

Optimizing Traffic Flow in Austin Using Clustering and Analysis of Collision Trends

Carly Carroll

The University of Rhode Island

BAI 457: Visualizing Data for Business Intelligence

May 5, 2025

Abstract

This project investigates patterns in traffic incidents across Austin, Texas, with a focus on understanding how speed, volume, time and incident severity contribute to collision trends. Using real-world traffic data, the analysis applies various techniques in order to classify and analyze traffic patterns in a meaningful way. Multiple structured visualizations were developed in both Python and Power BI to analyze incident frequency, time, cluster-specific risk profiles, severity distribution, and average traffic conditions. The results indicate that collision frequency and severity are closely associated with peak traffic hours and congestion levels, especially during late afternoons and weekends. The study highlights the importance of awareness during certain times, high volumes, and high speeds to support smarter traffic management and public safety initiatives. These findings can inform city planners, law enforcement, and local agencies in implementing various strategies to reduce traffic risks and optimize urban mobility.

Introduction

Understanding the dynamics behind traffic incidents is essential for improving public safety and urban infrastructure. In metropolitan areas like Austin, Texas, where population growth and rapid development have increased traffic volumes, analyzing traffic data can provide valuable insights into when and where collisions are likely to occur. This report focuses on uncovering key patterns in traffic incidents by analyzing various factors including speed, volume, time, and severity.

The overarching goal of this study is to identify conditions in which traffic collisions are most frequent and severe, using a dataset of real time radar and incident report data. By clustering traffic conditions based on speed, volume, and time, and categorizing the severity of each incident, the analysis reveals underlying trends that might otherwise go unnoticed.

The central question driving this project is: *Under what traffic conditions, and at what points in time, are collisions most likely to occur in Austin?* The project also explores how severity varies across time and cluster categories, and whether patterns in volume and speed can serve as predictors of high-risk conditions. The ultimate objective is to provide evidence that supports more efficient traffic planning, real-time risk detection, and data driven policy decisions for urban safety.

Data Collection and Preprocessing

The datasets used for this project contain detailed traffic incident records collected throughout Austin, Texas. Radar-based traffic metrics were merged with incident report data using a common feature column to ensure each entry had complete contextual information. The final merged data includes both real-time traffic metrics from radar sensors and descriptive incident reports. The key features of this dataset include:

- Speed (mph): measured by radar data at the time of incident
- Volume: the number of vehicles recorded in a given interval
- Time_Hour: timestamp of the incident
- Issue Reported: text description of the incident type
- Location, Latitude, and Longitude: provides location information

To begin the analysis, the dataset was imported into a Python environment for preprocessing. Dates were parsed to exact time-based features which were then used for time-based analyses. Additionally, incident descriptions were categorized into broader groups using custom logic based on keyword matching. This step helped reduce noise within the data and focus the analysis on meaningful categories. The Issues Reported feature was mapped into five main categories:

- Collision: includes reports with key words like “crash”, “collision”, or “accident”
- Hazard/Obstruction: includes key words like “debris”, “blocked”, or “hazard”
- Vehicle Issue: stalled vehicles or vehicle fires
- Fatality/Injury: contains indicators of injury or fatality
- Other: all remaining entries not fitting the above

The dataset was cleaned and preprocessed to ensure models and visualizations were based on accurate records. The cleaned dataset was exported and used for both analysis in Python and visual dashboard development in Power BI.

Methodology

This analysis combines machine learning, statistical grouping, and descriptive visualization to examine traffic conditions in Austin. The methodology consists of three core components: Clustering traffic conditions using K-means, categorizing incident severity, and visualizing patterns across time, speed, and volume.

Traffic Condition Clustering with K-Means

To understand which traffic conditions are most associated with collisions, unsupervised learning was used to group traffic data based on three numeric features: volume, speed, and time. The K-means clustering algorithm was applied to records labeled as collisions. The algorithm produced three meaningful traffic condition clusters:

- High Congestion: high volume, low speed, typically in late afternoon and evening hours
- Rush Hour Flow: moderate speed and volume, primarily during commute times
- High Speed/Low Volume: low volume and high speed, common during late night and early morning hours.

Each collision record was assigned to one of these clusters based on its traffic conditions. This classification enabled targeted analysis of when and where dangerous conditions tend to arise.

Incident Severity Grouping

To analyze impact severity, the Issue Reported field was mapped to three simple levels:

- Major: fatal or injury-related incidents
- Moderate: typical collisions and crashes
- Minor: stalled vehicles or minor obstructions

This grouping allowed for easier comparisons of severity patterns across time and traffic clusters. The grouped severity data was used to generate both summary statistics and visualizations.

