Traffic Capstone Report

I. Proposal

Traffic jams are one of life's most annoying obstacles with people voting them as the 14th (news.com.au, 2013) and 19th (thesun.co.uk, 2018) most irritating things about life. In fact, a study by the Texas A&M Transportation Institute found that the average American commuter spends 42 hours per year waiting in traffic jams (New York Times, 2019).

This waiting is not only frustrating, but also dangerous. Road rage is a leading cause of traffic deaths and according to a study questionnaire of 200 people, traffic jams are one of the biggest factors leading to this type of aggressive driving (Suliman, 2003). Other studies have shown that rush hour between 3pm-6pm is the most dangerous time of day to drive and that Saturday is the most dangerous day of the week to drive due to the number of cars driving and the number of cars on the road (BacTrack, 2016). Dangers due to traffic are not limited to the road either as a study found that domestic violence is 9% more likely to occur after an evening commute on the two busiest highways from 2011-2015 (New York Times, 2019).

Traffic not only costs lives, but also has a financial impact on residents as well. The average increased cost of traffic is $1700 in the United States, $2500 in France, and is as high as $6000 in Los Angeles. These costs are highest for the trucking industry with traffic costing $74.5 billion for the industry in 2015 and 1.2 billion hours in 2016 (Fleetowner, 2018). Truck drivers are also paid per mile rather than per time, so traffic jams not only cost them valuable time, but also directly impacts their earnings per week too.

With traffic affecting so many aspects of people's lives, one would hope that the problem is being addressed. However, the problem continues to worsen as the number of drivers continues to climb due to increases in population size and increases in the average number of cars per person. The number of drivers in the United States increased about 35% from 1990 to 2017 (Statista, 2019) and the number of cars on the road rose 2.2% from 2018 to 2019 (thedrive, 2019). Throughout this time, road expansion has remained stagnant with only a 6% increase of 8.2 million miles to 8.7 million miles from 2000 to 2017 (bts.gov, 2017).

There are various different solutions proposed to address this problem which usually involve increasing spending on rebuilding the infrastructure or increasing efficiency of spending. However, analyzing currently available data is a low cost method of trying to address this problem. By looking at different variables, companies and people can better plan their commute to avoid traffic and accidents. A study was done to measure hourly traffic volume on the I84 highway near Minneapolis-St Paul, MN from 2012 to 2018. Factors such as holidays, dates and times, cloudiness, temperature, rain, and snow were all measured and compared. By performing analysis on these factors, people can understand how much of an effect these variables will truly have on traffic conditions.

II. Data Wrangling

The data was found from the UCI Machine Repository. It has 48,204 instances, has no missing values, the data was published to the site on May 9, 2019. There are 9 attributes: traffic volume, cloudiness, weather main categories, weather descriptions, rain, snow, holidays, temperatures, and dates and times. The data was formatted as a csv file and was formatted as a series of rows with a date and time that has the attributes listed before for each time. Below is an example:

holiday	temp	rain_1h	snow_1h	clouds_all	weather_r	weather_c	date_time	traffic_vol
None	288.28	0	0	40	Clouds	scattered	10/2/2012 9:00	5545
None	289.36	0	0	75	Clouds	broken clo	10/2/2012 10:00	4516

In my opinion, the most useful way of constructing the data was listing the traffic volumes of each category within the attributes one attribute at a time. Traffic volume is defined as the number of cars passing through the road within the hour. First, I wanted to group the data by temperature. The data given had temperature in Kelvins, but I converted the units to Fahreinheit. Then, I grouped the data by temperature and average traffic volume as below:

```
temp
-459.670    1318.200000
-21.568     1462.000000
```

Second, I wanted to compare the averages of the traffic volumes by holiday. To do this, I grouped the data together based on the holiday category and then calculated the average traffic volume for each holiday. Below are two rows for an example of this:

holiday	
Christmas Day	827.500000
Columbus Day	519.400000

I used the same group by method to find averages by weather description:

weather_main	
Clear	3055.908819
Clouds	3618.449749

Then, for weather description:

weather_description	
SQUALLS	2061.750000
Sky is Clear	3423.148899

I was also able to use these same methods to look at how rain and snow can affect the traffic conditions. The numbers below is the amount of mm of precipitation that was on the road within the hour that was recorded:

Rain		Snow	
0.00	3257.336321	0.00	3260.238508
0.25	3261.965190	0.05	2410.928571

The next attribute was cloud cover which is defined as the percent of the sky covered by clouds. I was able to show the data using the same group by mean data:

clouds_all	
0	3383.100604
1	2791.397539

I also wanted to determine the average traffic volume over months, years, and by hours of the day. To do this, I converted the day_time column into a panda date_time series. Then, first grouped the data by months. The result showed the year on the far left column and the numeric representation of the month in the middle column:

2012	10	3486.740373
	11	3198.431847
	12	2983.665635
2013	1	3153.654391
	2	3163.216179
	3	3178.204196
	4	3299.192758

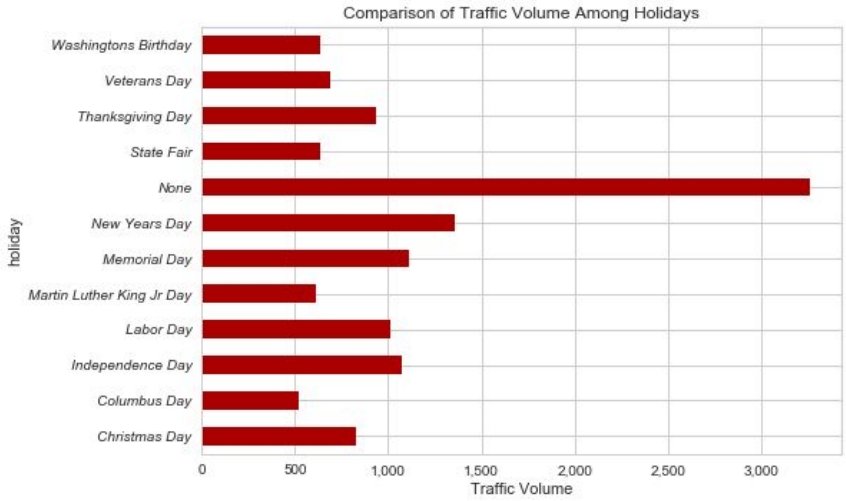
Finally, I grouped the panda data series by hour to find the average traffic volume for each hour of the day. The result was a table with the left column showing the military time hour and the right column showing the traffic volume:

```
date_time
0       834.781051
1       516.449000
2       388.353640
3       371.090864
4       702.551889
```

III. Data Visualization

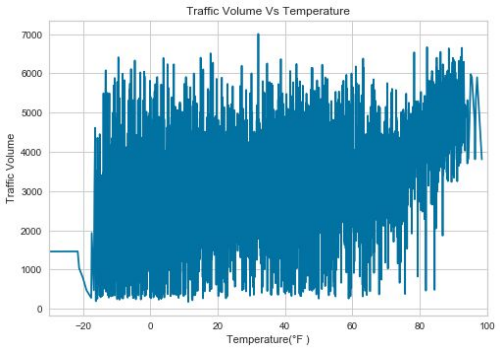
Now that the data is cleaned and in a usable form, I wanted to visualize the data in order to understand which attributes have promise for having an impact on traffic volume.

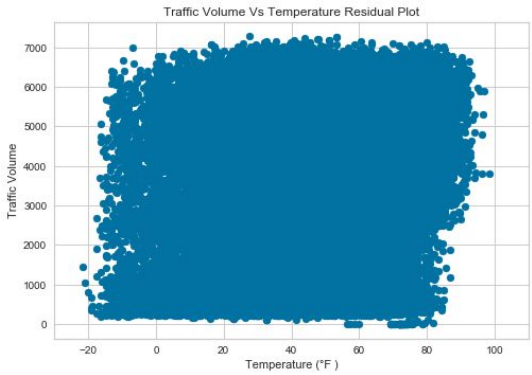
First, I wanted to show a visual for average traffic volumes among holidays. The best way to visualize this change was a bar graph since there were specific categories that I already had and wanted to compare. The bar graph is below. Based on this visual, I will continue to analyze the holiday data since there is a drastic visual difference between the 'None' category and the holidays.



