Midterm_Final_Copy

Kirstie Turnbull

March 21, 2017

The following data have been cleaned in preparation for a statistician to investigate and analyze the safety of workplaces in the state of Massachusetts. These data were retrieved from the Occupational Safety and Health Administration (OSHA) and are from April, 2005. The original set of data included information regarding both safety records and debt/payment records. Due to the fact that debt history data are not relavent to workplace safety, that data have been removed from the cleaned data represented in this document.

The first data set, CleanAccid, includes data about workplace accidents in the state of Massachusetts. The data were originally in code and the column titles were unclear, so in the cleaning process the coded data were replaced with English words and the column titles were clarified. The SITESTATE column was also removed, as all of the data relavent to the workplace safety question of interest are from Massachusetts. For commented code, see CleanAccid.R in the MA415MidtermFolder.

```
## ------------------------------------------------------------------------
```

```
## 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
```

```
##
## Attaching package: 'tidyr'
```

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

```
## ------------------------------------------------------------------------
```

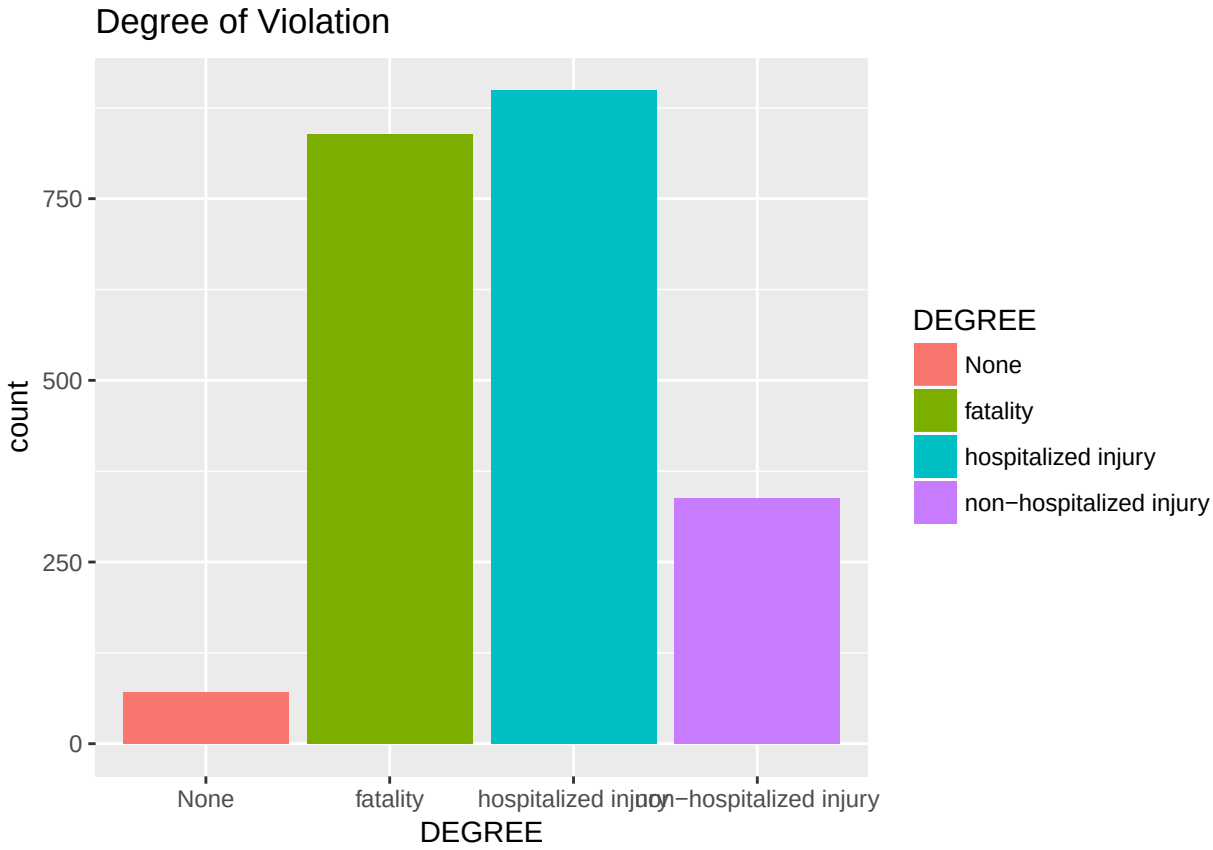
```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
```

```
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

##
##                   None               fatality     hospitalized injury
##                     71                    839                     899
## non-hospitalized injury
##                    338
```

