Lauren EDA

Lauren Mock

11/16/2021

```
library(tidyverse)
library(ggplot2)
```

Trends In Injuries Over Time

Read in and clean injury data

```
injuries <- read.csv("cleaneddata.csv", stringsAsFactors = FALSE)

# remove first two columns
injuries <- select(injuries, -c(X.1,X))

# add column for total injuries
injuries <- injuries %>%
  mutate(all_injuries = rowSums(injuries[,27:34]))

# filter out 2021 (partial data)
injuries <- injuries %>% filter(year != 2021)
```

Make a column for adjusted injuries

(adjusted by number of games in the season and number of players in this data set)

```
year_games <- injuries %>%
  distinct(year, games_in_season)
```

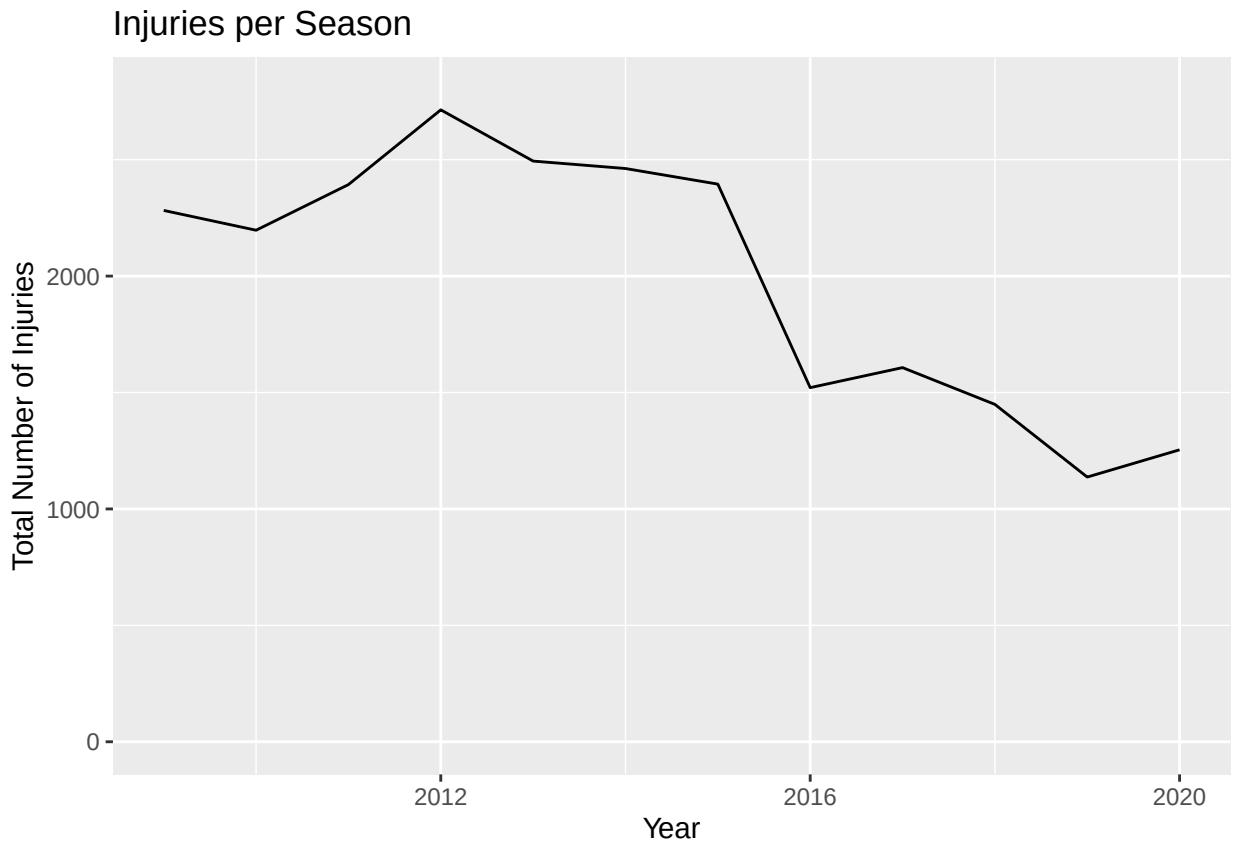
I'm confused!!!

also need to know–are years seasons?