Tools and Techniques

The Python visualizations were developed using:

- Plotly Express: interactive bar charts and scatter plots
- Matplotlib/Seaborn: used for heatmaps and boxplots
- Scikit-learn: K-means clustering
- Pandas/Numpy: data manipulation and transformation

Power BI was used to create a secondary set of visualizations from the same dataset.

These visualizations focused on simpler, aggregated insights to complement the structure of the Python plots. Measures such as average speed and incident count were calculated within Power BI using DAX and build-in aggregations. The combination of Python and Power BI enabled granular clustering of traffic behavior and high level exploration of trends across time and severity.

Uncovering Collision Trends Through Pattern Clustering

To uncover the patterns that define high-risk traffic conditions in Austin, this section focuses on evaluating how collision frequency and severity change over time and under various traffic flow clusters. The goal is to understand when and under what circumstances collisions are most likely to occur, how they vary across clusters, and what those patterns reveal about congestion, speed, and collision risk.

Traffic Pattern Clustering with K-Means

To segment traffic conditions into meaningful groups and to better understand the environments where collisions take place, K-means clustering was performed using three numeric features: vehicle speed, traffic volume, and hour of the day. This technique grouped collision events into clusters that reflect distinct traffic patterns and allowed the analysis to assign each incident to a specific traffic context for deeper analysis and examination.

Figure 1

Traffic Condition Clusters by Volume and Speed



The Clustering results revealed three interpretable traffic conditions: High Speed/Low Volume, Rush Hour Flow, and High Congestion. Each cluster represents a unique traffic environment. High Speed/Low Volume collisions, typically occurring late at night or early in the morning, are characterized by fast moving traffic and sparsely populated roads. Rush Hour Flow captures moderate speed incidents during commuter times, while High Congestion encompasses slower traffic commonly observed in afternoon and evening hours. This classification enabled the rest of the analysis to explore temporal and severity-based patterns within each traffic condition type.

Collision Frequency by Hour and Cluster

To determine when collisions are most likely to occur across each traffic cluster, the number of incidents was plotted against the hour of the day. This visualization provides insight into how traffic risk evolves throughout a 24 hour period and how it differs depending on traffic flow patterns.

Figure 2

Collision Frequency by Hour for each Traffic Cluster



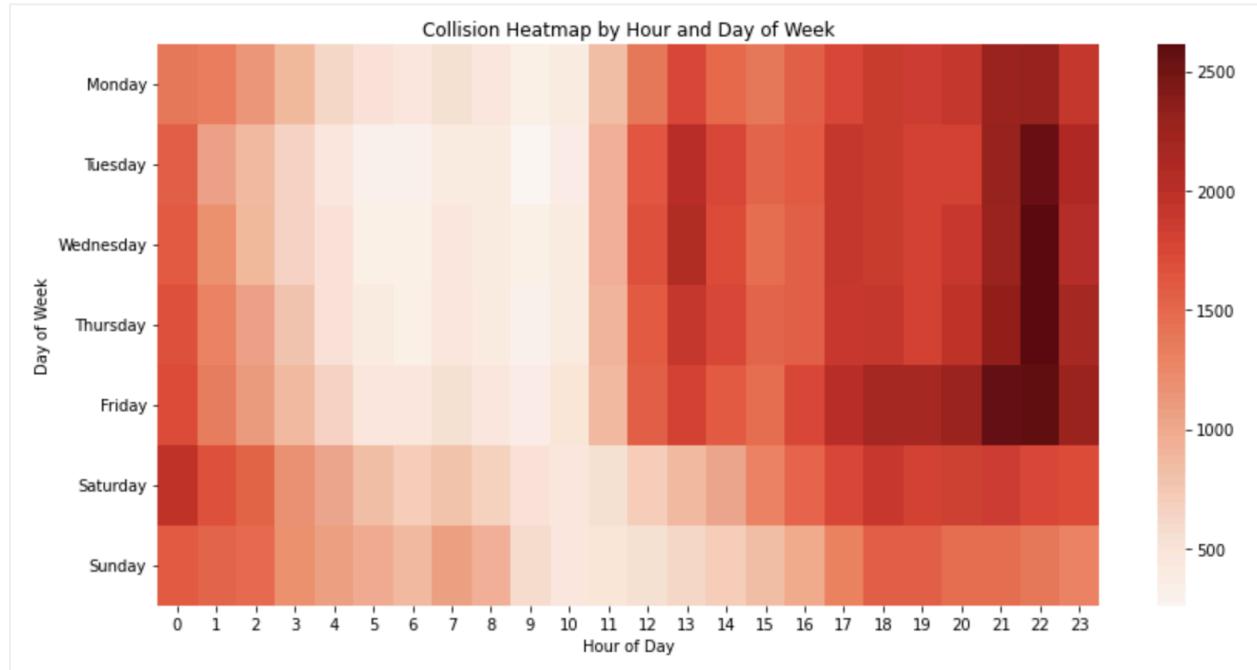
The analysis of the line chart reveals that High Speed/Low Volume collisions occur most frequently between midnight and dawn, when traffic is limited but speeds are at their highest. Rush hour flow incidents display two prominent peaks, corresponding to morning and evening commute hours. In contrast, High Congestion collisions steadily increase throughout the day, peaking in the late afternoon and early evening. These trends confirm that risk varies not only with the time of day but also with the surrounding traffic environment, suggesting that both speed and density contribute to the likelihood of a collision.

