DSGA-1007 Final Project User Guide

Jiayi Lu (jl6583)
Minzi Ji (mj1547)
Lei Lu (ll2850)

a. What the program does

This program will input a csv data set as a dataframe and conduct data analysis on this data set by providing users descriptive plots and regression analysis. The data used is the *NYPD Motor Vehicle Collisions* and the program will generate insightful plots about features such as the number of collisions, number of fatalities and types of vehicles, contributing factors given different boroughs and time period.

b. How to run the program

To run this program, simply use “python main.py” in terminal.

By running main.py a welcome page will pop up after the data preparation is accomplished, on welcome page you should see a brief instruction on how to get access to different functionalities, follow the instructions and the program will generate plots and tables

At any time while the program is running you can exit by using ctrl+C or ctrl+D, it is safe using these commands.

c. What input the program takes

This program will input a csv data set as dataframe and takes inputs from users. User’s input can determine which information the program will generate. Users are able to enter a date range on which the program will operate and display the data. The program provided as adequate format checking as possible so most of the possible date time entries are accepted. For example, the program will ask users’ interest of time period, and report the number of fatalities during this time period.

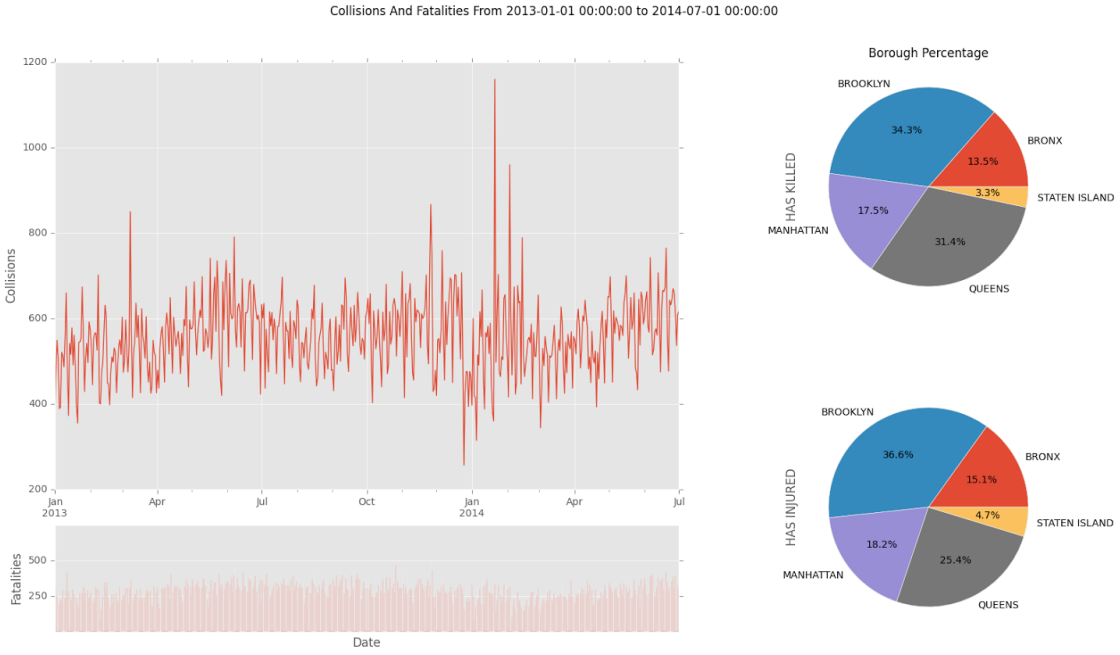
d. If external data sets are used, how to obtain them

The external data set “NYPD Motor Vehicle Collisions” used in this project is provided by New York Police Department(NYPD) and can be found at the following URL:
https://data.cityofnewyork.us/NYC-BigApps/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

