Congestion Analysis and Route Optimization for Buses in Nottingham

by Emaan Bashir - 20594616

Submitted to The University of Nottingham
in September 2024
in partial fulfilment of the conditions for the award of the degree of
Master of Science in Data Science

I declare that this dissertation is all my own work, except as indicated in the text

Contents

Abstract .... ***iv***

Acknowledgements .... ***v***

Table of Figures .... ***vi***

List of Tables .... ***vii***

Abbreviations .... ***viii***

Chapter 1: Introduction ***1***

1.1 Background and Motivation **1**

1.2 Problem Statement **1**

1.3 Objectives of the Research **2**

1.3.1 Identification of Congested Routes 2
1.3.2 Analysis of Congestion Patterns 2
1.3.3 Proposal of Alternative Routes 2
1.3.4 Evaluation of Alternative Routes 2

1.4 Scope of the Study **3**

1.5 Significance of the Study **3**

1.5.1 Improving Public Transport Efficiency 3
1.5.2 Reducing Traffic Congestion 3
1.5.3 Environmental Impact 3
1.5.4 Economic Benefits 3
1.5.5 Contribution to Research 3

1.6 Structure of the Dissertation **4**

Chapter 2: Literature Review ***5***

2.1 Traffic Congestion: Definitions, Characteristics and Impacts **5**

2.2 Real-Time Data and Congestion Analysis **5**

2.3 Predictive Modelling and Travel Times Estimation **6**

2.4 Optimization of Bus Routes and Schedules **7**

2.5 Evaluation of Route Adjustments **7**

2.6 Real-Time Adaptive Control Systems in Public Transportation **7**

2.7 Challenges in Implementing Adaptive Control Systems **8**

Chapter 3: Data Collection and Preprocessing ***9***

3.1 Data Collection **9**

3.1.1 Bus Location Data 9
3.1.2 Bus Timetable Data 9

3.2 Data Preprocessing **9**

3.2.1 Parsing XML Data 9
3.2.2 Data Cleaning 10
3.2.3 Data Transformation 10
3.2.4 Data Integration 11
3.2.5 Data Storage 12
3.2.6 Challenges in Data Collection and Preprocessing 12

Chapter 4: Identification of Congested Routes 13

4.1 Calculation of Average Speed and Bus Density 13

4.1.1 Average Speed Calculation 13

4.1.2 Bus Density Calculation 13

4.2 Exploratory Data Analysis..13

4.2.1 Distribution of Average Speed and Bus Density 13

4.2.2 Box Plot Analysis 14

4.2.3 Outlier Removal 15

4.3 Cluster Analysis 16

4.3.1 k-Means Clustering.. 16

4.3.2 DBSCAN Clustering 18

Chapter 5: Analysis of Congested Routes 20

5.1 Distribution of Congestion Levels 20

5.2 Visualization of Congestion Levels on the Map...21

5.3 Analysis of Congestion by Route Section 21

5.4 Congestion Analysis by Day of the Week 23

5.5 Congestion Analysis by Time Interval 25

Chapter 6: Route Optimization and Evaluation... 28

6.1 Proposing Alternatives 28

6.1.1 Creating a Graph for Route Analysis 28

6.1.2 Finding Alternative Routes... 29

6.2 Evaluating Alternative Routes..30

6.2.1 Run Time Comparison 30

6.2.2 Statistical Evaluation 30

Chapter 7: Conclusion and Future Work.. 32

7.1 Summary of Findings 32

7.1.1 Congestion Analysis 32

7.1.2 Route Optimization.. 32

7.2 Implications...32

7.3 Future Work 33

References 34

Appendix A: Raw Data Samples for Timetable Data: 36

Appendix B: Raw Data Samples for Location Data 41

Appendix C: Sample of Pre-processed Timetable Data 42

Appendix D: Integrating Location data with Timetable data.................................. 43

Appendix E: Average Speed and Bus Density... 44

Appendix F: Congestion Summary... 45

Abstract

Urban congestion is a significant challenge for public transportation systems, particularly in cities with dense traffic patterns like Nottingham. Congestion on bus routes leads to increased passenger waiting times, reduced service reliability and lower passenger satisfaction. This study focuses on analysing and optimizing bus routes in a congested urban environment using data-driven approaches. By utilizing data from GPS tracking over a period of two weeks, along with bus timetables and historical route information, various congestion hotspots were identified. Advanced data analysis techniques, including k-means clustering, DBSCAN and graph-based algorithms like Dijkstra's algorithm, were employed to categorize route links into low, medium, and high congestion levels and to propose optimized alternative routes.

The analysis revealed that 24.5% of route links experienced low congestion, 60.3% had medium congestion, and 15.3% were highly congested, with significant congestion occurring during weekday working hours. Visualization tools such as pie charts, maps and heatmaps illustrated the spatial and temporal distribution of congestion, pinpointing key congestion hotspots. By applying Dijkstra's algorithm, alternative routes were suggested for highly congested areas, achieving an average travel time reduction of 12.89%. A paired t-test confirmed the statistical significance of these improvements.

