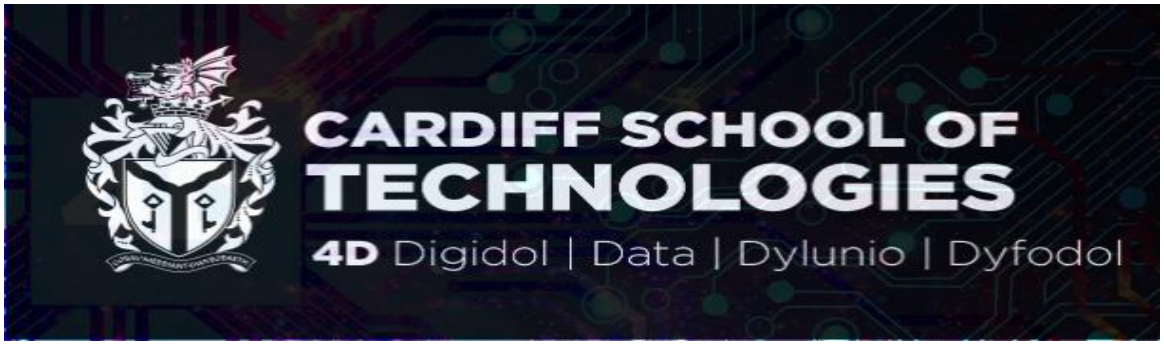



| Cardiff Metropolitan University | |
|---|---|
| **Cardiff School of Technologies** | |
| **Academic Year: 2024/2025** | |
| **Term: 2** | |
| **Module Name:** Big Data Technology | |
| **Module Code:** DSA7002 | |
| **Module Leader:** Dr Imtiaz Hussain Khan | |
| **MSc Programme:** Data Science | |
| | |
| **Assignment Title: UK Traffic Data Analysis using Big Data and Distributed Computing Approach** | |
| **Student Name:** Luke Strong | **Student ID:** 20121704 |

# Contents

# Introduction to the problem and objectives of the study

The advancements made in data analytics and algorithmic research can open new avenues for exploring traffic patterns, which would enable decision makers to be better informed. This can be achieved by designing smarter and more resilient transportation systems. This report explores a large-scale traffic dataset from the Annual Average Daily Flow (AADF) which consists of several key factors, such as road categories and types, differing vehicle types and traffic volumes, as well as regional and temporal data. This report can detail nuanced insights into how these factors can be used to determine traffic volumes and behaviours, informing decision makers into how best to adapt the current traffic systems to improve traffic flow.

## The Problem

Rapid growth has led to the increasingly congested roadways, which not only impedes economic productivity but also contribute to environmental degradation through higher emission levels. Marioni, (2024) details that the number of vehicles in use has increased exponentially, so much so that on average 61 hours is lost to each driver every year, with city areas being the highest in the country, especially in London. Which according to INRIX, (2024) London has the 4$^{th}$ largest number of delays due to traffic in the world and the largest in Europe. This all contributes heavily to environmental factors, such as the emission of greenhouse gases, as well as the increase in noise pollution and the degrading of air quality.

Regional disparities in transport infrastructure further complicate situations when evaluating the effectiveness of public transport versus private transport. This is largely due to population density discrepancies across the country. Bretter & Schulz, (2023) speaks to this issue and that it is probable that population density in more rural areas of the country is reliant on private transport, due to the absence of sufficient public transport infrastructure. This report can help to understand which portions of the country are most in need of public transport initiatives reviewed and revamped.

These challenges emphasise a need for new and more analytical approaches that can uncover less obvious patterns within a complex dataset, which can enable tailored interventions. Ramchandra & Rajabhushanam, (2022) researched the use of machine learning in traffic flow analysis, which found that the prediction of traffic flow can efficiently inform decision making when it comes to where flow is lowest and needs to be addressed.

## Purpose and Objectives

The predominant purpose of the study is to leverage advanced data analytics and clustering techniques to gain a meaningful understanding of traffic characteristics and transportation patterns. These objectives are to be achieved by applying distributed clustering methods using PySpark's machine learning libraries to uncover patterns thereby offering valuable insights that can inform infrastructure planning, policy decision-making, and sustainable transportation strategies.

In summary, this report is designed to explain the diverse nature of modern transportation systems. By bringing together large-scale data processing, rigorous analytical methodologies, and strategic insights, the study aims to contribute to a knowledgeable discussion about sustainable mobility and to support the development of more resilient transportation networks.

# Discussion of your approach to data preparation, EDA, clustering, and scalability challenges, critically evaluating your methodological choices

## Data Preparation

Once the dataset was loaded, it was initially very extensive, encompassing various traffic-related attributes. To enhance the efficiency and relevance of the analysis, select columns were kept while others were excluded. The included factors are traffic volume, vehicle types, road categories, and temporal aspects. Redundant or less informative columns, like geographical coordinates or specific vehicle axle configurations, were excluded. These essential columns were selected as the data is manageable and relevant for the tasks.

