UNIVERSITÀ DI PISA

Department of Computer Science
Artificial Intelligence
Human Language Technologies

Dawit Zemene Simie - [MAT. 682928]
Anson Johnson Madambi - [MAT. 681487]
Sounak Mukhopadhyay - [MAT. 684352]

Professor: Prof. Anna Monreale Professor: Prof. Mattia Setzu

Academic Year 2024/2025

Contents

Chapter 1

Data Understanding and Preparation

1.1 Data Overview

The Crash data are taken from the Chicago data portal and show information about each traffic crash on city streets within the City of Chicago; the dataset consists of 930701 row entries and a total attribute of 48 columns. The crashes_df dataset includes several key attributes that can significantly contribute to crash prediction models. The combined Chicago traffic crash dataset consists of three interconnected components—Crashes, Vehicles, and People—linked by a common identifier CRASH_RECORD_ID. The Crashes dataset captures the core event data, including the time, location (with GPS coordinates), environmental conditions, road features, speed limits, and injury counts. The Vehicles dataset supplements this by detailing each vehicle involved, such as its type, condition, movement, and damage, along with the actions and citations related to the driver. Meanwhile, the People dataset provides individual-level information about drivers, passengers, pedestrians, or cyclists involved in the crashes, including their demographics, injuries, safety equipment usage, and any signs of impairment. Together, these datasets allow for comprehensive incident reconstruction and risk assessment—enabling analysis of factors like high-risk areas, dangerous driver behaviors, and vehicle types most associated with severe outcomes. For instance, CRASH_DATE provides temporal patterns, such as rush hours or seasonal trends, while POSTED_SPEED_LIMIT may reveal the impact of speed regulations on crash likelihood. TRAFFIC_CONTROL_DEVICE, WEATHER_CONDITION, and LIGHTING_CONDITION capture environmental and situational factors that influence driver behavior and visibility. Variables like INJURIES_TOTAL and INJURIES_FATAL reflect the severity of the crash, which can help prioritize high-risk scenarios. DAMAGE indicates the material impact, which can be correlated with the intensity of the crash. LATITUDE and LONGITUDE support spatial analysis, enabling location-based risk assessments. BEAT_OF_OCCURRENCE adds a layer of administrative geography, helping identify patterns within specific enforcement zones. Finally, CRASH_TYPE and PRIM_CONTRIBUTORY_CAUSE offer direct insights into crash dynamics and causes, making them critical for targeted prevention strategies. Collectively, these characteristics are not only descriptive but can also serve as predictive variables in machine learning models aimed at forecasting and mitigating road crashes.

1.2 Variable Distribution Analysis

Here first we examined how key variables in the crash dataset are distributed:

1. Speed Limits The POSTED_SPEED_LIMIT variable showed a non-uniform distribution, with very clear peaks at common city speed limits such as 25, 30, and 35 mph. This indicates that most crashes happen in standard city-driving zones, as opposed to high-speed expressways. Extremely low or high speed limits (e.g., under 10 mph or over 50 mph) were rare, which suggests that those areas are either very safe or less traveled. The insight we have here is concentration

of crashes in mid-range speed zones aligns with areas of high pedestrian activity and stop-and-go traffic — common in residential and mixed-use zones.

2. Weather Conditions The WEATHER_CONDITION variable revealed that clear weather was the most frequent context for crashes. While this might seem surprising (we expect more accidents in rain or snow), it actually reflects exposure: drivers are on the road more often in good weather, increasing the chances of accidents during those periods.

Insight: Rather than bad weather causing more crashes, it seems that traffic volume is a more dominant factor. That said, a non-negligible number of crashes still occurred in rainy or snowy conditions, justifying a dedicated weather risk indicator.

3. Lighting Conditions The LIGHTING_CONDITION column indicated that most accidents occur in daylight, followed by dark conditions with street lights on. Few crashes happened in completely dark areas or under sunrise/sunset transitions.

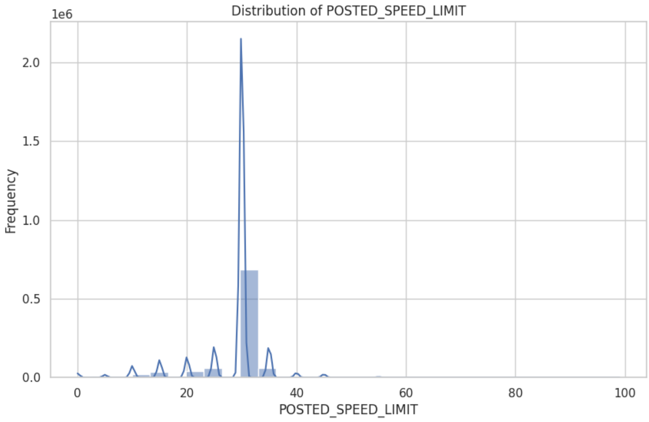
Insight: This once again highlights the role of traffic density. More people drive during the day, so naturally, more accidents occur during those hours — not necessarily because it's riskier, but because more cars are on the road.