Injuries per Season

```
time_series <- injuries %>%
  group_by(year) %>%
  summarize(all_injuries = sum(all_injuries))
```

```
time_series %>%
  ggplot(aes(x = year, y = all_injuries)) +
  geom_line() +
  ggtitle("Injuries per Season") +
  xlab("Year") +
  ylab("Total Number of Injuries") +
  ylim(0, 2800)
```

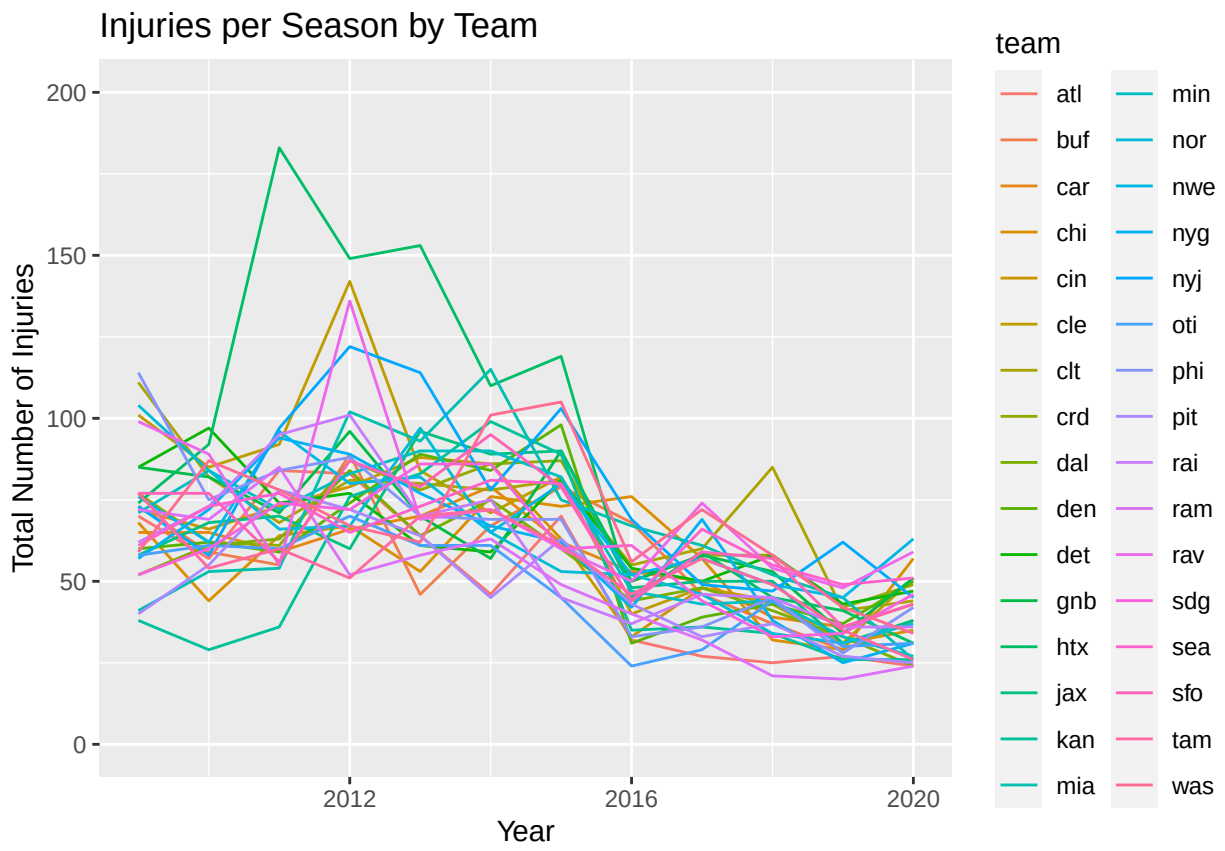


Injuries per Season by Team

```
time_series_teams <- injuries %>%
  group_by(year, team) %>%
  summarize(all_injuries = sum(all_injuries))
```