This research not only provides a robust methodology for addressing route congestion in Nottingham but also offers a scalable model that can be applied to other urban settings. The findings advocate for the continuous analysis and optimization of urban transit networks to promote better mobility, increased customer satisfaction, and reduced environmental impact. Future research could benefit from incorporating real-time traffic data, considering additional factors such as weather and events, integrating with other transportation modes and employing advanced machine learning techniques for predictive congestion management.

Overall, this dissertation contributes to the field of urban transportation planning by demonstrating how data-driven approaches can effectively address traffic congestion and improve public transportation systems.

Acknowledgements

I wish to express gratitude to God Almighty for the strength and grace to successfully complete this work. I also wish to appreciate my supervisor Colin Johnson for his unwavering support, patience and guidance throughout the duration of this project. My sincere thanks go to Adam Akram and Conrad Forbes from the Bus Open Data Service (BODS) team, who helped me in accessing and understanding the bus timetable and location data. I finally wish to express thanks to my family and friends for the support they offered during the course of this dissertation.

List of Figures

Figure 1.1: Problems caused by Traffic Congestion 1

Figure 1.2: Flow of the Study 2

Figure 1.3: Significance of the Study 4

Figure 4.1: Distribution of Average Bus Speed and Bus Density 14

Figure 4.2: Box plots for Average Bus Speed and Bus Density 15

Figure 4.3 (a): Histogram for Average Speed after Outlier Removal 16

Figure 4.3 (b): Box plot for Average Speed after Outlier Removal 16

Figure 4.4: Scatter Plot for k-means with k = 2 17

Figure 4.5: Scatter Plot for k-means with k = 3 17

Figure 4.6: Scatter Plot for k-means with k = 4 18

Figure 4.7: Scatter Plot for DBSCAN Clustering 19

Figure 5.1: Pie Chart to illustrate the proportion of route links under each congestion level 20

Figure 5.2: Visualization of the route links with different congestion levels 21

Figure 5.3: Bar Chart to illustrate the congestion percentage of each route 22

Figure 5.4: Visualization of congestion levels across Route 30:3 22

Figure 5.5: Heatmap for congestion intensity for each route across the week 23

Figure 5.6: Heatmap by weekdays for specific routes 24

Figure 5.7: Heatmap for congestion intensity across different intervals of the day 25

Figure 5.8: Line Graph for Congestion Percentage throughout the day 26

Figure 5.9: Heatmap by time interval for specific routes 26

Figure 5.10: Line plot for congestion patterns of Route 11:4 during different time intervals 27

Figure 6.1: Directed graph representation of bus routes, where node denotes a bus stop and each edge represents the travel time between stops 28

Figure 6.2: Visual representation of the route 36:1 and the proposed alternative route 30

List of Tables

Table 5.1: Number of route links in each congestion category 20

Table 6.1: Comparison of Main and Alternative Route for the Route 36:1 29

Table 6.2: Run Time Comparison for Main and Alternative Routes........................ 30

Table A.1: Sample data for Stop Points 36

Table A.2: Sample data for Route Links... 36

Table A.3: Sample data for Routes... 36

Table A.4: Sample data for Journey Pattern Timing Links 37

Table A.5: Sample data for Operators 38

Table A.6: Sample data for Service Lines 39

Table A.7: Sample data for Service Journey Patterns 39

Table A.8: Sample data for Vehicle Journeys.. 40

Table B.1: Sample Location Data... 41

Table C.1: Pre-processed Timetable data sample 42

Table D.1: Integrated Location and Timetable data 43

Table E.1: Sample of data after calculating speed and bus density 44

Table F.2: Congestion Summary of the Route Sections....................................... 45

Abbreviations

NCT	**Nottingham City Transport**
BODS	**Bus Open Data Service**
AVL	**Automatic Vehicle Location**
DBSCAN	**Density Based Spatial Clustering of Applications with Noise**
GPS	**Global Positioning System**
TXC	**TransXChange**
GMT	**Greenwich Mean Time**
KDE	**Kernel Density Estimation**
IQR	**Inter Quartile Range**

Chapter 1: Introduction

1.1 Background and Motivation

Urban congestion is a pervasive issue that has affected cities worldwide and has a significant impact on the efficiency of public transportation systems. As cities continue to grow and urbanize, the demand for efficient and reliable public transport increases. However, the number of vehicles on the roads also increases, which coupled with limited road infrastructure, often leads to traffic congestion, particularly during peak hours. This congestion not only causes delays in the bus arrival times, but also increases the operational costs for the transport services, decreases the service reliability and passenger satisfaction and contributes to environmental pollution.