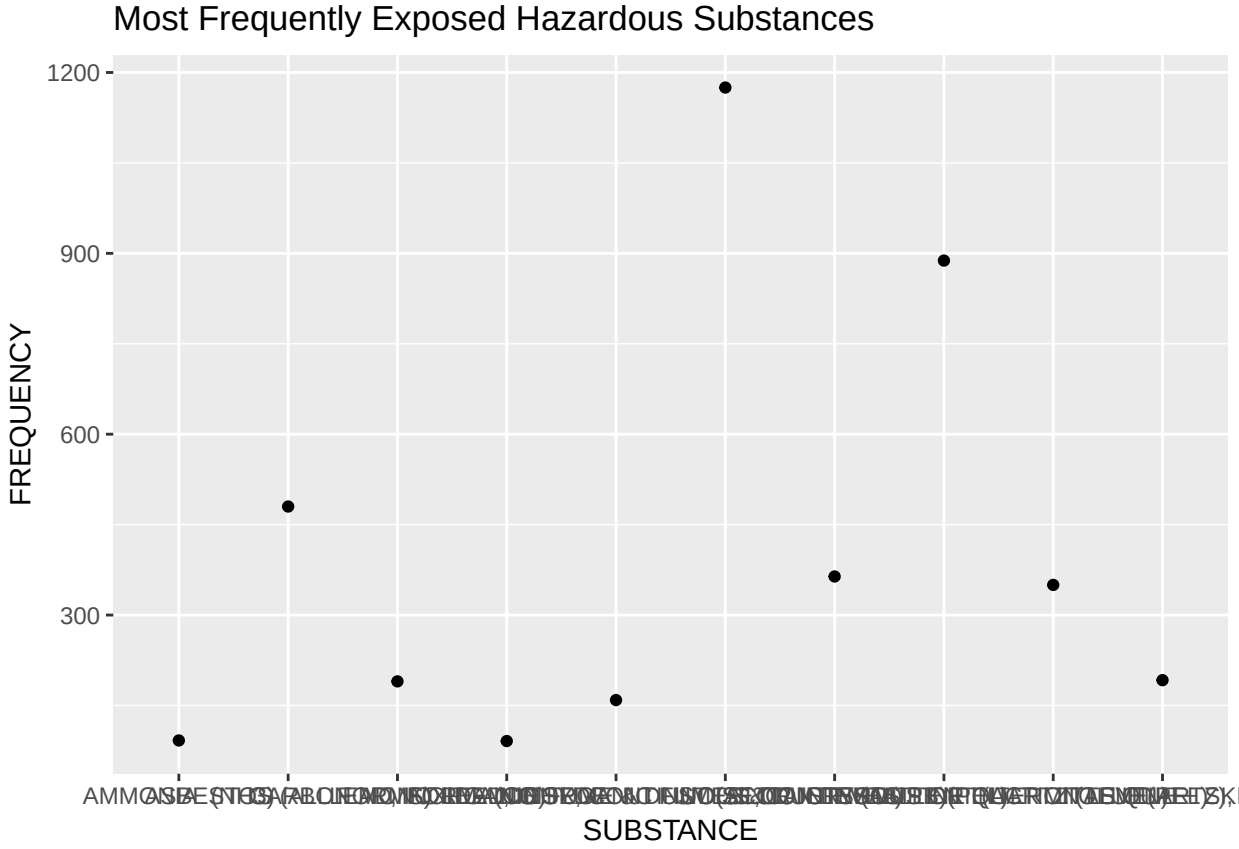


Degree of violation, as seen above in both table and plot format, is one example of interesting data from the CleanAccid dataframe that is relavent to the workplace safety question that is being investigated by statisticians.

