Final Report
Group 1

Eunah Choi

Mallika Choudhari

Prabhakar Elavala

Nikhil Gade

Tejashwini Nagendra Singh

College of Professional Studies, Northeastern University

ALY 6110 Data Management & Big Data
Dr. Donhoffner
February 13, 2025

TABLE OF CONTENTS

Introduction

Traffic safety remains a critical concern in Montgomery County, Maryland, significantly affecting the daily lives and well-being of its residents. The frequency and severity of traffic incidents not only disrupt public safety but also hinders mobility, creating challenges for the community. Motivated by the profound impact of these incidents and our personal ties to the county, our team embarked on this study to explore the root causes of traffic collisions. This analysis reflects our commitment to fostering a safer community and underscores the importance of ensuring the well-being of every resident.

Leveraging a comprehensive dataset from the Montgomery County Police Department, which includes over 190,000 traffic collision records from the past decade, this report aims to provide a detailed examination of the factors contributing to these incidents. The dataset features critical variables such as *Weather*, *Surface_Condition*, *Lighting*, and *Collision_Type*, each offering insights essential for understanding and mitigating the risks associated with traffic collisions.

Through this report, we aim to identify actionable insights that can lead to significant improvements in road safety. This endeavor is not just an academic exercise but a dedicated effort to address a critical community concern through rigorous analysis and strategic recommendations.

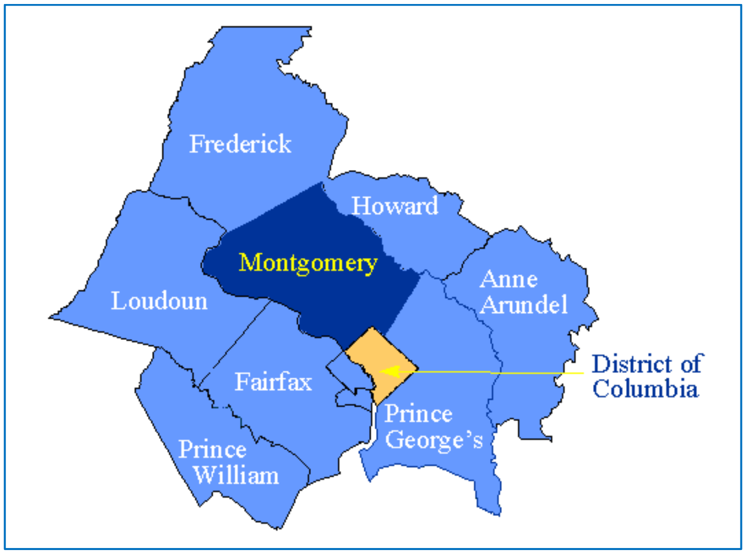


Figure 1(a): Map of Montgomery County

Research Question

Our study seeks to answer the following key question: How do lighting conditions impact traffic collisions in Montgomery County, and which locations require immediate streetlight installations to enhance road safety?

HYPOTHESIS TESTING

H_0 (Null Hypothesis): Lighting conditions have no significant impact on crash severity.

H_1 (Alternative Hypothesis): Lighting conditions significantly affect crash severity.

By analyzing high-risk crash zones, we aim to determine how inadequate lighting contributes to collision severity and identify priority areas where streetlight installation can mitigate accident risks.

Methodology

We utilized Python for data cleaning and analysis, focusing on standardizing critical variables such as Lighting Conditions and Road Surface Conditions to develop meaningful insights. Clustering techniques (K-Means and DBSCAN) were applied to identify high-risk crash zones, and statistical hypothesis testing (Chi-Square Test) was used to assess the significance of lighting conditions on crash severity.

K-Means clustering was implemented to segment crash-prone areas into high, moderate, and low-risk zones based on accident severity and frequency.

DBSCAN clustering was used to detect high-density crash hotspots in poorly lit areas, identifying locations where street lighting improvements are most needed.

Chi-Square testing confirmed a significant correlation between lighting conditions and crash severity, reinforcing the need for infrastructure upgrades in identified zones.

Geospatial visualizations, including heatmaps and scatter plots, were generated to illustrate patterns in accident distribution and pinpoint areas requiring intervention.

Real-World Problem

Traffic collisions in Montgomery County disrupt daily life and lead to significant health and economic burdens. This report leverages data to identify the conditions under which these collisions are most likely to occur and to understand the factors that contribute to their severity.

Figure 1 presents a heat map of Montgomery County, illustrating the geographic distribution of traffic collisions. This visual representation uses data points from the traffic incidents to highlight areas with higher frequencies of collisions. The heat intensity on the map correlates with the concentration of incidents, revealing hotspots where collisions are particularly prevalent.

