**Data Analysis of Traffic Violation in Denton**

By

Lakshmi Deepa Sunkara

lakshmideepasunkara@my.unt.edu

**Purpose:**

The main purpose of this project is to conduct a data analysis on traffic violation data for the City of Denton in 2016. The main goal of this project is to reduce the amount of risk of accidents caused by the traffic violations. Through the data analysis we can gather information regarding the common type of traffic violations that take place at Denton. It is very important to know about the frequent places or the geographical locations where most of the traffic violation takes place because through this data the associated police department can take some precautions and actions to reduce the traffic violations and accidents.

**Data and Data Collection:**

The data for this project is about the traffic violations in Denton and the subject domain is traffic violation. This required data is available through “traffic.csv” which refers to a large amount of data represented in several number of records. This CSV file contains information about listing of all cases through docket number, citation number, date and address at which the violation took place, status of the violation case and its date, and the description of violation. Finally, it also gives the information regarding the action taken by the police officer in a response of that violation. The data values which are present in the CSV file are text, Number and date format. The size of this file is 205 KB.

**Methodology:**

The development of this project involves four main steps. They are data collection, data processing, data analysis and report generation. The collection of data involves in the process of gathering the data which is then processed for finding the number of traffic violations, the geographical locations where the violations take place and types of traffic violations happening at Denton. The collected data is available in traffic.csv. The processing of data is all about modeling the data present in traffic.csv file to find the frequent areas and the common type of traffic violations by using the column names and the values in it. The data is then analyzed using the features of python to find at what place the traffic violation had taken place. Finally, a report is generated from the results of processed data in order to know about the most common type of violations, frequent places where the traffic violations occurred and the actions taken by police department in response to the violation.

**Function description:**

In this project, we have considered predefined modules like matplotlib, numpy, collections and CSV, where CSV module is responsible for importing the CSV file in order to perform various operations. Once the CSV file is imported, reading and writing operations are performed on that file. These operations are performed using csv.reader() and csv.writer() functions respectively. The required data is present in the “traffic.csv” file and this file contains information regarding the violations which takes place not only in Denton but also in few other cities of Texas. Hence, some lists are created to append the information about all the violation details like what type of violation had taken place and at which location and what is the necessary action that has been taken by the police officer for that violation particularly for Denton. This is done using append() function. Next, we have imported the "Counter" function from collections module and applied it on the violation list in order to know about unique violations and their frequencies. These unique violations and their frequency counts are stored in two separate lists by using keys() and values() functions. The function keys() returns the names of elements and values() returns the frequency counts of elements into lists. The same process is performed on the location list in order to find out the most common geography locations and their frequency counts where the violations had taken place. The procedure is once again repeated in order to find out the most common actions and their frequency counts that are taken by the officer regarding the violations. All these procedures are executed in a single for loop. Further, another CSV file is created which is named as “collect.csv” to store all the information regarding violations, geographical locations and actions taken by the police department in response to the violations which occur only at Denton. This file is created by zipping three lists which are found above using zip() function. Now each violation, location and action values are stored in a single row of the CSV file. After storing all the values in separate rows, we have again calculated the geographical location frequency count and the actions taken by police department in response to the top most violation which is nothing but “speeding in 30-mile hour zone”. Finally, we have used the matplotlib and numpy modules for generating the histogram reports which represents the most common violations in Denton, most geographical locations where more violations take place and for most common actions taken by police department in response of violations. The plotting library matplotlib is used for generating the histogram and numpy is also used in generating the histogram to return the evenly spaced values at a specified interval through np.arange(len()) function. The functions plt.bar() and plt.xticks() are used to plot the histogram and plt.title() to lable the histogram. The histogram is then displayed using plt.show() function, saved using plt.savefig(), and closed using plt.close() function.

**Evaluation Results:**

After processing the above-mentioned analysis, we have found the following results:

1) The total number of traffic violations which occur in Denton are 508.

2) The type and frequency count of the top four most common traffic violations in Denton are SPEEDING IN 30 MILE HOUR ZONE = 86 times, SPEEDING = 79 times, DL NO DRIVER'S LICENSE = 51 times, and FAIL TO MAINTAIN FINANCIAL RESPONSIBILIT = 51 times.

