HIGH-INCIDENT CRASH DATA

A PREPRINT

Jamie Tran
University of Virginia
Charlottesville, VA 22903
pby5br@virginia.edu

Brandon Yuan
University of Virginia
Charlottesville, VA 22903
shw3ht@virginia.edu

Aaron Yuan
University of Virginia
Charlottesville, VA 22903
tta5rc@virginia.edu

March 7, 2025

ABSTRACT

Traffic accidents are a major concern for public safety, leading to significant financial costs and loss of life. Understanding patterns and contributing factors in high-incident crash areas can aid in targeted interventions. This project applies the KMeans clustering algorithm to traffic crash data in Virginia to identify spatial clusters of incidents based on geographic proximity and related features. While DBSCAN was initially explored for its ability to detect arbitrarily shaped clusters, KMeans ultimately provided more coherent and interpretable results. Additionally, a correlation analysis investigates potential contributing factors such as light conditions, roadway surface conditions, and relation to the roadway. The findings aim to provide insights for traffic management and preventive measures.

1 Introduction

In today's rapidly urbanizing world, effective transportation management is becoming increasingly critical. As the population expands and the cities become larger, traffic congestion (especially in densely populated areas) intensifies, resulting in delays, increased air pollution, and increased costs of goods due to supply chain disruption. More alarmingly, a poor traffic management system contributes to a surge in roadway accidents– leading to not only significant losses in property damage but most tragically, the loss of many lives.

Addressing these challenges requires a data-driven approach to identify and mitigate factors in potentially dangerous situations that contribute to accidents in high-risk areas. It is essential to recognize spatial patterns in traffic incidents and contextualize them.

Previous research explored spatial clustering techniques to analyze traffic incidents. For example, Cheng and Washington [2008] applied K-means clustering to identify accident hotspots, while Yamada and Thill [2011] used kernel density estimation for similar purposes.

This project will apply clustering techniques, (building upon the aforementioned studies) to highlight areas with high incident rates. Specifically we will utilize both K-means and DBSCAN algorithms to look for and analyze spatial patterns in our crash data. While K-means is widely used for cluster identification, DBSCAN offers a unique advantage for this type of analysis: it does not require a specified number of clusters and can identify ambiguous shapes which is well suited for the uneven density distributions of our data.

Additionally beyond finding these spatial clusters, we will conduct a correlation analysis of the variables that may contribute to traffic incidents. We will primarily focus on environmental and roadway features such as lighting conditions, surface type, and the road-user relationship (e.g., intersection or not). By combining these two methods, we

hope to generate actionable insights to guide traffic safety improvements and enhance infrastructure planning, ultimately contributing to a safer and more efficient transportation system.

The dataset used in this study is publicly available from the Virginia Department of Transportation (VDOT) and can be accessed at:

`https://www.virginiaroads.org/datasets/VDOT::crashdata-basic-1/explore`

2 Method

So far, we have experimented primarily with clustering techniques to identify high-incident areas. Our initial approach involved DBSCAN clustering, where we used KneeLocator (from the kneed package) to find the optimal epsilon value. When this method failed to estimate a clear threshold, we resorted to a simple heuristic, identifying the point in the k-distance graph where the distances began to increase sharply.

The idea behind using DBSCAN is because of its ability to detect arbitrarily shaped clusters and filter out noise, which is ideal for geographic crash data. Since we are interested in grouping areas with a high density of crashes and understanding the contributing geographic and environmental factors, density-based clustering seemed like a natural fit.

After our initial DBSCAN experiments, we expanded our methods to include KMeans clustering, using the elbow method to determine the optimal number of clusters. This allowed us to compare how the choice of the clustering technique affects the quality of the groupings.

3 Experiments

Our initial focus was on feature engineering, particularly on selecting and encoding categorical variables. Since the dataset contains many categorical attributes (e.g., road types, lighting conditions, surface conditions), we prioritized dimensionality reduction to minimize noise and make the clustering more effective.

We selected the most relevant columns and used **Ordinal Encoding** for features with an inherent order (such as *Crash Severity*) and **One-Hot Encoding** for nominal features. This preprocessing step allowed us to create cleaner input for clustering algorithms.

