Analyzing the Traffic Crash Data of Chicago

Anisha Islam, CCID: aislam4

Abstract

The consequences of traffic accidents and crashes can be fatal and financially devastating to people. Analyzing the traffic crash data can help law enforcement authorities to take preventive measures and raise public awareness. Previous studies aimed to gain insights from the crash data by using various spatiotemporal tools. However, very few focused on clustering the crash data based on various categorical attributes of an accident. This study aims to prevent and reduce traffic accidents by analyzing the traffic crash data of Chicago. We explore the research questions on the causes, locations, and times of traffic accidents using different clustering algorithms suited for numerical and categorical attributes. Based on K-means, we found that most accidents happen from the evening to midnight. DBSCAN and OPTICS algorithms showed that the peak hours for accidents are 6 pm on Fridays and from 12 am to 4 am on Wednesdays to Saturdays, respectively. We also investigated the locations of accidents by their types. We identified that the leading causes of traffic accidents in Chicago's central community area are speeding and running red lights. Along the US 41 highway, the accidents are caused by poor road conditions and a lack of road markings. We also noticed that in the Greater Grand Crossing and the Grand Boulevard area, the most frequent cause of traffic crashes is not stopping at the stop sign. Moreover, we examined that most accidents occur in areas without traffic control devices and on undivided, one-way, or divided roads with non-raised medians. Our study can provide valuable insights for the authority to improve road safety and lower the chances of accidents.

1 Introduction

Traffic accidents can occur when a motor vehicle collides with another car, with pedestrians, animals, or with stationary objects such as trees and poles [1]. According to a report from the World Health Organization, approximately 1.3 million people die each year because of traffic accidents [2]. Analyzing the traffic crash data can help law enforcement authorities to identify the reasons and potential locations for traffic accidents, develop infrastructures, and allocate resources accordingly to reduce the severity of accidents. Additionally, finding the times and places of traffic accidents in a particular area and identifying the major contributing factors to the accidents can help drivers exercise caution when driving through those high-risk areas.

Previous studies have tried to find hotspots for traffic accidents using Geographic Information System (GIS), spatiotemporal models, hotspot identification techniques, and different clustering algorithms [3, 4, 5, 6, 7, 8]. As traffic accidents depend on multiple factors such as location, time, weather conditions, driving style, and road defects, only analyzing the crash data spatially or temporally can not capture the pattern between different relevant factors related to the accidents. Clustering algorithms can analyze the multidimensional traffic crash data and extract patterns between different factors. For example, clustering the crash data based on time gives us helpful information on the peak hours for accidents. Similarly, clustering based on locations can help us identify the hotspots of traffic accidents. Furthermore, clustering based on other relevant auxiliary factors can help us extract valuable patterns that can be beneficial for promoting road safety among drivers.

This paper aims to analyze the traffic crash data of Chicago [9] and determine the effects of the contributing factors, such as weather conditions, time of day, road surface conditions, road defects, and location. Using different clustering algorithms, we extract relevant clusters and information to identify the common reasons and conditions for traffic accidents in Chicago. We also examine the patterns between different contributing factors of traffic accidents.

We investigate three research questions in this paper:

RQ1: When do the crashes happen more often?

RQ2: What are the most accident-prone areas of the city?

RQ3: Is there any pattern between different contributing factors?

We have selected two types of clustering algorithms to answer the research questions. The first type of algorithms works well with categorical attributes, such as K-modes [10], and DBSCAN [11] with a modified distance measure called Gower Dissimilarity [12]. The second type of clustering algorithms performs well for numerical attributes like latitudes, longitudes, and crash hours. Examples of such algorithms are K-means [13], Hierarchical Clustering [14], DBSCAN [11], and OPTICS [15]. We compare the performances of these algorithms for each research question and validate the clusters using Silhouette Co-efficent [16].

Our results show that, according to the K-means algorithm, most accidents happen from evening to midnight. According to DBSCAN and OPTICS, the most common accident times are 6 pm Friday and Wednesday to Saturday from 12 am to 4 am, respectively. We also discover significant differences in accident locations based on the accident types. Additionally, we discover that traffic accidents usually happen in areas with no traffic control devices and with undivided, one-way, and divided traffic ways with medians that are not raised.

2 Related Works

Previous studies have tried to find the hotspots of traffic accidents in different areas using various spatial, temporal, spatiotemporal, GIS, and clustering methods. Zubaidi *et al.* [6] tried to explore the relationship between crash types at intersections, injury severity, and roundabout configurations. They used GIS on the traffic crash data of Oregon to identify the high-risk accident areas, the factors responsible for the accidents, and the countermeasures needed to be performed. They found out that the factors responsible for accidents at the intersections are not yielding right of way, improper change of traffic lanes, and following too closely. Khan *et al.* [5] analyzed the single-vehicle lane departure crashes in North Dakota using Global Moran's I, local Moran's I, and network kernel density estimation (NetKDE). However, they only focused on crashes related to single-vehicle lane departures and did not utilize any clustering algorithms.

Islam *et al.* [3] examined the clustering performances of different clustering algo-

rithms like K-means, Mini batch k-means, OPTICS, DBSCAN on only the location columns of a crash dataset and did not consider the additional categorical contributing factors like weather and road condition. According to their study, DBSCAN and OPTICS outperformed K-means and Mini batch k-means regarding the clustering performance. Shariff *et al.* [4] examined the traffic crash data of Malaysia using spatial analysis techniques such as Neighborhood Hierarchical (NNH) Clustering and Spatial-Temporal Clustering using tools like CrimeStat and ArcGIS and reported that the Spatial-Temporal Clustering performs better than the Neighborhood Hierarchical clustering.

