MA 415 Midterm Project

Emma Brown

Due: March 22, 2017

MA 415 Midterm Project: OSHA Data

This project cleans and prepares the OSHA datset collected by NICAR to for analysis to report on the most dangerous places to work in Massachusetts.

The necessary libraries were loaded:

```
require(data.table)
```

```
## Loading required package: data.table
```

```
require(foreign)
```

```
## Loading required package: foreign
```

```
require(dplyr)
```

```
## Loading required package: dplyr

## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(magrittr)
```

```
## Loading required package: magrittr
```

```
require(tidyr)
```

```
## Loading required package: tidyr

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```r
require(sqldf)
```

```
## Loading required package: sqldf
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(xtable)
```

```
## Loading required package: xtable
```

```r
require(knitr)
```

```
## Loading required package: knitr
```

Reading the database files we want:

```r
accid <- data.table(read.dbf("accid.dbf"))
acc <- data.table(read.dbf("lookups/acc.dbf"))
osha <- data.table(read.dbf("osha.dbf"))
scc <- data.table(read.dbf("lookups/scc.dbf"))
scc <- filter(scc, STATE=="MA")
occ <- data.table(read.dbf("lookups/occ.dbf"))
hzs <- data.table(read.dbf("lookups/hzs.dbf"))
# How many unique inspection numbers are there?
length(unique(accid$ACTIVITYNO))
```

```
## [1] 1570
```

```r
# We will group them by number to get all unique inspections
accid <- sqldf("select *
               from accid
               group by ACTIVITYNO")
```

```
## Loading required package: tcltk
```

```
## Warning: Quoted identifiers should have class SQL, use DBI::SQL() if the
## caller performs the quoting.
```

Let's join some data:

```r
# Joining all occupations into the accid table
accid <- data.table(accid)
setkey(accid, OCC_CODE)
setkey(occ, CODE)
leftCols <- colnames(accid)
leftCols <- sub("OCC_CODE", "OCCUPATION", leftCols)
accid <- occ[accid][, leftCols, with=FALSE]

# But we don't care about occupations not reported
accid <- sqldf("select *
               from accid
               where OCCUPATION != 'OCCUPATION NOT REPORTED'")
```

