Midterm Project

Jingwen Zhang

Project Topic:
"What is the least safe place in MA"

Summary:
My table is created from 2 data files: Osha and Viol files. Together I have 10 variables that are considered to be potentially helpful to the study of the topic. In the R code, the dataframe called "final" is the final table that I get. I use some techniques that are mentioned in Table 4 and Table 6 from the documents written by Hadley Wickham when I need to find the frequency of the activities. One thing I want to mention is that, although I have the code that cleans the data for Accid, I do not use Accid because there are only 2147 observations and if I combine these observations to the osha data, it will create many NA data, which is not very helpful to build the model.

Variables:

1. name: Name of the area in MA.
2. ACITIVITYNO: Unique activity number of the inspection.
3. Zipcode: Zip code for the place.
4. Address: Street address.
5. Date: Latest data activity applied against record.
6. Preptime: Time spent preparing for an inspection.
7. Industry: Classification of industry
8. EstabName: Name of establish men
9. violType: Violation types(other, serious, repeat, unclassified and willful)
10. viol_Freq: Number of times certain type of violations happen.

Steps:

1. After looking at the Readme.dbf, I only choose to use **Osha**, **Viol**, **Accid** dbf files based on our project topic.
   (I do not use **Debt**, **Admpay, Prog, Relact, History** and **Optinfo** because they are unrelated to our project topic.
   I do not use **Hazub** because we do not care too much about the hazardous substances involved.)

2. Looking at the Layout for each dbf file (Osha, Viol, Accid) and do some operations in R, I select several potentially effective variables from these dbf file.

3. Clean the data for each file, I create 3 tables. For Viol data, I replace the code by the type of violations they are corresponding to. Also, in the osha data, I find the area name base on the city and county code and add the names to the osha data.
   Combining the 3 table by the variable **ACTIVITYNO**.

4. After left_join all the tables, I get 14 variables. I look through each variable again and find that there are a lot of NA in my variables: **envir_fac**, **human_fac**, **sourc and nature**. I decide not to use these 4 variables, and only combine the two data: Osha3 and violations.

5. Then I rename all the variable names and change the viol_Freq into numbers instead of characters.