NEISS_Aanlytics

Bowei.Zhang

November 16, 2016

Introduction

In my spare time I will do some analytics projects using public data just for fun, this NEISS Analytics is one of them.

CPSC's National Electronic Injury Surveillance System (NEISS) is a national probability sample of hospitals in the U.S. and its territories. Patient information is collected from each NEISS hospital for every emergency visit involving an injury associated with consumer products. From this sample, the total number of product-related injuries treated in hospital emergency rooms nationwide can be estimated. This web access to NEISS allows certain estimates to be retrieved on-line. These estimates can be focused by setting some or all of the following variables (and an example of each):

Date (one year maximum range; e.g., how many injuries were treated in 1996)
Product (e.g., how many bicycle injuries occurred)
Sex (e.g., how many injuries occurred to women)
Age (e.g., how many injuries occurred to people aged 35-55)
Diagnosis (e.g., how many lacerations occurred)
Disposition (e.g., how many people were admitted to the hospital)
Locale (e.g., how many injuries occurred at a school)
Body part (e.g., how many injuries involved the knee)

This data was gathered from the United States Consumer Product Safety Commission's National Electronic Injury Surveillance System (NEISS) available here (http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/)

Full methodology is included at the above link, this archive also includes the NEISS coding manual with further descriptions of the data and methodology.

I also included all of the data that have been used in this analytics in **data** folder.

Required Packages:

```r
require(dplyr)    # For data wrangling
require(ggplot2)  # For plots
require(cowplot)  # For better arrange plots
```

Data Preperation

Input dataset

```
# Set working dictionary to target folder
setwd("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/")

# read sample data, and change all blanks into 'NA'
neiss <- read.csv("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/data/NEISS201
4.csv", header=T, na.strings=c("","NA"))

# Look at the size of the data and data type of each variables
str(neiss)
```

```
## 'data.frame':    65499 obs. of  18 variables:
##  $ CPSC.Case..: int  141200216 140117851 150144993 150230176 141220717 150151229 150216968 14
1200989 141241629 150226548 ...
##  $ trmt_date  : Factor w/ 365 levels "1/1/14","1/10/14",..: 47 26 313 293 364 308 280 79 95 2
79 ...
##  $ psu        : int  63 63 63 63 63 63 63 63 63 63 ...
##  $ weight     : num  99.7 81.6 99.7 99.7 99.7 ...
##  $ stratum    : Factor w/ 5 levels "C","L","M","S",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ age        : int  21 62 21 30 16 22 92 89 75 68 ...
##  $ sex        : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 2 2 2 1 2 ...
##  $ race       : Factor w/ 7 levels "American Indian/Alaska Native",..: 5 7 5 5 5 5 5 5 5 5
 ...
##  $ race_other : Factor w/ 40 levels "`","~","AMERICAN INDIAN",..: NA NA NA NA NA NA NA NA NA
 NA ...
##  $ diag       : int  62 57 57 64 57 64 72 57 59 64 ...
##  $ diag_other : Factor w/ 1239 levels "? INSECT BITE",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ body_part  : int  75 79 83 79 82 79 36 79 75 35 ...
##  $ disposition: int  1 1 1 1 1 1 1 4 1 1 ...
##  $ location   : int  1 1 0 1 0 1 1 1 1 4 ...
##  $ fmv        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prod1      : int  679 1807 1333 4076 1893 1871 214 4074 3251 5040 ...
##  $ prod2      : int  1807 115 NA NA NA NA NA NA NA NA ...
##  $ narrative  : Factor w/ 65422 levels "- LAC. TO BOTTOM LIP3 YOM  FELL & HIT HIS MOUTH  ON A
 STAIR DX; SMALL LAC. TO CORNER OF MOUTH",..: 17688 43945 17667 25215 11162 18680 55994 54281 49
263 46018 ...
```

```
# Take a look at typical value of each variables
summary(neiss)
```

```
##   CPSC.Case..          trmt_date          psu             weight      
##  Min.   :140104670   5/25/14:  259   Min.   :  1.00   Min.   :  5.717  
##  1st Qu.:140436040   5/26/14:  259   1st Qu.: 23.00   1st Qu.: 14.309  
##  Median :140725351   5/24/14:  251   Median : 42.00   Median : 37.415  
##  Mean   :141132777   5/19/14:  248   Mean   : 46.55   Mean   : 46.628  
##  3rd Qu.:141031376   6/8/14 :  247   3rd Qu.: 66.00   3rd Qu.: 81.576  
##  Max.   :150331961   5/11/14:  243   Max.   :101.00   Max.   :112.167  
##                      (Other):63992                                     
##  stratum        age            sex       
##  C:10457   Min.   :  0.00   Female:29996  
##  L: 6079   1st Qu.: 12.50   Male  :35503  
##  M:11197   Median : 28.00                 
##  S:15814   Mean   : 43.22                 
##  V:21952   3rd Qu.: 57.00                 
##            Max.   :223.00                 
##                                           
##                                 race              race_other   
##  American Indian/Alaska Native   :  249   HISPANIC    : 2549  
##  Asian                           :  621   UNKNOWN     :  410  
##  Black/African American          : 9935   MULTI-RACIAL:  252  
##  Native Hawaiian/Pacific Islander:   36   HISP        :  161  
##  None listed                     :19593   NS          :   52  
##  Other / Mixed Race              : 3389   (Other)     :  153  
##  White                           :31676   NA's        :61922  
##      diag              diag_other      body_part      disposition   
##  Min.   :41.00   PAIN         : 2655   Min.   : 0.00   Min.   :1.000  
##  1st Qu.:57.00   BACK PAIN    :  271   1st Qu.:35.00   1st Qu.:1.000  
##  Median :59.00   INJURY       :  223   Median :75.00   Median :1.000  
##  Mean   :59.98   LOW BACK PAIN:  201   Mean   :64.19   Mean   :1.268  
##  3rd Qu.:64.00   CHEST PAIN   :  195   3rd Qu.:82.00   3rd Qu.:1.000  
##  Max.   :74.00   (Other)      : 4575   Max.   :94.00   Max.   :8.000  
##                  NA's         :57379                                  
##     location          fmv                prod1          prod2      
##  Min.   :0.000   Min.   :0.000000   Min.   : 106   Min.   : 102   
##  1st Qu.:0.000   1st Qu.:0.000000   1st Qu.:1211   1st Qu.:1141   
##  Median :1.000   Median :0.000000   Median :1807   Median :1807   
##  Mean   :2.404   Mean   :0.008122   Mean   :2108   Mean   :1846   
##  3rd Qu.:4.000   3rd Qu.:0.000000   3rd Qu.:3265   3rd Qu.:1871   
##  Max.   :9.000   Max.   :3.000000   Max.   :5555   Max.   :5555   
##                                                    NA's   :57053  
##                                                  narrative    
##  2 YO FEMALE FELL DOWN STEPS.  DX HEAD INJURY          :    5  
##  17 YO MALE PLAYING BASKETBALL.  DX ANKLE SPRAIN       :    4  
##  10 YO MALE JAMMED FINGER PLAYING FOOTBALL.  DX FX     :    2  
##  10 YOM FELL WHILE SKATEBOARDING.  DX: FRACTURE ANKLE. :    2  
##  11 YO MALE PLAYING FOOTBALL.  DX FINGER FX            :    2  
##  11 YO MALE PLAYING FOOTBALL.  DX FINGER SPRAIN        :    2  
##  (Other)                                               :65482
```

```
# Loading lookup BodyParts lookup tables
bodypart <- read.csv("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/data/BodyParts.csv", header=T, na.strings=c("","NA"))


# Loading lookup Diagnosis lookup tables
diagnosis <- read.csv("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/data/DiagnosisCodes.csv", header=T, na.strings=c("","NA"))


# Loading lookup Disposition lookup tables
disposition <- read.csv("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/data/Disposition.csv", header=T, na.strings=c("","NA"))


# Join NEISS with lookup tables
neiss <- left_join(neiss,bodypart,by = c("body_part" = "Code"))
neiss <- left_join(neiss,diagnosis,by = c("diag" = "Code"))
neiss <- left_join(neiss,disposition,by = c("disposition" = "Code"))
```

Which parts of our body are safe and which are not?

What are the body parts most frequently represented in this dataset?

```
# Count Body Parts
bodyPart_count <- count(neiss,BodyPart)

# Change column name
colnames(bodyPart_count)[2] <- "count"

# Get top 5 most frequently body parts in the dataset
arrange(bodyPart_count,desc(count))[1:5,]
```

```
## # A tibble: 5 x 2
##       BodyPart count
##         <fctr> <int>
## 1         Head  9891
## 2         Face  5786
## 3       Finger  5783
## 4 Trunk, lower  5717
## 5 Trunk, upper  3868
```

What are the body parts that are least frequently represented?

```
# Get top 5 least frequently body parts in the dataset
arrange(bodyPart_count,count)[1:5,]
```

```
## # A tibble: 5 x 2
##         BodyPart count
##           <fctr> <int>
## 1 25-50% of body     4
## 2   Pubic region   286
## 3   Not Recorded   390
## 4       Internal   549
## 5     Arm, upper   745
```

Skateboard?

Sometimes it worth to take deeper look at a specific consumer prodcut, for example, one of my favorite: skateboard

```
# Filter injuries involve a skateboard, in order to do that, we define "narrative" column contains word 'SKATEBOARD' as injuries involve a skateboard
skateboard <- filter(neiss, grepl('SKATEBOARD', narrative))
# Count injuries involve a skateboard
nrow(skateboard)
```

```
## [1] 466
```

Of those injuries, what percentage were male and what percentage were female?

```
# Percentage of male and female
prop.table(table(skateboard$sex))
```

```
##
##    Female      Male
## 0.1759657 0.8240343
```

What was the average age of someone injured in an incident involving a skateboard?

```
# In order to calculate average age, first we need to make sure all age value represent actual age since we have several age value are not from NEISSCodingManual (for example 201 = less than 8 weeks, 206 = 6 months, 218 = 18 months)
max(skateboard$age)
```

```
## [1] 71
```

So now we confirm that there are no value larger than 100, let’s calculate average age:

```
# Average age of someone injured in an incident involving a skateboard
mean(skateboard$age)
```

```
## [1] 17.99142
```

First look at diagnosis

What diagnosis had the highest hospitalization rate?

```
# In terms of disposition, hospitalization could mean either "Treated and transferred to another hospital" (code 2) or "Treated and admitted for hospitalization (within same facility)" (code 4)

# Create a new variable to represent hospitalization, in this variable, we use 1 to represent hospitalization and 0 to represent not hospitalization.
neiss$hospitalization <- ifelse(neiss$disposition == 2|neiss$disposition == 4,
c(1), c(0))

#Now we group NEISS data by Diagnosis, and within each group, the
head(arrange(neiss %>% group_by(Diagnosis) %>% summarise(hospitalization_rate=mean(hospitalization)),desc(hospitalization_rate)))
```

```
## # A tibble: 6 x 2
##                         Diagnosis hospitalization_rate
##                            <fctr>                <dbl>
## 1 Submersion (including Drowning)            0.4259259
## 2                      Amputation            0.2560000
## 3                        Fracture            0.2134566
## 4          Ingested foreign object            0.1532091
## 5                       Poisoning            0.1464088
## 6            Internal organ injury            0.1377686
```