Weekly Collision Patterns

To expand the temporal analysis, collision density was examined by both the hour of the day and the day of the week. This dual axis view offers a more clear understanding of which specific time blocks throughout both a single day and day of the week carry the highest risk of traffic incidents.

Figure 3

Collision Density by Hour and Day of the Week



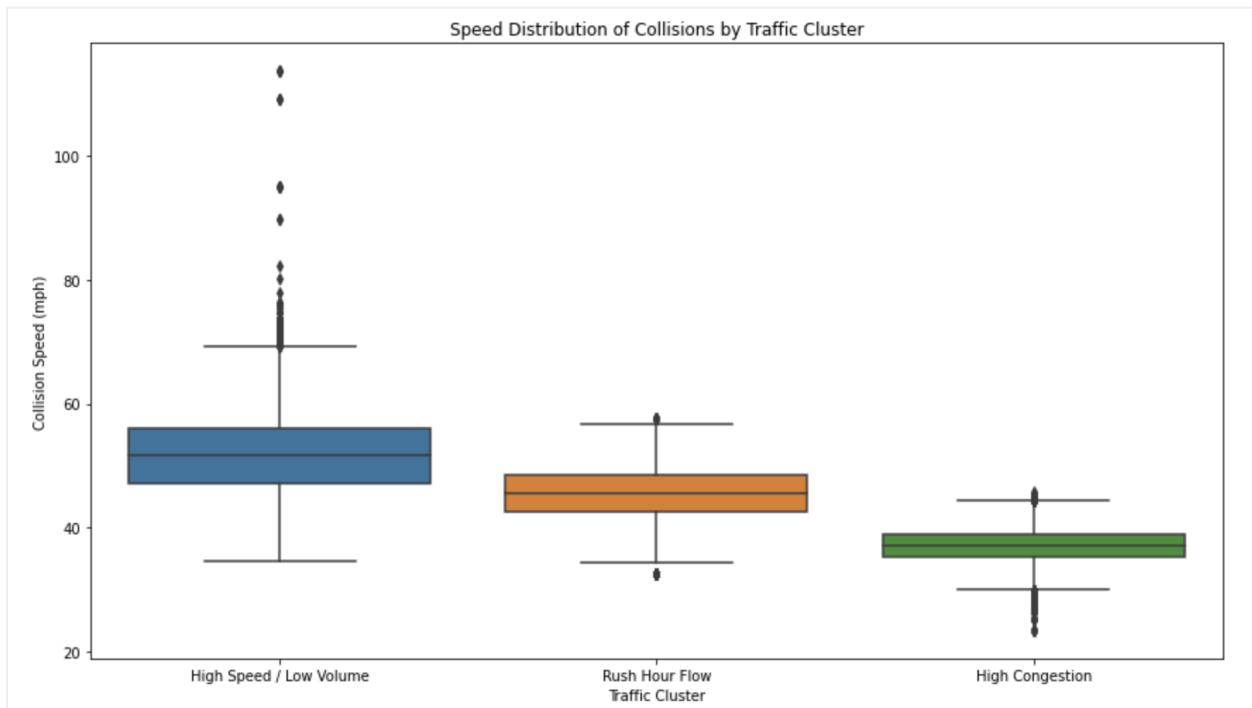
The heatmap reveals that Fridays and Saturdays experience the highest overall collision volume, particularly between 4 pm and 10 pm. These elevated evening risks are likely driven by increased weekend leisure travel, high congestion, and decreased attentiveness. Weekdays show a smaller but consistent spike during afternoon commute hours. Sunday presents with more moderate collisions overall, though small peaks are still visible. This visualization highlights the effect of traffic flow patterns, pointing to late week evenings as particularly hazardous times on Austin's roads.

