Assessment 2

Traffic Hotspot Detection using K-Means and GMM Clustering Techniques

Subject: MSE806 Intelligent Transportation Systems

Submitted By: Butch Dela Cruz and Amarsanaa Togtokhbaatar

Submitted To: Arun Kumar

Table of Contents

Introduction

Background

Urban traffic congestion is a pervasive issue in modern cities, significantly impacting the quality of life, economic productivity, and environmental sustainability. As urban populations continue to grow, the demand for efficient traffic management systems becomes increasingly critical. Effective traffic management aims to optimize traffic flow, reduce travel time, decrease fuel consumption, and minimize emissions. However, the complexity of urban traffic networks, coupled with dynamic and often unpredictable traffic patterns, poses significant challenges to traffic managers and urban planners.

Purpose

The primary aim of this study is to identify traffic congestion hotspots within the Auckland region using advanced machine learning techniques. By leveraging data-driven, unsupervised learning models such as K-Means and Gaussian Mixture Models (GMM), we aim to uncover patterns and trends in Auckland traffic data that may not be immediately apparent through traditional analysis. Identifying these hotspots is beneficial for targeted interventions, infrastructure improvements, and strategic traffic management, ultimately leading to more efficient and sustainable urban mobility.

Objectives

The objectives of this study are as follows:

1. **Data Collection and Preprocessing**: Gather comprehensive traffic data from relevant sources, including traffic volume, road name, and GPS coordinate, and preprocess the data to ensure its quality and usability for clustering analysis.
2. **Application of KMeans Clustering**: Implement the KMeans clustering algorithm to partition the traffic data into distinct clusters, identifying areas with high congestion levels and understanding their characteristics.
3. **Application of GMM Clustering**: Apply the Gaussian Mixture Model clustering technique to the traffic data to capture more complex, probabilistic patterns and compare its effectiveness with the KMeans approach.
4. **Comparison and Evaluation**: Evaluate the performance and accuracy of KMeans and GMM clustering techniques in identifying traffic hotspots, considering factors such as cluster coherence, interpretability, and computational efficiency.
5. **Visualization**: Design and develop effective visualization methods including interactive charts to display identified hotspots on map.

Through this study, we aim to contribute valuable insights into Auckland traffic management, demonstrating the potential of clustering techniques in addressing one of the most pressing challenges faced by modern cities.

Literature Review

Traffic congestion is a critical challenge in urban areas, adversely affecting economic productivity, environmental health, and overall quality of life. The integration of advanced technologies, particularly artificial intelligence (AI) algorithms and machine learning (ML) frameworks, is revolutionizing the field of Intelligent Transportation Systems (ITS). These innovations are instrumental in identifying congestion hotspots, enabling more efficient traffic management. This review examines the existing studies on these topics, with a particular focus on the diverse methodologies and technologies employed for detecting and managing traffic congestion through ML techniques.

Methods and Techniques

- **Statistical Methods**: These methods involve analyzing historical traffic data to identify patterns indicative of congestion. Techniques such as regression analysis and clustering are commonly used. For example, analyzing traffic volume data using regression models to predict congestion periods has been widely adopted (*How Can Regression Analysis Identify Traffic Trends?*, 2024).
- **Machine Learning Approaches**: Machine learning models, including supervised and unsupervised learning, have been employed to detect congestion hotspots. Techniques such as Support Vector Machines (SVM), Random Forests, and Neural Networks are utilized for predicting congestion based on traffic flow data (*The Machine Learning Framework for Traffic Management in Smart Cities: Optimizing Flows for a Sustainable Future | LinkedIn*, 2024).
- **Spatio-Temporal Analysis**: This involves the use of geographic information systems (GIS) and spatio-temporal data to detect congestion hotspots. Combining GPS data from vehicles with GIS to identify areas with frequent traffic slowdowns is a common approach (Ding et al., 2020).
- **Crowdsourcing and social media**: Leveraging data from social media and crowdsourcing platforms to detect real-time congestion hotspots is another innovative method. Analyzing Twitter stream data to identify real-time traffic is proposed by D'Andrea and his team, achieved 95.75% accuracy (D'Andrea et al., 2015).

Technologies

- **Sensors and IoT**: The deployment of traffic sensors and Internet of Things (IoT) devices to collect real-time traffic data is prevalent. Using inductive loop sensors and CCTV cameras to monitor traffic flow exemplifies this technology. Sonnleitner and his team proposed a system that applies raw video data from CCTV to a pretrained neural network framework, YOLO, which is a real-time object detection and classification suite, to count and classify passing vehicles. (Sonnleitner et al., 2020).

- **Mobile Data**: Utilizing data from mobile devices and GPS to monitor traffic patterns is also common. Aggregating anonymized GPS data from smartphones to detect congestion is a typical application (Ricciato et al., 2015).

Clustering Techniques in Traffic Management

- **K-Means Clustering**: A popular algorithm for partitioning data into clusters based on similarity. It is used to identify groups of congestion points, such as grouping intersections with similar congestion patterns.
- **Gaussian Mixture Model (GMM):** Probabilistic model that assume all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMMs are a form of soft clustering, where each data point can belong to multiple clusters with certain probabilities.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Effective for identifying clusters of varying shapes and sizes, especially in dense urban traffic data. Detecting irregular traffic congestion patterns that do not conform to a single shape demonstrates its application.
- **Hierarchical Clustering**: This technique builds a hierarchy of clusters and is useful for understanding the structure of traffic congestion. Creating a dendrogram to visualize the hierarchical relationship between different congestion hotspots is an example.
- **Fuzzy C-Means Clustering**: This allows each data point to belong to multiple clusters, providing a more nuanced understanding of traffic congestion. Identifying zones with varying degrees of congestion rather than distinct hotspots illustrates this method.

Proposed Clustering Algorithms

K-Means Clustering

K-Means Clustering is a widely used unsupervised machine learning algorithm for partitioning a dataset into K distinct, non-overlapping clusters. Each cluster is defined by its centroid, and the algorithm works iteratively to minimize the variance within each cluster.

Principles:

1. **Initialization**: Select K initial centroids randomly.
2. **Assignment**: Assign each data point to the nearest centroid based on the Euclidean distance.
3. **Update**: Calculate the new centroids by taking the mean of all data points assigned to each cluster.
4. **Iteration**: Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.

