Trends in NFL Injuries Over Time (2009-2020)

Lauren Mock

12/15/2021

I prepared this material independently for a group project in a data science course at Harvard. This file contains my exploratory data analysis to answer the following question: how have NFL injuries changed over time? Each member of my group addressed a different question.

```
# load libraries
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(pander)
library(forcats)
library(RColorBrewer)
```

Read in and clean NFL injury data (web scraped by my team member)

```
# read in data
injuries <- read.csv("all_injuries_clean.csv", stringsAsFactors = FALSE)

# new column for total injuries (sum over all injury types)
injuries <- injuries %>%
  mutate(total_injuries = select(injuries, c("head", "shoulder", "upper_torso",
                                             "lower_torso", "arm", "hand",
                                             "leg", "foot")) %>% rowSums)

injuries <- injuries %>%
  filter(year != 2021) %>% # filter out 2021 (incomplete data for current season)
  filter(total_injuries != 0) # filter out people who missed games due to illness, etc.

# print column names
names(injuries)
```

```
##  [1] "name"              "full_team"         "team"
##  [4] "year"              "games_in_season"   "num_games_injured"
##  [7] "num_games_missing" "injury_types"      "earliest_injury"
## [10] "latest_injury"     "head"              "shoulder"
## [13] "upper_torso"       "lower_torso"       "arm"
```

```
## [16] "hand"              "leg"               "foot"
## [19] "total_injuries"
```

Each row has a unique combination of player name and season. The columns called "head," "shoulder," etc. include counts of the number of head/shoulder injuries a given player had during a given season. The "total_injuries" column is a sum of all injuries a given player had in a given season. It is important to note that due to incomplete data among non-injured players, this data set only includes players who were *injured*.

Total injuries per season

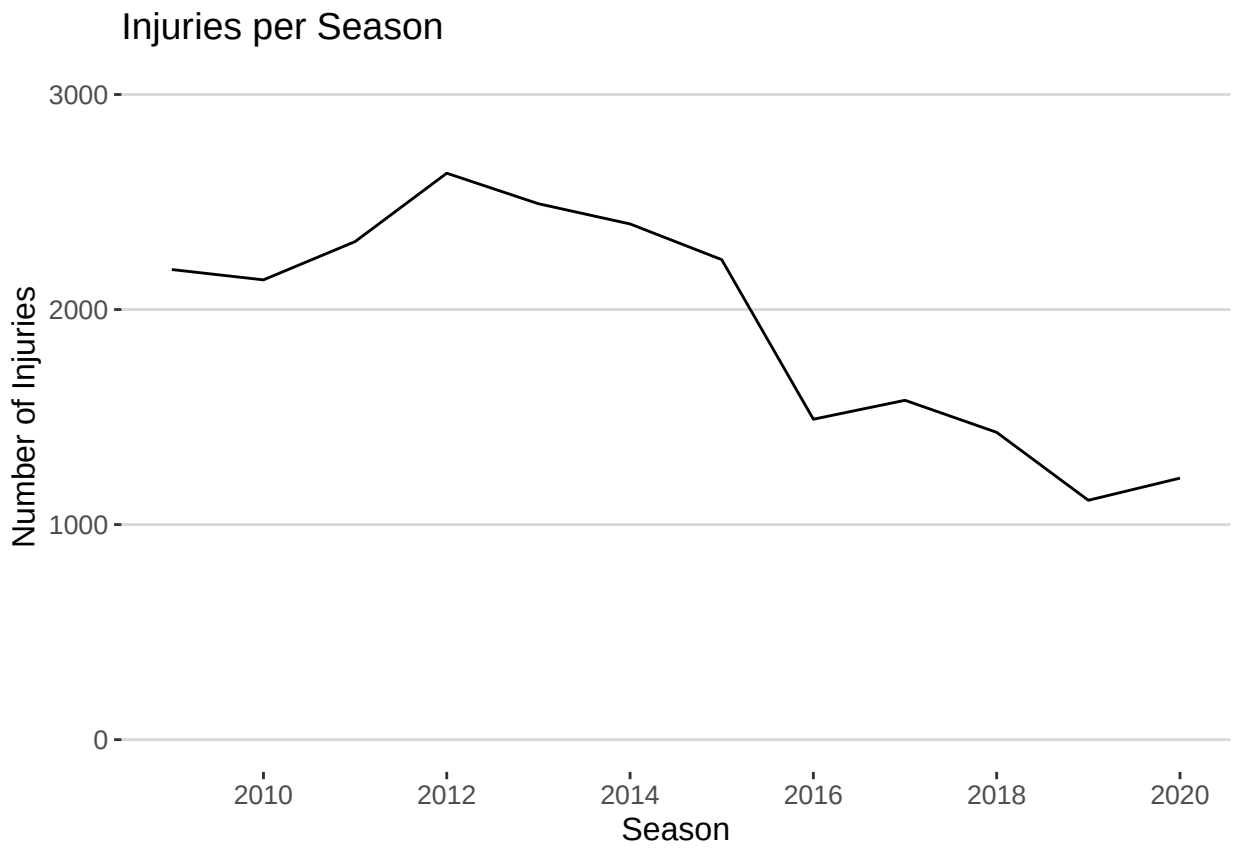
First, let's look at the total number of injuries per season.

```
# create summary table
inj_by_year <- injuries %>%
  group_by(year) %>%
  dplyr::summarize(total_injuries = sum(total_injuries))

# view table
pander(inj_by_year)
```

year	total_injuries
2009	2186
2010	2138
2011	2316
2012	2634
2013	2492
2014	2398
2015	2232
2016	1490
2017	1578
2018	1429
2019	1113
2020	1216

```
# plot
inj_by_year %>%
  ggplot(aes(x = year, y = total_injuries)) +
  geom_line() +
  ggtitle("Injuries per Season") +
  xlab("Season") +
  ylab("Number of Injuries") +
  ylim(0, 3000) +
  scale_x_continuous(breaks = seq(2010, 2020, by = 2)) + # change x-axis ticks
  theme_hc()
```

