Traffic Incident Prevention Report

Jean Eyeghe Obame*
jeey7674@colorado.edu
Boulder, Colorado, USA

Damien Puhill
dapu5387@colorado.edu
Boulder, Colorado, USA

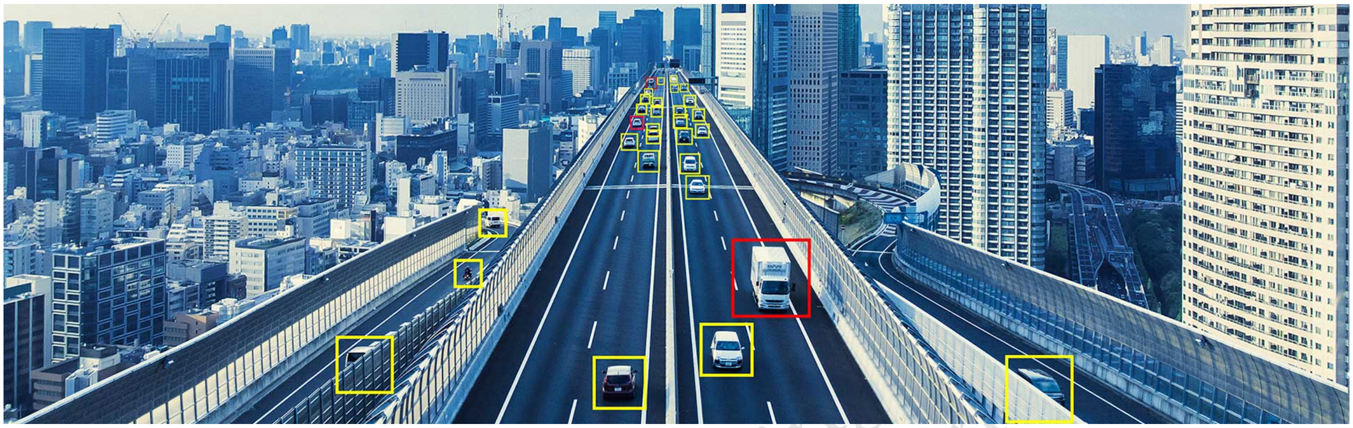


Figure 1: Traffic identification graphic from Blast Analytics

ABSTRACT

Over 76% of Americans drive to work everyday, while another 9% carpools with someone else (Adie Tomer) [1]. With a vast amount of people traveling each and every day, traffic has become a nightmare for many. In cities like Los Angeles, New York, Denver, traffic has become a large part of daily life. Although many see it as an annoyance, traffic congestion has become increasingly dangerous. With automobiles increasing by the millions each year, collisions have never been more prominent.

Austin, Texas has consistently ranked as one of the worst cities for traffic in the Western Hemisphere. Large amounts of these traffic congestions are due to bottlenecks caused by collisions. Within this analysis, I identified traffic collisions using a traffic database provided by the City of Austin and data mining techniques in order to find traffic patterns/hotspots. This is important because people travel by car more than ever and a single accident can cause major problems for both the driver's and those who will be delayed. The goal is to ultimately reduce accidents and traffic congestion by providing informed solutions to mitigate Austin's traffic issues.

KEYWORDS

datasets, data mining, traffic congestion study, accident prevention

*All authors contributed equally to this research.

ACM Reference Format:
Jean Eyeghe Obame and Damien Puhill. 2021. Traffic Incident Prevention Report. In . ACM, New York, NY, USA, 6 pages. https://doi.org/

1 INTRODUCTION

Traffic congestion usually results in bad judgments, some people may try to hurry and swerve into a lane not realizing there is a car that is in the blind spot of the driver or there may be a problem in the traffic control pushing vehicles to congest or more physical problems like potholes. All of these may happen to anyone and anywhere at any point of time.The collisions caused invoke road bottlenecks that will further delay commuters. With more people driving every year traffic is almost impossible to stop, however it can be mitigated with the informed solutions provided.

In this report, the goal is to utilize day to day traffic data from the City of Austin. By using this data, we can draw patterns and hot spots for collisions that create bottlenecks. This will then allow us to provide informed decisions on potential solutions for mitigating collisions and, by relationship, reduce traffic.