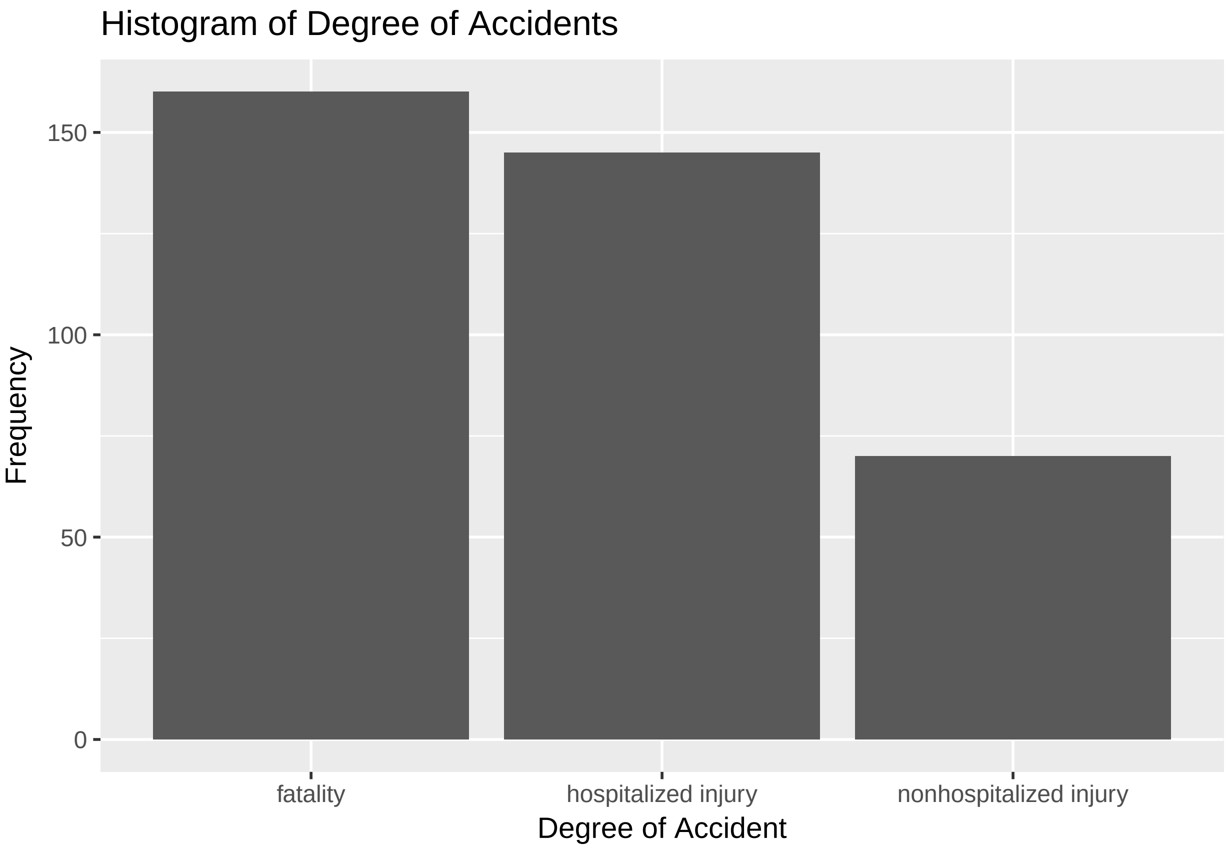
```
# Join hazardous substances into accid table
accid <- data.table(accid)
setkey(accid, HAZSUB)
hzs <- data.table(hzs)
setkey(hzs, CODE)
leftCols <- colnames(accid)
leftCols <- sub("HAZSUB", "TEXT", leftCols)
accid <- hzs[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="TEXT"] <- "HAZ-SUB"

# Join nature of injury into accid table
nature <- sqldf("select *
                from acc
                where CATEGORY == 'NATUR-INJ'")
setkey(accid, NATURE)
nature <- data.table(nature)
setkey(nature, CODE)
leftCols <- colnames(accid)
leftCols <- sub("NATURE", "VALUE", leftCols)
accid <- nature[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "NATURE-INJURY"

# Body parts
bodypart <- sqldf("select *
                from acc
                where CATEGORY == 'PART-BODY'")
accid <- data.table(accid)
setkey(accid, BODYPART)
bodypart <- data.table(bodypart)
setkey(bodypart, CODE)
leftCols <- colnames(accid)
leftCols <- sub("BODYPART", "VALUE", leftCols)
accid <- bodypart[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "BODY-PART"

# Event type
event <- sqldf("select *
                from acc
               where CATEGORY == 'EVENT-TYP'")
accid <- data.table(accid)
setkey(accid, EVENT)
event <- data.table(event)
setkey(event, CODE)
leftCols <- colnames(accid)
leftCols <- sub("EVENT", "VALUE", leftCols)
accid <- event[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "EVENT"

# Environment factor
environ <- sqldf("select *
                from acc
                  where CATEGORY == 'ENVIR-FAC'")
accid <- data.table(accid)
```

```
setkey(accid, ENVIRON)
environ <- data.table(environ)
setkey(environ, CODE)
leftCols <- colnames(accid)
leftCols <- sub("ENVIRON", "VALUE", leftCols)
accid <- environ[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "ENVIRON"

# Human Factor
human <- sqldf("select *
                from acc
                  where CATEGORY == 'HUMAN-FAC'")
accid <- data.table(accid)
setkey(accid, HUMAN)
human <- data.table(human)
setkey(human, CODE)
leftCols <- colnames(accid)
leftCols <- sub("HUMAN", "VALUE", leftCols)
accid <- human[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "HUMAN"

# Source injury
source <- sqldf("select *
                from acc
                  where CATEGORY == 'SOURC-INJ'")
accid <- data.table(accid)
setkey(accid, SOURCE)
source <- data.table(source)
setkey(source, CODE)
leftCols <- colnames(accid)
leftCols <- sub("SOURCE", "VALUE", leftCols)
accid <- source[accid][, leftCols, with=FALSE]
colnames(accid)[colnames(accid)=="VALUE"] <- "SOURCE-INJ"

# What kind of task were they doing?
levels(accid$TASK) <- c("N/A",
                        "Assigned",
                        "Not Assigned")

# Now let's fix those DEGREE categories
levels(accid$DEGREE) <- c("N/A",
                          "fatality",
                          "hospitalized injury",
                          "nonhospitalized injury")

# But we're only interested in the injuries
accid <- sqldf("select *
               from accid
               where DEGREE != 'N/A'")
```

Now that’s ready to work with Let’s look at some of this data:

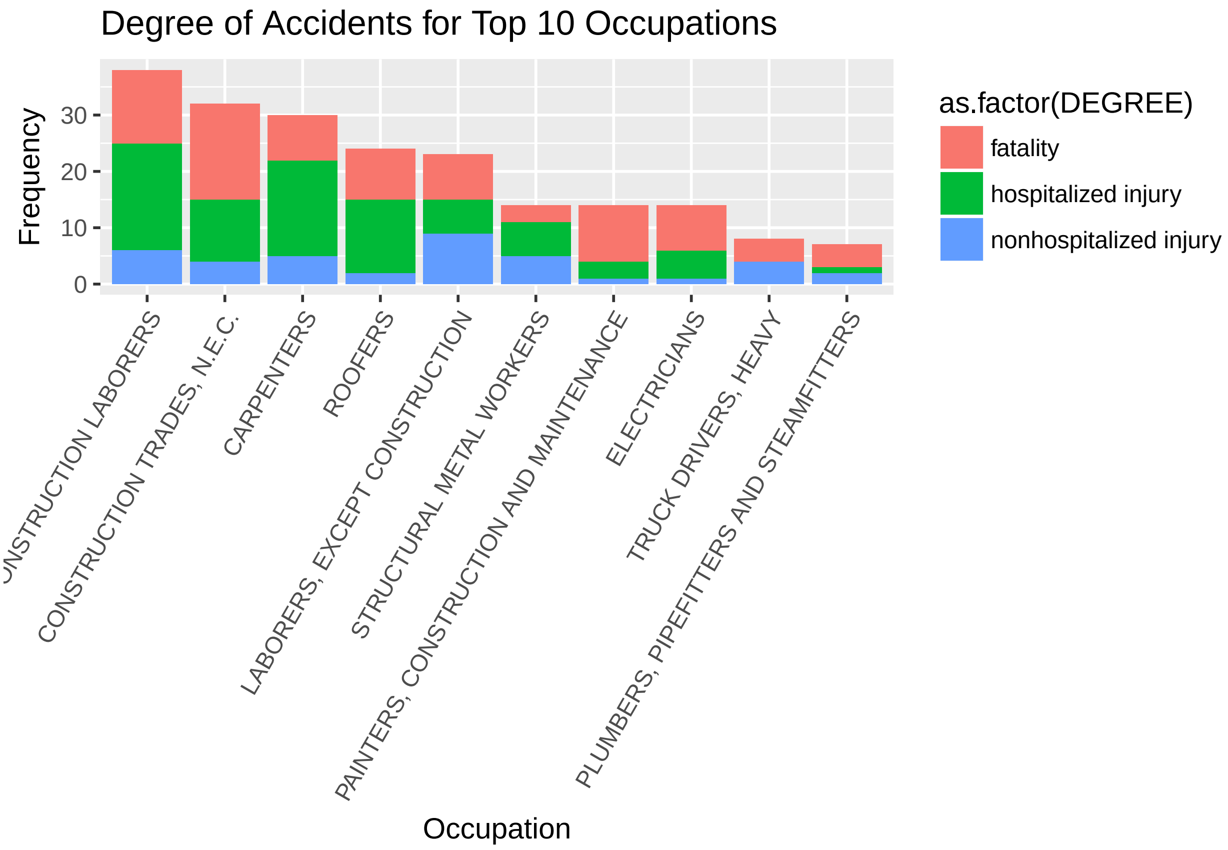
```
# Let's look at some of this data
# Histogram of Accidents by Degree
```

```r
x <- table(accid$DEGREE)
p <- qplot(accid$DEGREE)
p + scale_fill_brewer(palette="Blues") +
  xlab("Degree of Accident") +
  ylab("Frequency") +
  ggtitle("Histogram of Degree of Accidents")
```



As shown in the histogram of Degree of Accidents, most of the reported accidents are fatal. Let's see what occupations have the most accidents:

```r
# What are the most dangerous occupations?
table <- table(accid$OCCUPATION)
sorted <- sort(table)
top_10 <- tail(names(sorted), 10)
subset <- subset(accid, OCCUPATION %in% top_10)
subset$OCCUPATION <- factor(subset$OCCUPATION, levels=rev(top_10))
legend <- "Degree of Accident"
ggplot(subset, aes(x=OCCUPATION, fill=as.factor(DEGREE))) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  labs(title = "Degree of Accidents for Top 10 Occupations",
       x = "Occupation", y = "Frequency", col= "Degree")
```



The histogram shows Construction workers have the riskiest job.

