Baseball Data Visualization

Cody Zupnick, Max Mattioli, Priya Medberry, and Mohammad Radiyat

Introduction

Baseball, sometimes referred to as the Great American Pastime, is one of the most popular sports in America. There is an immense amount of baseball data freely available for analysis and is effectively a statistician's playground.

We chose the topic of baseball for our final project for a few reasons. One of our group members is involved in fantasy baseball, and we thought that exploring baseball data in-depth would probably help him in future fantasy bracket player selections. We also hoped that this analysis, or a similar analysis, would be potentially applicable to other sports including but not limited to basketball and American football. Lastly we wanted to get insight into a core element of American pop culture, and look into some interesting questions such as injuries and home runs trends.

Overview

For those who are less familiar with the sport, we've included a quick overview of the game. Baseball is a two-team competitive sport. There are 9 innings (or rounds) in total, and during each inning both teams play offense and defense, alternating until the game is over.

On defense, the pitcher tries to throw the ball into the strike zone and make the batter miss. If the batter misses the ball three times when the ball is in the strike zone, the batter gets a strike. If the batter gets three strikes, they are out. If the batter's team gets three outs, the team roles reverse. One cycle of offense-defenses is one inning.

On offense, the batter starts at home plate and tries to hit the ball into the outfield. If successful, they run towards first base. Depending on how they hit the ball they may go to second base, third base, or back to home plate. If they are tagged by the ball while not on a base they are out. Once they reach a base and stop moving, the next batter goes an the cycle continues.

Questions

We had a lot of questions that we thought of, and have listed these below:

- Are there any obvious trends between pitching volume and effort in a given year and injuries in the next year?
- How severe are different kinds of injuries to a baseball player's career?
- Are pitchers who throw different kinds of pitches more susceptible to getting injured? Do certain kinds of pitches lead to different kinds of injuries?
- How do different kinds of injuries affect hitting power?
- How does likelihood of future injury change with different kinds of injuries?
- How did a ball manufacturing change affect number of home runs?
- How may a robot umpire affect the strike zone?

Our group members are:

- Cody Zupnick - data collection, plotting, interactive, and report writing
- Mohammad Radiyat - plotting, interactive and report writing
- Priya Medberry - plotting and report writing
- Max Mattioli - data quality analysis and report writing

Description of data:

The data was collected from a few different sources. Pitcher data was collected from the following link. The batter data was collected from here.

The injury data was the most difficult to collect. Injury data was obtained via scraping, and the scripts to do so can be found at this Github link.

We look at a quick summary of the raw data files that we'll be using. First we begin with pitch data:

```r
library(dplyr)
library(tidyverse)
library(ggplot2)
library(zoo)
library(ggthemes)
pitches_2015 <- read_csv('./data/Pitchers2015.csv')
head(pitches_2015)
```

```
## # A tibble: 6 x 16
##   name     Team     Age Pitches `FA%`  `FT%` `FC%` `FS%` `SI%` `SL%` `CU%`
##   <chr>    <chr>  <int>   <int> <chr>  <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Dalier ~ - - -     29     429 20.20% 43.8~ 1.20% <NA>  <NA>  22.8~ <NA>
## 2 Jeff Ma~ India~    30     550 2.20%  54.2~ <NA>  <NA>  <NA>  <NA>  42.4~
## 3 Wade Da~ Royals    29    1050 53.80% 1.00% 26.5~ <NA>  <NA>  <NA>  <NA>
## 4 Carter ~ Marli~    24     445 64.60% <NA>  <NA>  <NA>  <NA>  13.5~ <NA>
## 5 Matt Al~ White~    32     566 0.90%  <NA>  <NA>  <NA>  68.0~ 12.0~ 2.00%
## 6 Jason F~ - - -     37     543 61.50% 0.20% <NA>  <NA>  <NA>  18.5~ <NA>
## # ... with 5 more variables: `KC%` <chr>, `CH%` <chr>, vFA <dbl>,
## #   IP <dbl>, playerid <int>
```

The first few columns are player name, what team the player is on, player age as of 2015, and total number of pitches in the 2015 season. The next columns (all of which end in %) represent how frequently the player threw a certain kind of pitch. For example, Dalier Hinojosa threw 429 pitches in 2015, 20.20% of which were *FA* (which stands for normal fastballs) and 43.80% of which were *FT* (which stands for two-seam fastballs). The different kinds of pitches and their code are outlined below:

- FA = normal (four-seam) fastball that goes straight and fast
- FT = two-seam fastball, which is similar to a normal fastball but has a little more movement
- FC = cutter, which is designed to shift a little bit when it gets close to home plate (or near where the batter is standing)
- FS = splitter, which breaks down suddenly near home plate
- SI = sinker, which has significant horizintal and downward movement
- SL = slider, which breaks down and away near home plate
- CU = curveball, where the ball has enough forward spin to dive downwards as it reaches home plate.
- KC = knuckle curve, which is a variant of the curveball that spins less than a normal curveball
- CH = changeup, which is a slow throw that looks like it's a normal fastball; used to throw off the batter's swing timing

The last few columns are *vFA* which is average fastball speed, and *IP* which is number of innings pitched. Note that innings pitched are not integers because pitchers can substitute within innings.

Next we can look at an example of the injury data:

```r
injuries_2016 <- read_csv('./data/injuries_2016.csv')
head(injuries_2016)
```

```
## # A tibble: 6 x 4
##   name               pos   type            days
##   <chr>              <chr> <chr>          <int>
## 1 Francisco Cervelli C     Wrist/Hand        48
## 2 Carlos Carrasco    SP    Hand/Hamstring    54
## 3 Starling Marte     CF    Back             566
## 4 Wilson Ramos       C     Knee ACL         567
## 5 Mac Williamson     RF    Quad/Shoulder    599
## 6 Josh Harrison      2B    Groin            568
```

The data includes the player, their position, kind of injury, and how long that injury lasted in days. Row examples of raw batter data is shown below:

```
df2014 <- read_csv('./data/batters2014.csv')
head(df2014)
```

```
## # A tibble: 6 x 22
##   name   Team      G    PA    HR     R   RBI    SB `BB%` `K%`    ISO BABIP
##   <chr>  <chr> <int> <int> <int> <int> <int> <int> <chr> <chr> <dbl> <dbl>
## 1 Mike ~ Ange~   157   705    36   115   111    16 11.8~ 26.1~ 0.274 0.349
## 2 Andre~ Pira~   146   648    25    89    83    18 13.0~ 17.7~ 0.228 0.355
## 3 Gianc~ Marl~   145   638    37    89   105    13 14.7~ 26.6~ 0.267 0.353
## 4 Micha~ Indi~   156   676    20    94    97    23 7.70% 8.30% 0.178 0.333
## 5 Antho~ Nati~   153   683    21   111    83    17 8.50% 15.2~ 0.186 0.314
## 6 Jonat~ Brew~   153   655    13    73    69     4 10.1~ 10.8~ 0.164 0.324
## # ... with 10 more variables: AVG <dbl>, OBP <dbl>, SLG <dbl>, wOBA <dbl>,
## #   `wRC+` <int>, BsR <dbl>, Off <dbl>, Def <dbl>, WAR <dbl>,
## #   playerid <int>
```

This data includes player name, team, and several yearly metrics. For more info on these metrics, feel free to read this documentation. The important quantity for us is *SLG*, or slugging percentage which is average number of bases a player earns per at-bat. Each time a player bats, he either hits or does not hit; on a hit, a player can get 1, 2, 3, or 4 bases (4 in the case of a home run). *SLG* captures total number of bases earned this way divided by total number of at-bats.

Next, we look at aggregate seasonal data:

```
homeruns <- read_csv('./data/homeRuns.csv')
head(homeruns)
```

```
## # A tibble: 6 x 50
##   Season League Month        G    PA    AB     H  `1B`  `2B`  `3B`    HR
##    <int> <chr>  <chr>    <int> <int> <int> <int> <int> <int> <int> <int>
## 1   2012 MLB    Mar/Apr   7033 25433 22753  5677  3754  1154   136   633
## 2   2012 MLB    May       8951 32385 28942  7377  4836  1514   168   859
## 3   2012 MLB    Jun       8455 30589 27517  7068  4705  1366   147   850
## 4   2012 MLB    Jul       8067 29028 26049  6668  4406  1306   145   811
## 5   2012 MLB    Aug       8816 32036 28884  7421  4946  1462   152   861
## 6   2012 MLB    Sept/Oct 10091 34564 30968  7827  5279  1454   179   915
## # ... with 39 more variables: R <int>, RBI <int>, BB <int>, IBB <int>,
## #   SO <int>, HBP <int>, SF <int>, SH <int>, GDP <int>, SB <int>,
## #   CS <int>, AVG <dbl>, `BB%` <dbl>, `K%` <dbl>, `BB/K` <dbl>,
## #   AVG1 <dbl>, OBP <dbl>, SLG <dbl>, OPS <dbl>, ISO <dbl>, BABIP <dbl>,
## #   `w RC` <dbl>, `w RAA` <dbl>, `w OBA` <dbl>, `wRC+` <dbl>,
## #   `GB/FB` <dbl>, `LD%` <dbl>, `GB%` <dbl>, `FB%` <dbl>, `IFFB%` <dbl>,
## #   `HR/FB` <dbl>, `IFH%` <dbl>, `BUH%` <dbl>, `Pull%` <dbl>,
## #   `Cent%` <dbl>, `Oppo%` <dbl>, `Soft%` <dbl>, `Med%` <dbl>,
## #   `Hard%` <dbl>
```

