ANLY501 Project Assignment1

Yanchen Wang
Jun Wang
Junke Wang

Data Science Problem

The problem we plan to investigate is the relationship between the weather conditions and the number of motor vehicle crashes, injuries, and fatalities that occurred in Washington, D.C. between year 2013 and 2016. It is meaningful to study this relationship because the explicit model we conduct would allow relative people to predict or forecast the number of car accidents based on the specific weather condition. As a result, people such as police officers and emergency medical service personals would be better prepared accordingly to different daily weather condition.

There many are precious researches studied the effect of weather conditions on car crashes during several years' period over different regions of the United States. Most of the studies used monthly or yearly data (Brijs, Karlis, and Wets, *Studying the effect of Weather Conditions On Daily Crash Counts Using a Discrete Time Series Model*). In our investigation, we are planning to use and analysis the impact of weather conditions on the daily number of motor vehicle crashes, injuries, and fatalities. In addition, instead of studying the entire country, we plan to focus on Washington D.C. as a representation of the urban area, since we believe that there is a significant difference in results between urban area and rural area.

Potential Analyzes that Can Be Conducted Using Collected Data

For our investigation, we plan to collect data set from the official government websites (http://opendata.dc.gov/datasets/crashes-in-dc/data and https://www.ncdc.noaa.gov) for both our daily car accidents and weather condition, since the official government websites would provide relatively reliable and complete data set which allow us to conduct future data cleaning and analysis. Also, the API for the data pages are also available to the public, so the data is accessible compare to other sources.

For the car accidents data set, the variables should contain the date of occurrence, the location of the accident, number of car, number of injury, and number of death. On the other hand, for the weather conditions data set, the variables should contain date, location of the weather stations, minimum temperature, maximum temperature, precipitation, and snowfall amount. All the variables should be useful since they would help us to have a better understanding of the relationship we want to investigate. First of all, the date and location variables allow us to match the weather condition with the car accidents. Then, the rest of the variables for car accidents data (number of car, number of injury, and number of death) could let us define or verified the scale of the accident, whether it is a major accident or a minor accident. Similarly, the rest of the variables for weather condition data (max/min temperature, precipitation, and snowfall) would allow us to interpret and estimate the specific weather condition during that date.

The possible directions and hypotheses that we may be able to investigate with the data we collected are as follow:

- The weather conditions have great impact on the number of car crashes, injuries, and fatalities.
- Weather conditions and number of car crashes, injuries, and fatalities have a positive correlation.
- Higher temperature has a negative influence to drivers' condition.
- Extreme temperature has a positive correlation to number of accidents.
- When the rainfall amount increase, the number of car crashes, injuries, and fatalities also will increase.
- The rainfall has a more negative influence to drivers' condition than the snowfall.
- When the same number of amount, rainfall will cause more number of car crashes, injuries, and fatalities than sow.
- The relationship between road condition and number of car accident.

Data Issues

The first problem we encountered is that the number of row for the weather condition data in year 2013-2016 is around 1,460 which is far less than the 5,000 requirement. In order to solve this problem, instead of using one location for the weather station, we plan to use three different locations of weather stations in Washington, D.C. area. The three weather stations we plan to use are GHCND: USC00186350: NOAA Station at National Arboretum, GHCND: USC00182325: NOAA Station at Dalecarlia Reservior, and GHCND:US1VAFX0063: NOAA Weather Station at Alexandria, VA. For those three stations, National Arboretum is located at east of Washington, D.C, Dalecarlia Reservior is located at Northwest corner of DC, and Alexandria, VA is close to southern corner of DC.

Except not enough entries for the weather condition data, following is the list of noise and missing values in our data set:

- Some entries are missing the amount of rainfall
- Some entries are missing the amount of snowfall
- Some entries are missing the minimum or maximum temperature values
- Some entries are missing the location for car accident
- For some entries, the amounts of rainfall are not consistence among three stations (one of the station has unreasonable value)
- For some entries, the amounts of snowfall are not consistence among three stations (one of the station has unreasonable value)
- For some entries, the min/max temperature are not consistence among three stations (one of the station has unreasonable value)
- For some entries, the number of car evolve in the accident is zero
- For some entries, the date of report accident is ahead of the date of occurrence
- For some entries, the sum of injuries adds up incorrectly
- For some entries, the accidents might be duplicated

Collecting New Data

As mentioned in the second section, the car accident and weather data set we collected is from "http://opendata.dc.gov/datasets/crashes-in-dc/data" and "https://www.ncdc.noaa.gov/cdo-

web/webservices/v2", where the car accident data contains 28 attributes and weather condition data contains 4 attributes (one of the attributes is called "data type" that contains 15 types of data which are untimely serve as 4 different attributes after we rearrange and clean the data).

Data Cleanliness

For each of the attributes in both data sets, we compute the fraction of missing values, as well as the fraction of noise values. As a result, for the car accident data, it rarely has missing values, however, many records are considered noise values. Especially, some of the entries contain potential incorrect dates where the reporting date of the accident was prior to the date that the accident was occurred. In this data set, around 20% of records has this issue. In the future, we will look into it and find out if they are really incorrect values.

On the other hand, for the weather condition data set, the attribute "snow" has a lot of missing values. We fixed this issue by using the minimum temperature for that day. Since snowing cannot occur when the minimum temperature is above freezing temperature (32 degrees Fahrenheit). After adding this restriction, many missing values are replaced and there are less than 5% of records in each attribute missing. To access the fraction of noise values, since Washington, D.C. is a fairly small area, we expect the amount of rainfall, the amount of snowfall and the minimum and maximum temperature do not have significant differences for the same day across three weather stations. So we looked into the data for each day to check the differences in precipitation, snowfall, and temperature across the three stations. If the difference is too big, we would look into that deeper and might take mean to replace the extreme value in further data cleaning step.

Data Cleaning

The three attributes we cleaned are "SPEEDING_INVOLVED" from car accident data, as well as the "PRCP" and "SNOW" from weather condition data.

For the attribute "SPEEDING_INVOLVED", in the original data set, except date, this attribute has a large number of noise values. In the data description, this attribute "indicates that if the reporting officer believed speeding was a factor for this car accident. This does not necessarily equate to participants being ticketed/cited for speeding" (DDOT, *Crashes in DC*). This attribute has some inconsistency when it was recorded. So, we changed this attribute into categorical (binary) data "0" or "1", where "0" mean speeding was not a factor and "1" means it was a factor. For the attribute "PRCP", we replaced missing values and noise values by using the mean of other two stations for that day. For the attribute "SNOW", we first set it to 0 if the minimum temperature was above 32 degrees (above freezing temperature) and then for those entries still missing value, we used the mean to replace the space.

After we cleaned the data and re-ran the data cleanliness program, all the three attributes were cleaner than before. For the attribute "SPEEDING_INVOLVED", the count for noise values was 296, and after cleaning, the count dropped to 0. For the attribute "PRCP", after cleaning, the count for missing values reduced from 1524 to 3 and the count of noise values reduced from 6 to 0. Finally, for the attribute "SNOW", after cleaning, the count for missing values reduced from 35 to 0, and the count for noise values reduced from 21 to 0. Therefore, after re-ran the cleaning program, the overall data quality improved.