Traffic Accident Clustering Methods

Jessica Wishart
jeswis@email.unc.edu

April 8, 2025

1 Introduction

Traffic accidents are a major public safety issue in America. Often, traffic accident data is analyzed through predictive models from supervised learning methods such as random forest or regression. However, these models typically only focus on a few features at a time (such as demographic data and alcohol use) and analysis is lead from prior assumptions on correlations of certain variables and accidents. Such analysis also tends to focus on broad trends, and how individual features contribute to traffic accidents and their severity. Traffic accident data is prone to noise, with many different pertinent features and high dimensionality and requires careful exploratory analysis.

This research explores the use of unsupervised learning techniques, particularly clustering techniques, to uncover hidden patterns in behavior, time or season, driver demographics, and environmental conditions that tend to result in accidents of various severity. The goal is to build the best model for identifying such patterns by comparing the performance of K-mean clustering, DBSCAN, and Gaussian Mixture Models. This research should be able to answer the question of if there exists patterns of accident type or severity at certain roads under certain conditions, or from demographically identified drivers at certain times of day, dates, or seasons.

Clustering techniques, such as K-means, DBSCAN, and Gaussian Mixture Models (GMM), have been widely used in customer segmentation and behavioral analysis, as such analysis is often conducted for the purpose of advertising and preparing speculative store inventory. In particular, this approach is based on methods typically used for customer behavioral analysis because such analysis can be translated to informing targeted public safety ads and emergency response preparation at hospitals, or use by government agencies in identifying particularly risky roads and in what ways their safety may be improved. The ability to predict and identify new accident patterns could allow targeted interventions that would save lives and inform public policy. In addition, the types of hyperspecific patterns likely to be uncovered by clustering methods could be useful to inform the investigation and building of predictive models of traffic accidents for use within a local scope, which could even allow the deployment of preemptive

emergency responses. As such deployment for traffic accidents can be expected to occur at hazardous times and conditions, this research could also assist the process to be more safe and efficient, such as through planning specific routes for emergency vehicles ahead of time.

2 Related Works

Many works that examine the occurences of traffic accidents do so based on the driver's features; alcohol, age, sex, and other demographic variables. McCarty and Kim (2024) did just that, connecting demographic factors to road accident risk through hexagonal gridding and a variety of regression models [1]. This avenue of research is largely motivated by informing "resource allocations of EMS and other accident-related services to insurance premiums," [1].

Likewise, the investigation into causes and patterns of traffic accidents in other works largely focused on regression models, such as logistic or forest regression. Wang and Zhang (2017), for instance, examined environmental and roadway factors and their influence on traffic crash severity through the use of logistic regression. They found that the probability of severe or fatal injury in a crash increased in order of the following factors: "urban expressway, urban interstate, urban minor arterial, urban major arterial, rural interstate, rural minor arterial, and rural major arterial," [2]. The presumption of this research is that a similar trend will be revealed through clustering methods, with rural areas being associated with higher severity. However, what Wang and Zhang did not uncover was the intercorrelation between very specific factors; such as whether urban roadways were safer than rural ones *except* at a certain time of day or specific season and at a certain type of intersection, whereupon the danger skyrockets. Such hidden patterns and neat coinciding of specific feature values can be more easily discovered by unsupervised learning methods such as clustering techniques rather than supervised learning methods. In essence, unsupervised learning techniques will be used in this research to discover new, unexpected patterns in the data, which can then be isolated and further investigated through supervised learning techniques in future works.

In particular, this research was inspired by the application of clustering techniques in market analysis, such as exemplified by Kashwan, who performed customer segmentation using clustering techniques [3]. In the same manner as customers were divided into groups based on their preferences in Kashman's research, traffic accident occurences will be clustered based on circumstantial and environmental factors. Such clustering can then be used to predict underlying dangers that could cause traffic accidents in specific areas, which may only become apparent when coinciding with other circumstantial features, such as weather.

3 Methods

As the goal of this research is to analyze environmental circumstances that contribute to crashes, this project examined Crash Report Sampling System data collected by NHTSA in the year 2022 [4]. Of this data, features decribing crash severity, number of involved vehicles (moving or parked), number of involved pedestrians, region, weekday, time of day, first impact of accident, manner of collision, type of intersection, car's relation to road, if accident occured in poximity to workzone, lighting conditions, weather conditions, urbanicity, alcohol involvement, schoolbus involvement, and highway involvement were selected for clustering analysis. Note that ChatGPT assistance was used in preparing the code [6]. See Figure 1.