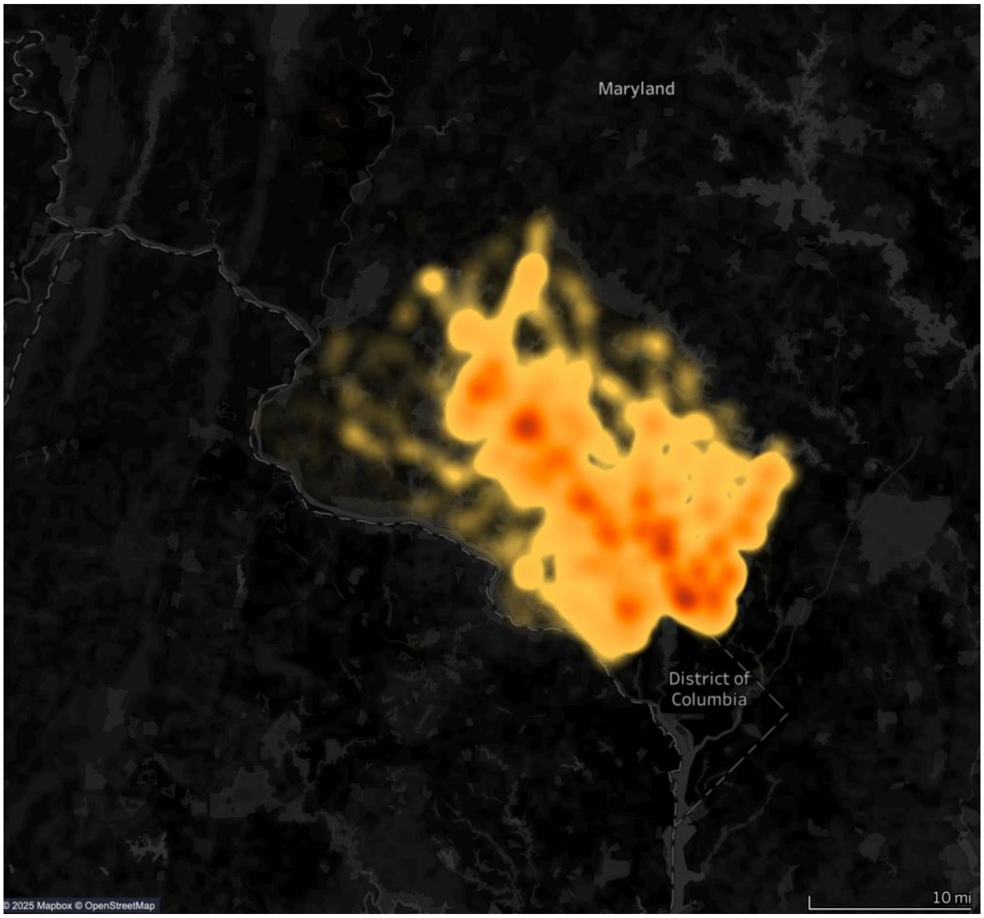


Figure 1(b): Heatmap of Crashes across Montgomery County from 2015 –2 2024

The map is instrumental in identifying areas where interventions might be most needed. For instance, the regions glowing brightest on the map suggest a higher occurrence of incidents, pointing to issues like inadequate lighting or poor road conditions. This geographic analysis is pivotal as it directs our focus towards specific areas that may require more detailed investigation and targeted measures to enhance road safety.

Environmental and Behavioral Insights

Our findings indicate that poor lighting conditions significantly contribute to high-severity collisions, particularly at intersections and high-speed roads. The bar chart analysis of lighting conditions confirmed that crashes occurring in "Dark (No Lights)" areas exhibit the highest severity levels, reinforcing the need for immediate streetlight installation in these zones. Additionally, statistical tests revealed a strong correlation between crash severity and inadequate lighting, providing a quantitative basis for intervention.

Analytical Focus and Key Questions

This study addresses several critical aspects of road safety, guiding our recommendations for infrastructure improvements and policy changes. A key focus is on lighting conditions and crash severity, exploring how inadequate lighting impacts collision severity and identifying poorly lit areas with the highest accident rates. Additionally, high-risk road identification is crucial in determining which specific streets require urgent streetlight installation, with geospatial mapping and clustering techniques helping to prioritize these locations. Finally, data-driven interventions play a vital role in mitigating risks in high-priority zones through measures such as smart lighting, improved signage, and enhanced traffic control mechanisms. Furthermore, a before-and-after crash analysis will be necessary to evaluate the effectiveness of lighting upgrades in reducing accident severity. By addressing these questions, this study provides concrete, data-backed recommendations for reducing severe crashes and enhancing road safety in Montgomery County.

Data Acquistion

Our data comes from DATA.gov, which is a United States government open database website. The datasets original name and source can be found as “Crash Reporting – Drivers Data.” and can be found at Link.The dataset is formatted as a csv file. It includes numerical (which also includes a floating time stamp for the crash date/time column), text, and location variables that include longitude and latitude coordinates. Each row of the dataset is displayed as a singular crash.

The dataset includes 192K rows, and 39 distinctive columns. The dataset is updated weekly. However, our dataset spans from January 1st, 2015, to January 21st, 2025. This range of dates is defined after data cleaning, as we do not have the direct history of when the data was downloaded from the website. This dataset includes structured data, where each row is defined as a singular crash or driver involved in a crash. The attributes of each crash in the columns provide insightful information regarding location, severity, type of crash, weather conditions, traffic conditions, and who is at fault.

Relevant Variables

In our evaluation and methods of our dataset we utilized variables such as:

Variable Name	Description	Unit of measurements
Weather	What type of weather condition at time of crash	Ex. Clear, Cloudy, Rain, Fog

Surface Condition	The state of the road	Ex. Dry, Ice, Wet
Light	Describing the time of day	Ex. Light, Dark, Dark (with lights on)
Traffic Control	What type of traffic control at site of crash	Ex. Stop Sign, Traffic Signal, no controls
Driver Substance Abuse	Was their substance evolved? What type?	Ex. None/None suspected, unknown, alcohol
Driver Distracted by	Was the driver distracted? What was the cause of the crash?	Ex. Not distracted, phone, unknown, inattentive
Speed Limit	Vehicle Circumstance – What was the speed limit posted in the area	Numerical: Miles per hour(mph)
Driver at Fault	Was the driver at fault?	Yes, No, unknown
Injury Severity	Was their injury and what was the severity?	Ex. No apparent injury, possible injury, suspected minor injury
Vehicle Body Type	What type of car?	Ex. Passenger car, van, bus, truck
Route Type	What type of roadway did the crash occur?	Ex. State (Maryland), County, Unknown,

We utilized these variables in our methods section for clustering and regression analysis as they provide concrete characteristics to track and direct attention to for future funding efforts. This data is representative of the population we intend to generalize, as it is census data of Montgomery County that we want to direct recommendations based off of our modeling. Some potential biases that come with working with this dataset is that this data is not being reported by one person but is being tracked by those that are present at the crash. Inherently, people will be describing each crash differently when it comes to some variables like vehicle damage extent and substance abuse. However, most of the variables in this dataset are descriptive of location and where damage has occurred to the vehicle, which limits the amount of bias that may come with working with this data set.

Data Cleaning and Preprocessing

To ensure data quality and integrity, we examined the dataset for missing values. The first step was removing columns that were not important for our analysis, specifically those with more than 100,000 missing values. This helped streamline the dataset while retaining relevant information. We categorized the remaining columns into high-priority and low-priority groups:

- High-priority columns: Instead of dropping rows with missing values, we replaced missing values with 'UNKNOWN' to preserve as much data as possible.
- Low-priority columns: For columns with less than 5% missing data, we dropped the rows to maintain data integrity without significant loss.

Regarding duplicated data that may be present in the dataset, we checked for duplicates using df.duplicated(). sum() and removed them using df.drop_duplicates(implace=True). Given that duplicate records can distort analysis and clustering results. This ensured the dataset contained only unique traffic collision instances. Given that data inconsistency arises when variations exist in categorical values from typos or improper formatting, we standardized the data. We standardized text-based columns: removing duplicate categories, group, grouping categories with very few occurrences into a single category called 'Other', correcting lower/uppercase inconsistencies, and updating categories with spelling mistakes to ensure uniformity.

After standardizing the dataset, we conducted variable transformation to improve model interpretability. We applied label encoding for categorical features such as ‘Weather Condition’ and ‘Collision Type’. We also applied binning for numerical features such as speed limits, grouping them into ranges (e.g., ‘low’, ‘moderate’, ‘high-speed zones’).

For feature engineering, new features were derived to enhance clustering effectiveness. The first feature includes *time of day classification,* which was created as a new categorical column categorizing time into morning, afternoon, evening, and night. The second feature includes *severity Index,* which is constructed as a severity metric using injury levels, vehicle damage, and road conditions. The last feature is a *weather-road condition interaction,* which combines weather conditions with road surface data to analyze impact on collisions.

