**Comprehensive data analysis of traffic violation data for the City of Denton in 2016**

*INFO 5731- Computational methods for Information Systems   
Spring 2017  
Dr. Jiangping Chen*

**Report by**

**Divya Yerramsetty(DivyaYerramsetty@my.unt.edu)**

**Amruta Mungikar (AmrutaMungikar@my.unt.edu)**

**(1) Purpose of the Project:**

Traffic violation is defined as the violations which occur when a traffic law is violated by a moving vehicle. The different kinds of traffic violations include drunk driving, speeding, running a stop signal or red light, no driving license, wrong way driving, unregistered motor, failed to yield, etc. The purpose of this report is to analyse the traffic violations dataset and draw out conclusions from the data. The analysis is done to understand the number and type of traffic violations that occur in Denton, geographical locations where the violations occur, and what were the actions taken by the Police Department. The purpose of analysing data involves finding the no of violations on each day which would be helpful for department to stricter the rules, number of violations based on time where it is used to estimate trends of violations during different timings of day, Finding out the geographical area where the violations are more, and knowing which status of violation is the most and which status is the least and what actions have been taken by police department. This analysis involves manual Data clean up, key phrase identification, statistical analysis, model building. The results will provide a clear idea about the traffic violations in Denton.

**(2) Data and Data collection:**

The dataset is a CSV file which consists of 1282 traffic violations tickets. These violations were recorded in Denton from the date 11/23/2016 to 12/31/2016.

For each ticket, it provides:

1. Violation Docket Number - Integer type
2. City Citation Number - Integer type
3. Violation date and time - String
4. Name of the City to which the vehicle belongs - String
5. Name of the State - String
6. Zip code of the area - Integer
7. Violation Status - String
8. Status Description - String
9. Violation Status date and time - String
10. Code Description - String

The Analysis is focussed only on the people from Denton, Texas, taking out the out-of-state people and analysing instate as a separate dataset and also considering the violations that have taken place only in the month of December i.e., from 12/01/2016 to 12/31/2016 . There are 11 different status decriptions of violations which gives us the information about the formal reading which are based on 55 various kinds of traffic violations.

The data collection involves finding out the number of records based on various contexts where we can calculate some statistics which provides information on what kind of violations are taking place in Denton. These analysed records can to used to plot graphs considering various findings.

By considering the zip code geographical representation can be done which will help department to know from which area there are more number of violations.

**(3) Methodology:**

The traffic violation in the month of December 2016 for the city of Denton is analysed from the data. Using python, a new csv file is created where the date and the time of the violations are separated and used for the analysis purpose. The following aspects are to be analysed from the data.

1. Date-wise violations: The number of violations based on each date starting from 12/01/2016 to 12/31/2016. This aspect is analysed by taking total number of violations on each single day. This gives a measure of how many violations are taking place in the city of Denton each day. Data is grouped based on date format and analysed. The data in the column of ct\_viol\_date is separated with date and time. The number of tickets corresponding to each date are grouped for all the dates and the number of violations for each day can be measured.
2. Day-wise violations: The total number of violations based on seven days of the week. From this number, we can know about which day in the week usually have more number of traffic violations. Based on the dates in the month, every date falls on either of the days of the week and the data based on seven days can be analysed. Calendar and datetime functions can be used to find out which date falls on what day and hence the violations for 7 days can be analysed.
3. Time-wise violations: The number of violations based on the time of 24 hours. The violations in the entire data is split into 4 main time periods, say morning hours from 06:00 hours to 12:59 hours, mid-day hours from 13:00 hours to 19:59 hours, night hours from 19:00 hours to 23:59 hours and early hours from 00:00 hours to 05:59 hours. By monitoring the data based on time we can estimate the trend of violations based on different time in the day. The range of 4 different time periods is given and the number of tickets corresponding to each time zone are grouped and analysed.
4. Place-wise violations: The number of violations in Denton from different locations based on pin codes. This gives us the information that people of which part of the city has made most number of violations across the month. Different pin codes and their corresponding tickets are grouped and the number of violations from a particular place in Denton can be analysed.
5. Status-wise violations: In the given data, there are a total number of 11 different kinds of status descriptions for the traffic violations. The number of violations for each status description is analysed to know about what status is given more number of times when traffic rules are violated. First each kind is separated from the main data and listed in different lists. Then by calculating the count in each list, the individual number of violations for each status can be measured.
6. Speed violations: There are 6 different kinds of speeding violations mentioned in the data. The number of each kind of speeding violation would be monitored and the highest number of violations for each kind can be found out. The number of tickets for each of the 6 speeding violations are grouped and the highest number among 6 can be analysed.

