# Traffic Relations

Extended Abstract*

David Nyberg
CU Boulder
Boulder, Colorado
dany3289@colorado.edu

Eric Ha
CU Boulder
Boulder, Colorado
erha5113@colorado.edu

Nicholas Sugarman
CU Boulder
Boulder, Colorado
nisu7560@colorado.edu

## ABSTRACT

This paper is in ACM SIG format and covers an overview of our data mining project completed in the Spring of 2018 at University of Colorado, Boulder. Data mining is useful tool in gaining knowledge from large data sets. For this project, we sought to analyze a data set that contains statistics regarding traffic information for the year 2015. The data set was retrieved from the website, Kaggle. They provided us with information about traffic ranging from the type of road, traffic volume per hour, and more. Through our knowledge of data mining, we hoped to find a relationship between volumes of traffic and whether seasons of the year affect this quantity. Furthermore, the data set provides information about traffic volume relative to the type of the road (e.g. Urban vs. Rural). In order to retrieve this information, it requires an analysis and transformation of a mass amount of data. Our team cleaned the dataset using Python, in order to evaluate only the data we needed to reach our answers. After compiling our findings, what we observed was that high traffic occurs during 4:00PM-5:00PM on average, for both urban and rural roads. We further did an analysis to compare the differences between the two types of roads. We found that urban roads handle a much higher volume of traffic than rural roads. Results further showed that seasons do have an effect on volume of traffic and they differ depending on the type of road. Additionally, we integrated classification and clustering to our data set to create a regression model to predict future traffic volumes. The following report provides an organized and thorough descriptions of the project's purpose along with the methods and processes used to reach our conclusions.

## KEYWORDS

ACM, LaTeX, Traffic, CU Boulder

**ACM Reference Format:**
David Nyberg, Eric Ha, and Nicholas Sugarman. 2018. Traffic Relations: Extended Abstract. In *Proceedings of Traffic Relations Data Mining 2018 (BOULDER, 18).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

*The full version of the author's guide is available as `acmart.pdf` document


*BOULDER, 18, May 2018, Boulder, Colorado USA*

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

## 1 INTRODUCTION

Through the access of a large traffic data set, we seek to find knowledge involving a correlation between seasonal differences and these differences' potential effects on traffic volume. An example of a finding would be that there is an increase in traffic during winter seasons. We hope to apply this knowledge towards building more efficient roadways which could reduce the traffic delays that occur during seasons of high traffic. Interesting knowledge that we also seek to find would be to find a correlation between traffic volume and different types of roads, such as urban and rural. To extend our previous example, we may find that there is a larger traffic volume at rural roads during the winter compared to other types of roads. We also hope to discover whether certain times of day have a higher traffic volume than others.

These questions are important to commuters attempting to beat major traffic rushes, city planners who are attempting to carry out new construction projects, or even construction companies trying to figure out the best time to repair and build roads. Our findings could be used not only to warn drivers to avoid certain types of roads at certain times, but they could help determine the optimal time to begin new construction projects so that traffic would be affected in the smallest way possible and when construction companies should send workers out to carry out these projects. These things would help minimize the delays caused by traffic and potentially even keep construction workers safe, since our findings would help ensure they worked during low-traffic periods where they are less likely to be in danger of getting hit by passing vehicles. In addition to answering these questions, we also want to gain insight into the behavior of people's daily commutes. For instance, if we evaluate the relationship between traffic volume and seasons of the year, we can see when people tend to drive more often. Our goal is to be able to form a regression model that would be a prediction for future traffic and to help inform users when is best to avoid high volumes of traffic.

## 2 RELATED WORK

Traffic issues are an ongoing problem in the United States and lots of previous work has been done to research and try to fix traffic around the world.

https://www.zmescience.com/research/technology/google-maps-traffic-05443/

This article talks about Google maps and how it uses predictive technology in tandem with traffic data to predict when and where traffic jams will occur. It can predict future traffic jams by referencing past traffic data and comparing it with current conditions to

see if there is a significant overlap. Similar work has been done to determine what days are more dangerous to drive on then others

https://www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road.