To optimize computational efficiency and remove redundant information, we removed highly correlated features using Pearson correlation analysis and applied Principal Component Analysis (PCA) (if applicable) to reduce dimensionality for clustering.

Outliers in traffic collision data often represent critical incidents, such as severe crashes, extreme weather conditions, or rare but significant events. Removing these outliers could lead to loss of valuable insights into high-impact accidents. Since clustering techniques can handle such variations effectively, we retained the outliers to preserve the natural structure of the data and improve model robustness.

Analytical Method

Clustering Analysis

Big Data Method: Clustering for Traffic Safety

In our study, we employed clustering, a big data methodology used to group similar data points based on shared characteristics. Specifically, we implemented K-Means clustering, a partitioning technique that divides data into distinct groups, helping us identify meaningful patterns. In the context of traffic accidents, clustering allows us to segment accident locations based on severity, frequency, and external factors such as road conditions and driver behavior. By grouping accident-prone areas into clusters, we can prioritize high-risk zones, enabling data-driven decision-making for road safety improvements. This approach ensures that resources such as traffic signals, lighting enhancements, law enforcement, and infrastructure upgrades are allocated efficiently to reduce fatalities and injuries.

How Clustering Was Applied to Identify High-Risk Zones

To implement clustering, we extracted latitude and longitude data from accident reports and applied K-Means clustering with four clusters using the KMeans(n_clusters=4, random_state=42) method. This approach divides accident locations into four distinct groups, each representing different levels of risk and severity. The key idea behind this segmentation is to differentiate between high-priority and low-priority areas based on accident patterns. High-risk clusters require immediate intervention, while lower-risk clusters may only need minor safety adjustments. By focusing first on the most dangerous areas, authorities can implement targeted interventions such as speed enforcement, DUI checkpoints, improved road design, and enhanced signage, significantly reducing accident severity in those zones.

Cluster Analysis: Understanding Accident Patterns

The clustering results reveal four distinct groups based on accident severity and frequency. The dark purple cluster represents the highest concentration of severe crashes, primarily occurring at busy intersections or congested roadways where frequent collisions take place. The blue cluster includes accidents on high-speed roads and highways, where accidents tend to be more severe due to high impact forces. The green cluster consists of accidents influenced by adverse weather conditions, such as rain or icy roads, where environmental factors significantly contribute to collision risks. Finally, the yellow cluster represents low-risk areas, typically found in residential or suburban locations, where accident severity is minimal and involves minor collisions or fender benders.

Prioritization of High-Risk Areas

We are concentrating on the high-priority areas, which include the dark purple and blue clusters, as they exhibit the highest severity and frequency of accidents. These zones require immediate attention as they pose the greatest risk to public safety. In these areas, we will analyze factors such as speed, traffic conditions, driver behavior, and environmental influences to determine the leading causes of severe collisions. Next, the moderate-priority areas correspond to the green cluster, where environmental factors like rain and snow contribute to accidents. These locations require intervention but not as urgently as high-priority zones. Lastly, the yellow cluster represents low-priority areas, where accident severity is minimal, and no immediate action is necessary. Our next steps will involve examining these clusters in detail to understand the key contributing factors affecting collision severity before moving toward specific recommendations.

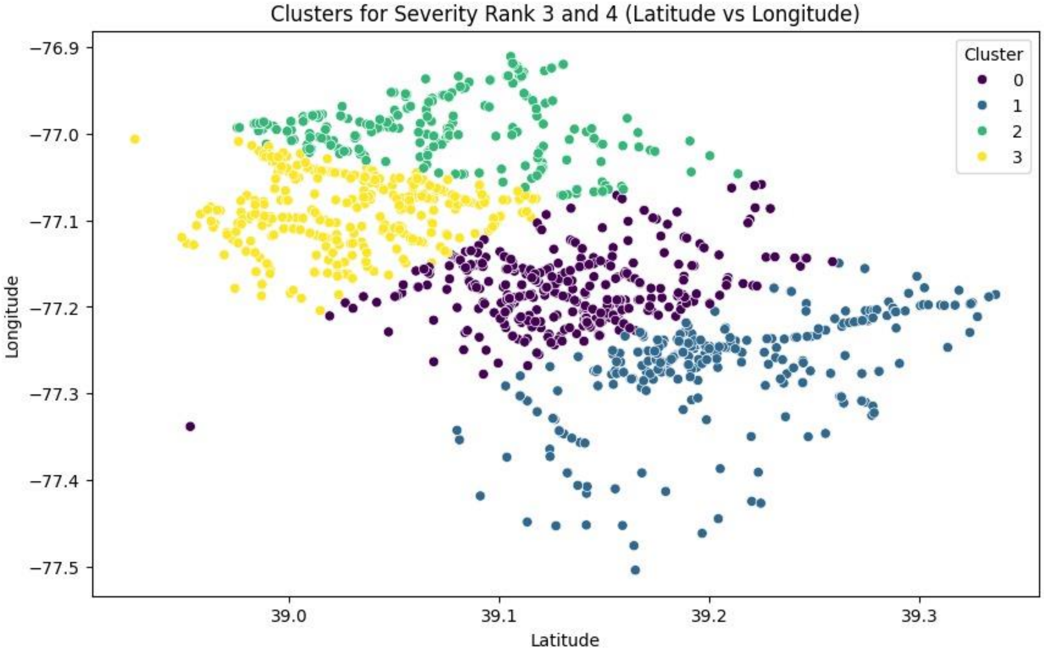


Figure 2: Clustering

Crash Severity Distribution by Speed Limit for High-Severity Crashes

The histogram titled "Crash Severity Distribution by Speed Limit (Severity Rank = 3 or 4)" provides valuable insights into the relationship between speed limits and high-severity crashes. The data reveals that most high-severity crashes (Rank 3 or 4) occur in speed zones between 30-50 mph, with a notable peak around 40 mph. In contrast, areas with speed limits below 20 mph show relatively few severe crashes, suggesting that lower speeds may contribute to reduced crash severity. Interestingly, there is also a decline in severe crashes for speed limits above 50 mph, which could be attributed to factors such as better road design, lower traffic volumes, or stricter safety regulations on high-speed roads.

The graph exhibits a bimodal distribution, with two distinct peaks—one around 35 mph and another around 45 mph—potentially indicating specific risk factors associated with these speed ranges. These findings have important policy implications, as targeted safety interventions, such as enhanced enforcement or road design modifications, could be implemented in 30-50 mph zones where severe crashes are most frequent. However, the relationship between speed limits and crash severity is complex, and this visualization suggests the need for deeper analysis into additional contributing factors, such as road conditions, driver behavior, and vehicle characteristics. Ultimately, this data serves as a strong foundation for further road safety research and policy development aimed at reducing high-severity crashes.