In contrast, our study aims to apply different clustering algorithms for categorical and numerical features on the traffic crash data of Chicago and tries to answer research questions related to the accidents' time, the accidents' locations, and the pattern between different contributing factors of the crashes.

3 Methodology

This section briefly discusses the methodology used for the research questions under consideration.

3.1 RQ1: Times of crashes

In the first research question, we tried to find the typical times of traffic crashes. For this purpose, we only chose the hour and the day of week information from our dataset and clustered the data using K-means, DBSCAN, and OPTICS.

3.2 RQ2: Locations of crashes

In the second research question, we tried to find out the locations of the traffic crashes. For this purpose, we only chose the latitude and longitude information from our dataset and clustered the data using K-means, DBSCAN, and Hierarchical clustering.

In addition, we considered subsets of the accident data with six different accident types to get a more precise insight into the crash locations. We tried to cluster the locations of those particular accident types using DBSCAN as DBSCAN can separate

the cluster points from noise points, unlike K-means and Hierarchical clustering. We selected the accident types that can be considered specific to certain regions rather than those that can occur at any location. The accident types that we considered are: FAILING TO REDUCE SPEED TO AVOID CRASH, DISREGARDING TRAFFIC SIGNALS, DISREGARDING STOP SIGN, VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.), EXCEEDING AUTHORIZED SPEED LIMIT, and ROAD ENGINEERING/SURFACE/MARKING DEFECTS.

3.3 RQ3: Pattern between different factors

For the third research question, we considered the categorical factors in our traffic crash dataset to find patterns between these relevant factors. We used two clustering algorithms for this purpose: K-modes and DBSCAN with Gower Dissimilarity.

K-modes: K-modes does not rely on distance measures to cluster data points but rather on dissimilarities, which are the number of mismatches between categorical values. The lower the dissimilarity, the more similar the data points are. For this reason, the K-modes algorithm can better cluster the categorical attributes of a dataset.

DBSCAN with Gower Dissimilarity: To measure how similar two data points are, the Gower similarity measure uses a partial similarity function that returns 1 if both data points have the same value for a particular categorical feature and 0 otherwise. This way, Gower similarity only considers exact matches and ignores any differences in the values of the categorical feature. Gower dissimilarity can be defined as 1 - Gower similarity. Since this measure does not depend on Euclidean distance, we can use this as a distance measure in our DBSCAN algorithm to generate clusters based on categorical attributes.

We consider eight categorical factors from our dataset, such as:

1. TRAFFIC_CONTROL_DEVICE
2. DEVICE_CONDITION
3. WEATHER_CONDITION
4. LIGHTING_CONDITION

5. TRAFFICWAY_TYPE
6. ROADWAY_SURFACE_COND
7. DAMAGE
8. PRIM_CONTRIBUTORY_CAUSE

We run the clustering algorithms mentioned above to generate clusters and get insights based on these categorical attributes.

3.4 Cluster Validation

We use the Silhouette Co-efficient for cluster validation, which gives a value between -1 to +1. According to this method, -1 means incorrect clustering, 0 means clusters with overlaps, and +1 represents correct distinct clusters.

4 Experiments & Results

In this section, we introduce the dataset used in our study, explain the preprocessing steps, answer our research questions, and derive patterns between different contributing factors of traffic crashes.

4.1 Dataset

For our study, we collected the traffic crash data of Chicago [9] from the Chicago data portal [17]. There were 49 columns in the dataset, and the total number of accident data in this dataset was 696,510. Each column represented one feature related to traffic accidents.

4.2 Preprocessing

Before answering our research questions, we preprocess our dataset in the following ways:

1. **Remove unnecessary columns:** As there are 49 columns in our dataset, we decided to drop some columns irrelevant to the research questions that we

considered and which can not be considered a possible cause of an accident. In this way, we reduced the dimensionality of our dataset. We removed 31 such columns. Some examples of the dropped columns are PHOTOS_TAKEN_I, STATEMENTS_TAKEN_I, and REPORT_TYPE.

2. **Remove missing and misleading latitudes and longitude:** Some data points had missing location information in our dataset. We removed those data points from our dataset. Also, there were some outlier points where the latitudes and longitudes were noted as 0. Chicago is not situated in the 0 latitude and 0 longitude regions, so we removed those misleading data points from our dataset.

3. **Filling missing values:** We identified the missing values in the columns of our dataset and filled the missing values using "-1" which denotes an unknown value.

4.3 Research Questions

4.3.1 RQ1: Time of the Crashes

Our first research question aims to find the most common times of accidents during the week according to different clustering algorithms. For this purpose, we only consider the crash hour and the day of the week information for each traffic accident data in our dataset. The clustering algorithms we chose for this scenario are K-means, DBSCAN, and OPTICS. The reason for choosing these algorithms is that the crash hour and day of the week can be considered sequential data.

K-means: We used the elbow method to find the most suitable value for the K-means algorithm. Figure 1 shows the change in the WCSS value with the varying number of clusters. Here, the WCSS value represents the "Sum of squared distances of samples to their closest cluster center, weighted by the sample weights if provided" [18]. We select the point after which the change in WCSS becomes minimal. For us, this value is at k = 5. The silhouette score in this scenario is 0.378, which represents clusters with overlaps. We run our K-means algorithm for five clusters. Figure 2 illustrates the clusters created by the K-means algorithm.