Applications in Traffic Data Analysis:

- **Identifying Congestion Zones**: K-Means can be used to group geographic locations (such as intersections) with similar traffic congestion patterns, helping to identify recurring congestion zones.
- **Traffic Flow Patterns**: By clustering traffic flow data, K-Means can reveal common patterns in vehicle movement, aiding in the development of traffic management strategies.

- **Temporal Analysis**: K-Means can cluster traffic data based on time intervals to identify peak congestion times and develop targeted solutions.

Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) are probabilistic models that assume all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMMs are a form of soft clustering, where each data point can belong to multiple clusters with certain probabilities.

Principles:

1. **Initialization**: Start with initial guesses for the parameters (means, covariances, and mixing coefficients) of the K Gaussian distributions.
2. **Expectation Step (E-Step)**: Calculate the probability that each data point belongs to each Gaussian distribution.
3. **Maximization Step (M-Step)**: Update the parameters of the Gaussian distributions to maximize the likelihood of the data given these assignments.
4. **Iteration**: Repeat the E-Step and M-Step until the model parameters converge.

Applications in Traffic Data Analysis:

- **Modeling Traffic Density**: GMM can model the distribution of traffic density across different regions, identifying areas with varying levels of congestion.
- **Anomaly Detection**: By understanding the normal distribution of traffic flow, GMM can help detect anomalies, such as unexpected traffic surges or drops.
- **Predictive Analysis**: GMM can be used to predict traffic conditions by modeling the probability distribution of traffic patterns over time.

Comparative Analysis

K-Means vs. GMM:

1. **Cluster Shape**: K-Means assumes clusters are spherical and equally sized, while GMM can model clusters with different shapes and sizes due to its probabilistic nature.
2. **Assignment**: K-Means assigns each data point to a single cluster, whereas GMM assigns probabilities to each data point belonging to multiple clusters.
3. **Complexity**: K-Means is computationally simpler and faster compared to GMM, which involves more complex calculations in the E-Step and M-Step.

Practical Considerations:

- **Scalability**: K-Means is more scalable for large datasets, making it suitable for real-time traffic data analysis.
- **Flexibility**: GMM provides more flexibility in modeling complex traffic patterns and distributions, which can be beneficial for detailed traffic studies and anomaly detection.

Both K-Means and GMM offer valuable insights into traffic data analysis, each with its strengths and limitations. The choice between them depends on the specific requirements of the traffic analysis task, such as the need for scalability, accuracy, and the nature of the traffic data being analyzed.

Methodology

Data Collection

Traffic count data plays a pivotal role in modern transportation planning, facilitating insights into traffic volumes, aiding in road design, prioritizing network improvements, assessing road safety risks, and evaluating the efficacy of previous interventions. Auckland Transport (AT) conducts a robust traffic flow counting program throughout the Auckland region, providing comprehensive data crucial for informed decision-making.

Data Source

The primary data source for this study is the Auckland Transport Traffic Count dataset, which includes traffic volume information collected from various sites within the Auckland region. The dataset covers the period from July 2012 to May 2024, and is available for download in XLSX format at https://at.govt.nz/about-us/reports-publications/traffic-counts. This data used in our study is intended solely for educational purposes within our Intelligent Transportation Systems (ITS) project and is not to be redistributed.

Description of Traffic Count Data

The traffic count dataset provides detailed information for each counting site, organized by road name, and further categorized by specific location details. Key columns within the dataset include:

- **Road name:** Name of the road where the traffic count was conducted.
- **Description:** Exact location description, including GPS coordinates since February 9, 2018.
- **Count start date:** Date when the 7-day traffic count period started.
- **5-day ADT:** Average Daily Traffic (vehicles per day) based on a Monday to Friday week.
- **7-day average:** Average Daily Traffic (vehicles per day) based on a Monday to Sunday week.
- **Sat:** Total number of vehicles counted on Saturday.
- **Sun:** Total number of vehicles counted on Sunday.
- **Direction:** Indicates the side of the road where the count was conducted.
- **AM peak volume:** Volume of traffic during the morning peak hour.
- **PM peak volume:** Volume of traffic during the afternoon peak hour.
- **Midday peak volume:** Volume of traffic during the midday peak hour.
- **HCV%:** Percentage of heavy commercial vehicles (HCV) among total vehicles counted.

Data Integrity and Usage Advancements

The Auckland Transport traffic count dataset is carefully curated to offer dependable insights, though it is provided "as is" without a warranty of accuracy. AT's dedication to data integrity ensures its reliability (Auckland Transport, 2024). While users are encouraged to complement this resource with others for informed decision-making, we find this dataset particularly suitable for our study on effectively identifying hotspots using K Means Clustering and Gaussian Mixture Models (GMM). This suitability arises from its extensive coverage across Auckland, detailed traffic volume metrics, and consistent data collection practices over time. It is important to note that fluctuations in traffic flow counts, influenced by factors like equipment sensitivity, counting methods, congestion, and

seasonal variations, highlight the dataset's role as a reliable approximation of traffic volumes at specific counting sites (Auckland Transport, 2024).

Data Preprocessing

The preprocessing of traffic count data from Auckland Transport (AT) involves several critical steps to ensure data reliability and suitability for subsequent analysis. This section outlines the systematic approach taken to clean and prepare the data for clustering analysis using Python.

Filtering Coordinates and Location Extraction

First, to focus exclusively on traffic data within the Auckland region, coordinates are filtered based on predefined geographical boundaries using latitude and longitude. This ensures that only data points within the city limits are included in the analysis. The filtering process helps maintain the geographic integrity of the dataset and eliminates outliers that may skew the analysis. Coordinates extracted from the traffic count descriptions are converted from NZTM (New Zealand Transverse Mercator) projection to the standard WGS84 (latitude, longitude) format. This conversion is essential for spatial analysis and visualization, aligning the dataset with widely used geographic information systems (GIS) standards.

Data Cleaning and Feature Engineering

The raw traffic count dataset is initially loaded from a CSV file, and columns that are entirely null are removed to streamline subsequent operations. The 'Count Start Date' column, which denotes the beginning of each traffic count period, is converted to datetime format. This conversion facilitates temporal analysis and ensures consistency in handling date-related operations. To enhance the dataset's utility, the average daily traffic (ADT) metrics are computed for weekends (Saturday and Sunday), consolidating them into a single 'Weekend Traffic ADT' feature. This feature provides a comprehensive view of traffic patterns over the entire week, accounting for variations between weekdays and weekends.

