Midterm Project - Dangerous Workplaces

Vikram Vishwanath

March 20, 2017

I. Introduction

What is the most dangerous workplace in Massachusetts? To answer this, we use the U.S. Dept. of Labor's OSHA database to compile, organize, and present accident and violation data for over 1000 workplaces in MA. The data span more than four decades of statistics, starting with the 1970s. The goal is to organize these data into meaningful sets that an analyst can manipulate without encountering errors. Such data cleaning mainly involves tidying data tables, removing/changing missing values, separating data into defined categories, etc. The following report contains the steps that have been taken to clean up the raw data. A short exploratory analysis is performed to gauge the cleanliness of the data.

II. Data Cleaning

A. Raw data

The files used for this preparation are:

1. *accid.dbf*
2. *viol.dbf*
3. *osha.dbf*
4. *acc.dbf*
5. *hazsub.dbf*
6. *hzs.dbf*
7. *sic.dbf*

The first three files provide the details of OSHA inspections. All of the data are coded (i.e. body parts are coded with numbers 1-31, injuries are coded 1-18, etc.). *acc.dbf* contains the references for these codes. *hazsub.dbf* and *hzs.dbf* give the codes for hazardous substances, and *sic.dbf* gives the industry type (fishing, repair, construction, etc.)

B. Cleaning *accid.dbf*

This file contains 2147 observations of accidents and related details. With 16 variables, the table holds a large amount of information; however, not all variables are needed.

After checking that every single value in the STATE column was the same, the column was removed. All of these incidents occurred in MA.

The DBF file has several columns but the ones that were appropriate for describing an accident are:

1. ACTIVITYNO: unique identifier for a given record
2. DEGREE: fatal or non-fatal
3. NATURE: type of injury
4. BODYPART: body part(s) affected
5. EVENT: circumstances of injury
6. SOURCE: source of injury (equipment, vehicle, etc.)
7. ENVIRON: location of injury

8. HUMAN: presence/absence of human error
9. TASK: regularly scheduled task or not
10. HAZSUB: hazardous substance present or not

Upon inspection, the last column, HAZSUB, seemed to have missing values. It was assumed that if HAZSUB = NA, then no hazardous substance was present during the accident. The total NA's in each column was counted. The table was filtered and all incidents containing hazardous substances was moved to a new DBF file. Hereafter, both the original accidents DBF (without hazardous substances) and the new file (with hazardous substances) are worked with simultaneously.

Both tables were arranged by DEGREE (1, 2, 3). It was found that the accidents without hazardous substances contained incidents of zero degree. From the README files for the original data sets, zero values were undefined. It was inferred that a degree of zero represents an incident without injuries. These rows were moved to a new table and removed from the original tables.

Using *hzs.dbf*, the coded HAZSUB column was replaced by its corresponding hazardous substances. Not all hazardous substances had a matching code so this joining procedure produced missing values. These rows cannot be omitted as they can be used for analysis; the table was arranged such that these missing values would be towards the end of the table.

C. Cleaning *viol.DBF*

This table contains the violations for each company as well as more than 30 other variables. To determine the hazardous conditions of a site, only the following columns were retained:

1. ACTIVITYNO: same unique identifier as in *accid.DBF*
2. ISSUEDATE: date of issue, useful for time series
3. ITEMNO: tally of repeat violations
4. GRAVITY: seriousness of violation
5. VIOLTYPE: type of violation (serious, willful, repeat, etc.)
6. INSTANCES: total number of violations

Some of the violations in this table were marked as "deleted due to some reason other than failure to submit information." These rows were removed.

After performing the above cleanup, all rows were checked for mising values; none were found, except in the GRAVITY column. Some duplicates were found, and were deleted. The GRAVITY column provides meaningful information so as we did with the accidents file, rows containing GRAVITY = NA were moved to the bottom of the table.

D. Cleaning *osha.DBF*

This is likely the most comprehensive of all of the tables. It contains summaries of all of the other tables. The columns of interest are:

1. ACTIVITYNO: unique identifier
2. ESTABNAME: name of workplace
3. SITEADD: address
4. SITEZIP: zipcode
5. SIC: industry code
6. ACCID_: number of accidents (in *accid.dbf*)
7. VIOLS_: number of violations (in *viol.dbf*)
8. HAZSUB_: number of hazardous substances

If a place has no violations, no accidents, and no hazardous substances, then we can assume it is a safe place. These rows were deleted because they would not assist in determining how dangerous other places are.

Using *sic.dbf*, the sic codes were replaced with their industry (i.e. the coded numbers were changed to strings such as "FISHING" or "REPAIR").

E. Joining tables

All three tables have a common identifier of ACTIVITYNO. Using this, the site names and addresses and industries (and all other relevant information) were attached to each table. The result was four large comprehensive tables:

1. Accident
2. Accidents (without hazardous substances)
3. Violations
4. Summary (from *osha.DBF*)

F. Replacing codes

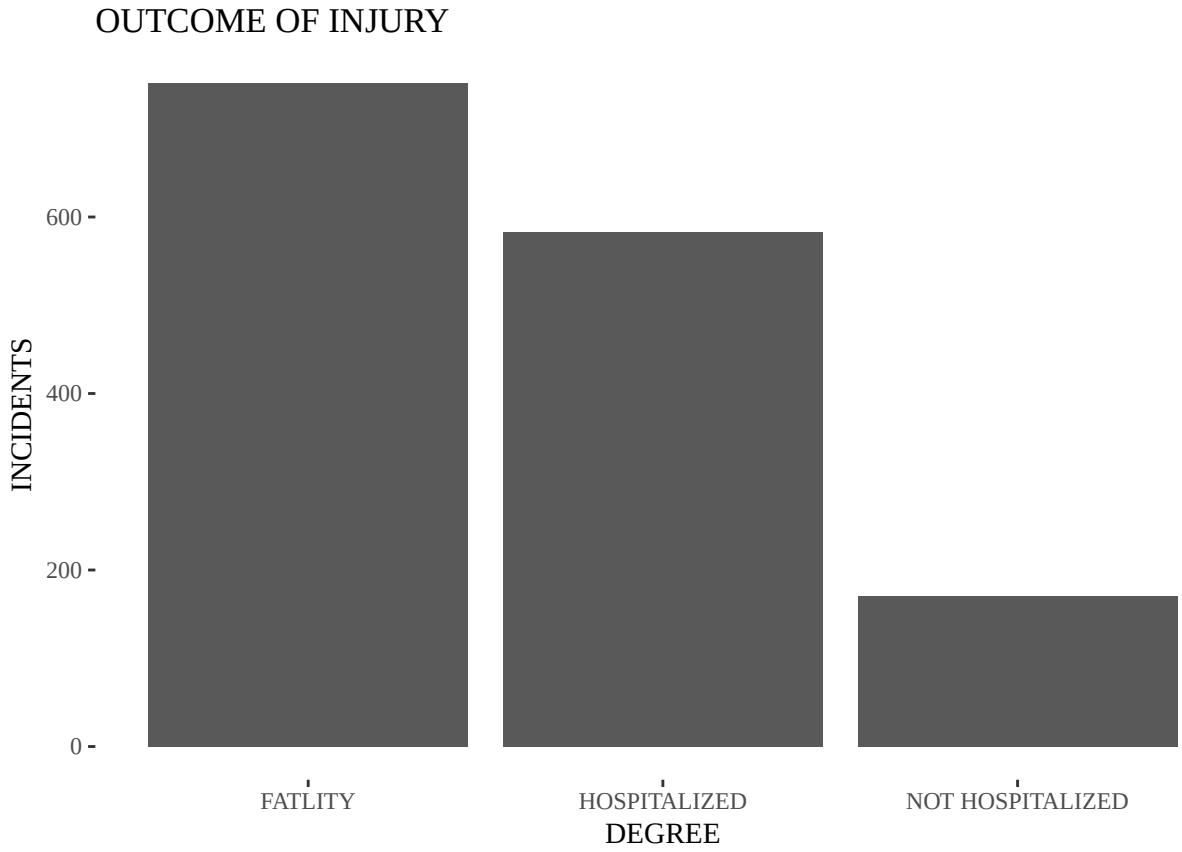
As we did with the SIC codes and HAZSUB codes, using the *acc.dbf* code lookup file, all coded values in the accidents and violations files were replaced by their corresponding strings. The data tables are now large tables of categorical data. This may make them difficult to work with, but one can easily factor each column and assign a numerical value to each level as needed.