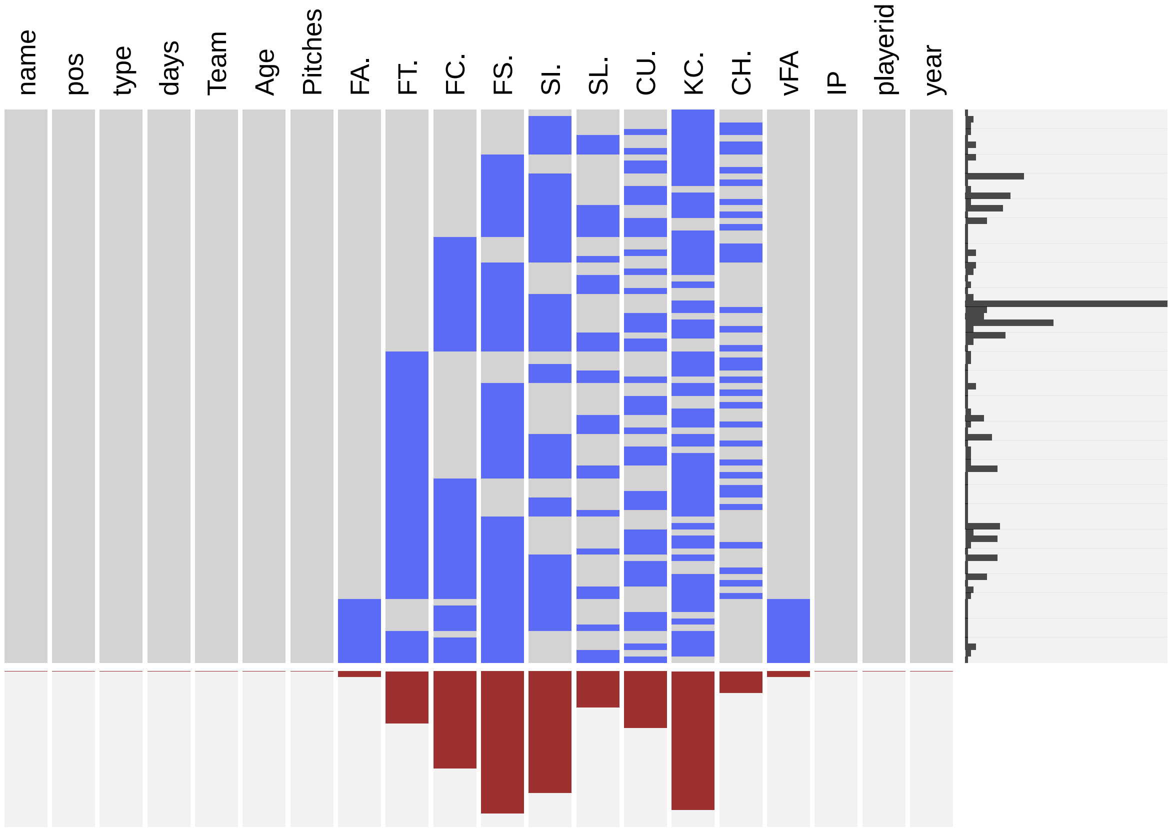
There are plenty of monthly stats here, and each column represents a different aggregate stat. Of use for us later is *HR* which represents aggregate home runs. Info about what the other stats are can also be found on the Baseball Almanac link above.

Analysis of Data Quality

Fortunately there is near complete public availability of the data we're exploring, and our scraping service ran without error for all the necessary data. There are no missing data in any of the injuries, batters, or homeruns sets, and the NA values in the Pitchers data all fall under the "pitch frequency" columns.

While these NA values represent 0% frequency of that type of pitch (instead of a true NA), it is still worthwhile to visualize their occurences across pitch-types to further understand pitcher preferences.

```r
library(extracat)
pitches_2015 <- read_csv('./data/Pitchers2015.csv')
injuries_2016 <- read_csv('./data/injuries_2016.csv')

joined1 <- merge(injuries_2016, pitches_2015, by.x = "name", by.y = "name", all.y = T)
joined1$year <- '2016'

pitches_2016 <- read_csv('./data/Pitchers2016.csv')
injuries_2017 <- read_csv('./data/injuries_2017.csv')

joined2 <- merge(injuries_2017, pitches_2016, by.x = "name", by.y = "name", all.y = T)
joined2$year <- '2017'

joinedQA <- rbind(joined1, joined2)
joined_pitchers <- filter(joinedQA, pos == "RP" | pos == "SP")

visna(joined_pitchers)
```

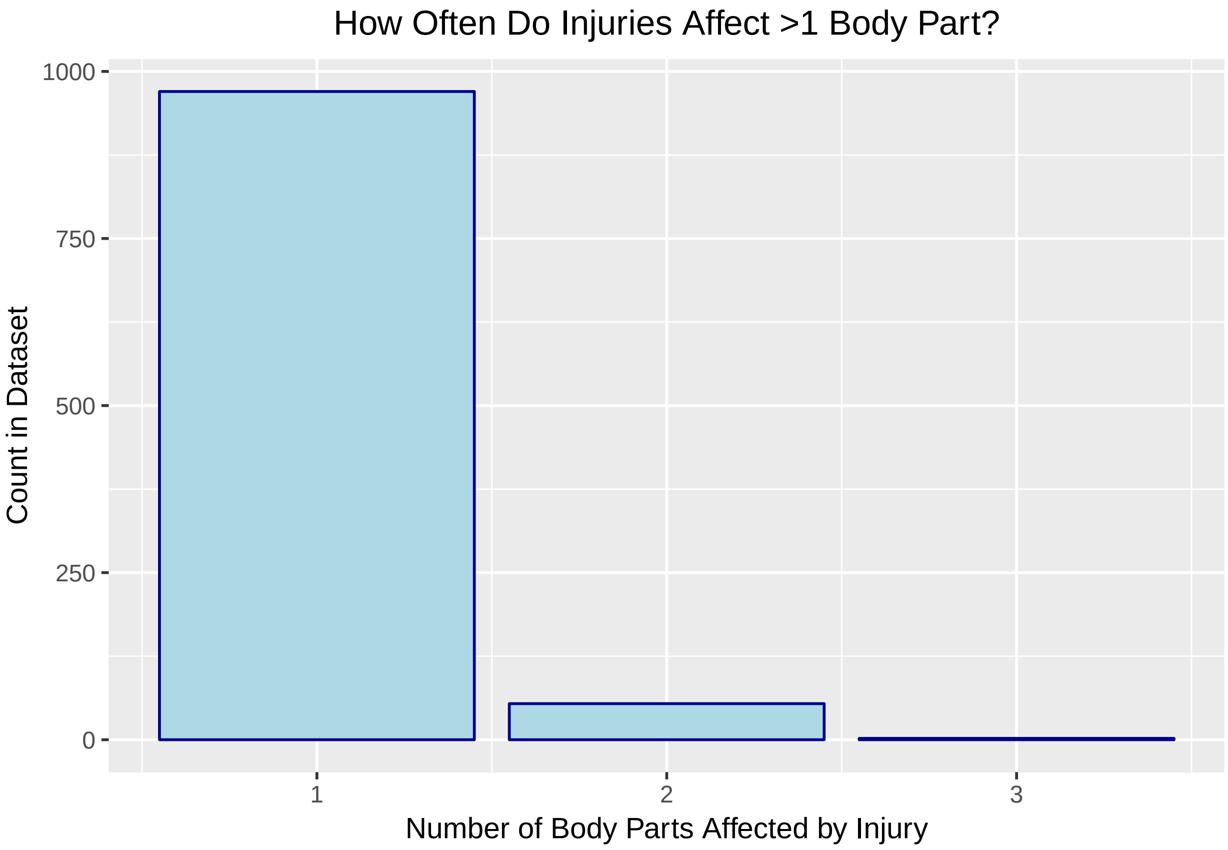


This graphic shows how different pitchers exhibit patterns of pitch-type specializations. Knuckle-curves (KC), splitters (FS), and sinkers (SI) are the most-avoided pitches in our set. This falls in line with the notion that those are some of the most difficult and physically taxing pitch-types. It is worth keeping this in mind for later analysis of injury by pitch type; Knuckle-curves, splitters, and sinkers have imbalanced representation in this dataset.

We can also see evidence of specialization for each row of data (pitcher), in that most pitchers have a few types that they will never throw. This makes sense because teams benefit from having "pointy"-skilled pitchers, as opposed to many jacks of all trade/masters of none.

The other data quality issue to address is one brought on by our preprocessing methods. Some injuries sustained by a player affect more than one part of the body, and in these cases we chose only one of the listed body parts to simplify our analysis. In an effort to "check ourselves", we use a bar plot to show the frequencies of injuries that affect 1, 2, and 3 parts of the body in our dataset.

```
extract_injury_count <- function(val) { length(strsplit(val, "/")[[1]]) }
joinedQA$num_injuries <-sapply(joinedQA$type, extract_injury_count)
ggplot(joinedQA, aes(x=num_injuries)) + geom_bar(color="darkblue", fill="lightblue") + xlab("Number of I
```



Most injuries only affect 1 part of the body, which means not too much information is cut out from the injuries with 2+ affected parts. Our preprocessing method of choosing the most-severe injury when there is 2 or 3 listed will not lead to misleading results.

Main Analysis

We start by importing the necessary libraries for data processing and plotting:

```
# Install all necessary packages
library(dplyr)
library(tidyverse)
library(ggplot2)
library(zoo)
```

Part 1: Pitching Volume, Effort, and Injuries

First we look at whether there exist any obvious trends can be seen between pitching volume in a given year and injuries in the following year. Our metrics of volume are seasonal pitch counts, pitches per game, and innings pitched per season. Our basic metric of effort is fastball velocity. All of these are imperfect metrics; a stronger pitcher can throw harder with less effort than a weaker one, for instance, and not all pitches are equally strenuous. But this is a good starting point.

For now we are not considering age, which we expect to be very important, since older players should get injured more frequently than younger ones.