The removal of many columns is for the reasoning of data relevance. Without relevance to the analysis that is to be performed, then the removal of these data points means that the computing workload decreases significantly while still maintaining integrity of the dataset (Günther et al., 2017).

Given the substantial size of the dataset, missing values were addressed with a combination of imputation and removal. For the columns involving vehicle types that had rows showing

null values, it was key to attribute these with genuine zero traffic instances, with imputation of true zero being employed to preserve information. The study of Li et al., (2019) stated that imputation of values is a favourable method of dealing with missing values, particularly when dealing with temporal traffic data, as it can break the link of year-to-year data.

For other columns in question, rows that had missing values were dropped from the analysis due to the size and nature of the dataset being vast enough that the dropping of columns doesn't affect accuracy of future results of analysis. This is because of the significant size of the original file, with the alternatives being less effective. Other methods of dealing with missing values seemed less effective in the attempt to not disrupt analysis. This correlates with the findings of the study by Platias & Petasis, (2020) who used the function of dropping rows with missing values, due to the size of the dataset that was used in their study.

## Exploratory Data Analysis (EDA)

The primary focus of the EDA phase was to uncover underlying patterns, trends, and relationships within the traffic data. This involved visualising temporal variations in traffic volume, examining regional differences, and investigating the influence of various vehicle types on overall traffic congestion.

The most notable characteristic across all different interpretations of the dataset was a sharp decline in traffic volume in 2020, coinciding with the COVID-19 pandemic. Analysis in this study can be correlated with the study of Singh et al., (2022), which went a step further to understand the environmental impacts of the reduction in traffic flow and volumes.

Regional disparities in traffic volumes and public transport usage were also evident, suggesting the influence of factors like population density and infrastructure availability. Vickerman, (2021) stated that with the reduction in public transportation usage due to the pandemic and the subsequent rebound was lower than pre-pandemic levels. This can put strains on the financial viability of the systems in place, with development and improvements becoming more difficult to implement.

The analysis further accentuated the substantial contribution of light and heavy goods vehicles (LGVs and HGVs) to overall traffic congestion on major road categories, particularly in relation to the pandemic, which Lin et al., (2022) found that mobility reduced by around 20%. It can be observed that the pandemic harshly affected the traffic flow of cars, while the

LGVs and HGVs were less affected. Furthermore, that more HGV type vehicles were used for delivery due to the expanding quantity of online ordering, with society being restricted on its movements.

Interactive visualisations using Plotly Express were employed to efficiently express perceptions. Interactive line charts were employed to illustrate trends in traffic volume over time, while stacked and grouped bar charts enabled comparisons of vehicle proportions across road categories. These visualisations allowed for a clear and innate comprehension of the complex traffic dynamics.

## Distributed Clustering

By grouping similar traffic patterns together, the analysis sought to uncover natural segments within the data, such as clustering by vehicle compositions, road characteristics or regional differences. This can inform both operational insights, as well as further model development. In a domain with diverse traffic conditions spanning urban and rural environments, clustering offers a valuable perspective on the underlying data structures that might otherwise be obscured by aggregated statistics.

Given the scale and noise complexity of the dataset, the K-Means algorithm was selected for its efficiency and scalability in distributed settings, particularly when dealing with compact hyper related clusters (Ahmed et al., 2020). K-Means partitions data points into k-clusters by iteratively assigning each point to the nearest centroid and recalculating centroids to minimise the within-cluster sum of squares (WCSS). This refinement process is well suited for large-scale traffic data, similarly, detailed in the research of Perera, (2025), which states that scaled features clustered efficiently can rapidly solve large complex problems, while ensuring computational efficiency.

Dynamic principal component analysis (PCA) was applied to select the minimum number of principal components necessary to explain a predefined portion of the variance. This reduced noise is discarding less informative components and improved both computational performance and the interpretability of the clustering results. Alongside this, categorical variables were encoded using One-Hot Encoding, this ensures that all features contributed appropriately to the computation of the K-Means algorithm.

## Scalability Challenges

When taking scalability into consideration, the primary challenge would be the sheer volume of the traffic dataset. To process and analyse a vast dataset, the computing power needed for efficient distributed computing techniques to ensure timely execution.

The distributed computing engine is tailor made for petabyte-scale data processing without the need for down sampling (Spark, 2018). Spark is useful for distributed computations across a cluster of machines significantly reducing processing time, this is done by leveraging parallel processing. As stated by Muvva, (2023) this framework helps manage the larger data volumes, providing the tools to perform complex data transformations, aggregations and machine learning tasks. This was critical for sustaining performance and scalability in environments.