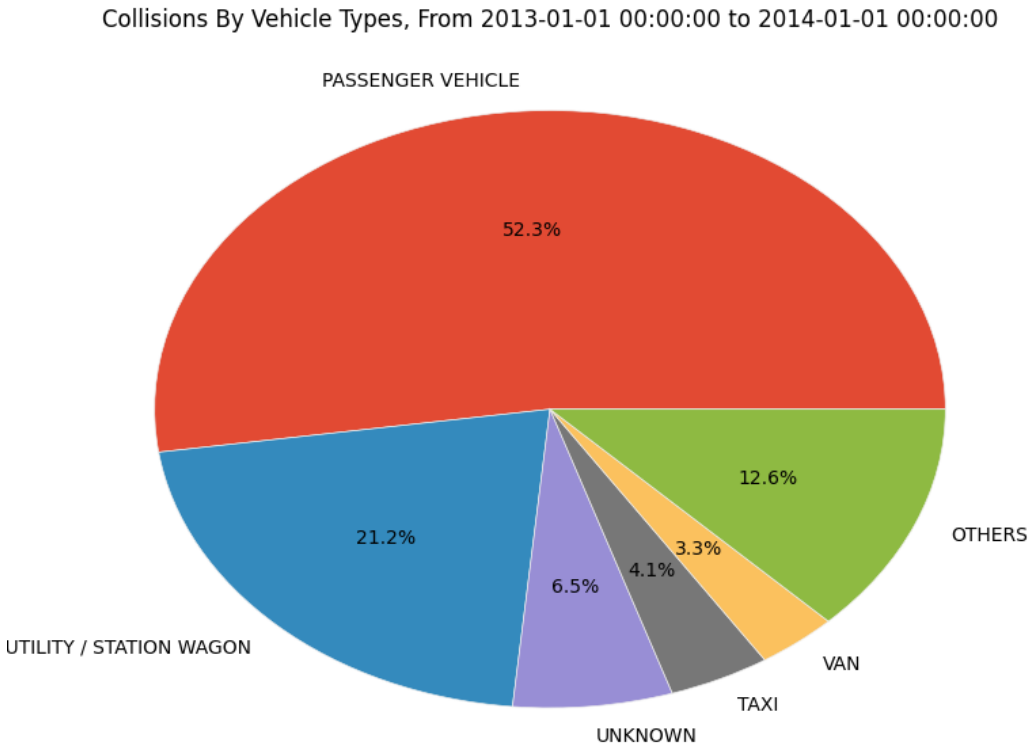
e. What output the program produces

This program a number of plots:

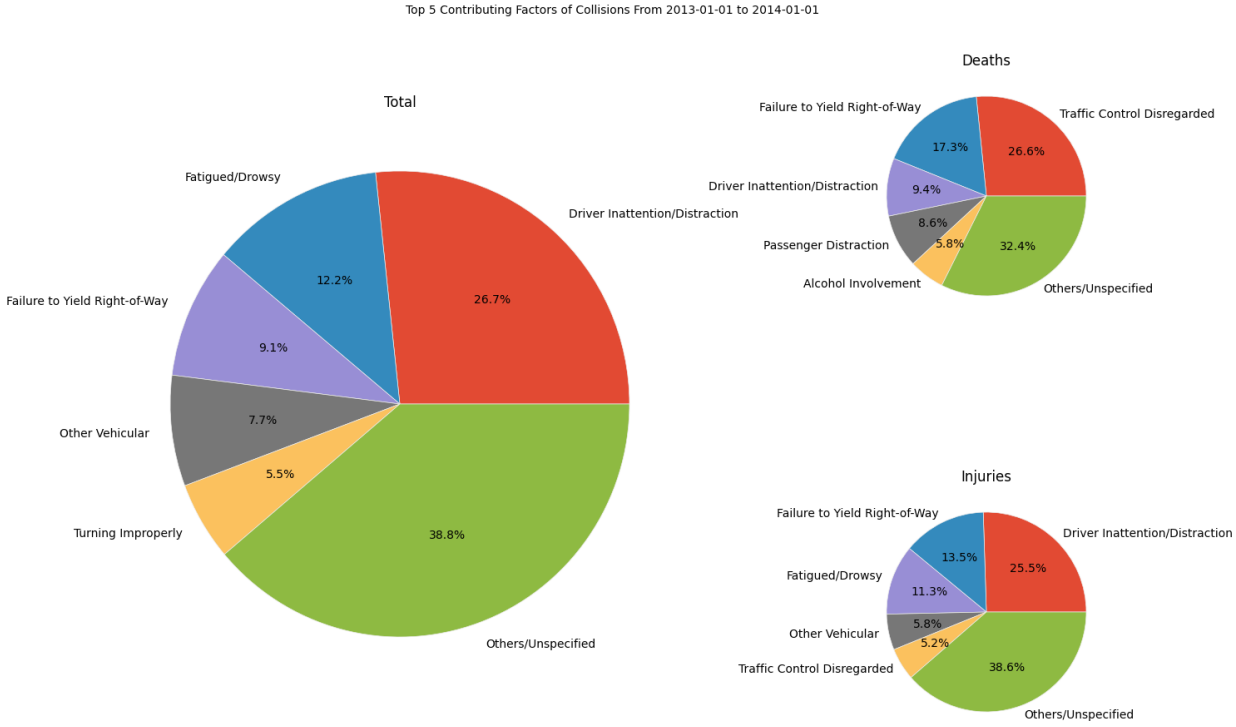
1. SUMMARY CHART ABOUT THE NYPD COLLISION DATA IN AN SPECIFIC PERIOD OF TIME



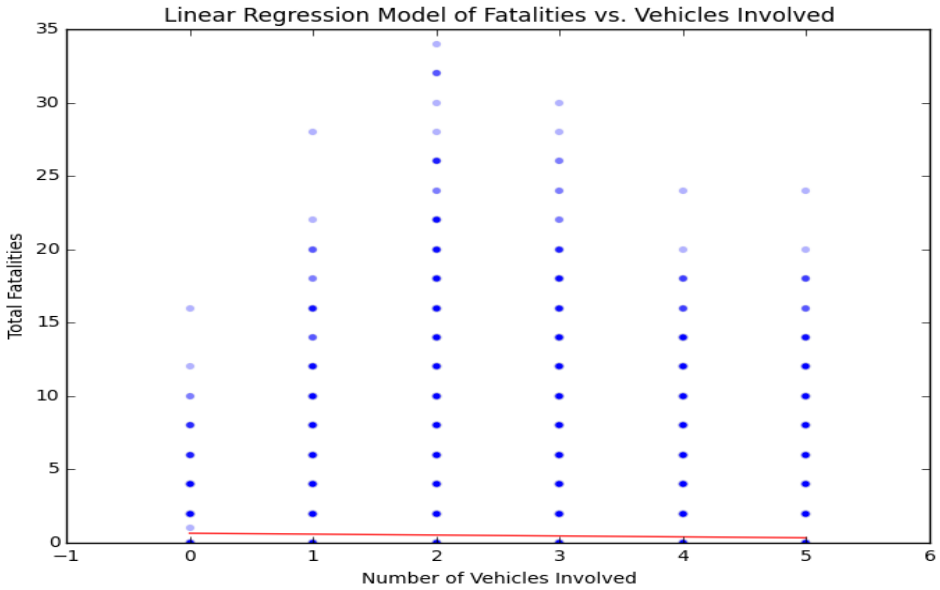
2. A PIE GRAPH SHOWING THE COMPOSITION OF COLLISIONS BY TYPE OF VEHICLES



3. THE TOP 5 CONTRIBUTING FACTORS OF COLLISIONS



4. REGRESSION ANALYSIS FOR THE FATALITY RATES AGAINST NUMBER OF INVOLVED VEHICLES IN A COLLISION



f. If there are any dependencies required to run the program, and if so, how to install them

The following python packages need to be installed in order to run the program：

1. numpy 1.8.1
2. pandas 0.14.0
3. matplotlib 1.3.1
4. statsmodels 0.5.0