This data is used to measure when drunk drivers are most likely to be on the road.

https://www.nature.com/articles/srep37300

Research has been done which investigates what weather conditions are most likely to put people in danger and where. The study also tries to take factors such as socioeconomic status into account, which is a little beyond the scope of our project but still fascinating.

http://kalw.org/post/driving-apps-waze-are-creating-new-traffic-problems

This article writes about how traffic phone applications such as Waze which is similar to Google maps showing users the fastest route to a location are actually creating more traffic issues. This is closely related to our project as we are going to look at when people use local roads depending on the time of year and how many lanes they have, this article mentions how applications similar to Waze are sending people onto these small rural roads in order to find a 'faster' path but it is actually slowing down traffic according to their studies.

## 3 DATA SET

Data provided from:
https://www.kaggle.com/jboysen/us-traffic-2015/feed

The dataset is a comprehensive view of various factors which affect the volume of traffic in a particular area at a particular time. Along with the date and times the data was taken at, the data assess factors such as the direction being traveled in, what kind of road is being driven on, the state code, and the traffic volume at various times. The data set contains mass amounts of data points that are capable of providing many different kinds of knowledge pertaining to traffic. There are over thirty attributes and 7.1 million data points within the set. As shown in **Figure 1.1**, this is a Python output of the data set's size of how many data objects it has by the number of each attributes.

For the purpose of our analysis, the variables in the dataset we will be focusing on most closely are the months the data was charted in, the type of road the data was charted on, and the traffic volume counted during each hour time frame. This can be achieved by data mining techniques such as data cleaning and data reduction. Through data transformation, a cleaned up data set will be available for analyzing more specific relationships lying within the initial data set. From there, we will be able to gauge what season each data point occurred in and what times traffic would be least likely to be affected by potential construction projects. We will also be able to discover whether there is a major difference between urban and rural roads that would potentially require them to be treated differently than each other.

**Figure 1.1 - Initial Data Set Size**

```
In [7]: traffic.shape
Out[7]: (7140391, 38)
```

## 4 TOOLS

Our project utilizes a variety of tools to clean, reduce, transform, and analyze the traffic data. The main tools that we used heavily were Python, Github, and RapidMiner.

### 4.1 Python

Python is the primary programming language we used to perform the data mining processes. We also used Anaconda and ipython notebooks as an easy way to create scripts and manage them. Anaconda also utilizes Jupyter to organize our ipython notebooks. They were helpful in executing specific commands in the code along with displaying outputs. Python is extremely useful in its collection of open source libraries for data mining. This includes pandas, matplotlib, and numpy as they provide programming methods to manipulate the data and display it visually. These libraries have the capability to read in a data set file and place it into a data frame that we were able to perform alterations on. Python is also excellent at handling mathematical equations, especially when open source libraries are used. This proved to be useful when we compiled results from our data. Our data set contained millions of data objects that Python was able to compute mathematical formulas on and output a desired result. Python is also flexible enough to store data which has previously been acquired, which will be useful for calculating the multi-dimensional correlations between different attributes.

### 4.2 Github

Github was also an important asset towards achieving our goal because it served as our version control as well as our software for project tracking. It also facilitated collaboration and ensured that we all had access to the same data programs, which smoothed the process of mining our data greatly. In addition, Github provided a way for the team to collaborate on the project in real-time while team members didn't need to be physically met together. This proved to be useful in finishing tasks for the project without compromising all of the team's individual schedules.

### 4.3 RapidMiner

Another tool we used was RapidMiner. RapidMiner is a software which is used to create machine learning models and visualize data. It is a paid software, but for the purposes of this project we used a free trial, that was able to provide us enough to reach our results. It was nice to use a GUI instead of raw python code, it made visuals very nice and easy to see as well as building an easy to use flow chart of analysis instead of debugging code.

### 4.4 Scikit-Learn Library

Scitkit Learn is an open source python library which is the main source of data mining and machine learning algorithms that we used. We used to clustering functions mainly from this python library.