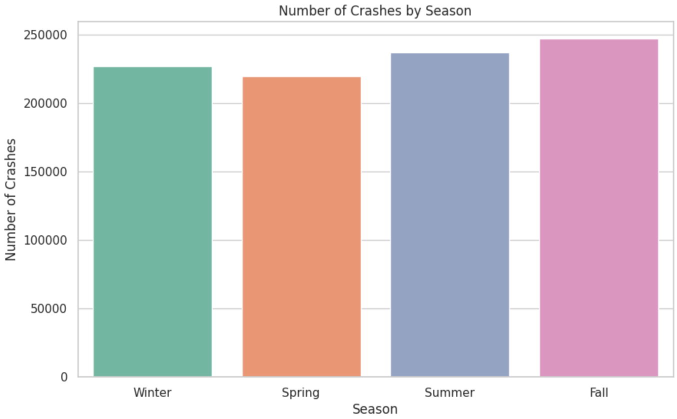
4. Injuries Injury-related fields (INJURIES_TOTAL, INJURIES_FATAL, etc.) showed a right-skewed distribution, meaning that:

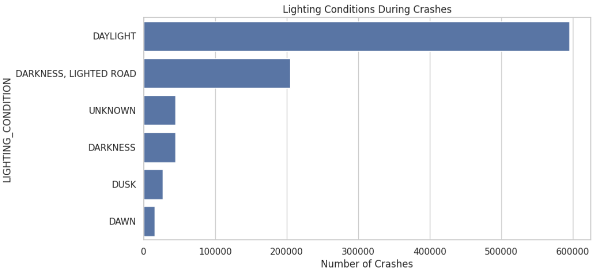
The majority of crashes caused 0–1 injuries.

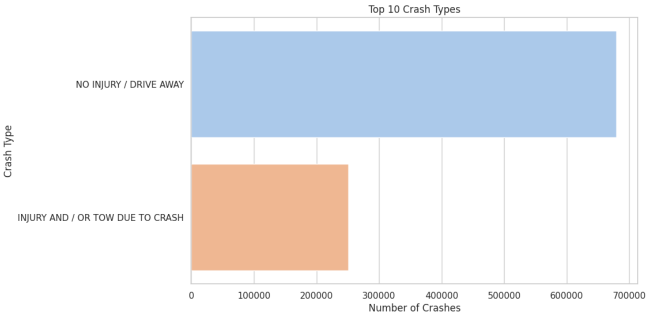
Only a small fraction involved severe or fatal injuries.

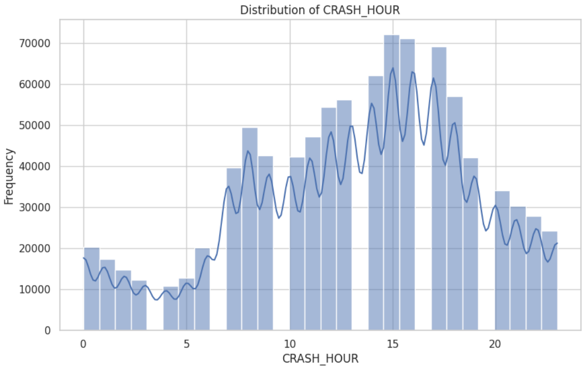
Insight: While most accidents are minor, the tail of the distribution (rare but severe cases) is especially important for insurance and emergency planning.