Standardization

To prepare the data for clustering algorithms such as K Means Clustering or Gaussian Mixture Models (GMM), standardization is applied to selected traffic volume metrics. This step involves scaling the data to a common mean and standard deviation, ensuring that each feature contributes equally to the clustering process. Missing values, if present, are handled by imputing zeros before scaling to maintain data integrity and consistency across all features. The final preprocessed dataset, now standardized and cleaned, is ready for advanced analytical techniques to uncover spatial and temporal patterns in Auckland's traffic flow. This rigorous preprocessing methodology ensures that the subsequent clustering analysis is based on robust, reliable data, facilitating informed decision-making in transportation planning and infrastructure development.

Outlier Detection and Removal

Outlier detection was also performed using Z-score analysis to enhance the reliability of the traffic count dataset used for subsequent clustering analyses. Z-scores were calculated for key traffic volume metrics such as '5-day ADT', 'Weekend Traffic ADT', 'AM peak volume', 'Midday peak volume', and 'PM peak volume', with a predefined threshold of 2 standard deviations to identify outliers.

Choosing a threshold of 2 standard deviations (2 SD) for outlier detection using Z-scores is based on statistical norms. It identifies data points that significantly deviate from the mean of the dataset. This threshold strikes a balance: it is sensitive enough to detect outliers that lie far outside normal variation but robust enough to avoid removing valid, albeit unusual, observations. This approach is widely accepted in statistical analysis for its clarity and effectiveness in identifying extreme values across diverse types of datasets (Z score for Outlier Detection – Python, 2024). Data points exceeding this threshold in any selected metric were removed from the dataset to mitigate the influence of extreme values on clustering results.

Filtered Data Shape

After completing the preprocessing steps, the traffic count dataset from Auckland Transport (AT) is refined to focus on relevant data points within the Auckland region. The filtered dataset, ready for subsequent analysis, exhibits the following characteristics:

- The filtered dataset comprises 9,289 rows and 26 rows.

This shape reflects the comprehensive nature of the dataset, encompassing a wide array of traffic volume metrics and spatial coordinates. The stringent filtering based on geographic boundaries and data quality criteria ensures that the dataset is tailored for accurate and insightful analysis.

Identifying Traffic Hotspots Using K-means Clustering

Traffic hotspots play a crucial role in urban planning and transportation management, as they represent areas with significant congestion or high traffic volumes that require targeted interventions. In this study, K-means clustering was employed as a powerful tool to identify and categorize traffic hotspots within the Auckland region based on detailed traffic count data collected by Auckland Transport (AT).

Algorithm Overview and Application

K-means clustering is an unsupervised machine learning algorithm that partitions a dataset into k clusters, where each data point is assigned to the cluster with the nearest centroid. To determine the optimal number of clusters k, the Elbow method was applied (Elbow Method for optimal value of k in KMeans, 2023). This method evaluates the within-cluster sum of squares, known as inertia, across a range of k values to find the point where adding more clusters provides diminishing returns in terms of explaining the variance in the data.

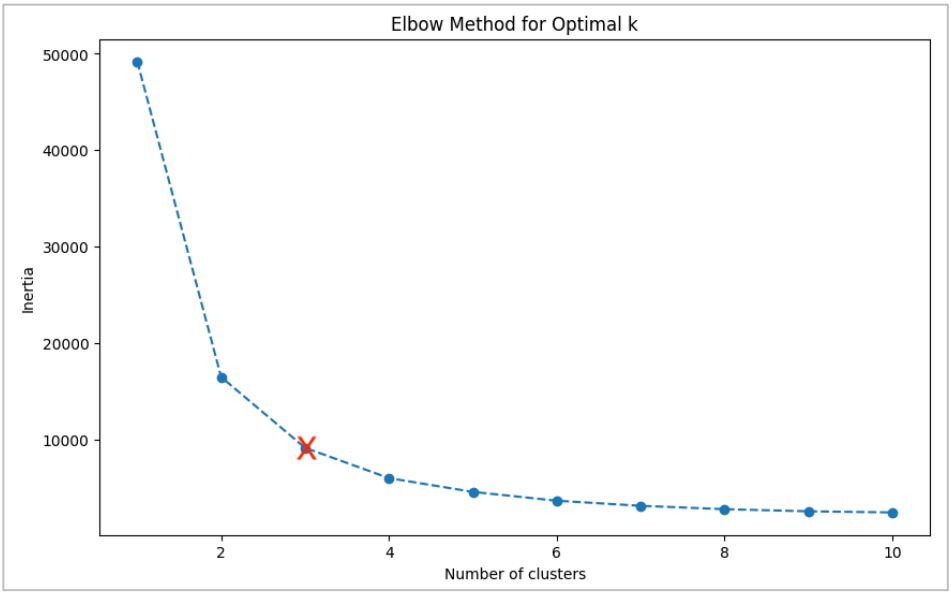

Figure 1: Elbow Method to find Optimal K for K-Means Clustering

The resulting plot from the Elbow method indicated that k=3 clusters were optimal for segmenting the traffic data effectively Figure 1: Elbow Method to find Optimal K for K-Means Clustering. Subsequently, K-means clustering was applied using k=3.