Nottingham, a historic city in the United Kingdom, is no exception to this trend. Known for its vibrant culture and economic activity, Nottingham faces significant traffic congestion challenges, especially in the central areas of the city. The Nottingham City Transport (NCT) operates extensively within Nottingham and provides a comprehensive service across the city with a network of about 300 buses. These buses are the backbone of the city's public transportation network and offer affordable and accessible means of travel. However, the efficiency of this service is often hampered by traffic congestion, which leads to non-adherence to the bus schedule and inconsistency in the service.

This research aims to address these challenges by conducting a detailed analysis of the congestion patterns affecting bus routes in Nottingham and proposing data-driven optimization strategies to enhance route efficiency. The motivation behind this research stems from the need to improve the overall quality of public transportation in Nottingham, ensuring that it remains a viable alternative to private car use, thereby reducing traffic congestion and its associated negative impacts.

1.2 Problem Statement

The primary problem addressed in this research is the inefficiency of bus routes in Nottingham due to traffic congestion. Congestion on the bus routes not only affects the punctuality of bus services but also leads to increased travel times, making public transportation less attractive to potential users. The situation worsens due to the fact that congestion is often unpredictable, with traffic patterns varying significantly across different days of the week and different times of the day.

The inefficiencies in the current bus routes can be attributed to several factors including the fact that most bus routes are designed based on static schedules and fixed paths and don't consider the traffic congestion. Despite the availability of large amounts of traffic and transit data, there is limited application of advanced data analytics and machine learning techniques in optimizing bus routes.

Given these challenges, there is a dire need for an approach that leverages bus location data to study the traffic congestion patterns and uses advanced data analytics to identify and optimize the congested routes, reducing the impact of congestion on service reliability and passenger satisfaction.

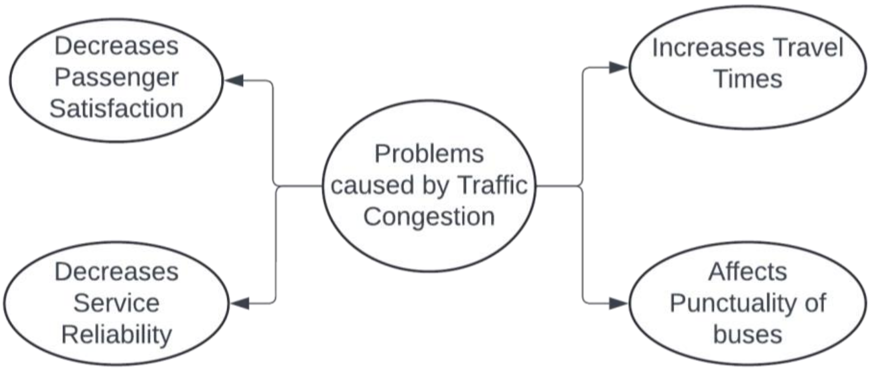

Decreases Passenger Satisfaction
Increases Travel Times
Problems caused by Traffic Congestion
Decreases Service Reliability
Affects Punctuality of buses

Figure 1.1: Problems caused by Traffic Congestion

1.3 Objectives of the Research

The main objective of this research is to develop and evaluate a congestion analysis and route optimization system for buses operating in Nottingham. The study is guided by the following specific objectives:

1.3.1 Identification of Congested Routes

The research seeks to identify the bus routes within Nottingham that experience higher levels of congestion compared to other routes, with a particular focus on those adversely impacting the bus service.

1.3.2 Analysis of Congestion Patterns

A comprehensive analysis of the congestion patterns along the bus routes in Nottingham will be conducted to identify key congestion hotspots and peak times that contribute to delays.

1.3.3 Proposal of Alternative Routes

The study will explore potential alternative routes for the most congested bus routes, prioritizing those that reduce the travel time and enhance the overall passenger experience.

1.3.4 Evaluation of Alternative Routes

The effectiveness of the proposed alternative routes will be evaluated by comparing the travel times of the suggested routes with those of the current routes, with the goal of demonstrating tangible improvements.

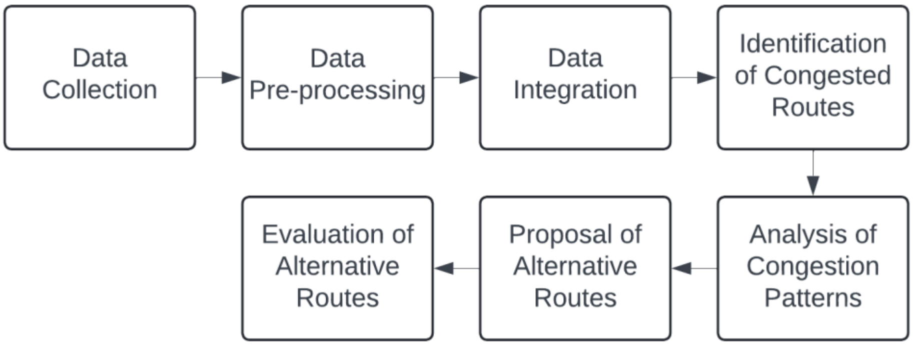

Data Collection
Data Pre-processing
Data Integration
Identification of Congested Routes
Analysis of Congestion Patterns
Proposal of Alternative Routes
Evaluation of Alternative Routes

Figure 1.2: Flow of the Study

1.4 Scope of the Study

The scope of this study is focused on bus routes within the city of Nottingham, with a particular emphasis on areas that are most affected by traffic congestion. The research considers the routes of buses operated by the Nottingham City Transport (NCT). While the primary focus is on identifying and analysing the traffic congestion on the bus routes, the study also suggests suitable alternatives for the congested routes.