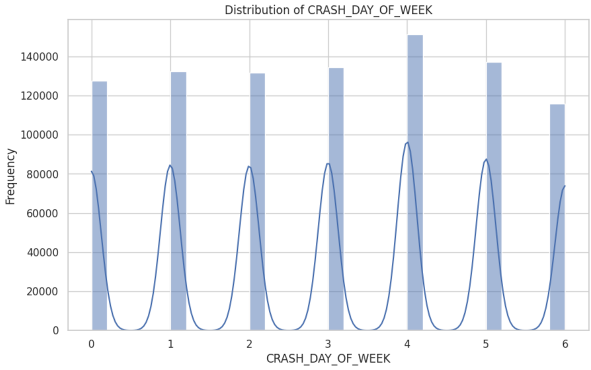






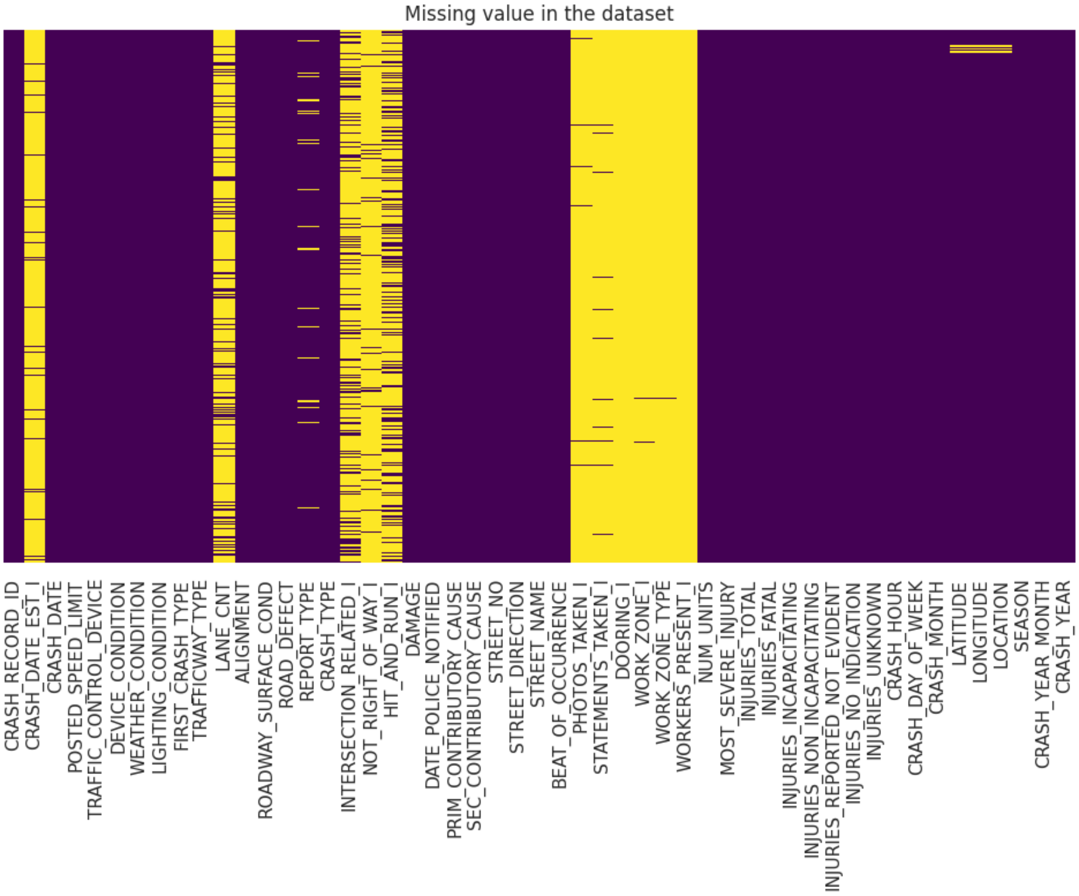






1.3 Data Quality Assessment

1.3.1 Missing values

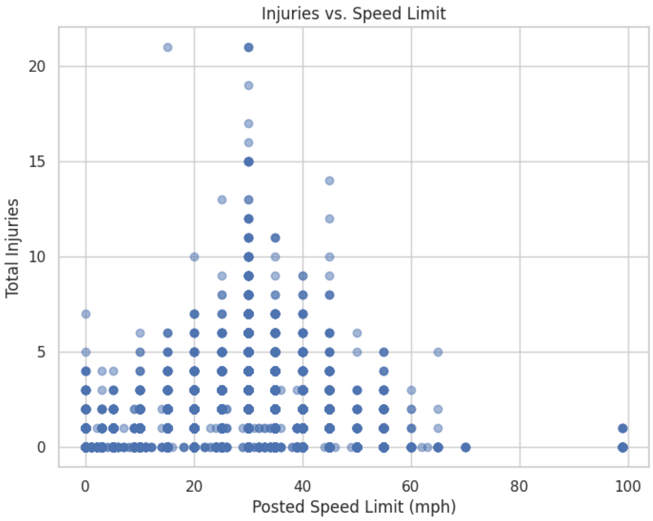


As seen from heatmap above each vertical stripe represents a column, and yellow marks indicate missing values. From the image, it's evident that while many key columns such as crash IDs, dates, and geographic coordinates are fully populated (shown in dark purple), a number of fields — particularly those related to road surface conditions, roadway features, and photo or statement records — exhibit substantial gaps. One column toward the center appears to be almost entirely missing. These insights are crucial because they guide our data cleaning strategy: variables with too much missing information may need to be dropped, while others with moderate gaps can be imputed or labeled as "Unknown" to retain their contextual value.

1.3.2 Handling Missing Values

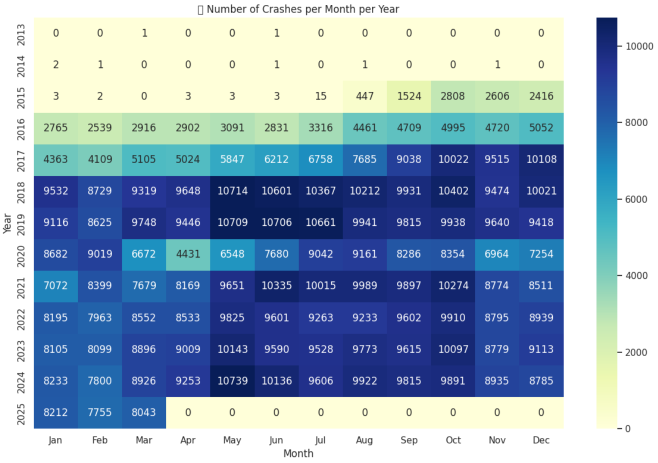
We have handled the missing values in the dataset using a combination of three well-established strategies tailored to the nature and importance of each variable. First, for categorical environmental attributes such as WEATHER_CONDITION and LIGHTING_CONDITION, we applied constant imputation by filling missing entries with the label "Unknown". This preserves all records while ensuring these variables remain interpretable and useful during analysis. Second, for numeric fields like POSTED_SPEED_LIMIT, we used the median value to impute missing entries. The median is a robust choice, particularly for skewed data, as it is less sensitive to extreme outliers than the mean and helps maintain a realistic distribution of values. Lastly, we dropped variables with excessive missingness, especially those whose names or context suggested low relevance to our predictive goals — for example, STATEMENTS_TAKEN or PHOTOS_TAKEN, which are administrative fields unlikely to influence crash severity or patterns. By doing so, we reduced noise and ensured our dataset remains focused on features that offer meaningful insight for downstream tasks like clustering and prediction.

