County Database of Traffic Violations

Christian Hill
Group #9
University of Colorado, Boulder
1+7203729470
chhi5098@colorado.edu

Chase Whyte
Group #9
University of Colorado, Boulder
1+4258948995
chwh9238@colorado.edu

Rachel Lewis
Group #9
University of Colorado, Boulder
1+7193391584
rale8469@colorado.edu

Ankhbayar Jansan
Group #9
University of Colorado, Boulder
1+3038345229
anja7469@colorado.edu

Figure 1: *Traffic Illustration* [1]

Categories and Subject Descriptors

1 [**Introduction**]: Dataset details – *Objectives.*

2 [**Problem Statement/Motivation**]: Motivation - *Reason for research in topic; What can be gained from this dataset.*

3 [**Literature Survey**]: Previous work on Topic/Dataset - *Rami Kumar's work;*

4 [**Proposed Work**]: Data preprocessing and data collection

5 [**Data Set**]: Dataset used for project - *Collected from Data.gov the government's official data website.*

6 [**Evaluation Methods**]: Research - *Characterization of data*

7 [**Tools**]: Tools used for evaluation - *Python; Orange; Weka;*

Panda; Batchgeo; Matplotlib; Data cleaning and scrubbing.

8 [**Milestones**]: Objectives - *Preprocessing; Scrubbing: analysis of results;*

9 [**Results**]:Data - *Seasonal correlation; Citation plotting; Histogram of traffic violation;*

10 [**References**]

CS CONCEPTS

▫ **Computer science →Data Mining**; Data Visualization, dataset maintenance, dataset evaluation

▫ **Programming Languages→Python→Tools**; Orange; Matlibplot; Weka; excel;

Additional Keywords

Montgomery, scrub, project, cars, traffic, data set, dataset, correlation, Orange, Python, Matplotlib.

1. INTRODUCTION

For this data mining assignment our team has decided to analyze a dataset of collected traffic violations in the county of Montgomery, Maryland. From this data set we want to determine correlations between the people who drive in Montgomery and the traffic violations they receive as well as the locations where most crashes occur. This allows us to see what environmental factors cause people to get tickets. To achieve this goal, we will utilize different data mining techniques, through preprocessing and analysis, discovering correlations between different attributes, and finally a publishing of our findings.

2. PROBLEM STATEMENT/MOTIVATION

Driving is a task that nearly everyone will have to perform at some point in their lives, and for many it is a daily task. However, it is also one of the largest sources of potential danger most people in the modern world encounter. Tens of millions of people are injured or disabled in car crashes every year. Cars are only going to become more and more utilized as the world becomes more technologically advanced. So, it is important now more than ever to obtain as much data as possible about traffic violations and traffic safety statistics. By mining this data set, the roads that have the most dangerous conditions can be determined, and these conditions can be analyzed to improve the general safety of public. Additionally, traffic violations in general can be mined to obtain correlations between the various data attributes and to determine how they affect each other. These correlations can be used to make observations and inferences concerning traffic data in Montgomery, and then generalized to the general population. More specifically, characteristics about the driver, their car, the day/time, and the location of the traffic violation can provide information on how these factors influence the likelihood of being stopped and ticketed.

3. LITERATURE SURVEY

The Director of Data Science at Microsoft, Srini Kumar, published an article about work done with the dataset we will be using[4]. SQL and R were used to show patterns in violations based on the time of day on a map of the county. The type of vehicles committing violations was visualized as shown below, but the data was not normalized so the visualization just showed that the most popular vehicles are responsible for most of the violations.

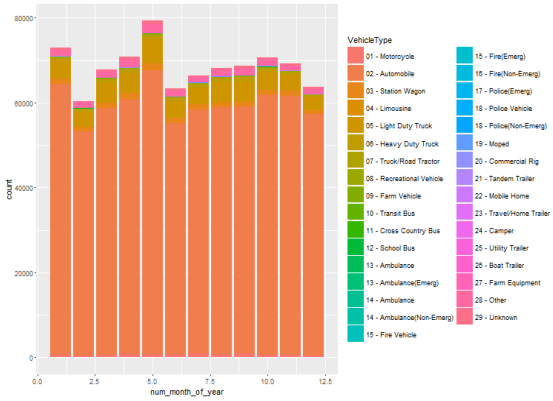

Figure 2: *Bar graph of vehicle types involved in traffic violations*

Data from this area is also being used to look at how insurance is related to traffic violations[3]. This article looks at how characteristics such as age, gender, and whether or not the car is a status brand correlate to traffic violations. This data is then compared to the insurance policy information to conclude whether or not insurance data provides accurate information about traffic violations.