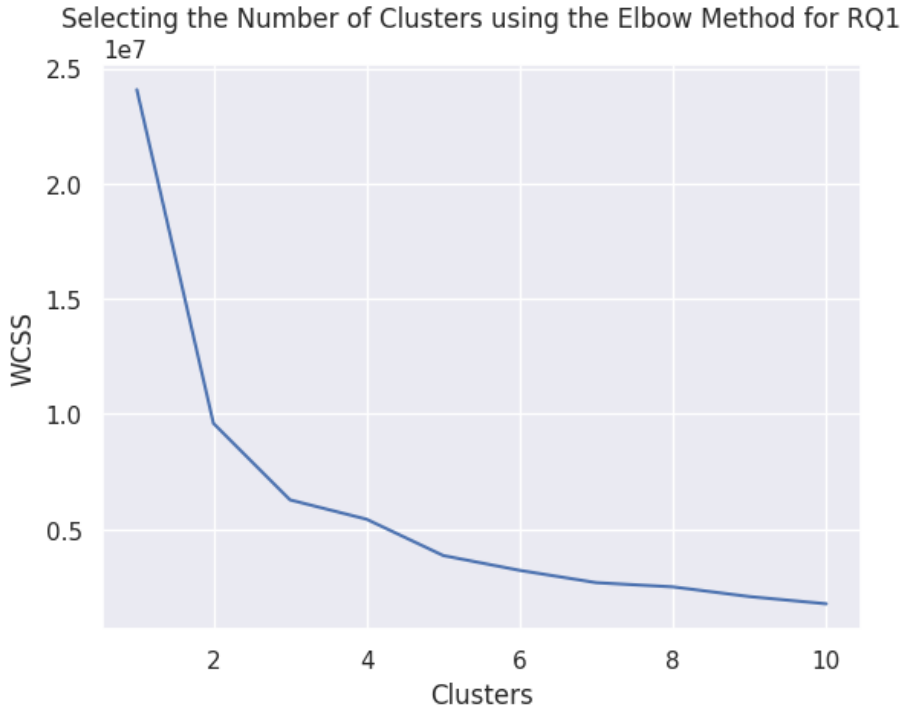

Figure 1: Selecting an optimal k-value for K-means

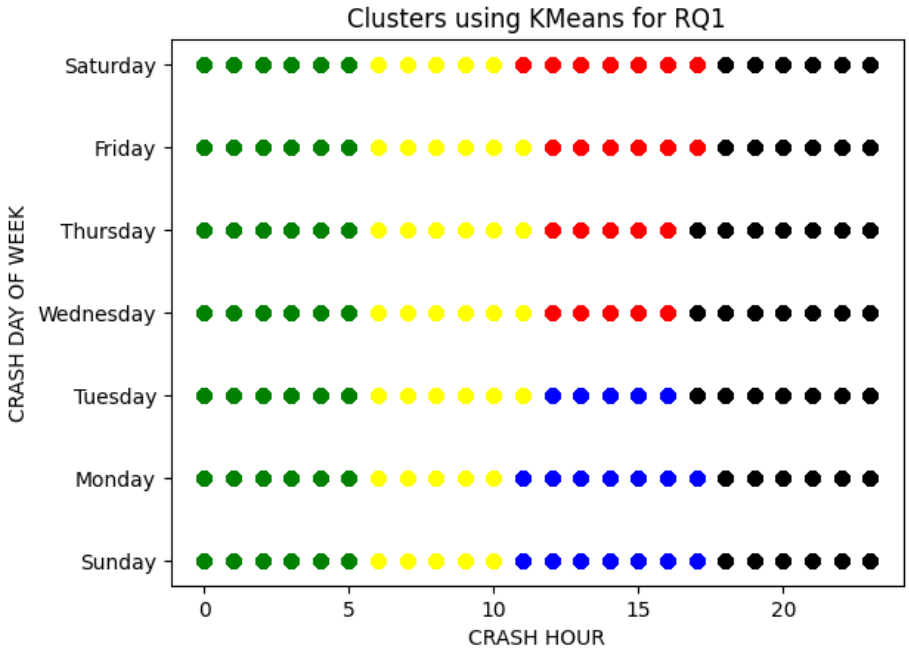

Figure 2: Clusters generated by K-means considering the time of the crashes

We can see that, using K-means, the more prominent clusters are formed during the following times.

1. Sunday to Saturday from midnight to 5 am (Before dawn)

2. Sunday to Saturday from 5 am to 11 am (Morning)
3. Sunday to Tuesday from 11 am to 7 pm (Office Hours + after office hours)
4. Wednesday to Saturday from 12 pm to 4 pm (Late office hours)
5. Sunday to Saturday from 6 pm to midnight (Evening)

Considering the frequency of the clusters generated by the K-means algorithm, we can say that **most crashes happen from evening to midnight.**

DBSCAN: We randomly selected 1000 data points from our dataset and chose epsilon = 0.005 and min_samples = 15. In this setting, DBSCAN generated 6 clusters with 894 noise points. The silhouette score for the generated clusters for DBSCAN is 1.0, representing correct distinct clusters. **According to DBSCAN, the most frequent cluster was formed on Friday at 6 pm.** Figure 3 illustrates the clusters generated by DBSCAN.