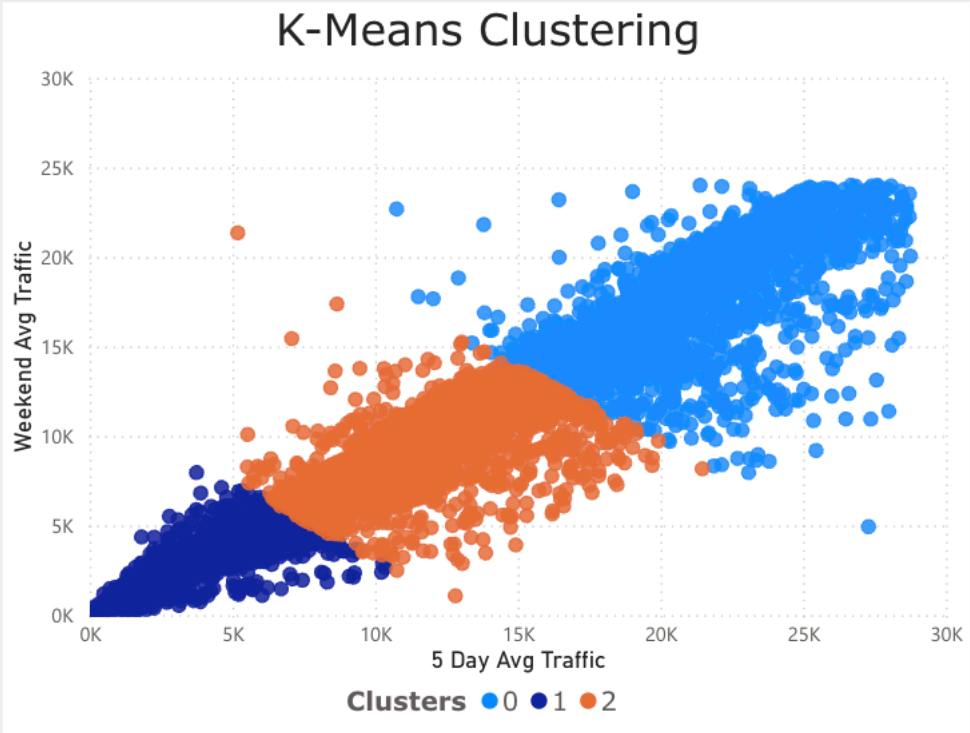

Figure 2: Scatter plot of the K-Means Clustering Results

Interpretation and Traffic Hotspot Identification

Based on the K-means clustering results using 5-Day Average Traffic (5 Day Avg Traffic) and Weekend Average Daily Traffic (Weekend Traffic ADT) Figure 2: Scatter plot of the K-Means Clustering Results, the interpretation reveals distinct patterns across different clusters of traffic locations within Auckland (Table 1: Interpretation of K-Means Clustering Results).

Cluster	Weekdays Avg Traffic	Weekends Avg Traffic	Interpretation
0	20,622.97	16,918.85	Represents areas with relatively high traffic volumes during weekdays and moderate traffic on weekends.
1	2,861.29	2,247.44	Indicates areas with moderate traffic volumes both during weekdays and weekends.
2	11,345.87	8,959.24	Shows areas with high traffic during weekdays and reduced traffic levels on weekends.

Table 1: Interpretation of K-Means Clustering Results

These clustering insights are instrumental in guiding traffic management strategies and urban planning initiatives. For Cluster 1, interventions may focus on alleviating congestion during peak hours through targeted traffic management measures and infrastructure improvements. In contrast, areas in Clusters 2 and 3 could benefit from policies that support sustainable urban mobility, such as promoting alternative transportation modes or enhancing local road networks to accommodate fluctuating traffic demands.

Identifying Traffic Hotspots Using Gaussian Mixture Models (GMM)

Traffic hotspots are critical focal points in urban planning and transportation management, highlighting areas with significant congestion or high traffic activity that require targeted interventions. In this analysis, Gaussian Mixture Models (GMM) were employed to identify and classify traffic hotspots within urban areas based on comprehensive traffic data collected by Auckland Transport (AT).

Algorithm Overview and Application

Gaussian Mixture Models (GMM) are probabilistic models that assume the data points are generated from a mixture of several Gaussian distributions. Unlike K-means clustering, which partitions data into distinct clusters based on proximity to centroids, GMM assigns probabilities to data points belonging to each cluster, accommodating more complex data distributions and potential overlap between clusters.

To determine the optimal number of clusters (components) for GMM, metrics such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) were utilized (Gaussian Mixture Model Selection, n.d.). These metrics assess model fit while penalizing complexity, guiding the selection of the most suitable number of clusters for interpreting traffic patterns.

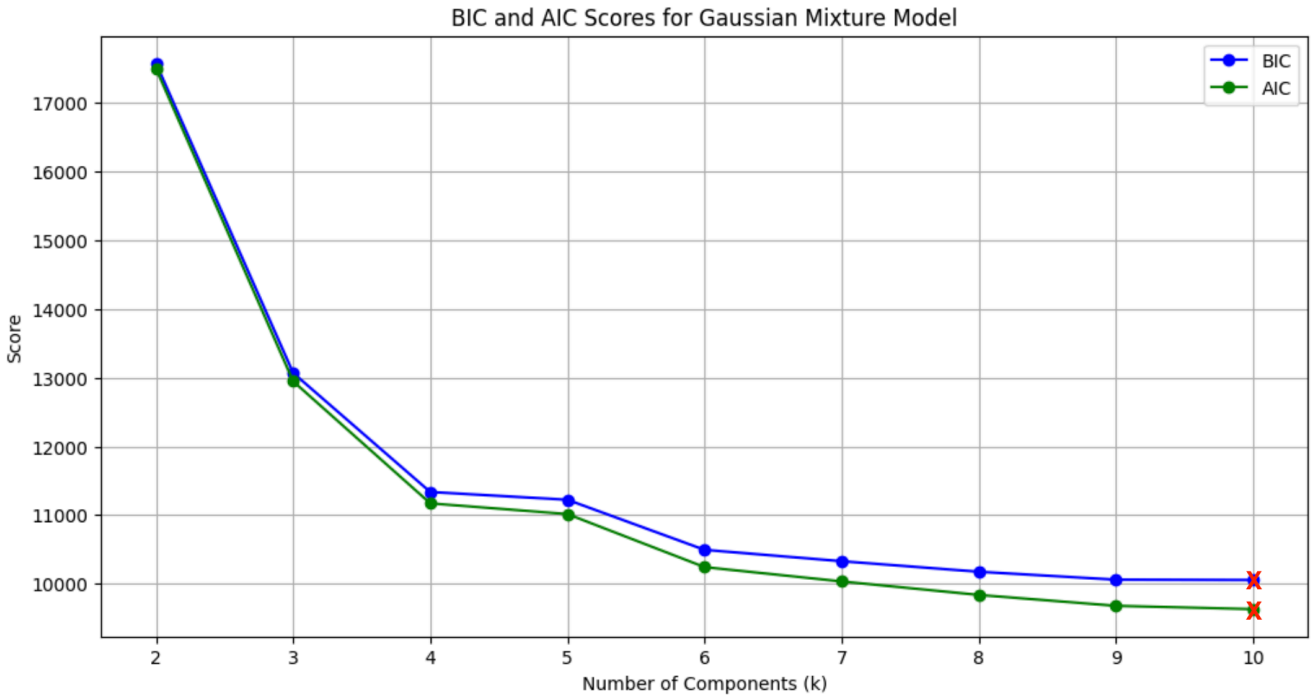

Figure 3: BIC and AIC Scores to find optimal k in GMM

The highest resulting BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) scores, obtained with 10 components in the Gaussian Mixture Model (GMM) analysis, suggest that a model with greater complexity is favored for clustering traffic data. Subsequently, GMM clustering was applied using k=10, optimizing the model for interpreting traffic dynamics.

