OSHA Data Cleanup and Exploratory Analysis

Lubov McKone

MA 415 Data Science in R

Cleanup

I started my analysis with some basic data cleanup – recoding of missing values, and deletion of superfluous variables. The three tables I chose to focus on were the accident table, the OSHA table, and the violation table.

In the accidents table, I retained only the variables related to the "story" of the accidents: The victim's name, age, sex, and the details of the accident such as which body part was affected, a hazardous substance if one was involved, and human and environmental factors that contributed to the accident.

I also retained the "ACTIVITYNO" field – a supposedly unique identifier for each inspection. I ran into some trouble with this field because the values are not actually unique – some inspection records were so long that OSHA had to break them up into multiple records with the same ACTIVITYNO. To solve this problem, I created a new table of distinct accidents by subsetting the original accidents table into a table that contained only the records related to the first occurrence of every unique ACTIVITYNO. I used the duplicated()function, which returns an list of boolean values specifying which of the indices have not already appeared in the list, to achieve this. It is important to note that not duplicated ACTIVITYNOs are meaningless – many exist because multiple people were hurt in a single accident. However, this is not the case for all duplicate ACTIVITYNO values, so I decided to make a new table of distinct "accidents" regardless of how many people were involved in each. The original table was retained so that if analysts wanted information about how many people were injured in each accident, they could retrieve it.

I used the OSHA table to record information about the establishments that were inspected, such as the establishment name, county and city it was located in, and industry. I also retained the ACTIVITYNO field in this table so that I could join it to other table.

Table Subsetting and Joining

I created a small table called estabs that gives the correlation between ACTIVITYNO and ESTABNAME. By left joining the table of distinct accidents to estabs, I created a table called estabaccidents in which I was able to see which establishment each accident had occurred at. I created another table called acount that counted the number of times each establishment name occurred in estabaccidents, yielding Table 1, which shows how many accidents occurred at each establishment. I arranged the table in descending order and retrieved the top 10 records:

ESTABNAME	numAccidents
BOSTON EDISON CO	7
GENERAL DYNAMICS QUINCY SHIPBU	7
GENERAL ELECTRIC CO	5
MODERN CONTINENTAL CONSTRUCTIO	4
UNITED PARCEL SERVICE	4
AVERY DENNISON	3
BOSTON EDISON COMPANY	3
KEN'S FOODS, INC.	3
MALDEN MILLS INDUSTRIES	3
MASSACHUSETTS ELECTRIC	3

Table 1: Number of accidents per establishment

I found the average number of accidents per establishment by taking the mean of the numAccidents column in Table 1, and found that the average number of accidents per establishment is 1.05, with a standard deviation of 0.349, so 7 accidents occurring in one establishment is a big deal.

Exploratory Analysis

To take a look at other factors that might be correlated with the number of accidents per establishment, I made some quick plots:

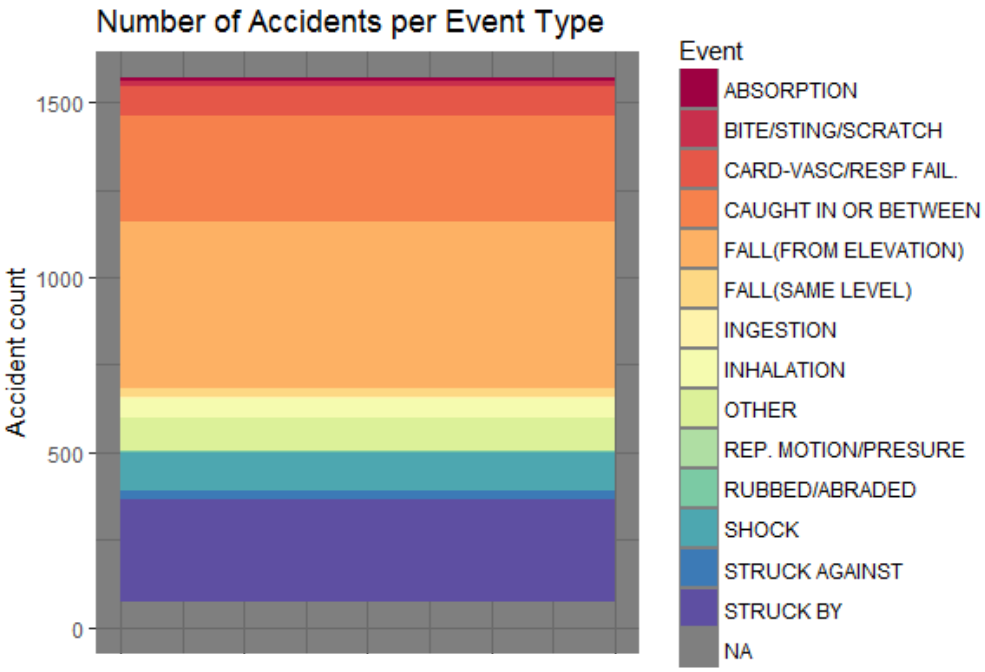

Figure 1: Causes of accidents in Massachusetts.

