MA415 Midterm Project: Safety in the workplace in Massachusetts

Making data ready for analysis

Sarah Gore- U72145380

This version: March 22, 2017

Contents

Available versions of this document: html–notebook version, pdf version, R script alone, html with full code version. These files can be found on my GitHub.

---

1 General purpose

1.1 Driving Question and Present Contribution

What is the safest workplace in MA and what can the data provided tell us about it?

I am cleaning the data provided in order to get it ready for analysis. Through doing this, I exclude data that I think is unnecessary,but at the same time I justify (to a certain extent) how and why this is done. This work is the important step between the data collection and data analysis.

This project will only focus on the second of the following steps in a data project:

1. Data collection
2. **Data cleaning**
3. Analysis

I acknowledge that there are many different ways to perform this project. I believe however, that this will prove useful.

2 Data

2.1 Source

The data was collected by OSHA and provided by NICAR.

The data provided covers the period 01/1972- 02/2006. It contains various tables ranging from the time that the inspections were carried out, accidents incurred, body parts injured, violations, fines etc…

The main data table is OSHA and the various supplementary tables provide key information for specific issues with the inspection such as accidents or violations, for example. There are also look-up tables that provide further sources of information about the coding for instance.

The links between the various data files can be made more clear by consulting the following file.

For more information about the variables used in the data, click on the following file.

3 Tables under consideration

I will be using OSHA as it is our main source. Furthermore, I will be using ACCID as well as VIOL. In this project I am concerned with safety in the workplace and it seems reasonable to analyze data related to accidents and violations.

I have excluded potentially useful data, such as company debt. However given the abundance of data, I find it is more relevance to choose the variables that I have.

4 Reading the data

```r
knitr::opts_chunk$set(echo = TRUE, tidy = TRUE, cache = TRUE)
require(ggplot2)
require(tidyr)
require(dplyr)
require(readr) # to read text
require(foreign) # library used to import the dbfs
require(lubridate) # library used for dates/times
```

First I load the relevant libraries that I use in the cleaning process. As mentioned above, I am working with three files, `osha`, `accid` and `viol`. In order to simplify the cleaning process and upcoming analysis, I am going to merge them into one file.

```r
# read the data and transform in tibble (for speed and better behaviour of
# the data frame)
osha <- read.dbf("osha.dbf", as.is = TRUE) %>% tbl_df
accid <- read.dbf("accid.dbf", as.is = TRUE) %>% tbl_df
viol <- read.dbf("viol.dbf", as.is = TRUE) %>% tbl_df

join_1 <- left_join(osha, accid, by = "ACTIVITYNO")
osha <- left_join(join_1, viol, by = "ACTIVITYNO")
rm(list = setdiff(ls(), "osha"))  # I erase my environment and now only have osha in my environment
```

I merge the three variables into one larger data frame and call it `osha`. This file contains all the information that was in `osha`, `accid` and `viol`.

However, after careful consideration, I realized that this goes against the principles of tidy data. So, I keep the three data frames separated.

```r
osha <- read.dbf("osha.dbf", as.is = TRUE) %>% tbl_df
accid <- read.dbf("accid.dbf", as.is = TRUE) %>% tbl_df
viol <- read.dbf("viol.dbf", as.is = TRUE) %>% tbl_df
```

4.1 Some notes on the data

The following is an expendable collection of notes that help the reader understand the data.

- Observations in `accid` are values for a single person in an accident. below, when counting accidents, I am referring to the number of people involved in accidents.
- After listening to the presentations of fellow students, I wanted to double check that the activity number's for each observation were unique before submitting my project and as I initally suspected, they were. There are 80445 rows in `osha` and 80444 different inspection identifiers in the data set. They equal up to a tiny difference. Therefore NO duplicates.
- All date variables are in the format of `package::lubridate`.

5 Removing variables

At this stage I want to remove some unnecessary variables. I explain below which variables I remove and the reason for doing so.

5.1 Single value variables

Single value variables are vectors that contain only one unique value or one value and 'NA's. In order to simplify the data sets, I remove these variables from the data frames.

For that purpose I wrote a specific function: `clean.from.one.value()`. This function takes a data frame as input and returns the same data frame without these variables.

```
clean.from.one.value <- function(df) {
    if (!is.data.frame(df)) {
        print("Not executed: the function requires a data frame object.")
        return()
    }
    one.value <- function(vec) {
        length(unique(vec)) == 1
    }
    clean.df <- df[, !sapply(df, one.value)]
    if (length(df) - length(clean.df) == 0) {
        print("No variable removed.")
    } else {
        print(paste(length(df) - length(clean.df), "variables removed:", length(df),
            "->", length(clean.df)))
    }
    return(clean.df)
}
```

I now apply the function defined above to my three data frames.

```
osha <- clean.from.one.value(osha)
accid <- clean.from.one.value(accid)
viol <- clean.from.one.value(viol)
```

```
## [1] "8 variables removed: 92 -> 84"
## [1] "1 variables removed: 16 -> 15"
## [1] "3 variables removed: 48 -> 45"
```

5.2 Irrelevant variables in osha

Here I make a choice of variables that I think are unnecessary and can therefore be removed.

5.2.1 NICAR dates

NICAR converted the original dates into a standard format. Therefore, I discard the original variables that contain the original dates.

```
irrelavant_variables <- c("OSHA1MOD", "OPENDATE", "CLOSEDATE", "CLOSEDATE2",
    "FRSTDENYN", "LSTREENTRN", "PENDUDATE", "FTADUDATE", "FRSTCONTST")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.2.2 Number of records

osha contains variables indicating counts in tables that I have discarded for this project. Therefore, I discard the variables relating to these tables.

```
irrelavant_variables <- c("PROG_", "RELACT_", "OPTINFO_", "DEBT_", "EVENT_",
    "HAZSUB_", "ADMPAY_")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.2.3 Paperwork times

Time spent on preparation, traveling, researching etc… are variables that I think are not going to be relevant in the analysis.

```r
irrelavant_variables <- c("PAPREP", "PATRAVEL", "PAONSITE", "PATECHSUPP", "PARPTPREP",
    "PAOTHRCNF", "PALITIGTN", "PADENIAL", "PASUMHOURS")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.2.4 Debt-related variables

Since I have discarded the tables relating to debt, I can remove these variables from `osha`.

```r
irrelavant_variables <- c("PENDUDATE", "PENDUDT", "FTADUDATE", "FTADUDT", "DUECODE")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.2.5 Penalty- related variables

The same argument as mentioned in the previous point applies to this category.

```r
irrelavant_variables <- c("PENREMIT", "FTAREMIT", "TOTPENLTY", "TOTALFTA")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.2.6 Other irrelevant variables

This category applies to administrative information that I cannot use.