4. PROPOSED WORK

While this dataset is already in excellent condition, some basic maintenance will need to be performed on the dataset. The data will be scrubbed to get rid of superfluous information. Normalizations will be performed on time of the year and demographics, as well as mining for different association rules and correlations. The team will then decide which patterns answer the questions asked of the dataset. Those will be used to create visualizations, such as bar plots

and graphs to show the trends found among the traffic data. More information on work to be completed is outlined in the Milestones (8) section.

5. DATA SET

The data set is found on the United States government's official website for datasets[2].

The dataset contains information from all traffic violations issued in the Montgomery County of Maryland. The data has been collected from 2012 to 2016, and contains identifiers like the time of stop, location, accident, injury (if any), alcohol and/or drug (if any), make/model of car, violation type, race, gender, and charge. The data set contains approximately one million objects, each with thirty-five attributes, and is being updated every fifteen minutes.

6. EVALUATION METHODS

People who commit traffic violations in the Montgomery County of Maryland, as well as the cars they drive can be characterized based on various attributes. Traffic data can be analyzed based on the driver's gender and ethnicity, and then compared to the corresponding distribution across the entire county. The fact that men drive more on average than women can also be taken into account to determine which gender is associated with a higher driving risk. Drivers can also be analyzed based on the color, make, and model of the car. Trends relating to different times of the year that the traffic violations and crashes occur will provide insight into how different weather conditions affect traffic data.

Mining trends concerning traffic data throughout the day will also provide the opportunity to analyze when most traffic violations and crashes happen. The roads that are most dangerous to travel on as well as the roads that drivers are most likely to be ticketed on can be mined as well. The roads that are considered most dangerous can then be further analyzed in order to determine what environmental and circumstantial factors might have influenced these results. This will in turn provide information regarding which road systems and infrastructures are associated with a lower level of safety. These conditions can then be avoided in the future to maximize road safety and decrease the number of injuries and fatalities.

Orange will be used to mine the reason for the traffic stop by searching for certain keywords. For example, the number of crashes that involve drugs and/or alcohol will be determined as well as the number of violations involving speeding tickets. Once the distribution of different traffic violations is determined, the likelihood that a particular violation will result in a ticket can be mined. For a particular violation, a correlation between gender and likelihood of being ticketed as well as ethnicity and likelihood of being ticketed can be mined. From all of these we can utilize this information to predict where crashes and traffic violations are most likely to occur and what environmental and situational factors influence this using Bayes' Theorem.

7. TOOLS

The primary language for programming in this project will be Python. In Python we will be using various tools, one of which being named Orange and is primarily used for text mining. We aim to use Orange to mine the various street names and the cars' type/color. Another Python tool that will be used is called Matplotlib, which will be used to create various charts and graphs to represent the data that has been mined. A majority of the dataset is usable and can be scrubbed using only excel. There are a couple of attributes that need to be removed such as the latitude and longitude, article, and the licence/drive state. Another python tool called pandas will be very helpful for reading in and working with large data sets. A latitude and longitude visualization application called 'batchgeo' will be heavily used as well.

8. MILESTONES

First the data will be scrubbed for any unfilled values, as well as certain unnecessary attributes. At this time these unnecessary attributes include the latitude and longitude, arrest type, driver's state, violation, and geo location. Next, a Python program will be written to process the data and begin visualizing the data. At this point the group will want to find correlations that will give us some information about our data and the questions that we want to answer. Bayes' theorem will be used to see what types of environmental factors affect traffic violations and compute the probability of traffic violations occurring at certain locations. Then we will find the correlation between certain days/years and how often violations are assigned. We will also find the most common demographics that get violations, warnings, and are involved in accidents. At this point we will be working on the progress report and will be updating our found results.