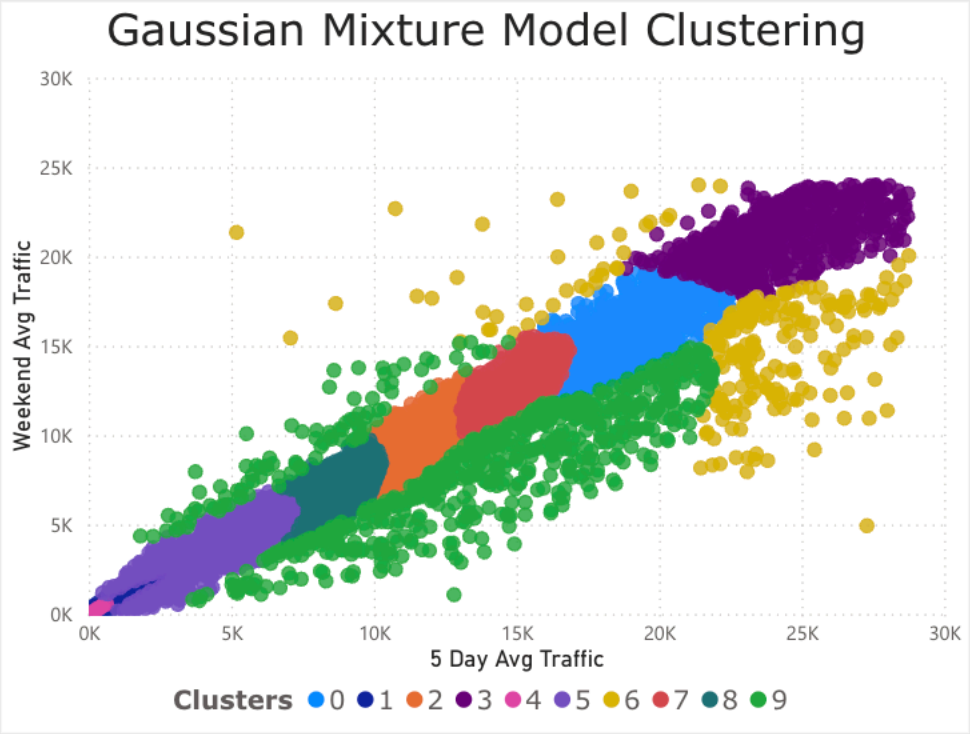

Figure 4: Visualization illustrating the results of Gaussian Mixture Models (GMM) clustering.

Interpretation and Traffic Hotspot Identification

The interpretation of GMM clusters based on the provided 5 Day Avg Traffic and Weekend Avg Traffic data (Figure 4: Visualization illustrating the results of Gaussian Mixture Models (GMM) clustering.) involves understanding the characteristics and patterns represented by each cluster (

Cluster	Weekdays Avg Traffic	Weekends Avg Traffic	Interpretation
0	18,872.83	15,750.55	Represents areas with consistently high traffic levels, likely urban or commercial centers.
1	1,521.76	1,236.27	Characterized by relatively low traffic volumes, indicating quieter residential or less frequented areas.
2	11,529.26	9,557.61	Represents mixed-use areas with varied traffic intensities.
3	24,086.93	20,911.50	Shows very high traffic volumes, possibly major highways or city centers.
4	263.60	230.67	Displays very low traffic volumes, possibly isolated or peripheral areas.
5	4,733.42	3,833.59	Exhibits moderate traffic volumes, likely near commercial centers or popular destinations.
6	22,725.75	15,430.96	Indicates areas with high traffic during weekdays and reduced traffic levels on weekends.
7	14,811.03	12,597.86	Displays moderate traffic volumes, possibly in areas with weekday-focused activities such as business districts.
8	8,728.26	7,044.69	Shows moderate traffic volumes, possibly in suburban or residential areas.
9	13,555.32	8,509.75	Represents areas with moderate traffic levels, possibly urban areas with consistent activity patterns.

Table 2: Interpretation of GMM Clustering Results).

Cluster	Weekdays Avg Traffic	Weekends Avg Traffic	Interpretation
0	18,872.83	15,750.55	Represents areas with consistently high traffic levels, likely urban or commercial centers.
1	1,521.76	1,236.27	Characterized by relatively low traffic volumes, indicating quieter residential or less frequented areas.
2	11,529.26	9,557.61	Represents mixed-use areas with varied traffic intensities.
3	24,086.93	20,911.50	Shows very high traffic volumes, possibly major highways or city centers.
4	263.60	230.67	Displays very low traffic volumes, possibly isolated or peripheral areas.
5	4,733.42	3,833.59	Exhibits moderate traffic volumes, likely near commercial centers or popular destinations.
6	22,725.75	15,430.96	Indicates areas with high traffic during weekdays and reduced traffic levels on weekends.
7	14,811.03	12,597.86	Displays moderate traffic volumes, possibly in areas with weekday-focused activities such as business districts.
8	8,728.26	7,044.69	Shows moderate traffic volumes, possibly in suburban or residential areas.
9	13,555.32	8,509.75	Represents areas with moderate traffic levels, possibly urban areas with consistent activity patterns.

Table 2: Interpretation of GMM Clustering Results

By understanding these distinct traffic patterns, city planners and policymakers can make informed decisions to optimize transportation efficiency, enhance infrastructure resilience, and improve overall quality of life for residents across different areas of Auckland. These findings underscore the importance of data-driven approaches in urban planning, ensuring that interventions are tailored to meet the specific needs and challenges of diverse traffic environments within a metropolitan region.