Collision Speeds by Traffic Cluster

To examine the relationship between collision speed and traffic environment, a box plot was created to compare speed distributions across the three identified clusters. This allows for the evaluation of how driving speed at the time of a collision varies depending on volume and time.

Figure 4

Speed Distribution of Collisions by Traffic Cluster



This plot shows that High Speed/Low Volume collisions occur at the fastest periods, with a wide range and frequent outliers, confirming the severity potential of these off-peak incidents. Rush Hour Flow collisions occur at moderate and relatively consistent speeds, while High Congestion collisions happen at overall, much lower speeds. These results support the idea that while congestion leads to more frequent collisions, high-speed environments carry a greater risk of serious impact, suggesting a trade-off between volume-related frequency and speed-related severity.

Total Incidents by Traffic Cluster

To summarize the overall impact of each traffic condition, total collision counts were aggregated for each cluster. This helped prioritize which traffic environments contribute most to the city's incident volume and should be targeted for invention.

Figure 5

Total Number of Collisions per Traffic Cluster



The bar chart confirms that High Congestion is responsible for the largest share of collisions, followed by Rush Hour Flow and then High Speed/Low Volume. While fewer collisions happen at higher speeds, the congestion cluster contributes significantly more to overall incident frequency. These results contribute to a two-way risk narrative: low speed, high volume environments lead to frequent incidents, while high speed environments, though less common, pose a greater risk for injury.

Citywide Incident Trends and Summary Insights

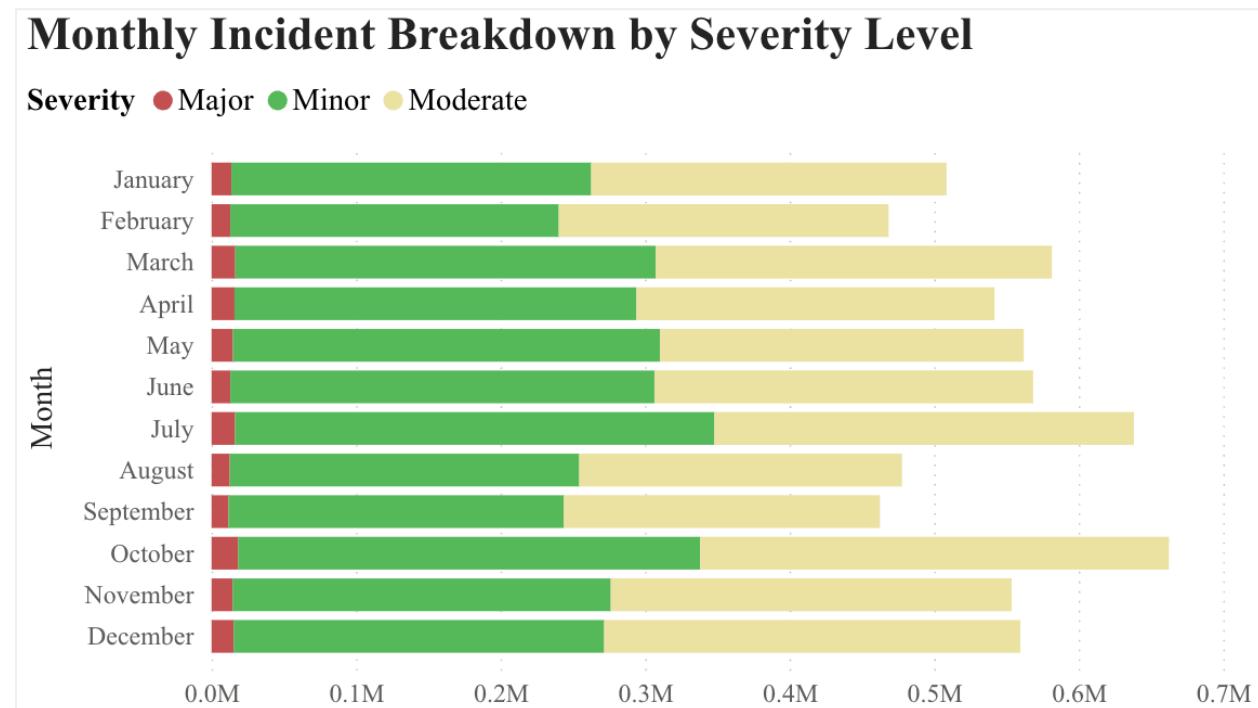
To complement the Python-based clustering and statistical analyses, a Power BI dashboard was developed to explore broader trends in collision timing and severity using interactive visuals. This section showcases multiple visualizations designed for interpretability, aiding a quick grasp of key insights related to collision frequency, incident type, and time-based risk factors for city planners. Each visualization utilizes different features from the same dataset, reinforcing the themes explored previously and helping to minimize the number of incidents throughout the city.

