the height, age, weight, number of games played in the season and more. The first set that was collected was from Regular Season 2021-2022 Player Bio Statistics for the overall season. There were 605 players in this set but since there were a total of 700 more players, the rest of the players were manually typed in by finding the missing age then going to each player’s statistics from their most recent season before 2023 and filling in their age, height and number of games played. At the end, the data set had 708 observations and 22 variables. Since excel was also used for this dataset creation, then while putting the height the format was changed to date, so this needed to be hand changed to height in inches.

When importing the data into R Studio, both data sets were imported. There were 4 variables that were also created in the data set. The first one created was ‘division’ which mutated the teams to select the teams into each of the divisions of Atlantic, Southeast, Central, and Northwest; this was looked up on NBA website to find divisions. The second variable that was created mutated the division into Eastern and Western Conferences. The third variable mutated the types of injuries. As each of the injuries were split by the side of the body, it mutated to the location of the body such as taking the 11 types of injuries into 5: Illness, Ankle, Back, Head, and Knee. The last variable that was created was Age group which was performing discretization on the continuous variable of age to split into bins in order to make the age more normal and account for the outliers such as players that play past the average retiring age of the NBA.

**Descriptive Statistics**

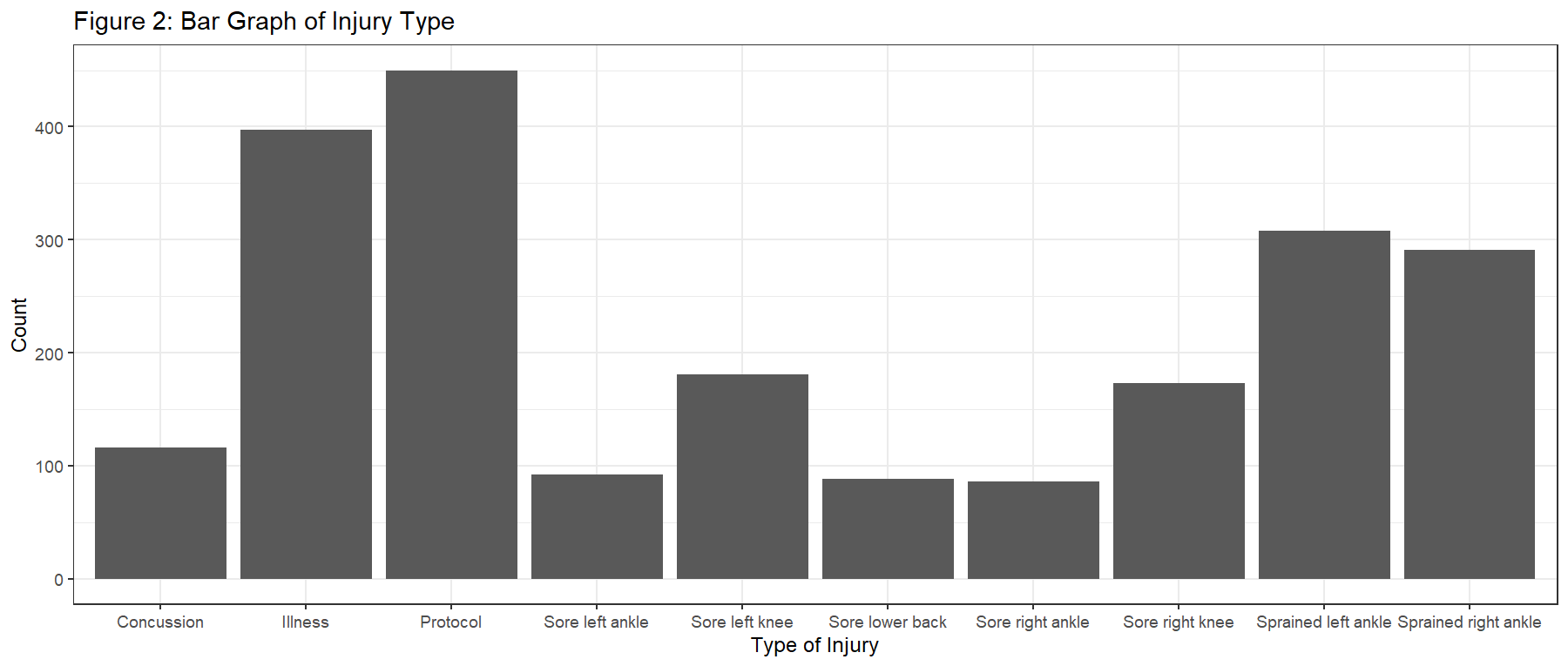
One of the first things I did with this data was looking at the data descriptively. I ran a summary of the variables which can be shown below in Figure 1.