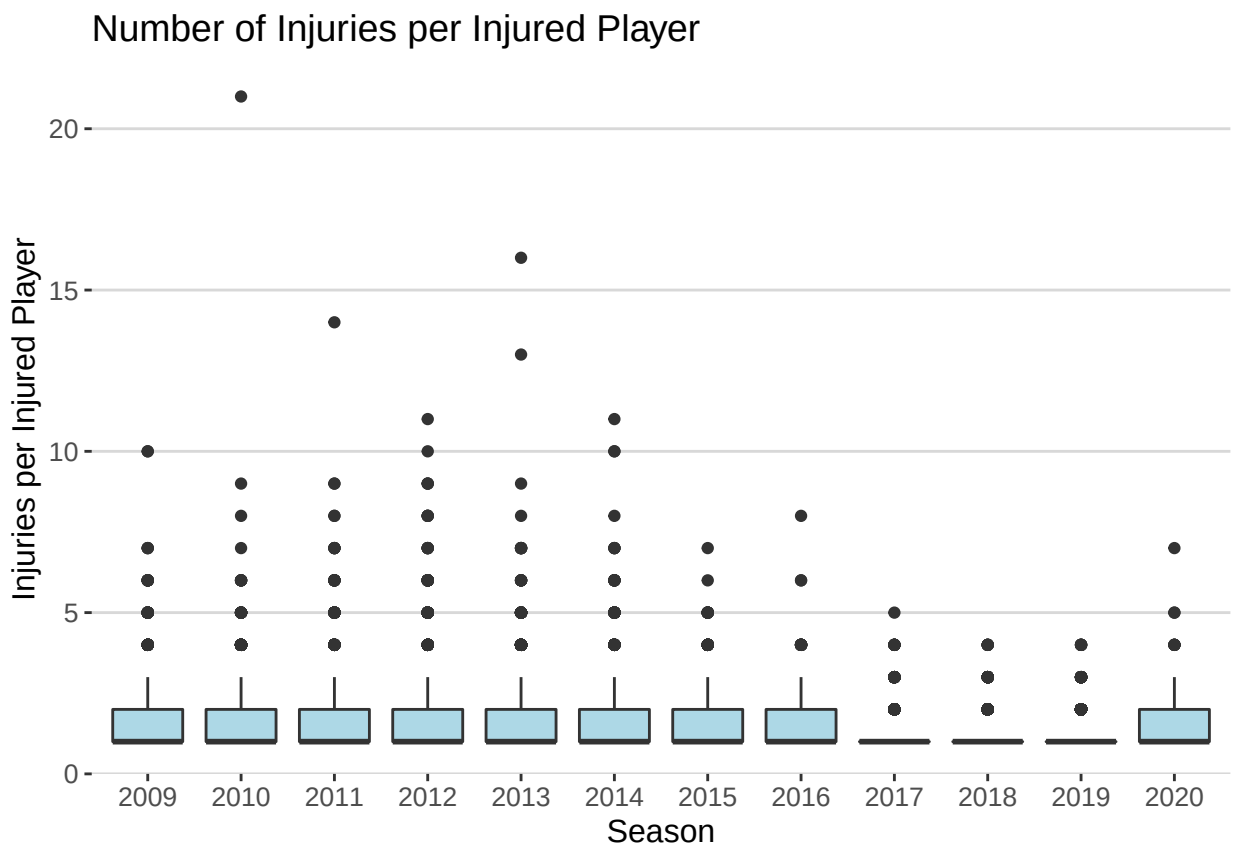


It looks like the total number of injuries per season declined over this time period, but there is a lot more to explore here.

Number of injuries per player

Next, let's look at the number of injuries per player per season. All players in this data set were injured, so we are looking at the number of injuries among players who were injured in a given season.

```
injuries %>%
  ggplot(aes(x = as.factor(year), y = total_injuries)) +
  geom_boxplot(fill = "lightblue") +
  ggtitle("Number of Injuries per Injured Player") +
  xlab("Season") +
  ylab("Injuries per Injured Player") +
  theme_hc()
```



The median number of injuries per player, among players who were injured, was 1 in every season.