In the plot above, we see that the most common events causing accidents are falling from an elevation, getting caught in or between something, or getting struck by something. This plot could be recreated using only accidents that resulted in fatality, hospitalization, etc. to get a better sense of accident causes for different degrees.

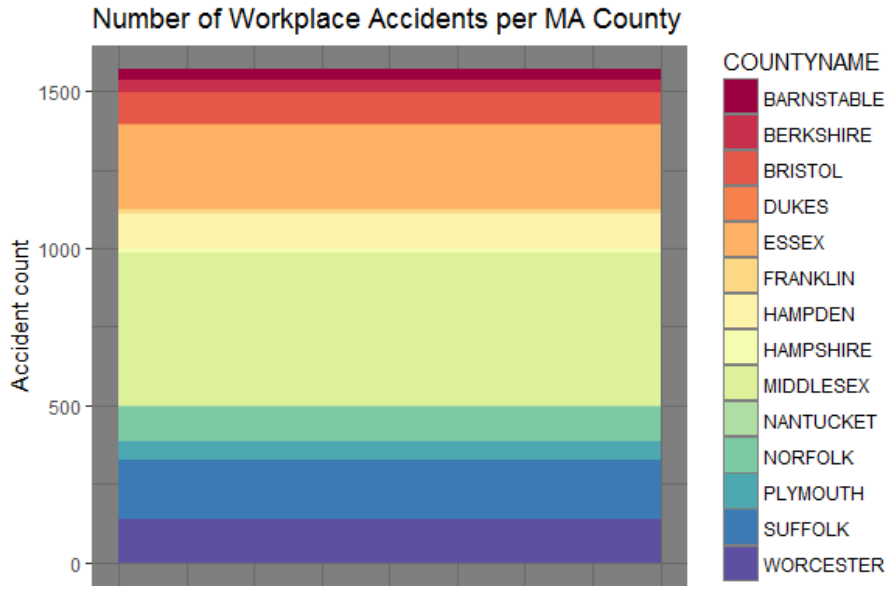

Figure 2: Number of workplace accidents per Massachusetts County

The above plot shows how many workplace accidents occur per county in Massachusetts. Note that the number of accidents is calculated from the distinct ACTIVITYNO values, so this is a count of accident "instances," not a measure of number injured or killed.

I wanted to check if there was any difference in the causes of accidents in establishments that had had a lot of accidents (>3) versus the causes of accidents in those that had very few (<=3).

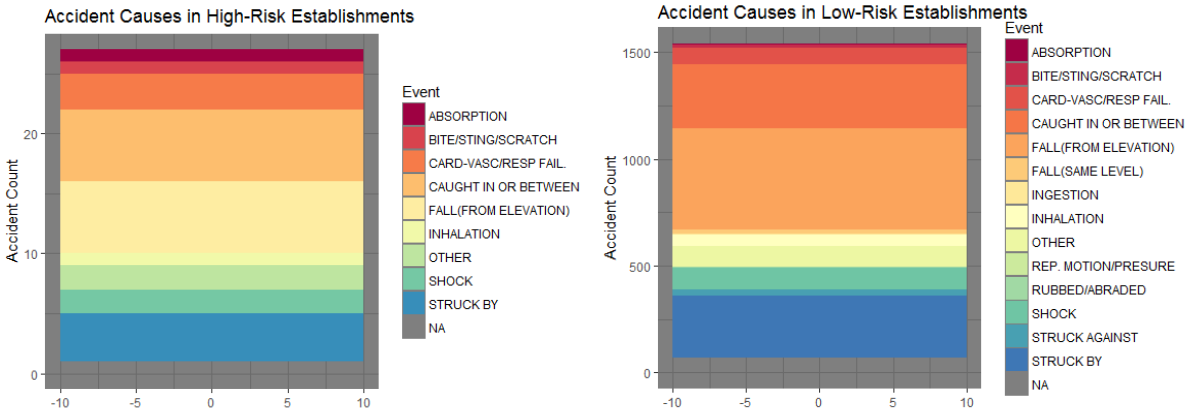

Figure 3: Accident causes in low- versus high-risk establishments

We can see that the plots look mostly the same, accounting for the color variation, but interestingly certain categories, such as ingestion and falling from the same level, were totally absent from the establishments that had more than 3 accidents.

Final Remarks

This dataset is extensive and versatile. It would be impossible to create all useful subtables, but the table I create in this data cleanup comprise a good, legible, meaningful starting set. The clean data are easy to subset and join with one another, creating a great starting point for analysis.

The one major problem in the data I didn't completely solve was the issue of non-distinct ACTIVITYNO records in the Accidents table. Since it seemed some of the duplicates were meaningful while others weren't, it was difficult to know how to treat the problem. However, I believe the table of distinct accident records is the best solution because it expresses how many distinct times an establishment had an accident. One establishment having 7 accidents that injured 1 person each might actually be more dangerous to work at than an establishment that had one major accident where 10 were injured – it's less likely that another accident will occur at the latter establishment.

Code

All code and documentation for this project can be found by cloning:

https://github.com/lmckone/Rmidterm