2 RELATED WORK

Data Mining related to traffic congestion has been studied in multiple ways in the past:

- A group from the Indian Institute of Technology Madras aimed to understand traffic data in an effort to supply drivers with information on the road. Using a data set which included location based sensors, which can measure volume and Time Mean Speed (TMS), Jitin Raj, Hareesh Bahuleyan, and Lelitha Vanajakshi created traffic density estimations/predictions that were then supplied to Advanced Traveller Information Systems to provide travelers with information that could be used to create an optimal route. These estimations were then

validated and strengthened through the use of more data. The team was able to conclude that the data driven techniques they had used were promising in predicting Indian traffic conditions. (Jithin Raj) [3]
- Shilpa Thaku from Lovely Professional University in Punjab studied traffic congestion patterns, using several data mining techniques including Naive Bayes, Fuzzy logic, and decision trees, to create an automated traffic management system. Using mass amounts of data from Indian traffic databases, Shilpa was able to create two algorithms that were able to decrease traffic congestion in 97.67% of proposed scenarios. She achieved this by creating automated systems that would lane shift (redistribute the amount of lanes going each way) according to bottleneck periods (found through data) when buildups would occur. (Shilpa Thakur) [4]
- Yuan Mei, Ting Hu and Li Chun Yang in their paper aimed to predict traffic congestion rather than solve it. These predictions can be immensely useful information to cities and governments to better understand when they will see congestion and account for it. Furthemore, like Jithin Raj's paper, these predictions can also be given to GPS navigation systems to help make ETA features more accurate. Using traffic data from Shenzen, the group created a Fuzzy Comprehensive Evaluation and Machine Learning algorithm that could "effectively analyze and predict the real-time traffic congestion. (Yuan Mei) [5]

The studies above were all conducted using detailed sets of data of traffic congestion abroad. These three papers, and most found online, are geared towards creating predictive models of traffic congestion and/or creating effective traffic management systems. Unlike the previous studies, we plan to look at traffic congestion within the United States with relation to incidents. Using patterns we find between the incidents and congestion, we hope to come up with methods that can be implemented to diminish incidents. Furthermore, we plan on using unique data sets that will provide different parameters than previous studies.

3 WORK STRUCTURE

Initially, I played around with the data set from Austin Texas in order to get familiar with it. Then I began to work towards building a predictive model, which was tested on different years.

The data set has 239k rows and 9 columns which is updated every 15 minutes with nominal and ordinal attributes. The oldest data ranges back to 2017 which was perfect since it allowed for more in depth analysis which leads finding patterns within the data using numerous testing methods. As we will see later in the report, I decided to stick with a K-means model.

After finding patterns/hotspots I doved into asking what could have caused these collisions. Could time (rush hour or weather) have effects on hotspots? If so, what solutions could we provide to mitigate collisions in these scenarios? After analyzing the causes/situations that invoke these collisions, I then went into researching and understanding solutions that could reduce collision counts.

As we will see in the conclusion, the analysis of the City of Austin provided various ways to effectively reduce collisions and traffic.

4 UML

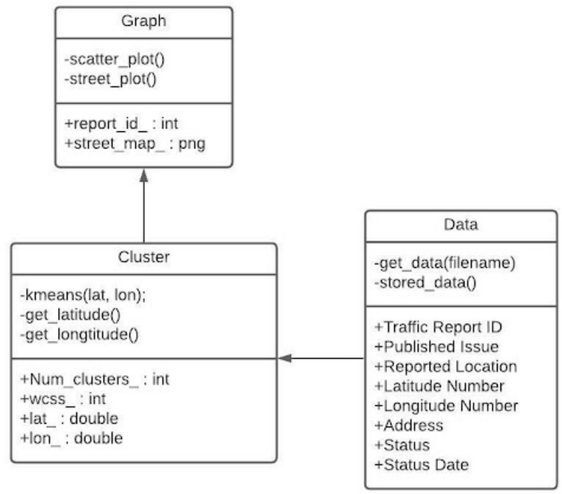
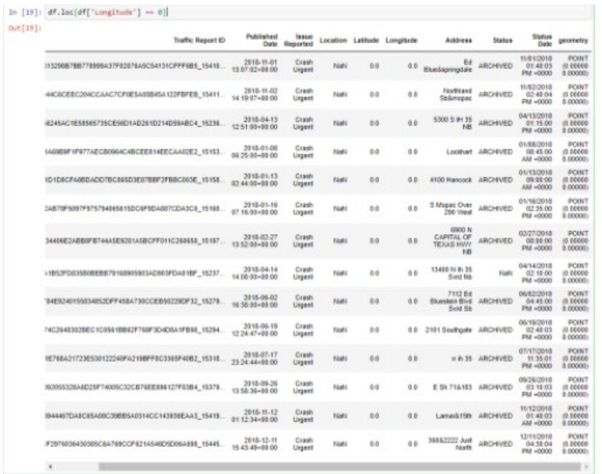