Additional optimisation strategies were implemented to future enhance scalability. A selection of data structures plays a key role in the process, ensuring that operations such as grouping, joining and aggregations could be achieved promptly and competently. Optimising these data structures directly contributed to minimising unnecessary data shuffling – a major source of performance bottlenecks in distributed computing environments (Sun et al., 2023). This is improved by the addition of nodes and more memory at a high computing cost and Spark is widely considered as the processing data engine of choice in this function.

## Critical Evaluation of Methodological Choices

While the data preparation process was vital for restructuring the analysis, potential limitations should be acknowledged. While the process of removing columns ensured consistency and facilitated downstream tasks by standardising data, the exclusion of certain columns might have inadvertently omitted valuable information. This method could have potentially risked bias by diminishing the significance of outliers and inherent data variability. Further exploration of alternative imputation techniques or strategies for handling missing data could be considered in future work, particularly when extreme values can hold critical information about traffic dynamics.

EDA involved grouping data along a variety of different dimensions, such as year, region and road type. The analysis successfully reveals temporal trends and regional patterns. Despite the insights gained from this, operations such as this can invoke considerable data shuffling, which can lead to network latency and resource contention. Liang et al., (2024) stipulates that

while Spark's functionality partially mitigates issues, through caching and partitioning techniques, there can be contention between processing efficiency and detailed data exploration. Furthermore, the necessity of exporting data for external visualisation can add a layer of complexity to the workflow.

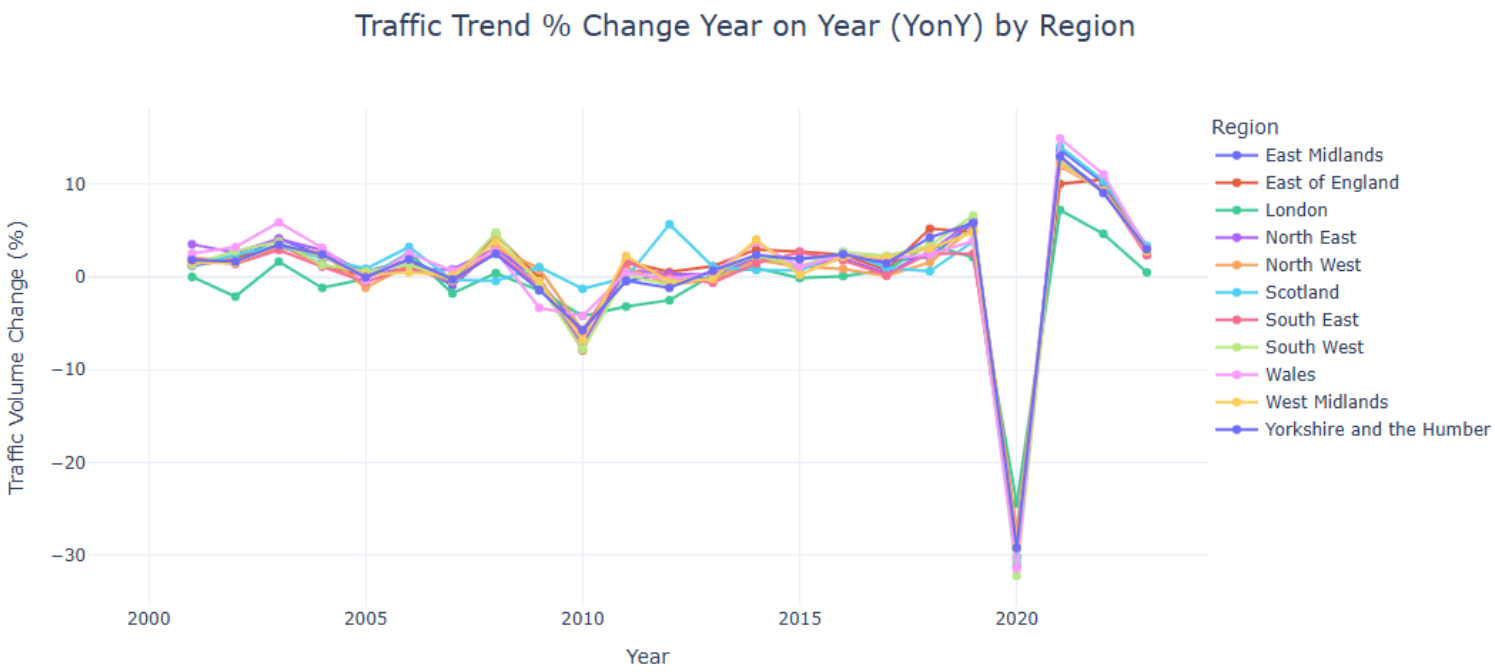
Even with dynamic PCA and advanced metrics, traffic patterns transitioned gradually, leading to overlapping clusters. This continuous nature made it difficult to distinctly separate into boundary regimes. Clustering-based pseudo-labels often resulting in dominant prediction clusters. Specifically, higher traffic classes were frequently misclassified into lower ones due to the dominant nature of the low traffic group in the underlying dataset and potential limitations in the features' discriminative power. Addressing this required more sophisticated sampling or cost-sensitive approaches, as well as refined feature engineering to capture subtler nuances.