## 5 MAIN TECHNIQUES APPLIED

In its raw form, our data was too large and broad to make sense of. Before we could mine the data, we had to clean, reduce, and transform it in such a way that we were able to easily pick through the variables we wanted to learn from. Furthermore, reducing the data set allowed for a quicker compile time to analyze our data. Once we got the data to the point where we were able to mine it, we applied classification and clustering techniques in an attempt to find patterns between the variables we wished to investigate. We also used RapidMiner to create a Regression Model, which helped glean even more information from our data.

### 5.1 Data Cleaning and Preprocessing

The initial form of our data contained a wide variety of variables for each individual data point. These variables include such varied factors as the direction in which a vehicle was traveling at any given time, the state code for each vehicle, and the functional classification for each vehicle. While these data points were interesting, we determined that they did not factor into our main goal of determining whether seasons, the time of day, or types of roads influenced traffic. That meant that these data points were extraneous and detrimental to our project because they ate up our computers' memory.

Python became our best tool in helping to filter our data. We ran our dataset through libraries such as pandas, numpy, and ipython notebooks in order to eliminate the unnecessary columns. An ipython notebook was made and was coded so that a user could go and edit the code to exclude any attributes they desired.

### 5.2 Data Reduction

In addition to reducing our data by removing unnecessary columns, we also removed data points with incomplete or corrupted data. This allowed all our data points to have the same number of variables and amount of information as any other data point. As well, we used regular expressions (i.e. regex) within our Python code to help reduce a clean data set even more to get specific information. For instance, we filtered out any road that was labeled as 'Rural' in order to have a data set that had only information of 'Urban' type roads and vice versa.

We also had to be careful when determining what was and was not an outlier. There were certain data points which initially seemed like an anomaly, but upon closer inspection we realized that it was actually an outlier.

### 5.3 Data Transformation

In addition to cleaning and reducing the data, an ipython notebook creates and outputs a new .csv file with the cleaned data set. This transformed set allowed for us to use that file and run an analysis that evaluates the initial data set in a more specific manner. This elimination of unnecessary data freed up some of our computers' memory, ensuring we would be able to sort through the necessary data in an efficient manner.

We also created two separate files for our .csv data. These two files each contained one type of road, either rural or urban. By splitting our data apart into two distinct files, this allowed us to get information on each individual type of road. This allowed us to determine whether the type of road had any impact on the volume of traffic during a particular time or season. These findings proved useful in helping us mine results from our dataset.

**Unfiltered Data Example**

| | date | day_of_week | functional_classification_name | month_of_data |
|---|---|---|---|---|
| 0 | 2015-04-07 | 3 | Rural: Principal Arterial - Other | 4 |
| 1 | 2015-09-26 | 7 | Urban: Principal Arterial - Interstate | 9 |
| 2 | 2015-06-16 | 3 | Urban: Principal Arterial - Interstate | 6 |
| 3 | 2015-04-26 | 1 | Urban: Principal Arterial - Interstate | 4 |
| 4 | 2015-05-23 | 7 | Rural: Minor Arterial | 5 |
| 5 | 2015-07-25 | 7 | Urban: Principal Arterial - Other Freeways or ... | 7 |
| 6 | 2015-09-10 | 5 | Urban: Principal Arterial - Other | 9 |
| 7 | 2015-10-27 | 3 | Urban: Minor Arterial | 10 |
| 8 | 2015-06-26 | 6 | Rural: Principal Arterial - Interstate | 6 |
| 9 | 2015-05-12 | 3 | Urban: Principal Arterial - Other Freeways or ... | 5 |

**Filtered Data Example**

| | date | day_of_week | functional_classification_name | month_of_data |
|---|---|---|---|---|
| 0 | 2015-04-07 | 3 | Rural: Principal Arterial - Other | 4 |
| 4 | 2015-05-23 | 7 | Rural: Minor Arterial | 5 |
| 8 | 2015-06-26 | 6 | Rural: Principal Arterial - Interstate | 6 |
| 10 | 2015-08-04 | 3 | Rural: Principal Arterial - Other | 8 |
| 11 | 2015-11-05 | 5 | Rural: Principal Arterial - Interstate | 11 |
| 13 | 2015-02-18 | 4 | Rural: Principal Arterial - Other | 2 |
| 19 | 2015-08-01 | 7 | Rural: Principal Arterial - Other | 8 |
| 21 | 2015-04-21 | 3 | Rural: Principal Arterial - Interstate | 4 |
| 23 | 2015-08-19 | 4 | Rural: Principal Arterial - Interstate | 8 |
| 24 | 2015-06-10 | 4 | Rural: Principal Arterial - Other | 6 |

