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, BatchGeo.

1. ABSTRACT

This paper covers the analysis of a dataset containing a list of traffic violations occurring in the county of Montgomery, Maryland. From this data, various attributes of traffic stops were analyzed to mine correlations in order to answer several questions about traffic data. Some interesting questions that are answered in this paper include the following: How does location affect the concentration of crashes and traffic violations? How do the time of day and time of year affect the number of traffic violations and crashes? How does the car color affect its likelihood of being pulled over, and is it true that red cars are more likely to be pulled over? How does gender affect the likelihood of receiving a warning or a ticket for a traffic violation? In summation, It was discovered through use of clustering and heatmaps what environmental factors impacted the highest density of crashes and citations. it was found that the frequency of crashes over a 24-hour-period peaks around rush hour and reaches a minimum at 5:00 A.M, while the frequency of all traffic violations peaks at midnight and also reaches a minimum at 5:00 A.M. For the car color, it was found that the percentage of red cars pulled over matched very closely with the average percentage of red cars owned. White and brown cars were less commonly pulled over while green, blue, gold/yellow, and other less common colors were more likely to be stopped. When comparing the percentage of citations received to total number of warnings and tickets combined, men received tickets 54% of the time while women only received tickets 43% of the time.

2. INTRODUCTION

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. In plotting the locations with the highest concentration of crashes, the most dangerous road conditions can be determined, and these conditions can be analyzed to improve the general safety of public. This was accomplished by using a mapping service called BatchGeo, which took location data to create a heat map of the traffic violations as well as accidents. To analyze the temporal trends, we extracted the month and the hour in which each traffic violation occurred and plotted them to look for peaks and valleys in the data. This is important because as drivers, we need to know what times we are more at risk and adjust accordingly. The data on car color can tell us which colors are most likely to be pulled over. Car color is fairly significant because a large majority of people consider it important when buying a car. The color trends mined can hopefully help make that decision a little easier. Lastly, the effect of gender on the likelihood of receiving a ticket can be mined by comparing the percentage of warnings and citations received by males with that of females. This might be important because it could show ways in which our traffic enforcement system is biased.

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.

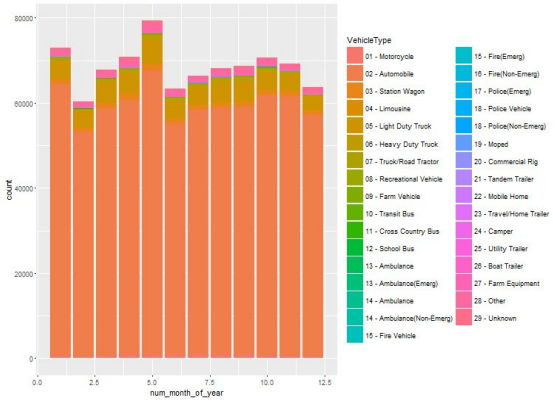

Figure 2: *Bar graph of vehicle types involved in traffic violations*

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.

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 was be scrubbed for any unfilled values, as well as certain unnecessary attributes. At the time these unnecessary attributes included the latitude and longitude, arrest type, driver's state, and violation type. Next, Python programs were written to process the data, and a program named BatchGeo was used to create heat maps and map based word-clustering to visualizing the data set. At this point the group wanted to find correlations that will give us some meaningful information about our data and the questions that we wanted to answer. The heat maps will be used to see what types of environmental factors affect traffic violations, as well as visualize the distribution of both female and male violations in the county, and observe environmental impacts occurring at certain locations. Then we found the correlation between certain days/years and how often violations are assigned. We also found the most common demographics that get violations, and their car colors, we then compared this information with other statistic, At this point we have answered all of our questions and have supported them with graphs, statistics, and maps.

