Analysis on Traffic Violation Data of Montgomery, Maryland

Kim Minsu, Kim Won Hyun, Liu Siyuan, Zhang Linghan, Tan Chin Won

Abstract

Violations in traffic laws are very common in the United States, the country with the world's most registered vehicles. Vehicle volumes and the number of traffic violations have been increasing drastically. In recent years, the government has been collecting data and publishing on traffic violations. This paper aims to discover characteristics and trends in traffic violations by analyzing government open data. The team adapts the traffic violations data in Montgomery County of Maryland in the United States published by the local government. The dataset contains 682353 observations from January 2012 to June 2015 within Montgomery County. The analysis focus on core features such as geolocation, gender and time of stop. The findings provide valuable insights on how the traffic violations distribute, different characteristics between violations by males and females and the number trend traffic violations over time.

Analysis on Traffic Violation Data of Montgomery, Maryland

The United States have the most number of cars in the world. As of the year 2012, there were some 254 million vehicles registered in the United States. The number of vehicles registered in the United States are steadily increasing over the last two decades[1]. With such a large number of cars, more traffic violations are expected to happen. This report explores some characteristics of traffic violation using the traffic violation data from Montgomery, Maryland. The questions to be answered are how the traffic violations distribute, differences on characteristics of traffic violations between males and females and the trend of the number of traffic violations over time.

Methods

Dataset

About 680,000 records of traffic violation in Montgomery, Maryland from January 2012 to June 2015 were analyzed during the analysis. Various information is provided with each record such as date, time, geolocation, gender, description, alcohol and fatality.

Apparatus

A laptop computer was used to download the dataset using its default web browser, Safari.

Procedure

The R programming language with its standard library and external libraries such as ggplot2 and ggmap was used for data analysis. The data were cleaned first by omitting incomplete entries. Several kinds of plots were used during analysis. Geolocation data were plotted on the map of Montgomery, Maryland to show their spatial distribution characteristics. Numerical data such as frequencies were plotted using bar plots to show the trends and characteristics. Besides plotting, analytical methods for both qualitative and quantitative data were applied for the analysis.

Qualitative method used was typology. And quantitative methods include linear regression, Student's t-test and k-means clustering.

Typology. The data were classified by date and gender for the analysis of trends over time and gender differences.

Linear regression. Linear regression was used to analyze the trends on the number of traffic violations over the years. Whole years of data are available for the years 2012, 2013 and 2014. And data in year 2015 are only available for the first six months of the year. The data of year 2012, 2013 and 2014 show strong linear relation. Therefore, a linear model was used to estimate the total number of traffic violations in 2015.

Student's t-test. From the plots of monthly number of traffic violations in year 2012, 2013 and 2014, they show similar distributions for these three years. To further test this hypothesis, Student's t-test was used and a significance level of 0.05 was adopted for the hypothesis testing.

K-means clustering. The plot of locations of traffic violations on the map shows some characteristics about the location distribution. To show the characteristics more in a clearer manner, k-means clustering was used to cluster geolocation data and find centers of clusters. The centers were then plotted on the map to see if there is any apparent characteristics about location distribution of data.

Results

Geolocation distribution

By plotting all geolocation data on the map of Montgomery, Maryland, a rough distribution pattern was shown. The pattern is that most of the traffic violations happened at the crossroads. To further prove such pattern, more than 20 centers of geolocations were selected by k-means

clustering algorithm and plotted on the same map area. From the plot of these centers, the above pattern became much more obvious. Figure 1 and figure 2 show the location distribution of all data and the centers.