The next data set, CleanHazsub, contains information regarding hazardous substances that employees have been to exposed to at different workplaces in Massachusetts. Again, because all of the data are from Massachusetts, the SITESTATE column was removed during cleaning. Additionally, the data were altered from code to the names of each substance. For commented code, see CleanHazsub.R in the MA415MidtermFolder.

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## [1] 363

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## [1] 217

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

##                                       SUBSTANCE FREQUENCY
## 299                             AMMONIA  (NH3)        92
## 320                          ASBESTOS (ALL FORMS)       480
```

```
## 473                               CARBON MONOXIDE (CO)       190
## 828                                       FORMALDEHYDE        91
## 998         LEAD, INORGANIC FUME & DUST (ACTION LEVEL)       159
## 999         LEAD, INORGANIC FUMES & DUSTS (AS PB)(PEL)      1175
## 1233  NOISE, CONTINUOUS OR INTERMITTENT (ACTION LEVEL)       364
## 1236             NOISE,CONTINUOUS OR INTERMITTENT(PEL)       888
## 1500 SILICA CRYSTALLINE QUARTZ (AS QUARTZ), RESP. DUST       350
## 1647                                    TOLUENE - SKIN       192
```



The most frequently exposed hazardous substances, such as lead and dangerous noise, are also relavent to workplace safety. These data are included in the final cleaned CleanHazsub dataframe.

The data in the CleanHistory data set describes information about past penalties or failures-to-abate that workplaces in Massachusetts faced because of safety violations. During the cleaning process, column titles were clarified and any empty columns and the SITESTATE column were removed. This leaves data regarding events, penalties, and the severity of the violation. These data were also originally encoded and when the data were cleaned, the code was changed into English words. For commented code, please see CleanHistory.R in the MA415MidtermFolder.

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:plyr':
##
##     here

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following object is masked from 'package:base':
##
##     date
```