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 also 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

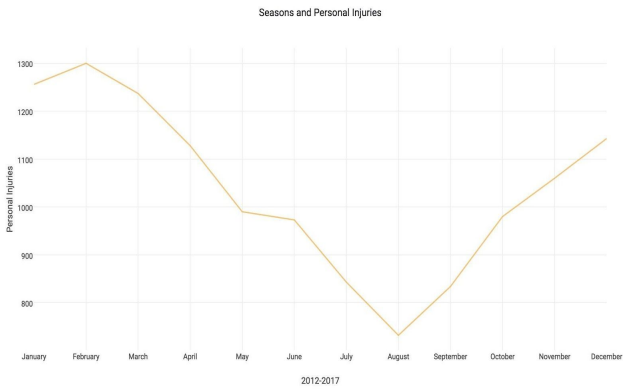

Figure 3: The seasonal effects on traffic violations with personal injuries

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.

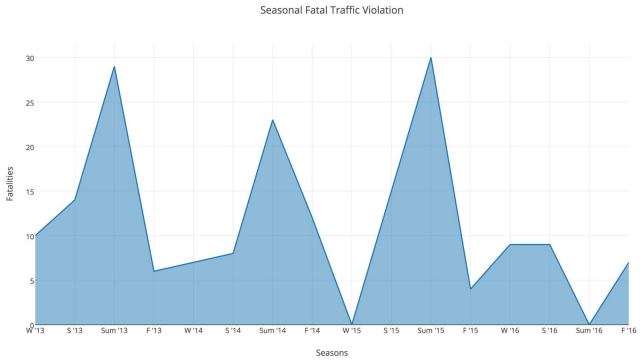

Figure 4: The seasonal effects on traffic violations with fatal injuries.

As illustrated in Figure 3, the traffic violations with personal injuries rise during the colder months and decline with the warmer

months. However, the inverse trend seems to be true for the more serious traffic violations that involve fatal injuries. There were 213 fatal injuries recorded by the county from 2012 to 2016. The data was sampled from the years 2013 to 2016 to locate the trend and draw the correlations. The trend can be seen in Figure 4, the number of fatal injuries increase as the months get warmer. The accidents reach its annual high every summer for three years in a row. Summer in 2016 seems to be an anomaly as it has no violations fatal injuries. The contrast between the two trends could lead to interesting conclusions. In terms of absolute numbers there could be no comparison between the two attributes. The spike in fatal injuries during the Summer months could indicate that people are more likely to drive recklessly during the warmer months. The weather plays a limited role in traffic violations during the warmer parts of the year, as compared to conditions present during the colder parts. Ice and snow are clear factors that play a role in the total number of traffic violations with personal injuries. The trends from Figure 3 seems to support the conclusion, but it's interesting that the opposite holds true for violations with fatal injuries. In 2016, the colder months register more fatal injuries than the warmer months. From the data, there is a higher overall chance of fatal injury occurring during summer, so 2016 must be an outlier in the trend. The correlations of seasonal effects are essentially the inverse between personal injuries and fatal injuries. Fatal injuries may be occurring more during the warmer months due to people driving increasingly reckless and careless with limited weather impact.