```r
irrelavant_variables <- c("DUNSNO", "HOSTESTKEY")

osha <- osha[, !(names(osha) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.3 Irrelevant variables in `accid`

5.3.1 Unused categories

This subset applies to variables in `accid` that I discard.

```r
irrelavant_variables <- c("SITESTATE", "NAME", "RELINSP", "SOURCE", "ENVIRON",
    "EVENT", "TASK", "HAZSUB", "OCC_CODE")

accid <- accid[, !(names(accid) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.4 Irrelevant variables in viol

Again, I only use a subset of the variables in viol and discard the others. Here, the justification is that I don't think that these variables are related to safety in the workplace, for instance abatement.

5.4.1 Dates, citations and abatements etc...

```
irrelavant_variables <- c("SITESTATE", "DELETE", "ISSUANCE", "ISSUEDATE", "CITATION",
    "ITEMNO", "ITEMGROUP", "EMPHASIS", "PENCURRENT", "PENINITIAL", "STD_LOOKUP",
    "STD", "ABATE", "DATE_ABATE", "ABATEDT", "ABATEDT2", "REC", "ABATEDONE")

viol <- viol[, !(names(viol) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.4.2 Violated Contested Information

```
irrelavant_variables <- c("ERCONTDT", "ERCONDATE", "VIOLCONT", "PENCONT", "EMPRCONT",
    "EMPECONT", "FINORDT", "FINORDATE", "PMA", "AMENDED", "ISA", "DISPEVT")

viol <- viol[, !(names(viol) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

5.4.3 Failure to abate

```
irrelavant_variables <- c("FTAINSP", "FTAPEN", "ISSUDT", "FTA_ISDT", "CONTDT",
    "CONTDATE", "FTA_AMN", "FTA_ISA", "FTA_DISP", "FTA_FIN", "FTAFINDT", "HAZCAT")

viol <- viol[, !(names(viol) %in% irrelavant_variables)]
rm(irrelavant_variables)
```

6 Adding information

In osha many variables come with coded-values; this is not very informative in particular when cleaning and graphing the data. In order to make the information more informative, I use the look-ups provided to add to translate the code to values I can use.

For instance, if I want to know about the places of inspections, I will use the name of the counties and I access these through the look-ups.

6.1 New variables to osha

I added information about the place of inspection to osha. *(Notice that this method is drastically different from what was shown in class)*

```
# I load the lookup from scc

scc <- read.dbf("lookups/scc.dbf") %>% tbl_df %>% filter(STATE == "MA")
counties <- read.dbf("lookups/scc.dbf") %>% tbl_df %>% filter(STATE == "MA") %>%
    filter(CITY == "0000", COUNTY != "000") %>% select(COUNTY, NAME) %>% rename(COUNTY_CODE = COUNTY,
    COUNTY_NAME = NAME)
osha <- left_join(osha, counties, by = c(SITECNTY = "COUNTY_CODE"))
cities <- read.dbf("lookups/scc.dbf") %>% tbl_df %>% filter(STATE == "MA") %>%
    filter(CITY != "0000", COUNTY != "000") %>% select(CITY, NAME) %>% rename(CITY_CODE = CITY,
    CITY_NAME = NAME)
osha <- left_join(osha, cities, by = c(SITECITY = "CITY_CODE"))
rm(scc, counties, cities)
```

6.2 New variables to accid

I added information about the body- parts, nature of injury and human contribution in the accidents to accid.

```
# I load the lookup from acc

acc <- read.dbf("lookups/acc.dbf") %>% tbl_df %>% filter(CATEGORY == "PART-BODY") %>%
    rename(BODYPART_NAME = VALUE)
accid <- left_join(accid, acc, by = c(BODYPART = "CODE"))

acc <- read.dbf("lookups/acc.dbf") %>% tbl_df %>% filter(CATEGORY == "NATURE-INJ") %>%
    rename(NATURE_NAME = VALUE)
accid <- left_join(accid, acc, by = c(NATURE = "CODE"))

acc <- read.dbf("lookups/acc.dbf") %>% tbl_df %>% filter(CATEGORY == "HUMAN-FAC") %>%
    rename(HUMAN_NAME = VALUE)
accid <- left_join(accid, acc, by = c(HUMAN = "CODE"))
rm(acc)
```

7 NA's and missing values

Many variables contain NA's. However there are various reasons for these NA's. In some of the cases they do represent missing values. However, when going through the descriptions of the osha data, it became clear that some NA's actually represent data and I want to account for that data.

7.1 NA's that represent actual data

Consider the following example taken from the description of the coding:

> WALKAROUND C 1 Employee representative present during inspection: B (blank) = no X = yes

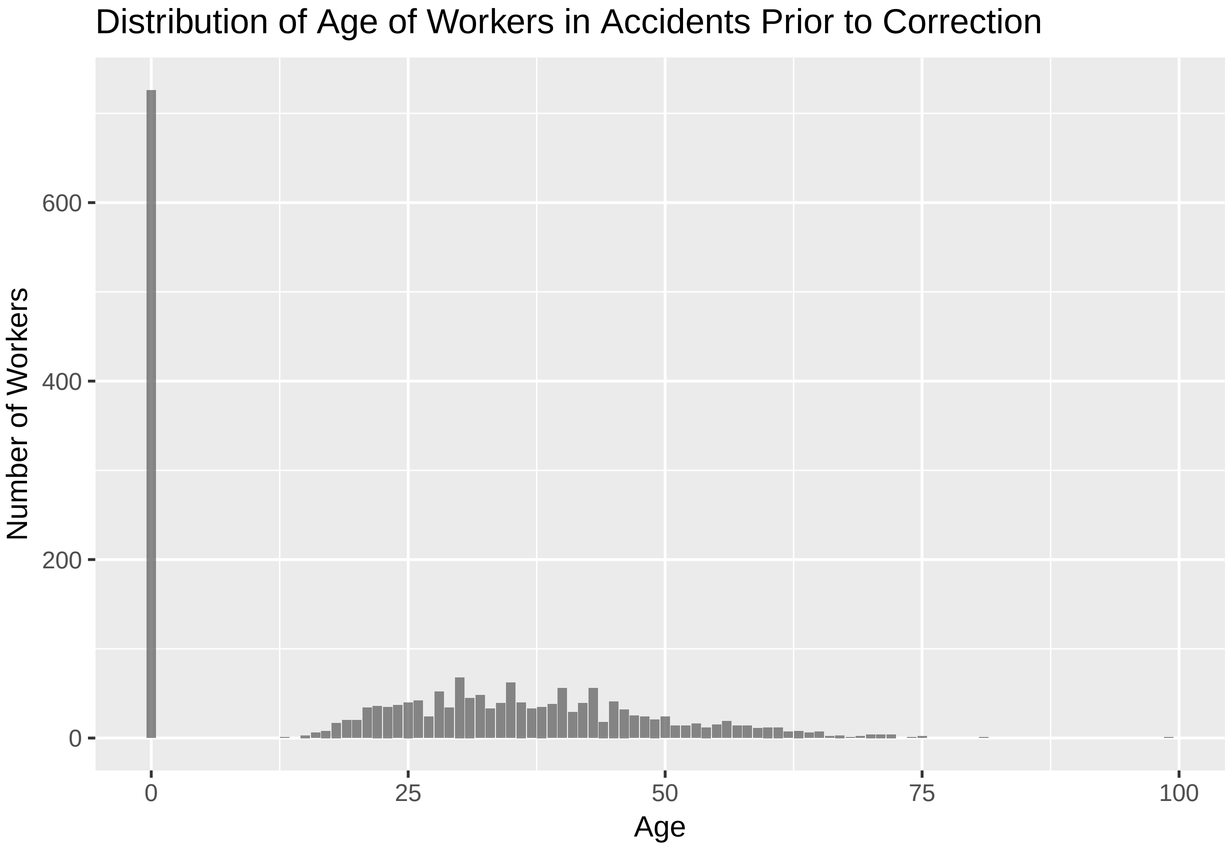
Clearly, there is no information missing. The NA seen in the data represents the value "no". Therefore in the variables that use this coding system, I replace the NA's by "no". My choice must be taken into account for future analyses.