1.3.3 Outliers



The plot suggests that moderate-speed zones (30–40 mph) — despite being safer than high-speed roads in theory — are where the most injuries occur, likely due to higher exposure and denser traffic. What stands out are a few vertical "spikes" in injury counts at moderate speed limits (e.g., 30 mph and 35 mph), where we observe incidents with 15 or even over 20 injuries. These are likely multi-vehicle collisions or crashes involving groups of pedestrians — rare but extremely impactful events.
Additionally, there's a data point at 100 mph with minimal injury, which appears to be a potential outlier and might reflect a data entry error or a highly unusual, controlled crash.

1.3.4 Pairwise Correlations



Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013	0	0	1	0	0	1	0	0	0	0	0	0
2014	2	1	0	0	0	1	0	1	0	0	1	0
2015	3	2	0	3	3	3	15	447	1524	2808	2606	2416
2016	2765	2539	2916	2902	3091	2831	3316	4461	4709	4995	4720	5052
2017	4363	4109	5105	5024	5847	6212	6758	7685	9038	10022	9515	10108
2018	9532	8729	9319	9648	10714	10601	10367	10212	9931	10402	9474	10021
2019	9116	8625	9748	9446	10709	10706	10661	9941	9815	9938	9640	9418
2020	8682	9019	6672	4431	6548	7680	9042	9161	8286	8354	6964	7254
2021	7072	8399	7679	8169	9651	10335	10015	9989	9897	10274	8774	8511
2022	8195	7963	8552	8533	9825	9601	9263	9233	9602	9910	8795	8939
2023	8105	8099	8896	9009	10143	9590	9528	9773	9615	10097	8779	9113
2024	8233	7800	8926	9253	10739	10136	9606	9922	9815	9891	8935	8785
2025	8212	7755	8043	0	0	0	0	0	0	0	0	0

Across nearly every year shown, there is a noticeable increase in crashes during the warmer months, particularly May through October. Months like June, July, and October consistently show the darkest shades, indicating the highest number of crashes. This likely reflects increased road usage during these months — due to better weather, summer travel, and higher pedestrian and cyclist activity.

In contrast, January and February consistently exhibit lower crash volumes, possibly due to fewer cars on the road in colder weather or more cautious driving in icy conditions. However, the drop is less pronounced in recent years, perhaps due to improved road infrastructure or changes in traffic behavior. tarting from 2016 onward, there's a significant jump in crash numbers, with peaks occurring between 2018 and 2019, where monthly crash counts regularly exceed 10,000. This surge could be related to increased vehicle registrations, population growth, or changes in reporting standards.

The year 2020 shows an unusual dip, especially around March and April, aligning with the early months of the COVID-19 pandemic lockdowns. Fewer cars on the road during that period resulted in historically low crash counts, clearly visible in the sharp drop for April 2020.

In the years following 2020, crash counts appear to rebound, with a gradual return to pre-pandemic levels by 2023–2024, though not quite reaching the peaks observed in 2018–2019.

Chapter 2

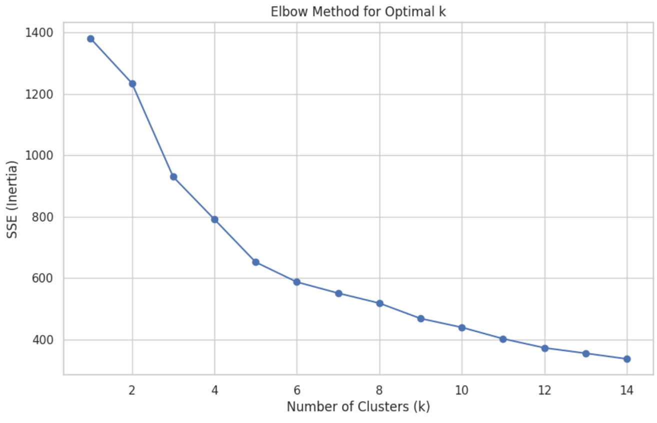
Clustering analysis

To prepare the crash data for clustering, the analysis focused on creating meaningful incident profiles for each police beat. These profiles were constructed by aggregating crash statistics and relevant features at a granular level, considering the day of the week and hour of the crash. Key features included total crashes, average injuries, average speed limit, the percentage of crashes influenced by weather (rain or snow), and the percentage occurring during nighttime conditions. This aggregation aimed to capture patterns and trends specific to each beat, enabling the identification of distinct crash characteristics.

For further improvement the profiles, feature engineering techniques were employed to derive more informative variables. A severity index was created by assigning weights to different injury types, providing a comprehensive measure of crash severity for each incident. In addition, time-of-day categories were introduced, classifying crashes into periods such as ”morning rush,” ”evening rush,” and ”nighttime” to account for potential temporal variations in risk. These engineered features aimed to enrich the profiles with relevant information for clustering.