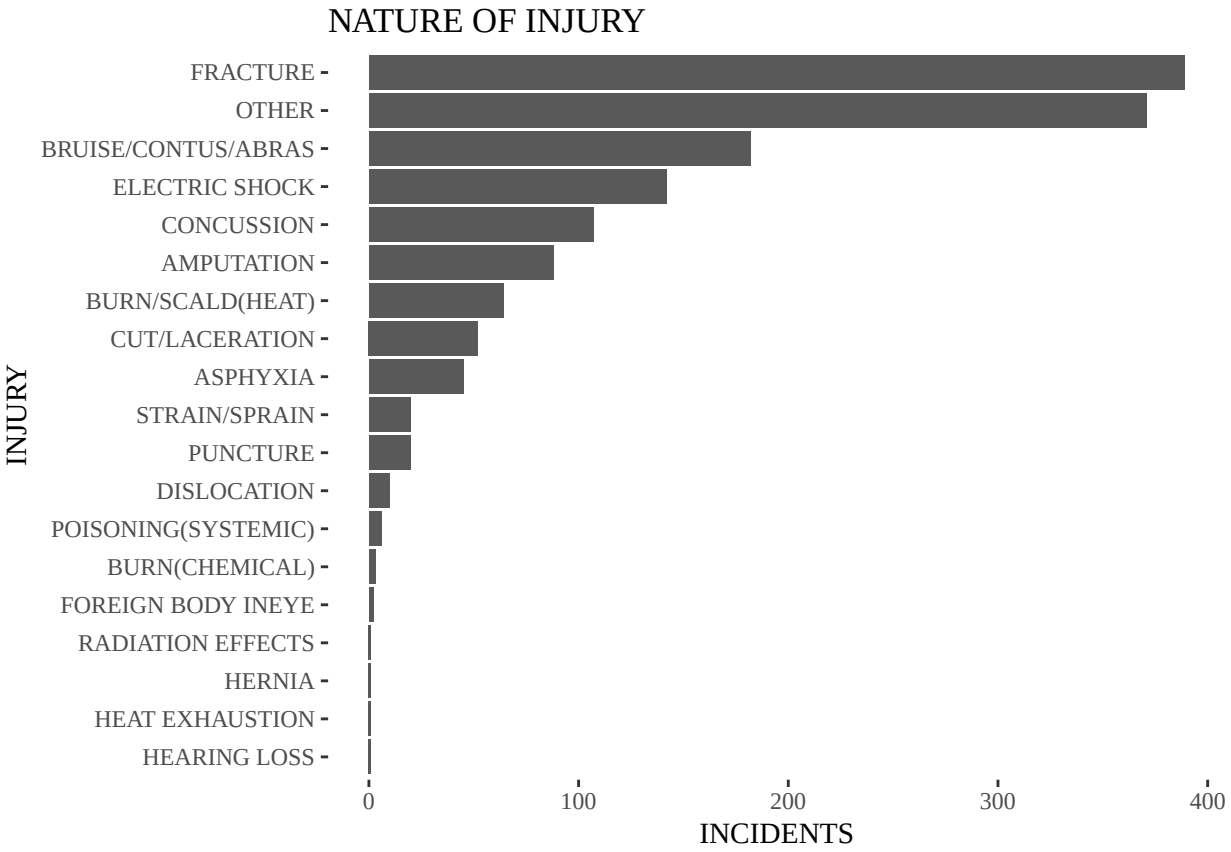
III. Exploratory Analysis

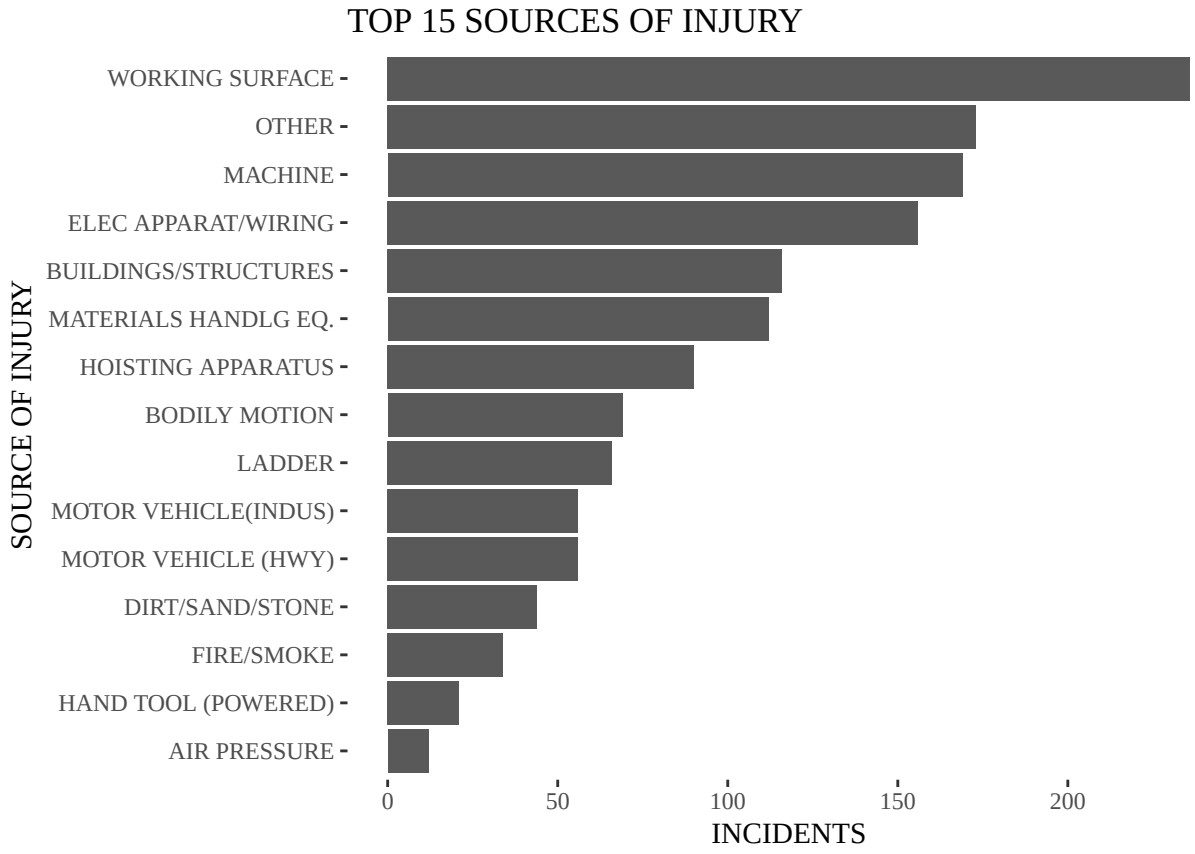
A. Frequency of appearance

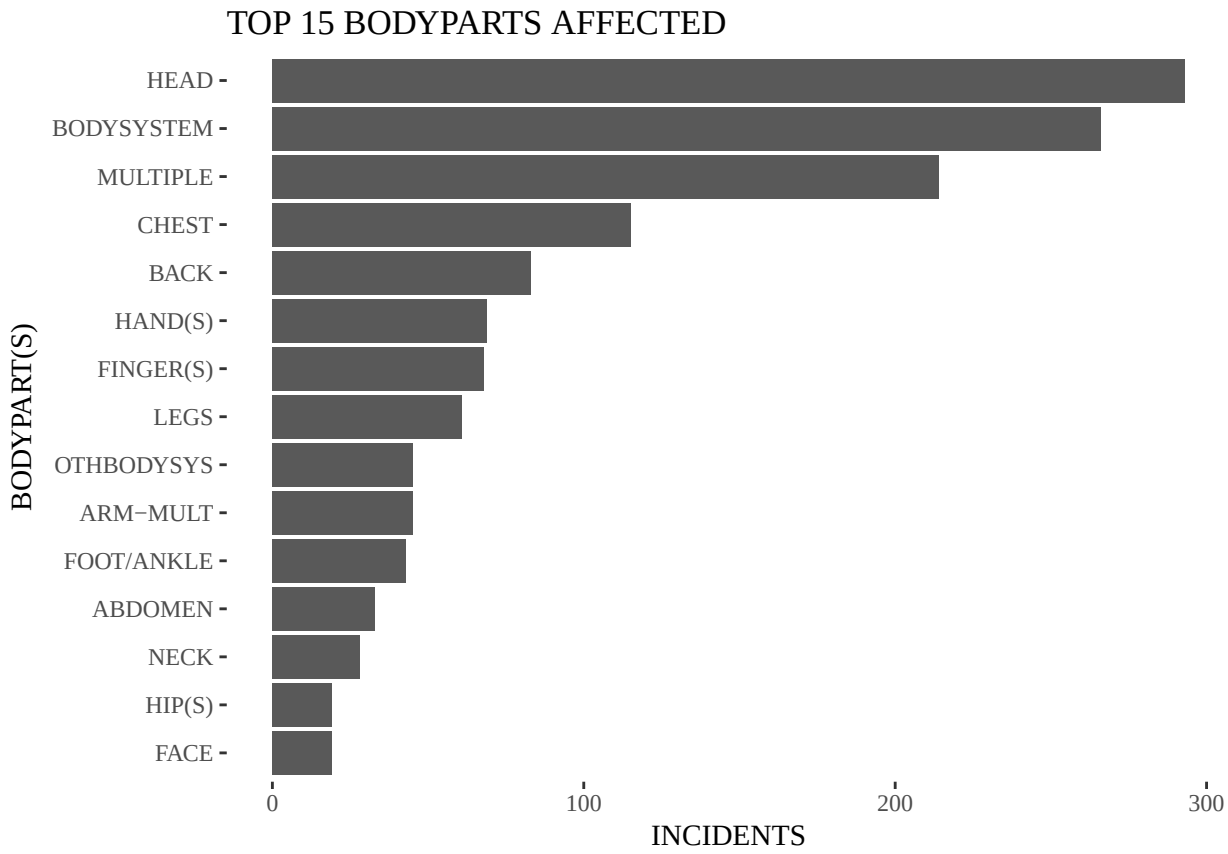
The frequency of appearance for the variables in *accid.dbf* were measured. Histograms of a few variables are shown below.