The research utilizes a combination of data sources, including real-time GPS data from the buses over a period of two weeks and timetable data from the Bus Open Data Service (BODS). Advanced data analysis techniques, including machine learning, optimization algorithms and statistical tests are employed to develop and validate the proposed solutions.

It is important to note that although the findings of this research are specific to Nottingham, the methodology and approaches developed can be generalized and applied to other areas facing similar congestion challenges.

1.5 Significance of the Study

This research is significant due to several reasons.

1.5.1 Improving Public Transport Efficiency

By optimizing bus routes, this study aims to enhance the reliability and efficiency of Nottingham's public transport system. This can lead to shorter travel times, more punctual services and increased passenger satisfaction, making public transport a more attractive option for commuters.

1.5.2 Reducing Traffic Congestion

Optimized bus routes can help reduce overall traffic congestion by encouraging more people to use public transport instead of private cars. More people will start travelling on buses due to the reduced travel times.

1.5.3 Environmental Impact

Reducing congestion and improving the efficiency of public transport can contribute to lower emissions, as buses spend less time stuck in traffic and more people prefer public transport over private vehicles. This aligns with the broader environmental goals, such as reducing the city's carbon footprint.

1.5.4 Economic Benefits

Efficient public transportation system is crucial for the economic importance of a city. By reducing the travel time and improving connectivity, optimized bus routes can support economic activities, making it easier for people to access jobs, services and amenities.

1.5.5 Contribution to Research

The methodology of this research and the insights generated, contribute to the academic field of transportation planning and data science. The study demonstrates the application of advanced data analytics in solving real-world transportation problems, providing a reference for future research in this area.

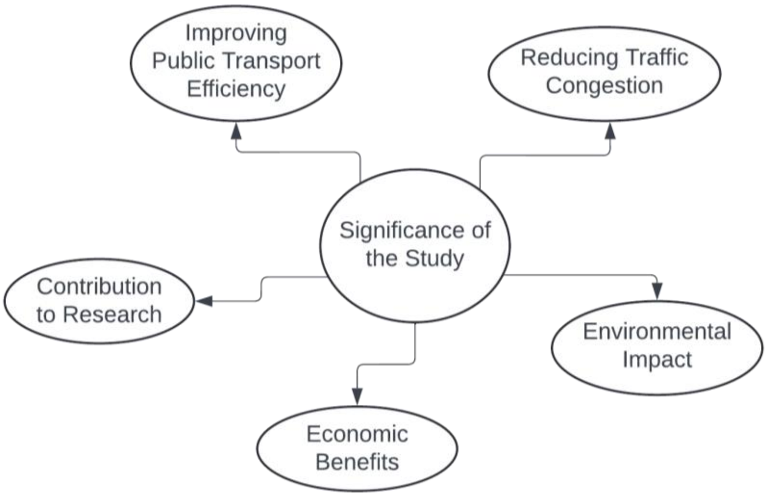

Improving Public Transport Efficiency
Reducing Traffic Congestion
Significance of the Study
Contribution to Research
Environmental Impact
Economic Benefits

Figure 1.3: Significance of the Study

1.6 Structure of the Dissertation

The following chapters explore key components of the research, beginning with a comprehensive review of the relevant literature. The structure of the remainder of this dissertation is as follows:

Chapter 2: Literature Review

This chapter provides an in-depth review of the existing literature on traffic congestion, public transport optimization, and the application of data analytics in transport planning.

Chapter 3: Data Collection and Preprocessing

This chapter outlines the data sources utilized in the study and describes the preprocessing techniques employed to ensure the quality and reliability of data. It also discusses the challenges encountered during data collection and the strategies implemented to overcome them.

Chapter 4: Identification of Congested Routes

This chapter details the methodology used to identify congested routes. The process and rationale behind selecting these routes are also discussed in detail.

Chapter 5: Analysis of Congested Routes

This chapter presents the findings of the congestion analysis, highlighting key congestion hotspots and peak congestion periods.

Chapter 6: Route Optimization and Evaluation

This chapter discusses the use of optimization algorithms to propose alternative routes aimed at reducing the travel time on congested routes.

Chapter 7: Conclusion and Future Work

This chapter summarizes the key findings of the study, reflects on the research objectives and concludes the report by identifying potential areas for future research.