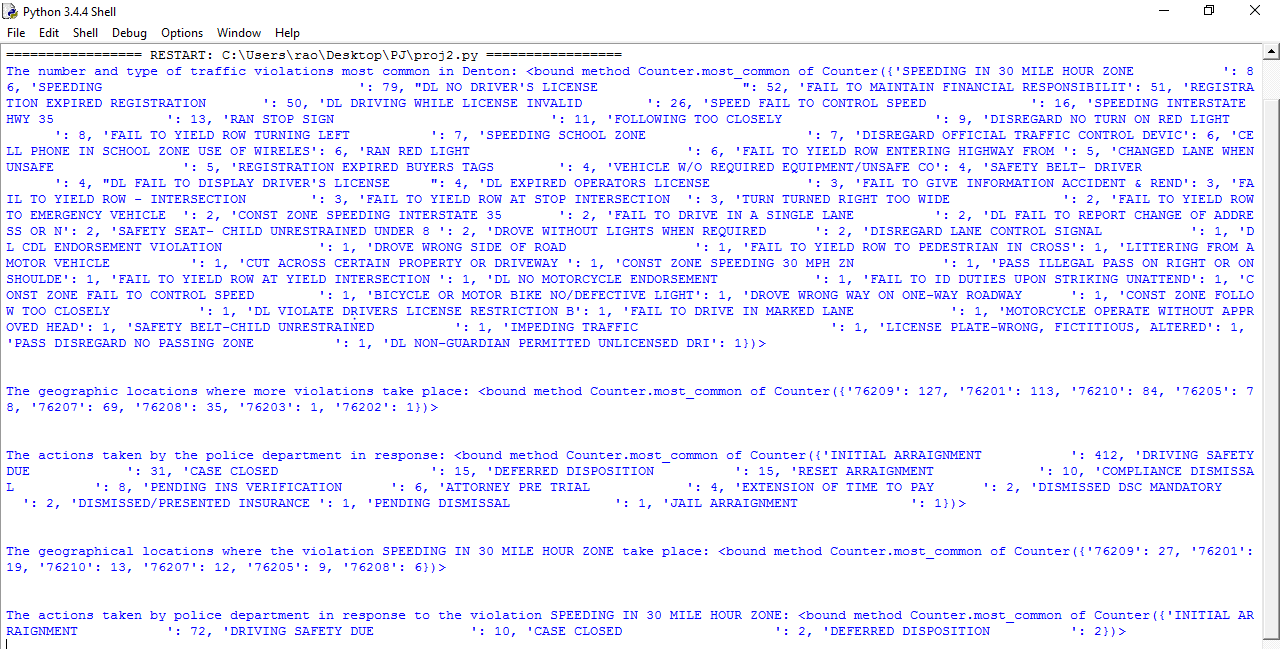
3) The frequency count of the top four geographical locations where more traffic violations take place are 76209 = 127 times, 76201 = 113 times, 76210 = 84 times, and 76205 = 78 times.

4) The frequency count of the top four actions taken by the police department in response are INITIAL ARRAIGNMENT = 412 times, DRIVING SAFETY DUE = 31 times, DEFERRED DISPOSITION = 15 times, and CASE CLOSED = 15 times

5) The top four geographical locations where the violation SPEEDING IN 30 MILE HOUR ZONE take place and their frequency counts are 76209 = 27 times, 76201 = 19 times, 76210 = 13 times, 76207 = 12 times.