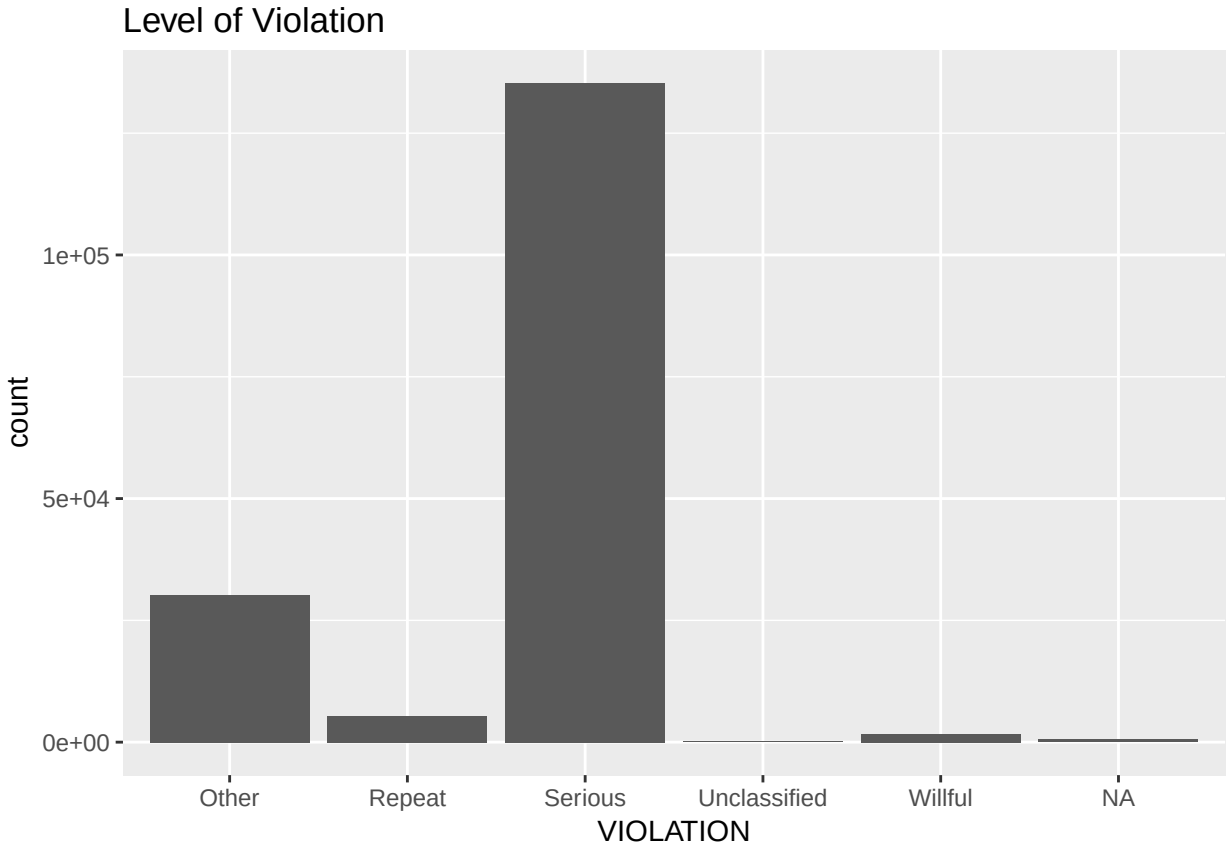
```
## Warning: 490 failed to parse.
```

```
## Warning: 512 failed to parse.
```

```
## The following `from` values were not present in `x`: Y, S, N, L, 1, 0
```

```
##
##        Other       Repeat      Serious Unclassified      Willful
##        30187         5400       135331          168         1704
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Level of Violation, as seen in the table and plot above, is another example from the cleaned history data that is relavent to the workplace safety question.

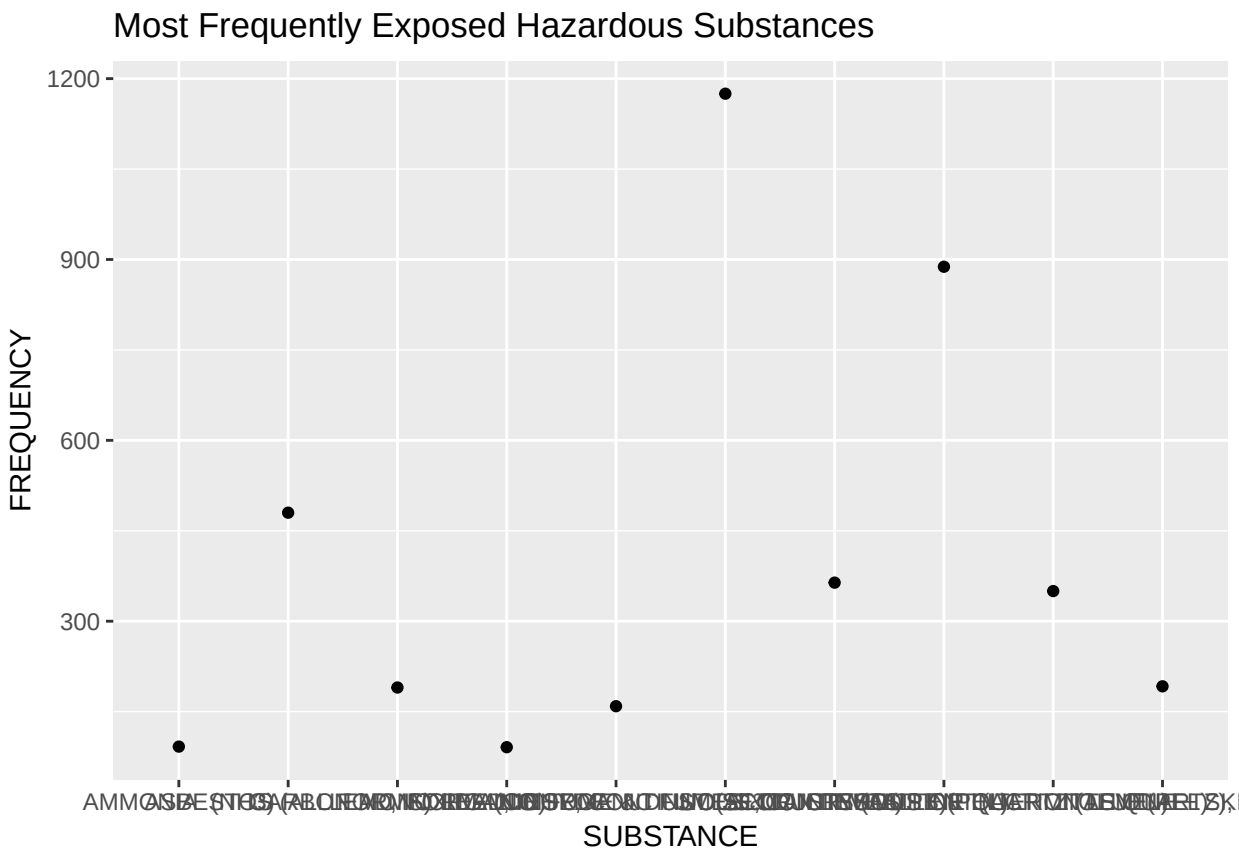
The OSHA data set contains data regarding workplace inspections, denial of entry when workplaces were meant to be inspected, and the specifics of how employees are involved in inspections and litigation. The original set also contained debt collection information, but that was removed because it was not relavent to the question of interest. Column titles were also clarified and encoded data was changed to English words or more obvious abbreviations (such as Y for Yes). Specific locations and names of workplaces are also included in this data set, as well as the number of records that are included in other data sets. For commented code, please see CleanOsha.R in the MA415MidtermFolder.