Monthly Incident Breakdown by Severity

To evaluate how incident patterns fluctuate overtime, a clustered column chart was created to show the number of incidents reported each month, grouped by severity level. This visualization helps identify seasonal or monthly shifts in collision types. Monthly trend analysis is useful for long-term traffic planning and anticipating high-risk periods. Incorporating severity into the timeline adds a deeper dimension to frequency patterns. This figure serves as a foundation for understanding how incident behavior evolves across the calendar year by analyzing month-to-month data. These insights can guide the timing when to implement traffic-related minimization strategies.

Figure 6

Monthly Distribution of Traffic Incidents by Severity



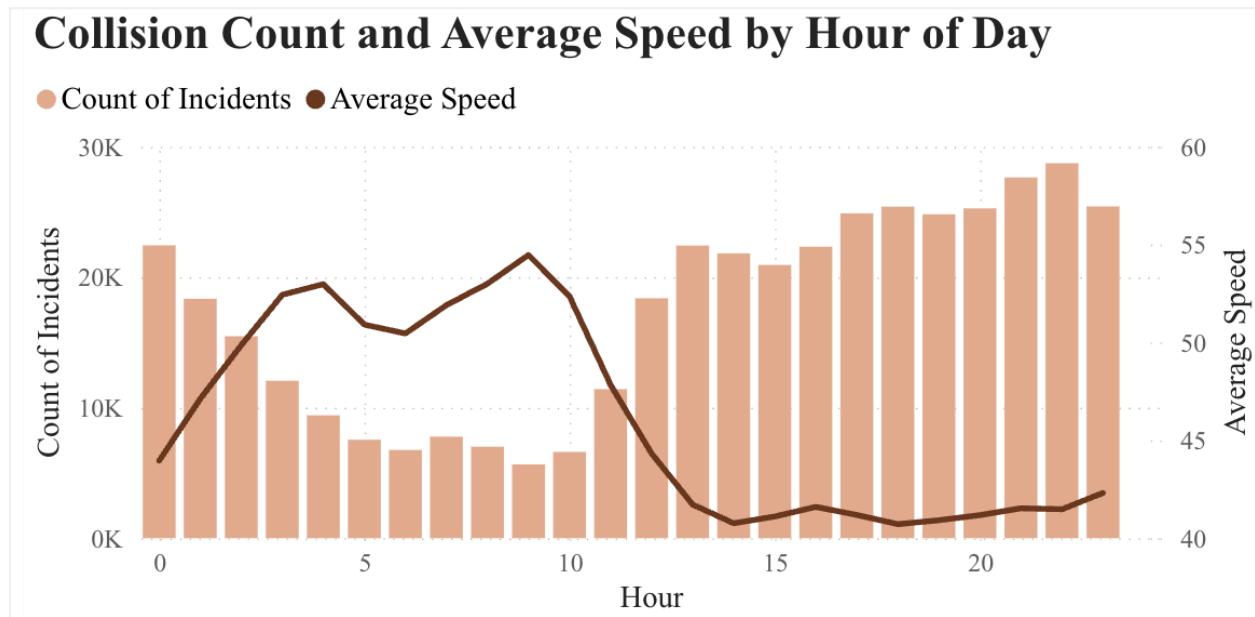
This visualization reveals a clear increase in collision reports during the spring and fall months. These seasonal peaks may reflect school schedules, event traffic, or simply, weather-related changes like “spring fever”. Most incidents in all months fall into the moderate and minor categories, with only a small proportion flagged as major. However, there are subtle increases in major incidents during months with higher collisions, suggesting that volume and severity can occasionally rise together.

Hourly Incident Volume vs Speed

To understand how traffic volume and average speed interact throughout the day, a combination chart was created. This visualization presents hourly incidents using vertical bars, while average speed is overlaid as a line. Both of these metrics are broken down by each hour to highlight key trends.