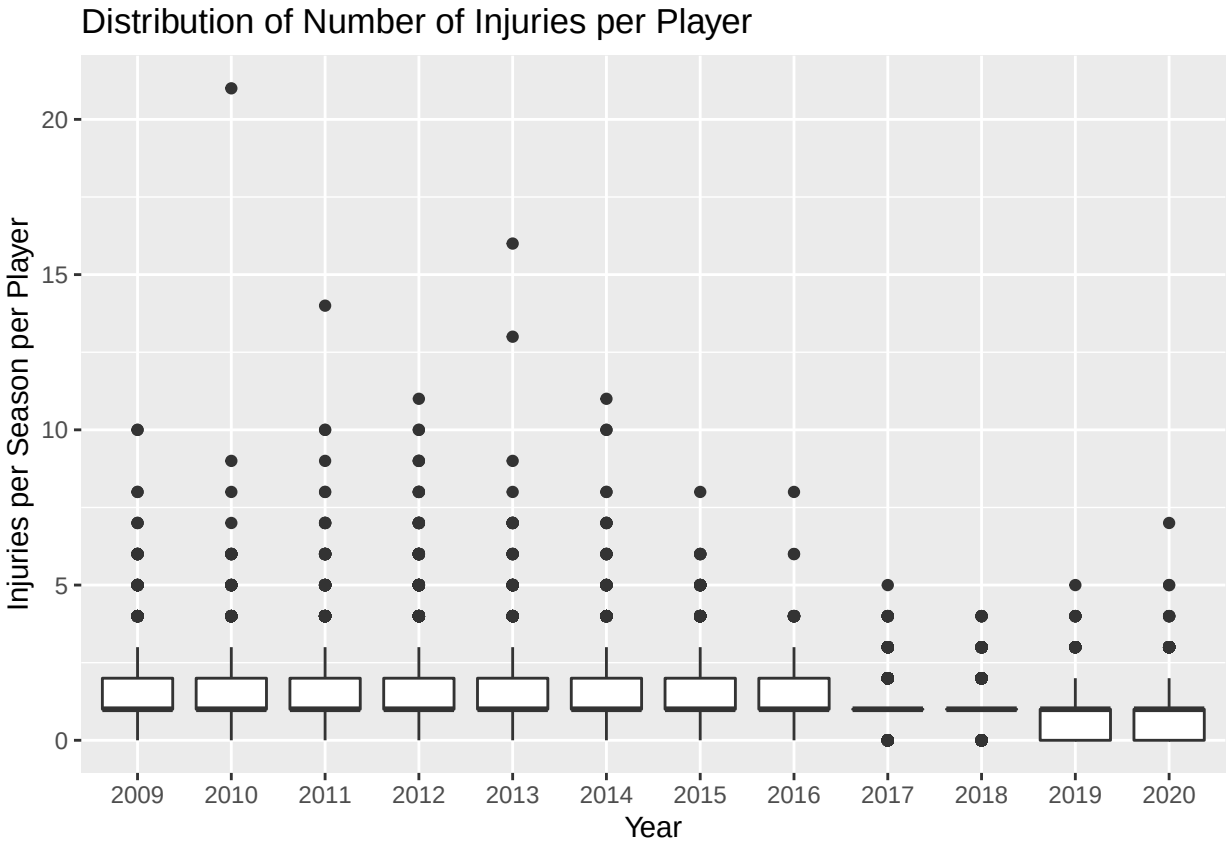
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
time_series_teams %>%
  ggplot(aes(x = year, y = all_injuries, color = team)) +
  geom_line() +
  ggtitle("Injuries per Season by Team") +
  xlab("Year") +
  ylab("Total Number of Injuries") +
  ylim(0, 200)
```



Boxplots

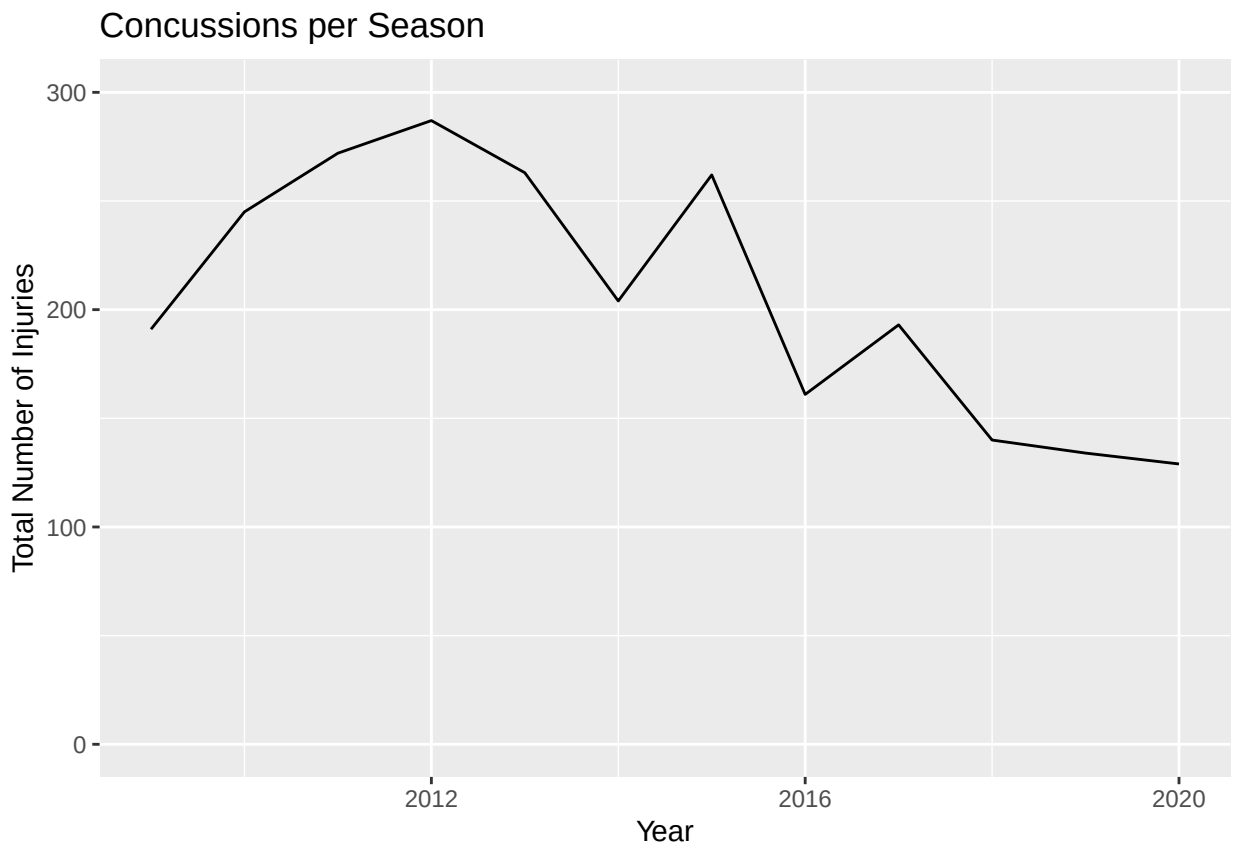
```
injuries %>%
  ggplot(aes(x = as.factor(year), y = all_injuries)) +
  geom_boxplot() +
  ggtitle("Distribution of Number of Injuries per Player") +
  xlab("Year") +
  ylab("Injuries per Season per Player")
```



Concussions per Season

