Enhancing Road Safety Through Cluster Analysis of US DOT Accident Data

CSC 460 Senior Seminar in Computer Science

Dr. Zohreh Safari

Iyana Jones, Computer Science & Coreen Mullen, Computer Science

4/27/2025

Table of Contents

1 **Abstract**

The purpose of this project is to identify traffic accident hotspots across North Carolina in both rural and urban areas by developing an interactive map that clusters traffic accidents to identify areas with a high frequency of accidents. We use both K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithms to be able to find hotspots and explore the role of spatio-temporal factors such as weather and time of day in the development of hot spots. Our data was obtained from the National Highway Traffic Safety Administration's Fatality Analysis Reporting System (NHTSA) and includes the location coordinates and other related data for all fatal accidents across the United States of America, where our data specifically narrows to accidents in North Carolina in the year 2021 and 2022. K-means and DBSCAN clustering have both been fully implemented into the software and both have undergone various developmental updates and verification tests. The results of the comparison highlight DBSCAN's benefits for finding significant clusters and handling of noise in the data. Overall the project aims to bring awareness to local governments on where hot spots are allowing for traffic infrastructure to know which areas are in need of the most assistance in relation to accidents. This project also hopes to inform the public on accident hotspots in their respective area by providing easy to read visualizations to the hot spot areas allowing for citizens to remain more cautious in these areas. By providing this information to both of these communities we hope to reduce the accident rates in North Carolina overall and potentially across the country.

2 Introduction

2.1 Project Description

This project represents the culmination of research and software development, with the intentions of significantly reducing traffic accidents throughout North Carolina and hopefully in other areas of the world as well. The number of traffic-related deaths in the United States of America in the year 2022 is calculated to be 42,795, a .3% decrease from 2021 (National Center for Statistics and Analysis, 2024). The amount of fatalities in the US is significantly higher than other countries, with the US having the most accidents compared to 29 other high-income counties (Yellman, 2022).

Driving is one of the most popular forms of transportation across the globe, in North Carolina around 25 cities have their own city specific public transportation system out of 552, public transport is more readily available in Urban areas across North Carolina (North Carolina Department of Transportation, 2024). Our dataset is pulled from the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS) 2022 and 2021 datasets. In this project we use k-means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to analyze and cluster the accidents. The end goal of the project is to create an interactive map to visualize accidents with hopes to increase public awareness about specific local areas where heightened road caution is necessary.

2.2 Objectives

Our specific goals for what this project would output includes:

1. Creating visual graphs that represent potentially influential factors on accidents gathered from our dataset like weather, time of day, etc.
2. Developing a map interface to represent accident hotspots across North Carolina

3. Explore different methods for clustering to ensure accurate hot spots are being detected.
4. Compare the k-means and DBSCAN clustering algorithms for this specific method of analysis to help future researchers understand how the results differ with each method.
5. Evaluate various methods for picking the optimal cluster number (K).

2.3 Scope

This project narrows the scope to specifically the Fatality Analysis Report of North Carolina in the year 2021 and 2022. It does not make any policy recommendations from a legal standpoint. The analysis focuses exclusively on fatal accidents due to data availability limitations, we take into account that this may skew the results of where hot spots are occurring for non-fatal accidents. We also acknowledge and take into consideration how factors such as population can skew the data at the county level.

3 Literature Review

3.1 Relevant Studies for Optimal K determination

3.1.1 Hybridization of K-means with Improved Firefly Algorithm

Alam and Muqeem's (2023) work on hybridizing K-means clustering with the Firefly algorithm takes its designs by the nature based Firefly algorithm that is used for testing the behavior of real fireflies. Their research demonstrates how this hybridization can significantly improve clustering accuracy for high-dimensional datasets. We considered using the Firefly algorithm for determining the number of clusters (K), and although we initially relied on the elbow and Silhouette method, we plan to explore this hybrid technique in future iterations to improve the robustness of the clustering.