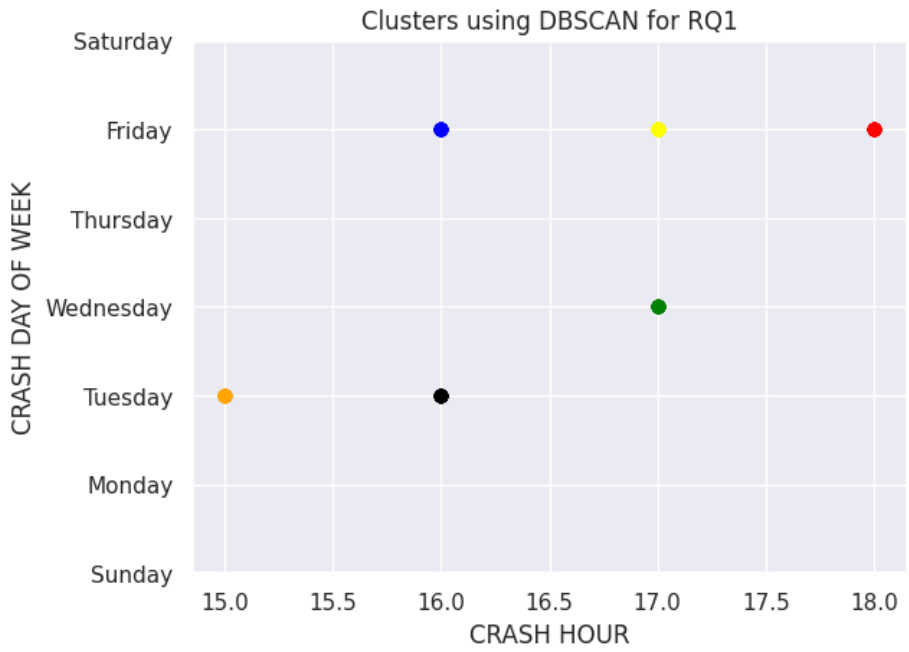

Figure 3: Clusters generated by DBSCAN considering the time of the crashes

OPTICS: We randomly selected 1000 data points from our dataset and chose epsilon = 0.005 and min_samples = 15 for running the OPTICS algorithm. In this setting, OPTICS generated 8 clusters with 811 noise points. The silhouette score for the generated clusters for OPTICS is 0.772, representing distinct, well-formed clusters. **According to OPTICS, the most frequent cluster was formed on Wednesday to Saturday from 12 am to 4 am.** Figure 4 illustrates the clusters

generated by OPTICS.

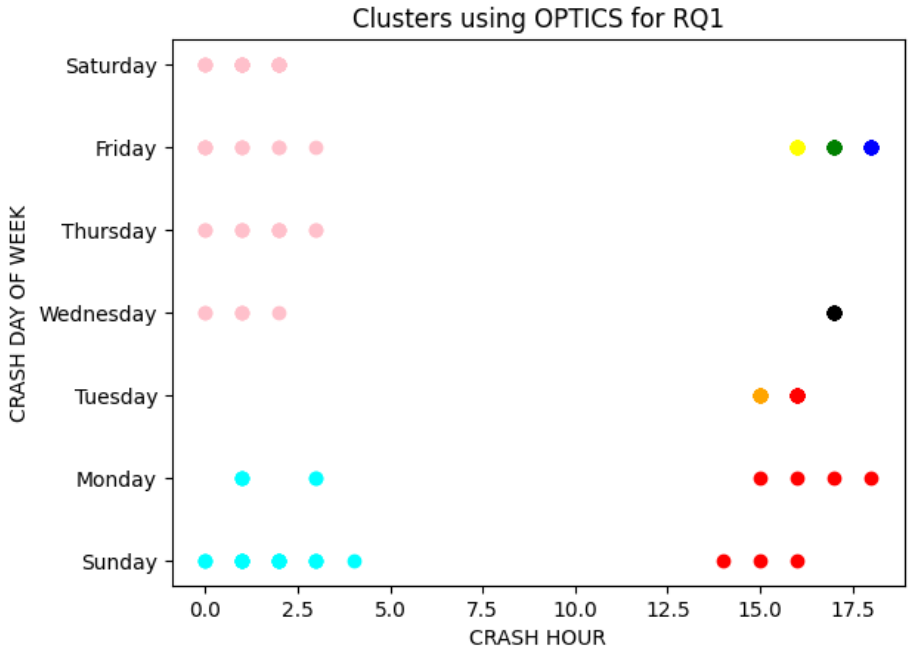

Figure 4: Clusters generated by OPTICS considering the time of the crashes

4.3.2 RQ2: Location of the Crashes

In this research question, we aim to find the most accident-prone areas of Chicago City. For this reason, we considered only the latitude and longitude information of the crashes and used three clustering algorithms: K-means, DBSCAN, and Hierarchical Clustering.

At first, we randomly selected 10,000 crash data for running the clustering algorithms. We applied Hierarchical Clustering with 7 clusters on the sampled data and visualized the output. Figure 5 represents the clusters generated by the Hierarchical clustering algorithm. Similarly, we ran DBSCAN and K-means (number of clusters = 7) on the sampled data and visualized the clusters. Figure 6 and Figure 7 represent these two cases, respectively.

We used Sihouette Co-efficient to validate the clusters generated by these algorithms. For Hierarchical clustering, the score is 0.38, which represents clustering with overlaps. For DBSCAN and K-means, the values were 0.46 and 0.41, respectively, representing distinct clusters with slight overlaps.

We can see that K-means and Hierarchical clustering consider the entire data while

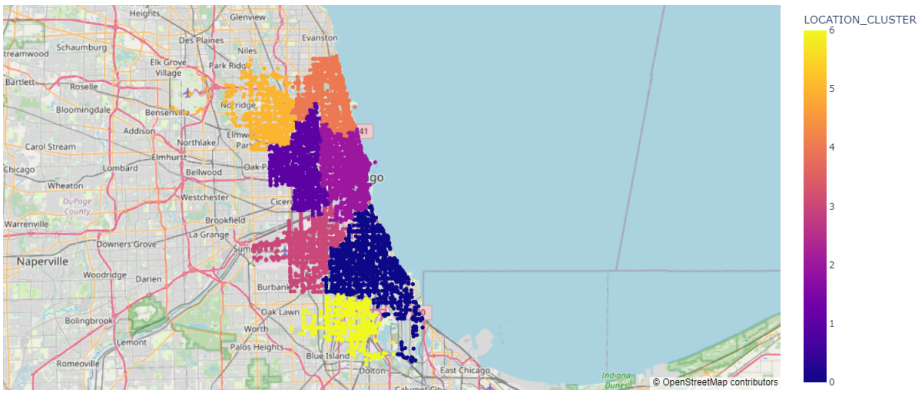

Figure 5: Clusters generated by Hierarchical Clustering considering the locations of the crashes