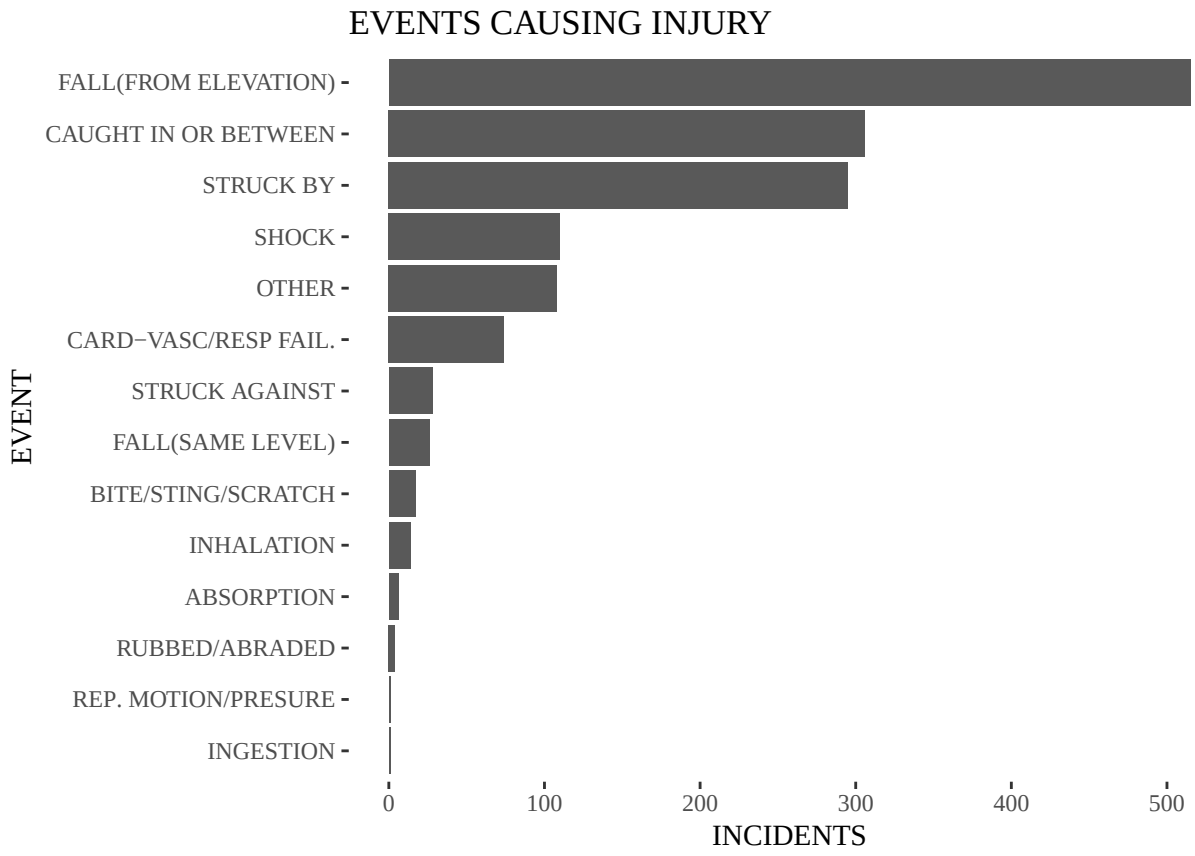
Note: For variables with more than 30 levels, only the top 15 are shown.