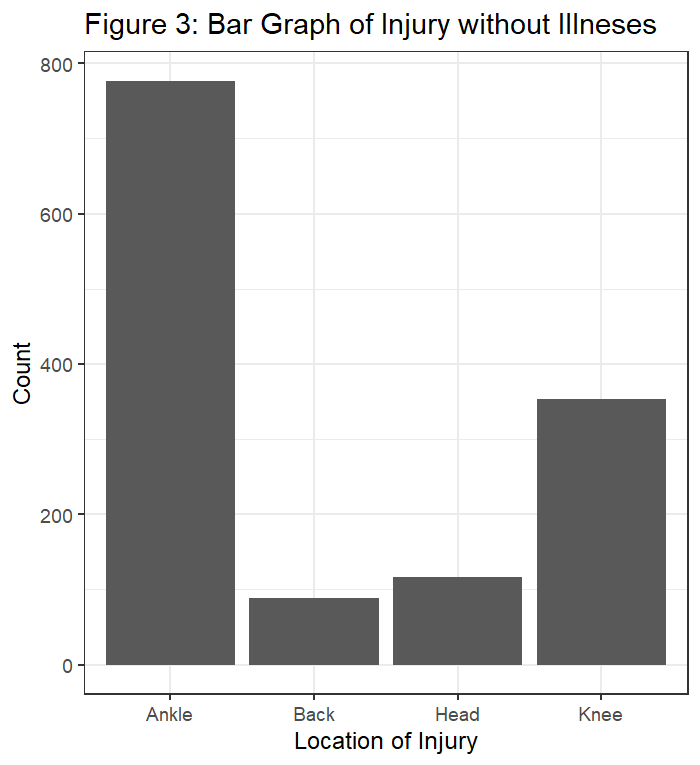
**Figure 1: Descriptive Statistics for Datasets**



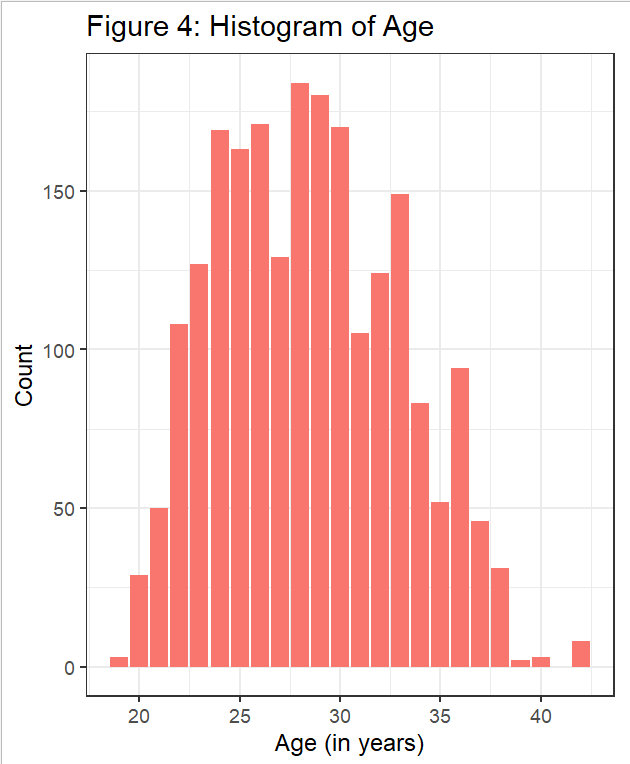
Most of the variables in the original data set were categorical variables for the expectation of Injured On, Returned, and Days Missed. Most of the variables in the NBA Statistics dataset were numerical variables except for player, season, team, college, country, draft year, draft round and draft number.