```
concussions <- injuries %>%
  group_by(year) %>%
  summarize(head = sum(head))


concussions %>%
  ggplot(aes(x = year, y = head)) +
  geom_line() +
  ggtitle("Concussions per Season") +
  xlab("Year") +
  ylab("Total Number of Injuries") +
  ylim(0, 300)
```

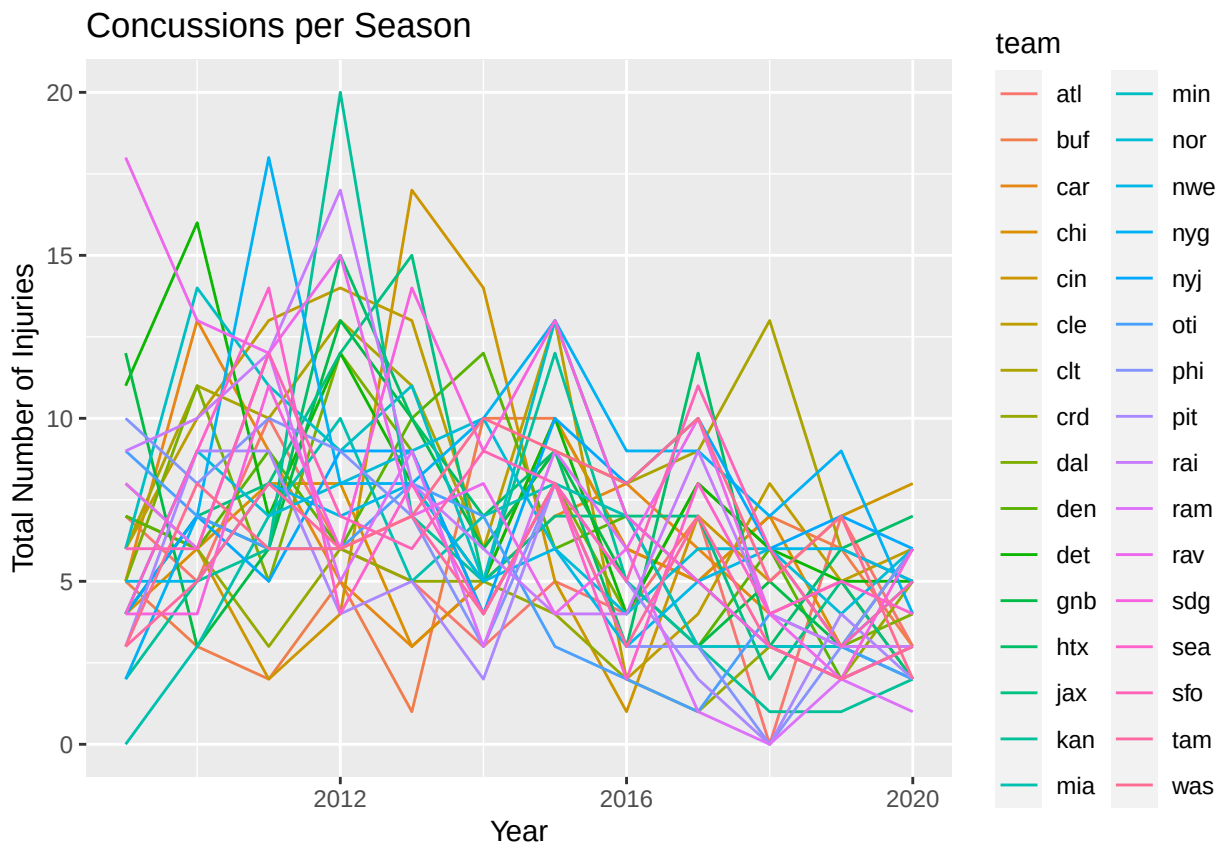


Concussions per Season by Team

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

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
concussions_team %>%
  ggplot(aes(x = year, y = head, color = team)) +
  geom_line() +
  ggtitle("Concussions per Season") +
  xlab("Year") +
  ylab("Total Number of Injuries") +
  ylim(0, 20)
```