We can see from the results that **Submersion** (42.59%) had the highest hospitalization rate

What diagnosis most often concluded with the individual leaving without being seen?

```
# leaving without being seen is in "disposion" (code 6)
head(sort(tapply(neiss$disposition, neiss$Diagnosis, function(x) prop.table(table(x))["6"]), decreasing = TRUE))
```

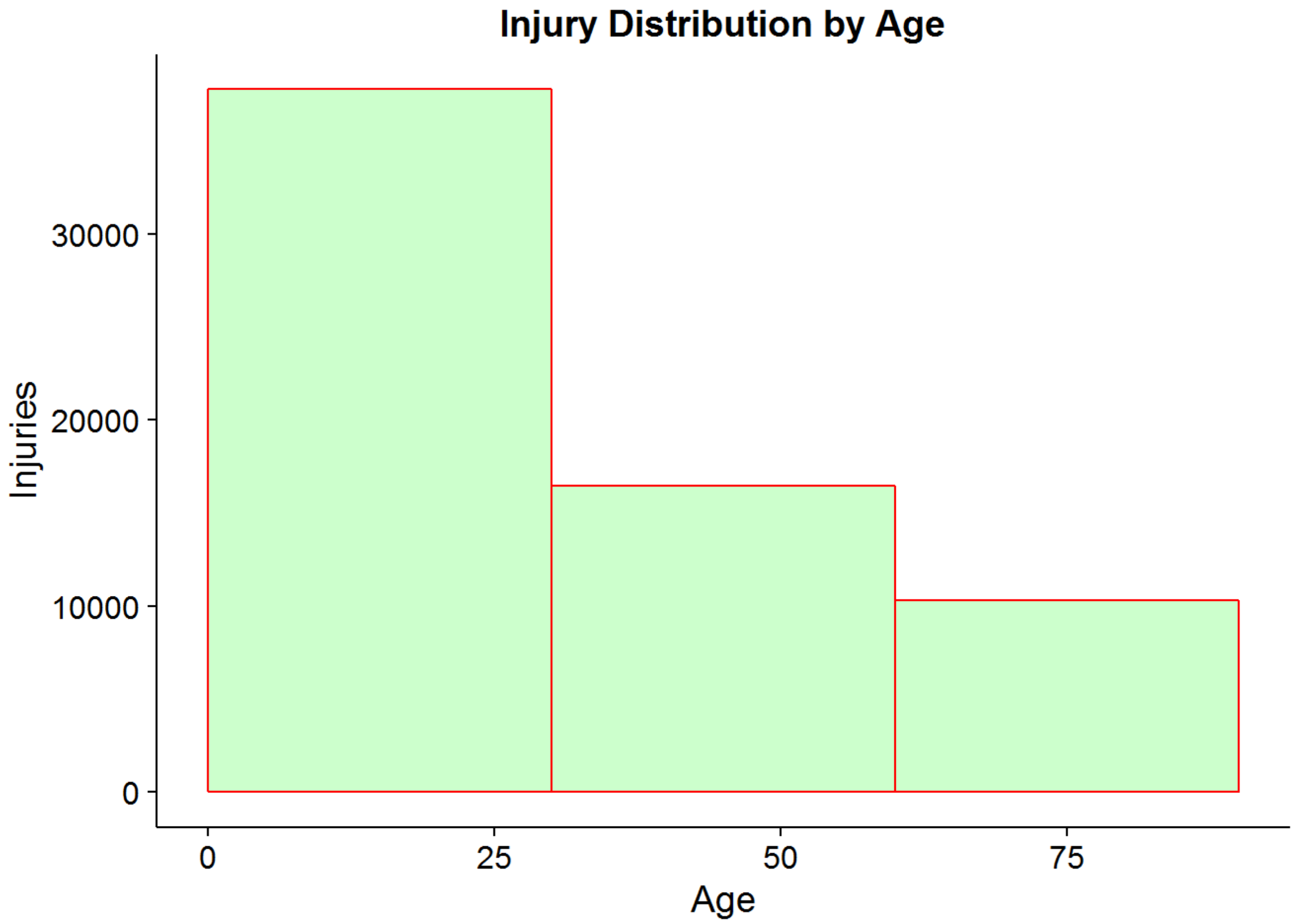
```
##
                  Poisoning
##
                 0.03314917
##
          Other/Not Stated
##
                 0.03165025
##
   Aspirated foreign object
##
                 0.03030303
## Burns, radiation (includes all cell damage by ultraviolet, x- rays, microwaves, laser beam, r
adioactive materials, etc.)
##
                 0.02857143
##                                                                                           Burn
s, chemical (caustics, etc.)
##
                 0.02325581
##
                   Hematoma
##
                 0.02247191
```

From above that we know **Poisoning** most concluded with the individual leaving without being seen (3.31%).

Will age have an impact on reported injuries?

It will be easier to tell if we visualize their relationship!

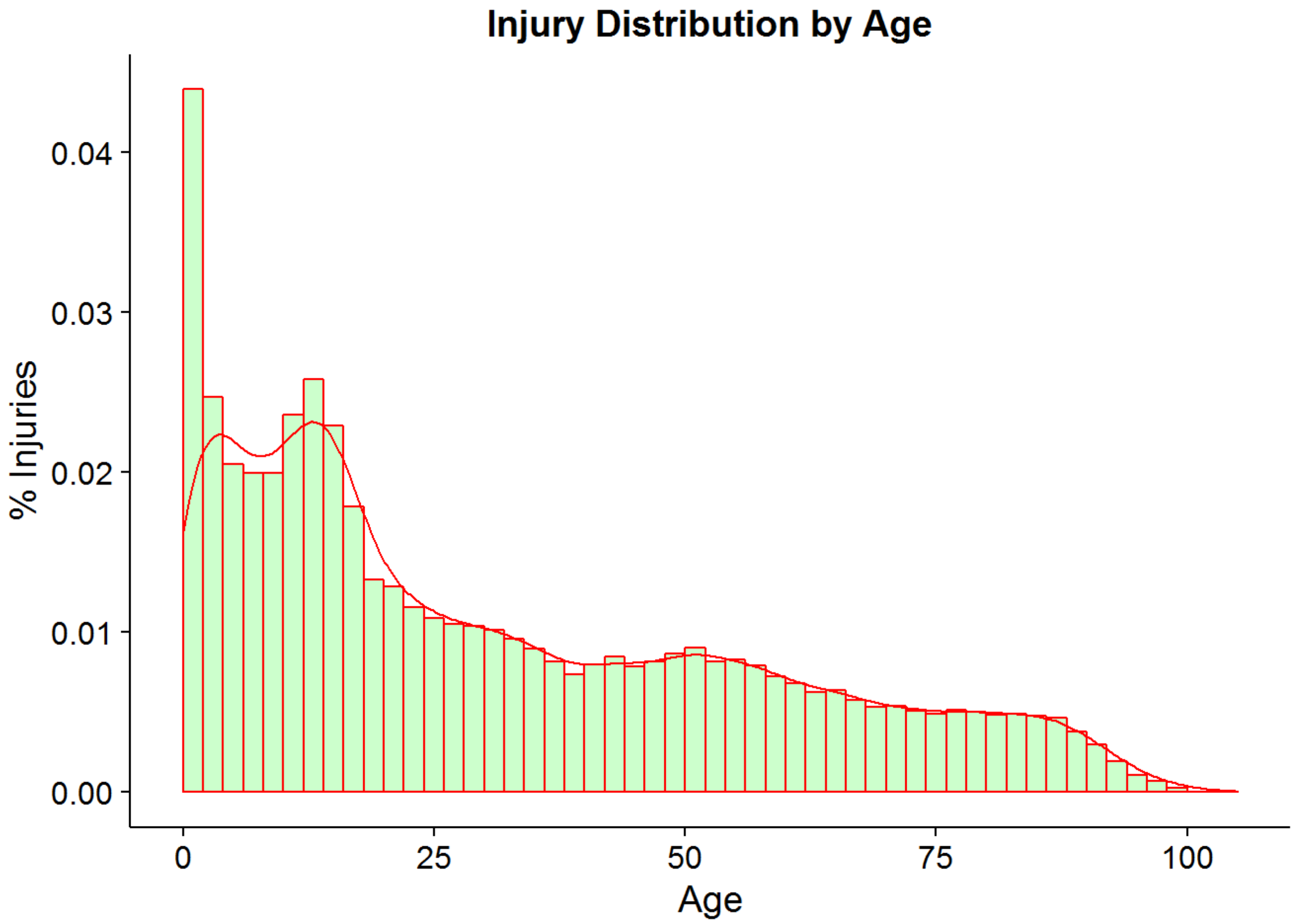
```
# First we need convert all values to represent the real age
neiss$plot_age <- ifelse(neiss$age > 200,
c(0), neiss$age)

ggplot(data=neiss, aes(neiss$plot_age)) +
  geom_histogram(breaks=seq(0, 100, by = 30),
                 col="red",
                 fill="green",
                 alpha = .2) +
  labs(title="Injury Distribution by Age") +
  labs(x="Age", y="Injuries")
```



I use 30 as bin width to divided age into 3 bins (Youth,Medium,Senior), from this chart we could know that people in **0-30 age group** (Youth) are more likely to get injury than medium or senior age group since their total injuries is larger than the sum of other 2 groups.