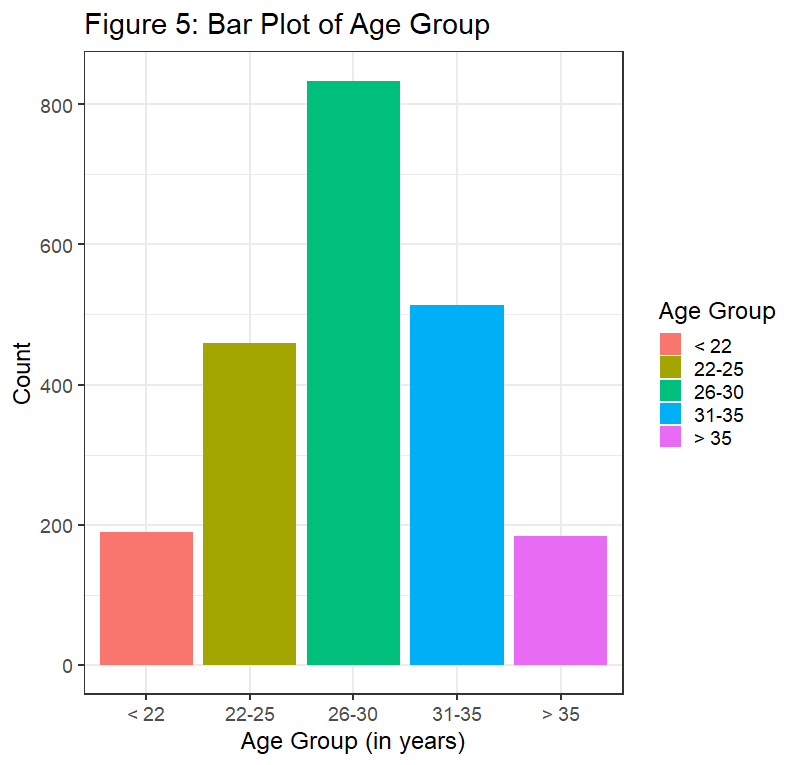
The first variable that was looked at in more detail was Injury Type. The injury that was listed as ‘NBA health and safety protocols’ was changed to Protocols. As seen in Figure 2, the most common injuries were Protocols and Illness. When we looked up on the NBA website what the meaning of protocols was it related to Covid safety protocols. Since this dataset was found during 2020-2022 the protocols and illnesses were related to Covid procedures. As this project was not focused on Covid but focused on injuries, the rest of the research was taking out observations related to Illness/Protocols. Figure 3 shows the distribution of the most common type of variables removing the covid observations and using injury from the location of the body. The most common injury is ankle injuries.





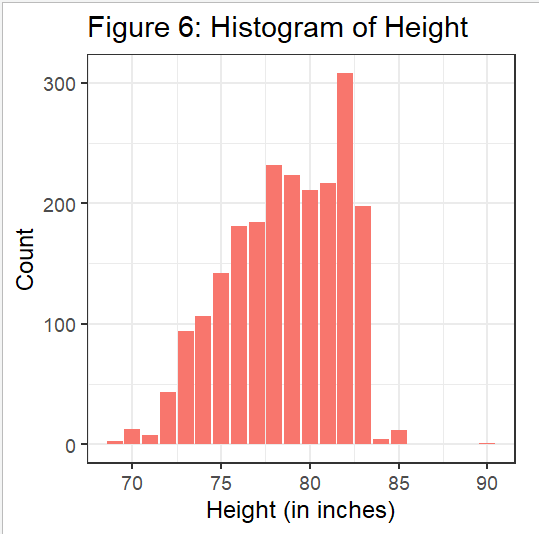
The next variable that was looked at was age. Figure 4 shows the histogram of the distribution of age. The distribution is normal, but those above the age 40 years old there are some outliers. In the NBA, the typical age of retirement is 33 years old. In order to account for the outliers, there was a new variable of age group created which is discretized into 5 groups. Looking at Figure 5, you can see the distribution of age group which looks more normally distributed.





Another interesting observation found when looking at the data was looking at players that had the greatest number of injuries, these two players being: "Giannis Antetokounmpo" and "Derrick Rose" with 26 injuries each.

Figure 6 below shows a Histogram of the Height in inches. As we can see, the distribution is left skewed with some outliers in the higher inches.

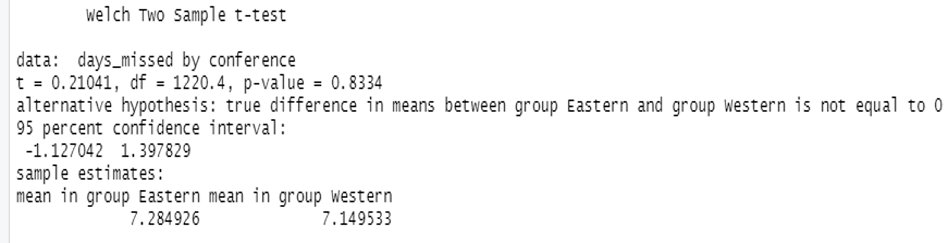


**Hypotheses and Results**

Hypothesis 1

The first hypothesis was: Mean number of days missed from injury in the Eastern Conference are different than the mean number of days missed in the Western Conference. There was no real reason as to why this hypothesis was studied; it was an initial hypothesis to look at the data of the set. A two-sample t-test was run for this hypothesis. The output is shown below in Figure 7. The null and alternative hypotheses are H0 (null): EC – WC =O and HA (alternative): EC – WC ≠ O.