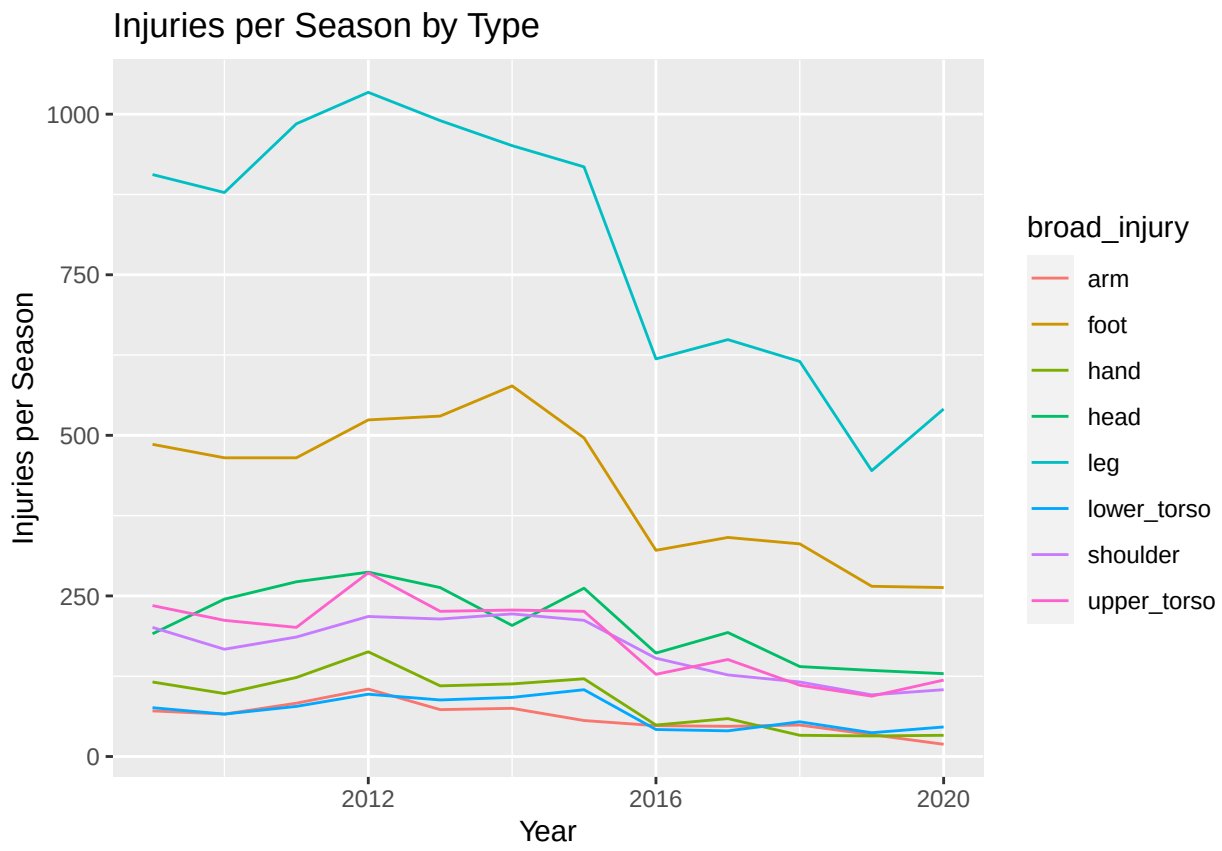
All Injury Types over Time

```
gathered <- injuries %>%
  gather(key = broad_injury, value = broad_count, 27:34)

gathered <- gathered %>%
  group_by(year, broad_injury) %>%
  summarise(broad_count = sum(broad_count))
```