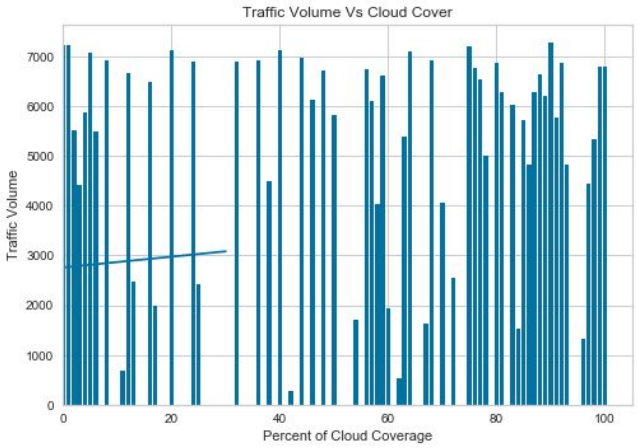
Next, I wanted to look at the temperature data. This visual was more complicated than the holiday since there was not a specific set of categories. Instead, it was 48,000 measurements given over a sequential time. First, I plotted a line graph of the data as below on the left. However, this was difficult to interepret on its own. So, I also performed a line regression and then created a residual plot

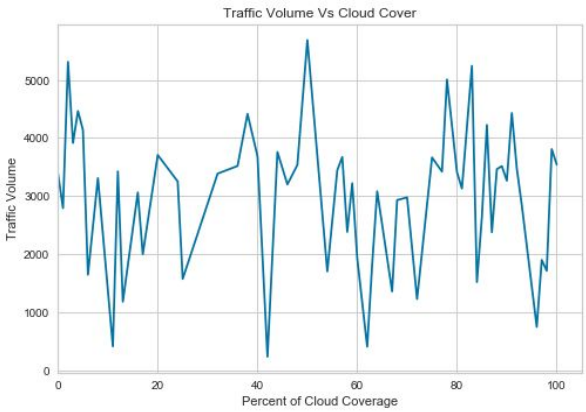
on the right. Based on this plot in addition to the difficult to interpret line graph, I felt that there was not a high chance of finding a correlation between temperature and traffic volume.



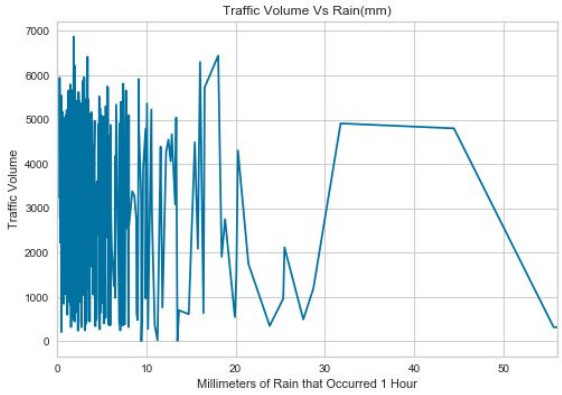


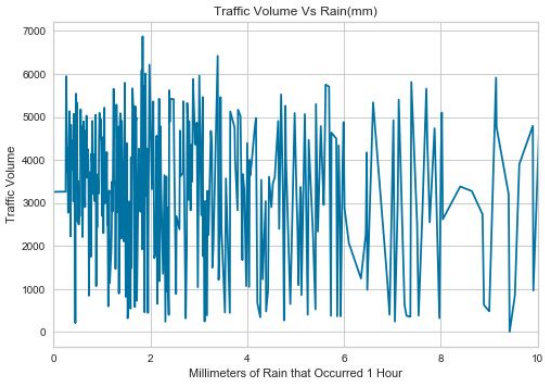
A line regression was also created for the percent of cloud coverage data and was visualized over a bar graph(left) and a line graph was made as well (right). There appear to be higher rates of traffic volumes at certain points of cloud coverage such as around 2%, 50%, and 80%. There also appeared to be low points at 10%, 42%, 62%, and 95%. These trends also seemed to be hard to create calculations, so this attribute will be left out as well.

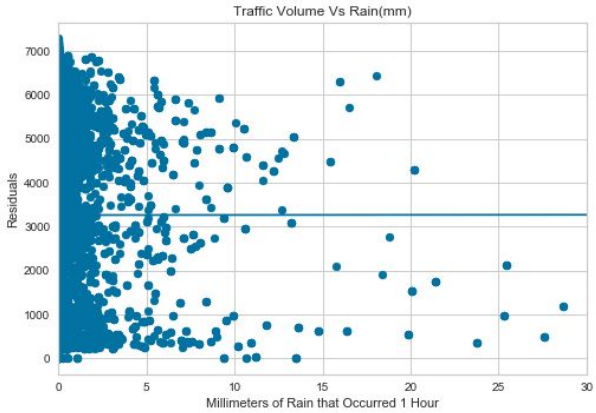




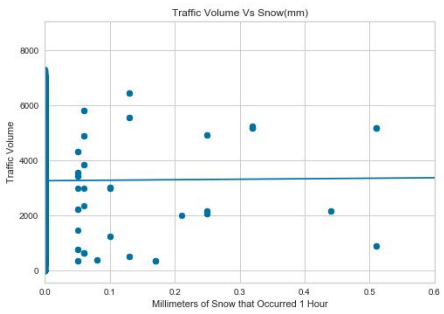
A residual scatter plot was also made for rain and snow as well as line graphs. Rain appeared to show now relationship until greater than 20mm where traffic tends to drop.