**Classifying Data Using the Month**

| month_of_data | Traffic_Volume_at_12:00AM-1:00AM | Traffic_Volume_at_1:00AM-2:00AM | Traffic_Volume_at_2:00AM-3:00AM |
|---|---|---|---|
| 1 | 48762744 | 33820546 | 28277165 |
| 2 | 44294818 | 29856754 | 25458348 |
| 3 | 50868743 | 34342078 | 28954283 |
| 4 | 54506161 | 36522808 | 30848015 |
| 5 | 60319064 | 40953714 | 33258624 |
| 6 | 53167997 | 34621091 | 28000089 |
| 7 | 60464696 | 39733390 | 32419499 |
| 8 | 57032453 | 37644543 | 30881406 |
| 9 | 52505517 | 34998119 | 29637718 |
| 10 | 52835944 | 35489208 | 30231867 |
| 11 | 50034611 | 34325573 | 28548735 |
| 12 | 57321181 | 39583826 | 33126975 |

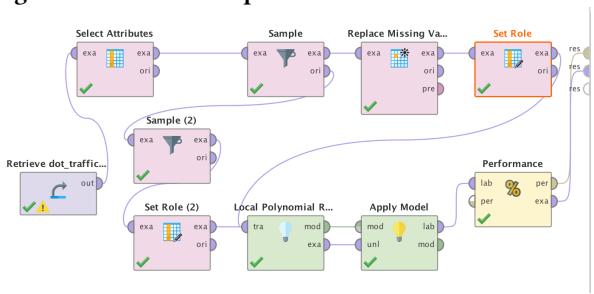
## 5.4 Classification/Clustering

In an effort to find which months had a high amount of traffic data and which months had a low amount of traffic data, we applied clustering techniques to our data. The clustering technique we used was a simple kmeans/kmedioids equation meant to isolate months that were noticeably higher or lower than the norm set by the other months. Even when kmeans was applied, however, the clustering technique proved mostly ineffective. Our research found that the datasets were actually fairly close together. The data was so close together that the clusters were almost irrelevant, no matter how many times we applied the kmedioids method.

## 5.5 Regression Model

Using RapidMiner, we attempted to predict future outcomes of traffic based off of the data we found. We did this in order to verify that the trends we noticed during the one year this dataset was collected in continued throughout multiple years. We also thought it would be useful having software that could predict future traffic patterns, as this information would be invaluable to construction projects. Below is a diagram of how a flowchart is set up using RapidMiner and how it makes a easy to use application of data mining techniques.

**Regression Model example**




Unfortunately, we ran into problems when using the RapidMiner program. The program required a high amount of memory and CPU power and our computers were unable to run it if we took a large sample size of data. This meant our sample sizes were small by necessity, negatively impacting the accuracy of our regression analysis. This lack of accuracy made us hesitant to use any data collected using this method, while not getting great results due to this limitation in RapidMiner applying OLS (ordinary least squares) in python from pandas was relatively easy and created another model. Also while these results may not be the best, the theoretical application of getting a very good model for future prediction is valuable.

## 5.6 Correlation

We applied some basic correlation analysis to the data to help determine if any of our variables were closely related. Application of this technique revealed there did not seem to be a substantial correlation between our data points. Correlation analysis was done in python using pandas corr() function and then was formatted into a table for an easy to read visual. Sadly the results of correlation were not very exciting as most of our attributes actually weren't directly correlated.