If we take a deeper look at distribution:

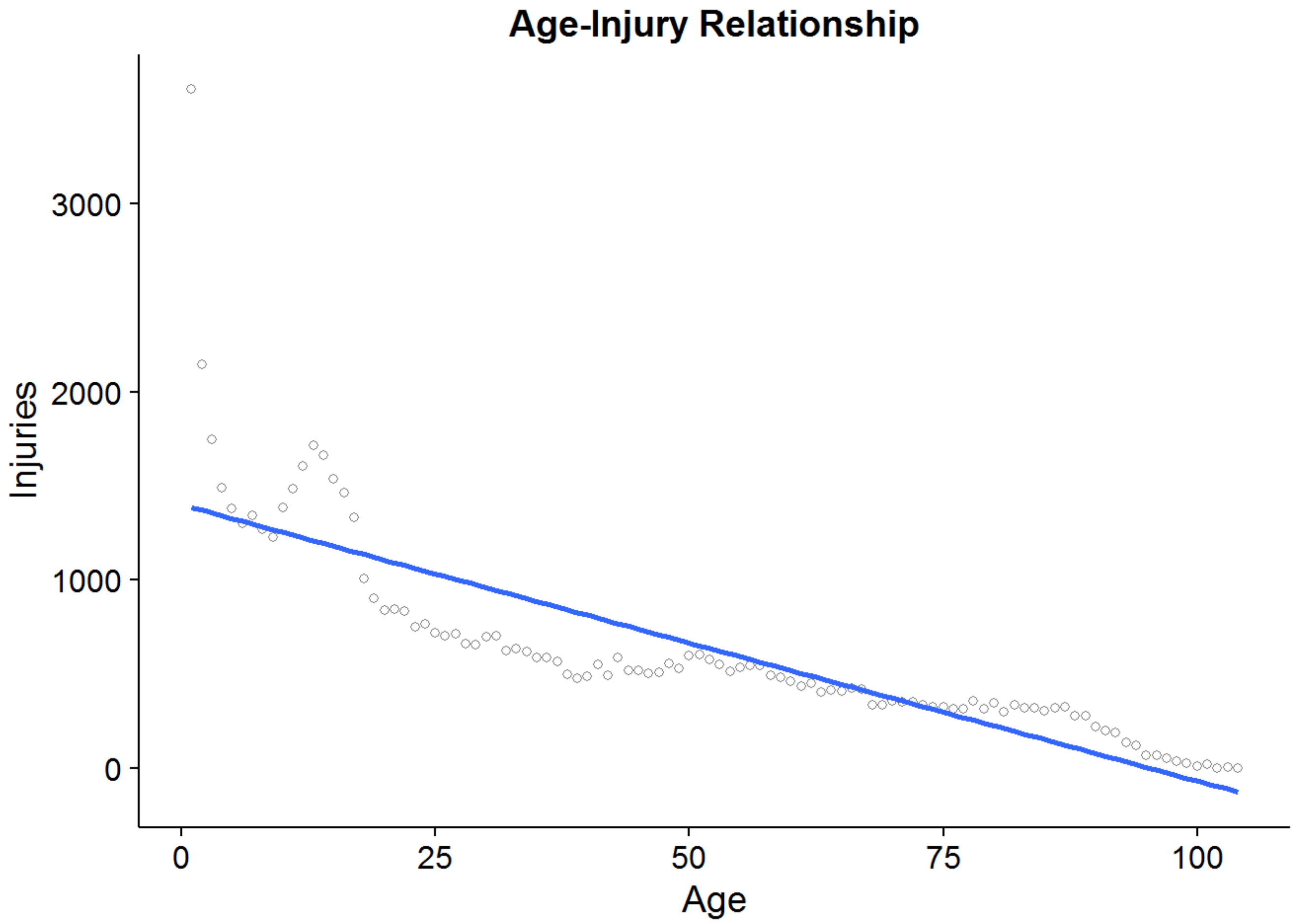
```
ggplot(data=neiss, aes(neiss$plot_age)) +
  geom_histogram(aes(y =..density..),breaks=seq(0, 100, by = 2),
                 col="red",
                 fill="green",
                 alpha = .2) +  geom_density(col=2) +
  labs(title="Injury Distribution by Age") +
  labs(x="Age", y="% Injuries")
```



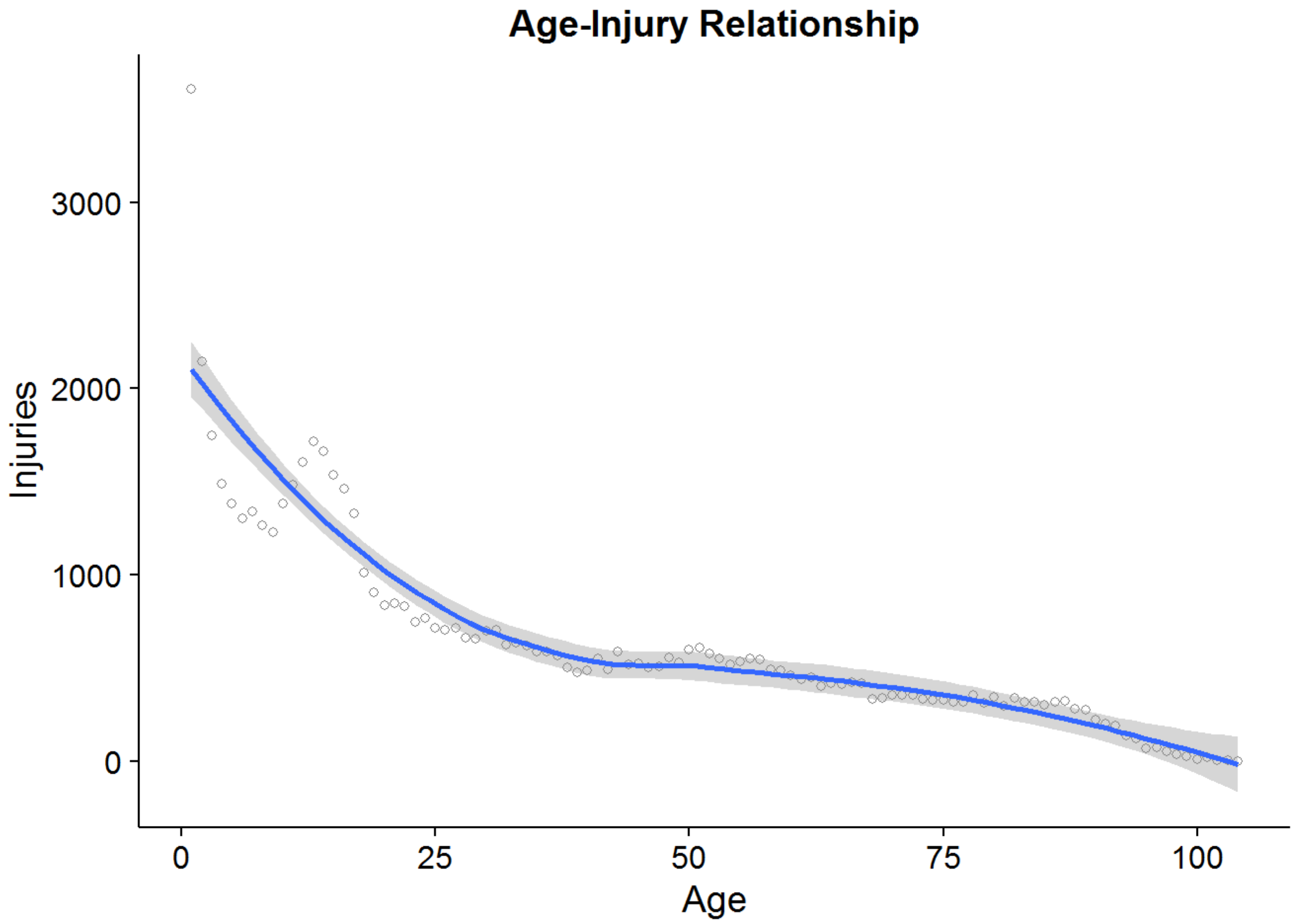
From the histogram above we could see that **newborn** (0-2) is the most dangerous age group in terms of getting injuries which make sense since they are not mature enough to detect dangers at that time.

Now let’s see if we could find the relationship between **Age** and **Injuries**

```
# Create a age frequency table:
age_freq <- as.data.frame(table(neiss$plot_age))
colnames(age_freq) <- c("age", "frequency")
age_freq$age <- as.numeric(age_freq$age)

# Create a scatterplot to visualize it

# Linear relationship
ggplot(age_freq, aes(x=age,y=frequency)) +
geom_point(shape=1,alpha =0.5) +
  labs(title="Age-Injury Relationship") +
  labs(x="Age", y="Injuries") +
  geom_smooth(method=lm,se=FALSE)
```



```
# A loess smoothed fit curve with confidence region
ggplot(age_freq, aes(x=age,y=frequency)) +
geom_point(shape=1,alpha =0.5) +
  labs(title="Age-Injury Relationship") +
  labs(x="Age", y="Injuries") +
  geom_smooth()
```

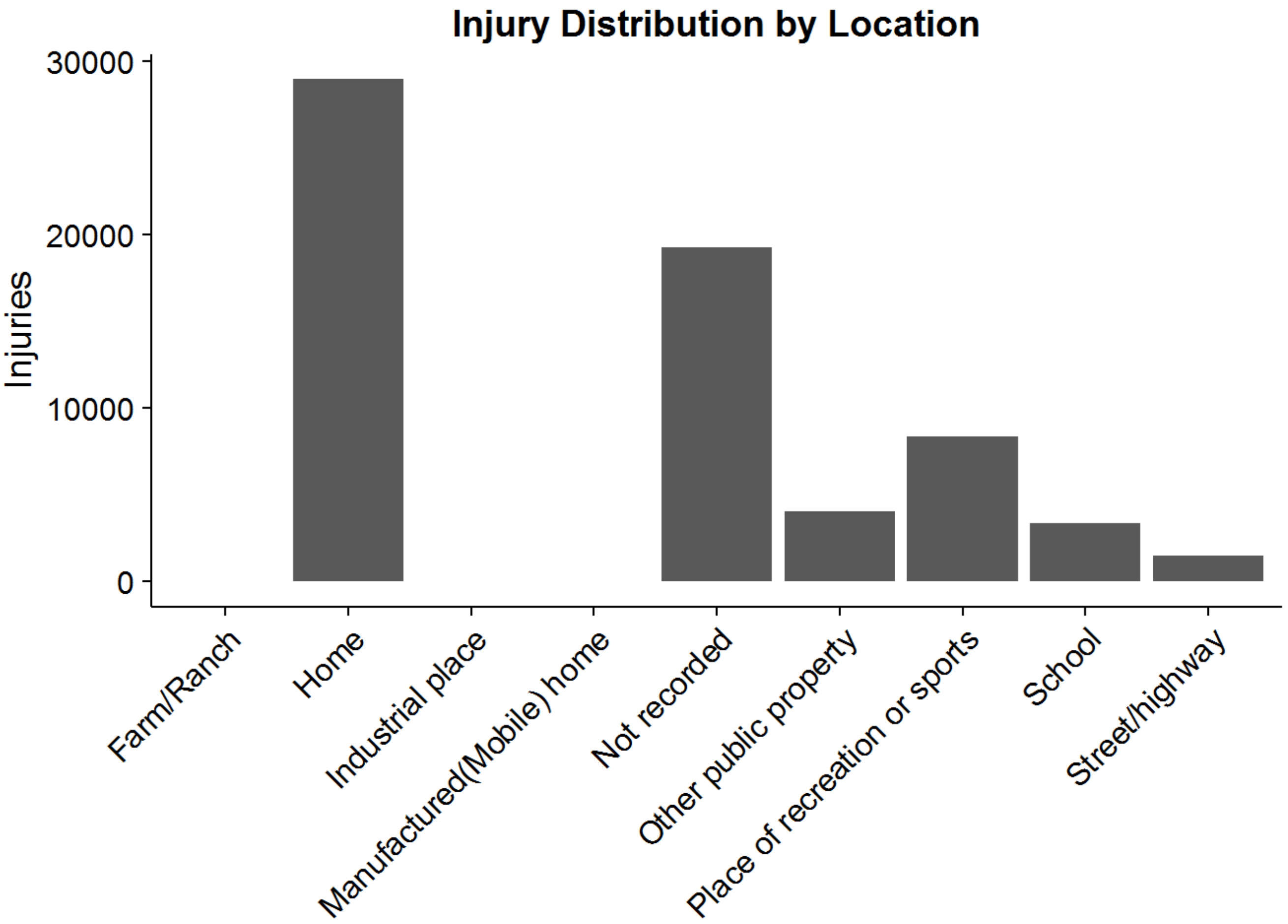


From the scatterplot we could see that injuries will decrease if age increase, it is almost a linear relatinoship given some outliers, there is a hill from **age 10** to **age 18**,which indicates that children in this age range are more active and risky on consumer product electric injuries.