The Firefly algorithm was originally proposed by Yang (2009), as an algorithm inspired by nature that imitates the social interactions of fireflies. In the context of clustering, It helps

determine the optimal number of clusters by considering each new value as a firefly where the brightness of the firefly determines how the points move to cluster together. This work by Alam and Muqueem attempts to circumvent the limitation of K-means' sensitivity for picking the initial value of K, allowing for better accuracy in the clusters, which is very useful for high dimensional datasets.

3.1.2 Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors

The work of Sabri et al. (2024) explores using dynamic clustering with adaptive distances and k-nearest neighbors method. Their method works with more flexible cluster boundaries that can adapt to natural patterns in the data. For each point, the algorithm will use the k-nearest neighbor method to aid in the cluster determination. After this step the work in the paper also uses adaptive distances to minimize dissimilarity in the clusters. The method works in a way where the clustering process runs different iterations allowing for the clusters to adapt to the new data.

This approach would be particularly useful for traffic accident data, where the data is irregularly shaped. The adaptability of this method was one of the strongest reasons for it to be a part of our literature review. Alongside the irregular shapes of where accidents are happening, roads are also not perfectly shaped as roadways can be complex in certain areas. Additionally, the iterative process would help for adapting the clustering as time goes on considering how traffic accident patterns would change as hopefully there would be improvements to the roadways, avoiding new clusters to form but removing existing ones.

3.2 Comparative Analysis and Justification of Chosen Method

After conducting our literature review, we found that the Firefly algorithm best suited our specific project. It uses Principal Component Analysis (PCA), Elbow method, and Silhouette

methods to determine the optimal k. The firefly algorithm works well for high dimensional datasets, which would closely align with the potential future work that could be made from this research. The ability of the algorithm to handle uniquely shaped clusters is essential to the way that traffic accidents form without defined patterns.

Though we were not able to implement the Firefly algorithm directly with our dataset, we incorporated the principles of PCA for dimensionality reduction and the combined Elbow and Silhouette methods for cluster validation. By using this hybrid approach we were better able to work with the complexity of all of the information that is included with accident data.

4 **Methodology**

4.1 Data Description, Collection, & Preprocessing

In our project we use two datasets: the NHTSA FARS(2022) and NHTSA FARS(2021) datasets. We specifically use the accident.csv located in these two folders made available by the NHTSA. These datasets include data for reported fatal accidents across all 50 states from 1975-2023 as of current. In the accident.csv for 2022 we have data that covers 39,222 rows representing individual accidents covering 85 columns that give information such as the state, latitude and longitude, number of fatalities, number of passengers in the vehicle, and other identifying information. The 2021 dataset contains similar information with a size of 39,786 rows x 85 columns

In the preprocessing of this data we removed columns that were not relevant to our project to reduce noise for the clustering and improve performance saving memory. We kept the features that were important to all of our different processes we ran. Another major step of the preprocessing was filtering the data to include only North Carolina and then in another step filtering by county to provide a more precise clustering. We also normalized numerical values in

the data to ensure that the data was processed equally, and handled missing values to ensure that empty values would not misconstrue the results.

4.2 Project Design

This project was designed from the beginning to be an interactive map allowing for a simple user interface. The interactivity of the design allows for the user to have a better experience and a more detailed experience, as with the zoom feature available in the design the user could look at roads they travel on often if they did not want to explore their whole county.

4.3 Tools & Technology

We used VSCode as our integrated development environment (IDE) in the development of the code for this project. We used the Python programming language and implemented the use of many python Packages in the code development.

01. Pandas: A very popular programming package, we use it in this script for reading in our dataset and managing the data frames such as merging the data for both years using the pandas concatenate feature.
02. Scikit-learn: Provided our DBSCAN and K-means functions.
03. Plotly: Provided us with the interactive county level maps.
04. OS package: Gave us access to operating system related tasks related to the file and folder directories.