# 6 RESULTS

For the sake of this results section, rural roads primarily refer to roads which are located away from major highways such as roads in neighborhoods, while urban roads refer mostly to highways, interstates, and freeways. Early results of our data showed most of the traffic numbers are actually in the bins of 0-3 lanes of travel. This was significant because it means that for most construction projects, traffic will become exponentially worse if even one lane is taken out of commission for repair purposes.

**Figure 1.2- Total lane usage**

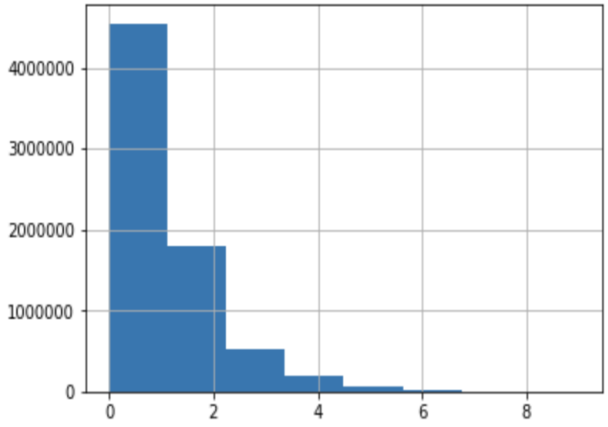



**Figure 1.3- Number of total lane usage**

| | |
|---|---|
| 1 | 2849316 |
| 2 | 1799970 |
| 0 | 1695513 |
| 3 | 527434 |
| 4 | 185915 |
| 5 | 53827 |
| 6 | 21265 |
| 7 | 6176 |
| 8 | 905 |
| 9 | 70 |

Our data revealed that on both urban and rural roads, the highest traffic volumes usually occur between 4 and 5 PM and that on average traffic is noticeably lower between the hours of roughly 8 PM and 7 AM. This data is significant because it reveals the optimal times for construction projects to occur, namely the eleven hour period between 8 PM and 7 AM where traffic is significantly lower.

## 6.1 Traffic Volumes & Time of Day

For both urban and rural roads, we discovered that there was a general period between 7 PM and 7 AM where traffic took a noticeable downward decline in overall volume, with the period between 11 PM and 5 AM being particularly low traffic periods. It is also noteworthy that on both rural and urban roads, traffic peaks at the same time: between 4 PM and 5 PM. Traffic volumes are close to the peak during both the hour before and hour after this peak period. Two hours after the peak period, there is a relatively sharp decline on both kinds of roads, followed by the beginning of the comparative "dead zone" which occurs between 7 PM and 7 AM.

One noticeable difference between urban and rural roads is that on rural roads, the curve of the data is predictable, while the urban road is a bit harder to predict. On rural roads, the data is at its lowest point between 2 and 3 AM. Once the data hits this point, the average volume of each subsequent hour will increase until it hits the 4-5 PM peak. Once this peak is reached, the average volume will decrease until it hits the 2-3 AM low point. On urban roads, however, there is an odd dip in traffic volume which occurs between 7 AM and 1 PM. While this is presumably due to the nature of jobs in urban environments keeping motorists off the road, it is still an interesting phenomenon worth noting.

**Figure 2.1- Average of Traffic Volume on Rural Roads Based on Hourly Intervals (Top)**

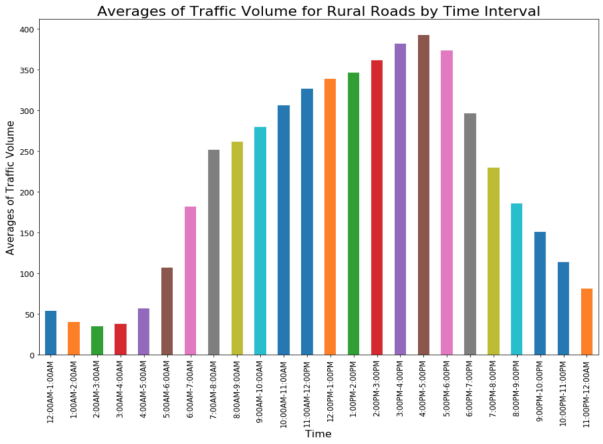
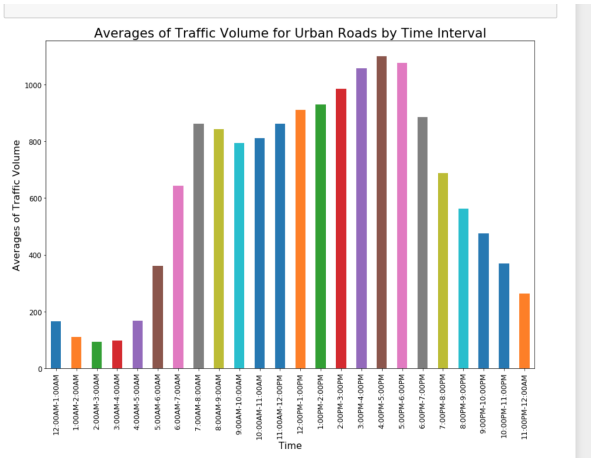