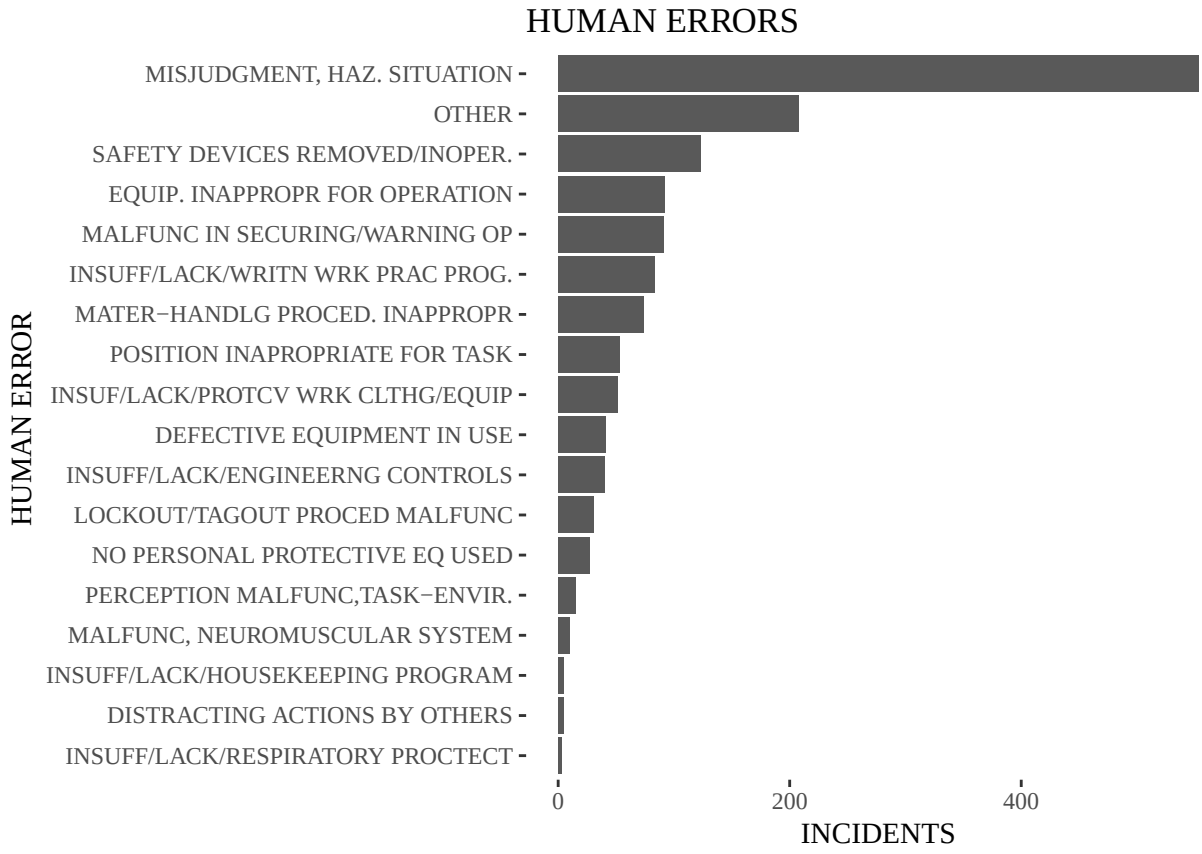

OUTCOME OF INJURY
INCIDENTS
600
400
200
0
FATLITY
HOSPITALIZED
NOT HOSPITALIZED
DEGREE

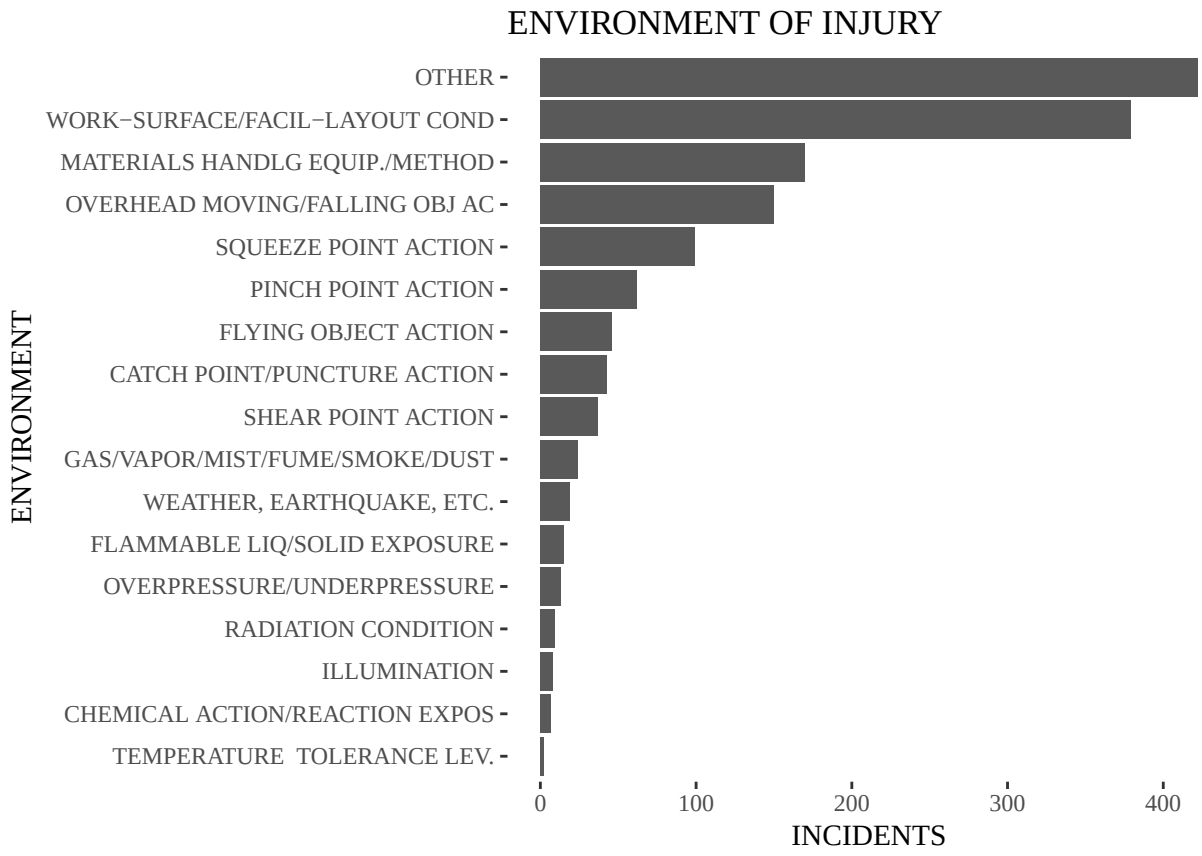

NATURE OF INJURY
FRACTURE
OTHER
BRUISE/CONTUS/ABRAS
ELECTRIC SHOCK
CONCUSSION
AMPUTATION
BURN/SCALD(HEAT)
CUT/LACERATION
ASPHYXIA
STRAIN/SPRAIN
PUNCTURE
DISLOCATION
POISONING(SYSTEMIC)
BURN(CHEMICAL)
FOREIGN BODY INEYE
RADIATION EFFECTS
HERNIA
HEAT EXHAUSTION
HEARING LOSS
INJURY
0
100
200
300
400
INCIDENTS