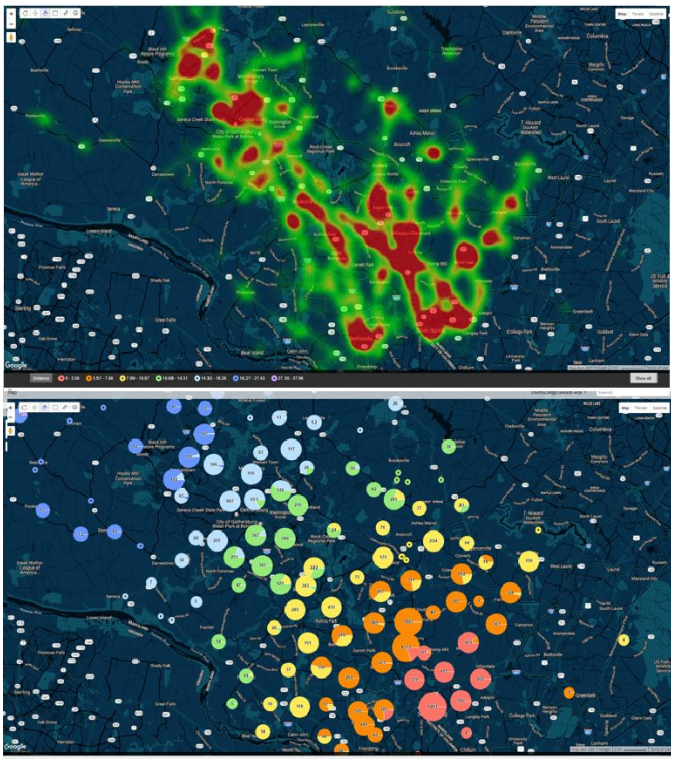

Figure 5: A heat map of citations given throughout the county of Montgomery, Maryland.

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.

Citation Plotting

Crash Plotting

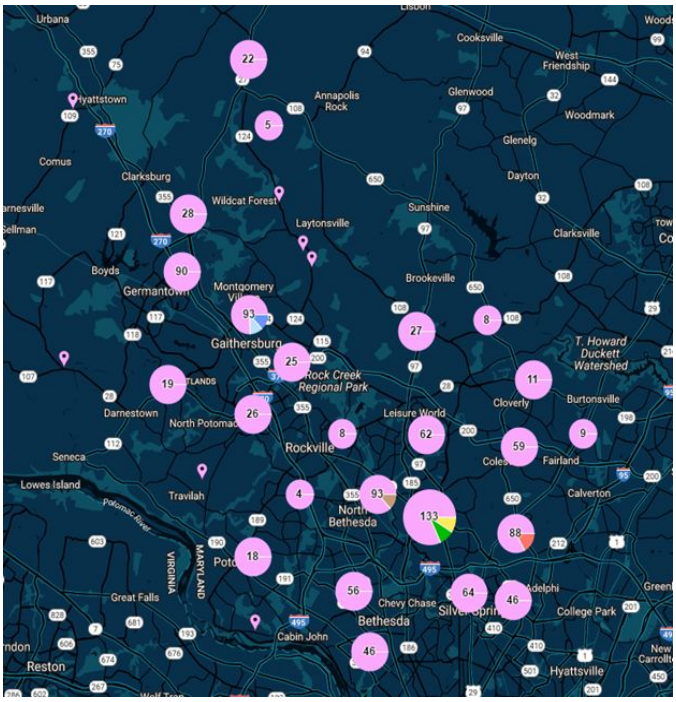

Figure 7: A clustering map of crashes throughout the county of Montgomery, Maryland.

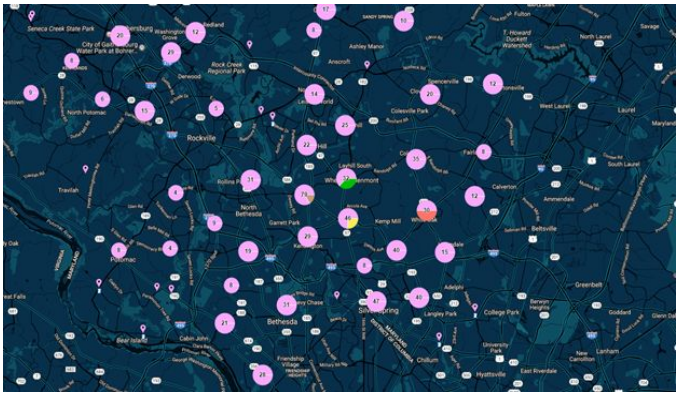

Figure 8: A closeup of the highest density of crashes in the county of Montgomery, Maryland, this is a continuation of figure 7.