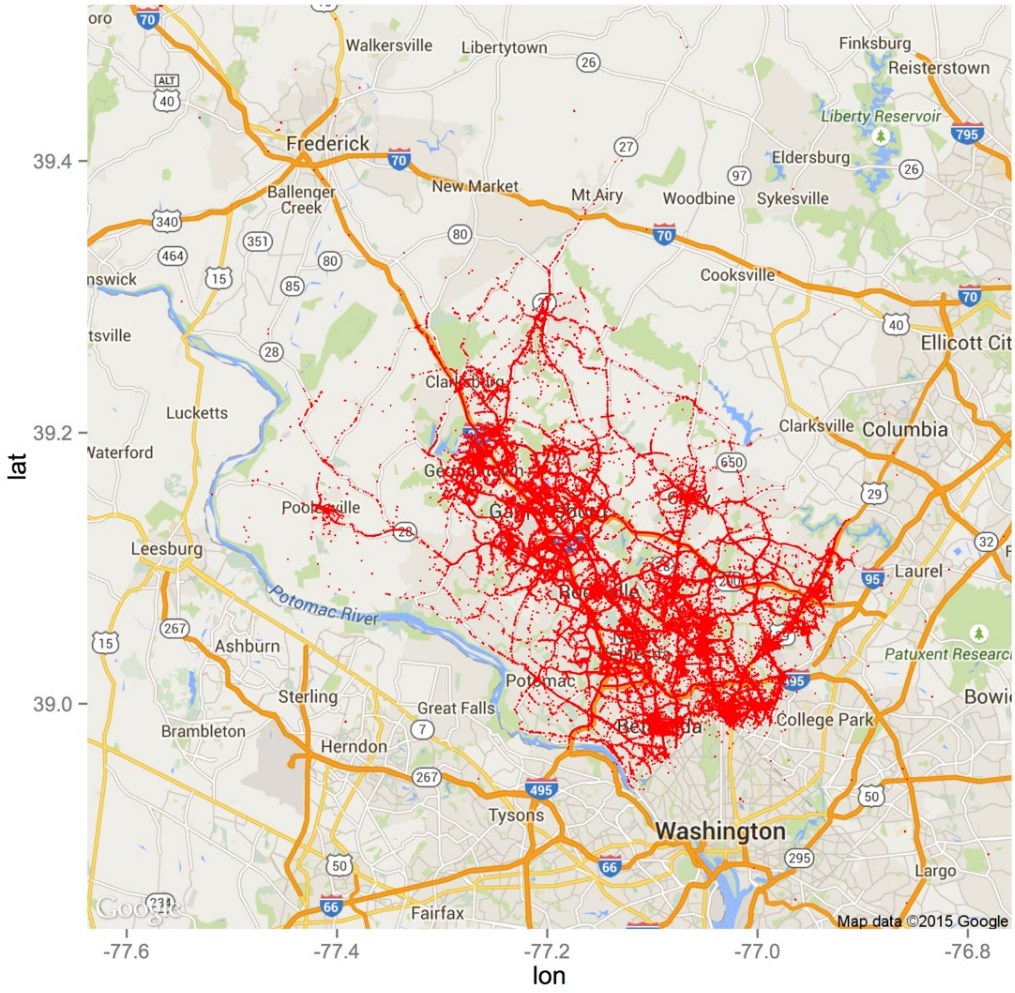

Figure 1. Plot of all traffic violation locations on the map of Montgomery, Maryland