**Figure 2.2- Average of Traffic Volume on Urban Roads Based on Hourly Intervals (Bottom)**

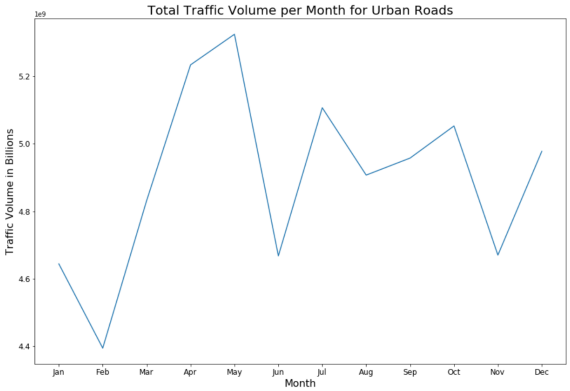



## 6.2 Traffic Volumes & Seasons

Once we determined the average traffic volumes during individual hours of the day, we examined the differences between urban and rural roads which emerged when the two were compared over a monthly basis. This data proved that there was a difference between when urban and rural roads were most commonly used.

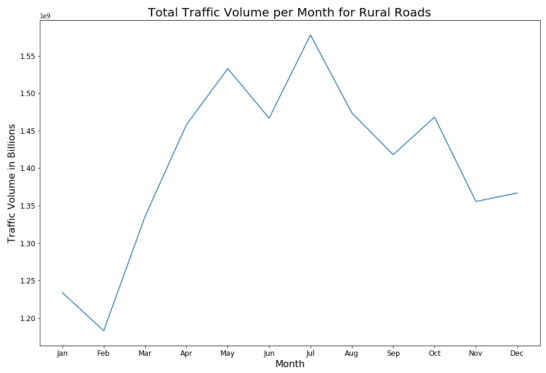
On urban roads, the months with the lowest average traffic volume were January and February. Winter was overall the least busy month for traffic, and spring was the season with the most traffic. Traffic volume peaked in May, followed by a noticeable dip in traffic volume in June.

**Figure 3.1- Urban Road Traffic Volume per Month**




On rural roads, however, the data is noticeably different. While January and February are still the least busy months and there is still a dip in traffic volume during June, the dip is nowhere near as dramatic and the data peaks in July. This makes summer the busiest season for traffic on rural roads, while winter remains the least busy season on this type of road.

**Figure 3.2- Rural Road Traffic Volume per Month**




This data indicates that winter is overall the season where traffic is at its least active, while the most active season for traffic depends on what kind of road is being driven on. This allows us to infer that during cold months with high likelihood for snow, people are less likely to get on the road. It also makes sense that summer would be the most active season in rural locations because of children being off school. Parents would take their kids to local restaurants, parks, etc. because of the increased amount of free time they have, increasing the traffic on local neighborhood roads.