Location?

We only have location code in dataset, but we could find specific location information in Manual, let’s explore:

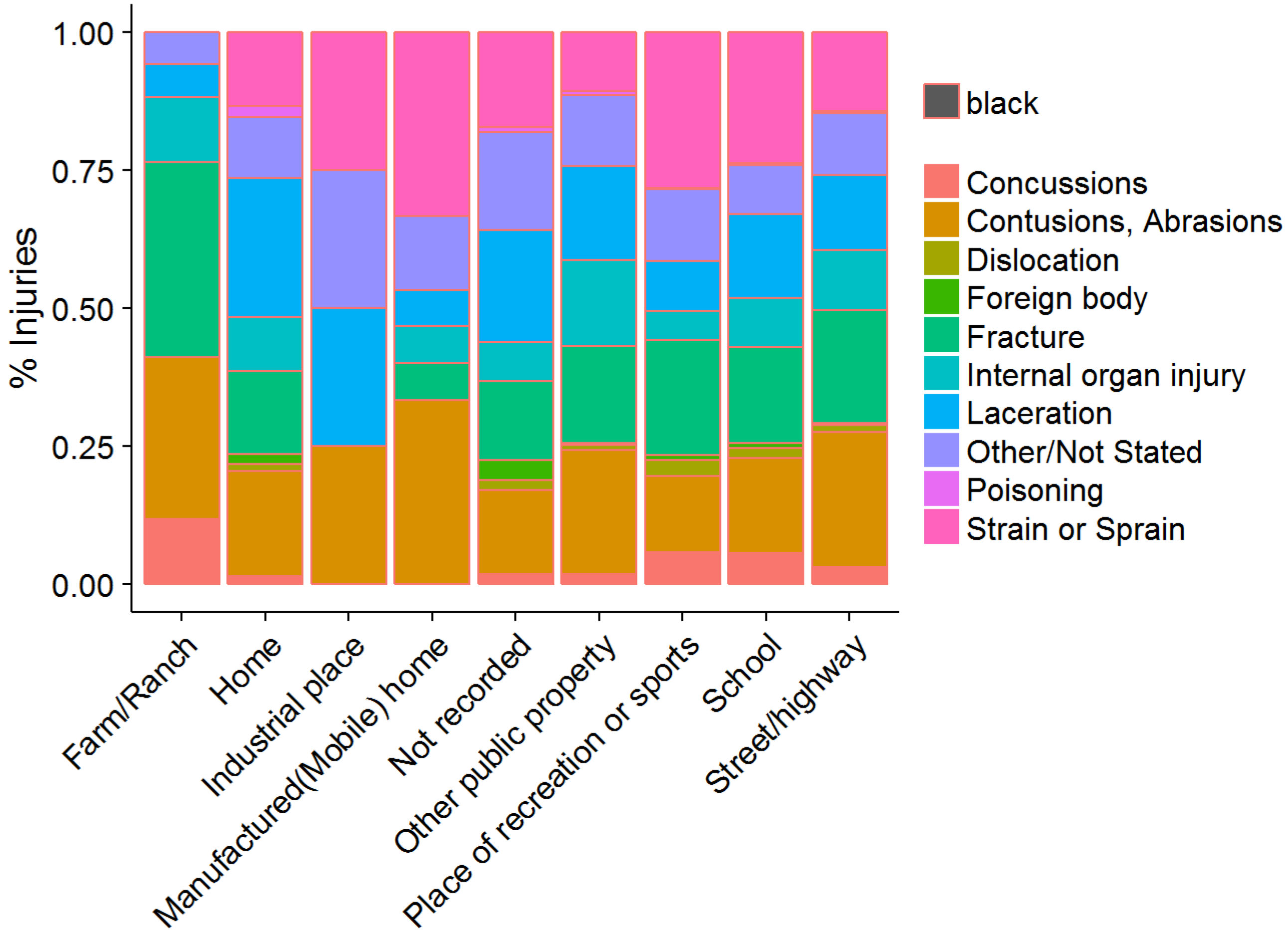
```
# Make location reference table first:
location <- c(1,2,4,5,6,7,8,9,0)
locationValue <- c('Home','Farm/Ranch','Street/highway','Other public property','Manufactured(Mobile) home','Industrial place','School','Place of recreation or sports','Not recorded')
location_lookup <- data.frame(location,locationValue)

# Add location information to main dataset
neiss <- left_join(neiss,location_lookup,by = c("location" = "location"))

# Let's find out where is the most dangerous place
ggplot(data=neiss, aes(locationValue)) +
   geom_bar() +
   labs(title="Injury Distribution by Location") +
  theme(axis.title.x = element_blank(), legend.title=element_blank(),
              axis.text.x= element_text(angle=45, hjust = 1)) +
   labs(x="Location", y="Injuries")
```



It seems like `home` is the most “dangerous” place, but it could due to we spend most time at home. `place of recreationor sports` and `school` are also on top 3.

Does location have impact on type of injuries?

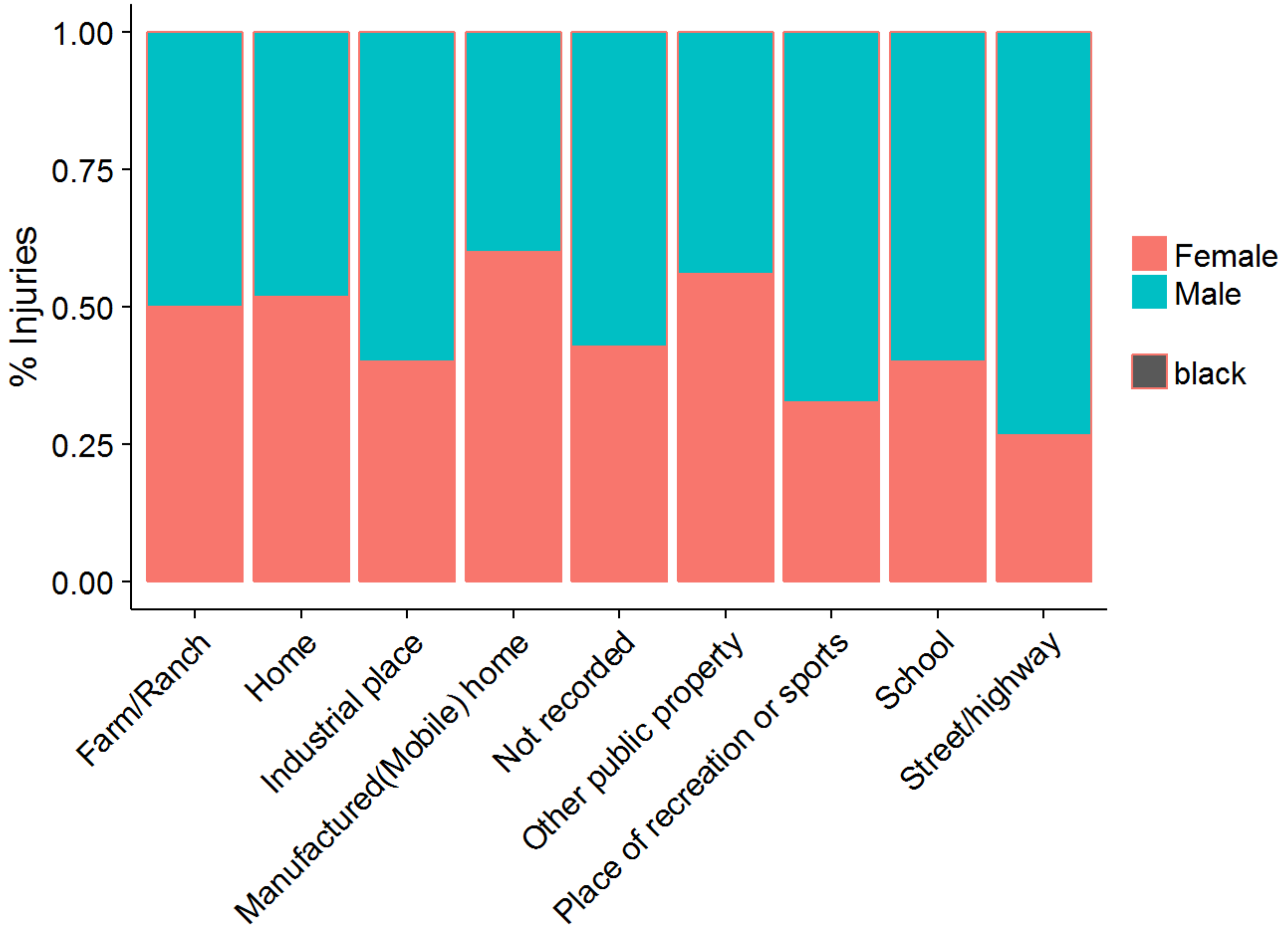
```
# Count Diagnosis
Diagnosis_count <- count(neiss,Diagnosis)

# Change column name
colnames(Diagnosis_count)[2] <- "count"

# Use most frequent diagnosis to find location differences
neiss_freq_Diagnosis <- subset(neiss, Diagnosis %in% arrange(Diagnosis_count,desc(count))
[1:10,]$Diagnosis)
ggplot(data =neiss_freq_Diagnosis, aes(x = locationValue, col="black",fill = Diagnosis)) + geom_
bar(aes(fill = Diagnosis), position = 'fill') +  theme(axis.title.x = element_blank(), legend.ti
tle=element_blank(),
              axis.text.x= element_text(angle=45, hjust = 1)) + labs(x = "location",y = "% Injur
ies")
```



From the chart we could see more `Fracture` at a `Farm/Ranch`; More `Strain` or `Contusions` at a `Manufactured (Mobile) home`; Also be careful to sharp object, especially at `home` since 1/4 of home injuries are coming from `Laceration`.