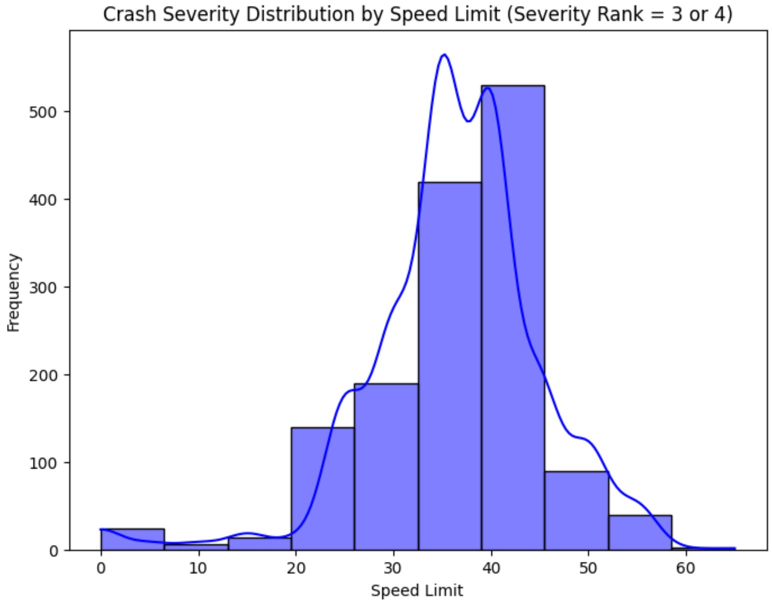


Figure 3: Crash Severity Distribution by Speed limit

Analysis of Lighting Conditions and Crash Severity in High-Priority Areas

Hypotheses Formulation

H₀ (Null Hypothesis): Lighting conditions have no significant impact on crash severity. *(Crash severity is independent of whether the accident occurred in daylight, dark with lights, or dark without lights.)*

H_1 (Alternative Hypothesis): Lighting conditions significantly affect crash severity. *(Crash severity is dependent on lighting conditions, meaning poor lighting leads to more severe crashes.)*

The bar chart illustrating the impact of lighting conditions on crash severity in high-priority zones highlights a significant correlation between poor visibility and severe accidents. The highest average crash severity is observed in areas classified as "Dark (No Lights)", reinforcing the hypothesis that insufficient lighting exacerbates accident severity. Even areas categorized as "Dark (With Lights On)" show relatively high crash severity, indicating that while artificial lighting helps, it does not completely mitigate the risks associated with nighttime driving. Comparatively, daylight conditions exhibit a much lower crash severity, suggesting that natural light provides a safer driving environment with better visibility and reduced reaction time impairments.

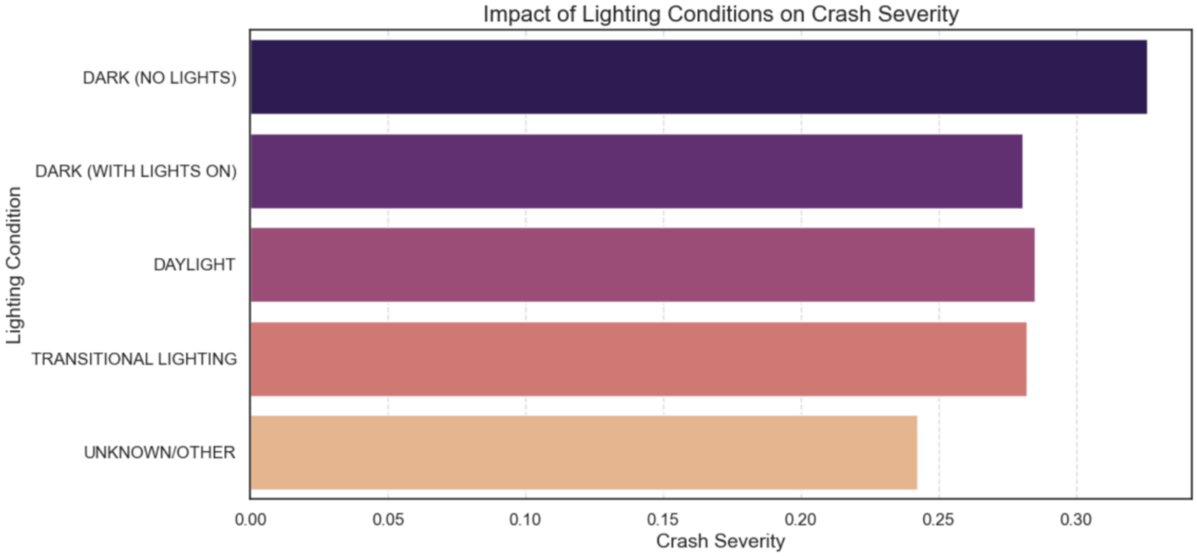
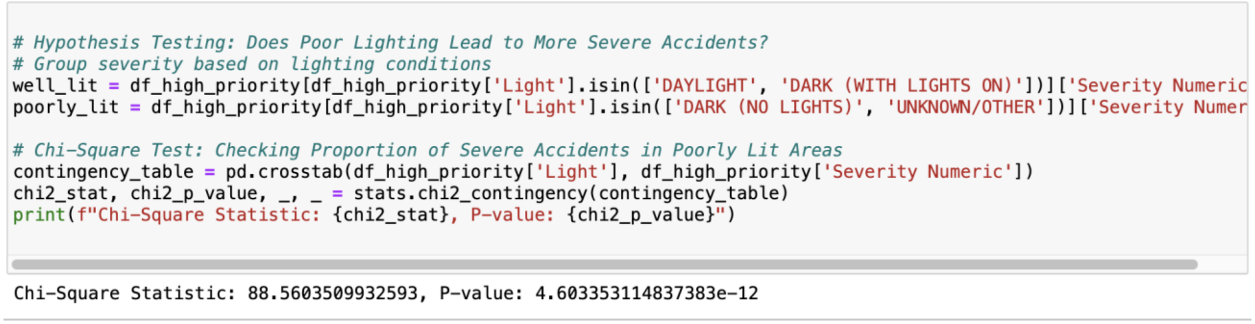