**Figure 7: Output of Two Sample T-Test for Number of Days Missed and Conference**

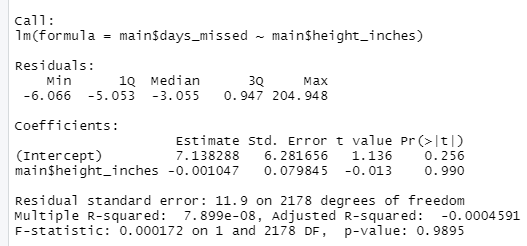


The p-value of the test is very high at 0.8334. The 95% confidence interval is (-1.13,1.40), which includes 0 in the interval. Also looking at the means the Eastern conference mean is only ~0.14 more than the Western Conference. Therefore, we fail to reject the null hypothesis, the average number of days missed for injury for the Eastern conference is not significantly different than for the Western conference.

Hypothesis 2

The second hypothesis researched was: taller players are more likely to miss more days while being injured. In order to research this, a linear regression was run. This linear regression was run not to predict the number of days that were missed from height of the player but to see if height was a significant variable to impact number of days missed of a player. The variable used in this regression was height in inches which was the variable made in excel changing the format of height recognized as a text date to height as a numerical value in inches. Figure 8 below shows the summary of the linear regression on height and number of days missed.

**Figure 8: Linear Regression Output of Days Missed and Height (in inches)**



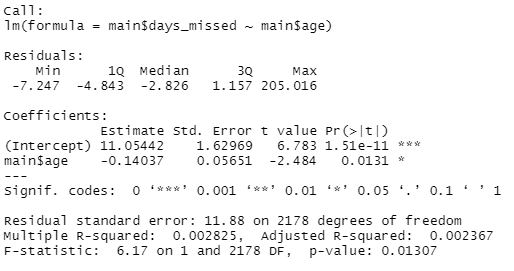
The R-squared value in this model is really small which shows this model is not good for prediction. The p-value for height in inches is very high which also suggests that we cannot use this variable to predict or count for significance on influencing the number of games the player missed when injured.

When looking at Figure 6, there are outliers in the graph. Anything higher than 85 inches was removed and the regression was re-run again. The p-value decreased to 0.936, which is still very high.

Hypothesis 3

The third hypothesis was: The older players are more likely to miss more days while being injured. In order to research this, a linear regression was run. This linear regression was run not to predict the number of days that were missed from age but to see if this age variable is significant and if there is a positive or negative slope to know older or younger players were to miss more days. Figure 9 below shows the summary of the linear regression ran on age and number of days missed.

**Figure 9: Linear Regression on Age and Days Missed**



The first thing to note from the output is the p-value of 0.0131. This means age is a significant enough variable to predict the days missed from an injury for a player. Another thing to mention is that the R-squared of this model is very low, which suggests this model is not a good predictor. In this type of analysis, it is okay because we are not here to use age to predict the number of days missed in a player’s season from injury, but rather we are just trying to see if age has an impact in general on how many days the player has missed.

As the age of the players increases by 1 year, the average number of days missed decreases by 0.14037 days. Since the slope is negative, this result contradicts the initial hypothesis that older players miss more days than younger players when they are injured. It was quite interesting to see this reverse, as for reason it could be this way, it will be discussed more in the next hypothesis.

Hypothesis 4

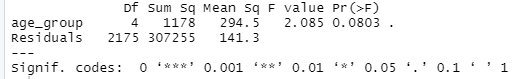
The fourth hypothesis that was analyzed was: There is a difference in the age groups and how many games the player missed. The reason the analysis was done was to compare how it was to analyze the age variable as the continuous variable vs. The categorical discretized variable. In order to look at this analysis an ANOVA test was run.