Chapter 2: Literature Review

Urban transportation systems face significant challenges due to traffic congestion, particularly in densely populated cities where public transportation is heavily relied upon. Traffic congestion not only leads to delays and increased travel times but also plays a part in environmental issues such as air and noise pollution (Afrin & Yodo, 2020) [6]. Therefore, optimizing public transport systems, especially bus routes and schedules, has become a subject of significant academic and practical interest. This chapter reviews the relevant literature on traffic congestion, its causes and effects, the role of real-time data in public transport optimization and the strategies that can be used to improve bus services in complex urban environments.

2.1 Traffic Congestion: Definitions, Characteristics and Impacts

Traffic congestion is a multifaceted issue that has been extensively studied, yet it remains challenging to define, quantify and mitigate. Over the years, many different definitions have been proposed, each reflecting different aspects of the phenomenon. For instance, Lomax and Levinson (1997) [7] define congestion in terms of travel time or delay, while Gao et al. (2016) [8] focus on the heterogeneity of vehicle speeds. The most commonly used metrics for characterizing traffic congestion are traffic flow, road occupancy, speed and road capacity, as highlighted in studies such as Chang and Herman (1978) [9] and Li et al. (2021) [10].

Traffic congestion can be classified into two recurrent and non-recurrent congestion. Recurrent congestion is typically predictable and occurs due to regular patterns of traffic flow during specific times of the day, such as rush hours. However, non-recurrent congestion can not be predicted and is caused by unexpected events like road blockage, accidents or adverse weather conditions (Gmira et al., 2020 [11]; Afrin & Yodo, 2020 [6]; Li et al., 2021 [10]). The distinction between the two types of congestion is crucial for developing targeted strategies to solve specific traffic issues.

The impact of traffic congestion is much more than mere delays. It contributes to increased fuel consumption, leading to higher emissions of carbon and other pollutants, which in turn degrade the air quality (Afrin & Yodo, 2020) [6]. In addition to this, congestion also increases noise pollution, which can negatively affect the well-being of the residents. The economic costs associated with congestion, including increased vehicle operating costs and lost productivity, further highlight the importance of effective mitigation strategies.

2.2 Real-Time Data and Congestion Analysis

The introduction of real-time data collection technologies, particularly Automatic Vehicle Location (AVL) systems, has revolutionized the monitoring and optimization of public transport systems. AVL data enables the continuous tracking of buses, providing valuable information about their performance and the prevailing traffic conditions. The integration of real time data into public transport systems allows for more accurate predictions of travel times, identification of congestion patterns, and dynamic adjustment of bus schedules and routes.

Clustering techniques have been widely used in traffic and congestion analysis to classify traffic conditions and detect patterns of congestion. These techniques offer a powerful tool for understanding the complex and dynamic nature of urban traffic. Almeida et al. (2023) [1] demonstrated the use of buses as mobile sensors to classify and monitor traffic congestion in Alveiro, Portugal. Their methodology includes finding the relationship between bus speed and density and then using this information to identify the congestion patterns. The study employed machine learning algorithms like k-means clustering and DBSCAN to identify consistent congestion patterns. This approach is particularly useful in cities like Nottingham, where congestion can vary significantly across different times and locations. The ability to detect and respond to congestion in real-time is crucial for maintaining the reliability and efficiency of public transport services and enhancing passenger satisfaction.

Aftabuzzaman (2007) [12] and Gmira et al. (2020) [11] utilized clustering algorithms to categorize different levels of congestion based on traffic speed and flow. These techniques allow for a better understanding of congestion patterns, enabling the relevant authorities to make targeted interventions. For instance, Fei and Gkoutouna (2018) [13] applied k-means clustering to estimated bus speed data, identifying specific road segments prone to congestion at different times of the day. Their work highlights the potential of clustering techniques to uncover hidden patterns in traffic data.

Similarly, Diker and Nasibov (2012) [14] employed a Fuzzy Neighbourhood Density Based Spatial Clustering of Applications with Noise (DBSCAN) cluster to detect varying levels of congestion in a city in Turkey using car speed data, demonstrating the versatility of clustering methods in traffic analysis. Almeida et al. (2023) [1] also used a similar approach where k-means and DBSCAN algorithms were used to identify congestion levels on specific road segments.

Yang et al. (2019) [15] used a combination of clustering techniques to detect traffic congestion in real-time, applying DBSCAN to locate congestion at specific time instances and performing distributed topology analysis to find congested areas. Their approach demonstrates the versatility of clustering techniques in traffic analysis, enabling the detection of both recurrent and non-recurrent congestion patterns.

2.3 Predictive Modelling and Travel Times Estimation

Accurate prediction of bus travel times is an important aspect of optimizing public transport systems. Travel time estimation models have evolved significantly, moving from simple historical data analysis to more complex predictive models that incorporate real time data and machine learning algorithms. These models aim to provide reliable travel time predictions, which are essential for maintaining schedule adherence and improving passenger satisfaction.