**(4) Function description:**

We used Anaconda 4.3.1 (64 -bit) for this project as it comes with many built-in libraries.

The libraries/functions/classes to be used for analysis of this project are:

1. Pandas: Pandas is used to operate table-like structures effectively. It is helpful to analyse huge amount of data. Also, date and timestamp can be separated using this module.

A function which splits the violation date and time is to be created to analyse about the violations as per date and time individually. Pandas is used to select desired columns from csv and sort results based on requirement.

1. Matplotlib: This is a plotting library. To have a visual representation the number of violations date wise, day wise, time wise can be plotted. This module is used to creating plots based on the obtained data from analysis. The number of violations is taken on Y-axis and the dates are taken on X-axis and plotted which gives the graphical representation of trend of number of violations on each day of the month. Similarly, graphical representation is made for day wise, time wise and speeding violations.
2. Numpy: NumPy is a general-purpose array-processing package designed to efficiently manipulate large multi-dimensional arrays of arbitrary records without sacrificing too much speed for small multi-dimensional arrays. It is used to create the arrays for plotting graphs by using Matplotlib.
3. datetime: In python, the datetime module gives classes for manipulating dates and times in simple and complex ways. While date and time arithmetic is supported, the focus is on efficient attribute extraction for output formatting and manipulation. In the python program, we use datetime.datetime which is a combination of date and time and then the date and time are separated and written to a new csv.
4. Basemap: The matplotlib basemap toolkit is a library for plotting 2D data on maps in Python.Basemap is specially used by earth scientists, particular oceanographers and meteorologists for plotting different kinds of data on the geographical map. Basemap was used to show the different zipcodes of Denton city which have different number of violations.

**(5) Evaluation Results:**

Evaluation is needed to check if the system is showing the accurate results. Evaluation is done for all the aspects of the analysis.

1. Action: Used datetime package to separate a column in the csv which had date and time in a single column.

Evaluation: This package helped to create a new csv with two separate columns for date and time. To examine this csv, it was opened in Excel.

1. Action: The total number of violations are calculated only for the vehicles coming from the city of Denton and in the month of Decemeber.

Evaluation: After the csv have been modified in the dataframes with the required month and city, the total number of violations were checked by the excel row count and it is same as that of the violation count. So from the given CSV file, considering only the city of Denton and in the month of December there are 478 violations in total. Both excel and python output matched this number.

1. Action: Used Pandas package to create dataframes.

Evaluation: Checked each new created dataframe by printing in the output console window and proceeded to the next step with the modified dataframe. (Every modified dataframe is not printed in the output as it is not necessary, but checked while writing the program.)

1. Action: Used basemap package to plot the geographical representation of the Zipcodes.

Evaluation: Examined the output by looking at the areas in the googlemaps. They were found to be accurate.

1. Action: The dates of the month and corresponding number of violations were displayed.

Evaluation: By searching in excel it was seen that only 29 days in December have violations in Denton as per given csv. The same output was produced by python.

1. Action: The days were listed based on dates.

Evaluation: Checked in the calender whether the dt.weekday\_name performed well in the dataframe and all the days were listed accurately with the corresponding dates.

1. Action: The number of violations based on pincodes, status wise and speeding violations are recorded.

Evaluation: Each count is verified from the csv files and matched with the count displayed in the python result.