3.1 DBSCAN Clustering

As a baseline, we ran **DBSCAN** on a mixed dataset containing both categorical and numerical features without any finetuning. This resulted in a large number of clusters with limited interpretability. We then limited the input to only spatial coordinates (longitude and latitude), and finetuned the hyperparameters using `KneeLocator` to find the best epsilon value before rerunning the DBSCAN. However, this spatial-only approach produced one main cluster that encompassed most of Virginia, suggesting that DBSCAN struggled to differentiate regional crash patterns effectively.

3.2 KMeans Clustering

After these results of DBSCAN we decided to run **KMeans** on the same data which only included spatial coordinates as features, and evaluated KMeans clustering for $k \in [2, 20]$, using the elbow method to identify the optimal number of clusters. The elbow point was selected based on inertia minimization, supported visually by plotting inertia against k. The chosen value of k also yielded favorable silhouette, Calinski-Harabasz, and Davies-Bouldin scores, indicating well-separated and compact clusters.

3.3 Cluster Interpretation and Visualization

We then visualized KMeans results using scatter plots and interactive Folium maps. KMeans clusters were spatially coherent and evenly distributed, with clearly identifiable centers. Each cluster's internal characteristics, such as crash severity distribution and categorical features (e.g., crash type), were analyzed using bar plots and severity histograms. These distributions further affirmed the interpretability of KMeans clusters.

Then, we looked into the features of each cluster, like crash severity, conditions of the road at a crash (lighting, rain, etc), and where the crash happened (interesection, nonintersection, entrance/exit ramp, etc), to find out which features were most at risk of crashes.

3.4 Subsectioning the Data

The results of Experiment 3.3 (discussed in results section) gave us a reasonable suspicion of bias in our data, which we believed was because there were more data points for crashes in normal conditions, leading to the results of Experiment 3.3.

To counteract this bias, we sub-sectioned the data based on certain features, like Day time Crashes, Night time Crashes, Dry Roads, and Wet Roads. Additionally to see if personal actions were more influential, we considered a few more features such as if alcohol was involved in the crash, and the effect of seat-belts on crash severity. We then reran KMeans on these subsections of the dataset and compared results of juxtaposed conditions like Day time and Night time, and Dry roads and Wet roads against each other in order to get better insights into what conditions lead to higher number of and more severe crashes.

4 Results

Link to colab:

`https://colab.research.google.com/drive/1WS3bpLVrdNd5esMG6ht7vWjKFNtnfnCL?authuser=1#scrollTo=rGtThF9tVTCu`

4.1 Results of DBSCAN and Initial KMeans

As seen in Experiment 3.1, DBSCAN yielded these results.

- Number of clusters: 1
- Noise points: 5 (0.00%)

In contrast, the KMeans Experiment 3.2 yielded:

- Number of clusters: 6
- Silhouette score: .60
- Calinski-Harabasz index: $334,000$
- Davies-Bouldin index: .57

4.2 Results of Cluster Analysis on KMeans

In Experiment 3.3, we were able to identify the clusters where most crashes happened, as well as what the conditions were of each crash. Specifically, we found that Northern Virginia, Richmond, and the Chesapeake Bay Area had the greatest number of crashes, and most crashes happened during the daytime, not at intersections, and on dry roads.

Number of crashes in each area:

- Northern Virginia: 34872 crashes
- Richmond: 25038 crashes
- Chesapeake Bay Area: 24358 crashes

4.3 Results of Subsectioning the Data

4.3.1 Day Time Crashes vs. Night Time Crashes

Day Time Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 0.59%
- suspected serious injury crashes was 4.59%
- suspected minor injury crashes was 19.06%
- possible injury crashes was 7.80%
- property damage only crashes was 67.97%

Night Time Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 1.00%
- suspected serious injury crashes was 5.35%
- suspected minor injury crashes was 16.03%
- possible injury crashes was 6.60%
- property damage only crashes was 71.02%

4.3.2 Dry Road Crashes vs. Wet Road Crashes

Dry Road Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 0.77%
- suspected serious injury crashes was 5.06%
- suspected minor injury crashes was 18.19%
- possible injury crashes was 7.36%
- property damage only crashes was 68.63%

Wet Road Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 0.59%
- suspected serious injury crashes was 4.18%
- suspected minor injury crashes was 17.48%
- possible injury crashes was 7.68%
- property damage only crashes was 70.07%

4.3.3 Effect of Alcohol on Crash Severity

Across the 6 clusters, the average percentage of:

- fatal crashes was 4.72%
- suspected serious injury crashes was 12.04%
- suspected minor injury crashes was 22.38%
- possible injury crashes was 6.30%
- property damage only crashes was 54.56%