Injuries per season by team

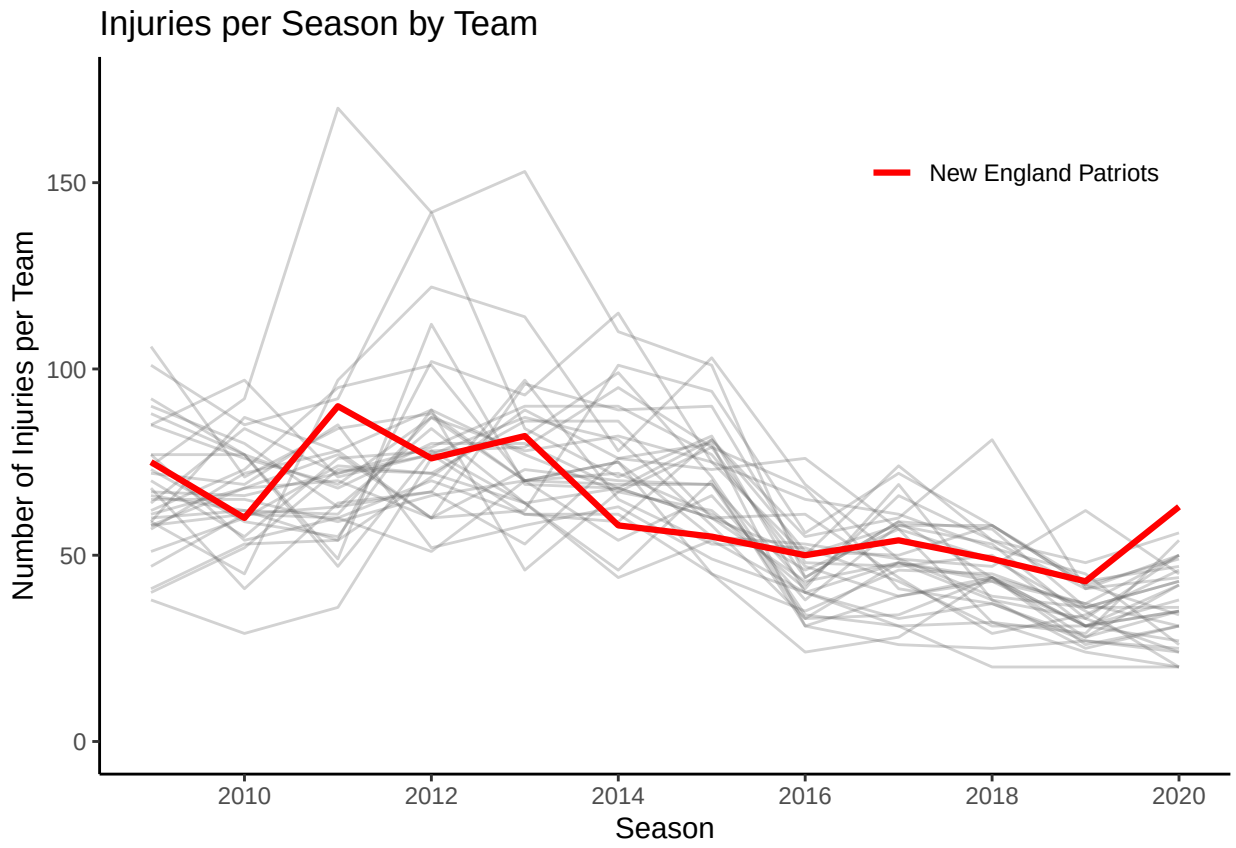
Are these trends consistent across all NFL teams?

```
# create summary table
inj_by_team <- injuries %>%
  group_by(year, full_team) %>%
  dplyr::summarize(total_injuries = sum(total_injuries))

# view first 6 rows of table
pander(head(inj_by_team))
```

year	full_team	total_injuries
2009	Arizona Cardinals	66
2009	Atlanta Falcons	70
2009	Baltimore Ravens	92
2009	Buffalo Bills	73
2009	Carolina Panthers	65
2009	Chicago Bears	68

```r
# plot
inj_by_team %>%
  ggplot() +
  geom_line(aes(x = year, y = total_injuries, group = full_team),
            color = "grey40", alpha = 0.3) +
  # highlight one team in red
  geom_line(aes(x = year, y = total_injuries, color = "New England Patriots"), size = 1.1,
            data = filter(inj_by_team, full_team == "New England Patriots")) +
  ggtitle("Injuries per Season by Team") +
  xlab("Season") +
  ylab("Number of Injuries per Team") +
  ylim(0, 175) +
  labs(color = "") + # remove legend title
  scale_color_manual(values = c("New England Patriots" = "red")) + # manual legend
  scale_x_continuous(breaks = seq(2010, 2020, by = 2)) +
  theme_classic() +
  theme(legend.position = c(.95, .95),
        legend.justification = c("right", "top")) # legend position
```



I highlighted the New England Patriots to see how injuries among players on this team compare to injuries among players on other teams, but this plot could easily be made interactive with an R Shiny app. There is substantial variation in injury counts by team, but the overall trends appear to be similar.