We import the data into R objects:

```
pitches_2015 <- read_csv('./data/Pitchers2015.csv')
injuries_2016 <- read_csv('./data/injuries_2016.csv')

pitches_2016 <- read_csv('./data/Pitchers2016.csv')
injuries_2017 <- read_csv('./data/injuries_2017.csv')
```

We combine the injury and pitch data frames together:

```
# join each year before concatting them
joined1 <- merge(injuries_2016, pitches_2015, by.x = "name", by.y = "name", all.y = T)
joined1$year <- '2016'

joined2 <- merge(injuries_2017, pitches_2016, by.x = "name", by.y = "name", all.y = T)
joined2$year <- '2017'

joined <- rbind(joined1, joined2)

joined_pitchers <- filter(joined, pos == "RP" | pos == "SP")
```

Some players are listed with multiple injuries, encoded as Injury1/Injury2/... In those cases the first injury is the most serious, so we will only consider the first listed injury per player for this analysis:

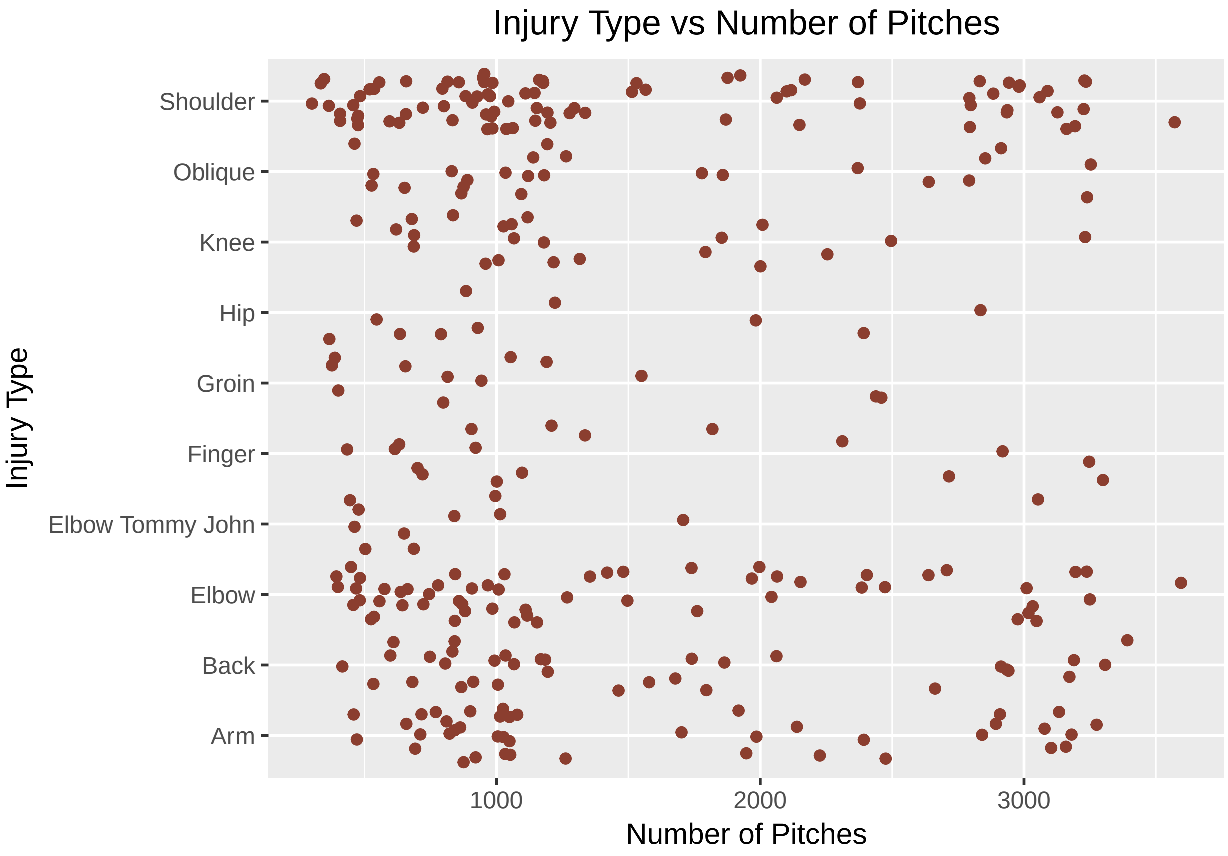
```
extract_injury <- function(val) { strsplit(val, "/")[[1]][[1]] }
```

Lastly we clean up the data, remove missing values and pitchers who never threw a pitch, and consider only injuries that at least ten players suffered. This makes the analysis cleaner, but inherently throws away information about rare injuries which might be valuable at some future point.

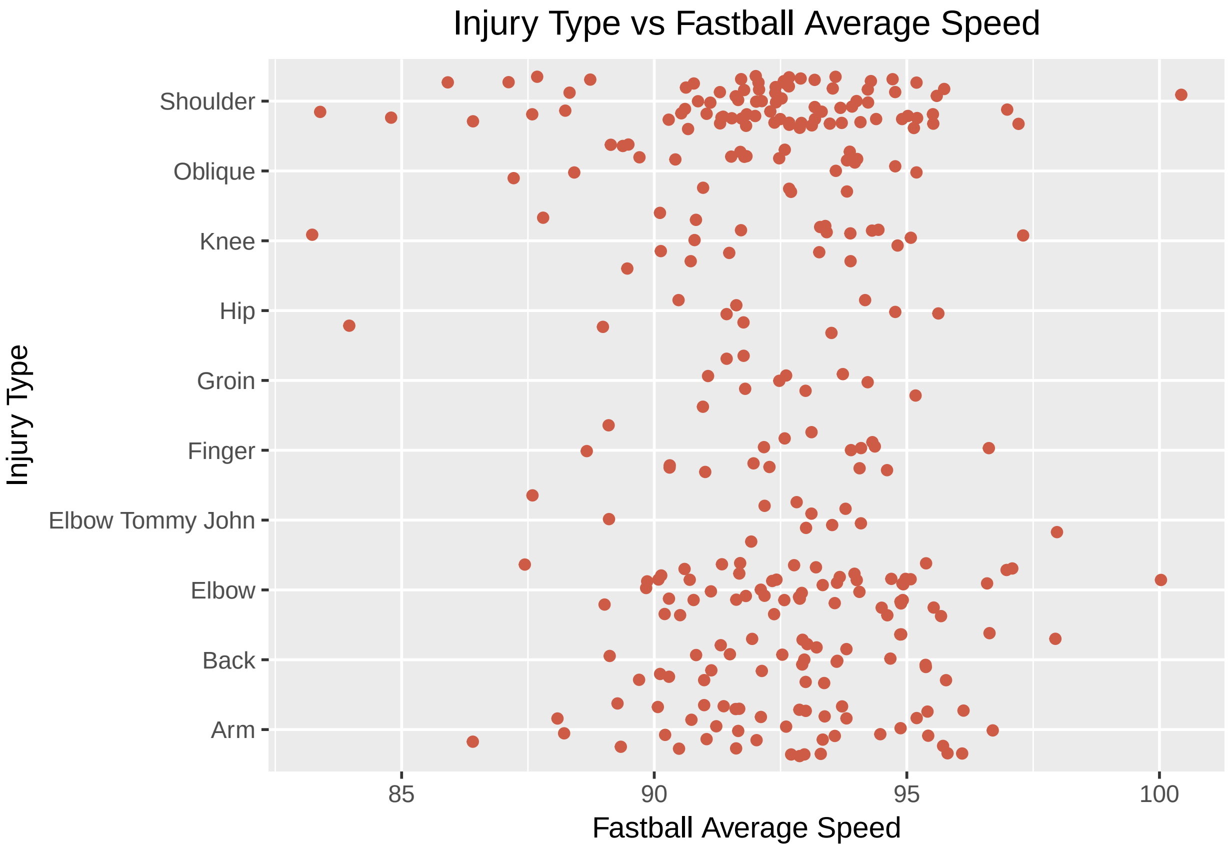
```
# @param df pitch dataframe joined with injury data
clean <- function(df) {
  percent_cols <- colnames(df)[8:16]
  f <- function(val) { sub("%", "", val) }
  df[percent_cols] <- sapply(df[percent_cols], f)
  df[percent_cols] <- sapply(df[percent_cols], as.numeric)

  # if a pitcher never throws a pitch
  df[percent_cols][is.na(df[percent_cols])] <- 0
  df$type[is.na(df$type)] <- "None"
  df$days[is.na(df$days)] <- 0

  df$type <- sapply(df$type, extract_injury)

  df
}

joined <- clean(joined)
joined_pitchers <- clean(joined_pitchers)
joined_pitchers <- mutate(group_by(joined_pitchers, type), cnt = n())
joined_pitchers <- filter(joined_pitchers, cnt >= 10)
```

Now we plot number of pitches in the previous season versus injury type and average fastball throw speed versus injury type:

```
ggplot(joined_pitchers, aes(x=type, y=Pitches)) + geom_jitter(color="coral4") + coord_flip() +
  ylab("Number of Pitches") + xlab("Injury Type") + ggtitle("Injury Type vs Number of Pitches") +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
ggplot(joined_pitchers, aes(x=type, y=vFA)) + geom_jitter(color="coral3") + coord_flip() +
  ylab("Fastball Average Speed") + xlab("Injury Type") + ggtitle("Injury Type vs Fastball Average Speed
  theme(plot.title = element_text(hjust = 0.5))
```

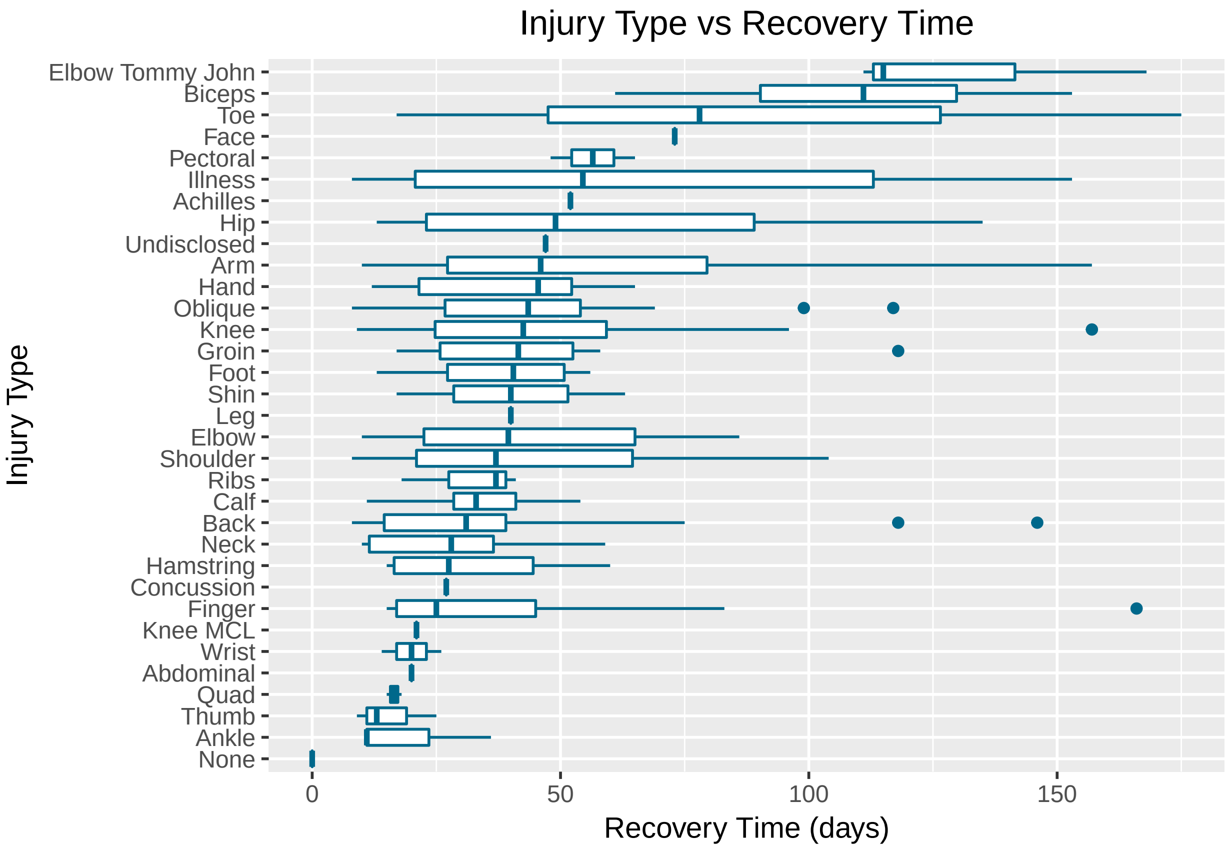


Not much jumps out in the scatterplots. It is conventional wisdom that high pitch total pitch volumes lead to injuries, but this expected result is either not visible over the course of one season or it's being swamped by other factors. There does seem to be a little bit of clustering of shoulder injuries at high pitch volume (around 3000 pitches in one season), implying that players who throw more pitches are more susceptible to shoulder injuries.