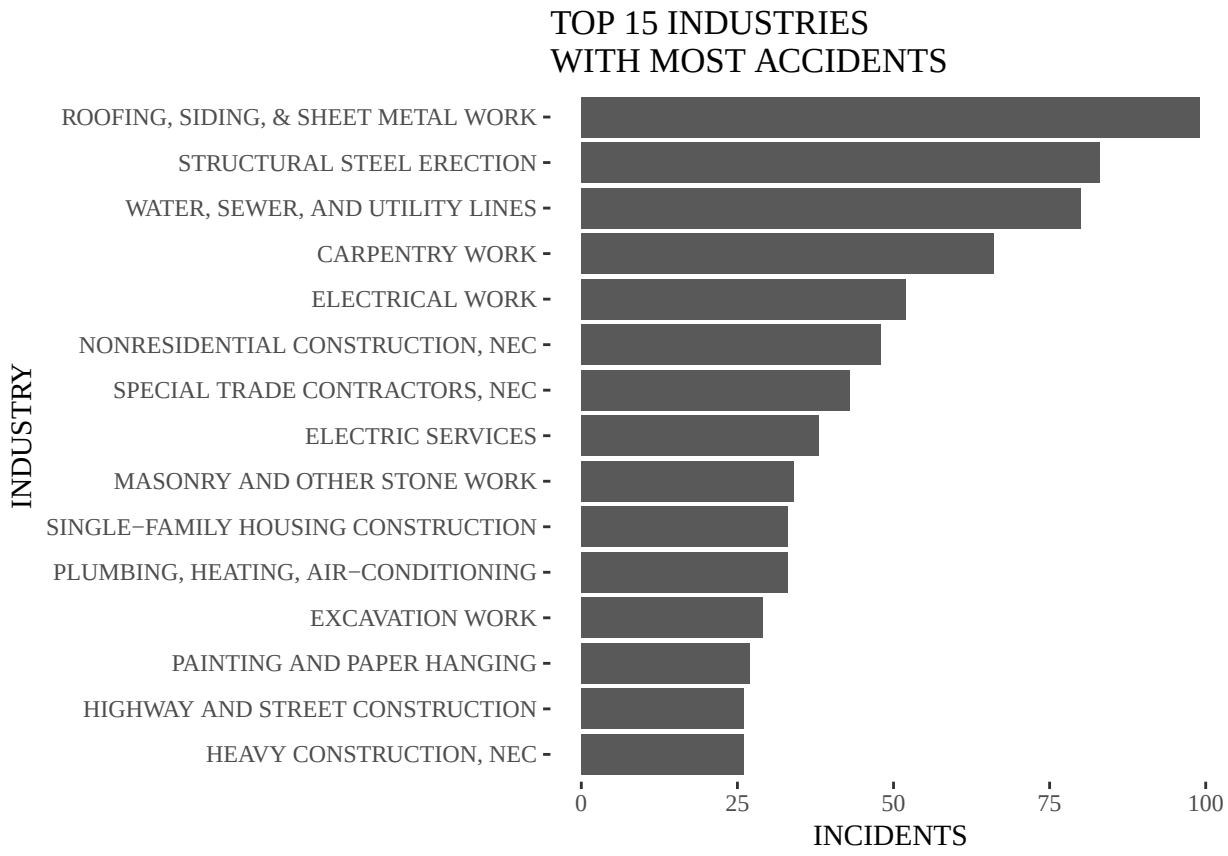

TOP 15 SOURCES OF INJURY
SOURCE OF INJURY
WORKING SURFACE
OTHER
MACHINE
ELEC APPARAT/WIRING
BUILDINGS/STRUCTURES
MATERIALS HANDLG EQ.
HOISTING APPARATUS
BODILY MOTION
LADDER
MOTOR VEHICLE(INDUS)
MOTOR VEHICLE (HWY)
DIRT/SAND/STONE
FIRE/SMOKE
HAND TOOL (POWERED)
AIR PRESSURE
0
50
100
150
200
INCIDENTS