The combined computational load of dynamic PCA, K-Means clustering, as well as grid search over hyperparameters was significant. Yang, (2024) states that this can be an issue, however running the analyses in a distributed environment can be of use, but resource constraints occasionally led to job cancellations or unexpected shutdowns. This can be put down to memory limits or processing power on local clusters. This dictated iterative adjustments in configuration, which would lead to increased driver memory and reducing parallelism, which would stabilise the environment.

# Implementation details (code snippets and discussion), presentation and analysis of results

## EDA:

**Figure 1. How Traffic Volumes in terms of percentage change Year-on-Year (YonY) has changed over Time? With a view to examine the impact of the Covid-19 Pandemic in 2020.**



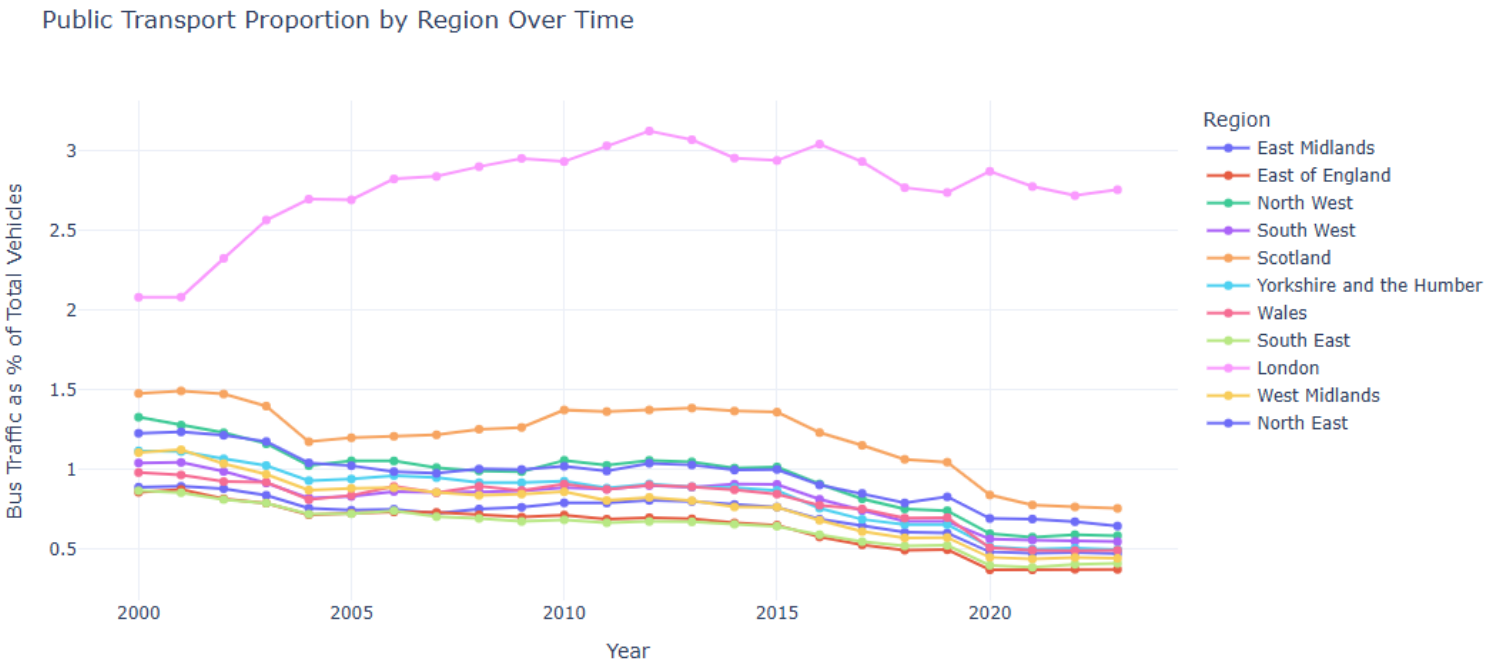

Discussion -

The sharp drop on the graph in 2020 demonstrates the dramatic effect of the COVID-19 pandemic on mobility and the significant increase in 'Work from Home' policies. The policymakers could use this insight to evaluate recovery efforts or formulate better plans for future disruptions. The implications of this shuffling of the data shows how each region was affected by the pandemic, but furthermore the way regional traffic may have changed over time. Which could link to relocation destinations, tourist ideals and such. Most regions experienced relatively minor YoY changes before 2020. This suggests traffic demand was stable and predictable, likely aligning with economic growth and population trends.