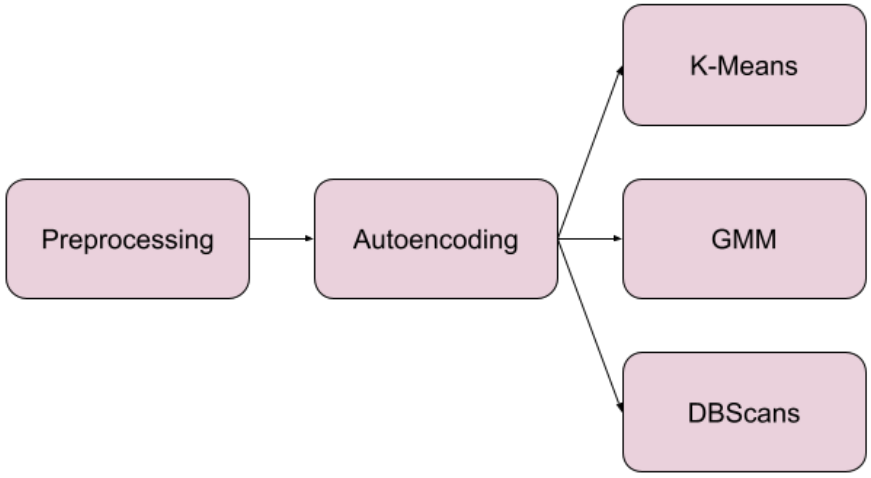

Figure 1: Workflow

3.1 Preprocessing

In the CRSS data, when a value is missing or unknown (such as "not reported"), it was replaced with a single imputed value. This is different from other techniques where a range of plausible values might be substituted. The imputed value is a single estimate of what the value likely should be. Of note, 34.9 percent of the 'alcohol involved in crash' variable was unkown/not reported and thus benefitted from imputing. Imputed data was used for clustering techniques, as such techniques require complete data [5]. Of the dataset, particular features were selected in an attempt to avoid collinearity. The number of vehicles in the crash who were moving was counted separately from the number of vehicles parked or that were work vehicles, such as those involved in construction. A deeper explanation of each variable involved in the code can be found in the

CRSS manual.

Data was cleaned for unknown values and features were converted to binary when appropriate. Categorical data was one hot encoded in preparation for clustering. Continuous features were scaled with StandardScaler(), as the results were preferred over MinMaxScaler(), and for the features describing the number if involved vehicles or pedestrians, outliers above 3 were capped in order to prevent skew to data that does not reflect the underlying logic of accident consequence. Thus, these features are considered as a scale from none, one, a couple, to 'many'. Time of day was split into intervals of 3. Rows with unknown values were dropped. Sequential order for the sequences of specific events in a crash was ignored in order to instead examine patterns of event co-occurence. 20 percent of the dataset was sampled in order to reduce processing time, as DBScan does not scale with as much processing efficiency as K-means Clustering. This resulted in approximately 17,900 rows. To finish the preprocessing, the sparse matrix was converted to a dense numpy array.

The selection of features went through trial and error, and presented a major problem for clustering analysis, as the dataset contained mixed feature types; continous, categorical, and binary. Clustering was attempted disregarding all continous features and treating max severity as a categorical variable, but the transformation of max severity to a continous variable and the inclusion of other continous features presented improved clustering results based off of metrics such as the Silhouette Score and Davies-Bouldin Index.

3.2 Autoencoding

The data was autoencoded to reduce dimensionality and then scaled again in order to prepare for clustering. Due to the process of autoencoding and scaling, the datapoints became very densely packed, which influenced how distance-based algorithmn analyzed the data. Parameters were finetuned, and ideal number of latent dimensions was selected based off the lowest mean reconstruction error. For the AutoEncoder model, two dense layers were used for encoding. The Dense layers utilized ReLU activation, and dimensionality was reduced to 47 latent dimensions. In addition, the model used two dense layers for decoding with sigmoid activation.

Other dimensionality methods were attempted, but Autoencoding performed the best. Of the other methods explored, PCA, and truncated SVD were attempted but eliminated based on poor results.