Figure 2: The UML. This is simple as we will be focusing on the collection of important data and storing it to visualize.

5 DATA

I decided to use the Real-Time Traffic Incident dataset provided by the official City of Austin open data portal. This dataset is in real time and updated every 15 minutes. The data is updated from the traffic incident information from the Austin-Travis County. Each entry in the dataset has the following attributes associated with it: Traffic Report ID, Date, Issue Reported, Location, Latitude, Longitude, Address, and Status. Each row in the data set is a traffic incident. The data starts from September 25, 2017 continues to the current date, accumulating for a total of 250k rows of traffic incidents.

For the purposes of this project I decided to filter the data and only include data labeled under "Crash Urgent" in the "Issue Reported" column. This is because I wanted to focus on incidents relating to human error, ignoring other unpredictable events such as "Loose Livestock" or "Vehicle Fire". However, when working with the data, I noticed several unusable fields, such as missing dates, issues, and locations (as seen below).



```
In [19]: df.loc[df['Longitude'] == 0]
Out[19]:
```

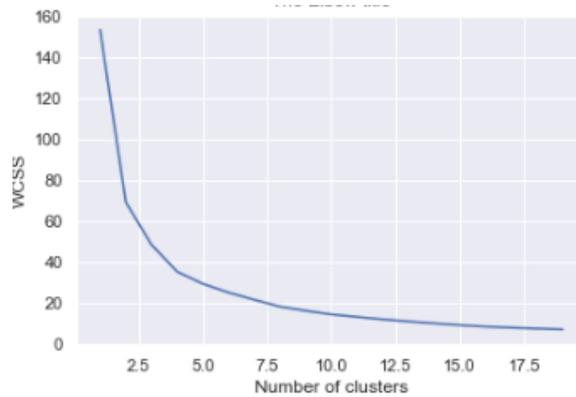
	Traffic Report ID	Published Date	Issue Reported	Location	Latitude	Longitude	Address	Status	Status Date	geometry
	[illegible]	2018-11-01 13:07:02+00:00	Crash Urgent	NaN	0.0	0.0	Ed Blue&springdale	ARCHIVED	11/01/2018 01:40:03 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-11-02 14:19:07+00:00	Crash Urgent	NaN	0.0	0.0	Northland Sb&mopac	ARCHIVED	11/02/2018 02:40:04 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-04-13 12:51:00+00:00	Crash Urgent	NaN	0.0	0.0	5300 S IH 35 NB	ARCHIVED	04/13/2018 01:15:00 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-01-08 06:25:00+00:00	Crash Urgent	NaN	0.0	0.0	Lockhart	ARCHIVED	01/08/2018 08:45:00 AM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-01-13 02:44:00+00:00	Crash Urgent	NaN	0.0	0.0	4100 Hancock	ARCHIVED	01/13/2018 09:00:00 AM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-01-16 07:16:00+00:00	Crash Urgent	NaN	0.0	0.0	S Mopac Over 290 West	ARCHIVED	01/16/2018 02:35:00 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-02-27 13:52:00+00:00	Crash Urgent	NaN	0.0	0.0	6900 N CAPITAL OF TEXAS HWY NB	ARCHIVED	02/27/2018 08:00:00 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-04-14 14:00:00+00:00	Crash Urgent	NaN	0.0	0.0	13400 N Ih 35 Svrd Nb	NaN	04/14/2018 02:10:00 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-06-02 16:38:00+00:00	Crash Urgent	NaN	0.0	0.0	7112 Ed Bluestein Blvd Svrd Sb	ARCHIVED	06/02/2018 04:45:00 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-06-19 12:24:47+00:00	Crash Urgent	NaN	0.0	0.0	2101 Southgate	ARCHIVED	06/19/2018 02:40:03 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-07-17 23:24:44+00:00	Crash Urgent	NaN	0.0	0.0	n ih 35	ARCHIVED	07/17/2018 11:35:01 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-09-26 13:50:36+00:00	Crash Urgent	NaN	0.0	0.0	E Sh 71&183	ARCHIVED	09/26/2018 03:10:03 PM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-11-12 01:12:34+00:00	Crash Urgent	NaN	0.0	0.0	Lamar&15th	ARCHIVED	11/12/2018 01:40:03 AM +0000	POINT (0.00000 0.00000)
	[illegible]	2018-12-11 15:43:49+00:00	Crash Urgent	NaN	0.0	0.0	360&2222 Just North	ARCHIVED	12/11/2018 04:38:04 PM +0000	POINT (0.00000 0.00000)