Before running the ANOVA test, a bar graph and combined scatter plot was made to look at the variable of days missed on the y axis, and the age group that is on the x axis. Figure 10 below shows this graph. Looking at the age groups, the age group of less than 22 years old has the highest mean of number of days missed from injury than any other age group. This also being said how this age group as seen back in Figure 5 has the second to lowest observations in this variable.



Figure 11 below shows the output from the ANOVA test. From the p-value of 0.0803, this test is significant under the alpha level of 0.10. Thus, we can reject the null hypothesis. There is enough statistically significant evidence that at least one of the means of the age groups are different from the others. Looking at Figure 10, it can be assumed the mean number of days missed that is different is the age group of ages less than 22 years old.

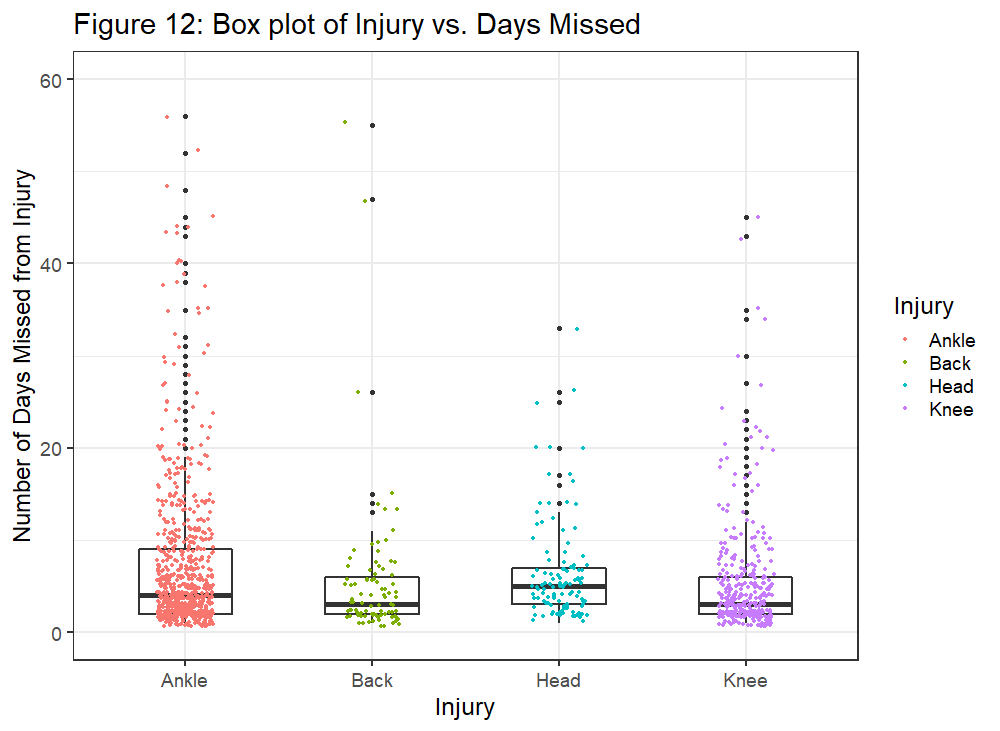
**Figure 11: ANOVA Test Output for Age Group and Number of Days Missed on Injury**



Hypothesis 5

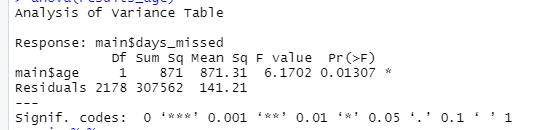
The third hypothesis is: There is a difference in the location of injury with how many games the players missed. This hypothesis is like the previous hypothesis. It was analyzed the same way by looking first at a bar graph combined with scatter plot of the location of the injury and number of days missed on the y axis. For the type of injury, the new variable of general location of the injury was used as the results for the actual type of injury were not as conclusive to see when more distributly spread out.

Figure 12 shows the graph of the location of injury and number of days missed for the players. As shown below, the injury with the highest mean number of days missed is injuries that occurred to the head. Even though the group has a fairly lower number of observations compared to the other locations of injuries on the body.



After running the ANOVA (results) are shown below in Figure 13. The p-value of 0.0107 is significant at the 0.05 alpha level. This means that we can reject the null hypothesis; there is enough statistical evidence to show that at least one of the means of the number of days missed is different than the others.