8.1 PROGRESS

We have already preprocessed the dataset by removing the attributes that were unnecessary. These included the agency and subagency issuing the traffic violation, as well as the description of the violation. Instead of the description, we decided to use the numeric code for each specific charge. This allows us to generalize violations, so for example we can group all charges that involved a licensing issue. We also scrubbed the accident attribute since the attribute value was a no for every single violation. However, many violations still included personal injury, property damage, and even fatalities, so we included these in our list of traffic violations involving crashes. The rest of the attributes we scrubbed because we could not mine anything interesting from them. These include whether the violation involved commercial licensing or a commercial vehicle, HAZMAT, and a work zone. We also scrubbed the state issuing vehicle registration, the year, make, and model of the vehicle, the article of state law, the driver's city and state of residence, the state that issued the driver's license, arrest type, and the geolocation of the violation. Afterwards, we decided what we need to do with missing attribute values, in particular with longitude and latitude. Since using a mean value or a global constant would place a large number of violations in an arbitrary location when mapping them out, we decided to just throw out the data that were missing latitudes and longitudes. After the data was preprocessed, we began by mining for a correlation between the time of the year and the chance of getting personal injury. We also

looked at the time of year of all traffic violations occurring in the county to see if there is a similar correlation to the temporal change of the number of violations involving personal injury. The full data set could not be read in due to encoding errors, so we were only able to use test data with around 10,000 points. We then separated these points based on the year that they occurred in and created histograms for the different years represented in the test data set.

8.2 Proposed Work (Updated)

There are several more different patterns we want to mine from this dataset. First of all, we still need to create histograms for each year using the entire dataset based on the time of year that violations occurred in. Also, we can create a histogram by combining data from each of the five years together in order to account for any fluctuations from year to year. We also want to mine whether gender or ethnicity can influence how likely someone is to receive a ticket for the same violation.

We can mine if certain ethnicities are more likely to get pulled over by finding a percentage distribution of traffic violations and comparing this with the demographic distribution of the county. This data can also tell us whether certain ethnicities or genders are more likely to drive a certain car type and what kind of infractions they are most likely to commit. In addition, we will determine whether vehicle type or vehicle color has any influence over how likely someone is to get pulled over. This will be based on the number of times the vehicle type or color is found in the data set, which can then be compared with the probable distribution of that vehicle type or color across the county. Lastly, we will attempt to find correlations between different attributes using Naive Bayes. The main attributes we want to target are ethnicity, location, and type of citation. Luckily, Montgomery is a relatively diverse community making our classifier accurate, without sacrificing accuracy.

9. RESULTS

Seasonal Correlation

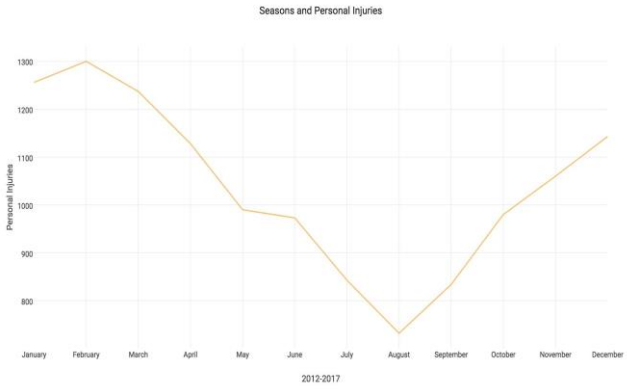

Figure 3: The seasonal effects on traffic violations with personal injuries

There were around 12,476 traffic violations with a personal injury, dating from 2012 to 2017. From the five-year period of data collected in Montgomery County, it shows that there is a clear correlation between season changes and traffic violations. During the winter season, the personal injuries from traffic violations steadily rise until around February. When winter transitions to spring in March, the violations with personal injuries begin to decline. We can begin to see the correlation of the road conditions and traffic violations. The number violations continue to decline until reaching the yearly low, in the summer around July to August. The summer months represent better road conditions as opposed to fall and winter as seen in the decline. However, this is still incomplete without completely comparing against the other attributes. There are many more factors contributing to traffic accidents, however there is a clear trend related to seasons. Extrapolating from this five year data, we could assume that barring any major changes, the correlation between seasons and violations will hold.