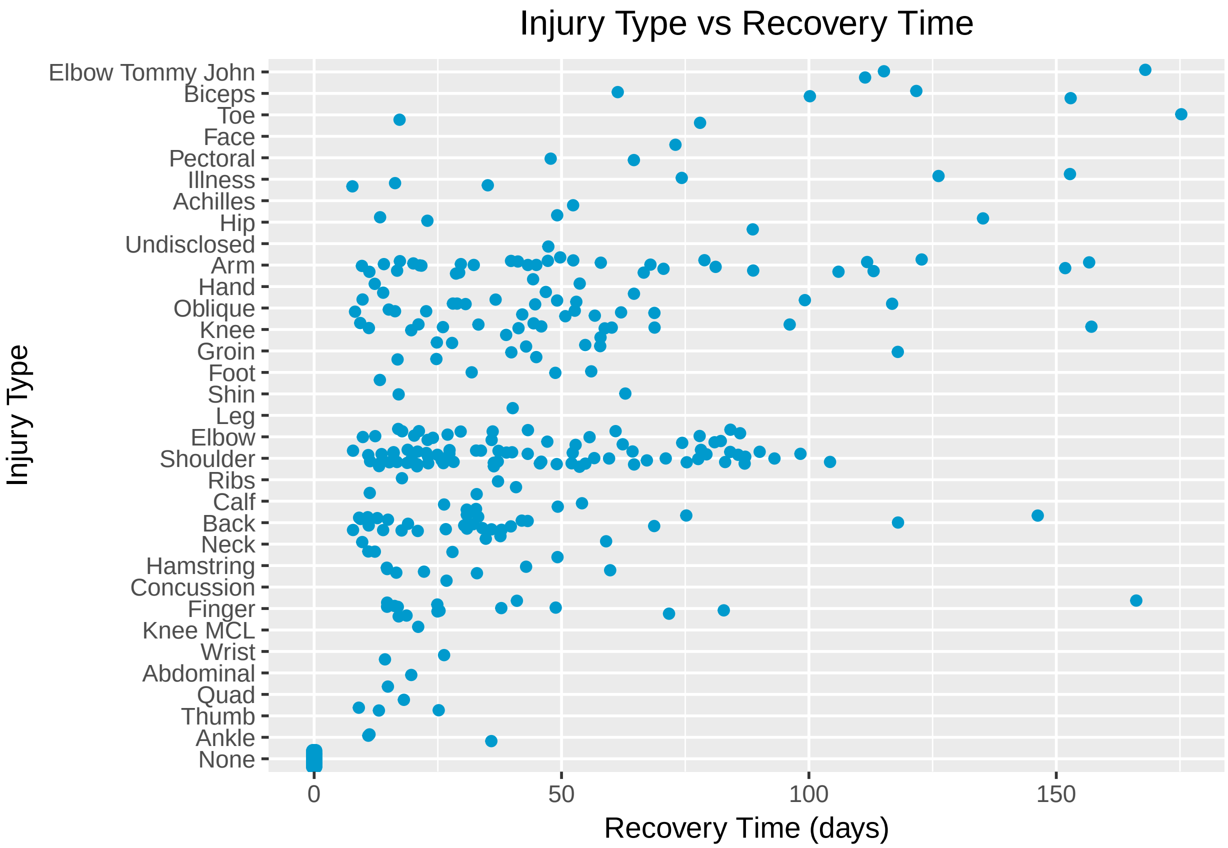
Injury Type

As a measure of injury severity, we look at how many days each injury lasted in the data for 2016 and 2017. We use both boxplots and scatterplots.

```
ggplot(filter(joined, days < 180), aes(x=reorder(type, days, FUN = median), y=days)) +
  geom_boxplot(color="deepskyblue4") + coord_flip() + ylab("Recovery Time (days)") + xlab("Injury Type"
  ggtitle("Injury Type vs Recovery Time") + theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(filter(joined, days < 180), aes(x=reorder(type, days, FUN = median), y=days)) +
  geom_jitter(color="deepskyblue3") + coord_flip() + ylab("Recovery Time (days)") + xlab("Injury Type")
  ggtitle("Injury Type vs Recovery Time") + theme(plot.title = element_text(hjust = 0.5))
```



By using median days values, the most severe injuries seem to be elbow Tommy John injuries, bicep injuries, and toe injuries. Elbow Tommy John injuries being the highest on this list makes sense. This kind of injury refers to a special elbow injury that involves a torn UCL, which is an extremely important elbow ligament.

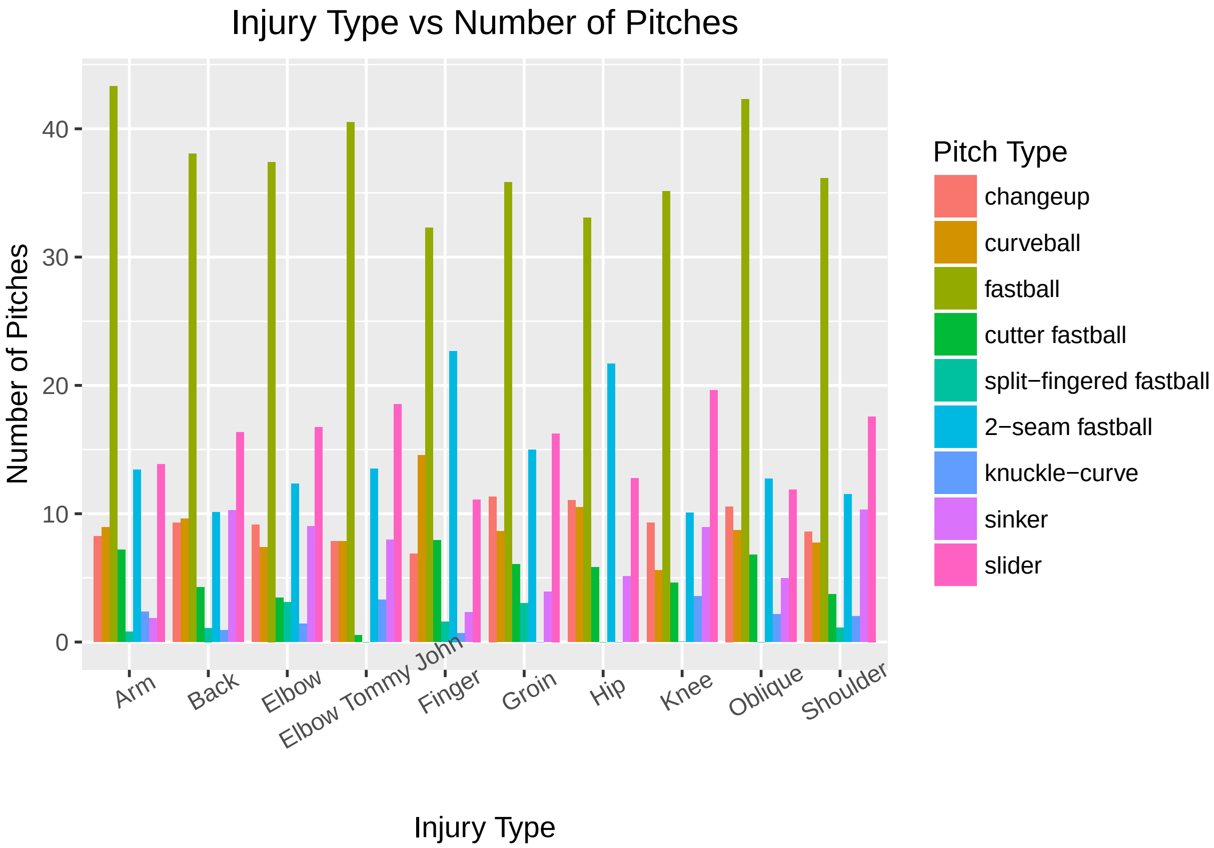
The median injury length values may be important, but arguably more important is frequency of injuries. The scatterplot indicates that injuries pitching injuries (shoulder, elbow, and arm) are among the most frequent.

Types of Pitches And Injuries

Since it seems that pitching injuries are quite frequent, it may be worth exploring pitching injuries in-depth. We look at whether players who throw certain types of pitches are more susceptible to getting injured, and whether certain kinds of pitches lead to specific types on injuries. This has been a holy grail of baseball research and is still hotly debated. While it is unlikely that anything will be settled without modeling, some trends may be visible using exploratory analysis.