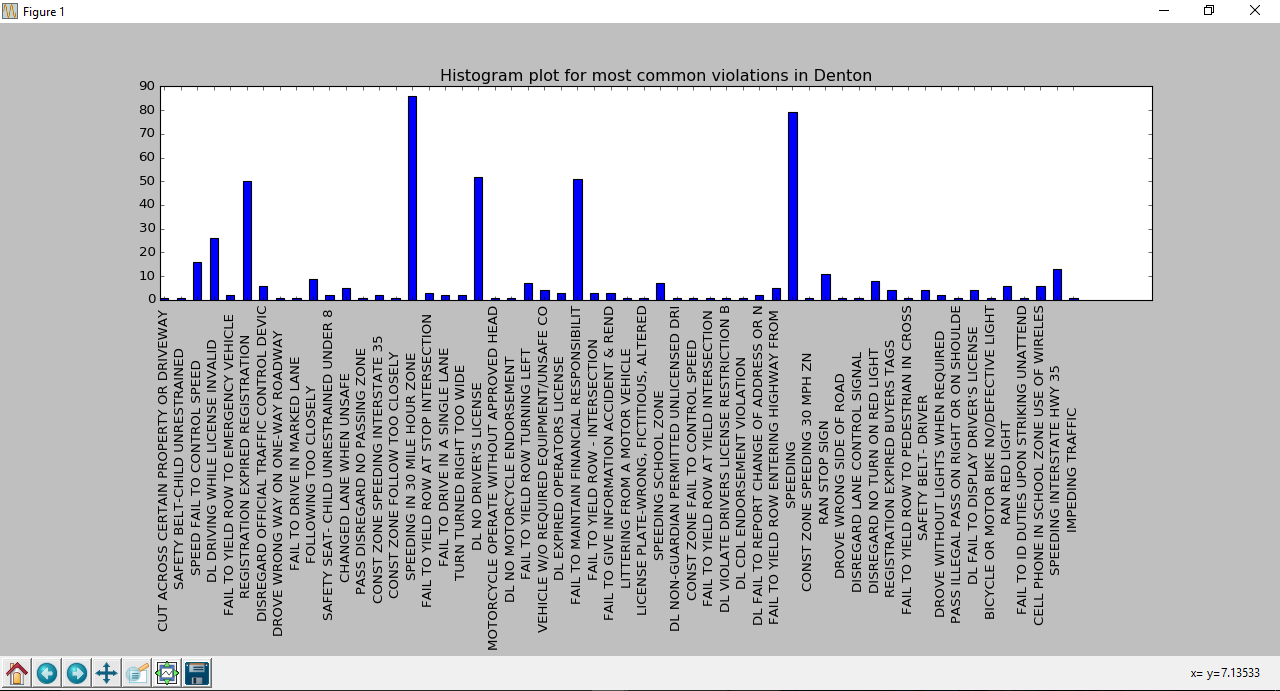
6) The top four actions taken by police department in response to the violation SPEEDING IN 30 MILE HOUR ZONE and their frequency counts are INITIAL ARRAIGNMENT = 72 times, DRIVING SAFETY DUE = 10 times, CASE CLOSED = 2 times, DEFERRED DISPOSITION = 2 times.

The screen shot of the evaluation results is:

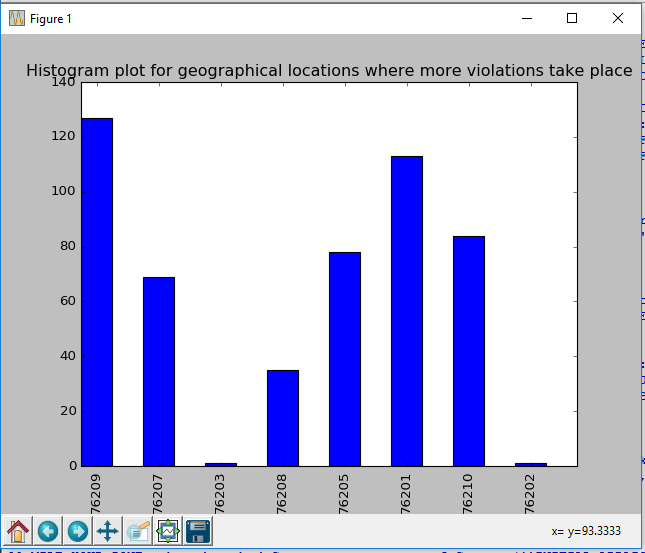


Now, the histogram diagrams are generated for representing the most common violations in Denton, most geographical locations where more violations take place and for most common actions taken by police department in response of violations.