Comparison Metrics

Metric	K-means Clustering	Gaussian Mixture Models (GMM)
Accuracy of Clustering	Measures data point assignment based on centroid proximity.	Assesses model fit to data using Gaussian distributions.
Interpretability of Clusters	Provides clear boundaries for clusters.	Offers probabilistic cluster assignments.
Handling of Cluster Shape and Variance	Assumes spherical clusters with equal variance.	Models clusters as Gaussian distributions, accommodating non-spherical shapes and varying variances.
Scalability	Scales well with large datasets and is computationally efficient.	Can be slower and less scalable, especially with increasing Gaussian components.
Robustness to Outliers	Sensitive to outliers due to centroid distance calculations.	Less affected by outliers due to probabilistic assignments.
Computational Efficiency	Generally faster due to simpler calculations.	Slower due to complex Gaussian parameter computations.

Model Selection Criteria	Uses inertia and the Elbow method to determine optimal clusters.	Uses BIC or AIC to determine optimal Gaussian components.
Application Suitability	Suitable for well-separated clusters and spherical data distributions.	Suitable for complex data distributions and overlapping clusters.

The choice between K-means clustering and Gaussian Mixture Models (GMM) depends on the specific characteristics of the dataset and the goals of the analysis. K-means is advantageous for its simplicity, speed, and clear cluster boundaries but assumes spherical clusters with equal variance. On the other hand, GMM offers flexibility in modeling complex data distributions and probabilistic cluster assignments, making it suitable for scenarios where data distributions are non-spherical, or clusters overlap. Consider these factors carefully when selecting the appropriate clustering method for urban planning, traffic management, or any similar applications.

Data Visualization Techniques for Traffic Hotspot Detection

Data visualization plays a crucial role in understanding and communicating traffic patterns identified through clustering algorithms like K-means and Gaussian Mixture Models (GMM). In this study, Power BI was utilized to create informative and insightful visualizations (Microsoft, 2024), aiding in the identification and analysis of traffic hotspots within the Auckland region (Figure 5: Power BI Dashboard: Visualizing Traffic Hotspot Detection with K-means and GMM Clustering).

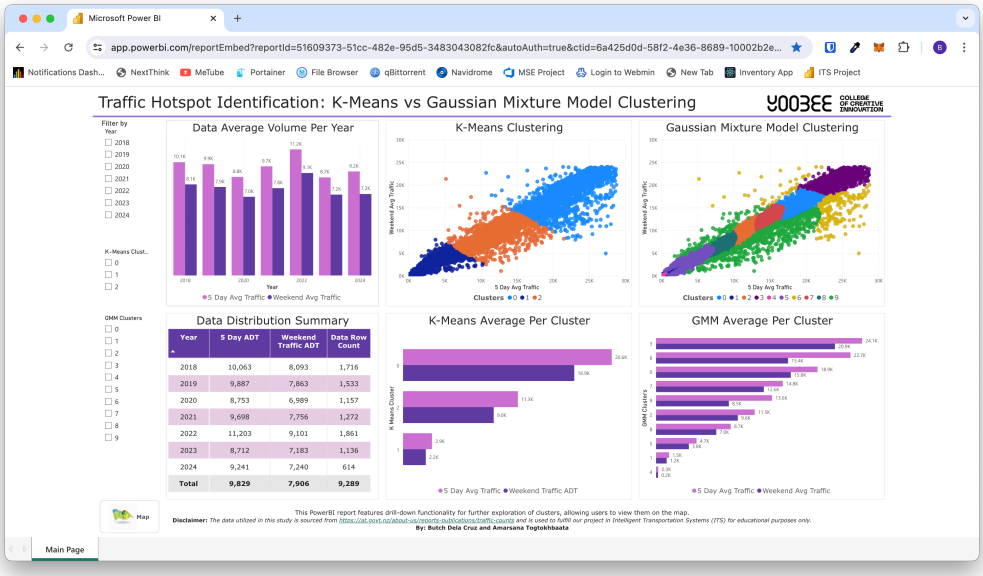

Figure 5: Power BI Dashboard: Visualizing Traffic Hotspot Detection with K-means and GMM Clustering

Visualization Components

1. *Scatter Charts for Clustering Results*

   Scatter charts are powerful tools for visually representing the results of K-means clustering and GMM. Each data point is plotted based on its feature values (5 Day Avg Traffic and Weekend Avg Traffic), with different clusters distinguished by color or marker shape. This visualization technique allows for an immediate understanding of how traffic points are grouped into clusters, revealing spatial patterns and potential overlap between clusters.

Scatter plots showing K-means and GMM clusters with distinct colors for each cluster. This helps in visualizing areas of high, moderate, and low traffic intensity across Auckland.

2. *Bar Charts for Average Traffic Analysis*

Bar charts are effective for comparing average traffic volumes across different clusters or over time periods. They provide a clear visual representation of traffic trends and variations, enabling stakeholders to identify areas with consistently high traffic volumes or seasonal fluctuations. Bar charts displaying average traffic per year or per identified cluster (K-means and GMM). This visualization helps in identifying trends and anomalies in traffic patterns that may require specific interventions.

3. *Interactive Maps with Drill-Down Features*

Interactive maps enhance spatial understanding by allowing users to explore traffic hotspots dynamically. These maps can incorporate drill-down features, enabling users to zoom into specific regions or clusters of interest. Clickable data points on the map provide detailed information about traffic conditions at specific locations. An interactive map of Auckland displaying traffic hotspots identified by K-means and GMM. Users can zoom in to view detailed traffic data at street level, facilitating targeted urban planning and management decisions (Figure 6: Interactive Maps with Drill-Down Features).