Ma et al. (2019) [2] proposed a novel segment-based approach for predicting bus travel times, which uses real time data from both buses and taxis. The model accounts for the different factors influencing travel times in different segments of the bus routes by dividing the routes into transit and dwelling segments. This results in more accurate and reliable predictions. This model addresses the inherent variability in urban traffic, which can be influenced by a range of factors including traffic signals, passenger boarding times, and road conditions. This method is particularly effective in urban environments with highly variable traffic conditions where each route segment can have a different traffic situation.

In addition to segment-based approaches, other studies have explored the use of clustering techniques for travel time estimation. For example, Ton and Ng (2020) [16] combined vehicle speed data with driving characteristics and used clustering techniques to classify and identify driving patterns, which can then be used to predict travel times more accurately. Similarly, Li et al. (2021) [10] used traffic flow, speed and road occupancy data to detect non-recurrent congestion, which is crucial for making reliable travel time estimations in the presence of unexpected traffic events.

Reliable prediction of travel time is not only important for schedule adherence. It also plays a crucial role in optimizing bus routes. Accurate predictions make it possible to make dynamic adjustments to bus schedules, ensuring that the services remain efficient even under changing traffic conditions. This is particularly important in cities, where traffic patterns can vary significantly throughout the day.

2.4 Optimization of Bus Routes and Schedules

Optimizing bus routes and schedules is a complex task and requires a deep understanding of urban traffic patterns and passenger demands. Traditional approaches to route planning often focus on minimizing travel times and maximizing coverage, but these methods may not fully account for the dynamic nature of urban traffic. Recent advances in data analytics and machine learning have enabled more sophisticated optimization models that can adapt to real time traffic conditions and passenger needs.

Chien, Dimitrijevic, and Spasovic (2003) [3] developed a model that optimizes bus routes, headways and fleet sizes in urban areas with irregular grid networks. Their approach involves transforming irregular grid networks into pure grid structures, which allows for the application of grid-based optimization models. By minimizing total system costs, which include both operator and user costs, their model offers a comprehensive solution for optimizing bus routes in complex urban environments.

Huang et al. (2020) [4] extended his work by developing a real-time Customized Bus (CB) route optimization framework. CB services offer flexible routing and scheduling based on real-time user requests, providing amore responsive and personalized transportation option. Huang et al.'s [4] framework addresses the complexities of real-time data utilization, enabling rapid adjustments to schedules and routes in response to new information. This approach is valuable in areas where traffic conditions can change rapidly, making it difficult to maintain fixed schedules.

2.5 Evaluation of Route Adjustments

The optimization of bus routes and schedules also involves evaluating the impact of route adjustments and schedule changes on service efficiency and passenger satisfaction. Shi et al. (2021) [5] proposed a comprehensive evaluation model that considers the multidimensional effects of bus route optimization, including the impact on passenger waiting times, travel times and overall service reliability. The model integrates various data sources to provide a holistic assessment of how route changes affect the overall performance of the bus network.

Clustering techniques have also been used to evaluate the effectiveness of congestion mitigation strategies. For examples, Anbaroglu, Heydecker, and Cheng (2014) [17] used spatio-temporal clustering to detect non-recurrent congestion by analyzing high link journey time estimates on adjacent road segments. This method allowed them to evaluate congestion levels over time and assess the impact of various traffic management interventions.

2.6 Real-Time Adaptive Control Systems in Public Transportation

The integration of real-time data into public transportation systems has paved the way for the development of adaptive control systems, which are capable of dynamically adjusting operations based on current traffic conditions. These systems are particularly valuable in mitigating the impacts of congestion on bus services, ensuring that schedules remain reliable even under adverse conditions.

Adaptive control systems use real-time data from various sources such as traffic sensors, passenger information systems and GPS equipped buses. These systems are capable of making real time adjustments to bus dispatching, routing and scheduling by continuously analysing this data. For instance, if a bus gets delayed due to traffic congestion, the system can automatically adjust the schedules of the other buses to prevent bunching at the same spot.

One approach to real-time adaptive control is the use of predictive models that predict traffic conditions and bus travel times based on current and historical data. Ma et al. (2019) [2] highlighted the importance of integrating segment-based travel time predictions into adaptive control systems. By accurately predicting travel times for different segments of the bus route, these systems can better predict delays and take proactive measures to reduce their impact.

2.7 Challenges in Implementing Adaptive Control Systems

Despite the potential benefits of real-time data and adaptive control systems, their implementation in public transportation systems faces several challenges. One of the primary challenges is the high cost associated with deployment and maintenance of the necessary data collection infrastructure, such as GPS devices, traffic sensors and communication networks. Additionally, integrating data from multiple sources, such as buses, traffic lights and passenger information systems can be technically complex and require significant coordination between different agencies.

Another challenge is the need for advanced algorithms and computing power to process and analyse the large volumes of data generated by these systems. Real-time data processing requires low latency computing systems that can quickly generate actionable insights and control signals. Developing and maintaining these algorithms can be resource intensive and require specialized expertise.