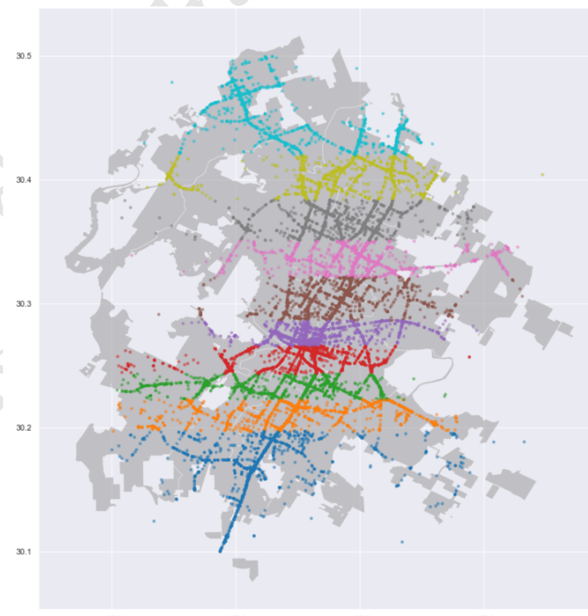
Thus, I continually cleaned the database by omitting missing fields. Eventually, reaching a point where only usable data was left.

6 EVALUATION

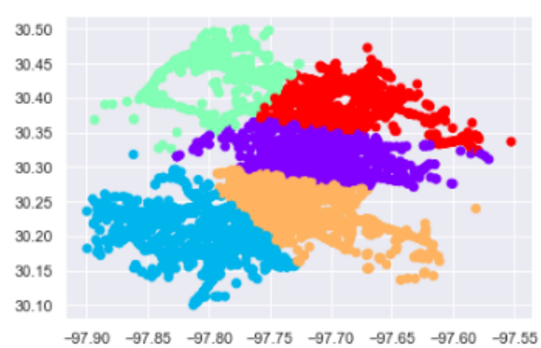
Once the data was clean I could start to see patterns. Initially I wanted to test the code and see if it would work with a smaller dataset of 293 rows. This was for just one week in early January. Using Geopandas I was able to map Austin and visualize where these accidents occurred as seen below.

This data was not as useful since it was so small but it gave me an idea as to what we should expect when we moved onto a larger dataset. Using K-Means was the obvious way to go as you can already see some clusters and it would allow us to recognize where most of the accidents occurred. Once I was comfortable analyzing the smaller dataset, I moved onto the larger dataset. In determining how many clusters there should be I used the Elbow method and saw that for each year (2018, 2019, 2020) that the most optimal clusters was 5.





You may notice the colors and think those are the clusters but it is not. We were having issues color coding the clusters so we just kept it default. The actual clusters can be seen below with latitude and longitude as the axis.