The OS package itself is built into Python; the other packages mentioned are open source.

4.4 Clustering Algorithms

4.4.1 K-Means

The K-means algorithm that we use comes from sci-kit learn programming package (Pedregosa et al., 2011). The underlying algorithm for K-means works by grouping data into a user-defined

number of clusters, where each data point is assigned to the nearest cluster center based on distance calculations.

The K-means algorithm follows these simple steps:

1. Initialize k centers randomly.
2. Assign each data point to the nearest center.
3. Recalculate centers as the mean of all points assigned to that center.
4. Repeat steps 2-3 until convergence (centers no longer change significantly).

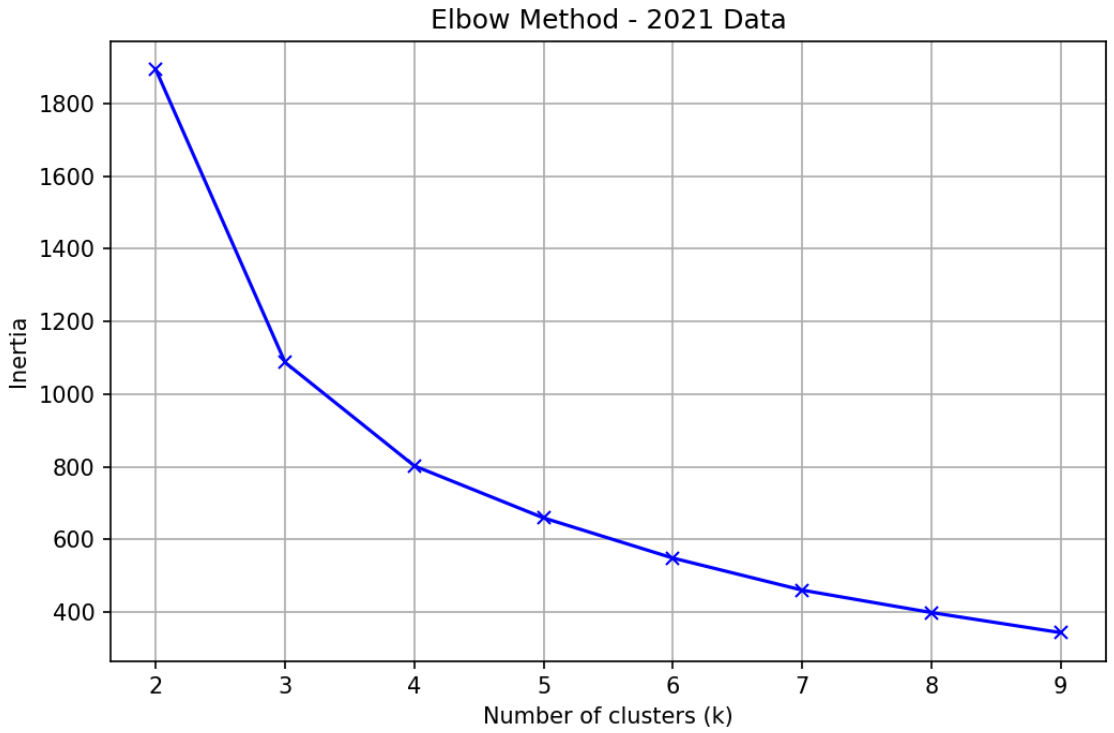

Fig. 1: A graph visualizing the results of the Elbow method being used on the 2021 FARS data for North Carolina.

In this case we use the Elbow method to try to determine the optimal number of clusters (k) [8]. The Elbow method works by plotting the variation as a function of the number of clusters, it then

locates where there is a sharp decrease on the plot representing that the results don't have a significant change after this point. In our testing we were returned an optimal value of K at 3 for the year 2021, where there is a sharp decrease, as seen in Fig. 1. The value of 3 was also given for the 2022 dataset, based off of the full state level data.