```
blank_variables <- c("CONTFLAG", "STFLAG", "PREVCTTYP", "WALKAROUND", "INTRVIEWD",
    "CLOSECASE", "SAFETYMANF", "SFTYCONST", "STFYMARIT", "HELTHMANF", "HELTHCONST",
    "HELTHMARIT", "MIGRANT", "ANTCSRVD")
blank_variables <- blank_variables[blank_variables %in% colnames(osha)]
osha[, blank_variables][is.na(osha[blank_variables])] <- "NO"
rm(blank_variables)
```

7.2 Values that are actually NA's

As a good practice, I plot the some of the data to look for outliers. Within the `accid` data under the age category, I encounter a bunch of zero's which does not make sense; indeed as such a large number of zeros is an indication that something is wrong.

```
accid %>% ggplot(aes(x = AGE)) + geom_bar(alpha = 0.7) + labs(title = "Distribution of Age of Workers i
    x = "Age", y = "Number of Workers")
```

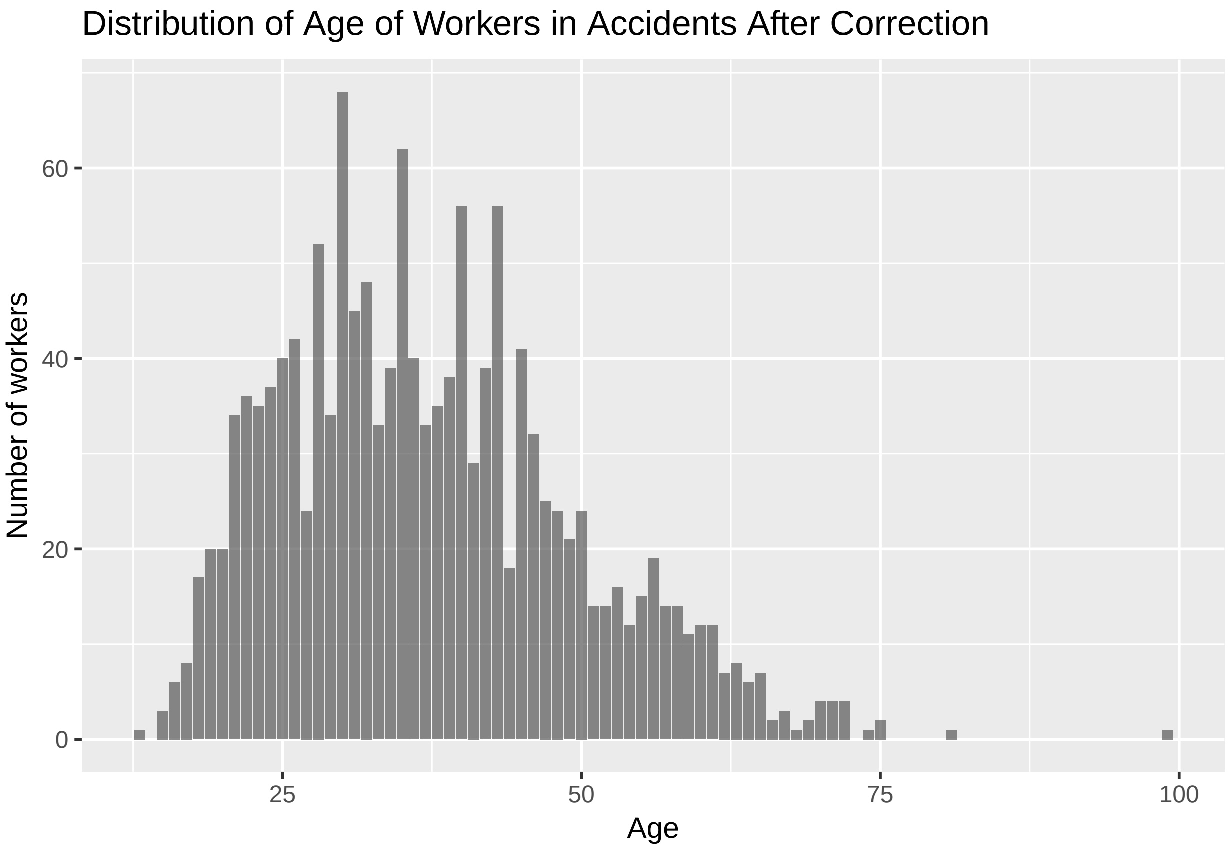


I correct for this by specifically assigning the value NA to these zeros.

```
accid$AGE[accid$AGE == 0] <- NA
```

In the amended distribution, there are still many extreme values, ranging from NA to NA. So can assume that these are outliers. For instance young/old people present when the accident occurred.