Figure 7

Hourly Traffic Speed and Incident Volume Trends



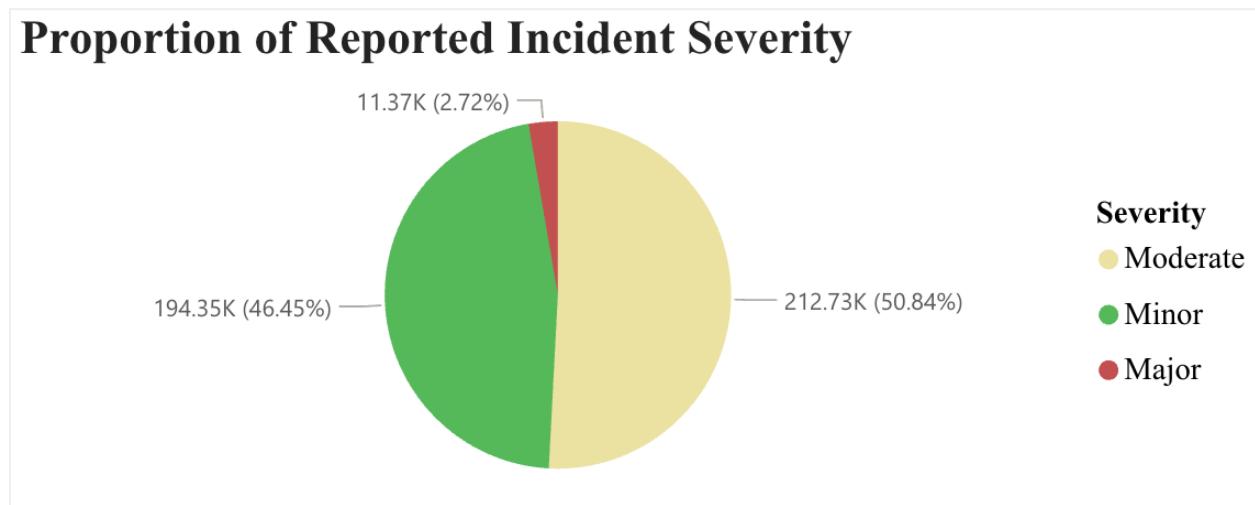
This visualization shows that incident volume tends to increase during the middle of the day and peak between 4 pm-6 pm, aligning with the evening rush. Additionally, the average speed decreases as the volume increases, particularly in the late afternoon, highlighting the typical effects of traffic congestion. During the early morning hours, speeds remain high but incident numbers are low, supporting earlier findings that fast and low-volume conditions are associated with fewer, but more severe collisions.

Severity Distribution

To provide an overview of the types of incidents occurring across the dataset, a pie chart was created to show the proportion of incidents by severity level. This offers a general view of the dominant incident types without focusing on the time or location. The chart divides the data into three severity levels: minor, moderate, and major. Each proportion of the pie reflects the count of incidents assigned to that category across the entire dataset.

Figure 8

Overall Distribution of Incident Severity Types



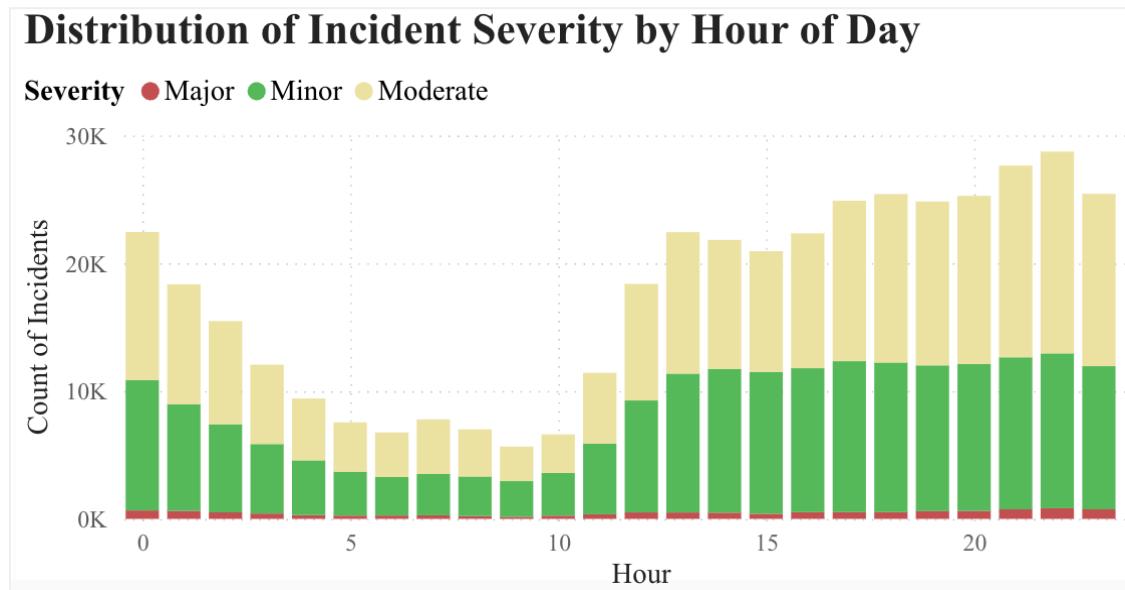
The pie chart confirms that minor and moderate incidents account for the vast majority of reported cases, reflecting conditions such as fender-benders, stalled vehicles, and minor roadway obstructions. Major incidents, while significantly less frequent, still represent a non-negligible portion of total traffic events. This reinforces the importance of addressing both volume and severity in traffic strategies, especially in areas or hours that show a mix of risk levels.