Total head injuries per season

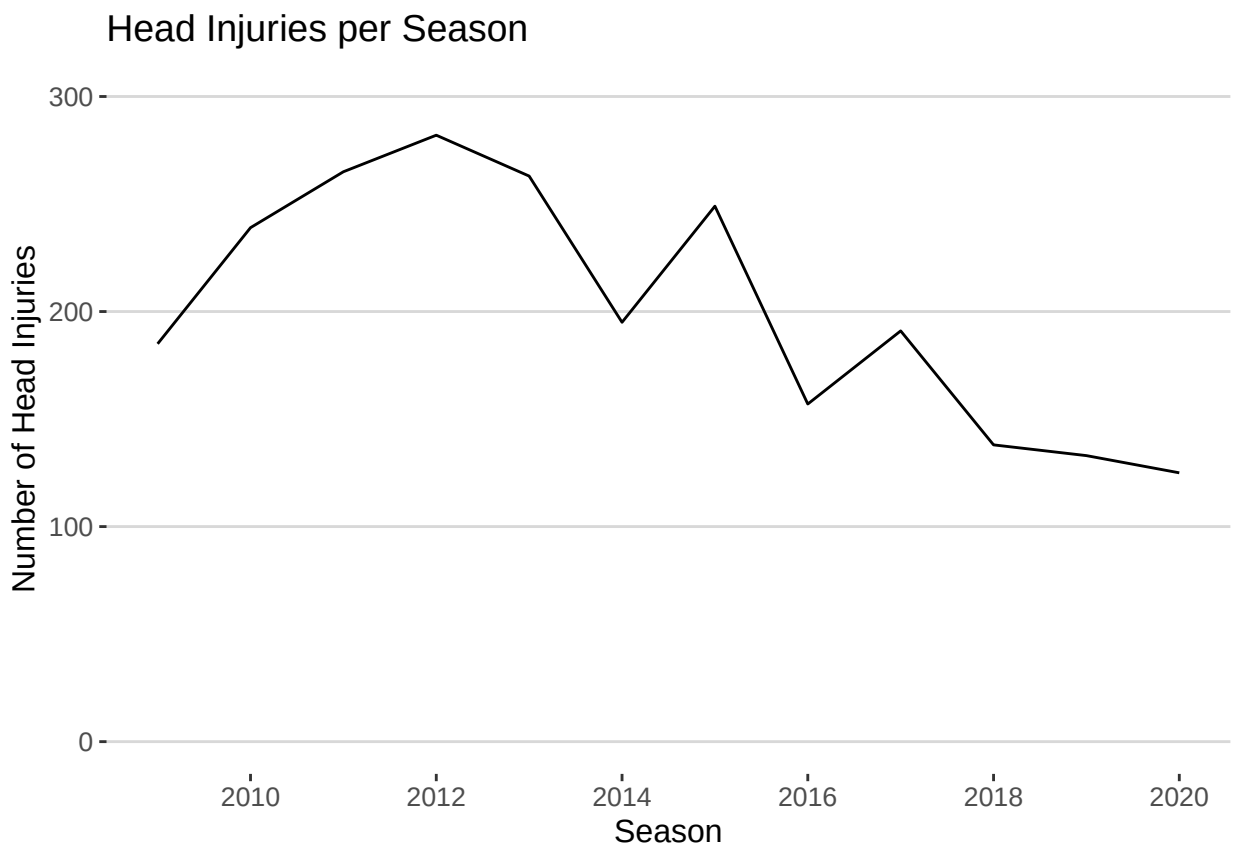
Concussions have been a major focus in recent conversations about football, so let's see if the number of head injuries per season has declined over time.

```
# create summary table
concussions <- injuries %>%
  group_by(year) %>%
  dplyr::summarize(head = sum(head))

# view table
pander(concussions)
```

year	head
2009	185
2010	239
2011	265
2012	282
2013	263
2014	195
2015	249
2016	157
2017	191
2018	138
2019	133
2020	125

```
# plot
concussions %>%
  ggplot(aes(x = year, y = head)) +
  geom_line() +
  ggtitle("Head Injuries per Season") +
  xlab("Season") +
  ylab("Number of Head Injuries") +
  ylim(0, 300) +
  scale_x_continuous(breaks = seq(2008, 2020, by = 2)) +
  theme_hc()
```



It does look like the number of head injuries per season has declined slightly over time.

Head injuries per season by team

Is this consistent across all teams in the NFL?

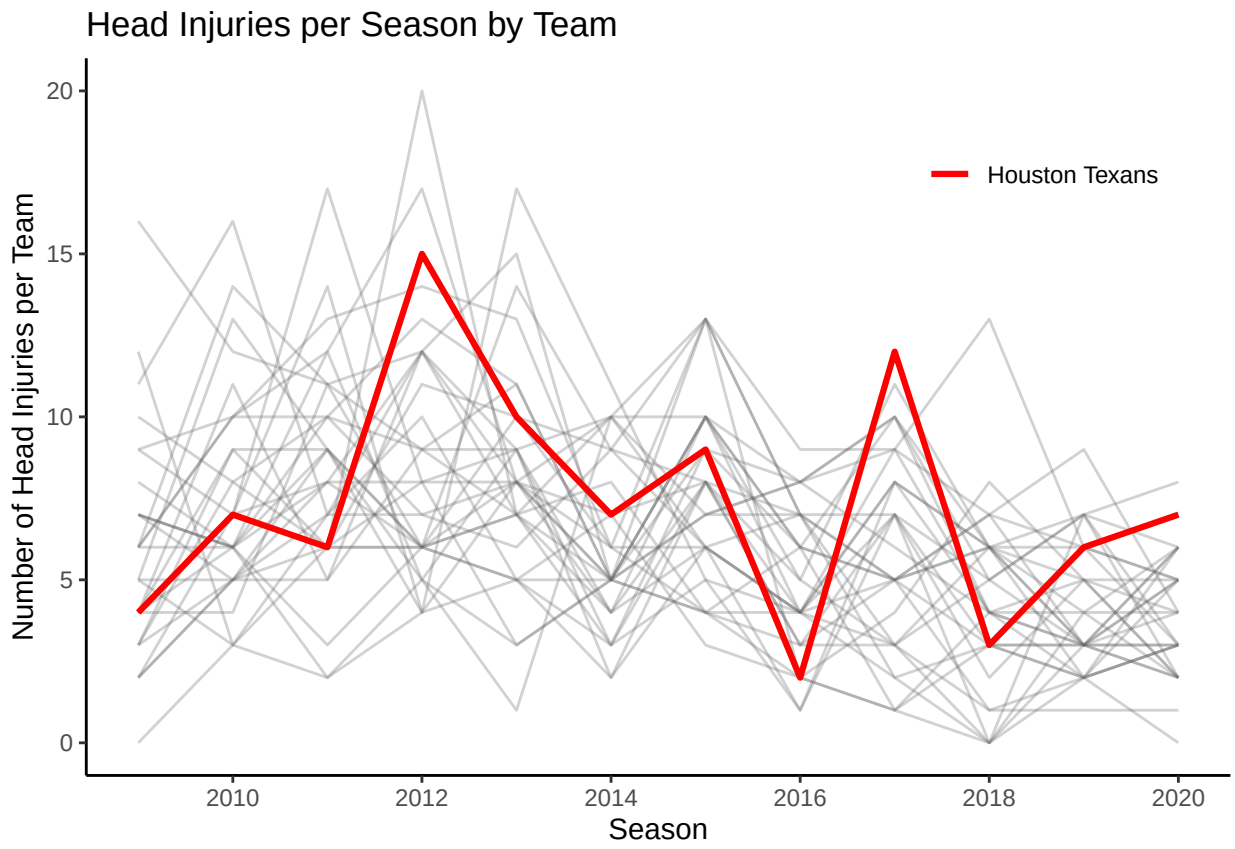
```
concussions_team <- injuries %>%
  group_by(year, full_team) %>%
  dplyr::summarize(head = sum(head))

# view first 6 rows of table
pander(head(concussions_team))
```