Does location have gender differences in terms of injuries?

```
ggplot(data =neiss, aes(x = locationValue, col="black",fill = sex)) + geom_bar(aes(fill = sex),
position = 'fill') +  theme(axis.title.x = element_blank(), legend.title=element_blank(),
              axis.text.x= element_text(angle=45, hjust = 1)) + labs(x = "location",y = "% Injur
ies")
```

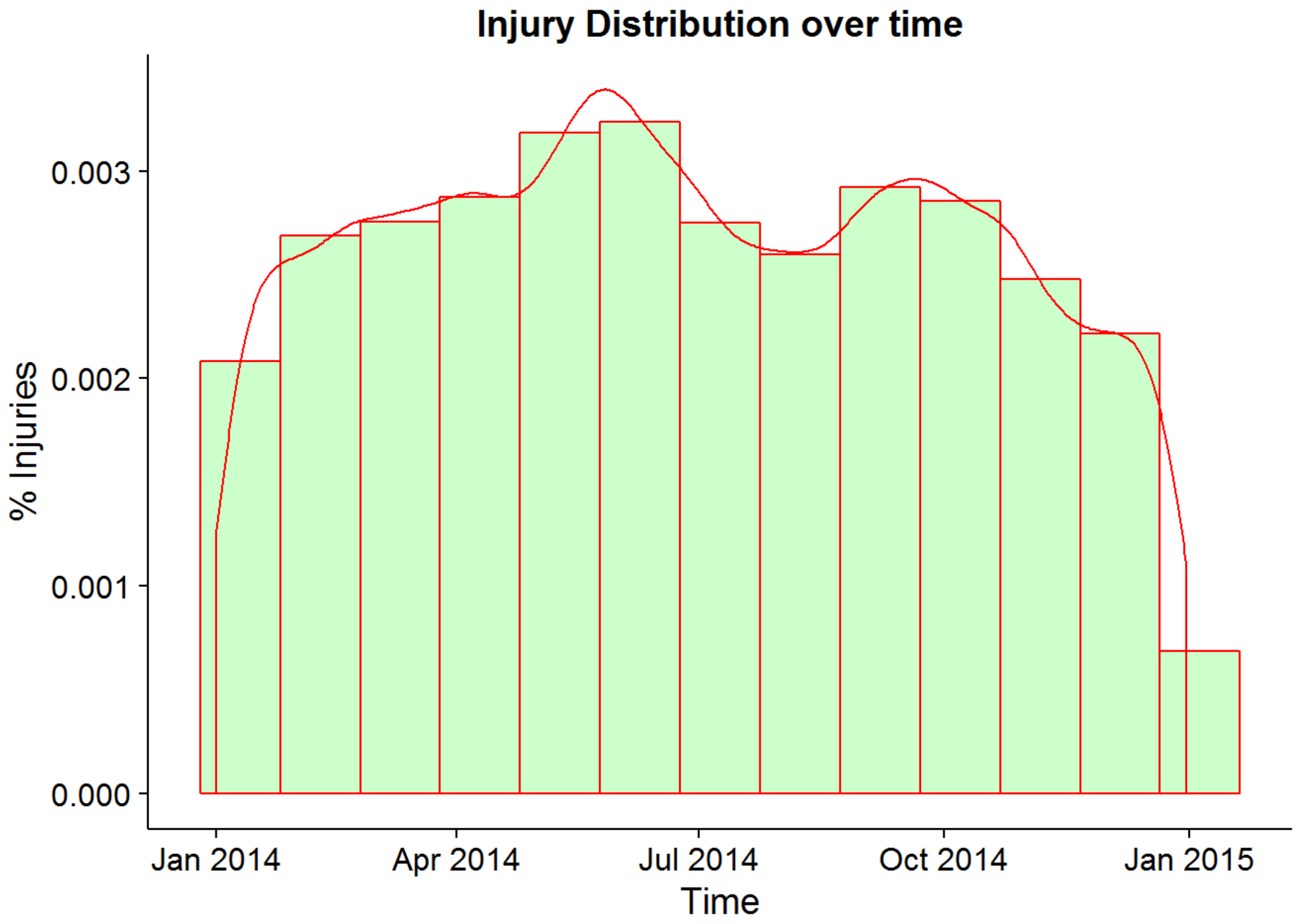


We do see gender differences here: `Street/highway` , `Place of recreation or sports` and `industrial place` do have more **male** injuries while `Manufactured (Mobile) home` and `Other public property` have more **female** injuries. Which make sense since there are more male sports fans and workers.

Does these injuries time sensitive?

Let’s take a look at injuries occured over time

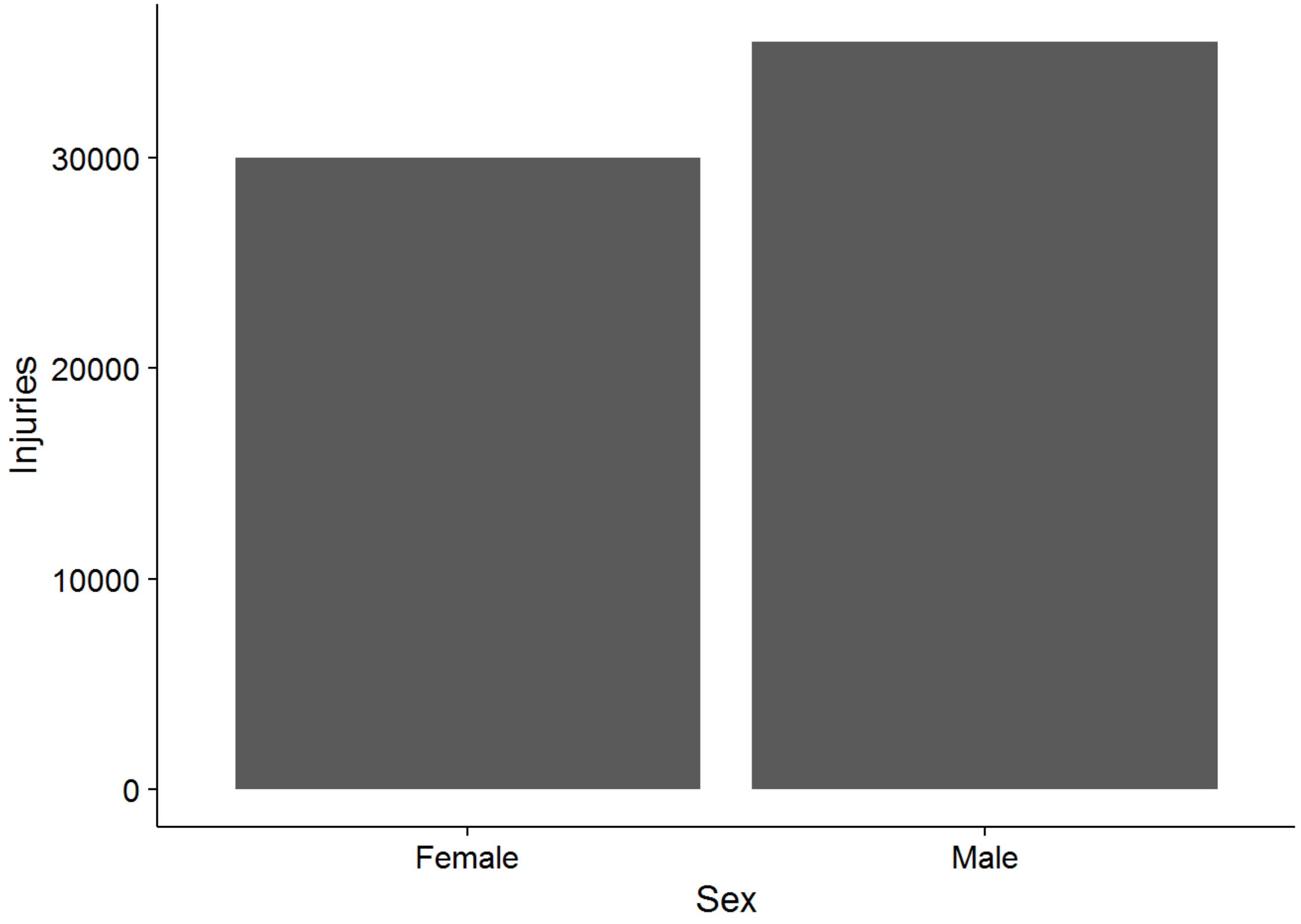
```
neiss$trmt_date <- as.Date(neiss$trmt_date, format = "%m/%d/%y")

ggplot(data=neiss, aes(neiss$trmt_date)) +
   geom_histogram(aes(y =..density..),binwidth = 30,
                  col="red",
                  fill="green",
                  alpha = .2) +  geom_density(col=2) +
   labs(title="Injury Distribution over time") +
   labs(x="Time", y="% Injuries")
```



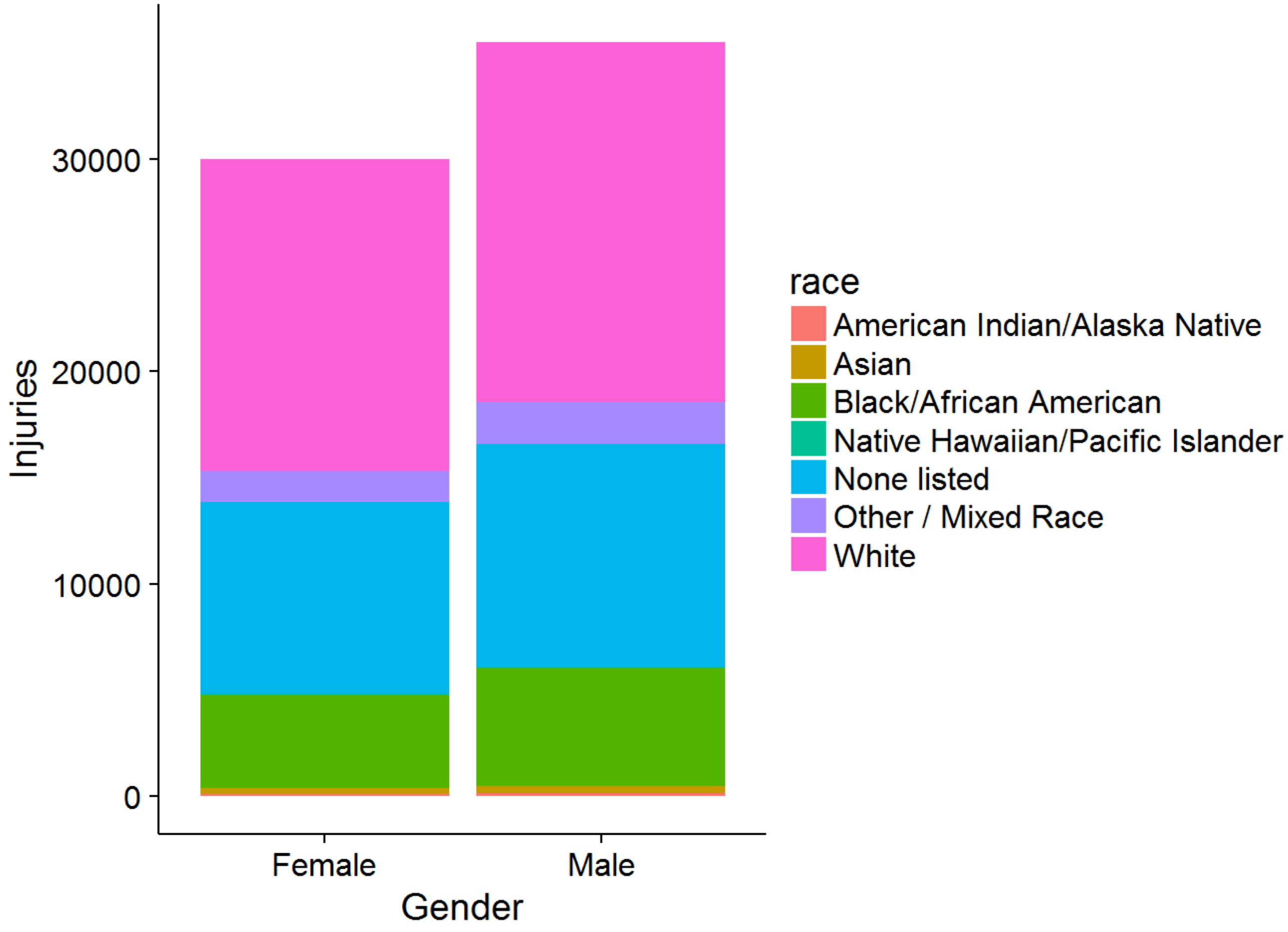
It look like that there are some seasonalities in the data, **spring** and **autumn** will have more injuries compare to summer and winter, it might due to temperature changes in **spring** and **autumn** are more dramastic.