To visualize the crash data we first used text clustering to discover what general locations had the highest density of crashes over the four years, then mined specific attributes to find crash locations. This information was very helpful in visualizing outliers as well, and showed that those crashes were mostly occurring on highways outside of major population centers. The map also shows where the most crashes happen in the county, which allows us to find certain environmental impacts that causes these locations to be more dense. After researching, the team discovered that the highest concentration of crashes in the data set (yellow) was surrounded by 5 driving schools that were either nearby or directly adjacent to the crash, The green is located near a pub and three parking garages, the pub appears to be very popular and a local favorite, with having numerous, very complementary yelp ratings and high Facebook traffic. The blue and light blue cluster are very highly populated areas in the county like a city center, mall, apartment buildings and two parking garages.

Gender Distributions of Citations And Violations:

Female Violations Ma

Figure 9: A clustering map based on the female traffic violations in Montgomery, Maryland.

Male Violations Map

Figure 10: A clustering map based on the female traffic violations in Montgomery, Maryland.

When frst comparing the two maps you'll notice that the distribution between the two sexes is actually very different, The male citation distribution is much more spread out than the female map, a reason for this distribution is because jobs located in the middle left and top left are manual labor jobs that have been historically male, there's also a notable jump in the number of citations/violations for males near college campuses and bars, this could possibly allude to an observation that males are more likely to drive while intoxicated in this county. The female clustering shows a higher density towards the center of the map, these locations trend to be parks and office buildings, there is also a slightly denser concentration of violations for females near by high schools, city centers, and malls.

These maps gives us interesting information about the traffic violators in the Montgomery county, While it would not be appropriate for the police to specifically target men/women in this area, it gives a general idea of where violations are given out the most in the area, and specific centers that the police should specifically monitor.

Histograms of Traffic Violations

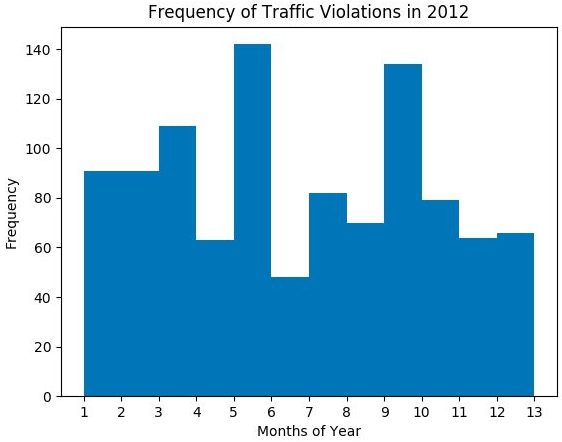

Figure 11: Frequency of traffic violations in 2012.

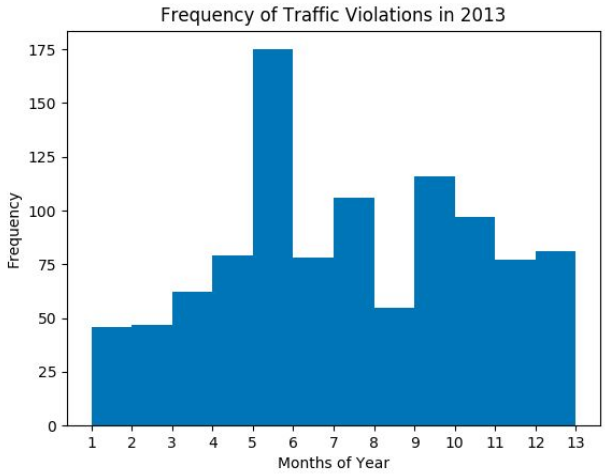

Figure 12: Frequency of traffic violations in 2013.

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.