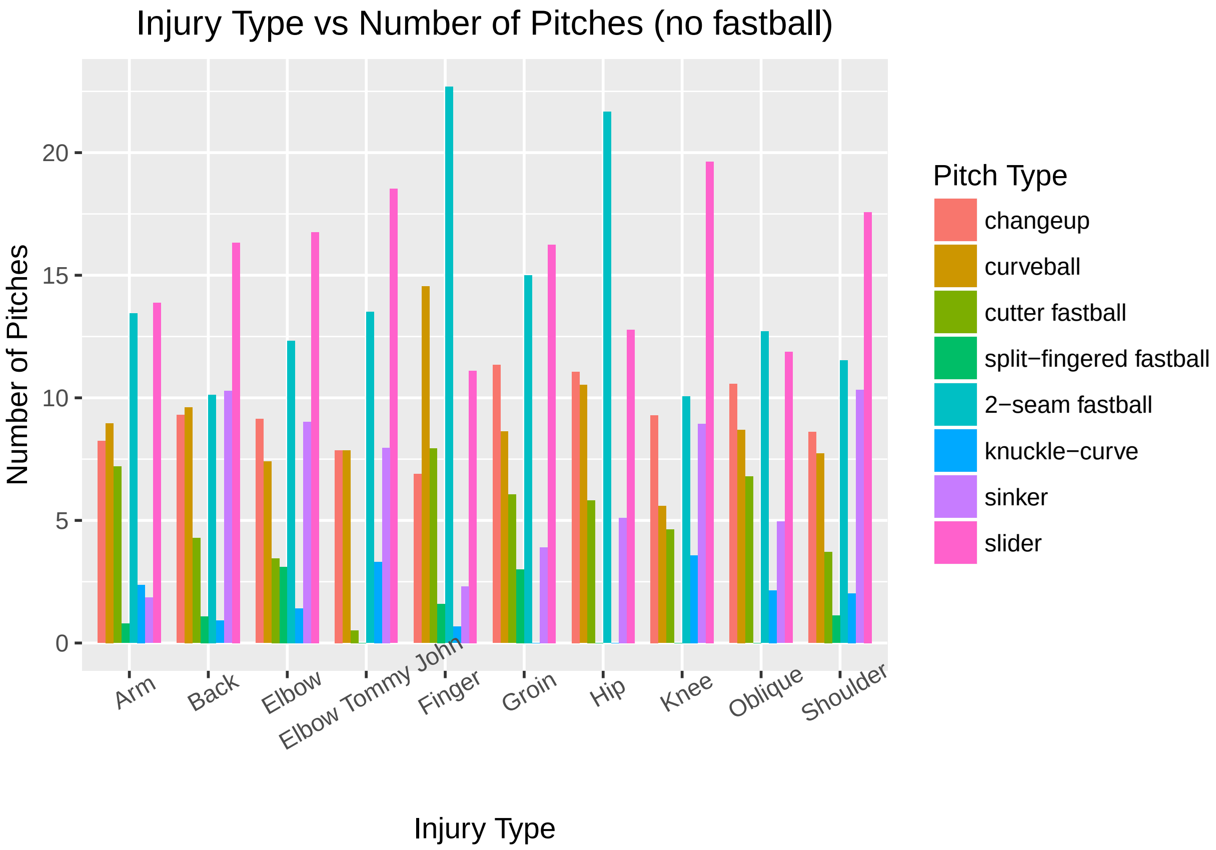
First we filter out every column except the 9 pitch percentage columns and plot the various kinds of injuries vs average pitch type percentage across all players with those injuries:

```
percent_cols <- colnames(joined)[8:16]
# tidy the data so we can facet on pitch types

g <- gather(filter(joined_pitchers), pt, pt_pct, percent_cols)
levels(g$pt) <- percent_cols
g <- arrange(g, pt)

ggplot(g, aes(x = type, y = pt_pct, fill = pt, order = pt)) + geom_bar(stat="summary", position = "dodg
  theme(axis.text.x = element_text(angle=30), plot.title = element_text(hjust = 0.5)) +
  ylab("Number of Pitches") + xlab("Injury Type") + ggtitle("Injury Type vs Number of Pitches") +
  scale_color_colorblind() +
  scale_fill_discrete(name="Pitch Type", labels = c("changeup","curveball","fastball","cutter fastball"
```



Fastballs dominate too much to clearly see what's happening with the rest of the pitches, so let's filter fastballs out and repeat the analysis:

```
percent_cols_no_FA <- colnames(joined)[9:16]
# tidy the data so we can facet on pitch types
g <- gather(filter(joined_pitchers), pt, pt_pct, percent_cols_no_FA)
levels(g$pt) <- percent_cols_no_FA
g <- arrange(g, pt)

ggplot(g, aes(x = type, y = pt_pct, fill = pt, order = pt, width=0.8)) + geom_bar(stat="summary", positi
  ylab("Number of Pitches") + xlab("Injury Type") +
  ggtitle("Injury Type vs Number of Pitches (no fastball)") + scale_color_colorblind() +
  scale_fill_discrete(name="Pitch Type", labels = c("changeup","curveball","cutter fastball","split-fing
```



Now, we see that players with finger injuries, oblique injuries, and hip injuries threw, on average, a higher percentage of two-seam fastballs than other pitches. For players who sustained all other injury types, the average percentage of sliders thrown was highest.

Part2: How Do Injuries Effect Power?

Now we will look at how different kinds of injuries affect hitters differently. Presumably some kinds of injuries hurt hitters more than others, and we will look at how different kinds of injuries affect power performance as measured by the aforementioned *SLG*. Because we're interested in the long-term effects of injuries, we're going to be looking at a one year lag, *ignoring* the season in which the player was injured. For each player we look at he season before and seasons after the injury to look for changes. First we prep the data:

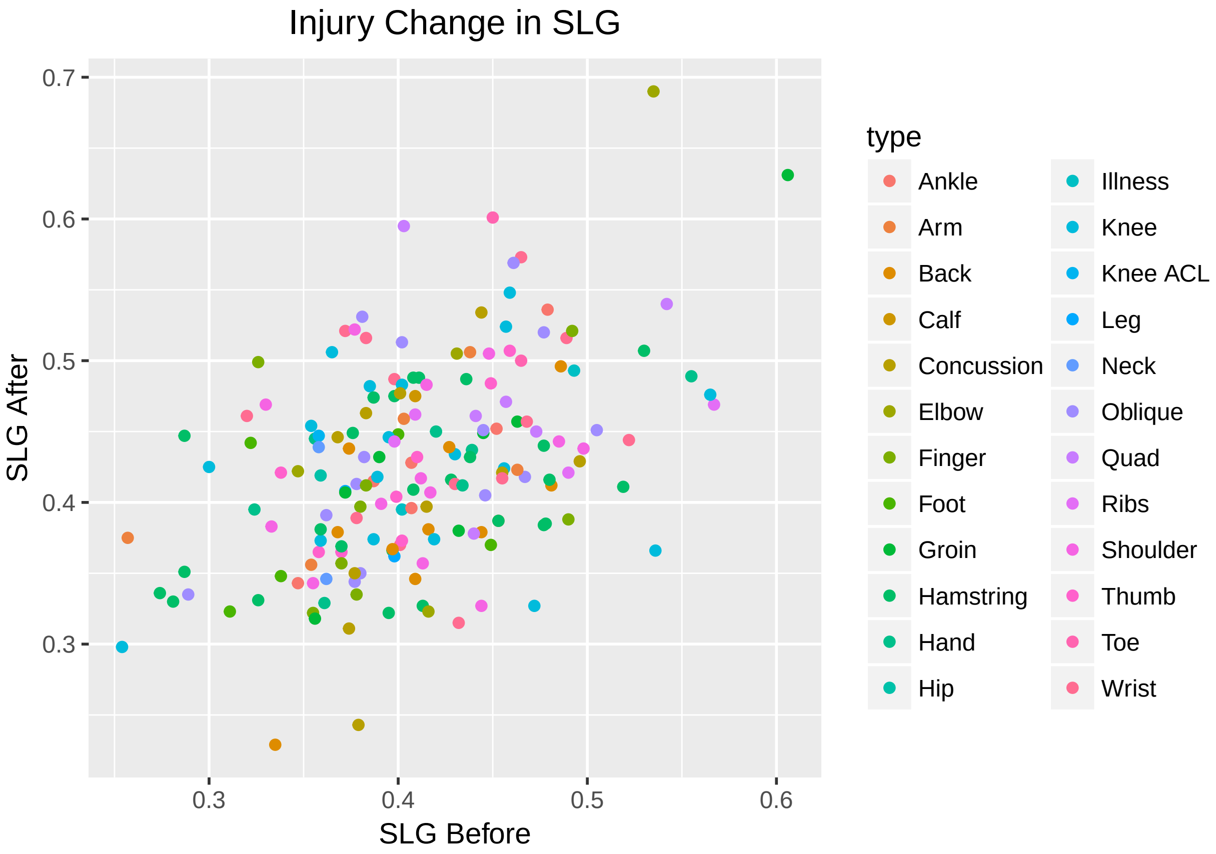
```
injuries2015 <- read_csv('./data/injuries_2015.csv')
injuries2016 <- read_csv('./data/injuries_2016.csv')

df2014 <- read_csv('./data/batters2014.csv')
df2015 <- read_csv('./data/batters2015.csv')
df2016 <- read_csv('./data/batters2016.csv')
df2017 <- read_csv('./data/batters2017.csv')

df2014$year <- 2014
df2015$year <- 2015
df2016$year <- 2016
df2017$year <- 2017
```

Next we tidy the data for a three-year period and finalize the data into one data frame:

```
process <- function(data, injuries) {
  # we only want players who played in each season, and batted at least 150 times in 2015/2017
  data <- mutate(group_by(data, name), cnt = n())
  first_yr <- min(data$year)
  third_yr <- max(data$year)
  scnd_year <- (first_yr + third_yr) / 2
  data <- filter(data, cnt == 3 & year != scnd_year)
  # tidy till before and after
  data <- mutate(data, occ = if_else(year == first_yr, "before", "after"))
  data <- data %>% select(name, occ, SLG)
  data <- spread(data, occ, SLG)
  # merge the injuries
  data <- merge(data, injuries, by.x = "name", by.y = "name", all.x = T)
  data$type[is.na(data$type)] <- "None"
  data$type <- sapply(data$type, extract_injury)
  data
}

bat1 <- rbind(df2015, df2016, df2017)
bat2 <- rbind(df2014, df2015, df2016)

bat1 <- process(bat1, injuries2016)
bat2 <- process(bat2, injuries2015)

bat <- rbind(bat1, bat2)
bat <- bat [!(bat$type=="None"),]
```