Finally, to ensure data quality and model performance, missing values in the incident profiles were addressed using imputation techniques. Numerical features, such as the average severity and speed limit, were imputed using the mean, while categorical features like traffic control device and road condition were imputed using the most frequent category. This imputation step ensured complete data for clustering and minimized potential bias introduced by missing values. After these preparations, the data was scaled using StandardScaler to normalize features and prevent dominance by features with larger scales, ensuring fair and accurate clustering results.

2.1 K-Means



Prior to applying the clustering algorithm, the features within the beat profiles were normalized using StandardScaler. This crucial step ensured that all features contributed equally to the distance calculations during clustering, preventing features with larger scales, like total crashes, from disproportionately influencing the results. StandardScaler transforms each feature by subtracting its mean and dividing by its standard deviation, resulting in features with zero mean and unit variance. This normalization ensured that the clustering algorithm focused on the relative differences between beats rather than the absolute magnitudes of the features.

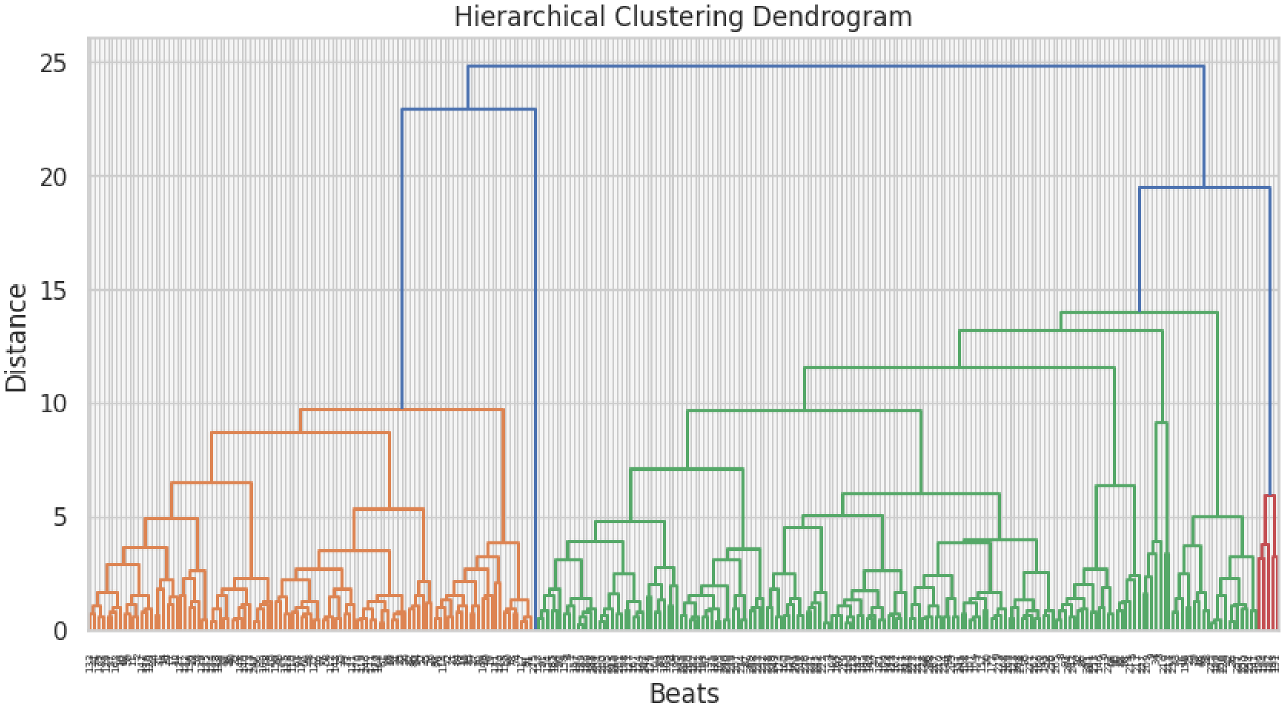
To determine the optimal number of clusters (k) for the KMeans algorithm, the elbow method was employed. This method involves calculating the sum of squared errors (SSE) for a range of k values and plotting the SSE against k. The elbow point on the resulting plot, where the rate of decrease in SSE slows down significantly, suggests an optimal number of clusters. This point represents a balance between minimizing within-cluster variance and avoiding excessive cluster granularity. By visually inspecting the elbow plot, an informed decision was made to select 4 as the most suitable k value for effectively grouping the police beats based on their crash characteristics, indicating the presence of four distinct crash risk profiles across the city.

2.2 Hierarchical clustering

Applying Hierarchical clustering to group police beats based on their crash characteristics using the Ward's linkage method. The resulting dendrogram provided a visual representation of the relationships between beats, highlighting natural groupings and potential clusters. Careful examination of the dendrogram, considering significant distance jumps and visual separation, suggested that four clusters were a sensible choice. This decision aligns with the elbow method applied to KMeans clustering, further strengthening its validity.

The selection of four clusters offers a balance between capturing significant crash patterns and maintaining interpretability for analysis and actionable insights. The clusters are expected to reveal distinct crash profiles across different police beats, enabling targeted interventions and strategies for improving road safety.