Citation Plotting

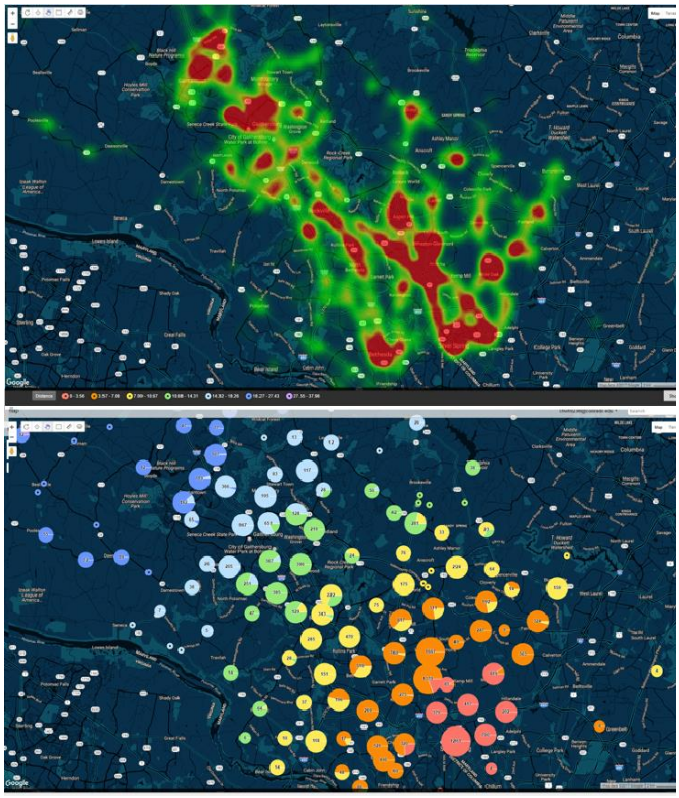

Figure 4: A heat map of citations given throughout the county of Montgomery, Maryland.

For this project our group decided to visualize the latitudes and longitudes of the traffic citations as a way to see what environmental factors may affect the amount, and types of citations given in Montgomery. The visualization of the Latitude and longitude will be done with a program called BatchGeo, with BatchGeo the project will include heat maps, as well as maps that tracks the number of citations given based on sections. First, we are able to see through the heat maps that the counties most affected are; Silver Spring, North Kensington, Germantown, and other surrounding towns[4]. While we haven't yet researched into depth about why these locations have a higher concentration of citations, at this time it's fair to assume that these locations either have: a higher populations than the other cities, there are speed traps located at certain locations in the 'hot' zones, and/or there may be a lot of traffic that passes through the area at certain times/months. This visualization of data will be used when locating density of crashes, traffic violations, and support other results discovered through research.

Histogram of Traffic Violations

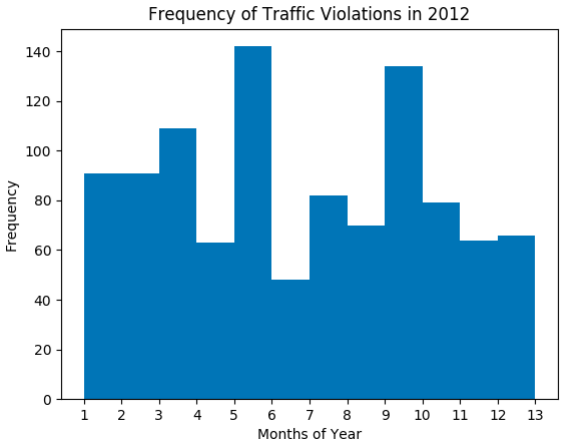

Figure 5: A histogram for the frequency of traffic violations in 2012.

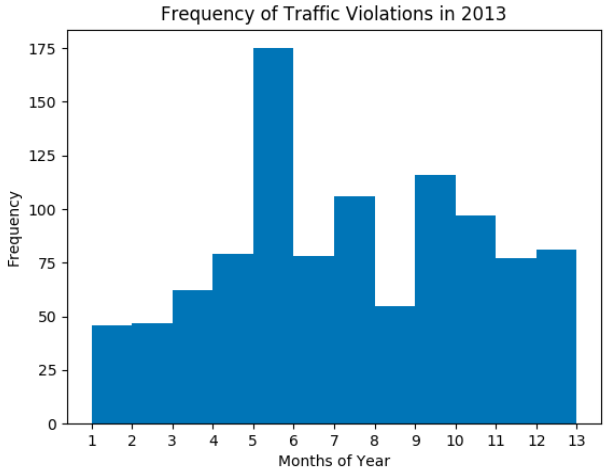

Figure 6: A histogram for the frequency of traffic violations in 2013.

In both of these histograms, the count of traffic violations for the month of May is much higher than in the other months. Additionally, in both years the number of traffic tickets jumped up during the month of September, most likely because school is just starting and more people are out on the road taking their kids to school. However, since only test data was used, it is difficult to determine whether the patterns displayed in these histograms accurately reflect the entire data set. There seems to be a lot of random fluctuations based on year, so more data mining is needed using the entire data set to determine a correlation actually exists for a certain time of year.