```
accid %>% ggplot(aes(x = AGE)) + geom_bar(alpha = 0.7) + labs(title = "Distribution of Age of Workers i
    x = "Age", y = "Number of workers")
```



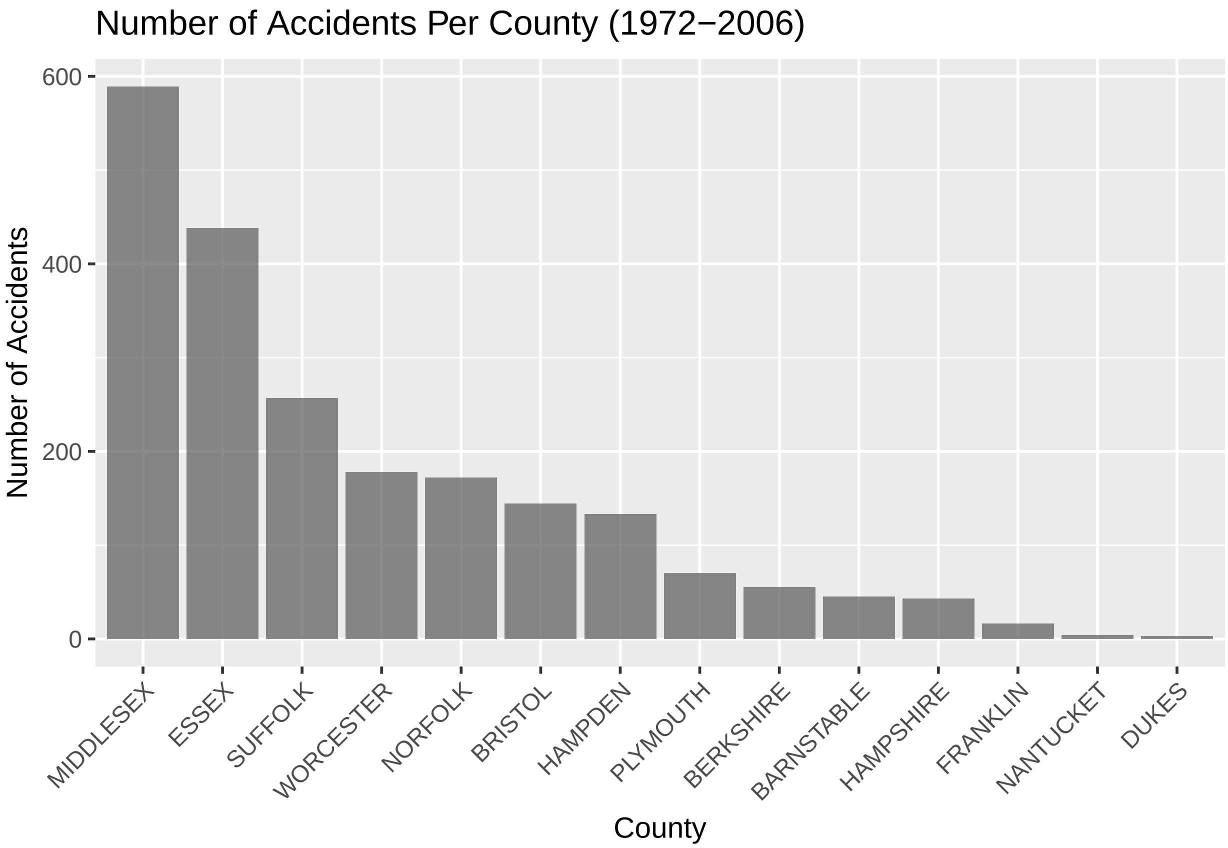
8 First glimpse at the data

There are very many ways of looking at the data that I have cleaned. Below are a few graphs that I feel give one a notion of the data.

8.1 The number of accidents per county

Since, I am looking for secure work environment, a natural perspective would be to break-up accidents by county.

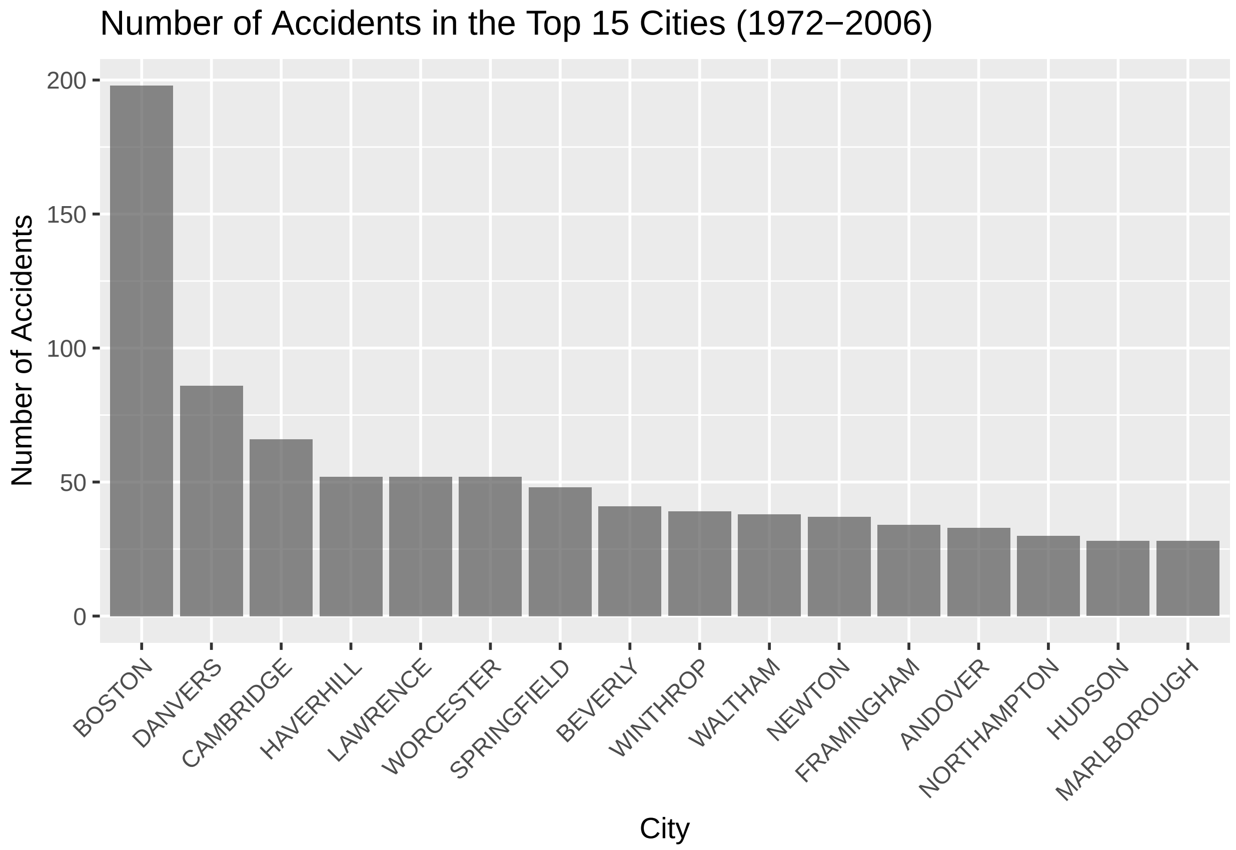
```
osha %>% group_by(COUNTY_NAME) %>% summarise(accidents = sum(ACCID_)) %>% filter(!is.na(COUNTY_NAME)) %>
    arrange(-accidents) %>% mutate(COUNTY_NAME = factor(COUNTY_NAME, COUNTY_NAME)) %>%
    ggplot(aes(x = COUNTY_NAME, y = accidents)) + geom_bar(alpha = 0.7, stat = "identity") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(title = "Number of Accidents Per Cou
    x = "County", y = "Number of Accidents")
```



8.2 The number of accidents per city

I can also break-up accidents by city. However, there are too many so I take the top 15 cities in MA.