**(6) Programming Strategies:**

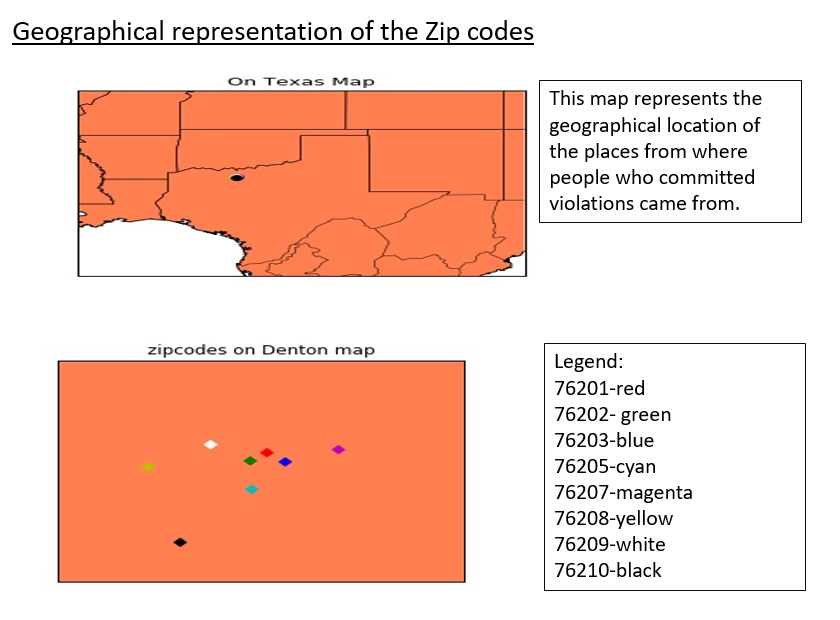
1. In pandas, dataframe is a basic table like structure with the required data. The entire data from the csv is read into pandas dataframe and modified as required.
2. The column header ‘ct\_viol\_date’ is modified into separate ‘Date’ and ‘Time’ columns for ease and has been used to sort data for the month of December. With the given original format of ct\_vio\_date, sorting date and time was not supported. So, the date format and time format has been changed.
3. The dataframe with the modified date and time was taken and further more sorted according to the city of ‘DENTON’ and dates corresponding only to the month of December.
4. List of each kind of violation as mentioned in the methodology was created and the violation data was displayed.
5. Plots were made for each type of violation mentioned in the methodology using matplotlib.
6. Used basemap package to plot the geographical map which represented the Zipcode areas of Denton.

**(7) Challenges and lessons learnt:**

1. The main lesson learnt by doing this project is the usage of pandas package. This package uses dataframe which is a 2D data structure with the required columns. Creating new columns, renaming the column headers, sorting the data based on specific strings or values with conditional logic from dataframes were learnt.
2. The major challenge encountered while performing this analysis is with the Date time format. Unless the date is filtered we could not proceed further. After series of trials, the date and time format was modified and listed out in separate columns. After this, the csv was sorted based on required city and dates.
3. To find out violations based on hours, the time has been converted into 24 hour format and then violations according to different hours were calculated.
4. Some learnings about matplotlib were incorporated while plotting data. Such as how to give x-axis and y-axis manually. If the X-axis have values as strings and Y-axis has values as integers, first the required data should be listed in arrays and the string values of the X-axis should be listed using ‘my\_ticks’ and the values are plotted.

**(8) Future improvement:**

1. Using more than a month’s data, the violation levels can be predicted.
2. An application could be made with the sorted data such as if a day and time are selected, the density of the traffic should be available on a map of the particular location. Also, folium in the anaconda package which will help to obtain a proper geographical map related to the density of the violations.



**(8) References used:**

1. <https://jeffdelaney.me/blog/useful-snippets-in-pandas/>

2. <https://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/>

3. <https://dev.socrata.com/blog/2014/11/04/data-visualization-with-python.html>

4. <http://stackoverflow.com>