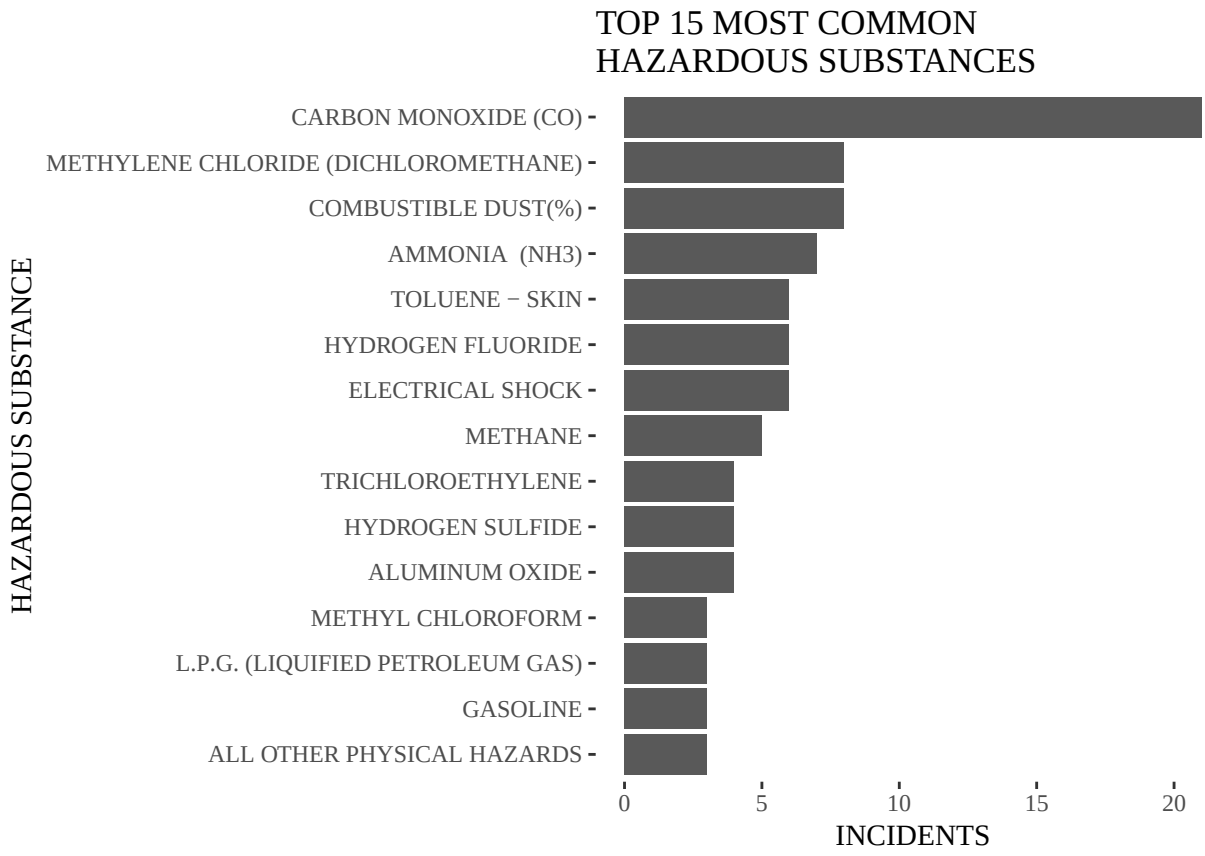

TOP 15 BODYPARTS AFFECTED
BODYPART(S)
HEAD
BODYSYSTEM
MULTIPLE
CHEST
BACK
HAND(S)
FINGER(S)
LEGS
OTHBODYSYS
ARM−MULT
FOOT/ANKLE
ABDOMEN
NECK
HIP(S)
FACE
0
100
200
300
INCIDENTS


EVENTS CAUSING INJURY
FALL(FROM ELEVATION)
CAUGHT IN OR BETWEEN
STRUCK BY
SHOCK
OTHER
CARD−VASC/RESP FAIL.
STRUCK AGAINST
FALL(SAME LEVEL)
BITE/STING/SCRATCH
INHALATION
ABSORPTION
RUBBED/ABRADED
REP. MOTION/PRESURE
INGESTION
EVENT
0
100
200
300
400
500
INCIDENTS

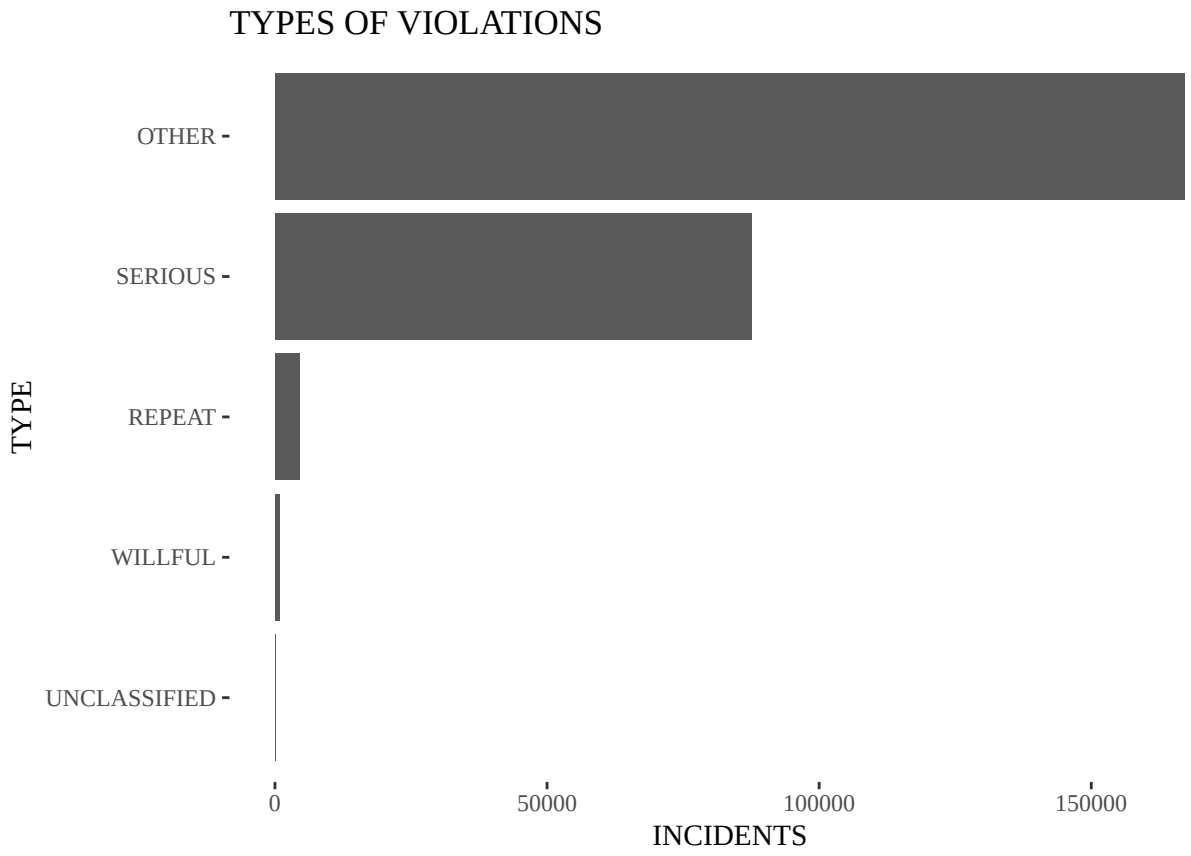

HUMAN ERRORS
MISJUDGMENT, HAZ. SITUATION
OTHER
SAFETY DEVICES REMOVED/INOPER.
EQUIP. INAPPROPR FOR OPERATION
MALFUNC IN SECURING/WARNING OP
INSUFF/LACK/WRITN WRK PRAC PROG.
MATER−HANDLG PROCED. INAPPROPR
POSITION INAPROPRIATE FOR TASK
INSUF/LACK/PROTCV WRK CLTHG/EQUIP
DEFECTIVE EQUIPMENT IN USE
INSUFF/LACK/ENGINEERNG CONTROLS
LOCKOUT/TAGOUT PROCED MALFUNC
NO PERSONAL PROTECTIVE EQ USED
PERCEPTION MALFUNC,TASK−ENVIR.
MALFUNC, NEUROMUSCULAR SYSTEM
INSUFF/LACK/HOUSEKEEPING PROGRAM
DISTRACTING ACTIONS BY OTHERS
INSUFF/LACK/RESPIRATORY PROCTECT
HUMAN ERROR
0
200
400
INCIDENTS