```
osha %>% group_by(CITY_NAME) %>% summarise(accidents = sum(ACCID_)) %>% filter(!is.na(CITY_NAME)) %>%
    arrange(-accidents) %>% top_n(n = 15, accidents) %>% mutate(CITY_NAME = factor(CITY_NAME,
    CITY_NAME)) %>% ggplot(aes(x = CITY_NAME, y = accidents)) + geom_bar(alpha = 0.7,
    stat = "identity") + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Number of Accidents in the Top 15 Cities (1972-2006)", x = "City",
        y = "Number of Accidents")
```



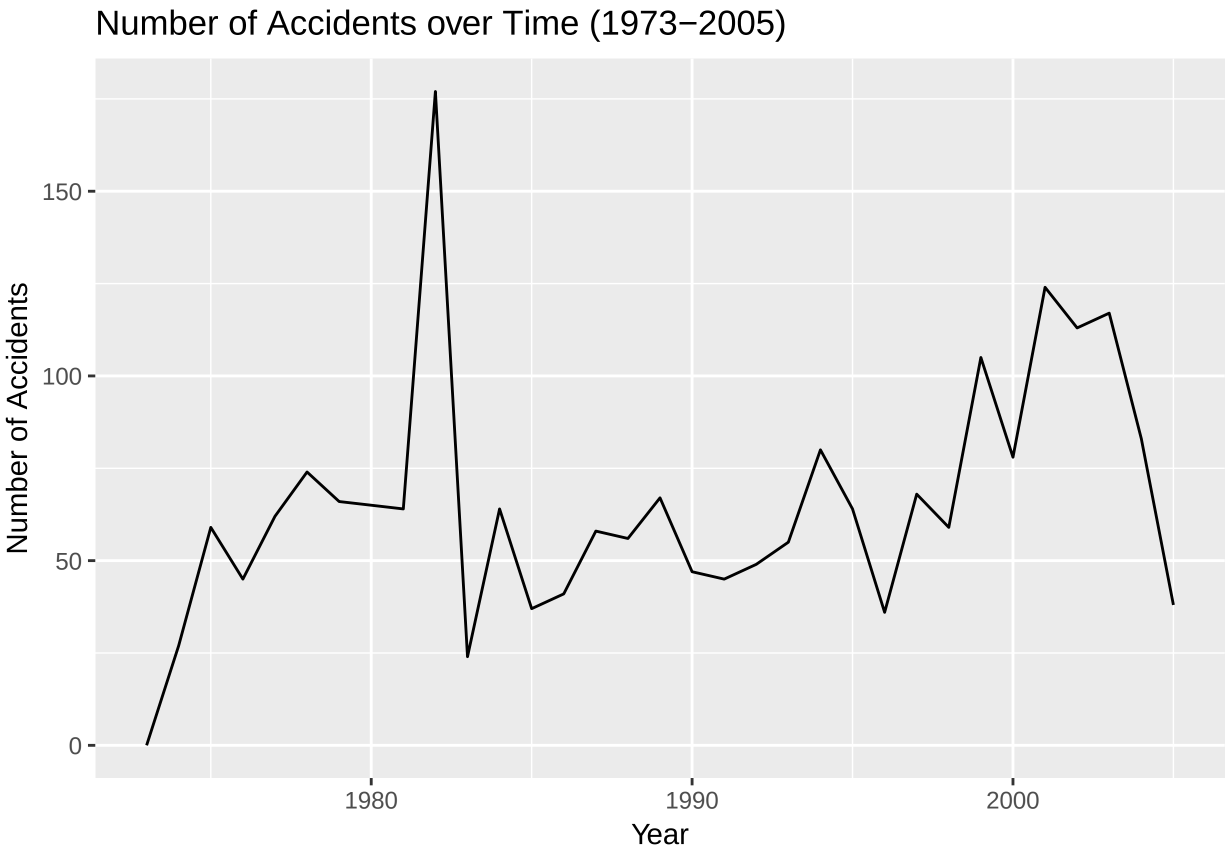
8.3 Accidents over time

I am interested in the evolution in the number of accidents over time. Here I plot the total number of accidents during that period in MA.

Notice that I exclude 1972 and 2006 because they are incomplete years.

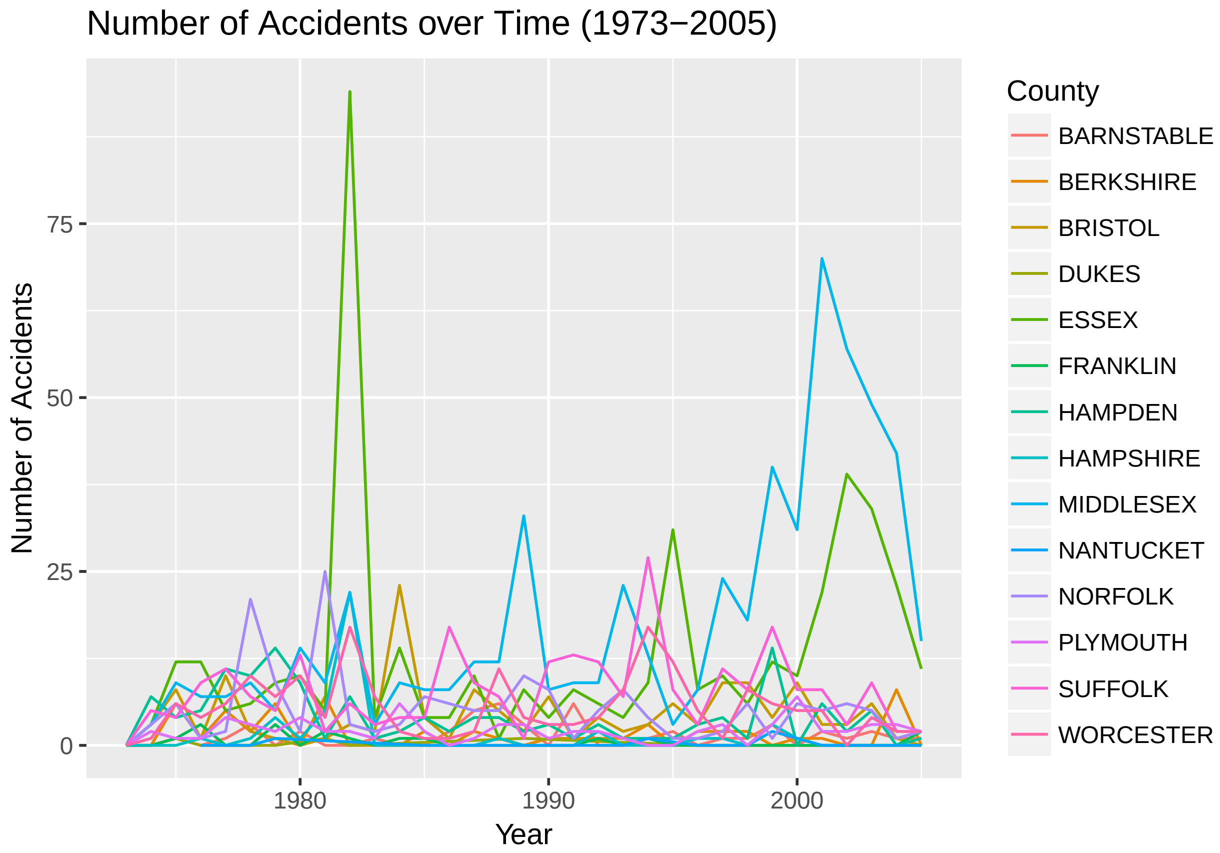
From the data I see that the max number of accidents in this time period was 72 and the min number of accidents in this time period was 0