**Figure 13: ANOVA Test Output for Age Group and Number of Days Missed on Injury**

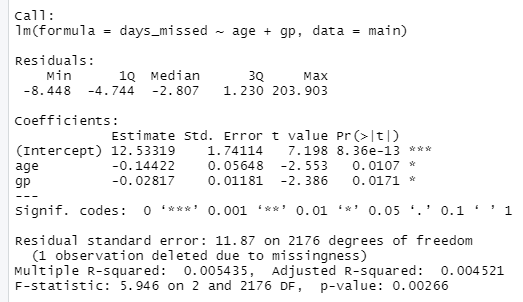


Hypothesis 6

The next part of the project is not much of a hypothesis but rather a look into the relationship between age, number of games played in the season and the number of days missed in the season through the lens of a multiple linear regression. Like Hypothesis 3, the regression was not run to predict the number of games that were going to be missed but rather to see if the variables of age and number of games played were significant to how many games the player missed from injury.

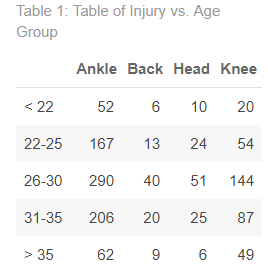
Figure 14 shows the output of the multiple linear regression. In the regression, the R-squared value is also very low meaning it would not be a good model to use for predicting, but both age and gp (which is number of games played) show significant p-values at the alpha level of 0.05. This means they have significant impacts on how many days are missed. As seen previously, age has a similar effect as seen in Hypothesis 3. In this regression, as the number of games played increases by 1 game, the number of games missed by the player decreases by 0.02817.

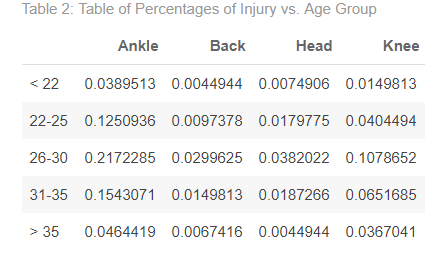
**Figure 14: Multiple Linear Regression of Days Missed from Age and Games Played**



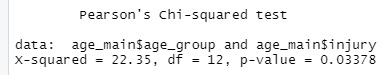
Hypothesis 7

The final hypothesis of this project was looking at the relationship between the age group variable and the location of the injury variable. In order to look at this, first a table was produced to see if there were any interesting cells that stood out. In Table 1, we see there is the cell of 290 observations in the age of 26-30 range under ankle injuries. This is the largest amount, but it was wondered if this was the most percentage wise because looking at Figures 3 and 5, these are also the top two categories of each variable. Looking at Table 2, the percentages show that this still is the largest amount with about 21.7%.





The next test that was run to analyze these variables was running a chi square test. In Figure 15, there is the output of the chi square test. The P-value of 0.03378 is less than 0.05, therefore it is significant. There is enough evidence to say that the probability of the type of injury a player receives is affected by the age (group) of the player.

**Figure 15: Output of Chi Square Test of Age Group and Injury** 

**Conclusions**

From hypothesis 1, looking at the conference or even division of the player who got injured, it is hard to use a potential predictor for how many days a player will miss.

From hypothesis 2, it is very hard to predict injuries days missed from height of players. It was thought taller players could miss more days because taller players can tend to get injured more due to joint movements and more, but it was the opposite effect.

There were a lot of different results that were opposite than the hypothesis. After analyzing hypotheses 3 and 4, it was really shocking to see the results of the lower ages having more days missed than the other ages. Thinking more about why this could be the case, there were many conclusions that came up.

Those conclusions consist of the following: Since NBA players are more in their prime at younger ages, if a younger player gets injured, it is potential for the coach of the team to have the player sit out longer to recover more in order to come back and play in more important games of the season. Another factor that can play into injury sit out time is the price of paying NBA players for the injuries. When players are injuries, some of them still can get salary if they return to play at the end of the season whereas some are forced to retire and deal with salary/pay parts later down the road. Thus, saying that, some older players might not be as concerned about the money as the younger players. The last factor to mention is from a biological standpoint. Typically, males do not reach full maturity until the age of 25. Therefore, if a player under 25 gets injured it may take their body to heal longer than older players who have reached full maturity physically.

After going over hypothesis 5, it makes sense to find the head to be the location of the injury with the greater number of days missed. With head injuries, observations such as concussion were listed under there. Concussions are very dangerous that if you do not seek help you will end up not making it, so if an NBA player gets one then it makes sense, they would have to sit for a long period of time to get the information. For injuries like ankle and knee, a lot of them were sore or sprained knee/ankle. These injuries are less severe than concussions, in which players can put on a brace or use RICE (Rest, Ice, Compression, Elevation) then return to the court sooner.