ENVIRONMENT OF INJURY
OTHER
WORK−SURFACE/FACIL−LAYOUT COND
MATERIALS HANDLG EQUIP./METHOD
OVERHEAD MOVING/FALLING OBJ AC
SQUEEZE POINT ACTION
PINCH POINT ACTION
FLYING OBJECT ACTION
CATCH POINT/PUNCTURE ACTION
SHEAR POINT ACTION
GAS/VAPOR/MIST/FUME/SMOKE/DUST
WEATHER, EARTHQUAKE, ETC.
FLAMMABLE LIQ/SOLID EXPOSURE
OVERPRESSURE/UNDERPRESSURE
RADIATION CONDITION
ILLUMINATION
CHEMICAL ACTION/REACTION EXPOS
TEMPERATURE TOLERANCE LEV.
ENVIRONMENT
0
100
200
300
400
INCIDENTS


TOP 15 INDUSTRIES
WITH MOST ACCIDENTS
ROOFING, SIDING, & SHEET METAL WORK
STRUCTURAL STEEL ERECTION
WATER, SEWER, AND UTILITY LINES
CARPENTRY WORK
ELECTRICAL WORK
NONRESIDENTIAL CONSTRUCTION, NEC
SPECIAL TRADE CONTRACTORS, NEC
ELECTRIC SERVICES
MASONRY AND OTHER STONE WORK
SINGLE−FAMILY HOUSING CONSTRUCTION
PLUMBING, HEATING, AIR−CONDITIONING
EXCAVATION WORK
PAINTING AND PAPER HANGING
HIGHWAY AND STREET CONSTRUCTION
HEAVY CONSTRUCTION, NEC
INDUSTRY
0
25
50
75
100
INCIDENTS


TOP 15 MOST COMMON
HAZARDOUS SUBSTANCES
HAZARDOUS SUBSTANCE
CARBON MONOXIDE (CO)
METHYLENE CHLORIDE (DICHLOROMETHANE)
COMBUSTIBLE DUST(%)
AMMONIA (NH3)
TOLUENE − SKIN
HYDROGEN FLUORIDE
ELECTRICAL SHOCK
METHANE
TRICHLOROETHYLENE
HYDROGEN SULFIDE
ALUMINUM OXIDE
METHYL CHLOROFORM
L.P.G. (LIQUIFIED PETROLEUM GAS)
GASOLINE
ALL OTHER PHYSICAL HAZARDS
0
5
10
15
20
INCIDENTS

The frequencies of the violation types are shown below. Most of the violations fall into the "OTHER" category in the raw data file. The file contains more than 250,000 observations.



B. Most dangerous places by total number of deaths

The top five establishments with the most deaths are shown below.

SITE	INCIDENTS
GENERAL DYNAMICS QUINCY SHIPBU	6
BOSTON EDISON CO	5
GENERAL ELECTRIC CO	4
MASSACHUSETTS ELECTRIC CO	3
PERINI CORP	3

The top five establishments with the most deaths not caused by human error are shwn below. This table was produced after filtering *accid.DBF* by human factors not related to human error. These were:

1. EQUIP. INAPPROPR FOR OPERATION
2. INSUF/LACK/PROTCV WRK CLTHG/EQUIP
3. INSUFF/LACK/ENGINEERNG CONTROLS
4. INSUFF/LACK/HOUSEKEEPING PROGRAM
5. INSUFF/LACK/RESPIRATORY PROCTECT
6. INSUFF/LACK/WRITN WRK PRAC PROG
7. DEFECTIVE EQUIPMENT IN USE
8. MALFUNC IN SECURING/WARNING OP
9. MATER-HANDLG PROCED. INAPPROPR
10. SAFETY DEVICES REMOVED/INOPER.

SITE	INCIDENTS
GENERAL DYNAMICS QUINCY SHIPBU	2
GENERAL ELECTRIC CO	2
PAR ELECTRICAL CONTRACTORS INC	2
PLYMOUTH & BROCKTON STREET RAI	2
A & M ROOFING & SHEETMETAL CO.	1

C. Interpretation

All of the distributions seem to be skewed. However, they are skewed towards the incident that is most logical in the context of these workplaces. Most of these companies handle building or repair or technical work that involve physical labor. As a result, one expects fractures, falls, damage to upper body/head, etc., to be most common. These are indeed the most common incidents. Accidents involving hazardous substances a slightly difficult to analyze because one cannot be sure what substance is most toxic. Still however, that carbon monoxide is the most toxic substance agrees with the fact that carbon monoxide is one of the more prevalent toxic gases in any establishment.

Based on the plots above the frequency of incidents, we may infer that most accidents are due to falls from elevated working surfaces that fracture mostly the upper body and head, and are due to workers' misjudgment of hazardous situations in companies that specialize in roofing, siding, and sheet metal work.

That a "misjudgement of hazardous situation" is by far the most common human error provides a great deal of insight as to whether a not a workplace is truly dangerous.

That the 'OTHER' violation type is the most common points to the incompleteness of the original raw data files. More data is needed to assign violation type(s) to these incidents.

The two tables on the previous page are an example of how the data set can be manipulated. One can filter deaths by removing human error and then calculate the proportion of deaths that are not caused by human error in determining the hazards in a given workplace. Based on how these data are cleaned, the analyst can choose from a variety of approaches to analyze the data.