How about gender?

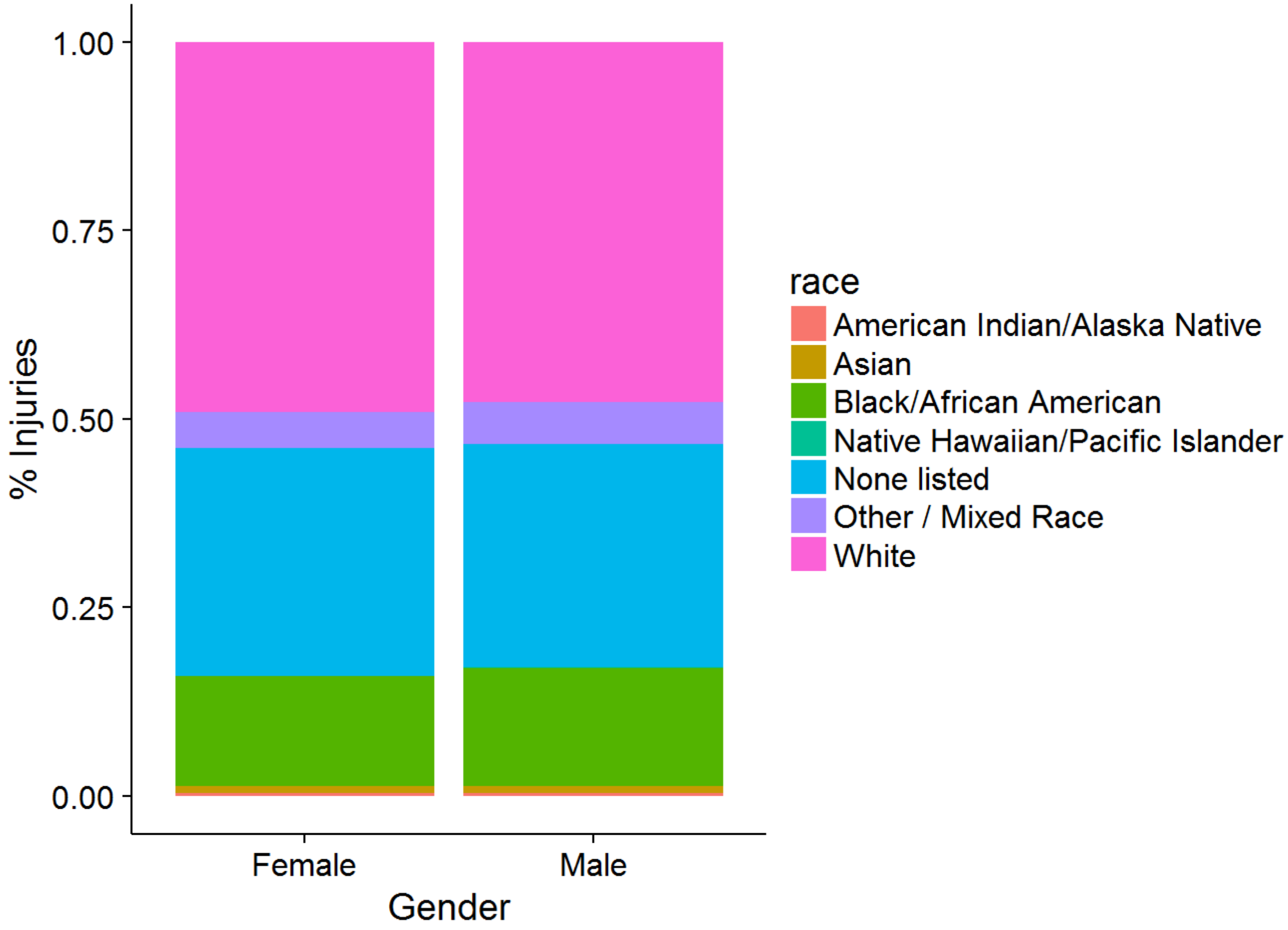
```
ggplot(data =neiss, aes(x = sex)) +
    geom_bar()+labs(x = "Sex",y = "Injuries")
```



```
ggplot(data =neiss, aes(x = sex, fill = race)) +
   geom_bar()+labs(x = "Gender",y = "Injuries")
```



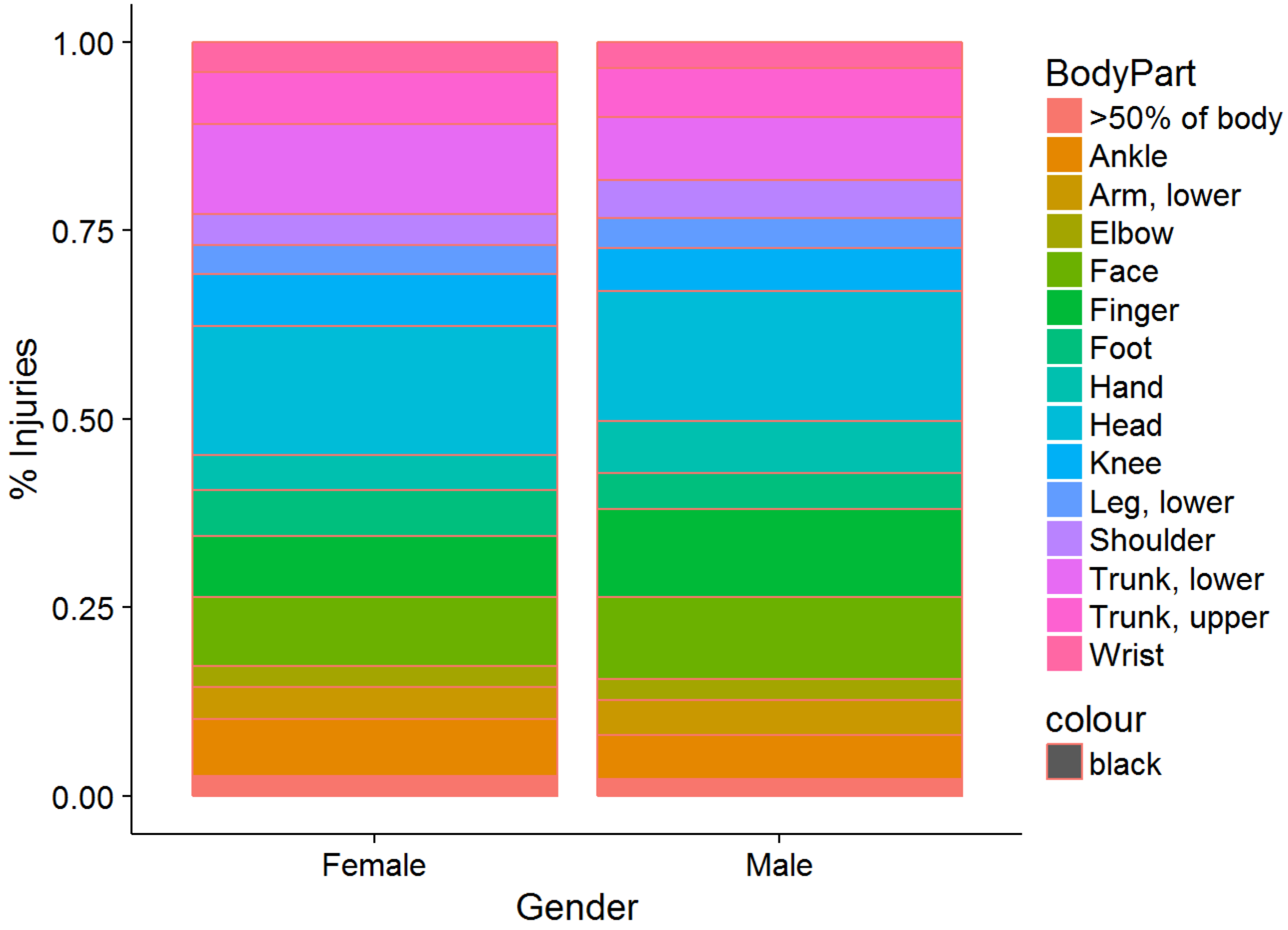
```
 ggplot(data =neiss, aes(x = sex, fill = race)) + geom_bar(aes(fill = race), position = 'fill')
+labs(x = "Gender",y = "% Injuries")
```



Basically `Male` is more likely get injuries compare to `Female` (1/6 chance higher) which also make sence since `Male` are more bold and careless than `Female` which will cause more potential injuries. The race distribution in 2 genders are pretty even.