```
## The following `from` values were not present in `x`: B, X/T
```

```
## Warning: 25 failed to parse.
```

```
## Warning: 1084 failed to parse.
```

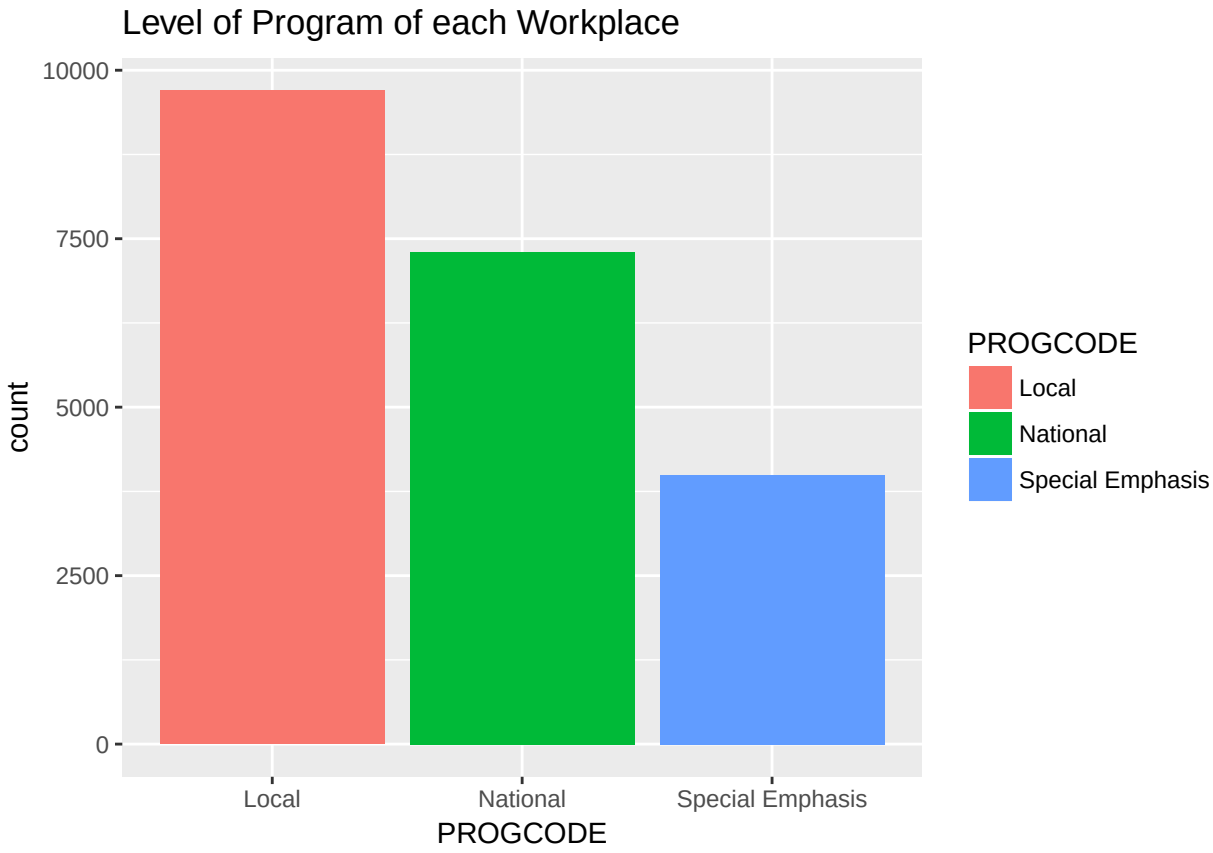
```
## Warning: 1263 failed to parse.