year	full_team	head
2009	Arizona Cardinals	7
2009	Atlanta Falcons	7
2009	Baltimore Ravens	16
2009	Buffalo Bills	5
2009	Carolina Panthers	5
2009	Chicago Bears	4

```
concussions_team %>%
  ggplot() +
```

```
  geom_line(aes(x = year, y = head, group = full_team), color = "grey40", alpha = 0.3) +
  geom_line(aes(x = year, y = head, color = "Houston Texans"), size = 1.1,
            data = filter(concussions_team, full_team == "Houston Texans")) +
  ggtitle("Head Injuries per Season by Team") +
  xlab("Season") +
  ylab("Number of Head Injuries per Team") +
  ylim(0, 20) +
  labs(color = "") + # remove legend title
  scale_color_manual(values = c("Houston Texans" = "red")) + # manual legend
  scale_x_continuous(breaks = seq(2008, 2020, by = 2)) +
  theme_classic() +
  theme(legend.position = c(.95, .95),
        legend.justification = c("right", "top")) # legend position
```



There appears to be substantial variation here, with the number of head injuries per team per season ranging from 0 to 20. When we look at one team, such as the Houston Texans, we do not see any clear trends over the time period.

Injuries over time by injury type

We can also see if the same trends hold true for all injury types.

```
# gather into long format for ggplot
gathered <- injuries %>%
  gather(key = broad_injury, value = broad_count, c("head", "shoulder", "upper_torso",
```

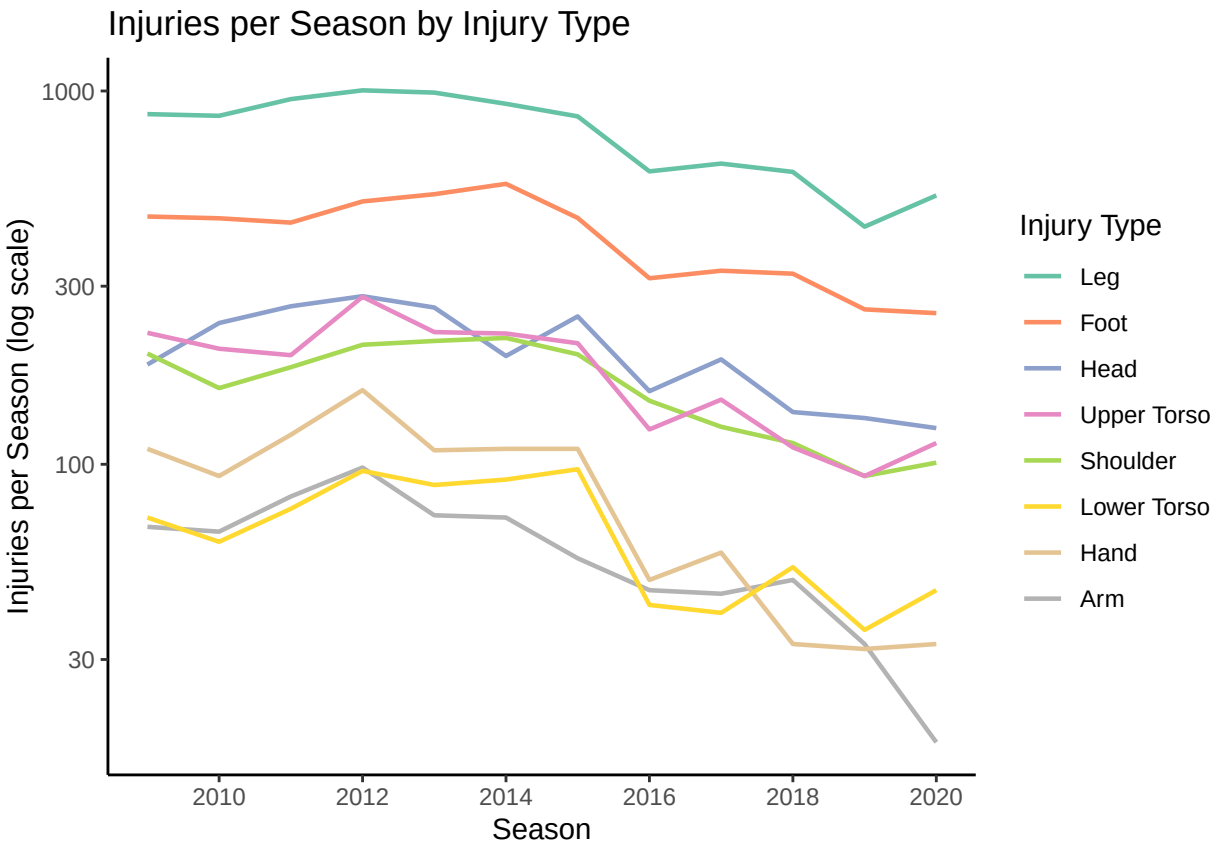
```
                                                    "lower_torso", "arm", "hand",
                                                    "leg", "foot"))

# create summary table
gathered <- gathered %>%
  group_by(year, broad_injury) %>%
  dplyr::summarise(broad_count = sum(broad_count))

# view first 6 rows of table
pander(head(gathered))
```