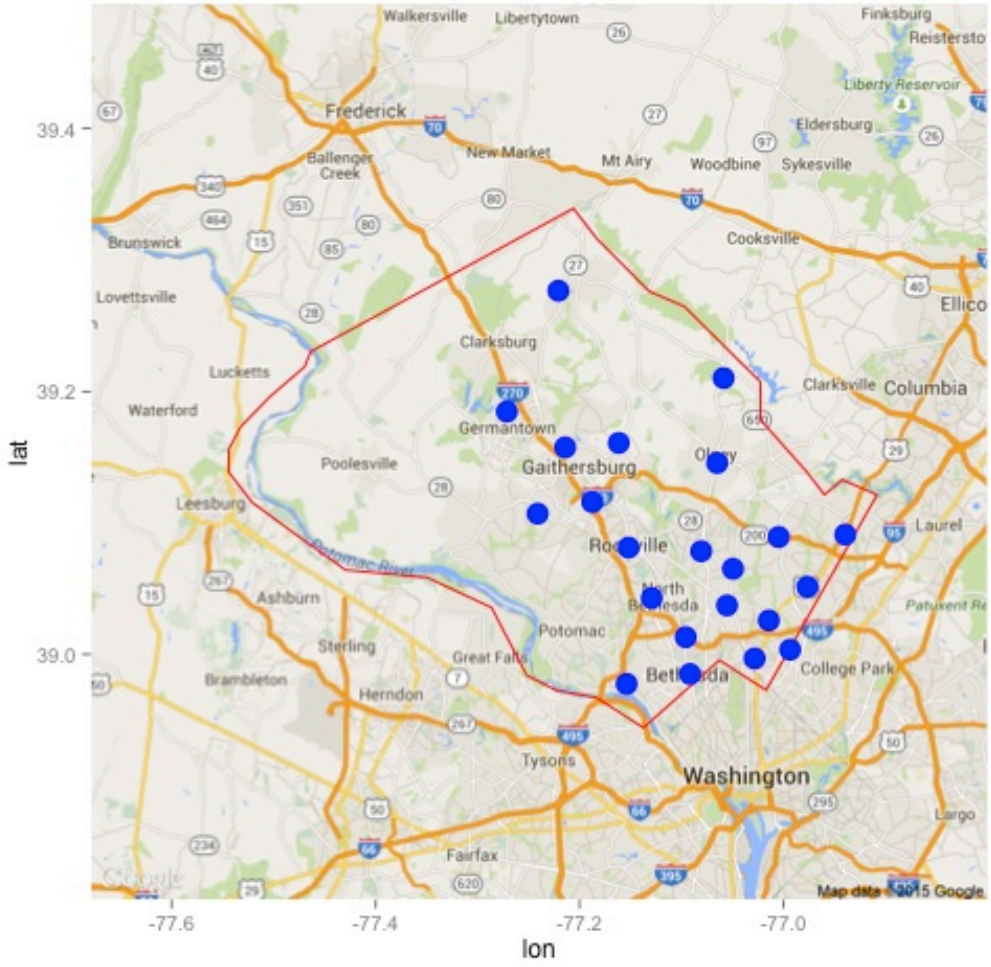

Figure 2. Plot of the traffic violation centers calculated by k-means clustering algorithm on the map of Montgomery, Maryland

Characteristic differences between genders

Comparison on several columns of the data was made against gender. Some of the comparison shows interesting characteristics about the differences between males and females. Figure 3 shows that the places where males were involved in traffic violations about alcohol are more widely distributed while these of females are more grouped in downtown areas.