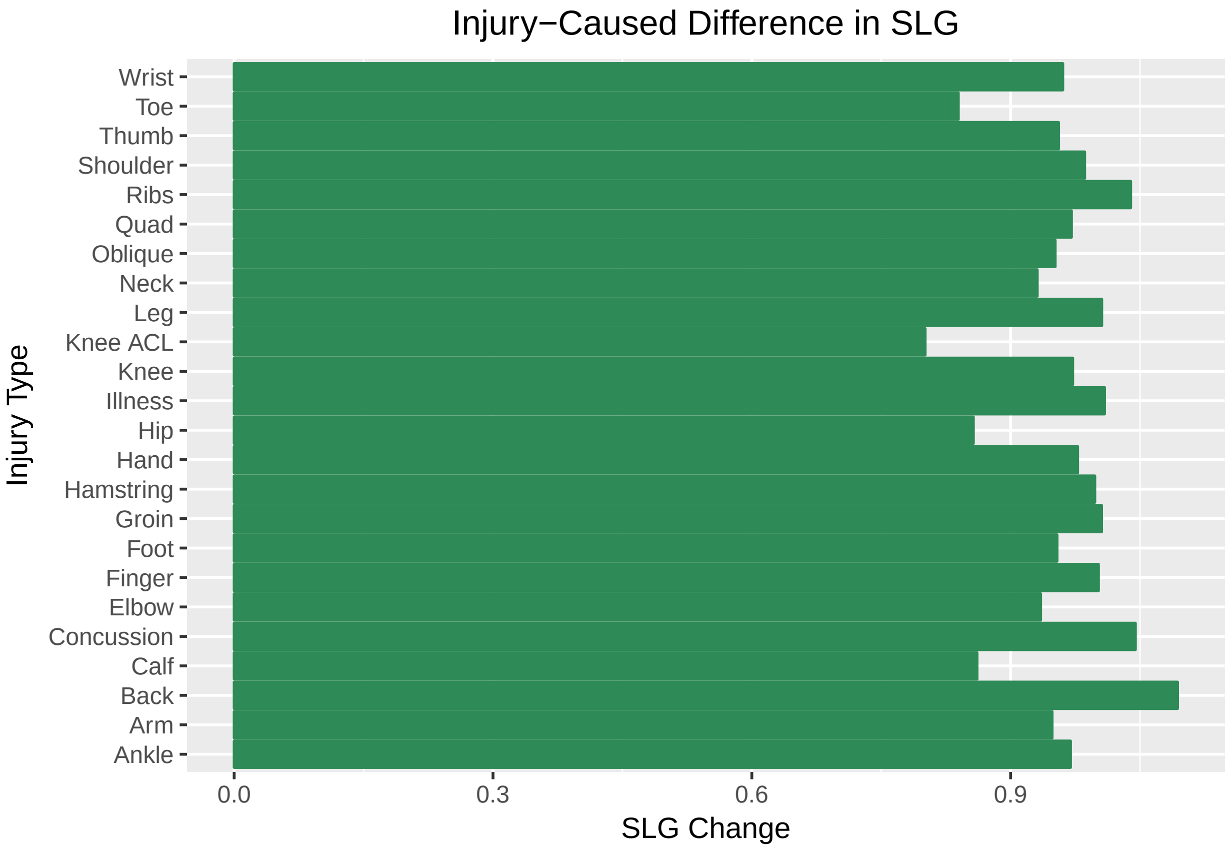
Lastly we use a scatterplot to plot *SLG* the year before injury vs *SLG* the year after injury colored by injury type:

```
ggplot(bat, aes(before, after)) + geom_point(aes(color = type)) + ylab("SLG After") + xlab("SLG Before")
```



Immediately, it is quite difficult to see much of anything, so instead we try a bar chart, plotting the ratios of *SLG* before and *SLG* after injury.

```
bat <- mutate(bat, chng = before / after)

ggplot(bat) + geom_bar(aes(x = type, y = chng), stat="summary", color="seagreen4", fill="seagreen4") + s
  coord_flip() + ylab("SLG Change") + xlab("Injury Type") + ggtitle("Injury-Caused Difference in SLG") +
  theme(plot.title = element_text(hjust = 0.5))
```

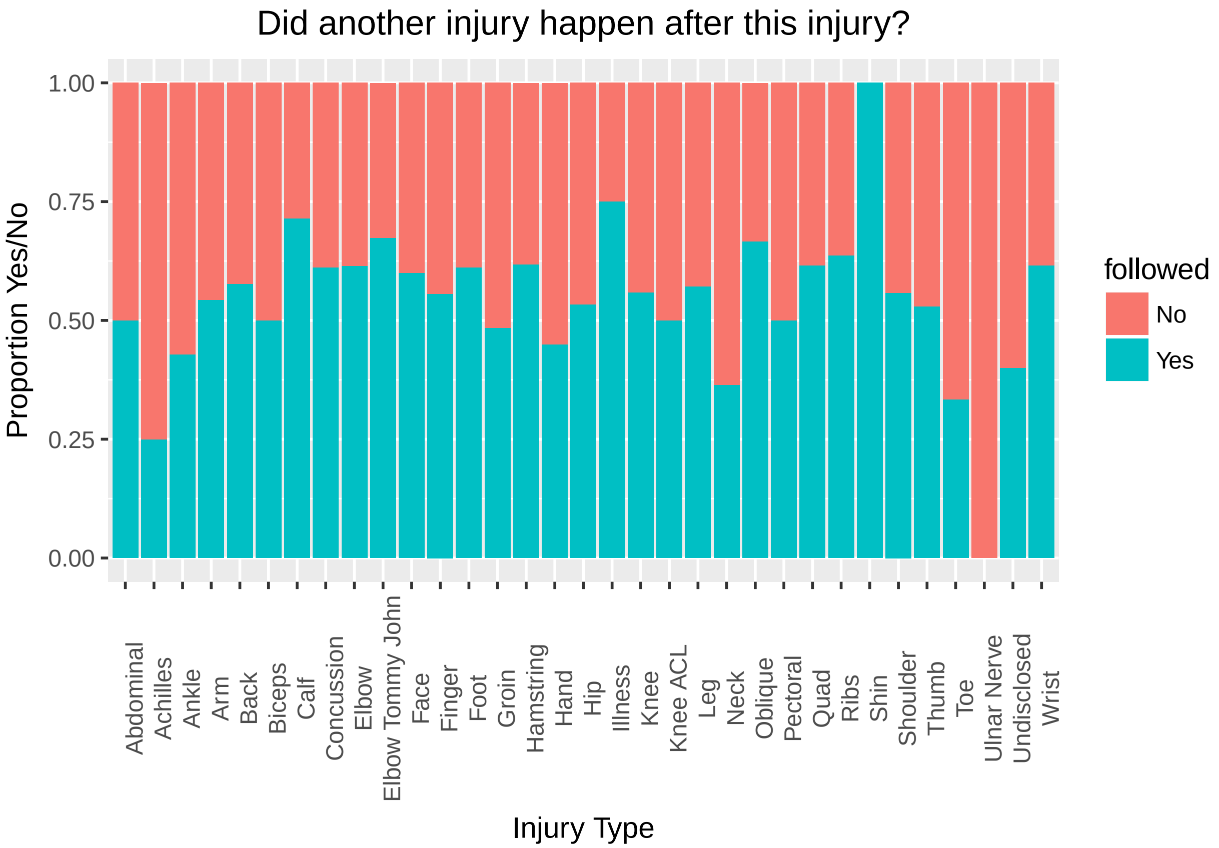


Now it's possible to see some patterns emerge. Lower body injuries seem to have the largest effect on power; ACL tears, knee injuries, toe injuries, hip, and calf injuries have the largest decreases in *SLG* the year after injury. This makes intuitive sense, since a large part of hitting involves building power from the lower half into the swing. However, some of the results are non-obvious; for example, it is remarkable that toe injuries have a larger effect on how hard a player hits the ball than wrist injuries. Of course this analysis uses only four seasons of data, and we'd want to revisit this on larger datasets. Nonetheless the lower body/upper body difference seems strong.

Part 3: How Well Do Injuries Predict Future Injuries?

Our next question involves the extent to which different injuries suggest future susceptibility to more injury. We combine all the injury data into one dataframe, and plot a stacked bar chart to see what percentage of players who sustained an injury got injured again:

```r
# process injury data
i1 <- read_csv('./data/injuries_2015.csv')
i2 <- read_csv('./data/injuries_2016.csv')
i3 <- read_csv('./data/injuries_2017.csv')

i1$year = 2015
i2$year = 2016
i3$year = 2017

iAll <- rbind(i1, i2, i3)
iAll <- arrange(iAll, name, year)
iAll <- mutate(group_by(iAll, name), cnt = n())
# data <- mutate(data, occ = if_else(year == first_yr, "before", "after"))
iAll <- iAll %>% mutate(was_followed = if_else(cnt > 1 & year != 2017, 1, 0))
# remove 2017
iAll <- filter(iAll, year != 2017)
iAll$type <- sapply(iAll$type, extract_injury)
iAll <- iAll %>% group_by(type, was_followed) %>%
          summarise(num_i = n()) %>%
          mutate(prop = num_i / sum(num_i))

iAll <- iAll %>% mutate(followed = if_else(was_followed == 0, "No", "Yes"))

ggplot(iAll, aes(x = type, y = prop, fill = followed)) + geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle=90), plot.title = element_text(hjust = 0.5)) +
  ylab("Proportion Yes/No") + xlab("Injury Type") + ggtitle("Did another injury happen after this injur
  scale_color_colorblind()
```



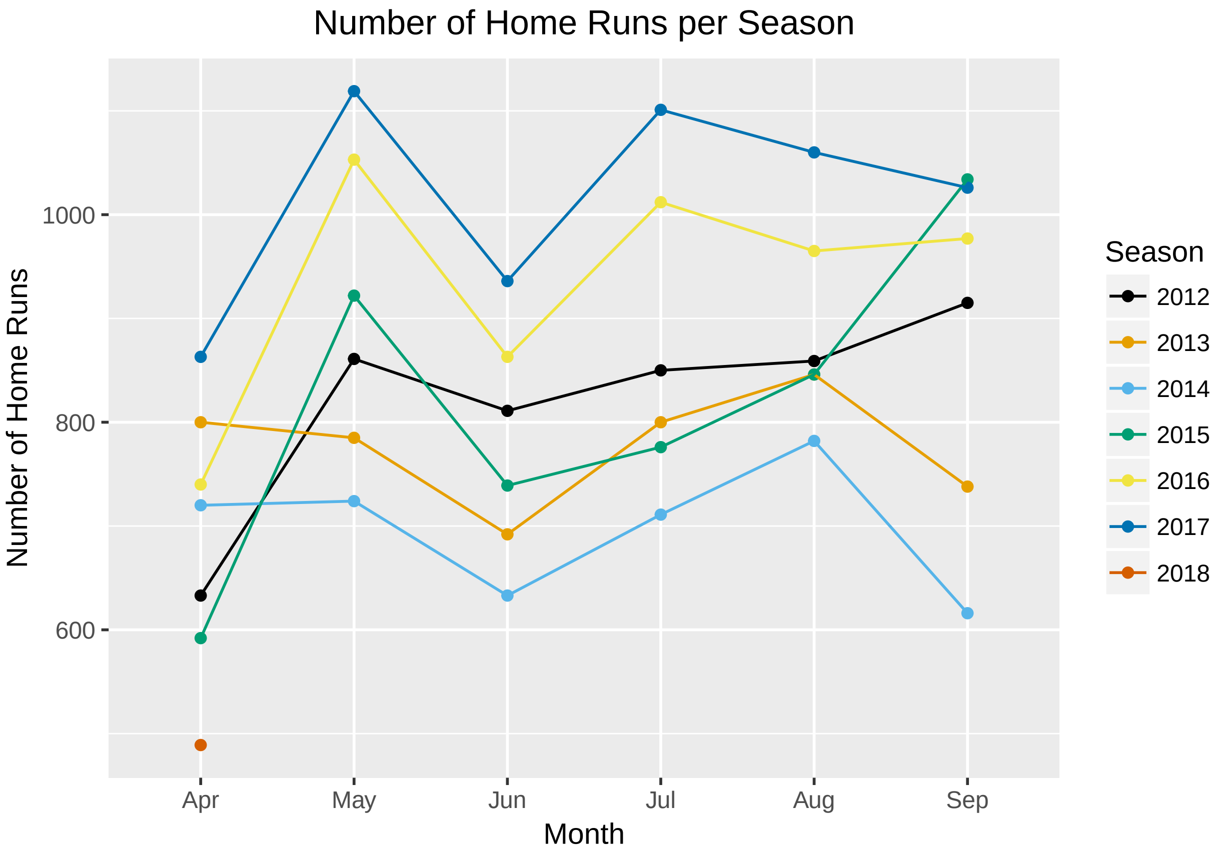
It seems that certain types on injuries are definitely more predictive of future injuries, but the caveat is that we don't have much data to work with. The numbers are not huge because finding injury data that goes back more than a few years is difficult. With a larger dataset, we could investigate if certain types of injuries could predict specific future injuries, but our slices would be too small for just three years of data.