year	broad_injury	broad_count
2009	arm	68
2009	foot	461
2009	hand	110
2009	head	185
2009	leg	867
2009	lower_torso	72

```
# plot
gathered %>%
  ggplot(aes(x = year, y = broad_count, group = broad_injury,
             color = fct_reorder2(broad_injury, year, broad_count))) + # reorder legend
  geom_line(size = 0.8) +
  ggtitle("Injuries per Season by Injury Type") +
  xlab("Season") +
  ylab("Injuries per Season (log scale)") +
  scale_y_log10() +
  scale_x_continuous(breaks = seq(2008, 2020, by = 2)) +
  labs(color = "Injury Type") + # legend title
  scale_color_brewer(palette = "Set2",
                     labels = c("Leg", "Foot", "Head",
                                "Upper Torso", "Shoulder",
                                "Lower Torso", "Hand", "Arm")) + # nice legend names
  theme_classic()
```

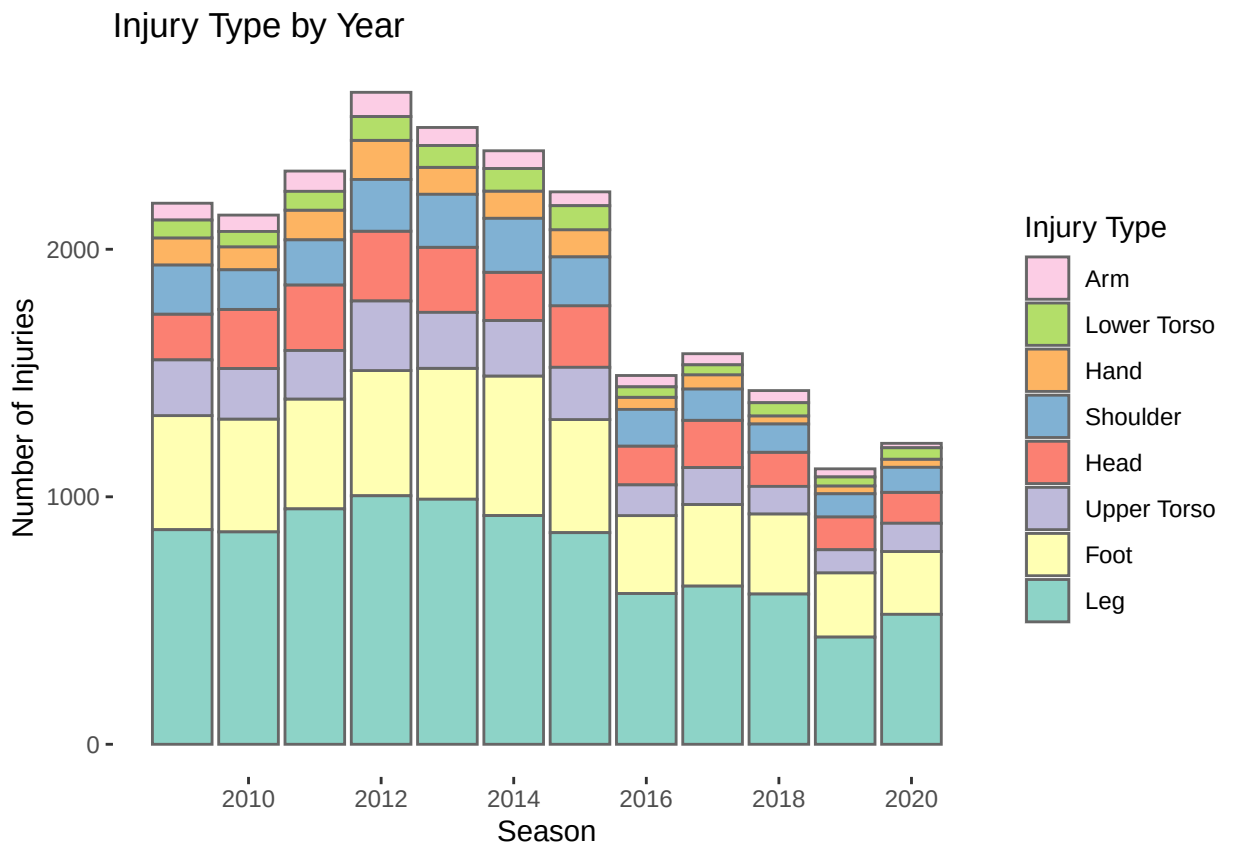


It looks like the prevalence of all injury types have generally declined over the time period. We can also see that leg and foot injuries were the most common across all seasons.