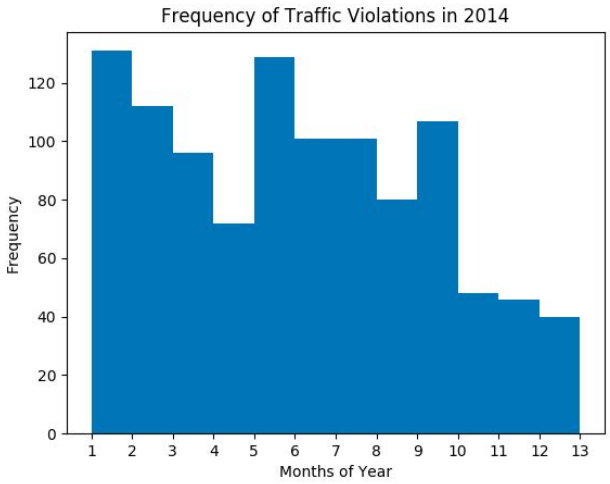

Figure 13: Frequency of traffic violations in 2014.

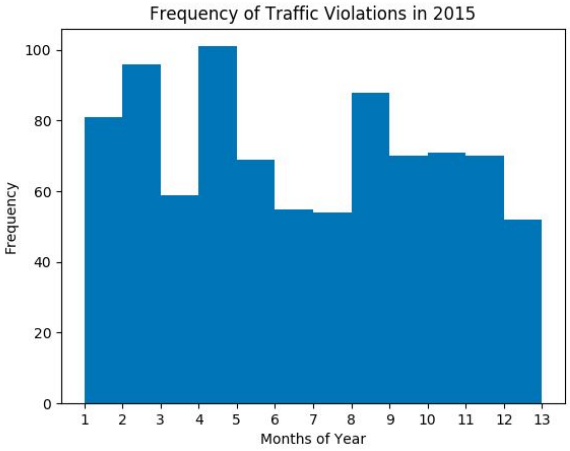

Figure 14: Frequency of traffic violations in 2015.

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.

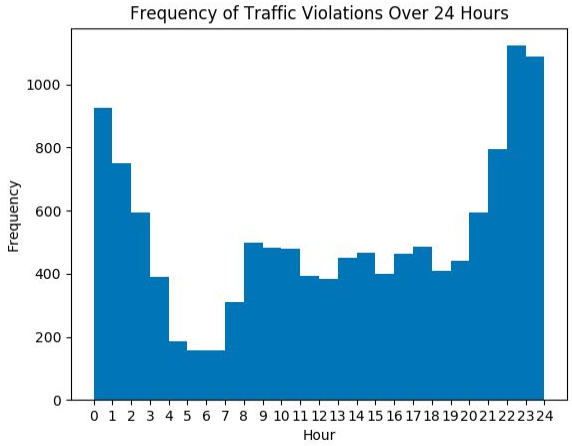

Figure 15: Frequency of traffic violations over 24 hours.

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.

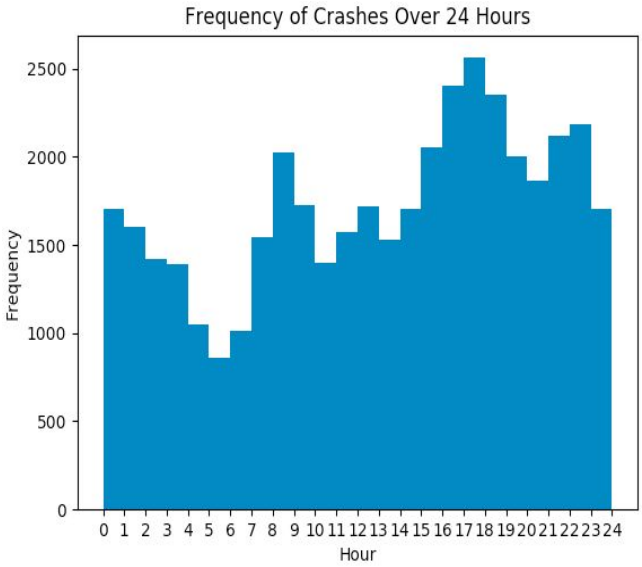

Figure 16: Frequency of crashes over 24 hours.