The other significant change in trend was in 2010 which saw a decrease in overall percentage change, which could be a factor of economic instability, due to the financial market crash of 2009 leading to a decrease in tourism across all regions, but in the seaside regions of the

Southwest, East or Northeast regions that see more tourism than other regions. The studies of Stapleton et al., (2017) and Song et al., (2013) identified that vehicle usage was affected by the rising costs of fuel caused by the economic recession, as well as the evolving socioeconomic status of the population causing changes in travel profiles.

**Figure 2. How does public transportation usage vary across regions, and what trends emerge?**




Discussion -

Regional disparities in transport infrastructure further complicate situations when evaluating the effectiveness of public transport against private transport. This is largely due to population density discrepancies across the UK. Bretter & Schulz, (2023) speaks to this issue and that it is probable that population density in more rural areas of the country is reliant on private transport, due to the absence of sufficient public transport infrastructure.

This shuffle shows this trend with the London area being clearly more reliant on public transport compared to others, with more rural regions such as the East of England and Southeast areas having lower proportion of public transport. This is due to population density, however more examination can be done on the impact of increased availability of public transport in the future.

**Figure 3. How do 'LGVs', 'HGVs' and 'Buses' influence traffic congestion compared to smaller vehicles across different road types?**

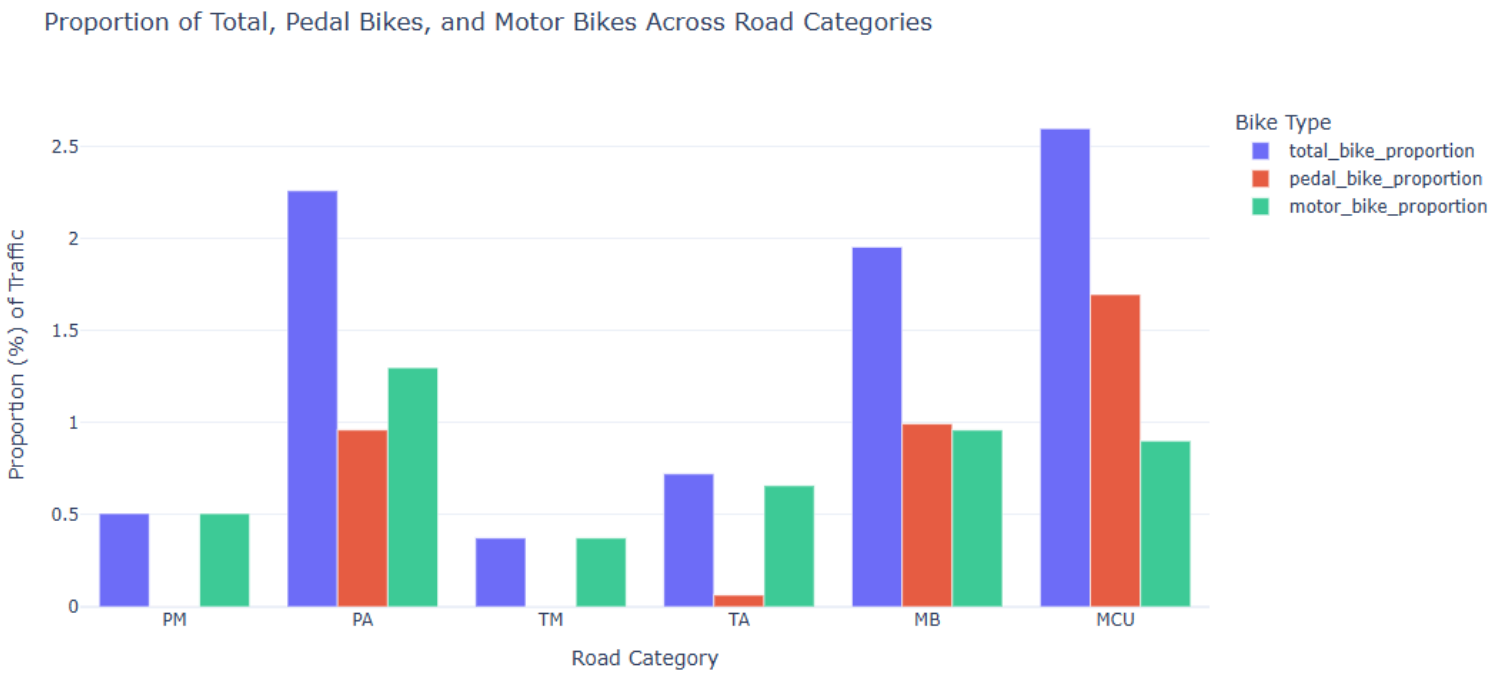

Discussion -

This bar chart provides a comprehensive view of how different types of vehicles contribute to traffic patterns across various road categories.