4.4.2 DBSCAN

DBSCAN is a clustering algorithm that groups together data points that are closely packed (dense regions), while marking points in low-density areas as noise (Ester et al., 1996). It does not require specifying the number of clusters beforehand and can detect clusters of arbitrary shape. Using two key parameters—eps (the neighborhood radius) and min_samples (minimum points to form a dense region)-DBSCAN classifies points as core, border, or noise.

The DBSCAN algorithm process works as follows:

1. For each point, find all points within distance eps.
2. Identify core points that have at least min_samples points within eps distance.
3. Form clusters by connecting core points that are within eps distance of each other.
4. Assign non-core points to clusters if they're within eps of a core point, or label them as noise.

This makes it especially useful for spatial data like accident locations, where clusters may vary in shape and density, and where identifying outliers is valuable.

5 Implementation

5.1 Development Structure & Code Format

Our development followed a modular approach with separate scripts for the multiple tasks needed to get our desired output:

1. Data preprocessing and cleaning
2. Exploratory data analysis
3. Clustering algorithm implementation
4. Visualization generation
5. Interactive map development

Each module was developed with clear documentation and comments to ensure maintainability and facilitate future extensions of the project.

5.2 Input and Output Representation

The input to our code is solely the filtered dataset from FARS. The program outputs multiple graphs as well as county-level maps into a folder after running the main script which show where the accidents land on a map of each county.

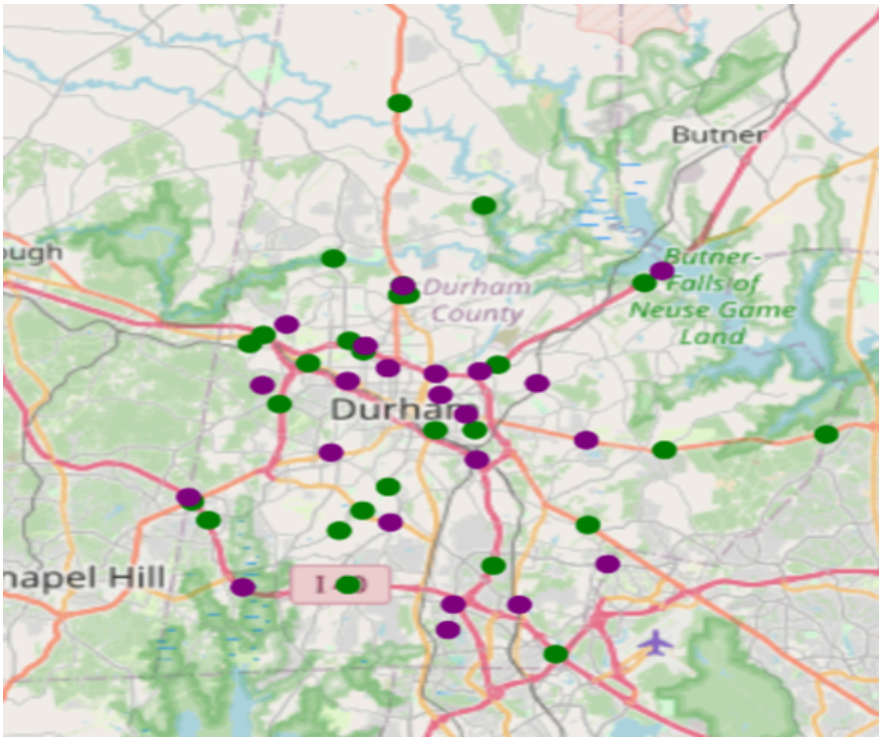

Fig. 2: A visualization of the location of accidents in Durham County, NC in 2021 (Purple) and 2022 (Green).

Other output includes a value for the number of optimal k determined by the Silhouette and Elbow method (Figure 1).