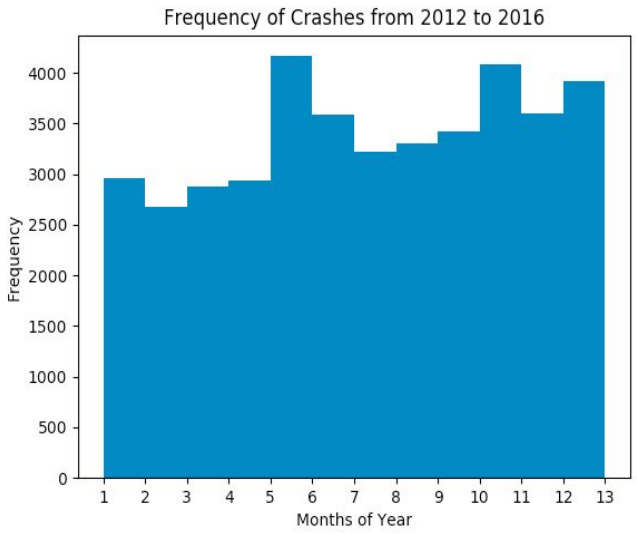

Figure 17: Frequency of crashes from years 2012 to 2016 in Montgomery County.

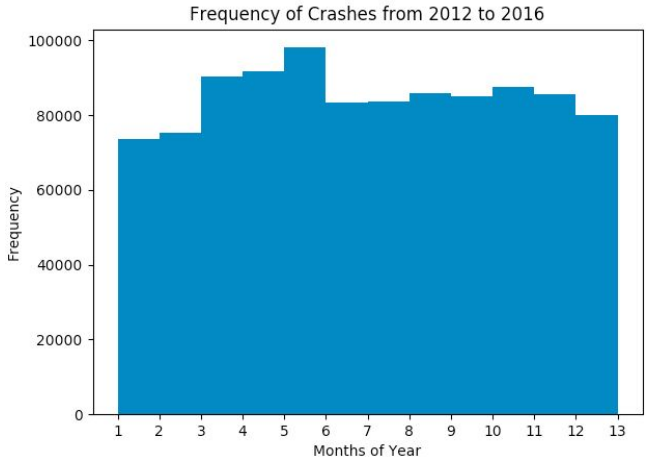

Figure 18: Frequency of traffic violations from years 2012 to 2016 in Montgomery County.

Both histograms show a peak

Pie chart for Plotting

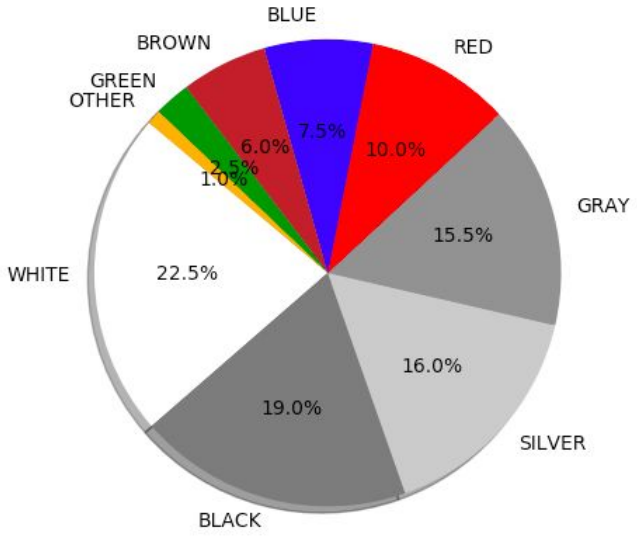

Figure 19: A pie chart of the average car color ownership based on two studies from American paint manufacturers PPG Industries and DuPont.

This pie chart is used to normalize our data. Since there is an unequal ratio of cars of each color being owned, the ratio of cars of each color being pulled over will not be equal.