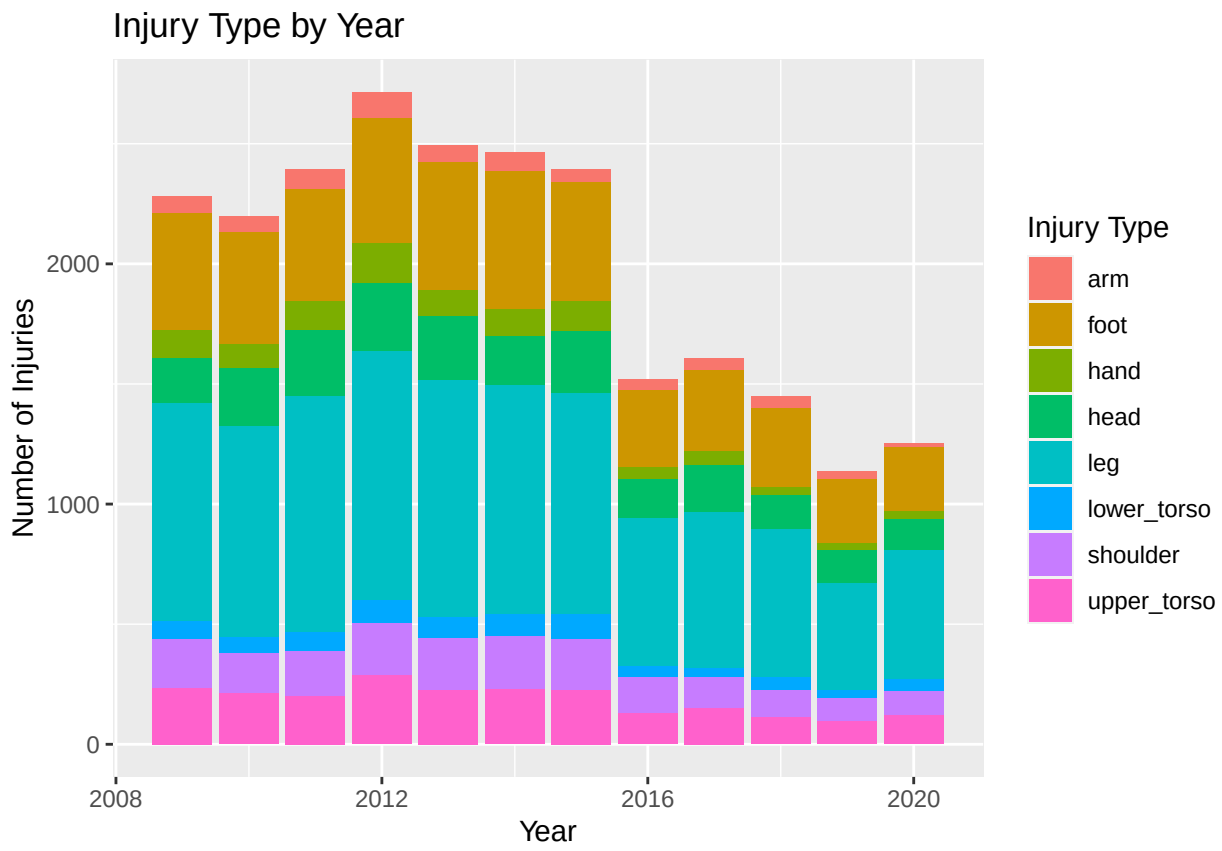
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
gathered %>%
  ggplot(aes(x = year, y = broad_count, color = broad_injury)) +
  geom_line() +
  ggtitle("Injuries per Season by Type") +
  xlab("Year") +
  ylab("Injuries per Season")
```

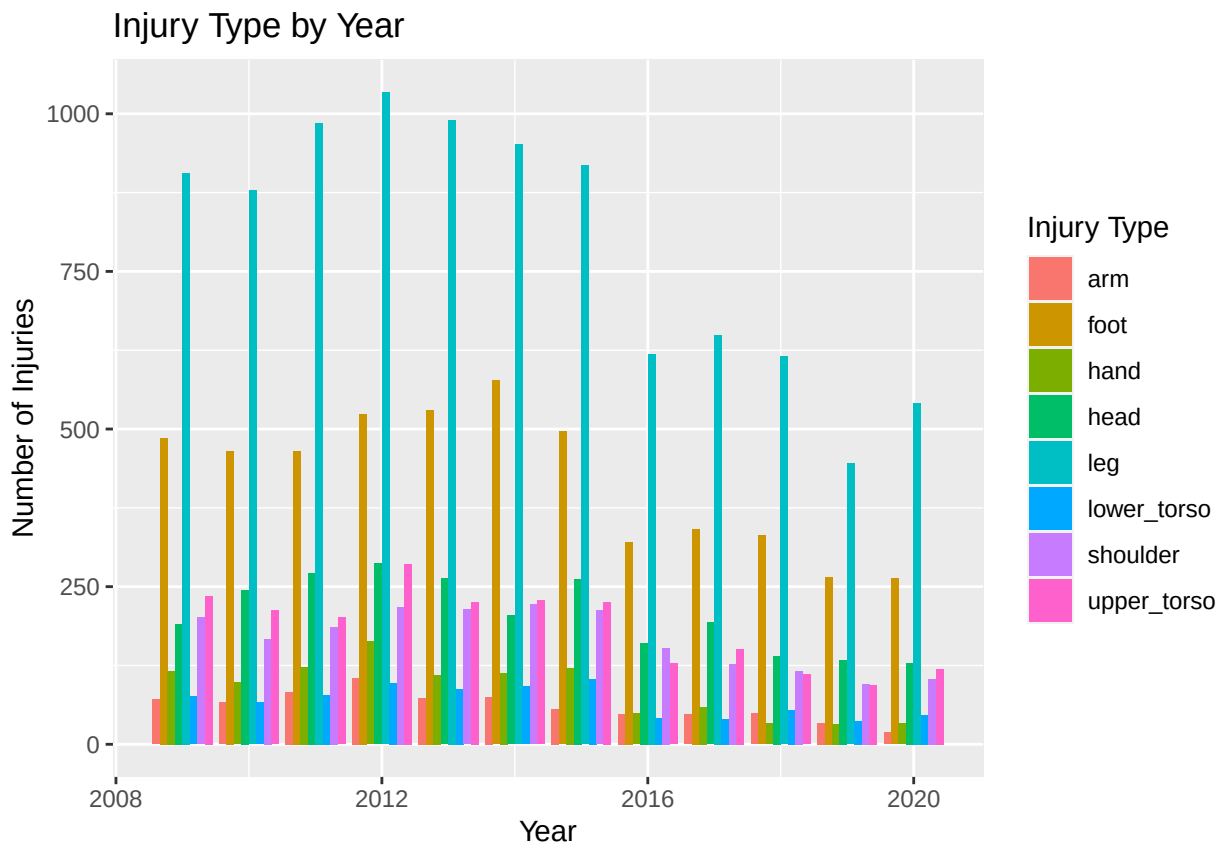


. . . as a barplot

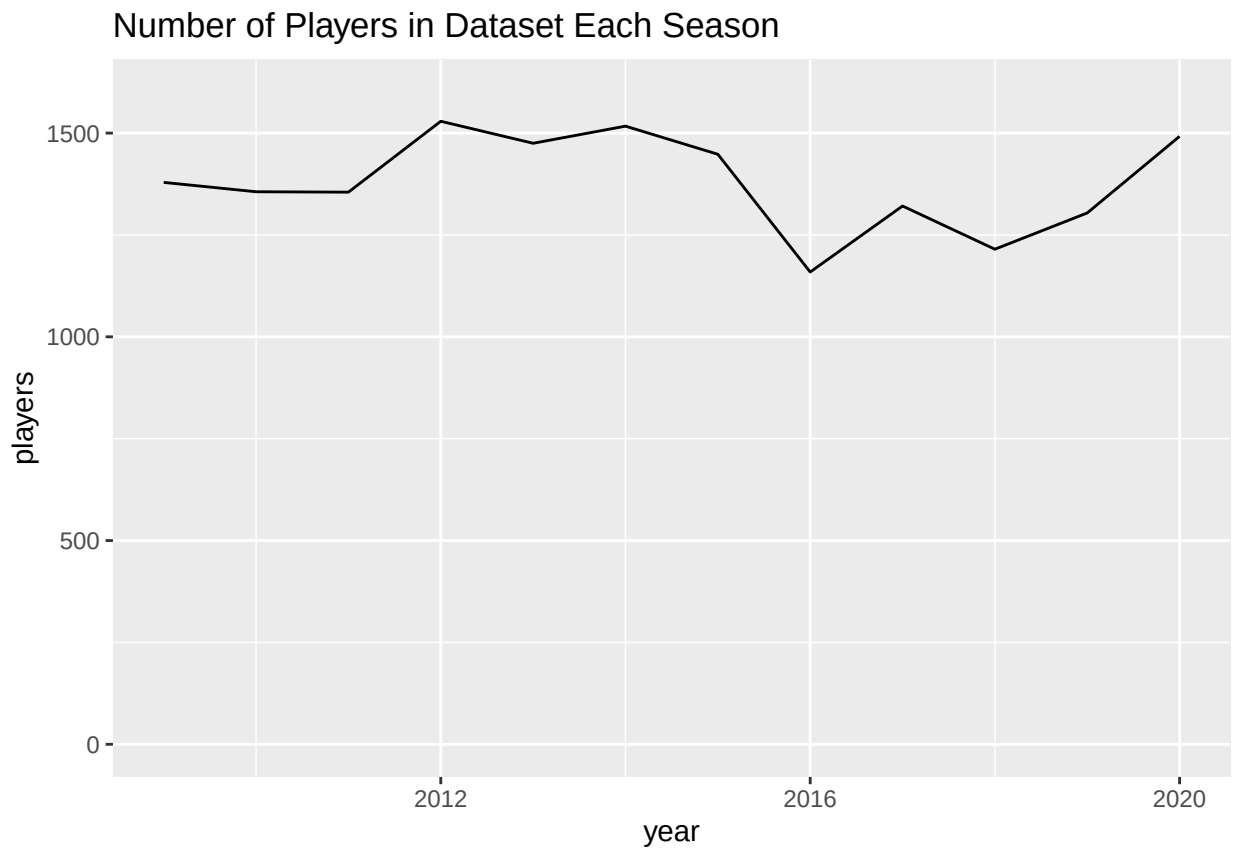
```
gathered %>%
  ggplot(aes(x = year, y = broad_count, fill = broad_injury)) +
  geom_bar(stat = "identity") +
  ggtitle("Injury Type by Year") +
  xlab("Year") +
  ylab("Number of Injuries") +
  labs(fill = "Injury Type") +
  scale_color_hue(labels = c("Arm", "Foot", "Hand", "Head", "Leg", "Lower Torso",
                             "Shoulder", "Upper Torso"))
```