In hypothesis 6, it was interesting to see how the number of games played had a decreasing effect on the number of games missed from an injury. Originally it was thought that the more games a player played in a season, the more likely they are to get injured and when injured they need to sit out more games, but this was not the case. As for the reason why, the results could be this way, more research would need to be done to explain why the results would have come out this way.

**Potential Future Research**

In overall conclusion, it's noted to say that trying to predict injuries from NBA players is not easy. Though there were some insignificant results and low chances of models being used, the project developed a lot of insight on what can potentially be looked at for future research.

If I were to continue this research, I would look at the data from one season of players in order to find more accurate time frames. In this, I would additionally look at not just the number of games played but the number of minutes played by the player in the whole season. The reason this could be more beneficial is because even though players can be counted for a certain number of games played, some of them can only play for a few seconds where some can play for the whole game. Thus, even looking at age it's potential that older players do not play as long in the games as younger players so the results could come out differently for how long they miss when they are injured or even how often they get injured in general.

The other factor I would investigate is the prominence of reoccurring injuries. Some players may get injured multiple times and the more times you get injured in one place, the more likely the player can miss out on games being injured from that injury.

In this current research, I opted to combine the injuries with the side of the body: left or right. Though this might be helpful to look more overall, it would be interesting to see how the side of the bodies can affect when it comes to the players dominant side. Possibly seeing are players left-handed or right-handed and if this side of the body has more injuries than their nondominant side. The last factor to think of checking is the position of the player. To wonder if certain positions receive more types of injuries than others or if they tend to stay out longer.

To end, there are many comments that can be made regarding this research and many avenues to travel. The purpose of this research and the future research is to look and see if there are patterns in NBA players' injuries in order to help see how to prevent the players from missing many days while being injured and eventually stop injuries.

**R Code**

#loading in the packages

library(readxl)

library(dplyr)

library(ggplot2)

library(outliers)

library(devtools)

library(cowplot)

library(dataRetrieval)

library(kableExtra)

library(knitr)

#loading in the datasets

injury <- read\_excel("NBA\_Injuries\_Since\_2010.xlsx")

NBA\_Stats <- read\_excel("NBA\_Stats.xlsx")

# changing column names

injury <- setNames(injury, c("player","team","injured\_on", "returned", "days\_missed", "injury\_type"))

#creating new variables (sorting teams, injury, age)

df <- injury %>%

na.omit() %>%

mutate(division = case\_when(team == 'Celtics' | team == 'Nets'| team == '76ers'| team == 'Raptors' ~ 'Atlantic',

team == 'Bulls' | team == 'Cavaliers' | team == 'Pistons' | team == 'Pacers' | team == 'Bucks' ~ 'Central',

team == 'Hawks' | team == 'Hornets' | team == 'Heat' | team == 'Magic' | team == 'Wizards' ~ 'Southeast',

team == 'Nuggets' | team == 'Timberwolves' | team == 'Thunder' | team == 'Blazers' | team == 'Jazz' ~ 'Northwest',

team == 'Warriors' | team == 'Clippers' | team == 'Lakers' | team == 'Suns' | team == 'Kings' ~ 'Pacific',

team == 'Mavericks' | team == 'Rockets' | team == 'Grizzlies' | team == 'Pelicans' | team == 'Spurs' ~ 'Southwest'

)) %>%

mutate(conference = case\_when(division == 'Atlantic' | division == 'Central' | division == 'Southeast' ~ 'Eastern',

division == 'Northwest' | division == 'Pacific' | division == 'Southwest' ~ 'Western')) %>%

mutate(injury = case\_when(injury\_type == 'Protocol' | injury\_type == 'Illness' ~ 'Covid',

injury\_type == 'Sprained left ankle' | injury\_type == 'Sprained right ankle'| injury\_type == 'Sore left ankle' | injury\_type == 'Sore right ankle' ~ 'Ankle' ,

injury\_type == 'Sore left knee' | injury\_type == 'Sore right knee' ~ 'Knee',

injury\_type == 'Concussion' ~ 'Head',

injury\_type == 'Sore lower back' ~ 'Back'))

colnames(NBA\_Stats)[which(colnames(NBA\_Stats) == 'PLAYER')] <- 'player'

main <- df %>% left\_join(NBA\_Stats) %>%

select(-'TEAM')

names(main) <- tolower(names(main))

main <- main %>%

mutate(

age\_group = dplyr::case\_when(

age <= 22 ~ "< 22", # age when people graduate college

age > 22 & age <= 25 ~ "22-25",

age > 25 & age <= 30 ~ "26-30",

age > 30 & age <= 35 ~ "31-35",

age > 35 ~ "> 35"