We can graph the same data in a bar plot:

```
gathered$broad_injury <- as.factor(gathered$broad_injury) # convert to factor

# stacked barplot
gathered %>%
  ggplot(aes(x = year, y = broad_count,
             fill = fct_reorder(broad_injury, broad_count))) + # reorder stacked bars
  geom_bar(stat = "identity", color = "gray40") +
  ggtitle("Injury Type by Year") +
  xlab("Season") +
  ylab("Number of Injuries") +
  labs(fill = "Injury Type") + # legend title
  scale_fill_brewer(palette = "Set3", direction = -1, # use palette in reverse order
                    labels = c("Arm", "Lower Torso", "Hand",
                               "Shoulder", "Head", "Upper Torso",
                               "Foot", "Leg")) + # fix legend colors and labels
  scale_x_continuous(breaks = seq(2010, 2020, by = 2)) +
  theme(panel.background = element_blank()) + # remove grid lines
  theme(legend.position = "right")
```



The breakdown by injury type is similar across all seasons. The declining trend in the number of injuries per season also appears consistent across all injury types. The sudden drop in injury counts beginning in 2016 may indicate, for example, a change in NFL policy or the way injuries are counted in the data.

Games missed and played injured per season

We can approximate injury severity with the average number of games missed per season due to injury, as well as the average number of games that occured while injured per season (which includes both games missed and games played while injured).

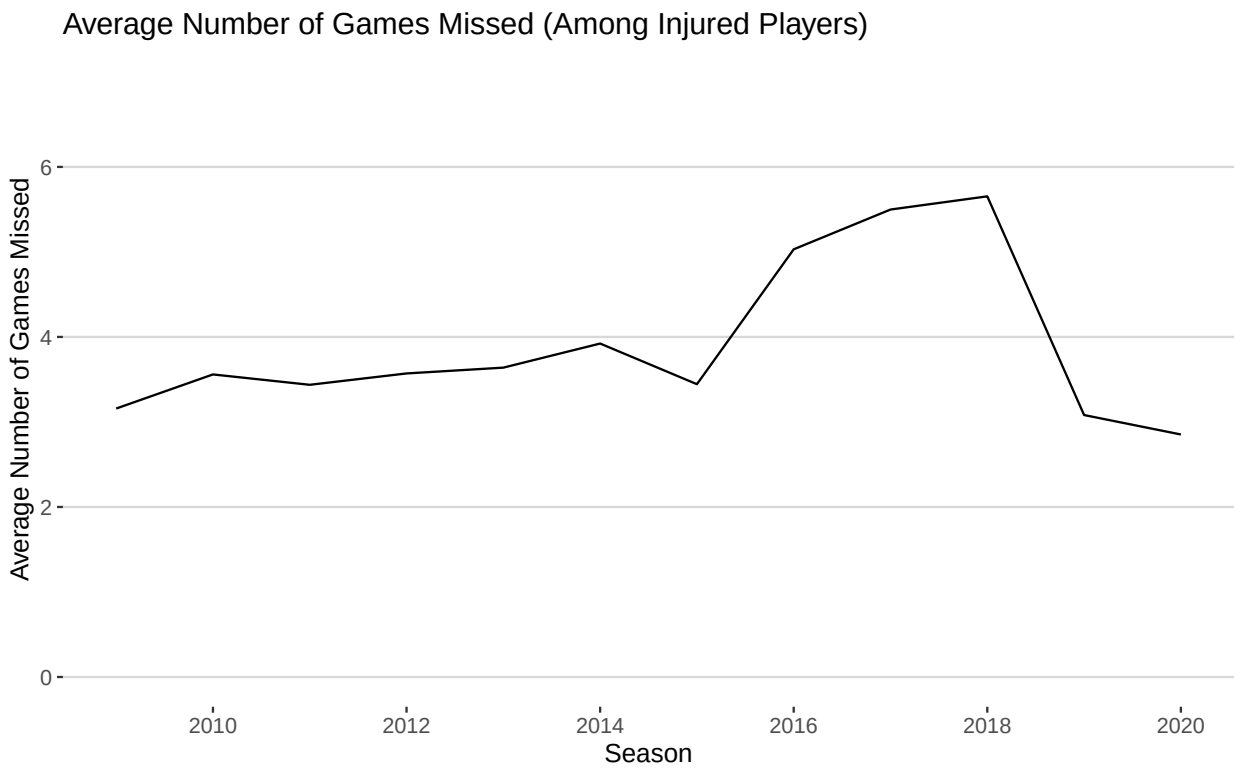
```
### games missed
# create summary table
games_missed <- injuries %>%
  group_by(year) %>%
  dplyr::summarize(avg_games_missed = mean(num_games_missing))

# view table
pander(games_missed)
```

year	avg_games_missed
2009	3.158
2010	3.559
2011	3.437
2012	3.571

year	avg_games_missed
2013	3.639
2014	3.922
2015	3.444
2016	5.031
2017	5.499
2018	5.654
2019	3.081
2020	2.853

```
# plot
games_missed %>%
  ggplot(aes(x = year, y = avg_games_missed)) +
  geom_line() +
  ylim(0, 7) +
  ggtitle("Average Number of Games Missed (Among Injured Players)") +
  xlab("Season") +
  ylab("Average Number of Games Missed") +
  scale_x_continuous(breaks = seq(2008, 2020, by = 2)) +
  theme_hc()
```