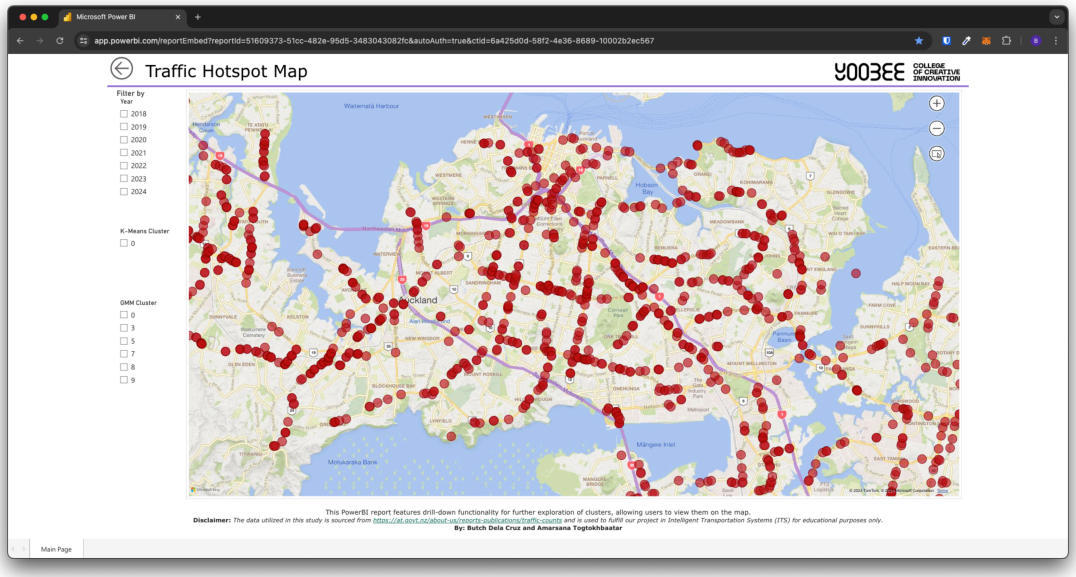

Figure 6: Interactive Maps with Drill-Down Features

4. *Tables for Detailed Data Insights*

Tables present detailed data distributions and statistical summaries, complementing other visualizations by providing specific numerical values and metrics. They are useful for presenting comprehensive information about traffic volumes, cluster characteristics, and statistical measures used in model evaluation. A table summarizing traffic data distributions per year or per cluster, including average daily traffic, peak hours, and traffic intensity variations. This tabular format aids in deeper analysis and validation of clustering results.

Advantages of Power BI for Visualization

Power BI offers several advantages for visualizing traffic hotspot detection results:

- **Interactivity**: Users can interact with visualizations, apply filters, and drill down into specific data points, enhancing engagement and exploration.
- **Integration**: Seamless integration of multiple visual components (charts, maps, tables) into cohesive dashboards allows for comprehensive data analysis and presentation.
- **Real-Time Updates**: Data refresh capabilities ensure that visualizations reflect the latest traffic data, supporting ongoing monitoring and decision-making.

Effective data visualization in Power BI facilitates the interpretation and communication of complex traffic patterns identified through K-means clustering and Gaussian Mixture Models. By leveraging scatter charts, bar charts, interactive maps, and tables, stakeholders can gain actionable insights into traffic hotspots, supporting informed urban planning and transportation management strategies in Auckland and similar metropolitan regions.

Incorporating these visualizations effectively showcases the capabilities of Power BI in transforming raw traffic data into actionable insights, enhancing the overall impact and utility of your study on traffic hotspot detection.

Feasibility Assessment

Scalability and Accuracy Evaluation

In evaluating the scalability and accuracy of K-Means and GMM in detecting traffic hotspots, it is evident that each algorithm offers unique advantages. K-Means stands out for its computational simplicity and speed, making it highly scalable for analyzing large datasets efficiently. This attribute enables K-Means to swiftly pinpoint congestion zones with consistent traffic patterns, providing a quick and practical approach to identifying areas of concern. On the other hand, Gaussian Mixture Models (GMM) showcase effectiveness in capturing the complex and probabilistic nature of traffic patterns. Despite being more intricate than K-Means, GMM's capability to model congestion areas of various shapes and sizes accurately enhances its accuracy in hotspot identification. By leveraging the strengths of both algorithms, traffic management systems can benefit from a comprehensive analysis that combines the scalability of K-Means with the nuanced accuracy of GMM for a more holistic approach to hotspot detection and traffic management strategies.

Applicability Assessment

The identification of traffic hotspots plays a crucial role in enhancing traffic management strategies and urban planning initiatives. By pinpointing congestion hotspots, authorities can strategically implement targeted interventions such as optimizing traffic signals or upgrading road infrastructure in these areas to effectively alleviate traffic issues. Understanding congestion areas also enables the development of traffic flow optimization strategies, allowing for the efficient rerouting of traffic or deployment of resources to enhance overall traffic flow and reduce congestion-related delays. Moreover, recommendations derived from hotspot analysis can inform urban planners in making informed decisions for sustainable urban mobility and infrastructure development, leading to long-

term improvements in the transportation network and urban livability. By integrating hotspot data into decision-making processes, cities can proactively address traffic challenges and enhance the efficiency and sustainability of their transportation systems.

Impact Measurement

Measuring the impact of hotspot detection on traffic congestion and efficiency involves assessing various key factors. One method is comparing traffic flow data before and after implementing interventions targeted at identified hotspots, providing insights into the effectiveness of congestion alleviation measures and their impact on traffic levels and flow patterns. Evaluating changes in travel time through hotspot areas post-implementation can offer tangible evidence of efficiency improvements, indicating the success of implemented strategies in reducing delays and enhancing overall travel experience for commuters. Furthermore, monitoring shifts in emissions and fuel consumption in hotspot areas can serve as a valuable metric for assessing the environmental impact of traffic management strategies, showcasing the effectiveness of initiatives aimed at mitigating congestion-related environmental concerns. By holistically measuring the impact of hotspot detection on traffic congestion, travel time, and environmental factors, authorities can gauge the success of their interventions and make data-driven decisions to optimize traffic management efforts and promote sustainable urban mobility.

Collaborative Approach

Collaboration with Experts

The involvement of traffic planners, engineers, and data analysts in hotspot detection is essential for a comprehensive approach to traffic management. Traffic planners contribute their expertise in designing targeted interventions based on hotspot analysis, ensuring efficient traffic flow and congestion alleviation. Engineers play a crucial role in implementing infrastructure improvements identified through hotspot detection, optimizing road networks to enhance mobility. Data analysts provide valuable insights by processing and interpreting traffic data, enabling evidence-based decision-making for effective traffic management strategies.