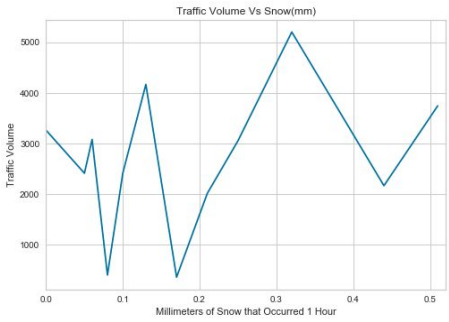




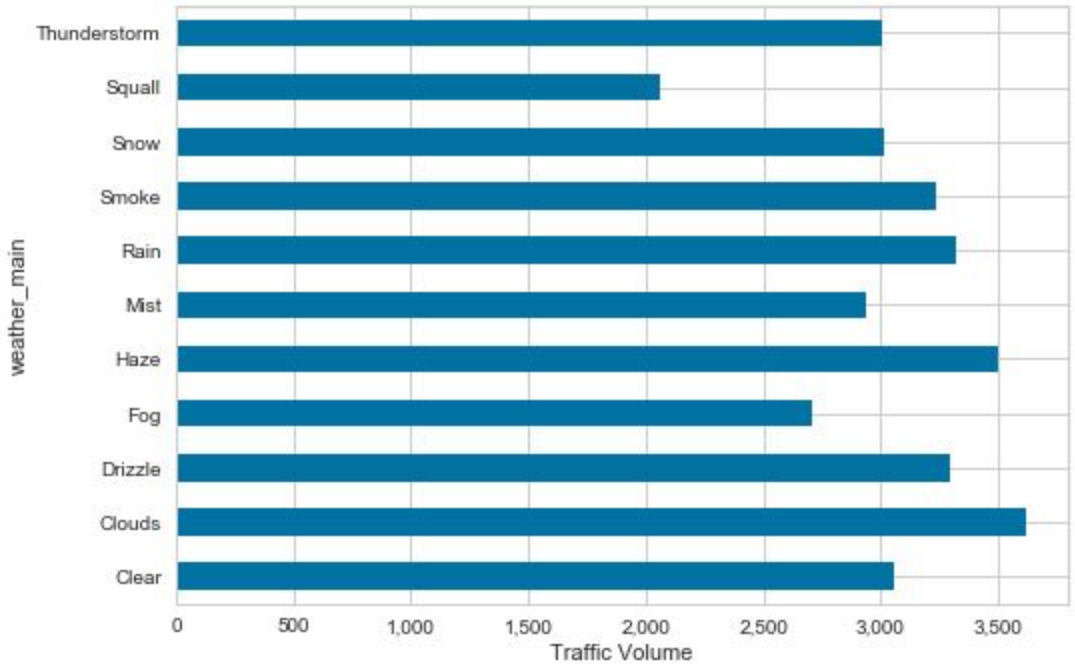


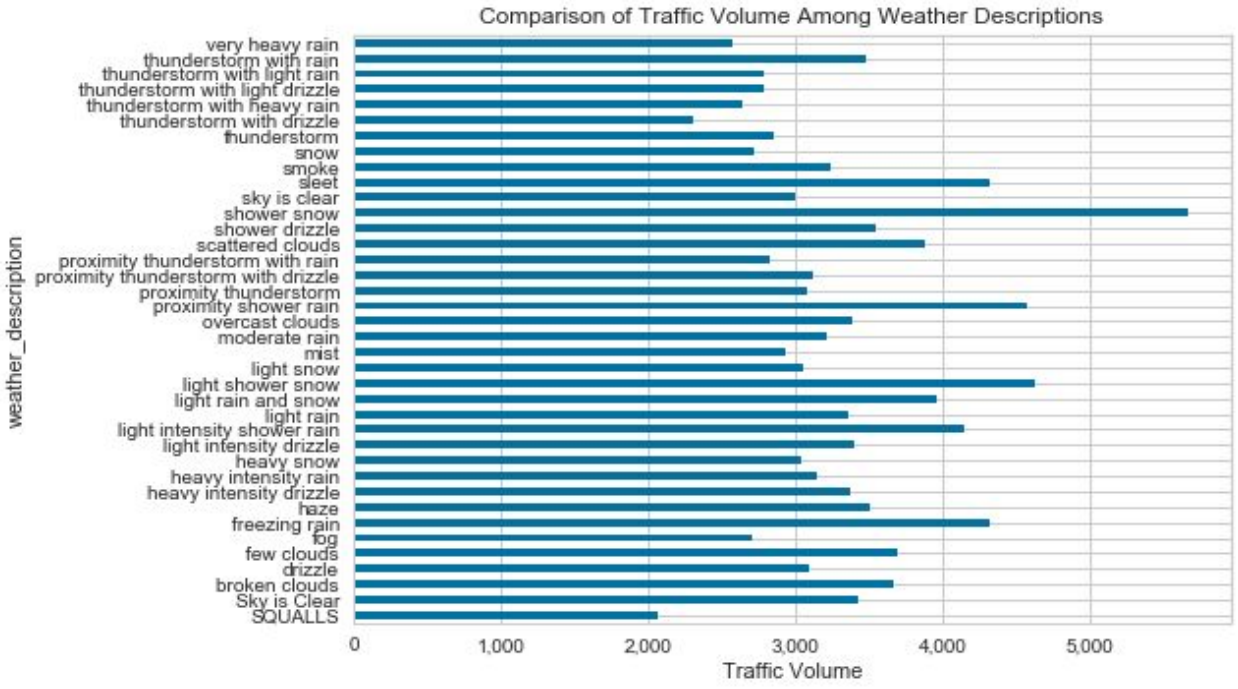
The snow plots below were difficult to find a clear relationship, and so were not further explored.



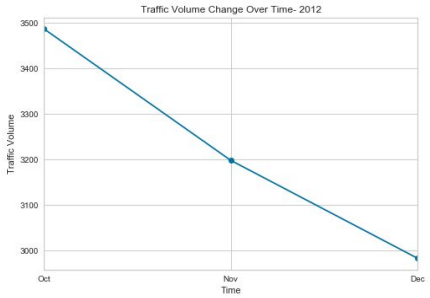


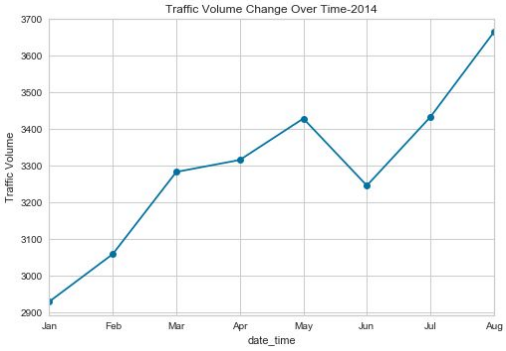
Weather descriptions were grouped and traffic volume was visualized with a bar graph. The bar graph clearly shows that weather descriptions have clear recognizable differences. It appeared snow showers and light snow had the highest traffic volumes. This relationship will be further explored.