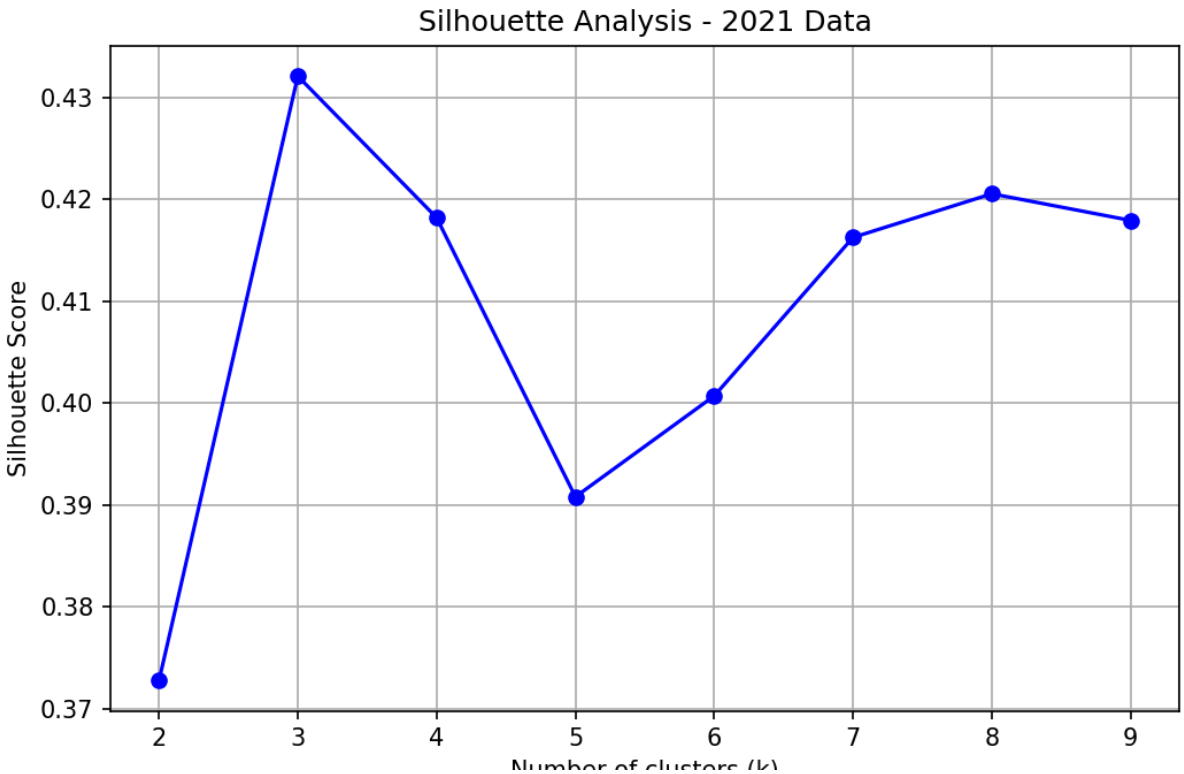

Fig. 3: A graph visualizing the results of the Silhouette method being used on the 2021 FARS data for North Carolina.

Visualizations of spatial-temporal factors like time of day and weather are also outputted with this code.