Hourly Breakdown of Incident Severity Levels

To analyze when the most serious incidents are likely to occur, a column chart was created to show incident severity by the hour. This visualization helps predict the severity profile of collision across different times of the day. Similar to previous visuals, the chart groups incidents into three severity levels minor, moderate, and major, and aligns them against a 24 hour clock. This allows for a side-by-side comparison of how the volume of each severity type fluctuates throughout the day, offering a time-aware lens into risk assessment. Such a breakdown can support targeted safety measures during periods when more severe collisions are historically more common.

Figure 9

Hourly Breakdown of Incident Severity Levels



The chart illustrates that moderate severity incidents are consistently the most common throughout the day, with a noticeable peak around 3 pm-6 pm. Major incidents, although significantly fewer, appear slightly more frequent during evening and nighttime hours, particularly between 9 pm and 1 am. This pattern may reflect riskier driving behavior such as speeding or impaired driving during late hours, when speeds are higher and traffic enforcement may be reduced.

Conclusion and Future Work

Summary of Findings

This study explored the relationship between real-time traffic conditions and collision patterns in Austin, Texas, using both machine learning and interactive visualizations. Through K-means clustering, collision incidents were divided into three distinct travel environments. These clusters helped contextualize incident behavior across different speeds, volumes, and times. This analysis revealed that congestion-related collisions occur most frequently, however

high-speed environments, though less common, are more likely to result in severe collisions. Temporal analyses further revealed that late afternoon and evening hours, particularly on Fridays and Saturdays, pose elevated collision risks. Moderate and minor incidents dominated the dataset overall, however major incidents spiked proportionally during late night hours, where faster travel and limited visibility are likely contributing factors. These findings were reinforced by using Python-based statistical analysis and Power BI visualizations, offering both granular and high level perspectives on how collisions occur in Austin.

Implications

The findings from this analysis have direct implications for urban mobility management, public safety, and policy development. Identifying the time periods and traffic conditions most associated with collisions allow city planners and traffic engineers to prioritize high risk time windows for various interventions such as law enforcement presence and adjusted timed signaling. The clustering of the traffic behaviors also suggest that context-aware traffic solutions, such as dynamic speed limits or targeted congestion relief strategies could be more effective. Additionally, categorizing incident severity and mapping across time introduces an important layer of risk profiling. This provides critical insight for emergency response planning and resource allocation, especially during late hours when fewer but more severe collisions occur.

Future Directions

While this project uncovered valuable trends, there are several ways it could be extended into future work. Incorporating weather conditions, road quality data, or event schedules could offer a more holistic view of external factors contributing to collision risk. Further enhancement of the machine learning component, including applying supervised models to predict severity, could also enhance the analytical depth.

Additionally, integrating geospatial analysis would allow for future researchers to assess not just *when*, but *where* collisions are likely to occur in Austin. This could support the creation of an interactive traffic risk heatmap that updates based on the time of day, cluster conditions, and severity likelihood. Moreover, with access to live traffic streams, a real time collision risk prediction system could be developed, bringing this research into practical deployment for cities seeking to minimize traffic incidents and smarter infrastructure.