Stakeholder Engagement

Interactions with local transportation authorities and stakeholders are vital for the successful implementation of hotspot detection strategies. Engaging with transportation authorities allows for the alignment of hotspot findings with existing traffic management initiatives and regulations, ensuring seamless integration into urban planning frameworks. Collaboration with stakeholders, such as businesses and community groups, facilitates the sharing of local knowledge and concerns, fostering a more inclusive approach to addressing traffic issues and garnering support for proposed interventions.

Research Partnerships

Collaborations with research institutions and industry experts bring valuable expertise and innovation to hotspot detection projects. Partnering with research institutions allows for the integration of cutting-edge methodologies and technologies in traffic analysis, enhancing the accuracy and effectiveness of hotspot identification processes. Working with industry experts

provides practical insights into real-world traffic management challenges, leading to the development of actionable recommendations and solutions tailored to the specific needs of urban mobility and infrastructure development. Research partnerships strengthen the foundation of hotspot detection projects, enabling a multidisciplinary approach to addressing complex traffic congestion issues and fostering sustainable urban transportation practices.

Results and Discussion

The application of K-Means and GMM for hotspot detection in Auckland provided valuable insights into traffic congestion patterns across the region. K-Means effectively identified distinct congestion zones characterized by similar traffic patterns, aggregating areas with consistently high traffic volumes into coherent clusters. In contrast, GMM offered a more nuanced analysis by accurately modeling complex congestion areas of various shapes and sizes, accommodating the diverse distribution of traffic intensities throughout Auckland. This comprehensive analysis revealed specific locations prone to persistent congestion and varying traffic dynamics, crucial for targeted interventions and strategic urban planning.

Comparing the performance of K-Means and GMM in hotspot detection highlighted their respective strengths. K-Means excelled in identifying congestion zones with uniform traffic patterns, leveraging its computational efficiency to manage large datasets effectively, making it suitable for real-time applications in traffic management systems. Conversely, GMM demonstrated superior capability in handling the diverse shapes and distributions of congestion areas, providing a detailed representation of overlapping or nonlinear traffic patterns.

Insights drawn from the hotspot detection using K-Means and GMM emphasize actionable strategies for traffic management. Implementing targeted interventions such as optimizing traffic signals and improving road infrastructure in identified hotspots can effectively alleviate congestion and enhance traffic flow. Strategies can further optimize resource allocation and reroute traffic to minimize delays, supporting overall urban mobility improvements. The data-driven approach facilitated by hotspot analysis enables informed decision-making in urban planning, prioritizing sustainable transportation solutions and infrastructure development initiatives. Continuous monitoring of hotspot areas ensures adaptive strategies that sustainably manage traffic and enhance the efficiency of urban transportation systems.

By leveraging the capabilities of K-Means and GMM in hotspot detection and translating these findings into practical insights, traffic management authorities can proactively address congestion challenges, optimize traffic flow, and foster the development of resilient, sustainable urban mobility solutions tailored to the dynamic needs of metropolitan areas such as Auckland.

Conclusion

In conclusion, the study has delved deep into the realm of traffic management by applying cutting-edge machine learning techniques such as K-Means and Gaussian Mixture Models (GMM) to detect

traffic congestion hotspots within Auckland. Through the utilization of unsupervised learning models, the research aimed to unveil intricate traffic patterns that might not be readily apparent through traditional analysis methods. These identified hotspots play a crucial role in guiding targeted interventions, infrastructure enhancements, and strategic traffic management initiatives, all geared towards fostering more efficient and sustainable urban mobility in Auckland.

Furthermore, the effective visualization of these hotspots through interactive charts and maps not only provides a clear depiction of congestion areas but also aids in the dissemination of information to stakeholders and decision-makers. By visually representing the clusters and traffic patterns, the study bridges the gap between data analysis and practical implementation, thereby offering actionable insights for traffic management strategies and urban planning initiatives. This comprehensive approach, combining robust data analysis with intuitive visualization, sets the stage for informed decision-making and paves the way for transformative changes in addressing urban traffic challenges.

In essence, the integration of advanced machine learning techniques and sophisticated data visualization methods showcased in this study represents a significant step towards enhancing the resilience and efficiency of urban transportation systems. By leveraging data-driven insights and visual representations, cities like Auckland can move towards a more sustainable future, where traffic congestion is mitigated, travel time is optimized, and overall urban mobility is enhanced for the benefit of all residents and stakeholders.

References

Z score for Outlier Detection – Python. (2024). Retrieved from Z score for Outlier Detection – Python: https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/

Auckland Transport. (2024). *Traffic counts*. Retrieved from Traffic counts: https://at.govt.nz/about-us/reports-publications/traffic-counts

Microsoft. (2024). *PowerBI - Data Visualization*. Retrieved from PowerBI - Data Visualization: https://www.microsoft.com/en-us/power-platform/products/power-bi#Resources

Gaussian Mixture Model Selection. (n.d.). Retrieved from Gaussian Mixture Model Selection: https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html

Elbow Method for optimal value of k in KMeans. (2023). Retrieved from Elbow Method for optimal value of k in KMeans: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

How can regression analysis identify traffic trends? (n.d.). Retrieved July 4, 2024, from https://www.linkedin.com/advice/0/how-can-regression-analysis-identify-traffic-qvndc

The Machine Learning Framework for Traffic Management in Smart Cities: Optimizing Flows for a Sustainable Future | LinkedIn. (n.d.). Retrieved July 4, 2024, from https://www.linkedin.com/pulse/machine-learning-framework-traffic-management-smart-cities-santosh-g-ym8oc/

Ding, W., Xia, Y., Wang, Z., Chen, Z., & Gao, X. (2020). An ensemble-learning method for potential traffic hotspots detection on heterogeneous spatio-temporal data in highway domain. *Journal of Cloud Computing*, *9*(1), 25. https://doi.org/10.1186/s13677-020-00170-1

Sonnleitner, E., Barth, O., Palmanshofer, A., & Kurz, M. (2020). Traffic Measurement and Congestion Detection Based on Real-Time Highway Video Data. *Applied Sciences*, *10*(18), Article 18. https://doi.org/10.3390/app10186270

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems*, *16*, 1–15. https://doi.org/10.1109/TITS.2015.2404431

Ricciato, F., Janecek, A., Valerio, D., Hummel, K., Ricciato, F., & Hlavacs, H. (2015). The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. *IEEE Transactions on Intelligent Transportation Systems*, *16*. https://doi.org/10.1109/TITS.2015.2413215