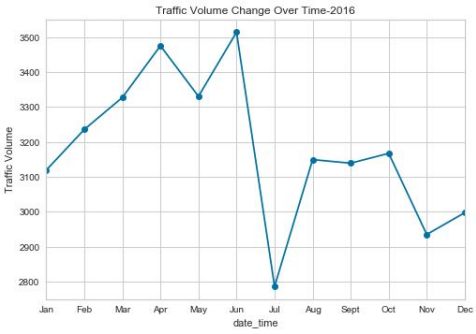


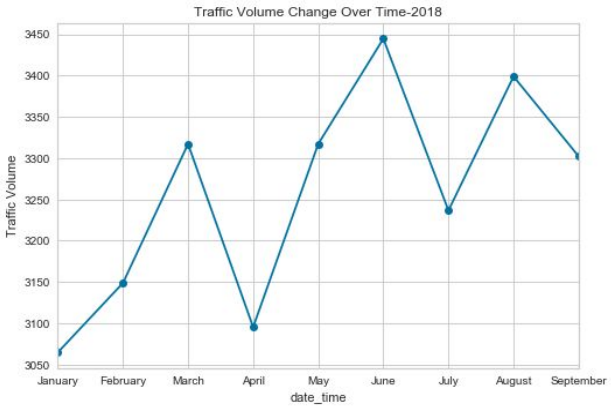


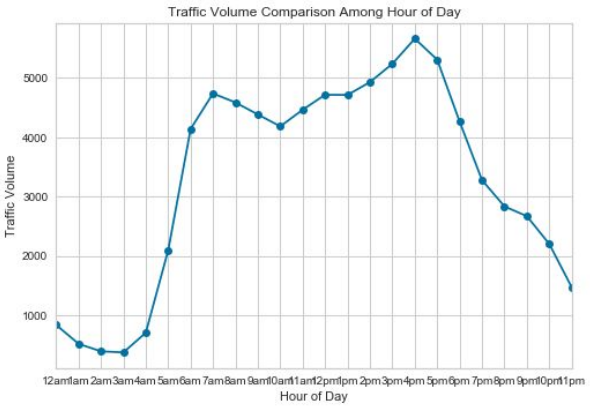
Line graphs were used to depict the traffic volumes over time both in months within years and times of day. Similar trends are seen throughout the plots with lows in the summer and December and highs in the fall and spring. Also, highs for the time of day were seen for around 8am and 5pm and lows toward the early morning and late night. These trends will be further explored.











IV. Statistical Analysis

In order to determine statistical significance, two tests were performed based on the types of data. The categorical data was had chi-square tests performed on the data in order to determine whether there was statistical significance among the groups. Statistical significance was confirmed between the groups via high P-value divergent values and very low P-values among the holidays, weather descriptions, and weather categories. Also, the chi-square test confirmed that there was statistical significance among the time series for months and hours.

The next test that was done was the Pearson Correlation test. This was to determine whether there was statistical significance between the two numerical categories for snow, rain, temperature, and cloudiness. Based on these tests, cloud cover and rain had statistically significant differences due to having lowing P-values(0.03 for cloud cover and 0.07 for rain). However, the snow and temperature showed P-values and were not statistically significant with values of 0.36 for temperature and 0.26 for snow.

Now that we have determined whether there is statistical significance among our groups, we will perform machine learning analysis to try to fit models and predict data based on these groups.

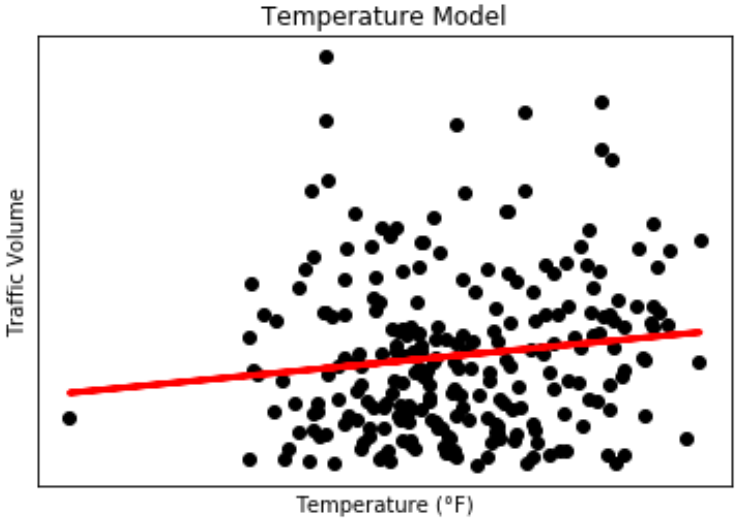
V. Machine Learning Analysis

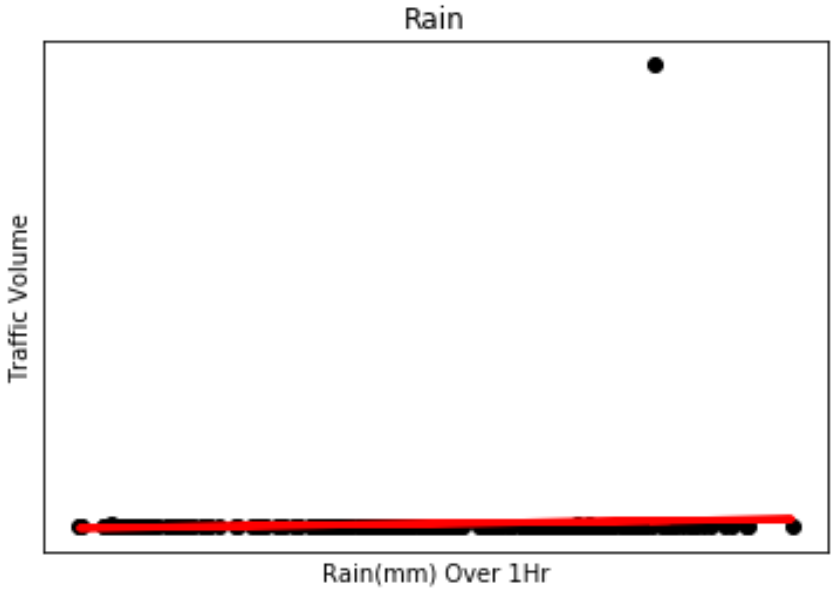
Regression analysis was performed and plotted for all of the attributes. To perform the regression analysis, the numerical data such as snow, rain, and temperatures were split into testing and training data. The training data was used to create the machine learning model after being reshaped and then the test data was created and plotted. The x variable for the plot was the attribute being measured while the Y variable was the Traffic Volume. A linear regression line was calculated and plotted on top of the line and the strength of the relationship can be measured based on how linear the plot points are around the line.