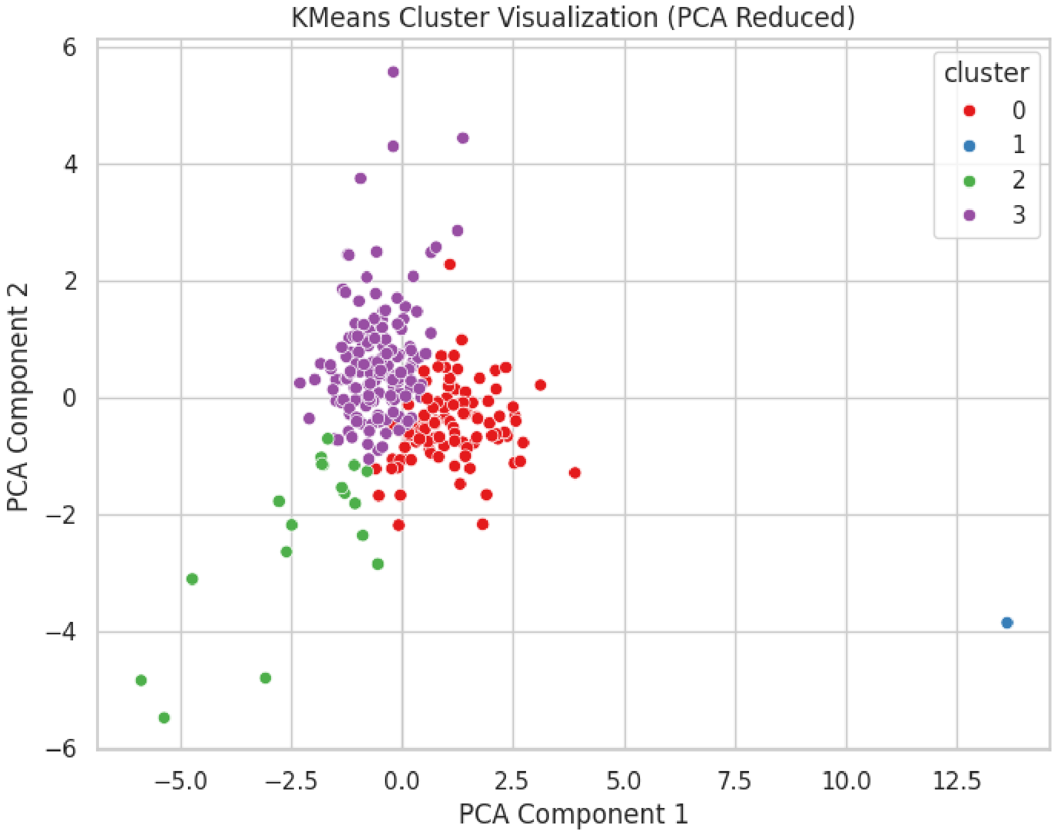
Further investigation will involve profiling each cluster based on its average crash attributes, mapping the clusters geographically to identify high-risk zones, and developing tailored interventions to address the specific crash patterns observed. This comprehensive approach, grounded in data-driven clustering, aims to advance our understanding of traffic crashes in Chicago and contribute to creating safer roads for all.



2.3 PCA

The scatter plot, resulting from PCA dimensionality reduction, visually showcases the classification of geographical areas (police beats) into distinct clusters based on their traffic crash risk profiles. Each cluster, distinguished by a unique color, reflects a specific combination of contributing factors, such as crash frequency, average injuries, speed limit, weather impact, and nighttime crash prevalence.

The blue cluster, as depicted in the visualization, represents one such grouping of police beats with shared risk characteristics. The spatial arrangement of points within this cluster indicates areas with similar crash patterns and contributing factors. By analyzing the features associated with the blue cluster – such as its average values for crash frequency, injury severity, etc. – we can understand the specific risk profile it represents.