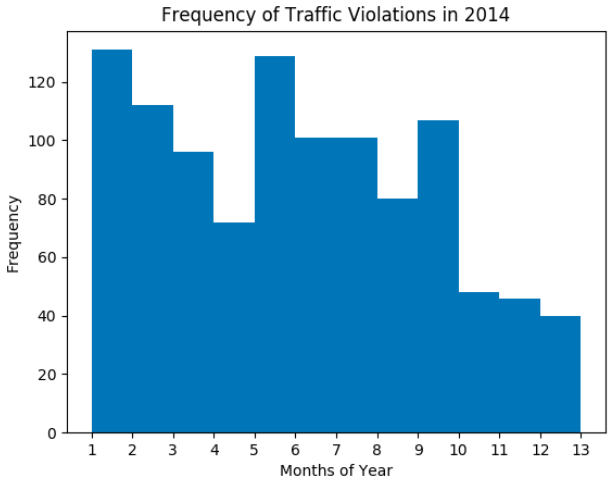

Figure 7: A histogram for the frequency of traffic violations in 2014.

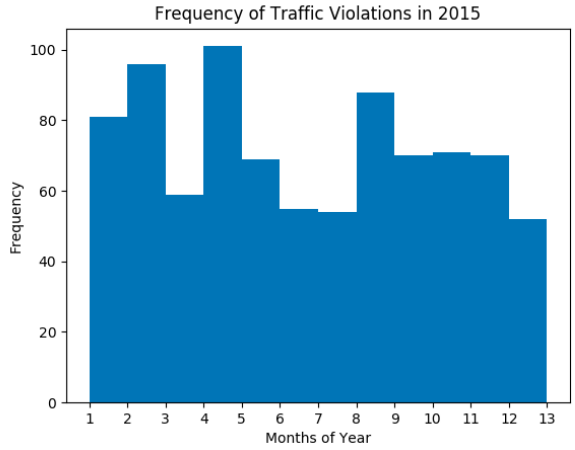

Figure 8: A histogram for the frequency of traffic violations in 2015.

The histograms from 2014 and 2015 also show fairly similar patterns, as both show peaks in violations during the first two months. Once again, May shows signs of a high number of violations on both histograms, but in 2015 violations were more concentrated in the month of April. Although the last few months for 2015 saw quite a few violations, 2014 did not see the same pattern. Again, we see the peak in September with 2014 and a relatively high count as well for September during 2015. The summer was also much busier in 2014 than 2015 for catching traffic violators.

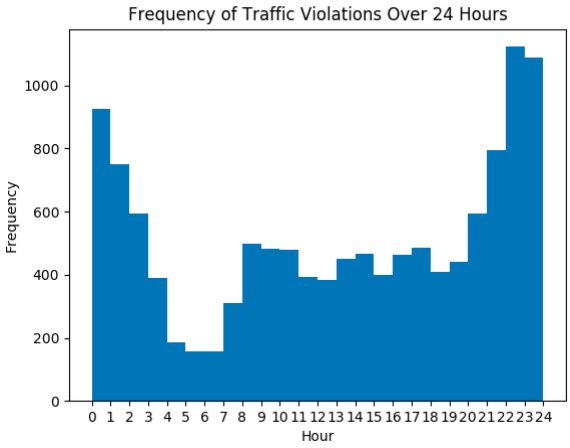

Figure 8: A histogram for the frequency of traffic violations over 24 hours.

This histogram shows the time of day in which each crash occurred over the 5-year period. It is similar to how we expected it to look, but we were expecting a much higher peak during the times that people are commuting to work. Instead, we only saw large peaks from around 10:00 PM to 2:00 AM. This is followed by a very low traffic violation count between 3:00 AM and 7:00 AM. In the morning you are at a slightly higher risk of receiving a ticket than throughout the rest of the afternoon.

10. REFERENCES

[1] Anon. 2013. How Google Tracks Traffic. (July 2013). Retrieved March 3, 2017 from https://www.ncta.com/platform/broadband-internet/how-google-tracks-traffic/

[2] Anon. 2017. Montgomery County of Maryland Traffic Violations. (March 2017). Retrieved March 2, 2017 from https://catalog.data.gov/dataset/traffic-violations-56dda

[3] Sara Arvidsson. 2011. Traffic Violations and Insurance Data. (2011). Retrieved March 2, 2017 from http://www.diva-portal.org/smash/get/diva2:669332/FULLTEXT02.pdf

[4] Srini Kumar. 2016. An Analysis of Traffic Violation Data with SQL Server and R. (April 2016). Retrieved March 2, 2017 from http://blog.revolutionanalytics.com/2016/04/an-analysis-of-traffic-violation-data-with-sql-server-and-r.html