4.3.4 Seatbelt Crashes vs No Seatbelt Crashes

Seatbelt Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 0.44%
- suspected serious injury crashes was 4.05%
- suspected minor injury crashes was 17.23%
- possible injury crashes was 7.44%
- property damage only crashes was 70.85%

No Seatbelt Crashes: Across the 6 clusters, the average percentage of:

- fatal crashes was 6.79%
- suspected serious injury crashes was 21.60%
- suspected minor injury crashes was 32.96%
- possible injury crashes was 5.97%
- property damage only crashes was 32.71%

5 Conclusion

This study set out to identify high-incident traffic crash areas in Virginia and examine the contributing environmental and situational factors using clustering and correlation analysis. Our hypothesis was that clustering algorithms, particularly DBSCAN, could reveal spatial patterns in crash data that, when coupled with categorical analysis, would provide actionable insights for public safety and traffic management efforts.

Though our results were not exactly as expected, we did supporting information that some of the variables exert influence over crashes rates. While DBSCAN was not as effective as initially believed, KMeans clustering, (especially when optimized through the elbow method) effectively segmented Virginia into meaningful spatial clusters of crash density. Northern Virginia, Richmond, and the Chesapeake Bay Area emerged as the most crash-prone regions. Further analysis revealed that most crashes occurred during the daytime, on dry roads, and away from intersections—suggesting that high visibility and favorable driving conditions do not inherently prevent crashes, possibly due to increased traffic volumes or driver complacency.

Our sub-sectioning approach offered deeper insight into how specific conditions dramatically affect crash severity. The data makes it alarmingly clear that alcohol consumption and failure to wear a seatbelt are two of the most dangerous behaviors on the road. Crashes involving alcohol had a fatality rate of 4.72%, more than seven times higher than the rate for typical daytime or dry-road crashes. Serious injuries were also significantly more frequent, with 12.04% of alcohol-related crashes resulting in suspected serious injury. Even more striking were the differences observed in crashes involving unbelted occupants: fatal crashes surged to 6.79%, and serious injuries rose to 21.60%, compared to just 0.44% and 4.05%, respectively, for crashes where seatbelts were worn. The data also showed that only 32.71% of crashes without seatbelts resulted in property damage alone, compared to 70.85% when seatbelts were used—highlighting how often unbelted crashes lead to injury or death. These findings underscore the life-saving importance of sober driving and consistent seatbelt use. They reaffirm the urgent need for intensified public awareness, stricter enforcement, and targeted interventions to reduce the devastating consequences of these preventable risk factors.

However, the analysis is not without limitations. The overrepresentation of crashes under normal conditions in the dataset likely introduced bias, potentially skewing our interpretation of risk factors. Moreover, DBSCAN's underperformance on our dataset—despite its theoretical strengths—points to limitations in its applicability when spatial distributions are relatively uniform or lack clear density gradients.

Future work could address these limitations by incorporating more detailed temporal data (e.g., exact time of day), vehicle information, and driver demographics. Additionally, exploring hybrid clustering methods or integrating supervised models could yield more nuanced insights. A predictive crash risk model informed by the clusters and correlations identified here could be developed to support real-time interventions by traffic authorities.

Ultimately, this project contributes to the broader goal of enhancing traffic safety in Virginia by highlighting geographic and conditional crash patterns. By focusing preventive efforts in identified high-risk zones and under hazardous conditions, policymakers and transportation planners can better allocate resources, design targeted safety campaigns, and implement infrastructural improvements—helping to reduce crash rates and protect the well-being of Virginia's residents.

6 Contribution

6.1 Brandon Yuan

- Wrote final technical report
- Worked on subsectioning the data for Experiment 3.4.
- Wrote script for presentation video.
- Voiced presentation video.

6.2 Jamie Tran

- Implemented data cleaning and preprocessing.
- Implemented initial DBSCAN
- Animated final presentation video.

6.3 Aaron Yuan

- Implemented final DBSCAN and K-means models.

- Worked on subsectioning the data for Experiment 3.4.
- Edited final presentation video.

This Work is part of the Machine Learning for Virginia (ML4VA) initiative at the University of Virginia

References

W. Cheng and S. Washington. Identifying accident hotspots: Comparing kernel density estimation and spatial autocorrelation analysis. *Accident Analysis & Prevention*, 2008.

I. Yamada and J.C. Thill. Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis*, 2011.