## The following `from` values were not present in `x`: N, O, P, T, U, W, X, Y

## The following `from` values were not present in `x`: X
## The following `from` values were not present in `x`: X

## Warning: 80043 failed to parse.

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector

##              CITY FREQUENCY
## 2209       BOSTON     11740
## 3030    CAMBRIDGE      2098
## 7102   FALL RIVER      1128
## 10007     HOLYOKE      1108
## 12827      LOWELL      1043
## 15241 NEW BEDFORD      1193
## 18167      QUINCY      1225
## 20990 SPRINGFIELD      2734
## 23326     WALTHAM      1112
## 24614   WORCESTER      2593
```



As seen in the above plot and table, the OSHA data set, while containing a lot of information that statisticians with the workplace safety question would find interesting, also includes information about location. This would allow investigators to uncover where the most dangerous workplaces tend to be.

The CleanProg data set includes information about the kind of program that each workplace inspection was a part of. These data were already mostly clean, but the SITESTATE column was removed and the National, Local, or Special Emphasis focus of each inspection was decoded into English words. For commented code, please see CleanProg.R in the MA415MidtermFolder.

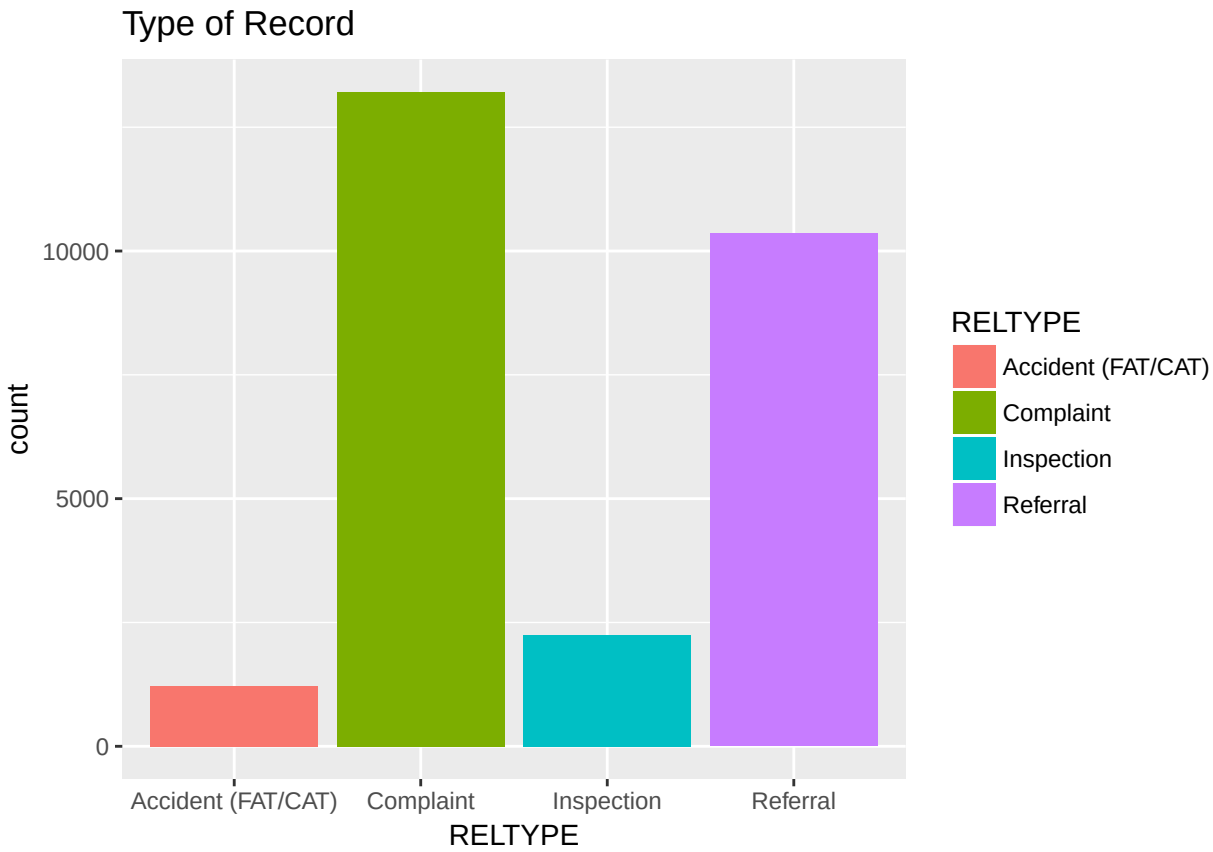
```
##
##           Local        National Special Emphasis
##            9695            7304            3992
```



In the cleaned prog data set, it was clear that each workplace investigated fell into a certain type of program. A statistician could investigate further, but the larger the program (i.e. National level instead of Local level), the more serious their violation could have been. This would directly relate to workplace safety.

The CleanRelact data set included information on events related to inspections. Primarily, it includes data regarding the type of relative event that caused the information to be recorded (i.e. complaint, referral, inspection) and whether or not there was a safety or health violation reported. The only cleaning required was the removal of the SITESTATE column and the decoding of the date from abbreviations to English words. For commented code, please see CleanRelact.R in the MA415MidtermFolder.