```
osha %>% group_by(year(OPENDT)) %>% summarise(accidents = sum(ACCID_), year = mean(year(OPENDT))) %>%
    filter(year > 1972, year < 2006) %>% ggplot(aes(x = year, y = accidents)) +
    geom_line() + labs(title = "Number of Accidents over Time (1973-2005)",
    x = "Year", y = "Number of Accidents")
```



I get further inside by breaking up the evolution over time by county.

```
osha %>% group_by(year(OPENDT), COUNTY_NAME) %>% filter(!is.na(COUNTY_NAME)) %>%
    summarise(accidents = sum(ACCID_), year = mean(year(OPENDT))) %>% filter(year >
    1972, year < 2006) %>% ggplot(aes(x = year, y = accidents, col = COUNTY_NAME,
    group = COUNTY_NAME)) + geom_line() + labs(title = "Number of Accidents over Time (1973-2005)",
    x = "Year", y = "Number of Accidents", col = "County")
```

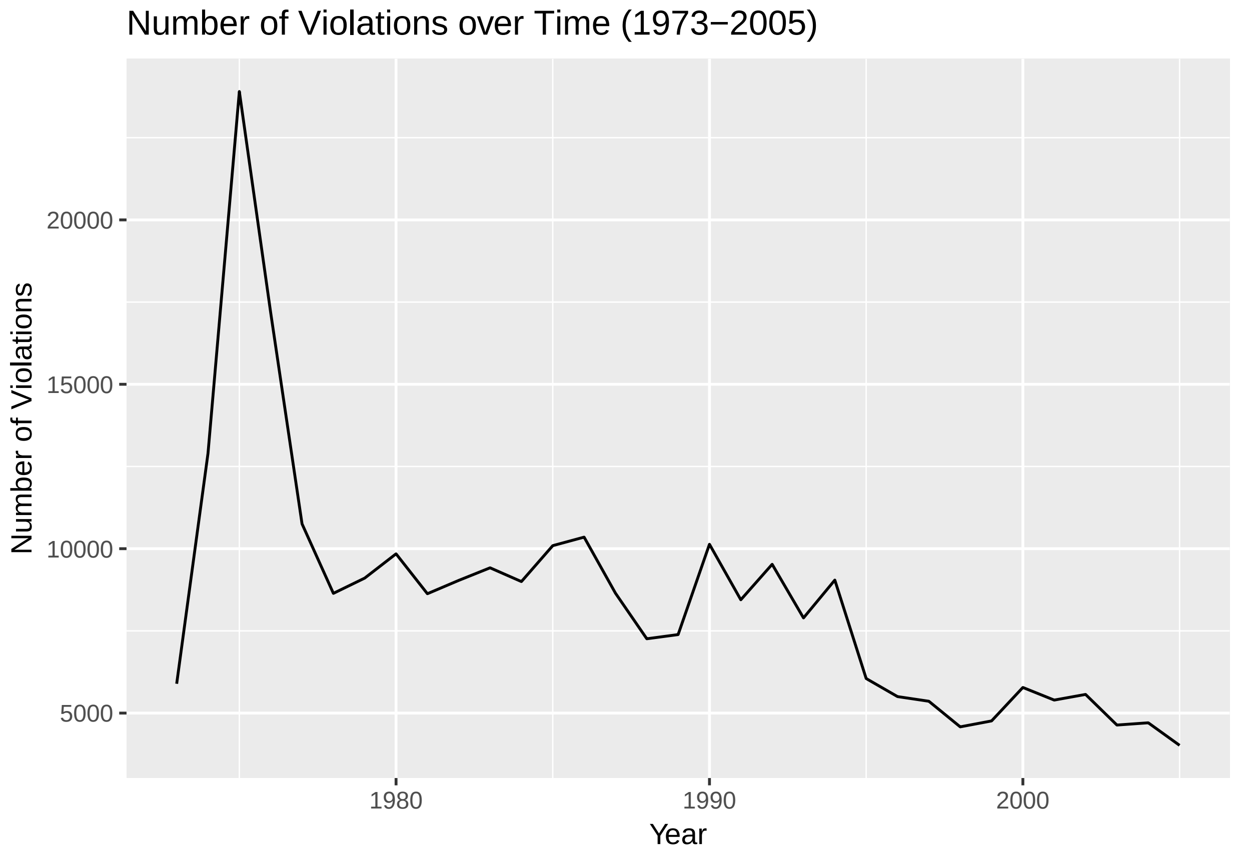


## Violations over time Violations tell us how well firms are complying with the rules of safety. This is why evolution over time is informative. Here I plot the total number of violations during that period in MA.

Again I exclude 1972 and 2006 because they are incomplete years.

From the data I see that the max number of accidents in this time period was 161 and the min number of accidents in this time period was 0

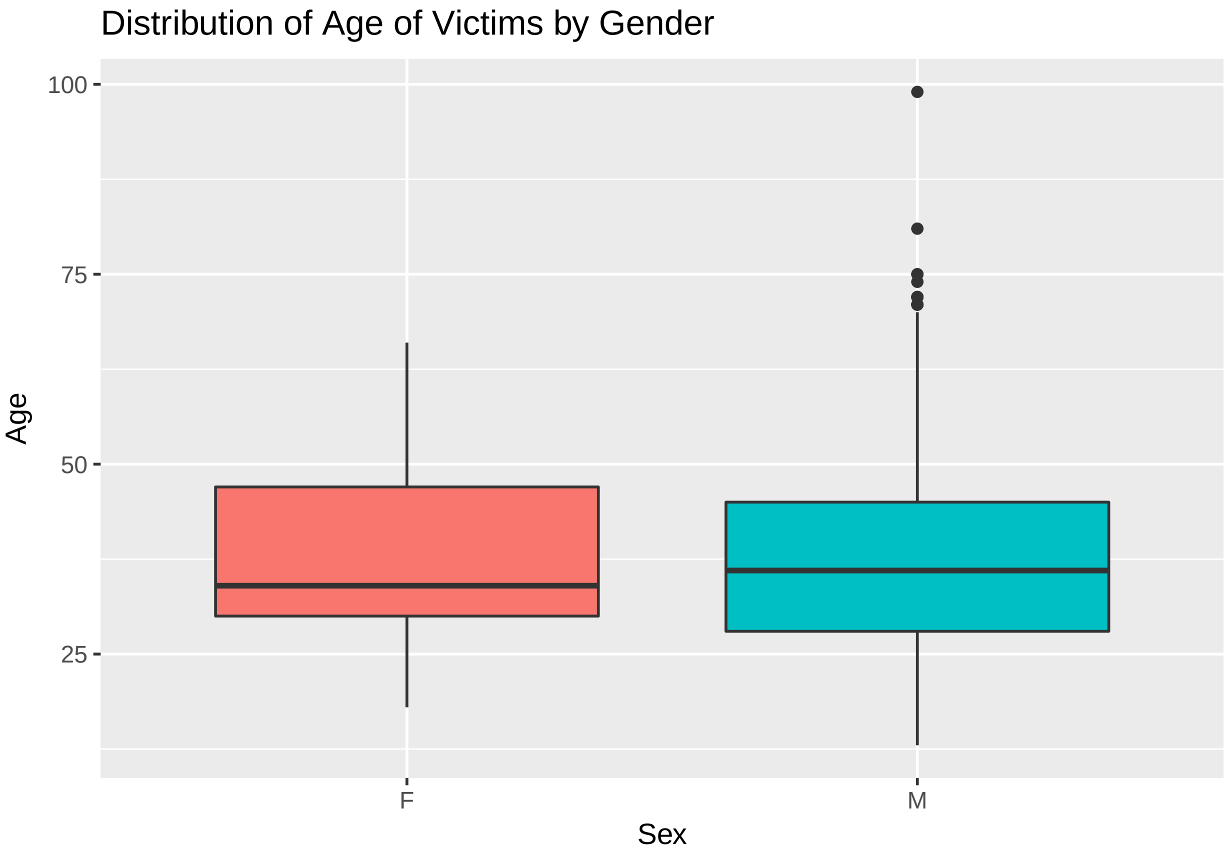
```
osha %>% group_by(year(OPENDT)) %>% summarise(violations = sum(VIOLS_), year = mean(year(OPENDT))) %>%
    filter(year > 1972, year < 2006) %>% ggplot(aes(x = year, y = violations)) +
    geom_line() + labs(title = "Number of Violations over Time (1973-2005)",
    x = "Year", y = "Number of Violations")
```



8.4 Age of the victims of accidents

Age usually plays a role in many work-place accidents. here I plot the distribution of age, differentiated by gender.