Figure 4: Impact of Lighting Conditions on Crash Severity

The Chi-Square test further validates these observations, showing a statistically significant relationship between lighting conditions and crash severity. The extremely low p-value suggests that the variations in crash severity across different lighting conditions are not random but rather influenced by the presence or absence of adequate lighting infrastructure. This confirms that inadequate lighting in high-risk areas directly contributes to more severe accidents, necessitating targeted interventions such as better street lighting in dark zones to mitigate crash severity.



```
# Hypothesis Testing: Does Poor Lighting Lead to More Severe Accidents?
# Group severity based on lighting conditions
well_lit = df_high_priority[df_high_priority['Light'].isin(['DAYLIGHT', 'DARK (WITH LIGHTS ON)'])]['Severity Numeric
poorly_lit = df_high_priority[df_high_priority['Light'].isin(['DARK (NO LIGHTS)', 'UNKNOWN/OTHER'])]['Severity Numer

# Chi-Square Test: Checking Proportion of Severe Accidents in Poorly Lit Areas
contingency_table = pd.crosstab(df_high_priority['Light'], df_high_priority['Severity Numeric'])
chi2_stat, chi2_p_value, _, _ = stats.chi2_contingency(contingency_table)
print(f"Chi-Square Statistic: {chi2_stat}, P-value: {chi2_p_value}")
```

```
Chi-Square Statistic: 88.5603509932593, P-value: 4.603353114837383e-12
```

Figure 5 Python Code for Hypothesis testing

In conclusion, the findings from both the bar chart and hypothesis testing indicate that improving street lighting in high-priority crash zones should be a top priority. Our next steps will involve identifying specific geographical locations within these high-risk clusters where lighting improvements could most effectively reduce crash severity. By integrating spatial mapping with crash data, we can pinpoint the worst-affected areas and prioritize infrastructure upgrades accordingly.

DBSCAN Analysis for High-Risk Crash Hotspots

The DBSCAN clustering technique was applied to identify severe crash hotspots in poorly lit conditions. The visualization highlights high-density crash zones, where darker red points represent locations with a higher concentration of severe crashes. This analysis allows us to prioritize intervention efforts by pinpointing areas where accidents frequently occur due to low visibility and the absence of adequate street lighting. By using this unsupervised learning method, we can isolate specific problem regions, helping decision-makers allocate resources efficiently.

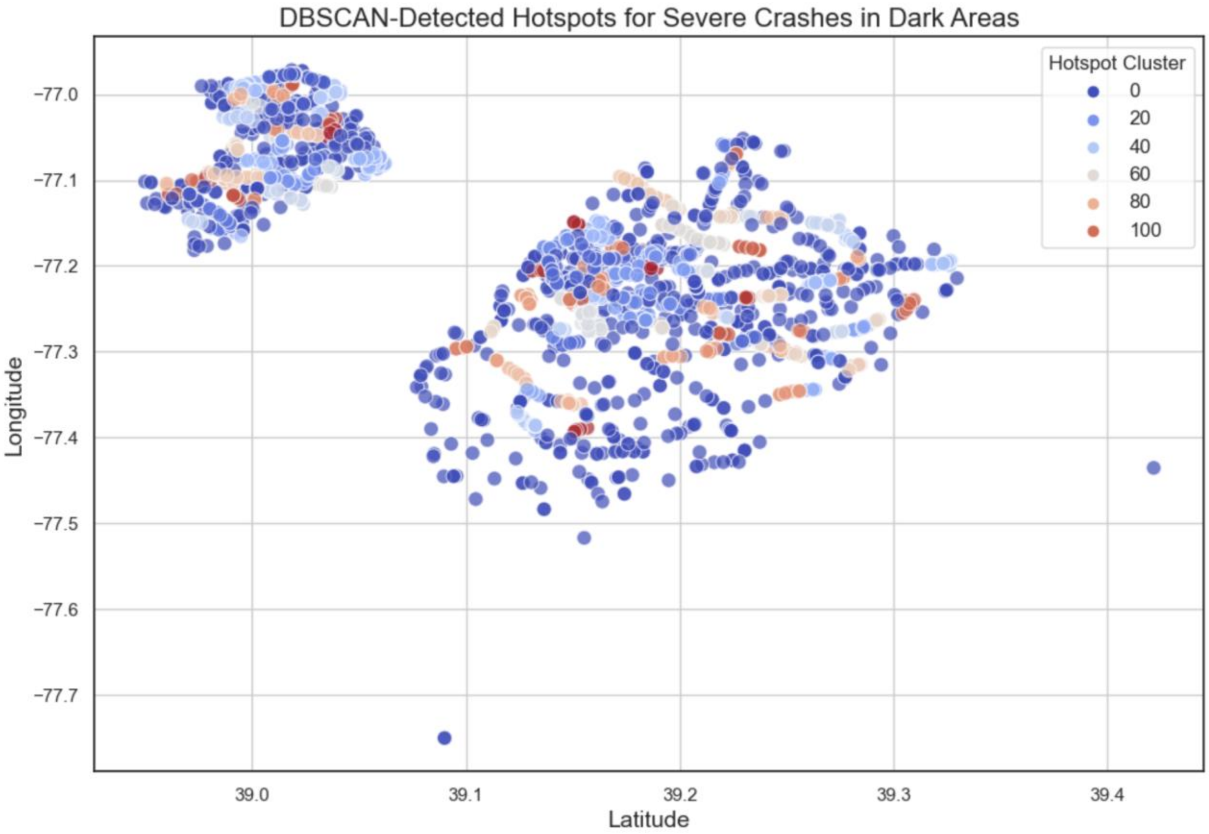


Figure 6: DBSCAN- Detected Hotspots for Severe Crashes in Dark Areas

Focusing further, the visualization below filters only clusters with severity values of 80 and above, ensuring that only the most critical hotspots are considered. These high-risk zones are primarily located in the northern and central regions, aligning with previously identified dark zones. The clustering reveals patterns of repeated accidents, emphasizing the need for urgent countermeasures. Roads within these clusters should be immediately evaluated for safety enhancements, including improved street lighting, signage, and traffic control measures.