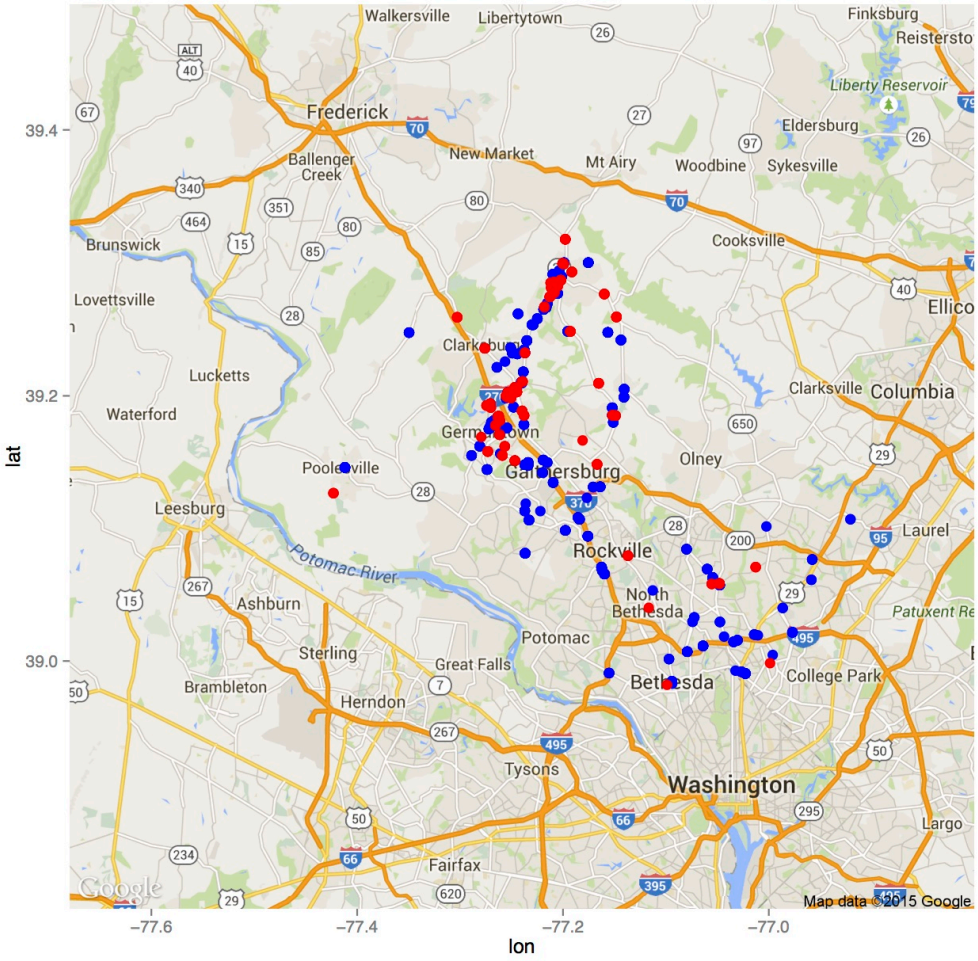

Figure 3. Plot of traffic violations with alcohol involved by gender. Males are represented by blue points and females are represented by red points.

Another noticeable characteristic is that traffic violations with fatality involved have a higher percentage in females' violations than in males'. Table 1 shows this characteristic with percentage numbers.

Table 1

Number of traffic violations in total and with fatality involved for males and females

	In Total	Fatality Involved	Percentage of Fatality Involved
Male	445413	92	0.020%
Female	228446	70	0.031%

Trends of the number of violations over time

The monthly distribution of the number of traffic violations were plotted for year 2012, 2013 and 2014. The plots show that the number of traffic violations in May in every year is particularly high compared to other months. Figure 4, figure 5 and figure 6 display the monthly distribution in bar plots.