```r
accid %>% filter(!is.na(SEX)) %>% ggplot(aes(x = SEX, y = AGE, fill = SEX)) +
    geom_boxplot() + labs(title = "Distribution of Age of Victims by Gender",
    x = "Sex", y = "Age") + guides(fill = FALSE)
```

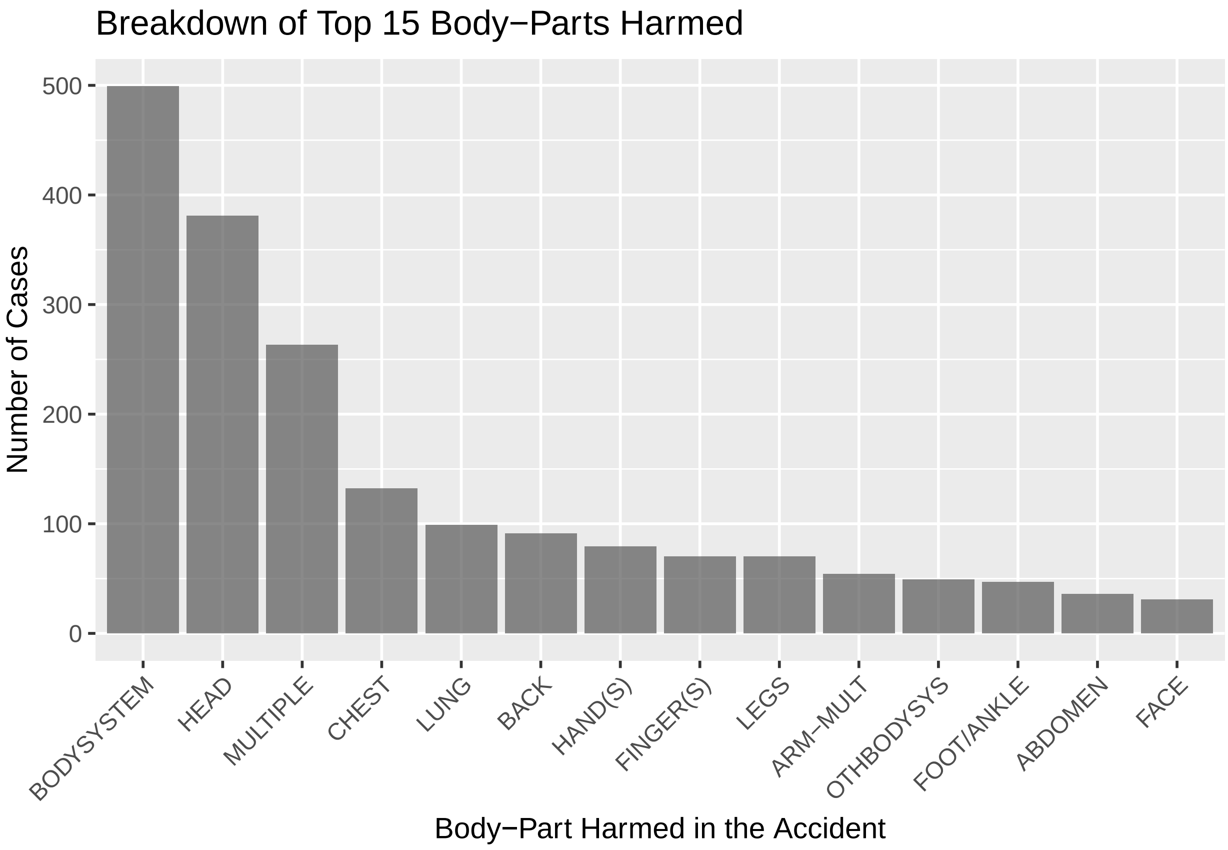


8.5 Nature of the injuries

The nature of the injury can provide useful informative in terms of possible prevention.

I chose the top 15 body-parts as there are 31 different categories of body-parts that were harmed in the accidents.