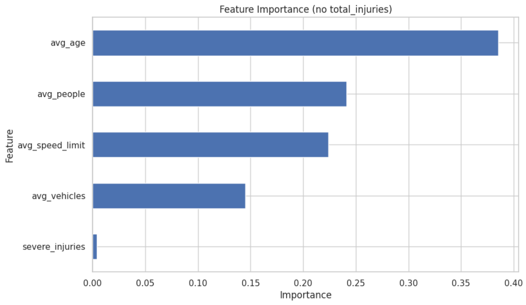
Chapter 3

Predictive Analysisg

3.1 Estimating Crash Damage Using Machine Learning

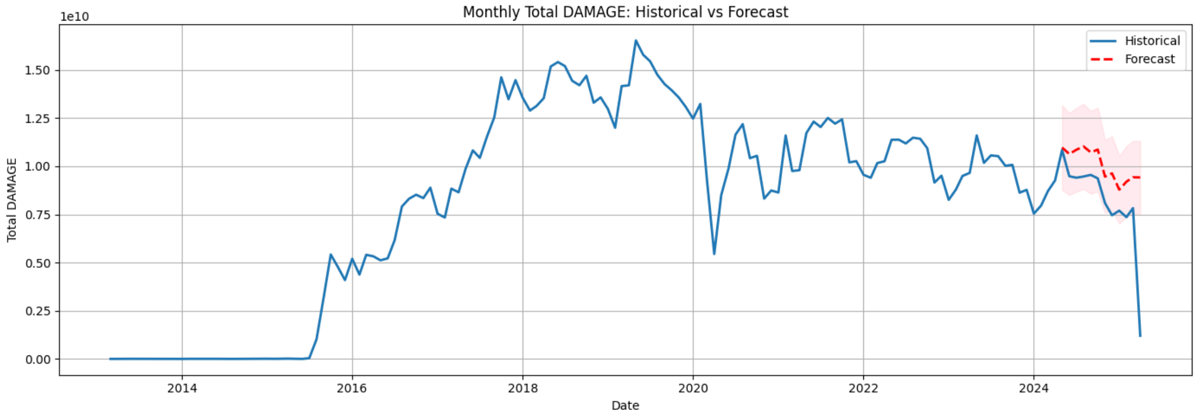
In the final phase of the project, a predictive analysis was carried out to estimate the average damage per crash using machine learning techniques. The goal was to model the financial impact of traffic incidents based on various contextual and behavioral indicators, thereby offering a valuable tool for insurance assessment and urban safety planning.

3.1.1 Methodology Overview

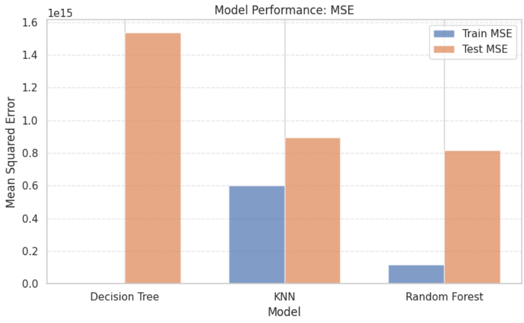


The analysis began with data merging and feature engineering, where multiple datasets — including crash details, vehicle records, and people involved — were integrated into a unified structure. This allowed for the creation of rich, aggregated features such as the average age of people involved, average speed limits, total injuries, number of vehicles and people involved, and counts of severe injuries. These engineered variables served as input features for the machine learning models, while the average damage per crash was selected as the target variable to be predicted.

Before training the models, the dataset was carefully split into training and testing sets to ensure robust performance evaluation. Feature scaling was applied specifically for the K-Nearest Neighbors (KNN) algorithm, which relies on distance metrics and is highly sensitive to differences in feature magnitudes.



3.1.2 Modeling and Evaluation



Three different models were implemented and compared: a Decision Tree, K-Nearest Neighbors, and a Random Forest. Each model was trained on the training data and then evaluated on the test set using Mean Squared Error as the performance metric. The results showed that the Random Forest model significantly outperformed the other two, achieving an MSE of approximately 713 million, compared to 1.2 billion for KNN and over 1.5 billion for the Decision Tree. This highlights the robustness of ensemble methods like Random Forest in capturing nonlinear relationships and interactions among features.

Chapter 4

Time Series Analysis

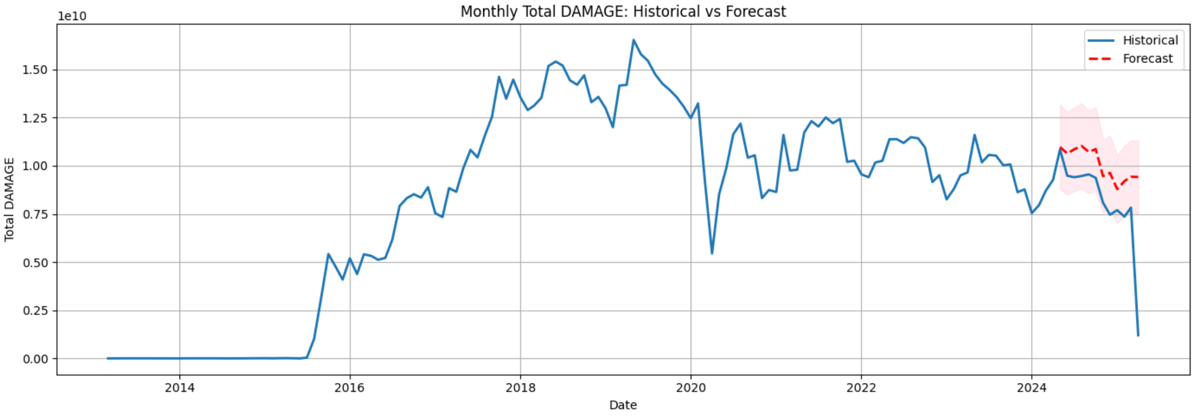
4.1 Forecasting Total Damage Using ARIMA, SARIMAX