# 7 APPLICATIONS OF RESULTS

Our results can easily be used to help influence construction planning. On both urban and rural roads, traffic is at its lowest average volume during the winter. This means that the bulk of major construction work should be moved to the winter, when motorists will not be affected by heavy construction like they would be during the summer. We can also take into consideration the high traffic periods for both rural and urban roads to further influence when construction occurs. On rural roads, the highest traffic volumes occur in the summer, so it would be best to plan for projects on those kinds of roads to occur during the fall-spring season, with most of the work taking place in winter. On urban roads where the traffic is highest in the spring but dips around June, projects can be started in the summer when the traffic dips to low volume, then work can continue during the summer-winter season, with most of the work occurring during the winter months with lower traffic.

## 7.1 Traffic and Daily Life

One obvious application for knowing when high-traffic and low-traffic periods are is helping citizens know the best time to drive to work, recreation, meetings, etc. While certain time frames are unavoidable and the times with the overall least amount of traffic would not be used often in daily life (it would be inconvenient to drive at 2 AM all the time, for example), these times can still help people time their daily commutes. For example, if someone works at 7 AM and drives on an urban road to get to work, there is a large increase in traffic during the 6 AM to 7 AM hour compared to the 5 AM to 6 AM hour. If this person rolled back his commute so he left the house between 5 and 6 AM, he would be able to get to his workplace sooner and potentially have time to fit in personal actions such as taking a walk, catching up on extra work, or a morning coffee.

## 7.2 Road Construction

Knowing when traffic is at its high and low points would also be vital for road construction and general city planning. Since we have learned winter is the season where traffic is generally at its lowest point, we can advise construction companies to try and do the bulk of their work during the winter, weather permitting.

Doing the bulk of the work on a construction project while traffic is at its lowest point has multiple benefits. One such benefit is that traffic could be less affected if construction companies considered this information. If companies avoid building during the seasons and months where traffic is at its highest, they will minimize their impact on traffic and keep the flow of traffic moving. Working during low-traffic periods will also keep construction workers safe. Working around moving vehicles is inherently dangerous, and there are multiple ways a vehicle could accidentally hurt a construction worker, especially if the worker was working on or near a freeway or interstate at the time. By listening to our data, construction companies would ensure that their employees stayed out of the way of traffic, reducing the threat of on-the-job accidents and fatalities.

Our models for what time of day is busiest would also helpful for road construction. Our data shows that there are particular hours which are incredibly busy, but other hours which comparatively have much less traffic.

## 7.3 Predicting Traffic

This dataset's results could also help news stations predict what the traffic will be like in a given month and hour. Assuming these general trends can be proven over a period of years, it would be safe to assume that the trends would not change drastically. This would allow predictive traffic data to be more accurate, which could help newscasters determine which roads they should pay attention to for traffic updates or warnings. This would also help commuters plan their routes, because they would be able to know when traffic would be at its worst and prepare accordingly.

## 7.4 Future Research

Potential future research could involve the variables which were ignored in this particular study. Data such as whether direction traveled had any impact on the flow of traffic could be interesting, and depending on the findings, might impact city planning. For example, if it was found that traffic got worse when heading in an eastern direction, a city could add an alternate route which runs north or south for a while, which could reduce the volume of traffic on heavy-traffic roads.

# 8 CONCLUSIONS

We set out to find interesting information about our USA traffic dataset. We successfully made great visuals and spent lots of time planning and cleaning our data out so we could do analysis and come up with interesting results. Mainly using python we achieved this and found some results we expected such as high traffic volumes around summer and less during winter, and some interesting results such as the traffic volumes on rural vs urban roads varying by season. Applying clusters and outlier analysis was deemed not very useful for us as it didn't give us much insight that we could work with, while it was able to find outlier, there was no direct

information we could achieve from seeing these outliers other than a high traffic volume day, no clear sign of why. We have the potential for a regression model that can predict the future traffic volumes which can be used for city planning and construction planning that can be easily applied in any location. In conclusion traffic analysis is an important aspect as traffic is an ongoing issues around the globe, and data mining information about this problem is a useful research topic.

## 9 REFERENCES

https://www.kaggle.com/jboysen/us-traffic-2015/feed
https://www.nature.com/articles/srep37300
https://www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road
https://www.zmescience.com/research/technology/google-maps-traffic-05443/
http://kalw.org/post/driving-apps-waze-are-creating-new-traffic-problems