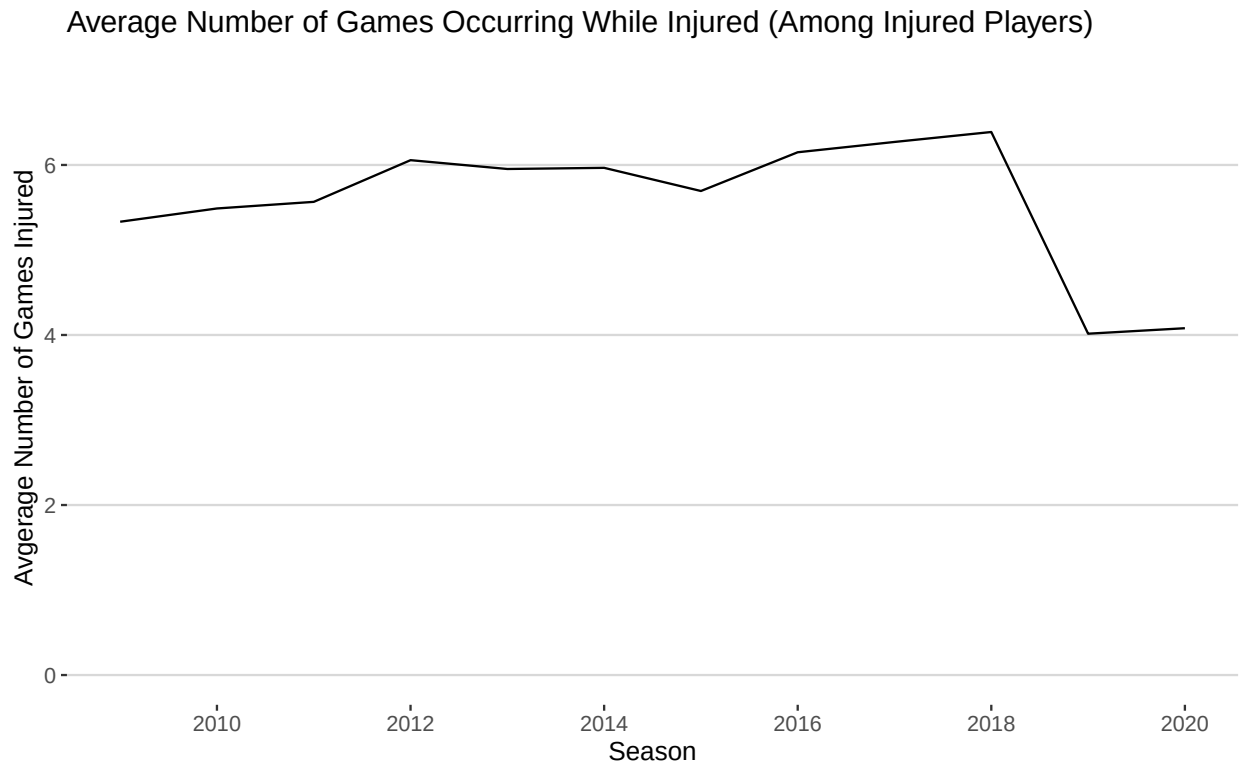



```
### games injured
# create summary table
games_injured <- injuries %>%
  group_by(year) %>%
  dplyr::summarize(avg_games_injured = mean(num_games_injured))
```

```r
# view table
pander(games_injured)
```

year	avg_games_injured
2009	5.332
2010	5.488
2011	5.566
2012	6.056
2013	5.952
2014	5.966
2015	5.694
2016	6.149
2017	6.271
2018	6.388
2019	4.015
2020	4.08

```r
# plot
games_injured %>%
  ggplot(aes(x = year, y = avg_games_injured)) +
  geom_line() +
  ylim(0, 7) +
  ggtitle("Average Number of Games Occurring While Injured (Among Injured Players)") +
  xlab("Season") +
  ylab("Avgerage Number of Games Injured") +
  scale_x_continuous(breaks = seq(2008, 2020, by = 2)) +
  theme_hc()
```



These last two plots show entirely different trends than the trends in the previous plots, highlighting the importance of exploring any data set extensively before drawing conclusions.

In the first plot, we see the there a spike in the number of games missed due to injury from 2016-2018, which corresponds with a sharp drop in raw injury counts in the other plots. It is possible that in these years players had fewer injuries than in prior years, but these injuries were more severe, leading to more games missed per injury. It is also possible that new rules forced injured players to sit out of games even when they wanted to play through their injuries.

The second plot depicts a more consistent trend in games occurring while injured with a steep drop-off in 2019. It is interesting to note that the trends in these two plots match reasonably well from 2016 to 2020, with the number of games that occur while injured only slightly higher than the number of games missed due to injury. Before 2016, it appears that there were many more games played while injured, because the number of games that occur while injured is much higher than the number of games missed. This may be due to changes in NFL rules that prevent injured players from injuring themselves further by continuing to play.