Data quality and reliability are also significant challenges. Real-time data systems are vulnerable to errors and inaccuracies, which can arise from various sources such as faulty sensors, GPS signal loss or data transmission issues. Inaccurate data can lead to incorrect predictions and decision-making, undermining the effectiveness of adaptive control systems.

Chapter 3: Data Collection and Preprocessing

The quality and relevance of the data is very important for any data driven research. In this study, the timetable and location data of the buses in Nottingham is used to analyse the traffic congestion patterns. This chapter details the data collection process, the types of data gathered, and the preprocessing steps taken to ensure the data's quality and usability. It also highlights the challenges faced during data collection and the techniques employed to address them.[1]

3.1 Data Collection

The data used in this project consists of timetable data and location data of the buses operating in Nottingham. This study specifically focuses on buses run by Nottingham City Transport (NCT).

3.1.1 Bus Location Data

The primary source of bus location data was the Bus Open Data Service's public API, which provides real-time bus location data in XML format. This data contains details such as vehicle locations, journey references and timestamps.

The data was fetched at 15 second intervals for two weeks. This frequency ensured the capture of high quality location data. The data was stored in the HDF5 format, to ensure efficient access to the data.

3.1.2 Bus Timetable Data

The timetable data was obtained from the Bus Open Data Service (BODS) in the TransXChange (TXC) format. TransXChange files are XML-based documents that contain detailed information about bus and coach services, including schedules, routes, stops, operators, and vehicle journeys.

This data was utilized to map the location data of the buses to the planned schedule. The timetable data was pre-processed and stored in an HDF5 format, facilitating efficient querying and integration with the location data.

3.2 Data Preprocessing

Before analysis, the data underwent several preprocessing steps to ensure its quality and suitability for modelling. The key preprocessing steps included:

3.2.1 Parsing XML Data

The XML data from the API was parsed using the python library, 'ElementTree'. The relevant namespaces were defined and key elements were extracted. To ensure data integrity, specific tags and attributes within the XML files were targeted, and the extraction process was carefully validated against the XML schema. The data was then transformed into pandas dataframes for further analysis. The information extracted from the data is as follows:

Parsing Location Data

The attributes extracted from the location data are response time, recorded time, item identifier, line reference, direction, line name, operator, destination reference, destination name, aimed destination

[1] The data for this study can be accessed via https://drive.google.com/drive/folders/19kGE2r_m_ghKlXYqvrnSLeGS3YpQb9_i?usp=sharing

arrival time, vehicle location, block reference, vehicle reference and the vehicle journey reference. A sample of the raw location data can be found in appendix B.

Parsing Timetable Data

The timetable data was divided into the several pandas data frames, each corresponding to different aspects of the bus service:

- **Stop Points:** This contains the stop reference, stop name and locality name for each bus stop.
- **Routes:** This contains the route id and description for each route.
- **Route Links:** Every route section is divided into multiple route links. This contains information about the start and end point of each route link, the track distance and path, route section id and the route link id.
- **Journey Patterns:** This contains information about each bus journey including the destination, direction, line description, route reference and journey pattern id.
- **Journey Pattern Timing Links:** This specifies the aimed travel time between two stop points for each journey pattern.
- **Operators:** This is a list of bus operators. For this study, data from only one operator NCT was used.
- **Service Lines:** This contains information about each bus, including the line id, line name, journey directions, operating periods.
- **Vehicle Journeys:** This contains information about each vehicle journey, including journey code, block number, departure day and time, dead run times and the line reference.

Refer to appendix A for a sample of each data frame.

3.2.2 Data Cleaning

- **Removing Duplicates:** To avoid redundancy and ensure data accuracy, duplicate records were identified and removed from the timetable data. This was particularly important for ensuring that each stop point, route link, bus line and journey was uniquely represented in the dataset.
- **Data Type conversion:** The timestamps were converted to a datetime format to facilitate time-based analysis. The run times were also converted from string to float, in order to specify the travel time in seconds, which can easily be used for detailed analysis.
- **Removing Inconsistent Data:** A filter was applied on the location data, to remove the records where the difference between the response time and the recorded time exceeded two minutes, since these records do now provide the exact location of the bus at the time when the information was received.
- **Handling Missing Data:** Missing values were addressed using different strategies depending on the nature of the data. For instance, missing values for the bus speeds on specific route links were handled by replacing them with the average speed of the bus during that journey.

3.2.3 Data Transformation

After cleaning, the data was transformed to make it suitable for analysis. This involved creating new features and restructuring the data to better align with the analytical goals of the project.

- **Time zone handling:** The timestamps in the location data were originally provided in GMT. These were converted to local UK time to facilitate accurate comparisons between the actual and scheduled departure and arrival times, as outlined in the timetable, which is based on the UK time zone.