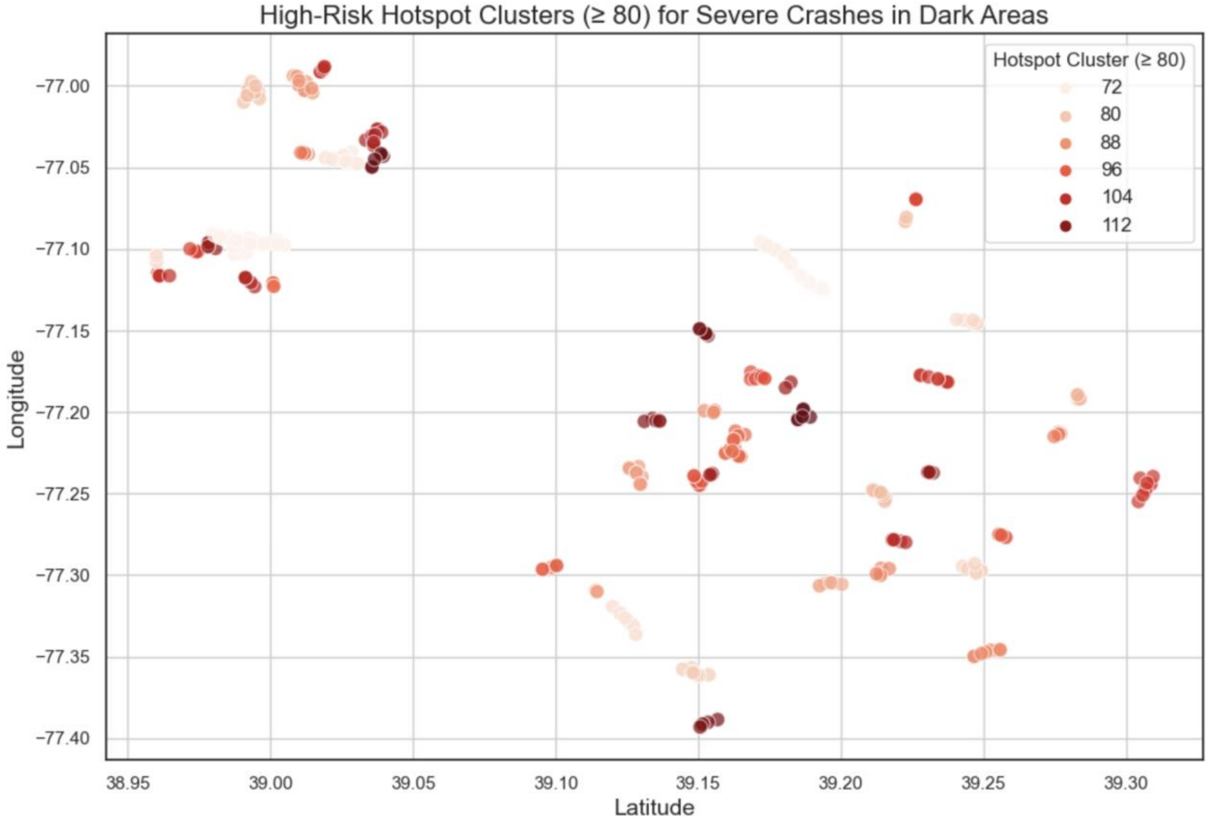
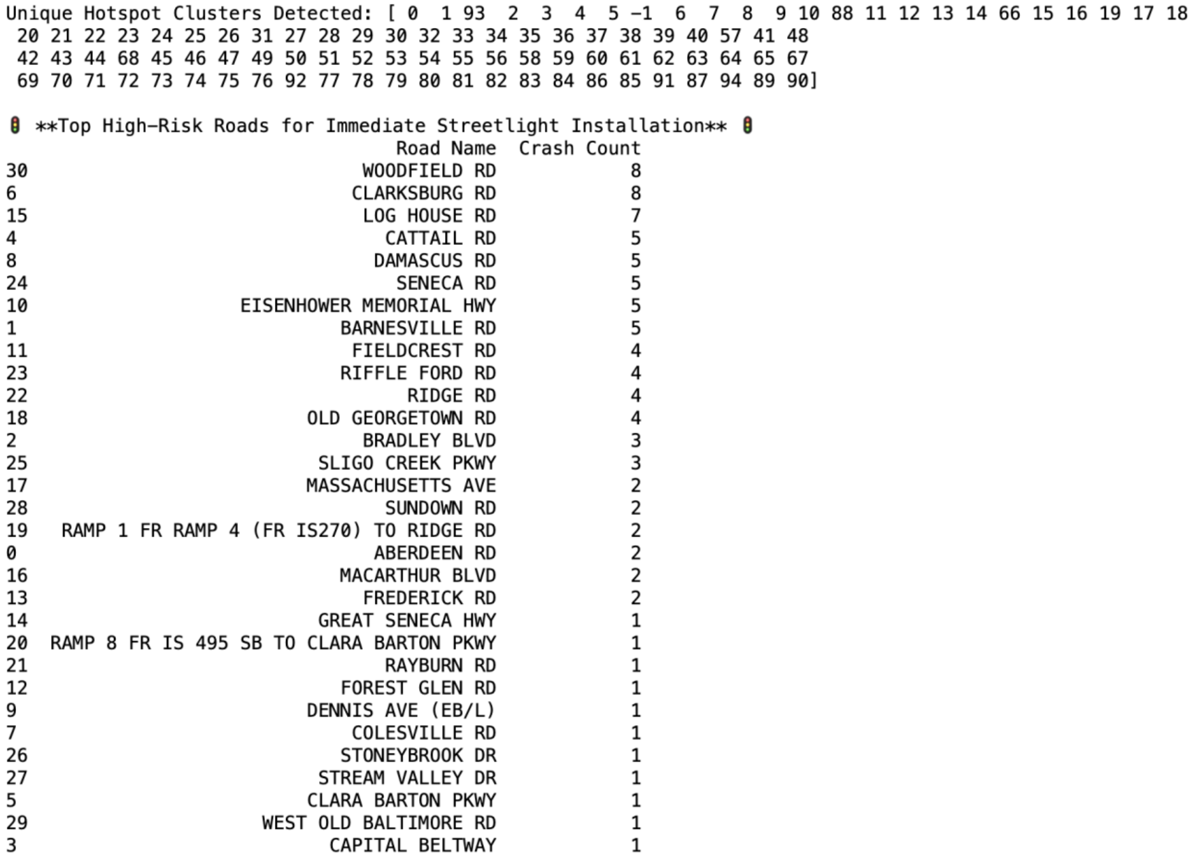


Figure 7: High Risk Hotspot Clusters

Prioritization of Roads for Immediate Streetlight Installation

Based on the DBSCAN clustering and high-risk severity analysis, we have identified specific roads where the absence of lighting is a critical factor in accident frequency. The roads listed in the high-risk hotspot clusters (≥80) represent areas where multiple severe crashes have occurred, making them prime candidates for immediate intervention.