Part 4: Did the Ball Change in the Middle of 2015?

Baseball fans and statisticians noticed that there was a spike in home runs after 2015. This wouldn't be unusual if the trend weren't sustained over time, but interestingly it seems that more home runs for the season is the new norm. We investigate this below. First we plot change in total home runs over time, with a different line for each year:

```
## structure the data so that months are orderd included, in the date
hr <- read_csv('./data/homeRuns.csv')
# just make Mar/April April and Sep/Oct September
hr$Month[hr$Month == 'Mar/Apr'] <- 'Apr'
hr$Month[hr$Month == 'Sept/Oct'] <- 'Sep'
hr$Month <- as.factor(hr$Month)
levels(hr$Month) = c("Apr", "May", "Jun", "Jul", "Aug", "Sep")
# zoo parser is much better
hr$Date <- as.yearmon(paste0(hr$Season, hr$Month), "%Y%b")
hr$Date <- as.Date(hr$Date)
```

```
label <- function(d) { format(d, "%b") }

ggplot(hr, aes(x = Month, y = HR, color = as.character(Season), group = as.character(Season))) + theme(
  labs(x = "Month", y = "Number of Home Runs") + ggtitle("Number of Home Runs per Season") +
  scale_color_colorblind(name = "Season")
```



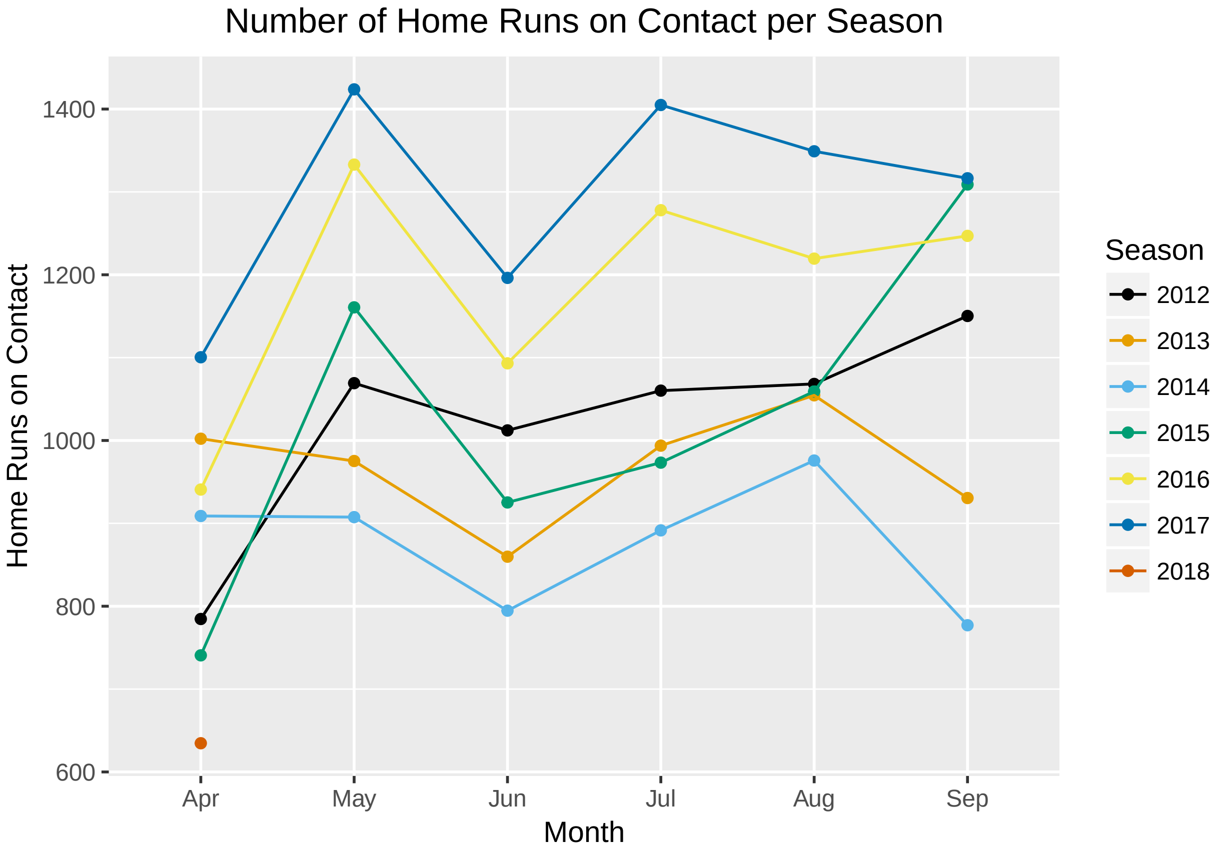
The first thing to note is how different and consistent the trends are for home runs across months. Fewer home runs are hit in colder months than warmer months. The reason for this is that colder weather makes air more dense, and denser air creates more drag on flying balls than thinner air. Also, certain months have fewer games and thus have fewer expected number of home runs.

As for yearly trends, clearly total home runs are up after 2015. The top two time series lines are for 2016 and 2017. Looking at the lines, it certainly looks like home runs spiked above the previous trend after July of 2015. This doesn't tell us anything about the ball, however.

A more informative thing to look at would be *home runs on contact*, which is a measurement of home run rate as a percentage of contact (contact meaning at-bats that weren't strikeouts) since how many strikeouts there are could be influencing the number of home runs. If fewer players are striking out, then home runs will naturally go up. We can strip out all strikeouts to see what would happen if every ball was put into play.

Home Runs on Contact

```
hr <- hr %>% mutate(onContact =  HR / (1 - `K%`))
levels(hr$Month) = c("Apr", "May", "Jun", "Jul", "Aug", "Sep")
ggplot(hr, aes(x = Month, y = onContact, color = as.character(Season), group = as.character(Season))) +
  theme(plot.title = element_text(hjust = 0.5)) + geom_point() + geom_line() +
  labs(x = "Month", y = "Home Runs on Contact") + ggtitle("Number of Home Runs on Contact per Season") +
  scale_color_colorblind(name = "Season")
```



Now, the difference looks even starker. The gap between May 2015 and May 2017 looks like it's a little under 300, from a baseline of around 1150, which corresponds to an increase of around 25%. That's an enormous number given the context; if a ball were put into play in May of 2017, it was 25% more likely to be hit for a home run than it was in 2015. There are of course other explanations - the composition of players isn't the same, ballparks in the league could have changed, or weather patterns could be changing in specific ways that influence home runs. But there are hundreds of players in the league and compisition changes slowly, and the trend increase appears to have begun in the middle of a season, when none of the parks had changed.

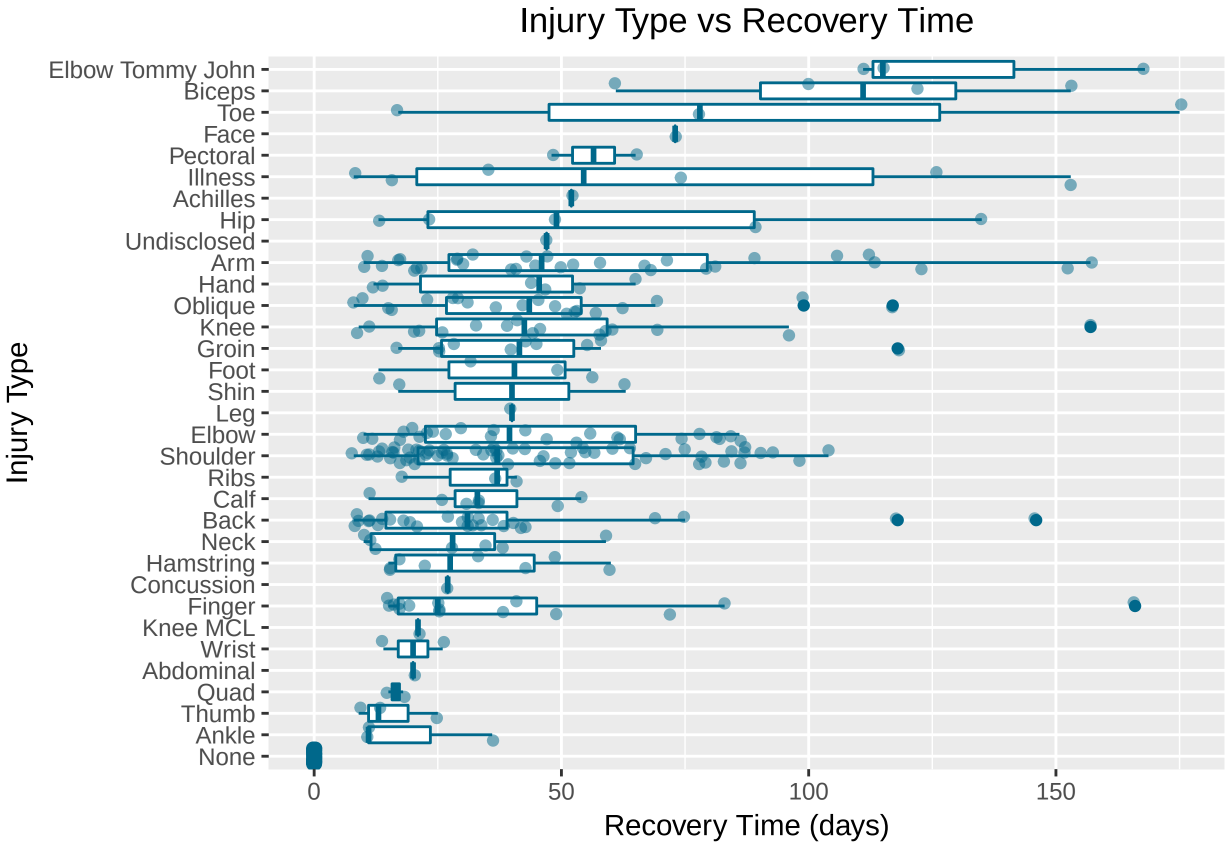
Interestingly, one small change was made; the location of a baseball manufacturing factory was changed in the middle of the 2015 season, and anecdotal comments from players suggest that the seams of the ball are generally slightly lower (though still within the strict range allowed by the MLB). Even a small decrease in seam height reduces drag on the ball and causes it to carry further. The league insists it has run tests and nothing has changed, but it may be worth trying to collect a set of balls from before and after the location change and see if seam height was is different.