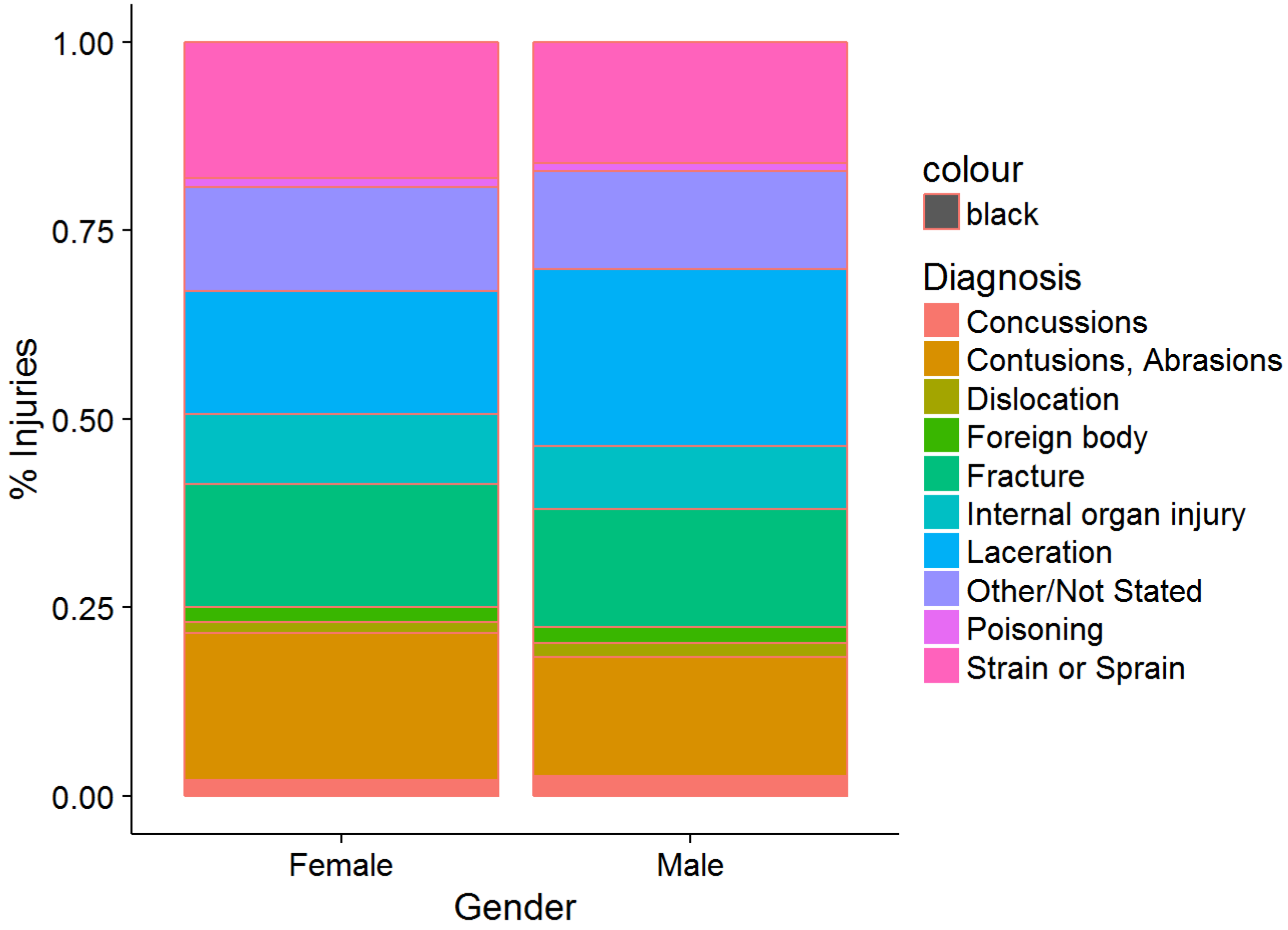
Let’s go deeper into other variables:

```
# Use most frequent body parts to find gender differences
neiss_freq_bodyPart <- subset(neiss, BodyPart %in% arrange(bodyPart_count,desc(count))[1:15,]$BodyPart)

ggplot(data =neiss_freq_bodyPart, aes(x = sex, col="black",fill = BodyPart)) + geom_bar(aes(fill = BodyPart), position = 'fill') +labs(x = "Gender",y = "% Injuries")
```



From the bar chart above we could find that **female** tend to have more injuries in their **lower trunk** (`Trunk,lower` , `Ankle`), while **male** tend to have more injuries in their **upper trunk** (`Face` , `Finger`)

```
ggplot(data =neiss_freq_Diagnosis, aes(x = sex, col="black",fill = Diagnosis)) + geom_bar(aes(fi
ll = Diagnosis), position = 'fill') +labs(x = "Gender",y = "% Injuries")
```

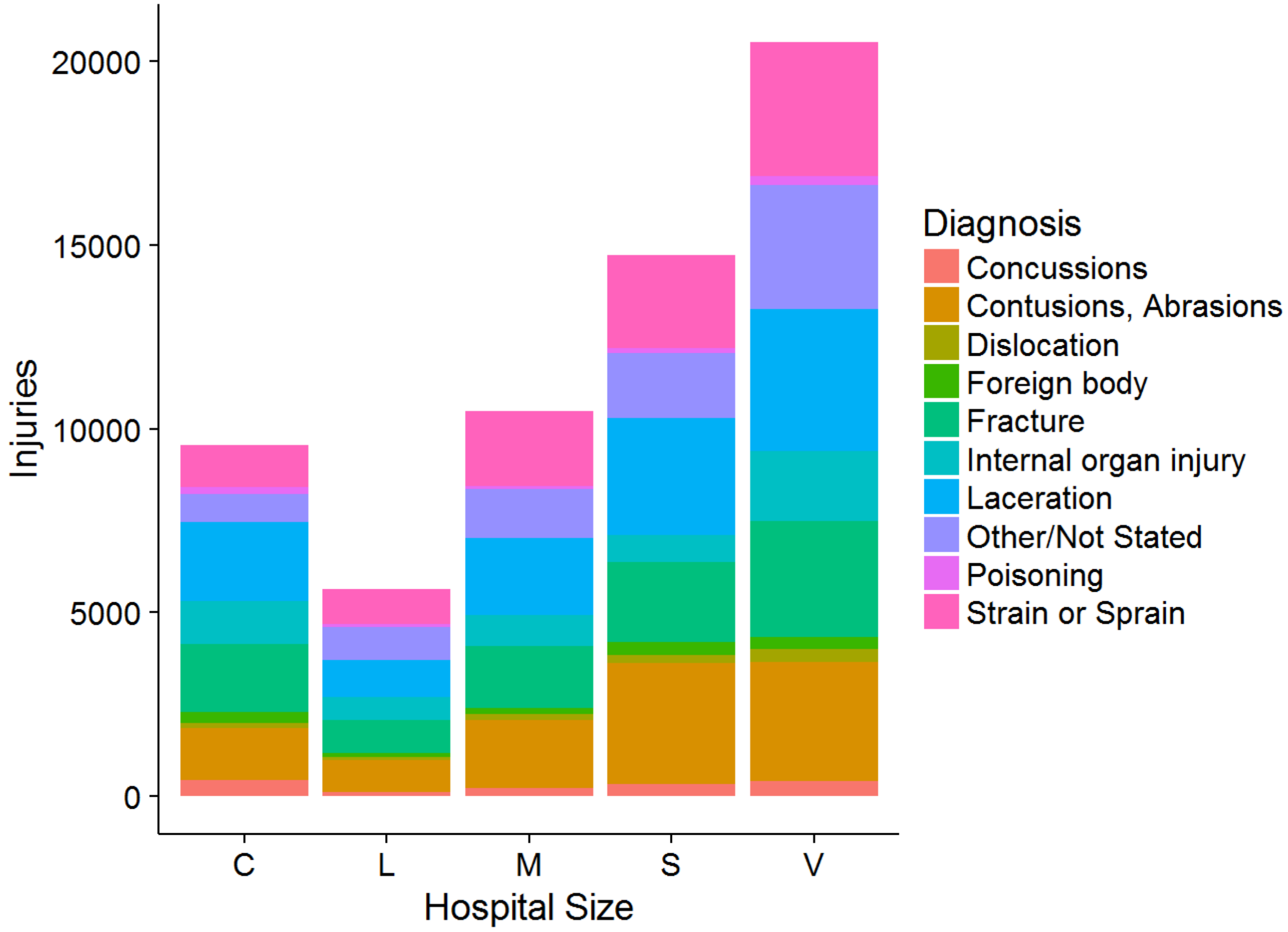


In terms of Diagnosis, **female** will have more `Contusions` and `Strain` while **male** will have more `Laceration`

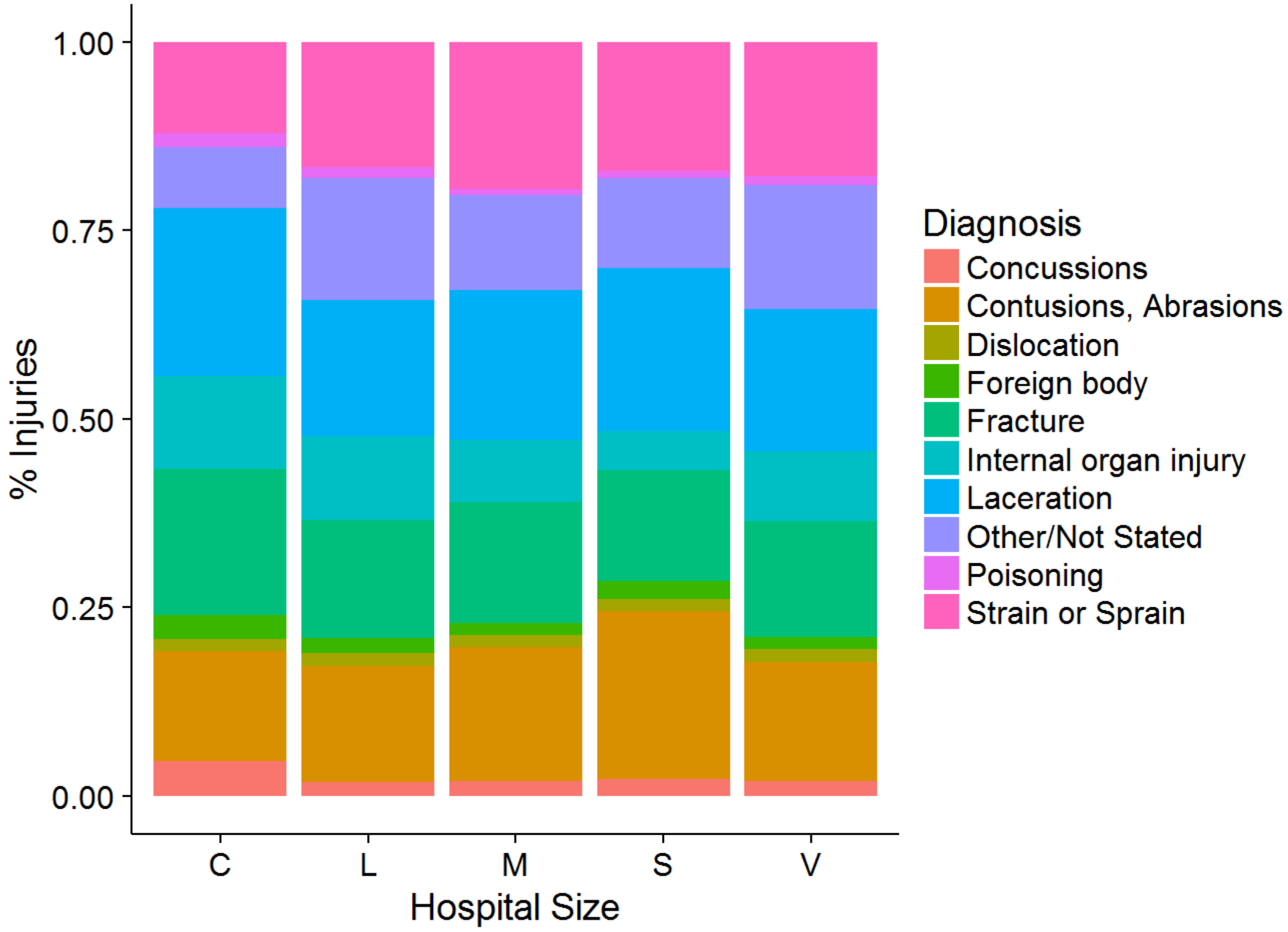
Hospital size?