```
accid %>% group_by(BODYPART_NAME) %>% summarise(cases = n()) %>% arrange(-cases) %>%
    top_n(n = 15, cases) %>% mutate(BODYPART_NAME = factor(BODYPART_NAME, BODYPART_NAME)) %>%
    filter(!is.na(BODYPART_NAME)) %>% ggplot(aes(x = BODYPART_NAME, y = cases)) +
    geom_bar(alpha = 0.7, stat = "identity") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1)) + labs(title = "Breakdown of Top 15 Body-Parts Harmed", x = "Body-Part Harmed in the Ac
    y = "Number of Cases")
```



- I acknowledge that I benefitted from the help of a tutor with whom I work regularly. He reviewed my code and offered some suggestions of improvement. I remain soley responsible of this work.
- I hope you found this as interesting as I did!

9 Appendix: Variable Description

The following tables provide information about the variables that remain in the data set after cleaning. For further information, please refer to the original document such as osha.txt.

9.1 Description of variables in osha

```
text <- read_lines("layouts/OSHA.txt", skip = 10)
text1 <- gsub("\\t\\t\\t.*", "", text)
text2 <- text1[grepl("\t", text1)]
text3 <- gsub("\\t\\t", "\t", text2) %>% strtrim(100) %>% tbl_df %>% separate(value,
    c("variable", "type", "length", "description"), sep = "\t")
text3$variable <- gsub(" ", "", text3$variable, fixed = TRUE)
text3 <- text3 %>% select(variable, description) %>% arrange(variable)
osha_descriptions <- text3 %>% filter(variable %in% names(osha))
knitr::kable(osha_descriptions, caption = "Description of Variables in `osha`")
```

Table 1: Description of Variables in osha

variable	description
ACCID_	Number of records that should be in ACCID.DBF for this inspection
ACTIVITYNO	Unique identifier for each inspection record.
ADVNOTICE	Advance notice given
ANTCSRVD	Warrant served on subpoena prior to start of inspection
CAT_SH	CODES:
CATSICGDE	SIC found in the planning guide for the establishment
CATSICINSP	SIC inspected if different from primary SIC
CLOSECASE	Codes:
CLOSEDT	NICAR-converted version of CLOSEDATE
CLOSEDT2	NICAR-converted version of CLOSEDATE2
CONTFLAG	Indicates whether record is a continuation of the previous record (occurs primarily whe
DATARQD	200 log data required per ISA
ESTABNAME	Name of establishment
FRST_DT	NICAR-converted version of FRSTDENYN
FRSTCONDT	NICAR-converted version of FRSTCONTST
FRSTDENY	Date of initial denial of entry
HELTHCONST	Health construction inspection
HELTHMANF	Health manufacturing inspection
HELTHMARIT	Health maritime inspection
HISTFLAG	Identifies the entry source and era of the record.
INSPSCOPE	Inspection scope
INSPTYPE	Inspection type
INTRVIEWD	Employee(s) interviewed during inspection (state only)
JOBTITLE	The job classification of primary compliance officer
LSTR_DT	NICAR-converted version of LSTREENTRN
LSTREENTR	Date compliance officer re-entered establishment inspection
MIGRANT	Inspection of migrant farm worker camp
MOD_DATE	NICAR-converted version of OSHA1MOD.
NAICS	North American Industry Classification System (See NAICS.DBF)
OPENDT	NICAR-converted version of OPENDATE
OWNERCODE	Agency code for owner-type = D only (see FDA.dbf)
OWNERTYPE	Type of owner
PREVACTNO	Unique OSHA number for this inspection
PREVCTTYP	The type of the most recent OSHA activity, if any.
REPORTID	Unique identification code, one for each federal and state
SAFETYMANF	Safety manufacturing inspection
SFTYCONST	Safety construction inspection
SFTYMARIT	Safety maritime inspection
SHPGM	Safety/Health program initiated as result of an
SIC	Standard industrial classification (see SIC.DBF)
SITEADD	Street address
SITECITY	Department of Commerce city code (see SCC.DBF)
SITECNTY	Department of Commerce county code (see SCC.DBF)
SITEZIP	Zip
TOTALVIOLS	Total number of violations issued
TOTSERIOUS	Total number of serious, willful, and repeat violations issued
UNION	Employees represented by union
VIOLS_	Number of records that should be in VIOLS.DBF for this inspection
WALKAROUND	Employee representative present during inspection
WHYNOINSP	If no inspection, INSPSCOPE = D

9.2 Description of variables in `accid`

```r
text <- read_lines("layouts/ACCID.txt", skip = 9)
text1 <- gsub("\\t\t\t\t.*", "", text)
text2 <- gsub("\\t\t", "\t", text1)
text2 <- gsub("\\t\t", "\t", text2)
text2 <- gsub("\t$", "", text2)
text3 <- text2[grepl("\t", text2)]
text3 <- text3 %>% tbl_df %>% separate(value, c("variable", "type", "length",
    "description"), sep = "\t") %>% select(variable, description) %>% arrange(variable)
accid_descriptions <- text3 %>% filter(variable %in% names(accid))
knitr::kable(accid_descriptions, caption = "Description of Variables in `accid`")
```

Table 2: Description of Variables in `accid`

variable	description
ACTIVITYNO	Unique identifier for each inspection record
AGE	Age of victim
BODYPART	Part of body (see ACC.DBF lookup for codes)
DEGREE	Extent of injury
HUMAN	Human factor (see ACC.DBF lookup for codes)
NATURE	Nature of injury (see ACC.DBF lookup for codes)
SEX	Gender of victim

9.3 Description of variables in `viol`

```r
text <- read_lines("layouts/VIOL.txt", skip = 7)
text1 <- gsub("\\t\t\t\t.*", "", text)
text2 <- gsub("\\t\t", "\t", text1)
text2 <- gsub("\\t\t", "\t", text2)
text2 <- gsub("\t$", "", text2)
text3 <- text2[grepl("\t", text2)]
text3 <- text3 %>% tbl_df %>% separate(value, c("variable", "type", "length",
    "description"), sep = "\t") %>% select(variable, description) %>% arrange(variable)
viol_descriptions <- text3 %>% filter(variable %in% names(viol))
knitr::kable(viol_descriptions, caption = "Description of Variables in `viol`")
```

Table 3: Description of Variables in `viol`

variable	description
ACTIVITYNO	Unique identifier for each inspection record
GRAVITY	Indicates the level of potential harm to worker(s), values B,
INSTANCES	Number of instances of violation of standard related event
NUMEXPOSED	Number of employees exposed to hazard violated
VIOLTYPE	Current type of violation
VIOLTYPEA	Initial assess violation type

```r
rm(text, text1, text2, text3)
```