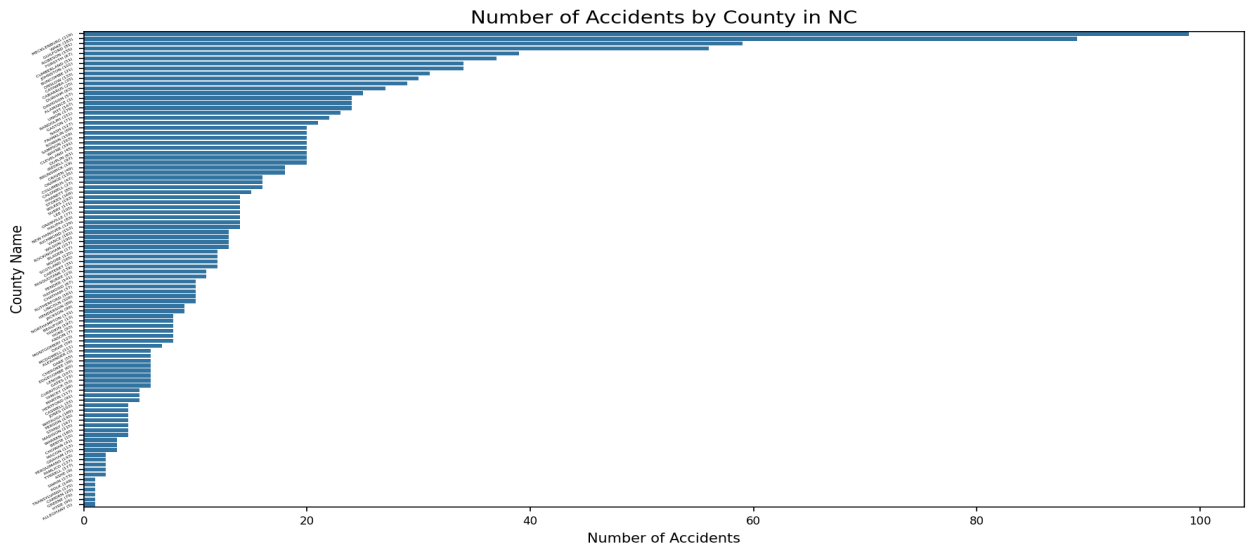

Fig. 4: A graph showing the number of accidents by county across North Carolina for the 2022 dataset.