Executive Summary

Since its inception, baseball has maintained a position of fond reverence in American culture, and from an analytical perspective is rich with potential, so much so that books, movies, even businesses have been devoted to the use of statistical techniques to inform decisions about the sport. Baseball is somewhat of an isolated team sport - every player is working towards a team goal, but by nature of the game, their contributions to the team goal are largely individual, which enables ease of analysis. Our report squares its aim on some of the lesser-analyzed areas supporting the game, such as the parameters surrounding pitcher injuries, how slightly changing the design of the ball may have caused an increase in home runs, and finally what would happen if we replaced umpires (human referees) with robots.

In our analysis of pitcher injuries, we didn't find any clear relationship between between injury type and number of pitches or average pitch speed. We were able to determine that the body parts most commonly injured were the arm, elbow, and shoulder as shown in the plot below.

```
ggplot(filter(joined, days < 180), aes(x=reorder(type, days, FUN = median), y=days)) +
  geom_boxplot(color="deepskyblue4") + geom_jitter(color="deepskyblue4", alpha = 0.5) + coord_flip() +
  ylab("Recovery Time (days)") + xlab("Injury Type") + ggtitle("Injury Type vs Recovery Time") + theme(
```

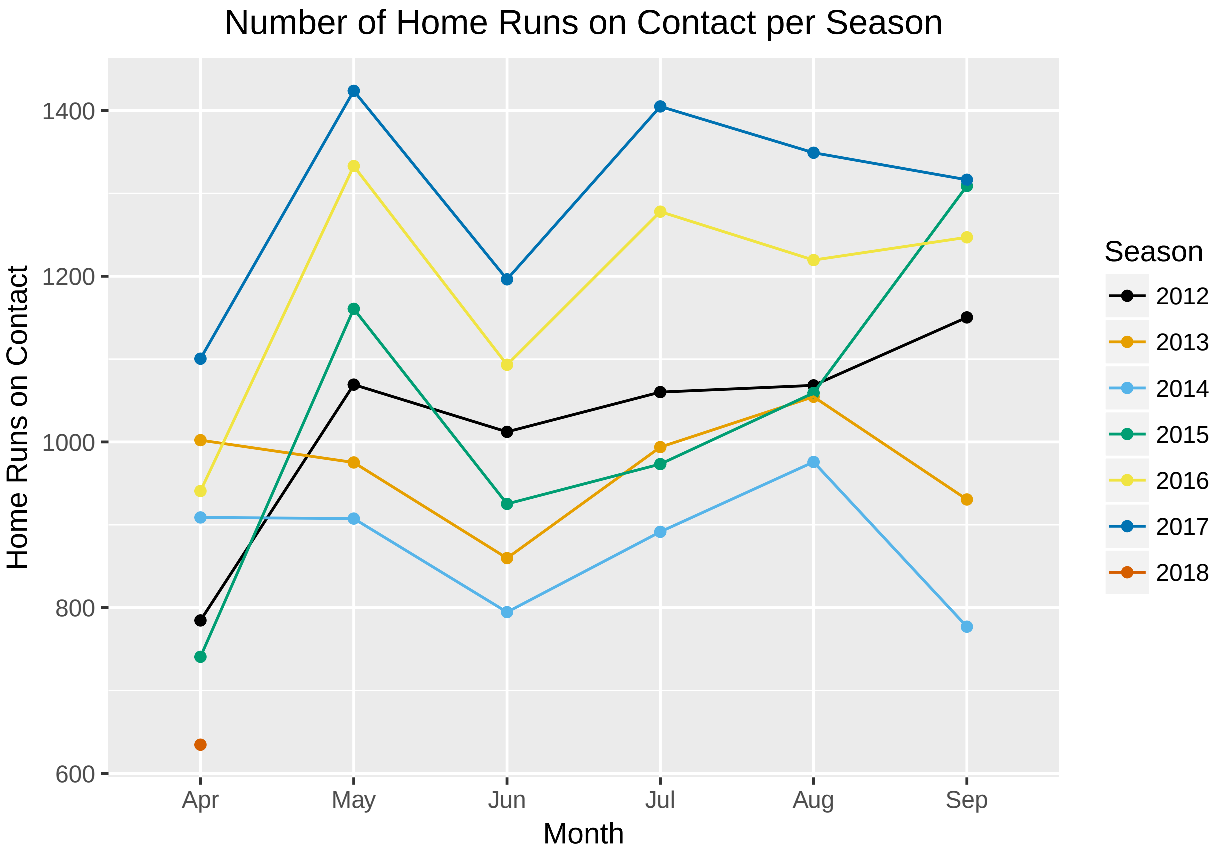


The dots represent individual injuries, while the boxplot shows the median and spread of recovery times for each location-based injury. The plot is ordered from longest recovery time at the top to shortest recovery time at the bottom. The Elbow Tommy John injury, also known as tearing the UCL (located in the elbow) is the most severe injury in terms of recovery time - so severe it usually renders the pitcher unable to play for the rest of the season. Given that pitchers use their arms more than the other parts of their body, it makes sense that the most common injuries are located in the shoulder/elbow/arm region of the body.

Analysis of pitch type and injury showed that players with injuries located on their fingers, obliques, and hips threw more two-seam fastballs than other types of pitches. For all other injury locations, the slider was the most commonly thrown pitch type. Examination of a pitcher's offensive performance (how well they hit) showed that on the whole, lower body injuries caused a greater decrease in offensive performance than upper body injuries. Finally, it seems that certain injuries are more likely to recur or cause new injuries to the same location, however these results varied based on location with no general trends.

Next we explored the theory that the 2015 change in the location of the baseball-manufacturer may have caused an increase in the number of home runs hit in following seasons.

```
ggplot(hr, aes(x = Month, y = onContact, color = as.character(Season), group = as.character(Season))) +
  theme(plot.title = element_text(hjust = 0.5)) + geom_point() + geom_line() +
  labs(x = "Month", y = "Home Runs on Contact") + ggtitle("Number of Home Runs on Contact per Season") +
  scale_color_colorblind(name = "Season")
```



No matter which way we looked at the data, two observations held true: more home runs are being hit every year, and more home runs are hit in warmer weather versus colder weather, as shown by the plot above. In fact, there appears to be a nearly 30% increase in homeruns hit during August 2017 compared to August 2015.

Finally, our interactive was a tool to explore how different pitchers use their various pitches. It lets you plot by pitch type for five of the best pitchers in baseball, colored by balls and strikes. So you can see that Max Scherzer's sliders tend to break away from righthanders, and that he gets a lot of swinging strikes on them out of the zone. You can also see how different pitchers use place them near different parts of the plate. We'd like to include all pitchers across many years, but pitch data is enormous and it would require a real backend API, since loading a season's worth of data into the browser for more than a few pitchers is a non-starter.

But that will probably be the next feature we add.

Conclusion

In this report we assessed all the parameters surrounding pitcher injuries, highlighted overall home run trends in recent years, and took a predictive look at how robot empires could affect the strike zone and therefore the future of the game.

As is common, we were limited by the scope of data that was readily accessible on the internet, but our assessment of data quality indicated that there were no inconsistencies or glaring gaps in the data that might have negatively affected our results. But, while the data we were able to find was of high quality, we still wish we had more data to work with, especially for our injury analysis. We are cognizant that extensive injury reports are a more recent trend, and becoming more popular especially with media attention on the intersection between sports and player health, but we were only able to find data going a few years back, which limited the scope of our inquiry. Furthermore, many of our injury graphs mention parts of the body twice (such as Knee ACL vs Knee) where one reference is more specific than the other, demonstrating a somewhat confusing lack of consistency (for example, does this mean that "Knee" references any other non-ACL knee injuries). We think the quality of injury data will improve as the sport increases its attention toward injuries.

In the future, we'd love to get our hands on more fine grained injury data across more seasons. Unfortunately, exploratory data analysis is incapable of answering many injury questions of interest. There are lots of interaction effects happening in baseball (for instance, more pitches might cause more injuries all things equal, but healthier pitchers will tend to throw more than unhealthy ones), and many of the relationships can only be teased out through modeling. But exploratory visualization is a great place to start looking for questions to ask. The data on lower body injuries for hitters was interesting, and that's another thing we'd like to explore further. See it here link.

Aside from the individual conclusions revealed throughout our report, we gained an understanding of the importance of letting the data drive the direction of our analysis and drawing insight from the data instead of relying on our opinions. For example, in our evaluation of pitch type versus injury type, we initially found that fastballs were highly correlated with every injury, which makes sense since it is the most common type of pitch. Once we removed fastball from the data, subtler trends became more pronounced such as the correlation between the two-seam fastball and finger/hip/oblique injuries. However, we also learned that trends are merely indicative, meaning they don't demonstrate a clear cause-effect relationship. In an ideal world, if a player gets injured, it's because they over-used a body part, but accidents happen and they happen frequently. And that's the price tag that comes with using real data: sometimes the trends don't explain or predict what you see, but with them you can get a bit closer to understanding.