3.3 K-Means Clustering

After the data was preprocessed, K-Means Clustering commensed. Ideal paramaters for K-Means Clustering were visually determined through the Elbow Method, and then narrowed down by best resulting silhouette score and Davies-Bouldin Index Score. See Figure 2. The utilization of k-means++ algorithm was experimented with, but the original K-Means algorithm performed sufficiently, and there were no significant improvements when k-means++ was attempted.

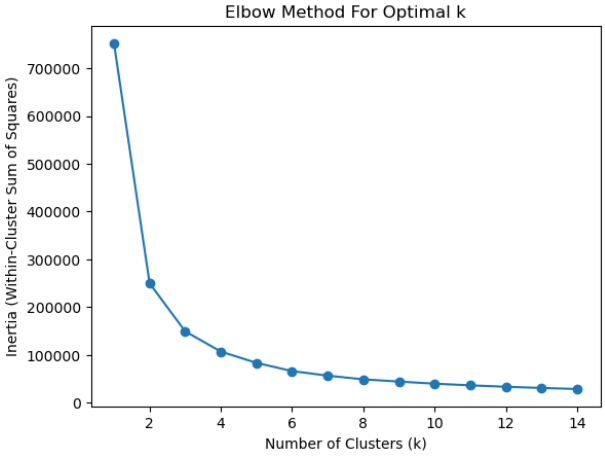

Figure 2: No. of clusters by inertia for choosing K-Means parameters

3.4 Gaussian Mixture Model

Ideal hyperparameters for the key factors of n_components, covariance type, and the choice of initialization parameters to be determined by kmeans or random were isolated through experimentation. Ideal parameters were chosen based off of the best results from a comparison of the clustering metrics Silhouette Scores and DBI Scores. This experimentation revealed the optimal configuration for the Gaussian Mixture Model on this dataset was diagonal covariance combined with kmeans selection with three components. It is likely that diagonal covariance performed best because of the process of autoencoding and scaling taken earlier, which resulted in a densely packed data cloud.

3.5 DBScan

Different parameters were tested to find ideal parameters resulting Silhouette Score and DBI Score. The ideal epsilon value was visually determined through a K-distance graph. See Figure 3.

However, crosschecking the results revealed high noise, and the ideal Silhouette Score were very close to 1 and 0, respectively, which raised concerns of overfitting. Increasing the parameters to obtain metric evaluation scores comparable to the results of K-Means and GMM resulted in a more meaningful distribution of clusters. The difference is visualized below with t-SNE dimensional reduction, however note that this is an imperfect 2D representation of higher dimensional data, and that the method of dimensional reduction performed before clustering was Autoencoding. See Figure 4, 5, and 6.

This overfitting based off of ideal Silhouette Score and DBI Score was likely due to the density of the data.

4 Experiments/Results

Table 1: Comparing Metrics

Method	Silhouette Score	Davies-Bouldin Index (DBI)	Clusters
K-Means	0.5789	0.5320	3
GMM	0.5691	0.5435	3
DBSCAN (eps=0.03)	0.7816	0.2908	9
DBSCAN (eps=0.07)	0.5192	0.4181	11

Based off these results, each method of clustering analysis seems to be a competitive candidate. Though K-Means performed with the highest Silhouette Score, GMM performed better in DBI. Likewise, DBScan managed to obtain higher metrics, but a closer look at the clustering result showed a high level of noise in the -1 cluster. It's worth noting that each of these clustering methods use different algorithms to obtain their result, and as such the evaluation metrics used to compare these methods are not perfectly comparable. Each model makes different assumptions on the underlying behavior of the data, such as how K-Means assumes spherical clusters of equal variances, even though the true clusters can be elliptical or overlapping [6]. This research began with the attempt to find the ideal clustering method for this analysis because of these differences in the models, however the results above imply each model performs competitively with each other, and can each make unique insights.

Though these results showed that clustering was fairly distinct

4.1 K-Means Clustering Results Analysis

This section briefly analyzes the results of the K-means Clustering analysis for the purpose of demonstrating the benefits of performing clustering algorithms on traffic accident data. See Figure 6.

K-Means obtained the best Silhouette Score and a competitive DBI Score when compared to GMM or the adjusted DBSCAN.