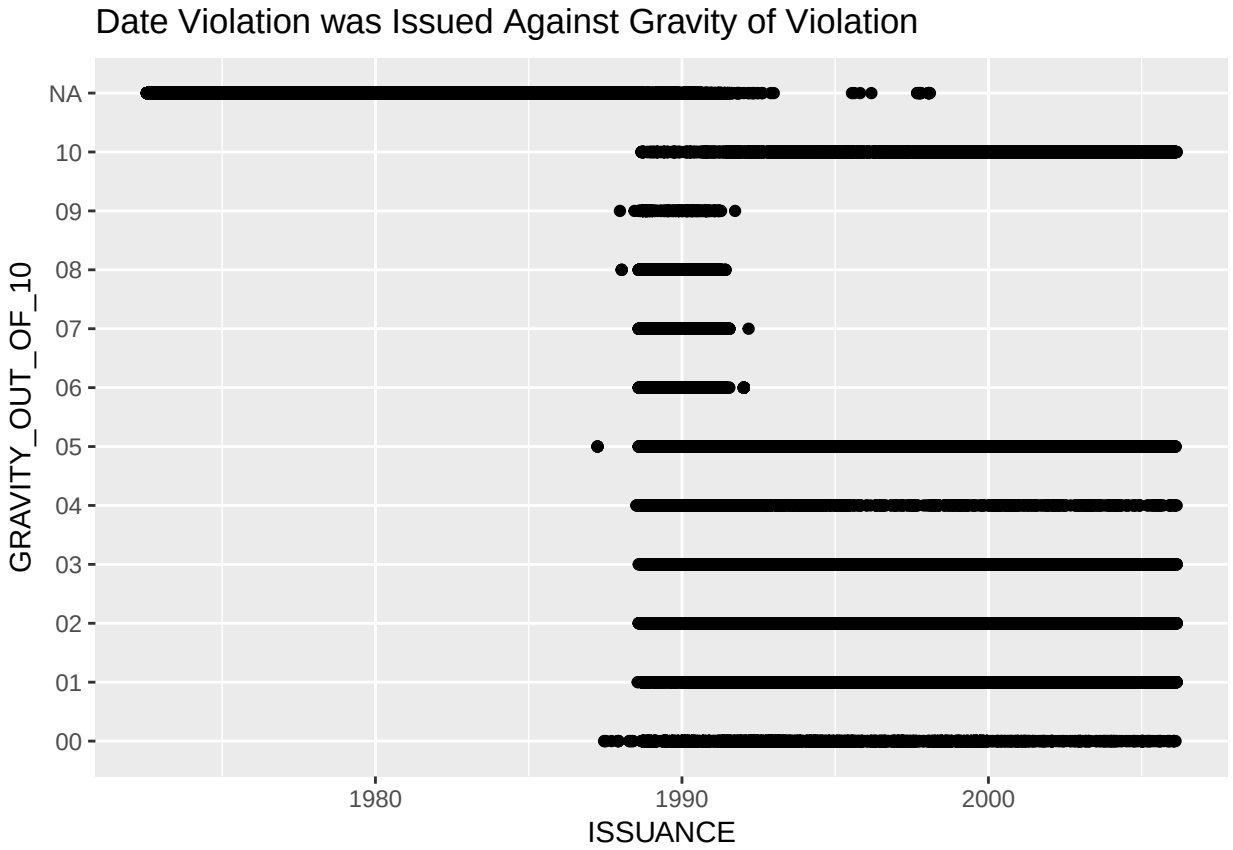
```
##
## Accident (FAT/CAT)          Complaint         Inspection
##               1215              13206               2251
##           Referral
##              10356
```



The cleaned relact data included information on the kind of record that reported health and safety code violations. This could be important to the statisticians that are investigating workplace safety because an accident is going to be more alarming than a complaint.

The CleanViol data set includes information about workplace violations. In order to clean the data, the SITESTATE column, any empty columns, and any non-relavent columns were removed. Encoded data were decoded (replaced with clearer English words), dates were made easier to read, and column titles were clarified. Now CleanViol primarily contains information regarding severity of violations, time of violations, penalties, and how different workplaces reacted to accusations (did they contest claims or not?). For commented code, please see CleanViol.R in the MA415MidtermFolder.

```
## Warning: 112 failed to parse.

## Warning: All formats failed to parse. No formats found.

## Warning: 262128 failed to parse.

## Warning: 269241 failed to parse.

## Warning: 279965 failed to parse.

## Warning: All formats failed to parse. No formats found.

## The following `from` values were not present in `x`: L, Y, 1

## The following `from` values were not present in `x`: W, L, Y, R, 1, 2, 3

##
##    00    01    02    03    04    05    06    07    08    09    10
##  5241 37252 19738 15793  3466  7996  2856  2338  1170   408 12355
```



There are a lot of data in the cleaned violations data set that could be useful regarding the workplace safety question. As seen in the above plot and table, the gravity of the violation out of ten is something that could definitely be interesting. The plot of the date of the violation against the gravity of the violation could be very interesting because it could help statisticians to see whether or not workplace safety has been getting better or worse over time.