The LGV_to_Car_Ratio is consistently the highest across all road categories. This implies that light goods vehicles make a substantial contribution to traffic congestion relative to cars, emphasizing their importance when traffic planning takes place. The HGV_to_Car_Ratio and HGV_to_Total_Ratio show large contributions, particularly in road categories related to freight transport like motorways and major roads. This highlights the logistical reliance on heavy goods vehicles and policy interventions, such as dedicated HGV lanes, could be explored to ease traffic contributions.

The Buses_to_Car_Ratio and Buses_to_Total_Ratio are the lowest across all road categories. This suggests that buses, as public transport options, play a smaller role in overall traffic volumes compared to LGVs and HGVs. While this might reflect efficient use of bus services, it also raises questions about bus utilisation and whether public transport needs enhancement in certain areas.

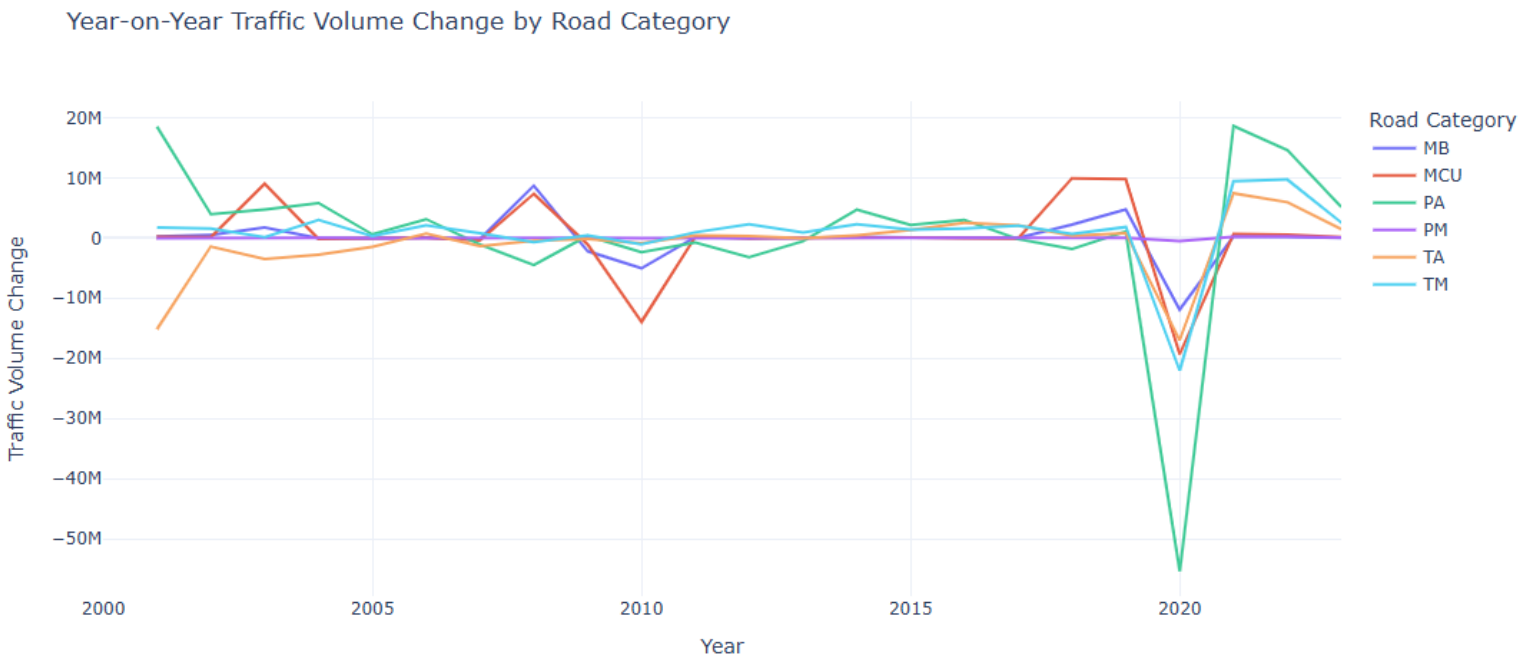
**Figure 4. How does the distribution of pedal bikes and two-wheeled motor vehicles vary across different road types, and what factors contribute to the variance of these vehicles on certain types of roads?**




Discussion –

Pedal bike proportions are particularly low across most categories (e.g., 0.00032% for PM). This might indicate inadequate cycling infrastructure or a preference for other modes of transport. High proportions of two-wheeled motor vehicles (e.g., MB and TA) suggest strong reliance on this mode. Factors such as road safety, accessibility, and commuting patterns may need further exploration.