Appendix

References

City of Austin, T.- data.austintexas.gov. (2021, September 9). *Radar traffic counts: Open data: City of austin Texas*. Data.AustinTexas.gov - The official City of Austin open data portal.
https://data.austintexas.gov/Transportation-and-Mobility/Radar-Traffic-Counts/i626-g7ub/about_data

City of Austin, T.- data.austintexas.gov. (2025, April 20). *Real-time traffic incident reports: Open data: City of austin Texas*. Data.AustinTexas.gov - The official City of Austin open data portal.
https://data.austintexas.gov/Transportation-and-Mobility/Real-Time-Traffic-Incident-Reports/dx9v-zd7x/about_data

Feature Descriptions

Feature Name	Description
Traffic Report ID	Unique identifier for each reported incident
Published Date	Date and time when the incident was officially reported
Issue Reported	Raw text describing the nature of the traffic incident

Location	General description of where the incident occurred
Latitude	Geographic latitude coordinate of the incident location
Longitude	Geographic longitude coordinate of the incident location
Address	Street level address provided in the incident report
Status	Current resolution status of the report
Agency	City department responsible for the report
Time_Hour	Rounded timestamp to the nearest hour
Volume	Number of vehicles detected at the location at time of the incident
Speed	Average speed of vehicles near the incident location (in mph)

Severity Categorization Mapping

Description Keywords	Description Keywords
Minor	“STALLED”, “BLOCKED”, “VEHICLE FIRE”, “HAZARD”
Moderate	“COLLISION”, “CRASH”, “ACCIDENT”, “AUTO/PED”
Major	“FATAL”, “INJURY”, “FTSRA”

Tools and Libraries

- Python: pandas, numpy, plotly.express, seaborn, sklearn, matplotlib
- Power BI: DAX, Power Query, clustered column chart, combo chart, pie chart, stacked bar chart
- Development Platforms: Jupyter Notebook (EDA & Modeling), Power BI Desktop

Figure Number	Description
1	Traffic Condition Clusters by Volume and Speed
2	Collision Frequency by Hour for each Traffic Cluster
3	Collision Density by Hour and Day of the Week
4	Speed Distribution of Collisions by Traffic Cluster
5	Total Number of Collisions per Traffic Cluster
6	Monthly Distribution of Traffic Incidents by Severity
7	Hourly Traffic Speed and Incident Volume Trends
8	Overall Distribution of Incident Severity Types
9	Hourly Breakdown of Incident Severity Levels

Supplemental Visualizations and Tables

The following visuals were explored during analysis but were not included in the main body due to redundancy or space constraints. They are retained here for reference and reproducibility.

Figure A1*Hourly Distribution of Incidents*

0	22431
1	18340
2	15467
3	12055
4	9410
5	7538
6	6755
7	7775
8	7005
9	5650
10	6605
11	11418
12	18379
13	22417
14	21817
15	20932
16	22327
17	24886
18	25394
19	24813
20	25263
21	27632
22	28724
23	25417

Figure A2*Weekly Distributions of Incidents*

Name: Hour, dtype: int64
Friday 66140
Thursday 62789
Wednesday 62042
Tuesday 60762
Monday 59870
Saturday 57695
Sunday 49152
Name: Weekday, dtype: int64

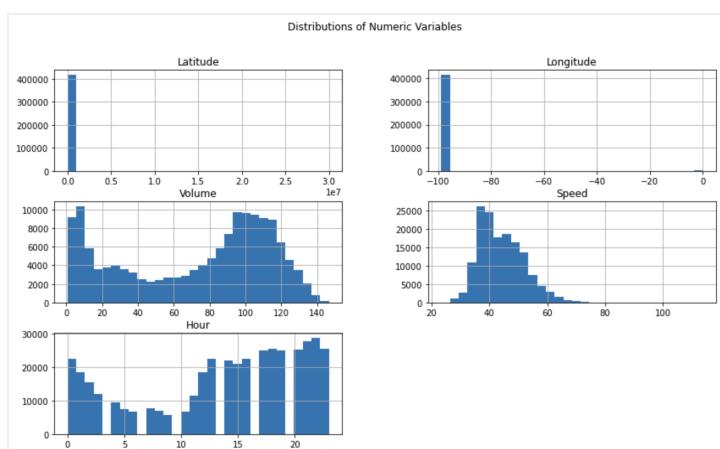
Figure A3*Distributions of Numeric Values*

Figure A4*Incident Category Counts*

Collision	213041
Hazard/Obstruction	168347
Vehicle Issue	19602
Fatality/Injury	11192
Other	6268
Name:	Issue_Category, dtype: int64

Figure A5*K-Means Cluster Centers*

Cluster Centers:			
	Volume	Speed	Hour
0	107.203528	37.064131	15.466552
1	22.353268	51.887231	5.484580
2	74.470187	45.533668	18.585769

Figure A6*Cluster Sizes by Label*

	Cluster_Name	Incident Count
0	High Congestion	28581
1	Rush Hour Flow	22950
2	High Speed / Low Volume	19880