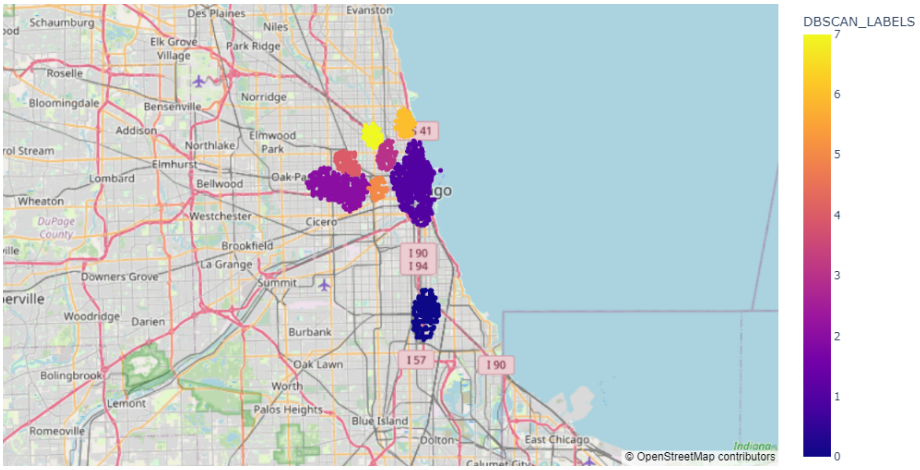

Figure 6: Clusters generated by DBSCAN considering the locations of the crashes

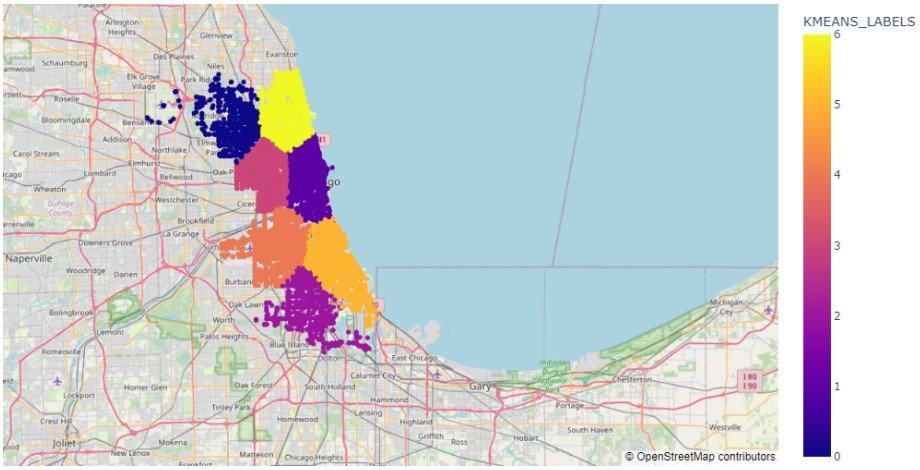

Figure 7: Clusters generated by K-means considering the locations of the crashes

clustering and can not differentiate between cluster points and noise points. DBSCAN, on the other hand, considered 6,884 points amongst the 10,000 points as noise points (epsilon = 0.01 and min_samples = 100) and generated 8 clusters.

However, we considered all types of accidents while creating the sample data. Now, we consider only specific types of accidents for further clustering to get a more precise insight into the accident locations. There are 40 different primary causes of accidents in our dataset. For this part of the research question, we only considered six accident types which are mentioned in Section 3.2.

We took a subset of data from our dataset based on the primary cause of the accidents mentioned in Section 3.2 and clustered the locations of the accidents using DBSCAN since DBSCAN can generate distinct clusters and separate noise points. Figure 8 to Figure 13 represent the locations of accidents caused by the different causes under consideration.