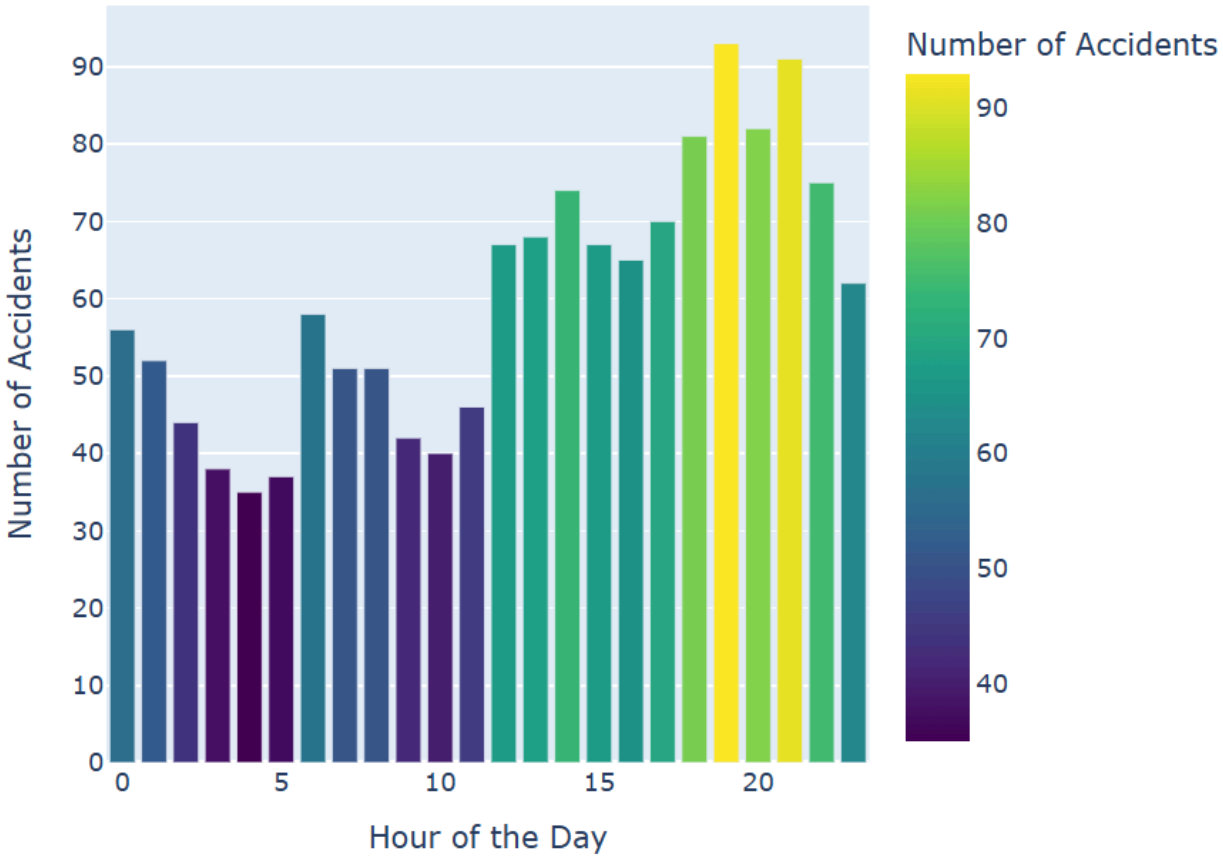

Fig. 5: A bar plot showing the number of accidents by time of day across North Carolina from the 2022 dataset.

5.3 Project Issues & Troubleshooting

Throughout the project, we faced two prominent challenges during development:

The initial issue occurred with our hopes to develop our literature review in more depth through implementing the algorithms in the papers into our datasets. However after contacting authors from both articles we learned that there was no project code available from the algorithms in the literature review available. Due to this we decided to take elements of the Firefly algorithm, after selecting it as the more aligned algorithm for our project, such as the PCA, Silhouette method, and Elbow method to perform our own comparison on these methods with our project. Even without an algorithm available to run based specifically on the firefly algorithm, we were still able to improve the review with the testing of Silhouette, PCA, and Elbow method.

The second major issue we encountered was a lack of memory notification we would receive when running our code as the output was generated. During the project we worked mainly with the FARS 2022 dataset, however we decided to also include 2021 which effectively doubled our dataset size. After this increase and separating the state dataset into county level maps, the task of running these files became too much for our computers to handle. But we were able to solve it by having our output maps be saved to the computer rather than autoloading after running the script. Even after running there are still memory space issues with having multiple maps open at the same time. Due to this error, we decided to only keep the two years merged so as to not encounter more memory issues and become unable to finish the project.

6 Description of Each Team Member's Responsibility

6.1 Iyana Jones

In this Project, I was primarily responsible for the K-means and Dbscan code that would close in on a specific city that had the most accidents. Then I would make that into a graph that shows visually how many clusters are in a specific spot. I was also in charge of seeing the comparisons and differences of the two algorithms to see which one works best with the database we have. I was also responsible for half of the side's development and presenting for the class and together we worked on the organization of the tasks.

6.2 Coreen Mullen

In this Project, I was primarily responsible for the literature review and development of the comparative analysis for the Principal Component Analysis, Elbow method, & Silhouette method to build upon our clustering approach. I was responsible for half of the slide development and presenting for the class. The writing of this report itself was something I collaborated on with my research partner. Together we also worked on the initial project selection, the planning and separation of tasks, and maintaining communication about changes and development during the project.

7 **Time Scheduling**

For scheduling our time spent on the project,we maintained communication with each other through text and email as well as using the dedicated time for the class working on the project. We set clear project goals at the beginning of the semester and were open to changing those deadlines if issues came along the way.

8 **Societal and Global Impact**

The societal and global impact of our work is seen through how research like this can help decrease the number of accidents not just in one place but all over the world. The overall output of our code is a visual diagram of where dangerous accidents seem to be occurring at a higher

rate, and that is not something that only happens in North Carolina or in just one part of the globe. According to the National Center for Statistics and Analysis (2024), there were 42,795 people who died in motor vehicle traffic crashes in the United States in 2022. Yellman and Sauber-Schatz (2022) found that the US had significantly higher motor vehicle crash death rates compared to other high-income countries, suggesting substantial room for improvement. By identifying accident hotspots and analyzing contributing factors, our research can help inform targeted safety interventions that could save lives both locally and, if applied more broadly, globally.