Now onto Osha!

```
osha <- data.table(read.dbf("osha.dbf"))
osha <- sqldf("select ACTIVITYNO, ESTABNAME,
              SITEADD, OPENDATE, SITECNTY, SITECITY, TOTALVIOLS
              from osha")

# Again group by ACTIVITYNO
osha <- sqldf("select *
              from osha
              group by ACTIVITYNO")
```

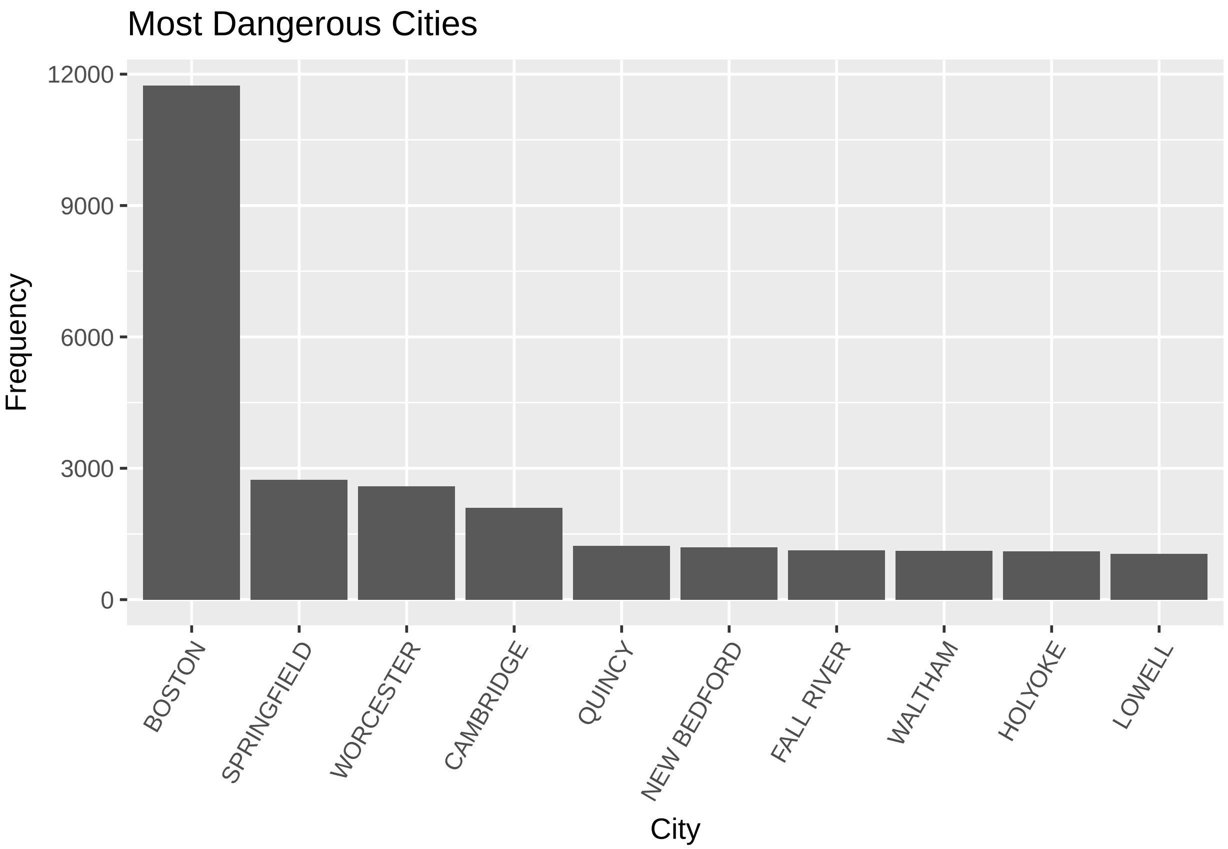
Join the city to the OSHA data

```
osha <- data.table(osha)
scc <- data.table(scc)
setkey(osha, SITECNTY)
setkey(scc, COUNTY)
setkey(osha, SITECITY)
setkey(scc, CITY)
leftCols <- colnames(osha)
leftCols <- sub("SITECITY", "NAME", leftCols)
osha <- scc[osha][, leftCols, with=FALSE]
```

Let's look at some of this…what are the top 10 most dangerous cities in Massachusetts?

```
# Most Dangerous Cities
table <- table(osha$NAME)
```

```r
sorted <- sort(table)
top_10 <- tail(names(sorted), 10)
subset <- subset(osha, NAME %in% top_10)
subset$NAME <- factor(subset$NAME, levels=rev(top_10))
ggplot(subset, aes(x=NAME)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  labs(title = "Most Dangerous Cities",
       x = "City", y = "Frequency")
```



This makes sense because Boston is the most densely populated city in Massachusetts. It would be interesting to see how this would compare by population, or perhaps on a nationwide scale? Where is the most dangerous place to work in the country?