**Figure 5. How has the YonY traffic volume changed for each road category, and what trends can be observed?**




Discussion -

When considering the traffic volume across the period there is a largely steady volume, except for the Covid-19 year of '2020'. This is particularly shown in the usage for 'PA' roads, or major A roads in the country. This can be down to the social distancing guidelines and the travel restrictions that were placed onto society during this period. This invites further exploration into its causes, such as economic disruptions or the pandemic. The pronounced decline in traffic volume change across most road categories (especially PA & TM) around 2020 warrants further investigation into its causes and the study of recovery trends. This graph is effective in visualizing broad traffic trends and anomalies over time.

## Clustering Analysis:

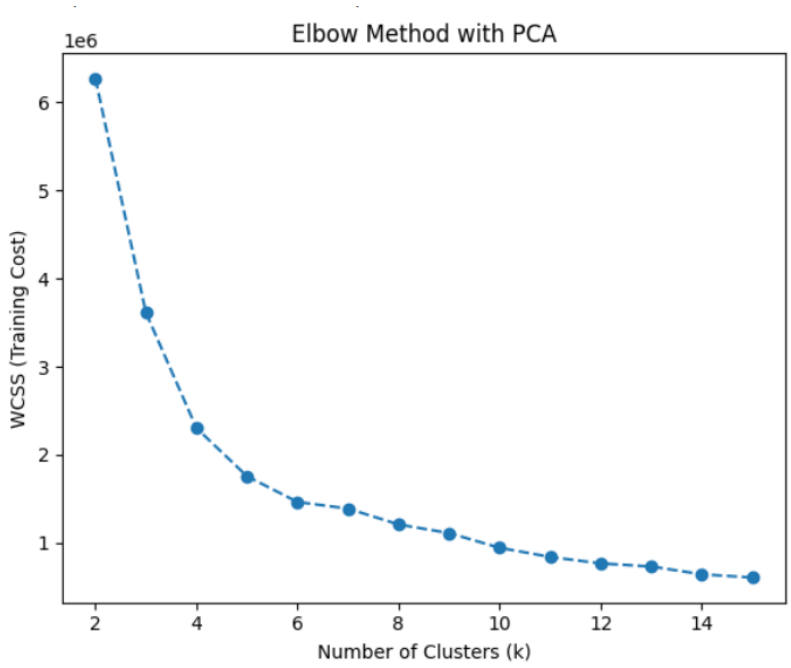
For the K-Means clustering analysis to take place, the study incorporates feature engineering, dimensionality reduction, and evaluation metrics to systematically determines an optimal number of clusters by evaluating how well-separated and compact they are. The Silhouette Score and WCSS help decide which k best segments the traffic data (Figure 6).

**Figure 6. Code Snippet to determine Elbow & WCSS Score for the most important features.**

```
assembler = VectorAssembler(inputCols=feature_columns, outputCol="features_unscaled")
df_vector = assembler.transform(df)

scaler = StandardScaler(inputCol="features_unscaled", outputCol="features", withMean=True, withStd=True)
scalerModel = scaler.fit(df_vector)
df_scaled = scalerModel.transform(df_vector)

pca = PCA(k=pca_components, inputCol="features", outputCol="pca_features")
pcaModel = pca.fit(df_scaled)
df_pca = pcaModel.transform(df_scaled)

evaluator = ClusteringEvaluator(featuresCol="pca_features", metricName="silhouette", distanceMeasure="squaredEuclidean")

silhouette_scores = []
wcss = []  # Within-Cluster Sum of Squares for elbow evaluation
models = {}

for k in range(k_min, k_max+1):
    kmeans = KMeans(featuresCol="pca_features", k=k, seed=42)
    model = kmeans.fit(df_pca)
    predictions = model.transform(df_pca)

    # Evaluate silhouette score
    silhouette = evaluator.evaluate(predictions)
    silhouette_scores.append(silhouette)

    # Obtain WCSS (training cost)
    current_wcss = model.summary.trainingCost
    wcss.append(current_wcss)

    models[k] = model

    print(f"K: {k} | Silhouette Score: {silhouette:.4f} | WCSS: {current_wcss:.2f}")
```

The results of this showed that when the major features of traffic data were paired with the One-Hot Encoded data of the additional features, that the ideal Elbow value with the highest Silhouette Score was K=6, with a score of 0.7559 (Figure 7).

**Figure 7. Elbow Method with PCA results**



When considering the distributed supervised learning methods to gain insights, due to the vast nature of the dataset. The results tended to be skewed, which would have been a result of:

- Class Imbalance
- Feature Dominance
- Overfitting
- Data Scaling issues
- Clustering concerns

This class imbalance was best seen when using the Random Forest and Logistic Regression Classification method, which showed a large disparity in cluster sizing (Figure 8).