Variable `stratum` record the size of hospital that the injury will be treated:
`S` is Small hospital
`M` is Medium hospital
`L` is Large hospital
`V` is Very large hospital
`C` is Chilren's hospital

```
ggplot(data =neiss_freq_Diagnosis, aes(x = stratum, fill = Diagnosis)) +
    geom_bar()+labs(x = "Hospital Size",y = "Injuries")
```



```
ggplot(data =neiss_freq_Diagnosis, aes(x = stratum, fill = Diagnosis)) + geom_bar(aes(fill = Dia
gnosis), position = 'fill') +labs(x = "Hospital Size",y = "% Injuries")
```



From the graph that we could see people are tend to go very large hospital. Specifically, `Contusions` tend to be treated in small hospital since it only require simple treatment and equipments while `Internal organ injury` always require complex medical procedure, even surgery, which only larger hospitals could satisfy the requirement.

Any dangerous products?

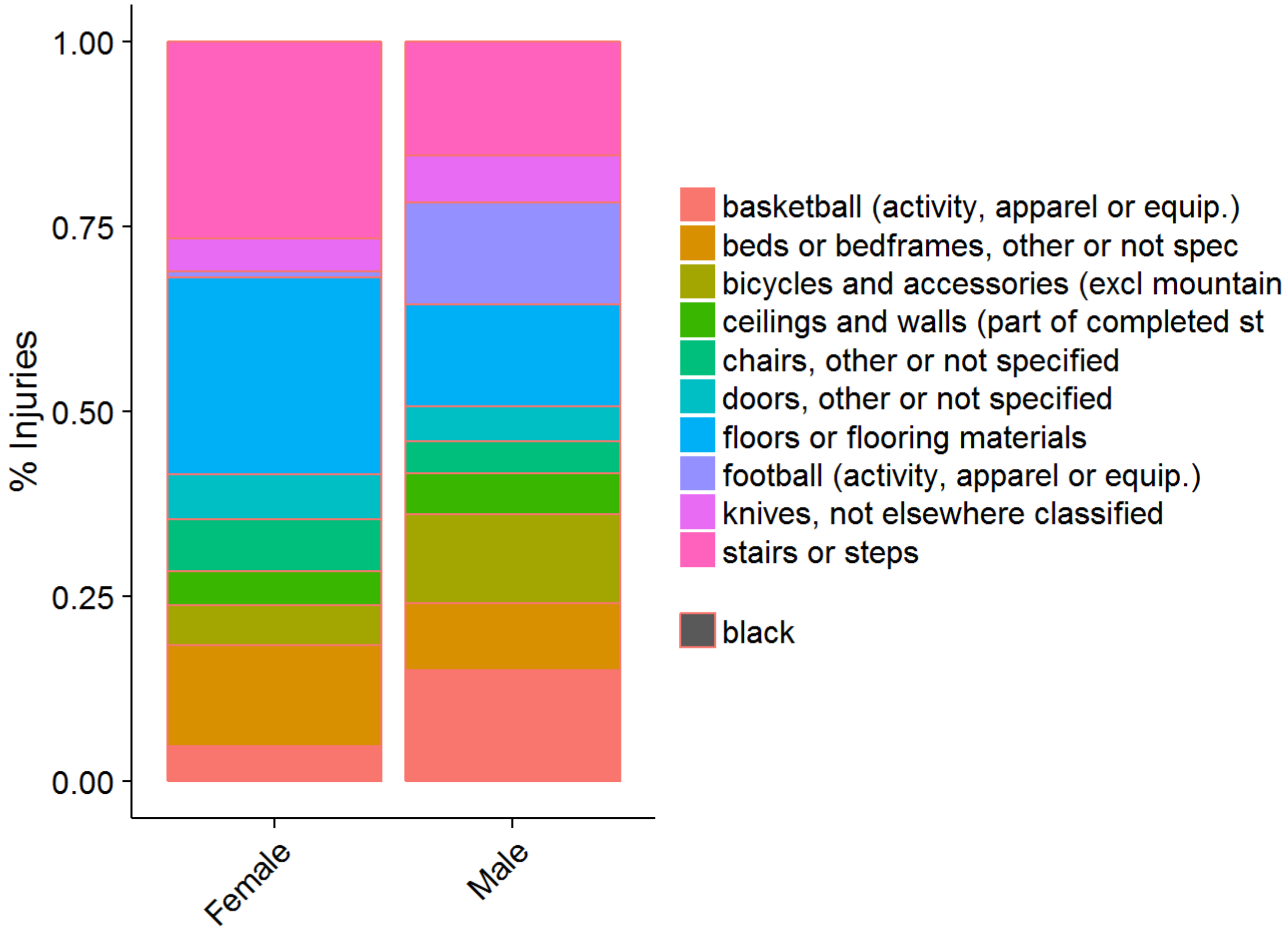
```
load("C:/Users/bowei.zhang/Desktop/ME/Career/Analytics_Example/NEISS/data/products.rda")
neiss <- left_join(neiss,products,by = c("prod1" = "code"))

# Let's find out top 10 frequent products that involved injuries
Product_count <- count(neiss,title)
colnames(Product_count)[2] <- "count"
arrange(Product_count,desc(count))[1:10,]
```

```
## # A tibble: 10 x 2
##                                       title count
##                                       <chr> <int>
## 1                            stairs or steps  5382
## 2                 floors or flooring materials  5162
## 3       beds or bedframes, other or not spec  2923
## 4   basketball (activity, apparel or equip.)  2680
## 5    bicycles and accessories (excl mountain  2343
## 6     football (activity, apparel or equip.)  2036
## 7             chairs, other or not specified  1475
## 8           knives, not elsewhere classified  1427
## 9              doors, other or not specified  1375
## 10 ceilings and walls (part of completed st  1326
```

It is not surprise that most dangerous products are most common ones: `stairs` , `floors` , `bed` , `basketball` , `bicycle` , `chairs` , `knives`

Does product have gender differences?

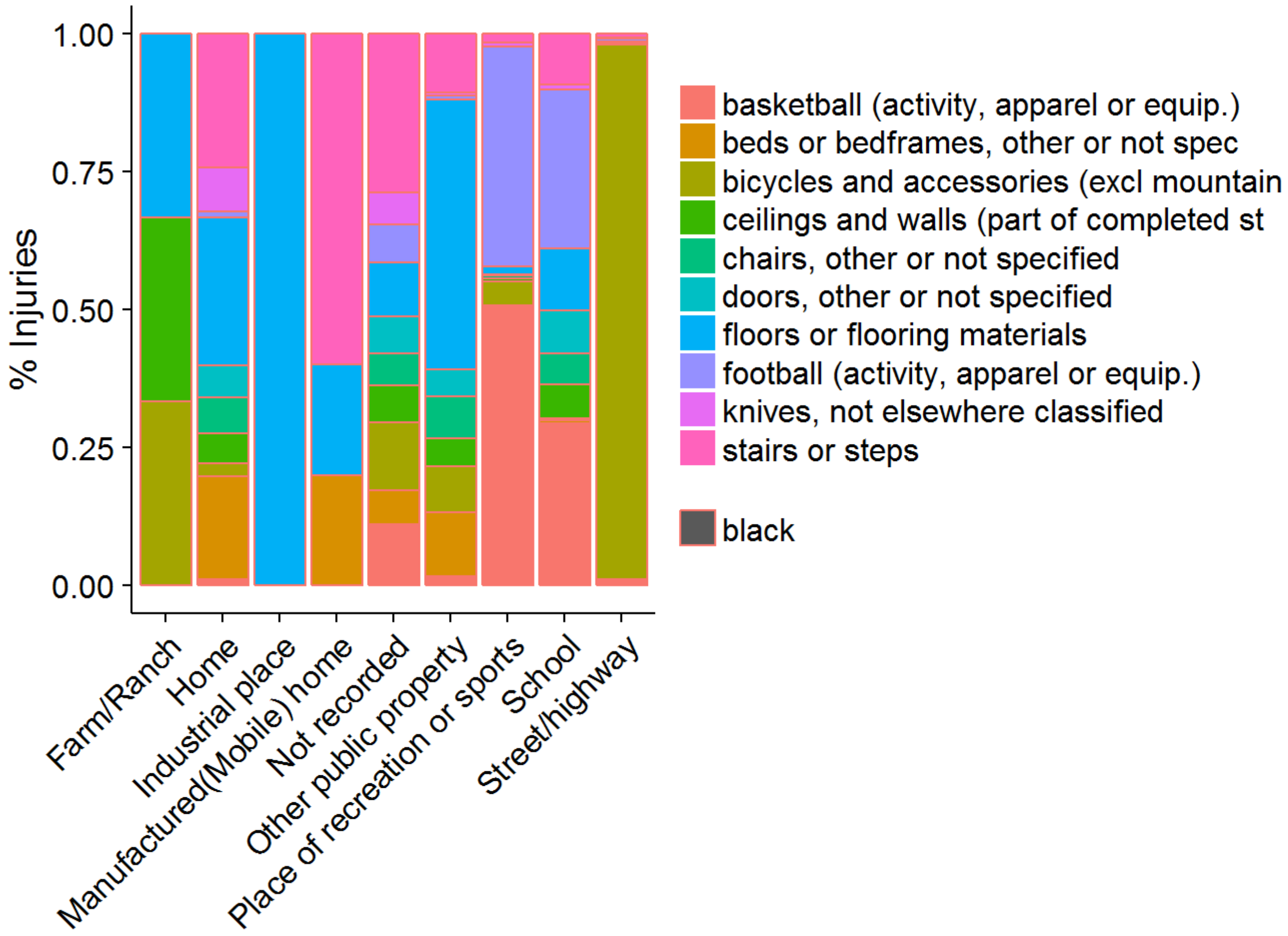
```
# Use most frequent products to find gender differences
neiss_freq_product <- subset(neiss, title %in% arrange(Product_count,desc(count))[1:10,]$title)
ggplot(data =neiss_freq_product, aes(x = sex, col="black",fill = title)) + geom_bar(aes(fill = t
itle), position = 'fill') +  theme(axis.title.x = element_blank(), legend.title=element_blank(),
              axis.text.x= element_text(angle=45, hjust = 1)) + labs(x = "Gender",y = "% Injurie
s")
```



Gender do have huge impact on products!
male has more injuries involved **sports** (`basketball` , `football` and `bicycles`) while **female** is more involved with **furnitures** (`bed` , `stairs`)

Does product have location differences?

```
# Use most frequent products to find location differences

ggplot(data =neiss_freq_product, aes(x = locationValue, col="black",fill = title)) + geom_bar(aes(fill = title), position = 'fill') +  theme(axis.title.x = element_blank(), legend.title=element_blank(),
               axis.text.x= element_text(angle=45, hjust = 1)) + labs(x = "Location",y = "% Injuries")
```



Of course it has!
More `basketball` and `football` injuries in `Place of recreation and sports` and `school` as well;
Most injuries happened in `street/highway` are related to `bicycles` or similar;
Most injuries happened in `industrial place` are related to `floors or flooring materials` or similar.