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; Matplotlib; Data cleaning and scrubbing.*

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

9 [**Summary of Peer Review Session**]: Review session - *Questions posed after presentation; Plans to work on data.*

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.

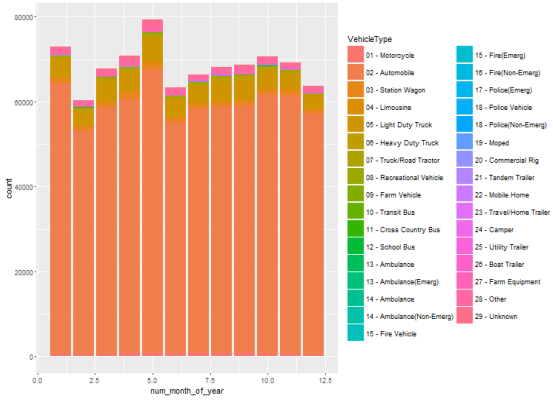

Figure 2: *Bar graph of vehicle types involved in traffic violations*

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 tools 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 license/drive state. We also want to use WEKA to gain some experience with machine learning as well as its data visualization capabilities.

8. MILESTONES

The milestones for this project are as follows:

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.

9. SUMMARY OF PEER REVIEW SESSION

The peer review session provided the team with more details to consider, most of which have been accounted for in this document. One suggestion was to normalize the percentage of violations based on certain demographics, so as to not have misleading results. Separating the information on accidents and violations such as speeding tickets was proposed, since these two categories give information that can lead to different preventative measures. City planners could then see which roads are more dangerous and make changes while police can post themselves in locations where they are most likely to catch people speeding. Finding corresponding weather data was also suggested to provide potential environmental context for road conditions that might affect the amount of accidents that occur.

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