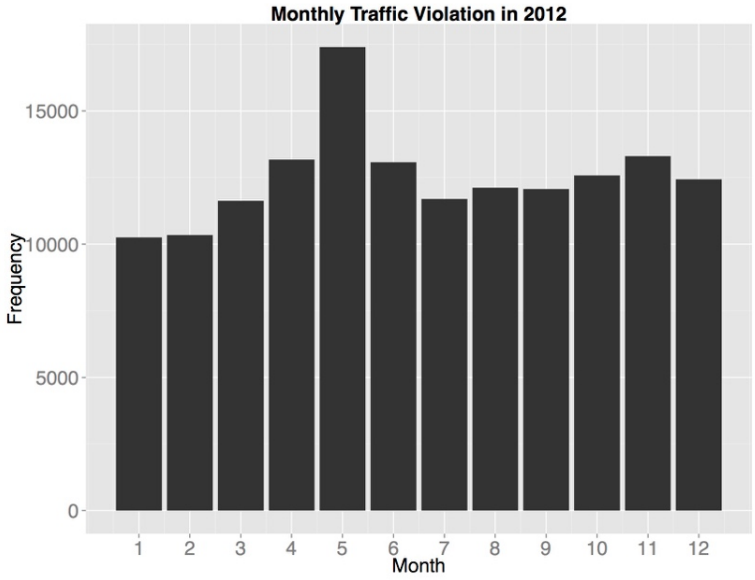

Figure 4. Bar plot of monthly traffic violations in 2012

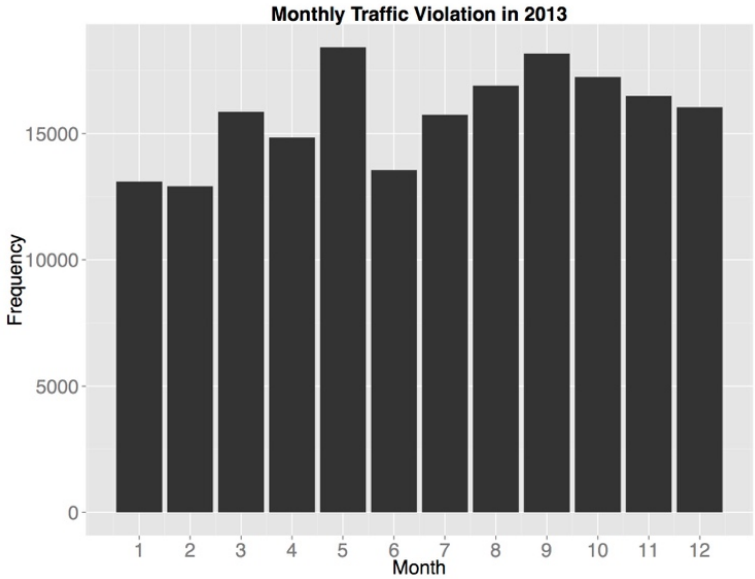

Figure 5. Bar plot of monthly traffic violations in 2013

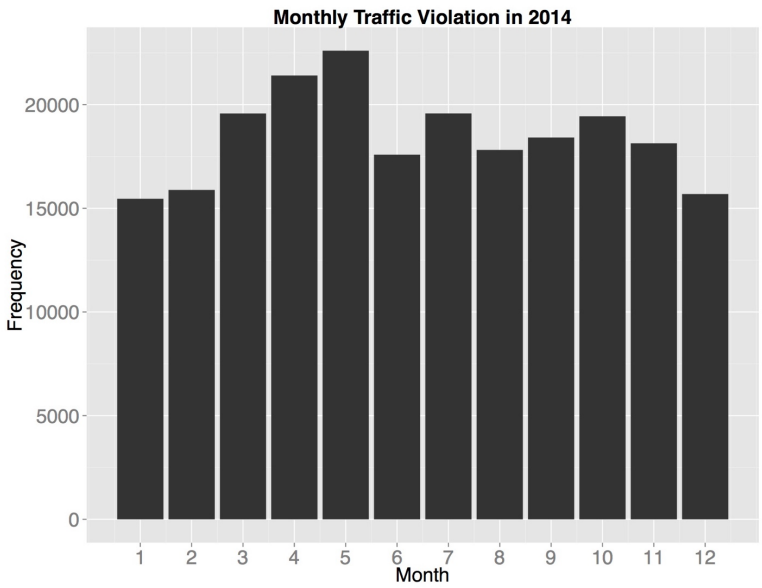

Figure 6. Bar plot of monthly traffic violations in 2014

Student's t-test was also run to test if there is a similar distribution pattern of traffic violations within a year with the significance level set to 0.05. The null hypothesis is that there is a similar pattern of monthly traffic violations distribution. The alternative hypothesis is that there is not such pattern. The test was run between 2012 and 2013, 2013 and 2014, 2012 and 2014. The p-values of these 3 tests are 0.0002609, 0.003938, 4.04e-07 respectively. Therefore, all three p-values are too small. The null hypothesis is rejected in favor of the alternative hypothesis.

The yearly traffic violations were also plotted using a bar plot. The plot shows that there is a strong linear relation. Therefore, a linear model was calculated. The slope is 36176 and the intercept is -72632751. The total number of traffic violations in 2015 is estimated as 261889. Figure 7 shows this yearly traffic violation distribution.