Top High-Risk Roads

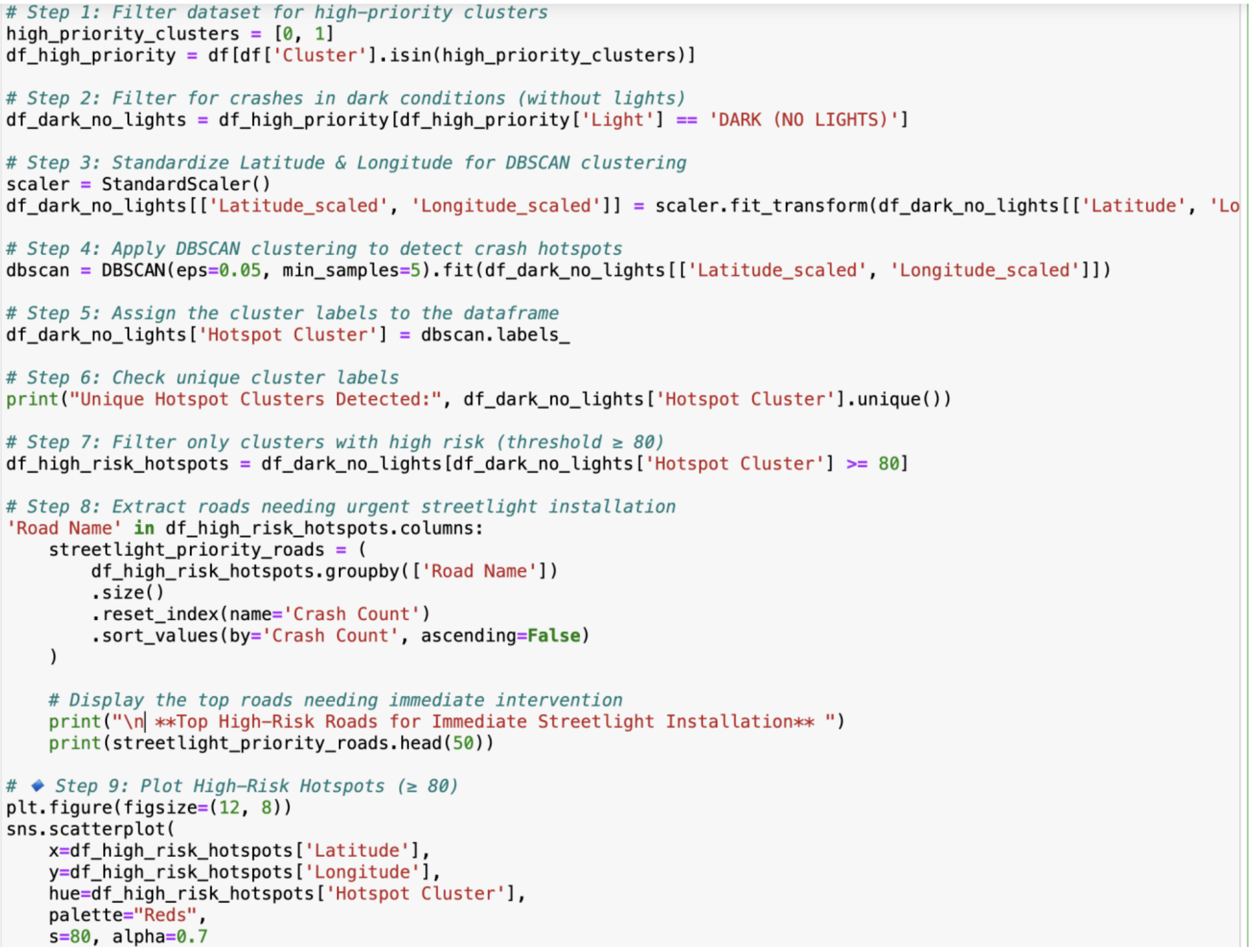


```
Unique Hotspot Clusters Detected: [ 0  1 93  2  3  4  5 -1  6  7  8  9 10 88 11 12 13 14 66 15 16 19 17 18
 20 21 22 23 24 25 26 31 27 28 29 30 32 33 34 35 36 37 38 39 40 57 41 48
 42 43 44 68 45 46 47 49 50 51 52 53 54 55 56 58 59 60 61 62 63 64 65 67
 69 70 71 72 73 74 75 76 92 77 78 79 80 81 82 83 84 86 85 91 87 94 89 90]

🚦 **Top High-Risk Roads for Immediate Streetlight Installation** 🚦
                                   Road Name  Crash Count
30                              WOODFIELD RD            8
6                              CLARKSBURG RD            8
15                              LOG HOUSE RD            7
4                                 CATTAIL RD            5
8                                DAMASCUS RD            5
24                                 SENECA RD            5
10                   EISENHOWER MEMORIAL HWY            5
1                             BARNESVILLE RD            5
11                             FIELDCREST RD            4
23                            RIFFLE FORD RD            4
22                                  RIDGE RD            4
18                         OLD GEORGETOWN RD            4
2                               BRADLEY BLVD            3
25                          SLIGO CREEK PKWY            3
17                         MASSACHUSETTS AVE            2
28                                SUNDOWN RD            2
19   RAMP 1 FR RAMP 4 (FR IS270) TO RIDGE RD            2
0                                ABERDEEN RD            2
16                            MACARTHUR BLVD            2
13                              FREDERICK RD            2
14                          GREAT SENECA HWY            1
20  RAMP 8 FR IS 495 SB TO CLARA BARTON PKWY            1
21                                RAYBURN RD            1
12                            FOREST GLEN RD            1
9                          DENNIS AVE (EB/L)            1
7                              COLESVILLE RD            1
26                            STONEYBROOK DR            1
27                          STREAM VALLEY DR            1
5                          CLARA BARTON PKWY            1
29                     WEST OLD BALTIMORE RD            1
3                            CAPITAL BELTWAY            1
```

Figure 8: High Risk Roads Where Street Light Installation is needed

The analysis highlights roads such as Woodfield Rd, Clarksburg Rd, Log House Rd, Damascus Rd, Eisenhower Memorial Hwy, Barnesville Rd, and Ridge Rd, which have recorded the highest number of crashes under poor lighting conditions. These roads require urgent streetlight installation to mitigate risks. The presence of these roads across different regions suggests that multiple zones in the county lack proper illumination, contributing significantly to accident severity.



```
# Step 1: Filter dataset for high-priority clusters
high_priority_clusters = [0, 1]
df_high_priority = df[df['Cluster'].isin(high_priority_clusters)]

# Step 2: Filter for crashes in dark conditions (without lights)
df_dark_no_lights = df_high_priority[df_high_priority['Light'] == 'DARK (NO LIGHTS)']

# Step 3: Standardize Latitude & Longitude for DBSCAN clustering
scaler = StandardScaler()
df_dark_no_lights[['Latitude_scaled', 'Longitude_scaled']] = scaler.fit_transform(df_dark_no_lights[['Latitude', 'Lo

# Step 4: Apply DBSCAN clustering to detect crash hotspots
dbscan = DBSCAN(eps=0.05, min_samples=5).fit(df_dark_no_lights[['Latitude_scaled', 'Longitude_scaled']])

# Step 5: Assign the cluster labels to the dataframe
df_dark_no_lights['Hotspot Cluster'] = dbscan.labels_

# Step 6: Check unique cluster labels
print("Unique Hotspot Clusters Detected:", df_dark_no_lights['Hotspot Cluster'].unique())

# Step 7: Filter only clusters with high risk (threshold ≥ 80)
df_high_risk_hotspots = df_dark_no_lights[df_dark_no_lights['Hotspot Cluster'] >= 80]

# Step 8: Extract roads needing urgent streetlight installation
'Road Name' in df_high_risk_hotspots.columns:
    streetlight_priority_roads = (
        df_high_risk_hotspots.groupby(['Road Name'])
        .size()
        .reset_index(name='Crash Count')
        .sort_values(by='Crash Count', ascending=False)
    )

    # Display the top roads needing immediate intervention
    print("\n **Top High-Risk Roads for Immediate Streetlight Installation** ")
    print(streetlight_priority_roads.head(50))

# ◆ Step 9: Plot High-Risk Hotspots (≥ 80)
plt.figure(figsize=(12, 8))
sns.scatterplot(
    x=df_high_risk_hotspots['Latitude'],
    y=df_high_risk_hotspots['Longitude'],
    hue=df_high_risk_hotspots['Hotspot Cluster'],
    palette="Reds",
    s=80, alpha=0.7
```