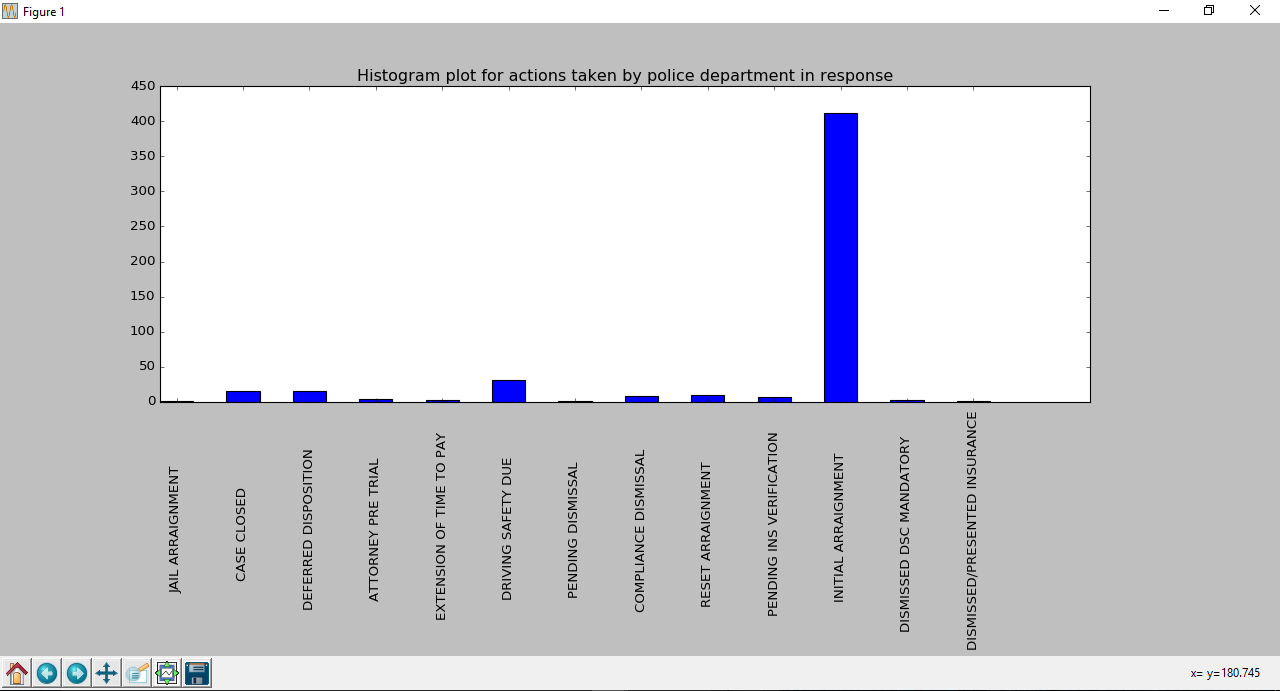
**Histogram diagram for the most common violations in Denton:**

****

**Histogram diagram for the most common geographical locations where more violations take place:**

****

**Histogram diagram for the most common actions taken by police department in response of violations:**

****

**Programming Strategies:**

The programming strategies of this project are:

1. We are making use of the predefined modules in order to easily perform some operations in our program to get the necessary evaluation results.
2. The custom code is written where the counter method is used to calculate the frequency count of common violations, geographical locations and the actions taken by the police department in the response to the violations.
3. Then after, we are using the predefined methods like keys() and values() so that we can easily separate the elements and their frequency counts into two lists. These lists will be helpful for plotting the histogram diagrams.
4. The frequency distribution of the plots is calculated and the plots are mapped for representing final report results.
5. As we want to analyze the three aspects like violations, locations and actions taken by the police department, we are fetching the data from previous created lists of each record.
6. Finally, we are making use of the matplotlib and numpy modules to plot the histogram diagrams for the evaluation results.

**Challenges:**

The following are some of the challenges faced while developing this project.

1. Extracting the related data from CSV files which is required for processing.
2. Calculating the frequency count of all three (violations, locations and actions taken by the police department) fields and then storing them into separate lists.
3. We also faced some problem when we tried to calculate the frequency count of the locations and actions taken in response to the top most violation, hence to avoid this problem we allocated space for newline while storing the zipped data into new CSV file.
4. While generating the histogram reports, our X-axis values are overlapping on each other, hence we used the rotation feature to tilt our values to 90 degrees so that they won't overlap.

**Lessons Learned:**

1. While developing the project we learned to make the proper use of the predefined modules in the python and also learned about few functions which comes with those modules.
2. Performing the Data Analysis on the huge amount of information might become a challenging task because first we need to trim down the huge data to avoid duplications and redundancies and then perform operations on it to get efficient results.
3. Also, we learned to plot the accurate results in the histogram report because it is one of the most important feature where the user can find an easy way to analyze the data by seeing pictures rather than reading the text.
4. We also learned about the different ways (histogram bar charts, pie charts, maps) the python language offers to present out result reports in an effective manner .
5. Last but not the least, we learned to write a small and an effective code instead of generating huge lines of code.

**Future Improvements:**

1. For future improvements of the project, we need to find a way to fetch large amount of real time data without depending on the CSV files.
2. Also, we need to increase our boundaries by conducting an analysis on the traffic violations data for all over the US in order to serve the maximum amount of population with the safety measures.
3. If possible, implementing the maps to plot the results along with the time and date in the output according to the user selected location.