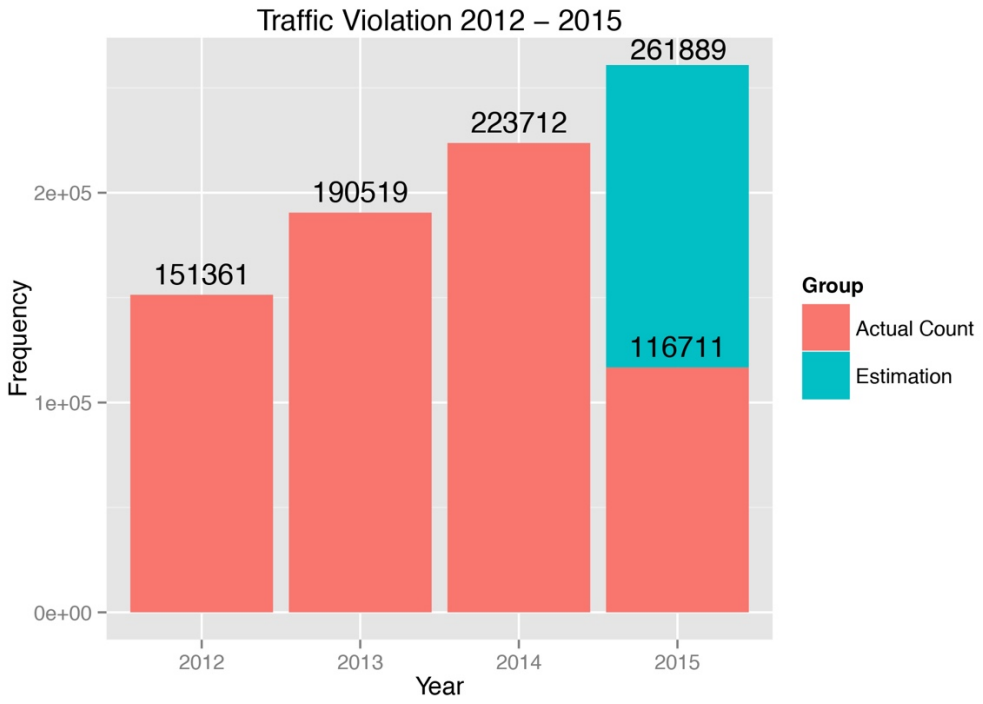

Figure 7. Bar plot of yearly traffic violations from 2012 to 2015. The blue part is the estimation for the last six months in 2015.

Discussion

The results obtained from traffic violation data of Montgomery, Maryland show some interesting insights.

Possible reasons of some of the characteristics discovered

Traffic violations happen at crossroads most. Although traffic violations happen on a lot of roads over the whole Montgomery county, the reason why most of them happen at crossroads is probably because there are more traffics going on which increases the chance of violating the traffic rules.

Different distribution pattern of alcohol involved violations between gender. The distribution of alcohol involved violations by males are all over the map of Montgomery,

Maryland. One possible reason may be that males tend to drink not only in downtown area but in their home or bars in the country side as well. Another possible reason would be that males tend to drive around after drinking which causes that wide distribution.

The number of traffic violations increases over the years. As stated in the introduction, although the United States already has a large number of vehicles registered, the number is still steadily increasing. More cars in the United States is probably the main cause that leads to more traffic violations happening.

Conclusion and Future Directions

Conclusion

In conclusion, some interesting characteristics are discovered by analyzing the traffic violation data of Montgomery, Maryland. Several analytical methods were used during the analysis such as k-means clustering algorithm and linear regression to help get better insight. For some of the characteristics, possible reasons are given basing on common sense. However, more formal conclusions should be given after deeper analysis of the data.

Future Directions

There are still many things in this dataset that can be explored. In the future, the analysis can focus on exploring the relation between traffic violation and road condition. Also, the relation between the types of traffic violation and the locations of traffic violations is another interesting question to be explored and answered.

References

1. United States - vehicle registrations 2012 | Statistic. (n.d.). Retrieved July 14, 2015, from http://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990/