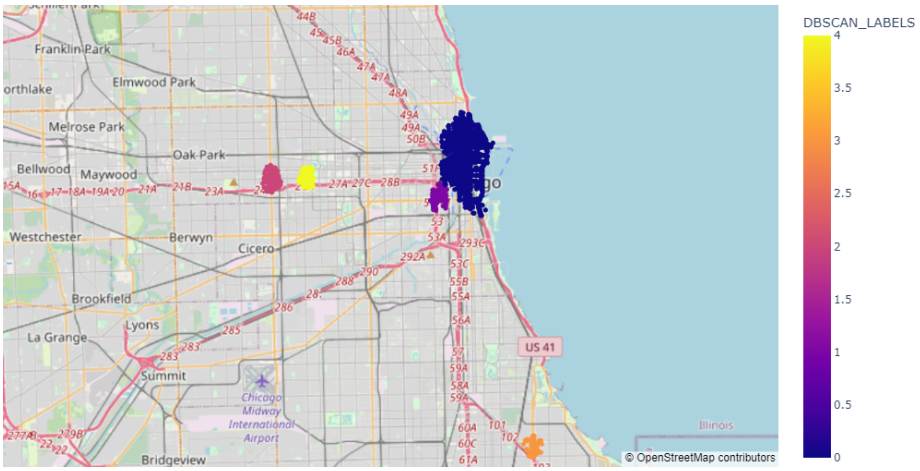

Figure 8: Primary Cause of Accident: FAILING TO REDUCE SPEED TO AVOID CRASH

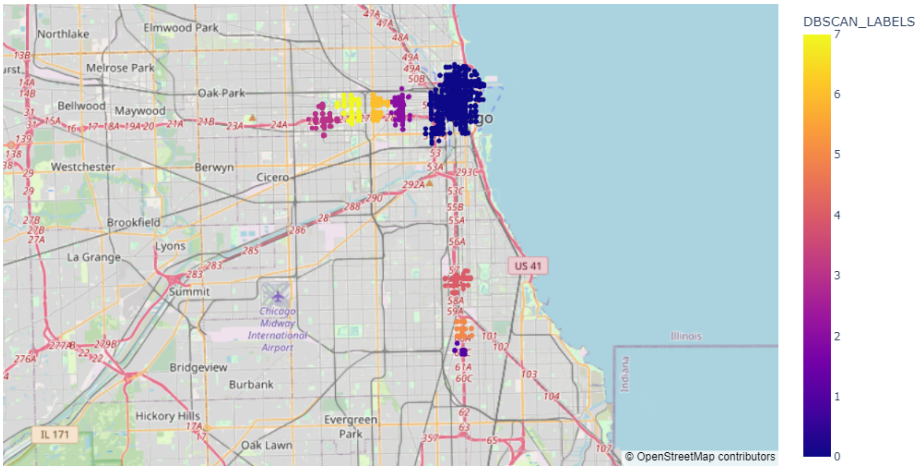

Figure 9: Primary Cause of Accident: DISREGARDING TRAFFIC SIGNALS

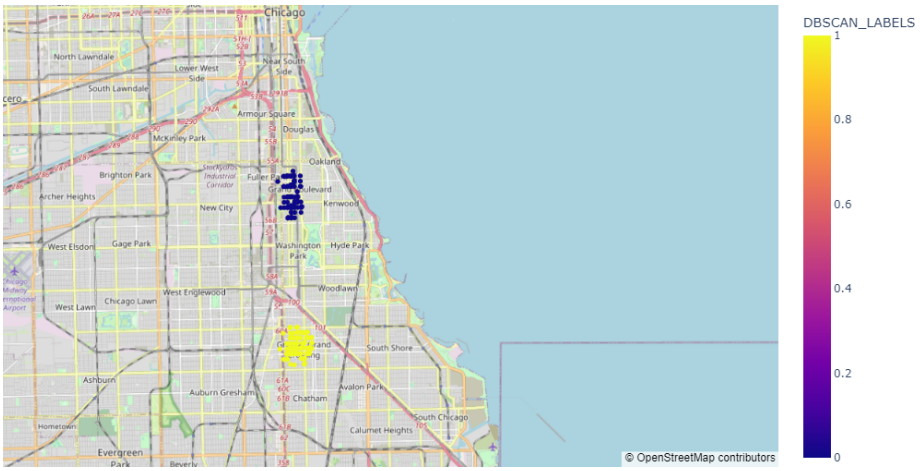

Figure 10: Primary Cause of Accident: DISREGARDING STOP SIGN

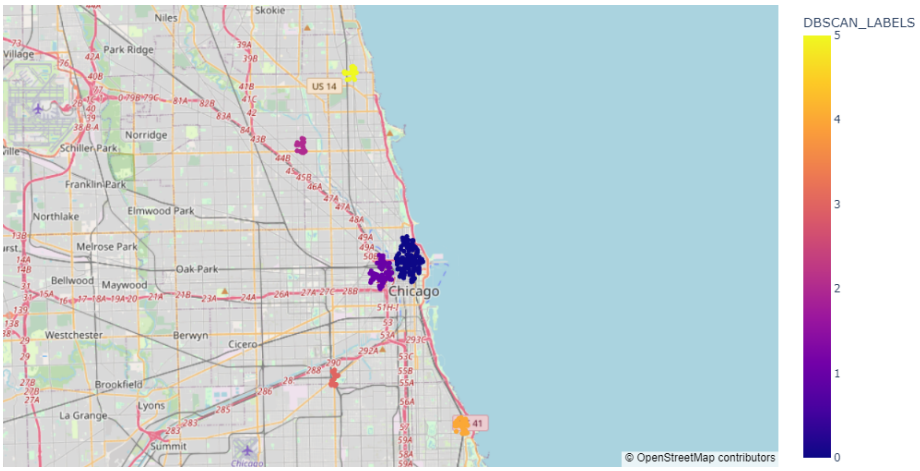

Figure 11: Primary Cause of Accident: VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)