Feature	Cluster 0	Cluster 1	Cluster 2
REGION	S (60%)	S 3 (41%)	S (51%)
Weekday	Peak on weekdays 5, 6	Monday (19%)	Monday (16%)
Hours	Afternoon/Evening	Evening	Peak 12–19h
EVENT1	Mostly Event 12 (95%)	Event 8 (17%), 12 (4.5%)	Event 12 (31%)
Collision	High in types 1, 6	Mostly type 0 (96%)	Mostly type 0 (68%), some 6 (12%)
Interchange Type	Type 1 (42%), 2 (41%)	Mostly type 1 (77%)	Type 1 (81%)
Road	Mostly 1 (95%)	Diverse: 1, 4, 7	4 (46%), 1 (42%)
Urban	83% Urban	69% Urban	64% Urban
Alcohol	96% no alcohol	20% alcohol-involved	12% alcohol presence
Severity	58% minor severity	28% fatal severity	Moderate severity mix
Moving vehicles	Type 2 (75%)	Type 1 (94%)	Type 1 (65%), 2 (30%)
Pedestrians	100% no pedestrians	31% with pedestrian	0.4% pedestrians
Weather	Clear weather (90%)	Clear (74%), some adverse (rain/snow)	Mostly clear (83%)
Highway	Low highway involvement (12%)	High (47%) on highway	Moderate (27%) on highway
Interchange Related	Low relation (5%)	22% related to interchange	12% related to interchange

Based off these results and observation of patterns, some suggestions could be made to public safety officials. For instance, Cluster 1 observes high alcohol risk, fatal severity, high highway involvement, and relation to rainy/snowy conditions that obscure visibility. Based off of this pattern, it seems that these features may be correlated, which bears further investigation with regression analysis. Otherwise, a deeper dive into geospatial data with clustering analysis could reveal specific highways or other roads and intersections that could benefit from changes to their infrastructure in order to lower risk for drivers. This can be done through salting roads and improving weather alerts and emergency responder resources. Another pattern is revealed by comparing the hours of greatest accident activity and pedestrian risk, and by comparing alcohol use to the manner of collision, as clusters with high alcohol involvement seem to show higher non-vehicle collisions.

5 Conclusion

In conclusion, the results of this exploratory analysis of traffic accident data with clustering methods has successfully raised questions on hidden underlying patterns in the dataset. Unique connections were made between environmental and situational factors such as weather or time of day or week, road types, and the manner of accident collisions. A closer look at the items within SOE, Sequence of Events, also connects certain unique events such as Rollovers to these different clusters on conditions. These findings suggest that certain conditions may be

more likely to contribute to specific kinds of accidents on certain road-types, and the clustering results can serve as a launching point for real-world efforts in risk mitigation. In addition, the results of this analysis can be used to inspire further regression analysis based off this data in order to isolate meaningful correlations, or to inform and direct data gathering efforts for the purpose of more specific clustering analysis based off of precise local geospatial data.