9 **Future Work & Recommendation**

The future work for this project would include:

1. Adding more data from FARS into the running of the code, specifically by including more years of data the declaration of the hotspots would become more accurate, as it would decrease the likelihood that the hotspot is just a coincidence.
2. Implementing the firefly algorithm described by Alam and Muqueem (2023) for more dynamic clustering.
3. Expanding the scope beyond North Carolina through the use of all the states available with FARS.
4. More development into spatial-temporal factors such as season based off of the dates given by the dataset.
5. Including non-fatal accident data collected from other resources.

10 **Conclusion**

In conclusion this research project was able to provide the results of an interactive map that represents where traffic accidents clusters are located in different counties across North Carolina.

In this research we also did a literature review comparison of two different methods of optimal k selection to assist in our K-means clustering research effort.

Our comparison of K-means and DBSCAN algorithms demonstrated that while both perform the same function of clustering, DBSCAN provided a major advantage for our specific project due to its ability to consider the diverse shapes of road networks across the state. The inclusion of the optimal k determination methods, Elbow & Silhouette, allowed for a more accurate approach to what numbers we would use with our clustering. The spatial-temporal analysis we performed highlighted the ways that everyday factors such as weather and time of day can play a role in where hot spots form.

By visualizing these accident clusters on interactive maps, we hope that we can provide a way for the local government of these different counties to explore the reasons that these areas might be a hotspot and find solutions for those reasons. We also hope that these interactive maps could be used to inform the public of high-risk areas allowing us to contribute to the development of a safer driving environment for people all across the state and hopefully further. While we were not able to complete all of the tasks possible for a project like this, we were able to achieve 3 of our major goals. We created spatio-temporal graphs to see how different factors affect accidents, create visualizations of traffic accidents onto maps, and perform analysis of the K-means and DBSCAN algorithms.

Bibliography

1. National Center for Statistics and Analysis. (2024, June). Overview of motor vehicle traffic crashes in 2022 (Report No. DOT HS 813 560, Revised). National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813560.
2. Yellman, M. A., & Sauber-Schatz, E. K. (2022). Motor vehicle crash deaths—United States and 28 other high-income countries, 2015 and 2019. Morbidity and Mortality Weekly Report, 71(26), 837–843. https://doi.org/10.15585/mmwr.mm7126a1
3. North Carolina Department of Transportation. (2024). *Public transit systems in North Carolina* [PDF]. NCDOT. https://www.ncdot.gov/divisions/integrated-mobility/public-transit-services/Documents/NC_public_transit.pdf
4. Alam, A., & Muqeem, M. (2023). *Hybridization of K-means with improved firefly algorithm for automatic clustering in high dimension*. arXiv. https://arxiv.org/abs/2302.10765
5. Sabri, M., Verde, R., Balzanella, A., & Vichi, M. (2024). A novel classification algorithm based on the synergy between dynamic clustering with adaptive distances and K-nearest neighbors. *Journal of Classification, 41*(2), 264–288. https://doi.org/10.1007/s00357-024-09471-5

6. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 226–231. AAAI Press.
7. National Transportation Atlas Database. (2014). Fatality Analysis Reporting System (FARS) 2014-Present Datasets. https://rosap.ntl.bts.gov/view/dot/56262
8. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
10. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering, 336*, 012017. https://doi.org/10.1088/1757-899X/336/1/012017
11. Pandas: McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 56-61. https://doi.org/10.25080/Majora-92bf1922-00a

12. Scikit-learn: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
13. Plotly: Plotly Technologies Inc. (2015). *Plotly's open source graphing library for Python*. Retrieved from https://plot.ly/python/

Appendix

Code available at: https://github.com/coreen-mullen/Senior_Project