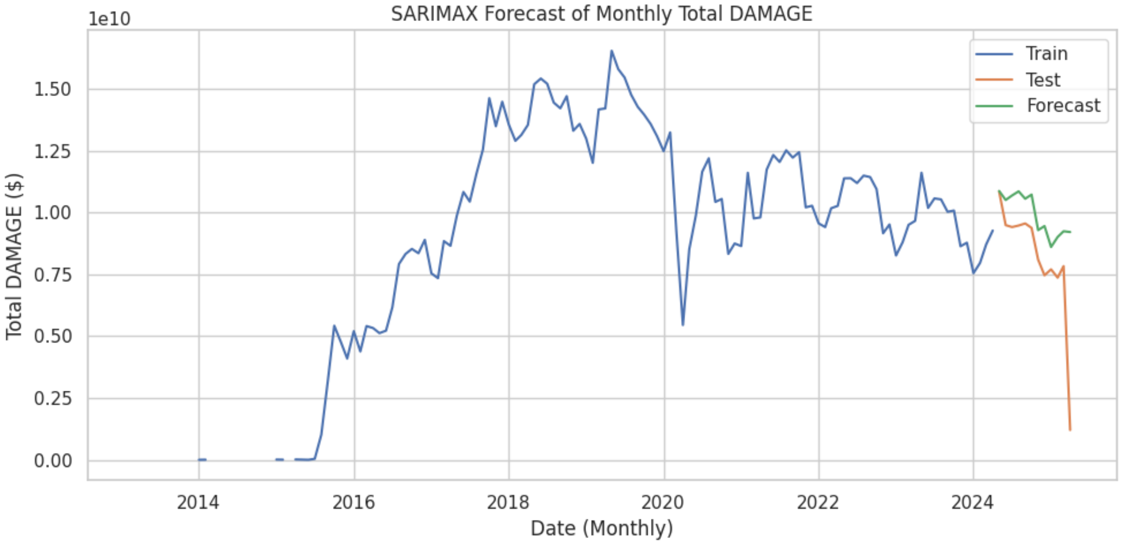
To better understand and anticipate trends in traffic-related financial loss, a time series forecasting approach was applied to the dataset. The process began with data aggregation, where crash records were grouped by date to compute the total reported damage per day. This transformation allowed the data to be treated as a continuous time series, laying the foundation for predictive modeling.

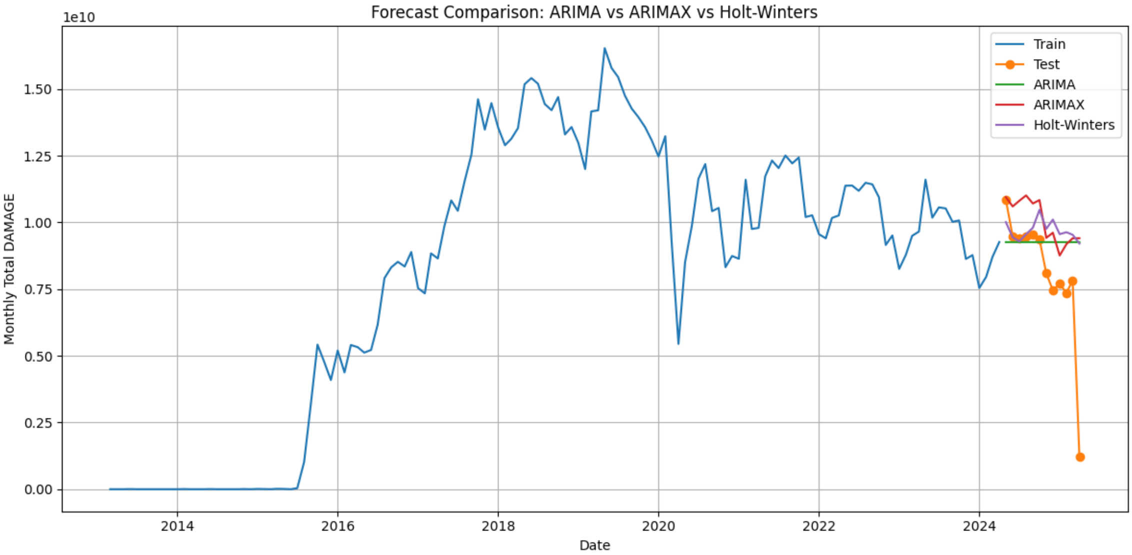
Three models were explored to forecast future damage values: ARIMA, SARIMAX, and Holt-Winters Exponential Smoothing. Among them, the SARIMAX model was particularly valuable because it supports the inclusion of exogenous variables — external factors that may influence the target variable, such as weather, average speed, or number of people involved in crashes. This allowed the forecasting process to go beyond mere trend extrapolation and account for real-world dynamics that contribute to variations in crash damage.

The models were trained on historical crash damage data, and forecasts were generated for the subsequent 12 months. For the SARIMAX model, relevant exogenous variables were selected and incorporated during training to improve forecast reliability. The resulting forecasts were then visualized alongside the actual historical damage values. These plots also included confidence intervals, providing an estimate of the uncertainty around future predictions and helping to identify potential periods of elevated risk.

Overall, each of the models successfully produced future projections, with SARIMAX offering the added advantage of context-aware predictions. These insights are particularly useful for planning insurance reserves, emergency response resources, or targeted road safety campaigns in anticipation of high-risk periods.







Based on the analyses, Random Forest was identified as a promising model for predicting damage costs associated with crashes. The models, including SARIMAX, were able to forecast city-wide total damage over time with acceptable accuracy. SARIMAX, with its ability to leverage exogenous factors, offers the potential for more nuanced and potentially more accurate forecasting.