),

# Convert to factor

age\_group = factor(

age\_group,

level = c("< 22","22-25","26-30", "31-35", "> 35")

)

)

# Descriptive Stats

summary(df)

#Injury

injury %>% ggplot(aes(injury\_type)) + geom\_bar() + theme\_bw(base\_size = 18) +

labs(title= "Figure 2: Bar Graph of Injury Type", x="Type of Injury",

y= "Count")

main %>%

filter(injury != 'Covid') %>%

ggplot(aes(injury)) + geom\_bar() + theme\_bw(base\_size = 18) +

labs(title= "Figure 3: Bar Graph of Injury without Illneses", x="Location of Injury",

y= "Count")

#Age

main %>%

ggplot(aes(age, fill = "chocolate1")) + geom\_bar() + theme\_bw(base\_size = 18) +

labs(title= "Figure 4: Histogram of Age", x="Age (in years)", y= "Count")

#Age Group

main %>%

ggplot(aes(age\_group, fill = age\_group)) + geom\_bar() + theme\_bw(base\_size = 18) +

labs(title= "Figure 5: Bar Plot of Age Group", x="Age Group (in years)", y= "Count", fill = 'Age Group')

#finding player with the most injuries

find\_mode <- function(x) {

u <- unique(x)

tab <- tabulate(match(x, u))

u[tab == max(tab)]}

find\_mode(injury$player)

# Height

main %>%

ggplot(aes(height\_inches, fill = "chocolate1")) + geom\_bar() + theme\_bw(base\_size = 18) +

labs(title= "Figure 6: Histogram of Height", x="Height (in inches)", y= "Count")

# Hypothesis 1: Mean number of days missed from injury in the Eastern Conference are different than the mean number of days missed in the Western Conference

t.test(data=df2, days\_missed ~ conference)

# Hypothesis 2: Taller players are more likely to miss more days while being injured

results <-lm(main$days\_missed~main$height\_inches)

summary(results)

# Hypothesis 3: The older players are more likely to miss more days while being injured

results\_age <-lm(main$days\_missed~main$age)

summary(results\_age)

# Hypothesis 4: There is a difference in the age groups with how many games the players missed

main %>%

ggplot(aes(x=age\_group, y=days\_missed)) +

geom\_boxplot(width=0.5,lwd=1.0) +

theme\_bw(base\_size = 18) +

geom\_jitter(width=0.15,size=1, aes(color=age\_group)) +

scale\_y\_continuous(limits = c(0,60)) +

labs(title= "Figure 10: Box plot of Age Group vs. Days Missed", x="Age Group (in years)",

y= "Number of Days Missed from injury", colour = 'Age Group')

results\_age\_group <-aov(days\_missed~age\_group, data=main)

summary(results\_age\_group)

# Hypothesis 5: There is a difference in the location of injury with how many games the players missed

main %>%

filter(injury != 'Covid') %>%

ggplot(aes(x=injury, y=days\_missed)) +

geom\_boxplot(width=0.5,lwd=1.0) +

theme\_bw(base\_size = 18) +

geom\_jitter(width=0.15,size=1, aes(color=injury)) +

scale\_y\_continuous(limits = c(0,60)) +

labs(title= "Figure 12: Box plot of Injury vs. Days Missed", x="Injury",

y= "Number of Days Missed from Injury", colour = 'Injury')

results\_injury <-aov(data=main, days\_missed ~ injury)

summary(results\_injury)

# Multiple Linear Regression for Days Missed for Age and Games Played

results\_both <- lm(days\_missed ~ age + gp , data=main)

summary(results\_both)

# Relationship between Age Group and Injury chi square test

age\_main <-main %>%

filter(injury != 'Covid')

table<- table(age\_main$age\_group, age\_main$injury)

table %>%

kbl(caption = "Table 1: Table of Injury vs. Age Group") %>%

kable\_paper("striped", full\_width = F) %>%

kable\_styling()

table2 <- prop.table(table)

table2 %>%

kbl(caption = "Table 2: Table of Percentages of Injury vs. Age Group") %>%

kable\_paper("striped", full\_width = F) %>%

kable\_styling()

chisq.test(age\_main$age\_group, age\_main$injury, correct=FALSE)