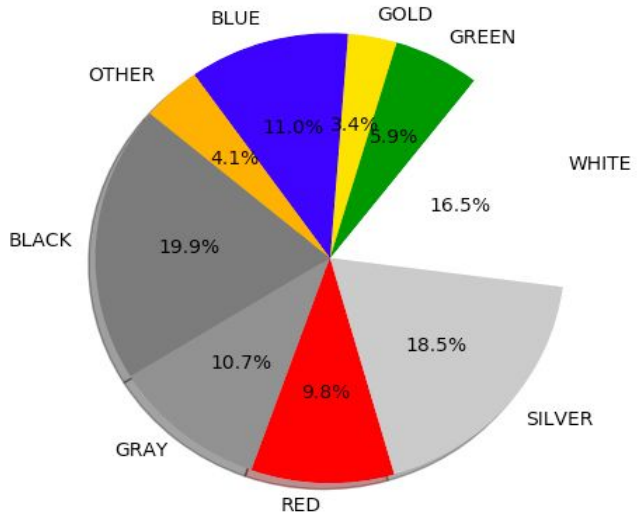

Figure 20: A pie chart of the colors of cars stopped in the county.

The pie charts show that there is minimal bias based on car color. By including both the distribution of cars pulled over for traffic violations and the average car ownership, bias is exposed. Most of the percentages vary slightly, meaning there is not much bias. Blue cars are the most significant change in being pulled over

more often, and white cars are the most significant change in being pulled over less often.

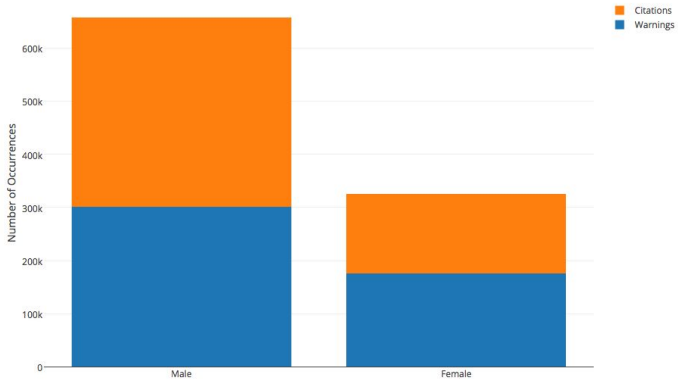

Figure 21: Stacked bar chart of the number of warnings and citations for males and females.

10. Applications

The data that has been collected and analyzed can be used in many ways.

One such way is to explore what traffic policies may be more effective than others. There was a dip in the average number of traffic violations in the year 2015. The next step in using this data would be analyzing the changes in policies or any other factors that could have affected the change.

The location with the highest traffic violation concentration was a major point of interest. Using this, police can know which areas are more likely to have violations and be in that area more often. City planners can also use this data to see what environmental factors may contribute to the number of violations. For example, more personal injury violations may occur on a certain bend of road where visibility is minimal. With the data layered on the map it is easy to see what may cause these. The city planners can then work with civil engineers to correct these hazards.

It was also found that personal injury accidents occur more in the winter than in the summer. One application of this finding is the police force can have more accident investigators available during the winter so they can be prepared for the higher numbers. Since inclement weather occurs more in winter, using the locations with the highest traffic violations with personal injury can show engineers what stretches of road are more dangerous in slick weather. This can be used to make the streets safer to drive on and inform future planners of what types of road planning is dangerous in winter weather in Maryland.

Mining this data with regards to gender gives insurance companies a great deal of information. Insurance companies give lower rates to drivers who are thought to be less likely to drive recklessly. Our data showed that many more warnings and citations were given to men than they are to women. This supports what many insurance companies have already established and is the reason women have lower insurance costs right away. This, of course, changes if the driver gets in an accident.

This data can also shine a light on bias. We explored the bias associated with car colors in our data. The findings showed that the distribution for car colors getting traffic violations is consistent with the distribution of car colors owned in the area. Blue cars are slightly more likely to be pulled over, but in general the percentages were pretty consistent. This shows there is not much bias regarding car color.

11. 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