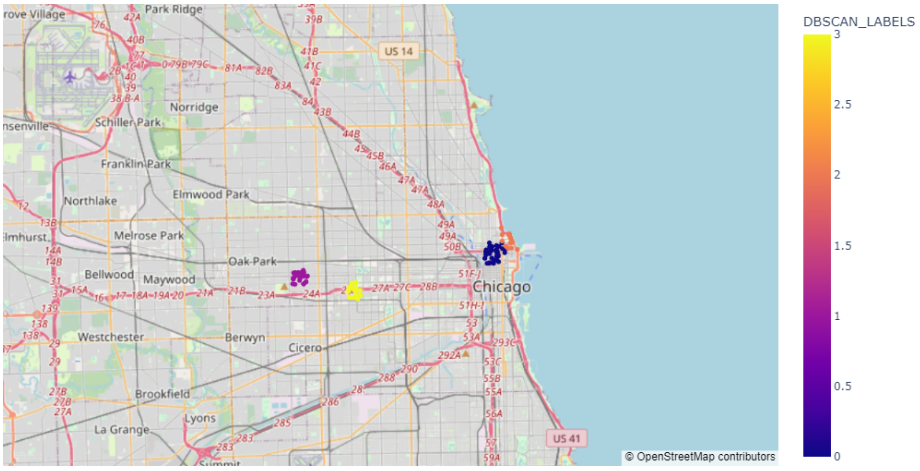

Figure 12: Primary Cause of Accident: EXCEEDING AUTHORIZED SPEED LIMIT

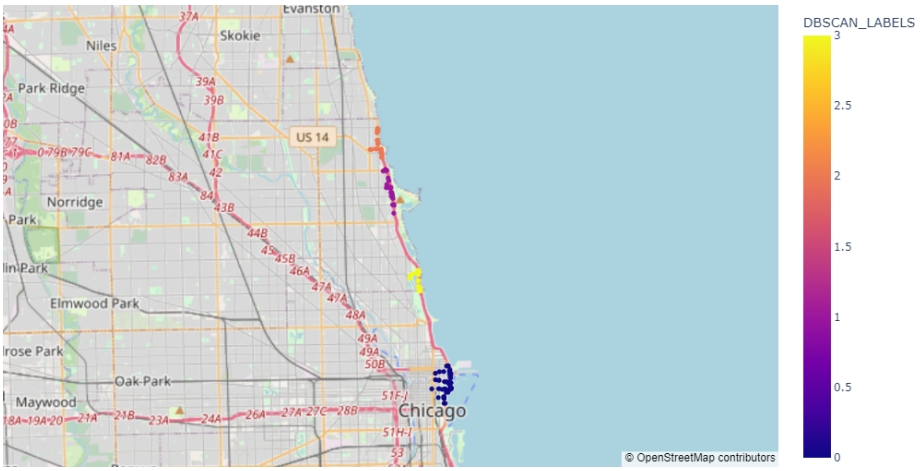

Figure 13: Primary Cause of Accident: ROAD ENGINEERING/SURFACE/MARKING DEFECTS

We can see that the clusters generated considering the different primary causes of the accidents are situated in distinct regions of Chicago. The locations for the particular types of accidents are:

- **FAILING TO REDUCE SPEED TO AVOID CRASH:** West Jackson Boulevard, West Madison Street, West Garfield Park, University of Illinois Chicago, Central side of Chicago
- **DISREGARDING TRAFFIC SIGNALS:** Eisenhower Expressway, Central area, Dan Ryan Expressway
- **DISREGARDING STOP SIGN:** Greater Grand Crossing, Grand Boulevard
- **VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.):** Hyde Park, McKinley Park, Near North side, Around Near West Side, North Kedzie Avenue, West Devon Avenue, Near Roger's Park
- **EXCEEDING AUTHORIZED SPEED LIMIT:** Northwestern University Chicago campus, West Madison Street, Eisenhower Expressway
- **ROAD ENGINEERING/SURFACE/MARKING DEFECTS:** Highway US 41

From the generated clusters, we can understand that, in the central community area of Chicago, the primary causes of accidents are FAILING TO REDUCE SPEED TO AVOID CRASH, and DISREGARDING TRAFFIC SIGNALS. More surveillance should be deployed in these areas to detect the traffic violations mentioned above. Additionally, along the US 41 highway, the cause of the accidents is ROAD ENGI-

NEERING/SURFACE/MARKING DEFECTS. There should be more signs on this issue along the highway so that the drivers can exercise more caution. Furthermore, we can also see that in the Greater Grand Crossing and the Grand Boulevard area, the primary cause of traffic crashes is DISREGARDING STOP SIGN.

In this way, by considering subsets of accident type, we can get a more precise insight into the locations of the traffic crashes. The authority can also take necessary steps to reduce the severity of accidents in these regions according to the common accident types.

4.3.3 RQ3: Pattern between Different Contributing Factors

When we only consider the time and location of the accidents for clustering, we can not get any insight into the additional relevant factors that also contribute to the causes of accidents. For this reason, in this research question, we only considered the categorical contributing factors in our dataset and tried to find patterns between them. As we considered primarily the categorical factors in this research question, we selected the K-modes algorithm and DBSCAN with Gower dissimilarity while clustering the categorical information. We used the categorical features mentioned in Section 3.3.

DBSCAN with Gower Dissimilarity: We used the DBSCAN algorithm with Gower Dissimilarity as a distance metric to deal with the categorical factors of an accident. We randomly selected 10,000 data points from our dataset, and DBSCAN generated eight clusters in this scenario (epsilon = 0.005, min_samples = 70) with 9016 noise points.

Analyzing the clusters further gave us insights into the patterns between different contributing factors. Some examples of such patterns are explained below.

1. Traffic way type: Not divided, Traffic control device: No control, Damage: Over $1500
2. Traffic way type: One way, Traffic control device: No control, Damage: Over $1500
3. Traffic way type: Not divided, Traffic control device: No control, Damage: Over $1500, Lighting condition: Darkness, lighted road

4. Traffic way type: Parking lot, Traffic control device: No control, Damage: Over $1500
5. Traffic way type: Divided w/median (not raised), Traffic control device: No control, Damage: Over $1500

We can see that DBSCAN mainly generated clusters based on the traffic way type and the traffic control device. In all cases, when there are no traffic control devices present in a traffic way, accidents occur in those regions, and the financial damage for those accidents is over 1500$. Traffic accidents usually happen in undivided, one-way, and divided traffic ways with medians that are not raised. The pattern generated from DBSCAN makes sense because the clusters represent severe accidents where no traffic control devices are present and roads where the two sides are not divided correctly or at all. The authority can take proper steps to install traffic control devices or cautionary boards in these regions.

K-modes: Using the K-modes algorithm, we also generated clusters using our categorical factors. We generated 8 clusters using K-modes, and the cluster centroids gave us helpful insight into the cluster types generated by K-modes.