```
gathered %>%
  ggplot(aes(x = year, y = broad_count, fill = broad_injury)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Injury Type by Year") +
  xlab("Year") +
  ylab("Number of Injuries") +
  labs(fill = "Injury Type") +
  scale_color_hue(labels = c("Arm", "Foot", "Hand", "Head", "Leg", "Lower Torso",
                             "Shoulder", "Upper Torso"))
```

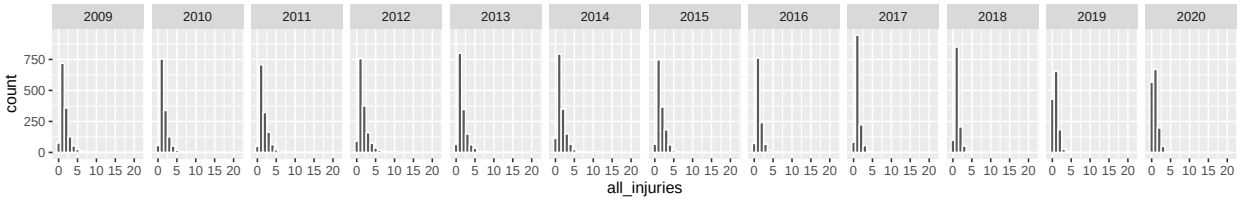



```
injuries %>%
  group_by(year) %>%
  summarize(players = n()) %>%
  ggplot(aes(x = year, y = players)) +
  geom_line() +
  ylim(0, 1600) +
  ggtitle("Number of Players in Dataset Each Season")
```



Can we use Poisson regression to predict # of injuries?

```
injuries %>%
  ggplot(aes(x = all_injuries)) +
  geom_histogram(binwidth = 1, col = "white") +
  facet_grid(. ~ year)
```



try Poisson regression

same number of players in the NFL every year? check as % of players tricky b/c we don't have everyone in this dataset - need to confirm that this isn't dependent on year, like we on

things to look at: - concussions by year - maybe look at other types of injury by year too - time series - avg. # of games missed by each player each season due to injuries - should adjust for # of games in season! - date of earliest injury

ARE YEARS SEASONS??