6 Appendix

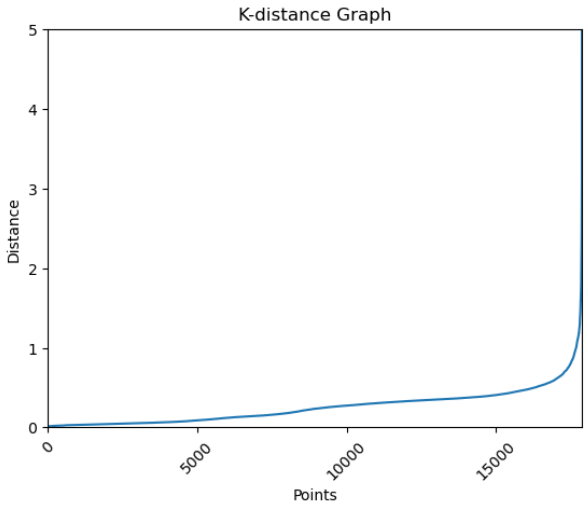

Figure 3: K-distance Graph for choosing DBScan epsilon parameter

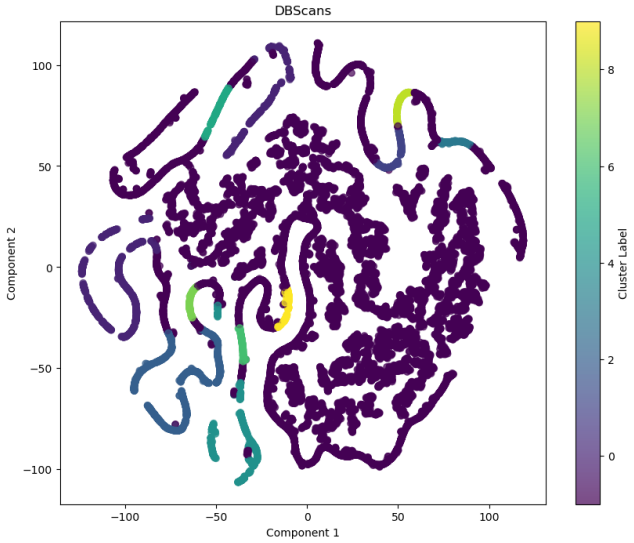

Figure 4: DBScans with ideal parameters

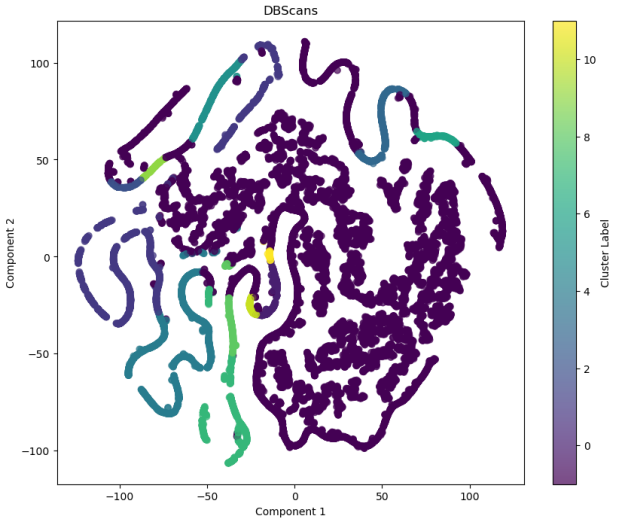

Figure 5: DBScans with adjusted parameters

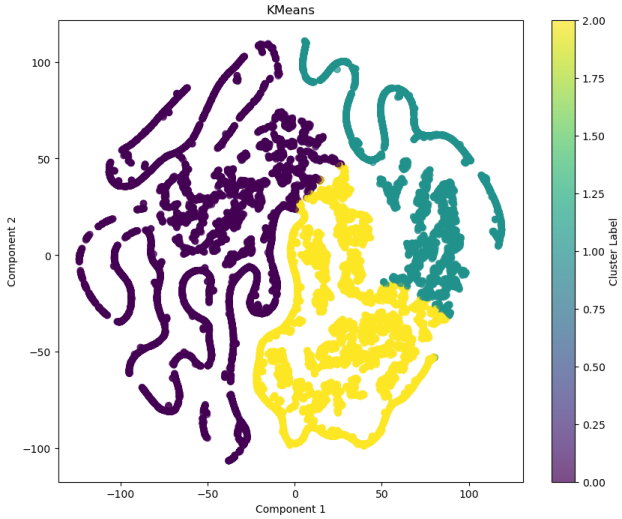

Figure 6: DBScans with ideal parameters

References

[1] McCarty, D., & Kim, H. W. (2024). Risky behaviors and road safety: An exploration of age and gender influences on road accident rates. *PloS one, 19*(1), e0296663. https://doi.org/10.1371/journal.pone.0296663

[2] Wang, Y., & Zhang, W. (2017). Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities. *Transportation Research Procedia, 25*, 21192125. https://doi.org/10.1016/j.trpro.2017.05.407

[3] Kashwan, K. R., & Velu, C. M. (2013). Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering, 5*(6), 509513.

[4] NHTSA, USDOT. (2022). *Crash Report Sampling System.* https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system

[5] National Center for Statistics and Analysis. (2025, March, Revised). *Crash Report Sampling System analytical users manual, 2016-2022* (Report No. DOT HS 813 557). National Highway Traffic Safety Administration.

[6] OpenAI. (2025). ChatGPT (April 8 version) [Large language model]. OpenAI. https://chatgpt.com/share/67f5e4e8-9544-8001-b9ff-aacfc6179632