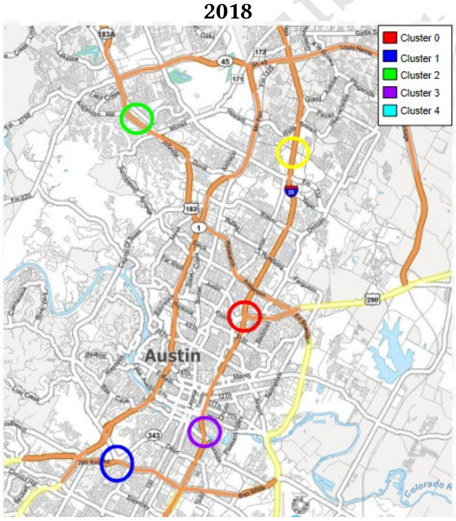
7 K-MEANS EVALUATION

As explained in the evaluation section, we chose to pursue a K-means approach. When analyzing the data, we first organized the data by year (2018, 2019, and 2020). We then ran the K-means algorithm to create five clusters for 2018, 2019, and 2020, as shown below:

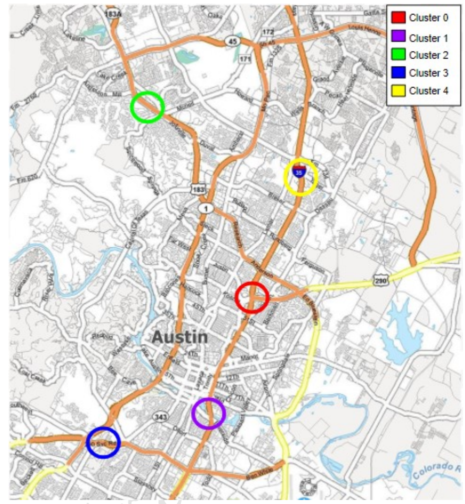
	2018	2019	2020
Cluster 0	Latitude = 30.319564702464753 Longitude = -97.71049424677231 Location = 6010 N Interstate Hwy 35, Austin, TX 78752	Latitude = 30.32105229031305 Longitude = -97.70128482575292 Location = 6225 U.S. 290 Frontage Rd, Austin, TX 78723	Latitude = 30.199924960712696 Longitude = -97.67167495344928 Location = Barton Hills, Austin, TX
Cluster 1	Latitude = 30.23029577540619 Longitude = -97.80314674482325 Location = 4514 West Gate Blvd, Austin, TX 78745	Latitude = 30.39450353917504 Longitude = -97.67611330412495 Location = Park 35, Austin, TX 78753	Latitude = 30.32388299423878 Longitude = -97.7054148526748 Location = 6121 N Interstate 35 Frontage Rd, Austin, TX 78752
Cluster 2	Latitude = 30.43848195540069 Longitude = -97.78933749547038 Location = 13376 Research Blvd #100, Austin, TX 78750	Latitude = 30.201173476603646 Longitude = -97.80013219026965 Location = 4514 West Gate Blvd, Austin, TX 78745	Latitude = 30.436454803118913 Longitude = -97.78869023879119 Location = 8805 Fairway Hill Dr, Austin, TX 78750
Cluster 3	Latitude = 30.247505295120387 Longitude = -97.72940765949289 Location = 1300 E Riverside Dr, Austin, TX 78741	Latitude = 30.43668376340484 Longitude = -97.79174852412851 Location = 8900 Jolly Hollow Dr, Austin, TX 78750	Latitude = 30.39246345076708 Longitude = -97.68422696499309 Location = 12049 Park 35 Cir, Austin, TX 78753
Cluster 4	Latitude = 30.39159148927477 Longitude = -97.67161212155242 Location = 12100 Park Thirty Five Cir, Austin, TX 78753	Latitude = 30.249270793185918 Longitude = -97.72874793405066 Location = 2101 Jesse E. Segovia St, Austin, TX 78702	Latitude = 30.244507721188363 Longitude = -97.72595198008473 Location = 1618 E Riverside Dr, Austin, TX 78741

The clusters above displays the five clusters of every year, which represent the five hotspots of every year. These hotspots were found by taking the averages of every cluster. When looking at the data, it is hard to tell how the hotspots varied between years. Thus, I decided to then graph each the five hot spots for 2018, 2019, and 2020:

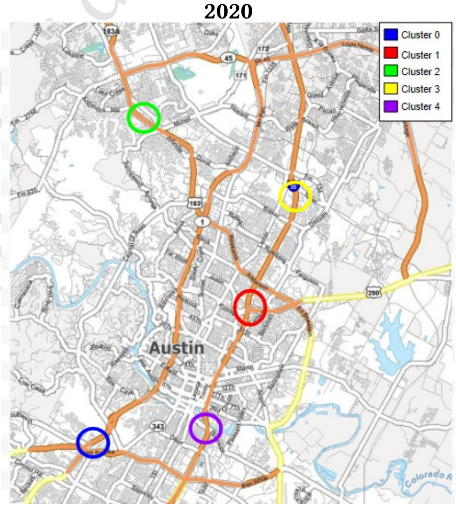
2018



2019



2020



As can be seen between the three maps of 2018, 2019, and 2020, the five hotspots were nearly the same every year. In fact, all five hopspots were within 0.5 miles of each other every year. Knowing this, we now changed focus to researching and coming up with valuable solutions. When doing this we found four key solutions:

(1) Improve infrastructure around five hotspots by adding more lanes. (Federal Highway Administration)
(2) Add extra police/ambulances at five hotspots to reduce response time and bottlenecks (Federal Highway Administration)
(3) Open extra lanes exiting Austin (I-35 specifically) and subtract from lanes entering the city during rush hour. Vice versa in the mornings (Federal Highway Administration)
(4) Reduce speed limits around hotspots to reduce collision induced bottlenecks (Federal Highway Administration)

8 CONCLUSION

As the K-Means Clustering showed, Austin has five trouble zones that have nearly been the exact same across multiple years. These areas, 3 of which land on I-35, reflect articles stating I-35 has is one of the United States' worst bottleneck problems (Knight, April 2021).

Knowing the issues, times, and areas of where most traffic accidents happen we are able to help maybe even nullify some problems ahead of time. Utilities is one important factor as most accidents seem to be set near the highway and as the article suggests, freight movement is important to see where and what level of investment should be made. Adding a few emergency lines to the highway may help those utility vehicles to arrive at the place of accident faster.

As the data suggests that if one part of the road is heavy in accidents than going back will only highlight more problems. Some solutions that might help change that is moving emergency vehicles closer to the clusters or at least lend higher priority. This will help alleviate traffic congestion faster as tow vehicles are closer to where most accidents happen. Another solution is to reduce speed limits around the clusters. Most of these issues are due to accidents that can be avoided if more time was given for drivers to react. Finally, opening lanes according to rush-hour can also help reduce lane merging so that vehicles won't be creating bottlenecks.

9 CHALLENGES FACED

The development of this model brought along various challenges. Distinguishing between which reported issues to target was not an obvious choice. There are seven different types of reported issues, some of these were traffic hazards, crash urgent situations, collisions, and stalled vehicles. I decided to set more attention towards the issues that were "Crash Urgent". Although some of the other issues such as stalled vehicles or traffic hazards were also a problem, the crash urgent situations were found to lead to heavier congestion. Crash urgent situations often involve law enforcement, ambulances, or some type of emergency responders to immediately come to the scene because these situations had the highest severity.

We found K-Means Clustering to be the most relevant technique for organizing the data. The issue with this was determining the number of clusters to appropriately model the data. Initially, I started with a random number of clusters then continuously increased and decreased that number in hopes of finding some improvements. It was very difficult to know if a particular number was optimal. By incorporating inertia and the elbow, finding the amount of clusters to use became much more apparent. Inertia, also referred to as within-clusters-sum-of-squares (WCCS) , measures the variability of data points within each centroid. This is done by calculating the sum of squares of the distances of all points in respect to their closest centroid. The goal is to have a low inertia and small number of clusters but not completely minimized because we would be losing a lot of the interpretability due to the tradeoff between number of clusters and inertia. Plotting WCCS vs. Number of clusters made it easier for us to determine the number that provided an optimal balance. We found that 5 was the number that met the threshold.