1. Traffic control device: Traffic signal, Weather condition: Rain, Road surface condition: Wet, Damage: Over $1500
2. Traffic way type: Not divided, Traffic control device: No control, Damage: Over $1500, Lighting Condition: Daylight or Dawn
3. Traffic way type: Parking lot, Traffic control device: No control, Damage: Over $1500
4. Traffic way type: Divided w/median (not raised), Traffic control device: No control, Damage: Over $1500

We can see that both K-modes and DBSCAN with Gower dissimilarity generate similar types of clusters. One additional insight that we can get from the K-modes clustering is that when the weather condition is rainy and the road surface condition is wet, there is a higher possibility of accidents even when traffic control devices are present. Drivers should be more careful on rainy days as there is a higher chance of accidents in this weather condition.

5 Threats to Validity

Some issues might question the validity of our study. The issues are mentioned below.

5.1 Using Silhouette Score for Density-Based Clustering

Silhouette score does not usually work well with density-based clusterings like DBSCAN and OPTICS. We tried to validate the clusters generated in the first two research questions using the Silhouette score, which might not give correct cluster validation results for the density-based clusters.

5.2 Sample of Data

Due to resource limitations, we took a sample of data for DBSCAN, OPTICS, and Hierarchical clustering. Considering the entire dataset would have given us a more accurate clustering.

5.3 Considering the Time Columns as Numerical Data

The crash hours and days of the week are represented by 0 to 23 and 1 to 7, respectively. There is a sequence between consecutive days and hours, so considering time columns as categorical data was not an option. However, keeping the data as it is meant that there is a higher distance between hour 0 and hour 23. Similarly, the distance between day seven and day one is considered a high distance according to Euclidean distance, whereas in reality, these values represent consecutive hours and days.

5.4 Limited Options for Categorical Clustering

There are only a few available clustering methods for categorical clustering. For this reason, we had to limit our study to K-modes and DBSCAN with a Gower distance measure for the categorical attributes.

6 Conclusion

Traffic accidents are a major cause of death and injury worldwide. To prevent and reduce traffic accidents, the authority needs to know the patterns in traffic accidents, such as the causes, locations, and times. By analyzing these patterns, the authority can identify the risk factors and design effective interventions, such as improving the road infrastructure, enforcing traffic laws, raising public awareness, and providing emergency services. In this study, we analyze the traffic crash data of Chicago and answer research questions on the time, location, and patterns between the causes of traffic accidents.

We applied different clustering algorithms to analyze the patterns of traffic accidents. According to K-means, most accidents occur from evening to midnight. We also used DBSCAN and OPTICS algorithms, which revealed that the peak times for accidents are 6 pm on Fridays and from 12 am to 4 am on Wednesdays to Saturdays, respectively.

Furthermore, we examined the spatial distribution of accidents by their types. We discovered that the main factors contributing to traffic accidents in Chicago's central community area are driving too fast to prevent a collision and ignoring traffic lights. These areas need more monitoring to catch these traffic offenses. Also, along the US 41 highway, the accidents are caused by poor road designs and lack of marking on road surfaces. There should be more warning signs along the highway to alert drivers to be more careful. Moreover, we observed that in the Greater Grand Crossing and the Grand Boulevard area, the most common cause of traffic crashes is not stopping at the stop sign.

Additionally, we discovered that most accidents happen in areas without traffic control devices and on undivided, one-way, or divided roads with non-raised medians by applying clustering algorithms suited to categorical attributes.

The findings of our study can help the authority improve road safety and reduce the risk of accidents.

References

[1] Wikipedia, "Traffic collision." https://en.wikipedia.org/wiki/Traffic_collision#External_links. Accessed: 2023-04-17.

[2] World Health Organization, "Road traffic injuries." https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed: 2023-04-17.

[3] M. R. Islam, I. J. Jenny, M. Nayon, M. R. Islam, M. Amiruzzaman, and M. Abdullah-Al-Wadud, "Clustering algorithms to analyze the road traffic crashes," in *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, pp. 1–6, IEEE, 2021.

[4] S. R. Shariff, H. A. Maad, N. N. A. Halim, and Z. Derasit, "Determining hotspots of road accidents using spatial analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 1, pp. 146–151, 2018.

[5] I. U. Khan, K. Vachal, S. Ebrahimi, and S. S. Wadhwa, "Hotspot analysis of single-vehicle lane departure crashes in North Dakota," *IATSS Research*, 2022.

[6] H. Zubaidi, I. Obaid, H. Mohammed, S. Das, and N. S. Al-Bdairi, "Hot spot analysis of the crash locations at the roundabouts through the application of gis," in *Journal of Physics: Conference Series*, vol. 1895, p. 012032, IOP Publishing, 2021.

[7] A. Soltani and S. Askari, "Exploring spatial autocorrelation of traffic crashes based on severity," *Injury*, vol. 48, no. 3, pp. 637–647, 2017.

[8] C. Liu and A. Sharma, "Exploring spatio-temporal effects in traffic crash trend analysis," *Analytic Methods in Accident Research*, vol. 16, pp. 104–116, 2017.

[9] City of Chicago, "Traffic Crashes - Crashes." https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if. Accessed: 2023-04-17.

[10] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, vol. 96, pp. 226–231, 1996.

[12] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[13] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications," *biometrics*, vol. 21, pp. 768–769, 1965.

[14] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[15] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM Sigmod record*, vol. 28, no. 2, pp. 49–60, 1999.

[16] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[17] City of Chicago, "Chicago Data Portal." `https://data.cityofchicago.org`. Accessed: 2023-04-17.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.