Figure 9: Python Code for the DBSCAN model

Findings

Our analysis has provided critical insights into the relationship between lighting conditions, road infrastructure, and accident severity. The clustering methods, K-Means and DBSCAN, have allowed us to segment high-risk crash zones effectively, prioritizing areas based on severity and frequency. The impact of lighting conditions on accident severity was validated through statistical hypothesis testing, reinforcing the need for immediate streetlight installation in dark zones. The DBSCAN clustering further refined our understanding by pinpointing specific hotspots where repeated severe accidents have occurred. The visualization of high-risk hotspot clusters with severity values of 85 and above has been instrumental in identifying problematic roads and intersections that require urgent safety interventions. Notably, roads such as Woodfield Rd, Clarksburg Rd, Log House Rd, Damascus Rd, Eisenhower Memorial Hwy, and Ridge Rd emerged as high-priority locations, requiring urgent improvements. The presence of clusters of severe crashes in these dark, unlit areas confirms that inadequate illumination is a significant contributing factor to road accidents.

Furthermore, our findings highlight a clear hierarchical structure of crash severity zones:

- High-priority clusters: Areas with the most severe crashes, requiring immediate intervention.
- Moderate-priority clusters: Areas influenced by environmental factors like rain or ice, necessitating targeted interventions but not as urgently.
- Low-priority clusters: Locations with minor fender benders and low accident severity, which do not require immediate action.

These findings provide a solid foundation for implementing targeted safety measures, ensuring that resources are allocated where they are most needed to prevent further fatalities and injuries.

Recommendations

To effectively mitigate the risks identified in our analysis, a comprehensive action plan is required, tailored to the severity levels of different accident-prone areas. In high-priority zones, immediate interventions such as installing high-intensity LED streetlights should be undertaken, particularly at dangerous intersections and curved roads where visibility is lowest. Additionally, traffic control enhancements like reflective markers, improved road signs, and speed bumps must be implemented to alert drivers and reduce the likelihood of high-speed collisions. These measures should be reinforced with law enforcement strategies, including increased police patrols and automated speed detection cameras, to deter reckless driving and ensure strict compliance with road safety regulations.

For moderate-priority zones, which are largely affected by environmental factors like rain, snow, or fog, targeted measures should focus on improving road surface conditions to enhance traction and prevent vehicles from skidding. Solutions like anti-skid coatings, better drainage systems, and the deployment of weather-sensitive traffic regulations can help reduce accident risks in these areas. Additionally, smart traffic signage that adjusts speed limits based on real-time weather conditions should be integrated to enhance driver awareness and safety. Enhanced visibility features, such as cat’s eyes reflectors and solar-powered road studs, should be introduced in locations where lighting infrastructure is not feasible, providing drivers with clear road guidance in low-visibility situations.

Low-priority zones, while not requiring immediate action, still benefit from long-term safety planning to maintain low accident rates. Regular road safety audits should be conducted to monitor traffic conditions and ensure that these areas remain safe over time. Public awareness campaigns focusing on safe driving habits should be promoted to educate drivers on accident prevention, particularly in residential areas where pedestrians and cyclists are at risk. Additionally, gradual infrastructure upgrades, such as lane separations, road widening, and improved road curvature design, should be considered for long-term safety improvements. These steps will ensure that even areas with lower accident severity continue to evolve in line with best practices in road safety management.

Future Steps for High-Risk Roads

For high-risk roads, geospatial mapping should identify the optimal locations for streetlights, using DBSCAN clustering and crash heatmaps. Smart lighting systems that adjust brightness based on traffic density and environmental conditions should be considered for efficiency.

Post-implementation, a before-and-after crash analysis will assess impact and refine strategies. Collaboration with local authorities, urban planners, and law enforcement will ensure effective execution. Engaging community stakeholders will provide insights into unreported hazards, refining future safety measures. By combining data-driven interventions with targeted infrastructure upgrades, these steps will significantly reduce severe crashes and save lives.

Conclusion

This study highlights the importance of a data-driven approach in improving road safety by identifying high-risk crash zones and prioritizing areas that require immediate intervention. By applying K-Means and DBSCAN clustering, we were able to categorize accident-prone locations based on crash severity, lighting conditions, and environmental factors. The results show that poorly lit areas experience a significantly higher number of severe crashes, emphasizing the need for better street lighting and infrastructure improvements in these locations. Through geospatial analysis and heatmaps, we have identified specific roads where crashes frequently occur under dark conditions, making them the top priority for streetlight installation.

Moving forward, implementing smart lighting solutions that adjust brightness based on real-time traffic density and environmental conditions could improve efficiency and safety. After installation, a before-and-after crash analysis should be conducted to assess the effectiveness of these interventions. Additionally, collaboration with local authorities, urban planners, and traffic enforcement teams will be essential in ensuring the proper execution of these improvements. Engaging with the community to gather feedback on unreported hazards and road safety concerns can also provide valuable insights for further refinement of safety measures.

By combining advanced data analytics with targeted infrastructure upgrades, this study provides a practical framework for reducing severe accidents and enhancing traffic safety. Implementing these recommendations will not only save lives but also contribute to a more efficient and well-managed urban transportation system. As cities continue to grow, integrating machine learning, AI-based predictive models, and real-time data analysis into urban planning can lead to proactive safety strategies that prevent accidents before they occur.

References

1. Data Montgomery. *Crash Reporting - Drivers Data*. (2025, February 7). Crash Reporting Drivers Data | Open Data Portal. *Data Montgomery*. https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632/about_data

2. Data.Gov. (2025, February 7). Crash Reporting – Drivers Data. *Data.Gov.* *https://catalog.data.gov/dataset/crash-reporting-drivers-data*