After properly clustering the data set I struggled with analyzing the causes of the congestion. Interpreting the results was a critical part of being able to construct a solution. Originally, I figured sorting the points by address and this showed which addresses encountered the most incidents. Although this showed the addresses with the most congestion, it was hard to get a sense of the general area because even if there were two addresses close to each other they would come as individuals within the grouping. It was also hard to get a sense of what caused certain addresses to get more congestion than others. I found that longitude and latitude coordinates would provide a better approximation of the data and better illustrate the causes. I took the average coordinates for each of the 5 clusters and linked the coordinates to a map and this allowed me to visualize the hot zones in more detail. With the precision of the coordinates it became much easier to pin-point each hot zone. Finding that many of the hot-zones came from highways, especially in the case where highways intersected.

10 FUTURE WORK

The results of my model indicates the congestion can best be broken down into 5 different zones within Austin, Texas. With this information the next step would be to prepare for running simulations to determine the best way to reduce these incidents. While implementing the simulations it is important to include different solution techniques, which can include adding more lanes to the hot zones, optimizing time increments of traffic lights to best enhance the flow of traffic, optimizing speed limits, etc. The only issue is this can come at an incredibly high cost due to the necessity of much more features than those included in the current data set. With a data set large enough to track many of the small-scale features (such as the length of a green/red light at a particular stop) we can simulate the effects of all potential techniques and determine those that work best.

In addition to working with a larger data set, breaking down the current data set and finding ways to split/sort the data by weekdays vs. weekends would undoubtedly give more insight and will allow for more correlations. Traffic activity varies greatly through the weekend and weekdays, and possibly even within the weekends or weekdays as well. The more detail I find will help provide more potential solutions to implement and minimize this issue. Another way to break down the data set involves splitting the data by dates where some major events (professional/collegiate sports, political events, concerts, festivals, etc) occurred, as I imagined there would be some rise in traffic levels. With the data set having the ability to update in real time, I have the luxury of continuously testing for improvements.

11 REFERENCES

[1] Tomer, Adie. "America's Commuting Choices: 5 Major Takeaways from 2016 Census Data." Brookings, Brookings, 9 Feb. 2018, https://www.brookings.edu/blog/the-avenue/2017/10/03/americans-commuting-choices-5-major-takeaways-from-2016-census-data/ #: :text=American%20commuters%20still%20largely%20depend%20on %20carstext=Over%2076%20percent%20of%20Americans,hitting %20American%20streets%20every%20day.

[2] "Real-Time Traffic Incident Reports: Open Data: City of Austin Texas." Data.austintexas.gov, 7 Oct. 2021, https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x.

[3] Jithin Raj, Hareesh Bahuleyan, Lelitha Devi Vanajakshi, Application of Data Mining Techniques for Traffic Density Estimation and Prediction,Transportation Research Procedia,Volume 17,2016, Pages 321-330, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2016.11.102.

[4] Thakur, Shilpa. DATA MINING BASED AUTOMATED SYSTEM FOR TRAFFIC CONGESTION MANAGEMENT USING TRAFFIC PATTERNS. May 2017, http://dspace.lpu.in:8080/jspui/bitstream/123456789/2550/1/11506615_5_6_2017%201_55_18%20PM_11506615complete.pdf.

[5] Mei, Yuan, et al. "Research on Short-Term Urban Traffic Congestion Based on Fuzzy Comprehensive Evaluation and Machine Learning." Data Mining and Big Data, 11 July 2020, pp. 95–107., https://doi.org/10.1007/978-981-15-7205-0_9.

[6] Federal Highway Administration. "Traffic Bottlenecks." Localized Bottleneck Reduction Program - Traffic Bottlenecks - FHWA Operations, https://ops.fhwa.dot.gov/bn/lbr.htm.

[7] Knight, Author: Drew. "I-35 In Austin Named One of Nation's Worst Bottlenecks." Kvue.com, 5 Apr. 2021, https://www.kvue.com/article/traffic/austin-traffic-i-35-worst-bottlenecks-us/269-b4fa54e3-3c64-43fc-ab13-ff197be87ca3.