**Figure 8. Cluster Framework after performing Distributed Supervised Learning (RF & LR)**

```
predictions.groupBy('indexedLabel', 'prediction').count().show(truncate=False)
```

```
+------------+----------+------+
|indexedLabel|prediction|count |
+------------+----------+------+
|2.0         |0.0       |46825 |
|1.0         |0.0       |90301 |
|0.0         |0.0       |177785|
+------------+----------+------+
```

Calculating the class weights based on the frequency of each traffic level category in the dataset confirms that uncommon classes receive elevated weighting, making their impact stronger during model training. When these class weights are integrated into supervised learning models, minority classes receive greater significance, thwarting them from being overshadowed by dominant classes. The model penalises misclassifications of rare categories more heavily, improving overall balance and accuracy across all classes.

This approach is particularly useful in traffic data analysis, where certain traffic patterns may be underrepresented but still crucial for decision-making.

# References


Ahmed, M., Seraj, R. and Islam, S.M. (2020) 'The K-Means Algorithm: A comprehensive survey and performance evaluation', *Electronics*, 9(8), p. 1295. doi:10.3390/electronics9081295.

Bretter, C. and Schulz, F. (2023) 'Public support for decarbonization policies in the UK: Exploring regional variations and policy instruments', *Climate Policy*, 24(1), pp. 117–137. doi:10.1080/14693062.2023.2273302.

Günther, W.A. *et al.* (2017) 'Debating big data: A literature review on realizing value from Big Data', *The Journal of Strategic Information Systems*, 26(3), pp. 191–209. doi:10.1016/j.jsis.2017.07.003.

INRIX (2024) *INRIX Scorecard*, *INRIX*. Available at: https://inrix.com/scorecard/#form-download-the-full-report (Accessed: 09 April 2025).

Li, L. *et al.* (2019) 'Missing value imputation for traffic-related time series data based on a multi-view learning method', *IEEE Transactions on Intelligent Transportation Systems*, 20(8), pp. 2933–2943. doi:10.1109/tits.2018.2869768.

Liang, H. *et al.* (2024) 'A survey on spatio-temporal big data analytics ecosystem: Resource Management, Processing Platform, and applications', *IEEE Transactions on Big Data*, 10(2), pp. 174–193. doi:10.1109/tbdata.2023.3342619.

Lin, S. *et al.* (2022) 'Impact of change in traffic flow on vehicle non-exhaust PM2.5 and PM10 emissions: A case study of the m25 motorway, UK', *Chemosphere*, 303, p. 135069. doi:10.1016/j.chemosphere.2022.135069.

Marioni, L. da S. (2024) 'Transport Connectivity in the UK: Regional Disparities and Policy Pathways', *National Institute of Economic and Social Research.*, 40.

Perera, N. (2025) 'Design of Cloud-Facilitated Data Repositories for Large-Scale Traffic Pattern Analyses', *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, 10(2), pp. 1–10.

Platias, C. and Petasis, G. (2020) 'A comparison of machine learning methods for data imputation', *11th Hellenic Conference on Artificial Intelligence*, pp. 150–159. doi:10.1145/3411408.3411465.

Ramchandra, N.R. and Rajabhushanam, C. (2022) 'Machine learning algorithms performance evaluation in Traffic flow prediction', *Materials Today: Proceedings*, 51, pp. 1046–1050. doi:10.1016/j.matpr.2021.07.087.

Song, Y., Preston, J.M. and Brand, C. (2013) 'What explains active travel behaviour? evidence from case studies in the UK', *Environment and Planning A: Economy and Space*, 45(12), pp. 2980–2998. doi:10.1068/a4669.

Spark (2018) *Apache sparkTM - unified engine for large-scale data analytics*, *Apache SparkTM - Unified Engine for large-scale data analytics*. Available at: https://spark.apache.org/ (Accessed: 26 April 2025).

Stapleton, L., Sorrell, S. and Schwanen, T. (2017) 'Peak car and increasing rebound: A closer look at car travel trends in Great Britain', *Transportation Research Part D: Transport and Environment*, 53, pp. 217–233. doi:10.1016/j.trd.2017.03.025.

Sun, X. *et al.* (2023) 'Survey of distributed computing frameworks for supporting Big Data Analysis', *Big Data Mining and Analytics*, 6(2), pp. 154–169. doi:10.26599/bdma.2022.9020014.

Vickerman, R. (2021) 'Will COVID-19 put the public back in public transport? A UK perspective', *Transport Policy*, 103, pp. 95–102. doi:10.1016/j.tranpol.2021.01.005.

Yang, W. (2024) 'Analysis and application of big data feature extraction based on improved K-means algorithm', *Scalable Computing: Practice and Experience*, 25(1), pp. 137–145. doi:10.12694/scpe.v25i1.2281.

# Appendix

## Table for Road Category

**Road types**

The following abbreviations are used in the 'Road Category' variable:

| Category | Category Description |
|---|---|
| PM | M or Class A Principal Motorway |
| PA | Class A Principal road |
| TM | M or Class A Trunk Motorway |
| TA | Class A Trunk road |
| M | Minor road |
| Of which… | |
| MB | Class B road |
| MCU | Class C road or Unclassified road |