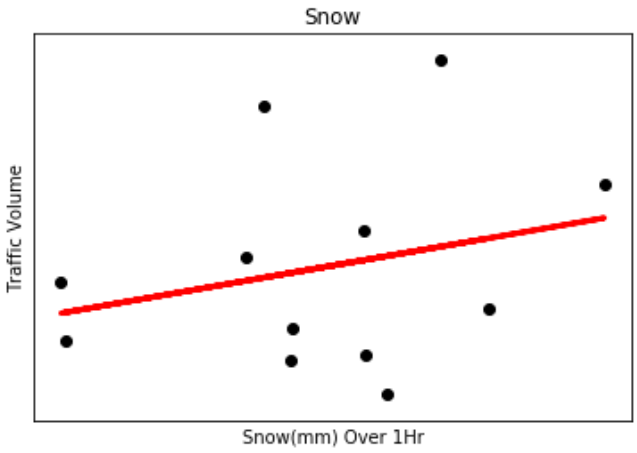
The categorical attributes were time, weather description and types, and holidays. Its difficult to create a machine learning model for this type of data, so a label encoder was used to convert them into numbers. Then, the same type of regression point and line modeling was applied between the labels and the traffic volume.

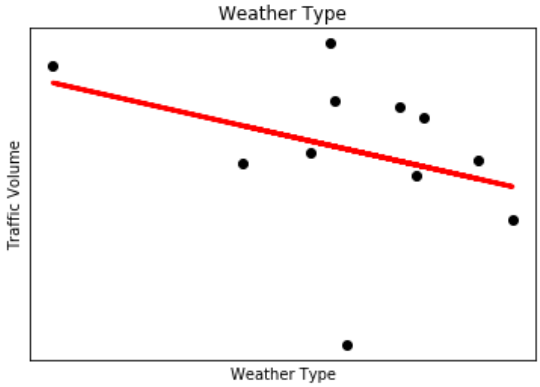
Based on the clustering of the points around the line, rain was by far the most predictable variable out of the attributes as the regression line was on top of almost all of the points.

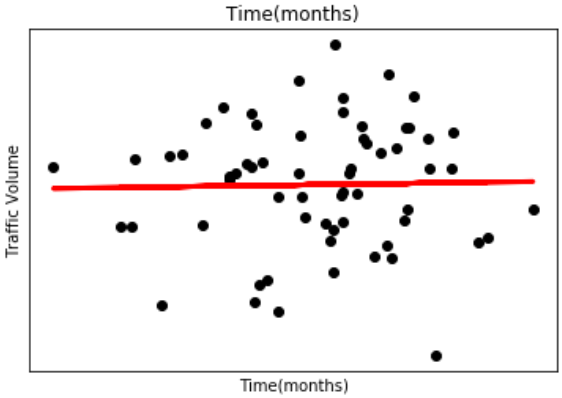
Holidays, weather type, and weather description were also strong predictors as well. To a lesser extent, temperature and time of day may also be useful for predictions as well due to the closeness of the points to the line, despite a segment of outlier points. Time over months, snow, and cloud coverage were the least predictive out of the models.

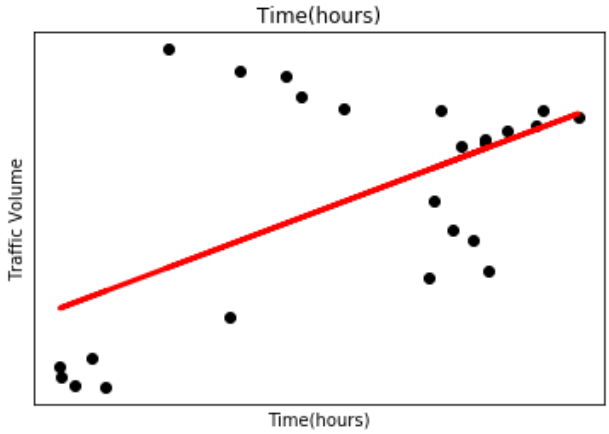












VI. Conclusion

In conclusion, we found some predictors for traffic conditions. The strongest predictors based on our machine learning model was train, so in order to avoid traffic congestion, one could decide to save certain tasks such as shopping for rainy days. Another useful predictor for everyday tasks is time of day. In order to avoid traffic, one should avoid rush hours like the

morning and early evening and instead travel during the early morning or late at night. However, this would not be a useful suggestion for those wanting to avoid traffic due to safety reasons as traveling in the dark and without proper sleep is even more dangerous than traffic.

Snow and cloud coverage are a little more difficult to use as predictors as snow does not have a strong relationship and cloud coverage is difficult to measure and see. Weather descriptions can generally predict whether or not there is traffic, however, it probably would not be safe to travel for most people especially in conditions such as thunderstorms or squalls. However, if a person's vehicle is equipped with all-wheel drive or other snow-specific enhancements, they may be able to take advantage of less traffic during snow.

For long-term planning, one could use the depictions and modeling from the time of year parameters. Based on the visualizations, one should try to plan road trips for the end of spring, the middle of summer, and the middle of winter.

However, when making these predictions it is important to understand the limitations of the data. The data is limited to five years and is limited to only one area of the country and a specific type of area. Further research will be needed on different parts of the country and a longer timeframe.

Works Cited

U.S. Department of Transportation, Bureau of Transportation Statistics, Transportation Statistics Annual Report 2017 (Washington, DC: 2017).

"The Cost of Traffic Jams." *The Economist*, The Economist Newspaper, 3 Nov. 2014, www.economist.com/the-economist-explains/2014/11/03/the-cost-of-traffic-jams.

Frakt, Austin. "Stuck and Stressed: The Health Costs of Traffic." *The New York Times*, The New York Times, 21 Jan. 2019, www.nytimes.com/2019/01/21/upshot/stuck-and-stressed-the-health-costs-of-traffic.html.

Murphy, Annabel. "Brits Reveal 40 Most Annoying Things from Late Trains to Cash Machine Charges." *The Sun*, The Sun, 28 Aug. 2018, www.thesun.co.uk/news/7119615/most-annoying-things-in-modern-life/.

NEWS.COM.AU, DAILY MAIL with. "The 50 Most Annoying Things in Life." *NewsComAu*, Daily Mail, 8 July 2013, www.news.com.au/lifestyle/the-50-most-annoying-things-in-life/news-story/eb0d7f0ec25ef746edc9101886a45e33.

Ray, Aaron. "The Most Dangerous Times on the Road." *BACtrack*, BACtrack, 30 June 2015, www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road.

"Road Rage Statistics Filled With Surprising Facts." *Elite Driving School*, 8 Mar. 2015, drivingschool.net/road-rage-statistics-filled-surprising-facts/.

StackPath, www.fleetowner.com/news/article/21703166/how-traffic-congestion-affects-the-trucking-industry.

Suliman, Mohd R, and Wa'El H 'Awad. "Aggressive Driving Is a Major Cause of Traffic Accidents and Road Rage in Jordan." *Proceedings of the 2nd International Driving*

Symposium on Human Factors in Driver Assessment, Training and Vehicle Design: Driving Assessment 2003, 21 June 2005, doi:10.17077/drivingassessment.1118.

"Truck Driving Per Mile Salary." *Truck Driving Per Mile Salary | AllTrucking.com*, www.alltrucking.com/faq/per-mile-trucking-salary/.

Tsui, Chris. "There's a Record Number of Cars on US Roads, and the Average Age Is Nearly 12 Years Old." *The Drive*, 28 June 2019, www.thedrive.com/news/28741/theres-a-record-number-of-cars-on-us-roads-and-the-average-age-is-nearly-12-years-old.

Wagner, I. "Number of Drivers Licensed in the U.S. 2017." *Statista*, 20 Aug. 2019, www.statista.com/statistics/191653/number-of-licensed-drivers-in-the-us-since-1988/.