- **Time Features Extraction:** The response timestamps in the location data were used to extract specific time-related features, such as the hour of the day or the day of the week. These features were then used for further analysis, focusing on particular days and times.
- **Run time calculation:** The run time for each route link in the timetable data was initially recorded in minutes. It was converted into seconds to enable more accurate speed calculation and analysis possible.
- **Track distance calculation:** The geospatial data for stop points and route links was extracted and processed to calculate the distances and map routes accurately. The 'geopy' library was utilized to compute straight-line distances between coordinates, which were then used to calculate the overall distance of each route link.
- **Aimed Departure and Arrival Time Computation:** The timetable data contains the time when each bus departs towards the first stop of its journey. The dead run time for the bus to reach the first stop and the time taken between all the stops by the bus are also specified in the data. By adding the dead run time to the initial departure time of the bus, the aimed departure time from the first stop was calculated. Furthermore, the aimed departure and arrival time between the stops were calculated by using the run time between those stop points.

3.2.4 Data Integration

Data integration is a critical step in transforming raw data into meaningful insights, especially when dealing with heterogeneous datasets. The integration process involves merging distinct data sources to create a unified dataset that allows for comprehensive analysis.

In this study, the bus location data was integrated with the timetable data to form a comprehensive dataset that can be used for analysis and modelling. The main goal was to synchronize the actual movement of the buses with their scheduled routes, enabling an accurate assessment of punctuality and adherence to timetables.

Processing the Timetable Data

The vehicle journey patterns were merged with the corresponding time taken for each journey pattern and the route links in order to create a comprehensive timetable dataset. The vehicle journey dataset was initially merged with the journey pattern timings. This dataset was then merged with the route links dataset, which provided the geographic track and distance information for each route link. This additional data enabled the calculation of speed for each segment of the journey. Refer to appendix C for a sample of the pre-processed timetable data.

The speed for each segment was calculated by dividing the track distance by the runtime. Special consideration was given to segments with zero runtime to avoid division errors. The missing values were filled with the mean speed for each journey, ensuring that they did not skew the analysis. Finally, the fully integrated and processed timetable dataset was saved in an HDF5 file format.

Integration of the location and timetable data

The core of this analysis involved synchronizing the bus location data with the timetable data. Each location record was mapped to the corresponding timetable route, and the arrival and departure times were estimated based on proximity calculations. Each bus location was compared to the expected locations from the timetable data to estimate the actual arrival and departure time. This involved creating a function to find the closest bus location to each stop point in the timetable route and recording that time as the arrival time of the bus at that stop.

After synchronization, the datasets were merged to form a comprehensive view of bus operations. The final dataset contained:

- The locations of the stop points connecting each route link.
- The scheduled arrival and departure times from the timetable
- The actual arrival and departure times derived from the location data.

Refer to appendix D for more detail.

3.2.5 Data Storage

After cleaning and preprocessing, the data was stored in a structured format suitable for analysis. The data was saved as HDF5 files, a format known for its efficiency in storing large datasets with complex structures. This format was chosen because it allows efficient data retrieval and manipulation during the analysis phase. It also requires less storage space compared to csv format.

3.2.6 Challenges in Data Collection and Preprocessing

The data collection and preprocessing stages presented several challenges, which are addressed as follows:

- **Real time data collection:** Collecting real time location data over a two week period posed significant challenges, as it required continuous script execution without interruptions. Maintaining uninterrupted data capture over such an extended period was difficult, with potential risks of data loss due to technical issues, such as system failures or network disruptions. Ensuring the reliability and completeness of the data collection process was a critical concern throughout this stage.
- **Resolving data inconsistencies:** There were inconsistencies in the data formats, units and timestamps, which complicated the integration and analysis process. These inconsistencies needed to be carefully standardized to create a uniform dataset. This involved converting various formats and units to a common standard and aligning timestamps to ensure accurate temporal analysis.
- **Time Discrepancies:** Discrepancies between the response time and recorded time were identified, which could lead to inaccuracies in the analysis. These discrepancies were handled by removing all the records where the difference between response and recorded time was more than two minutes.
- **Incorrect run time affecting speed calculation:** Some route segments recorded a travel time of zero between stops, which resulted in errors in speed calculations. These errors could skey the overall analysis of bus performance. To correct this, the affected speed values were replaced with the mean speed of the bus during that journey, ensuring that the speed data reflected more accurate and realistic values.
- **Addressing Spatial Discrepancies:** The spatial data presented challenges in accurately matching the GPS coordinates of moving buses with the fixed stop locations in the timetable. Since buses rarely stop at the exact GPS coordinates specified in the timetable, the closest recorded bus location to each stop point was used to calculate the bus arrival time.
- **Handling Large Datasets:** The large volume of data collected posed challenges in terms of storage and processing. To manage this, the data was stored in an HDF5 format, which requires less storage space than a csv file and is efficient for data retrieval. Synchronizing the location data with timetable data required considerable computational resources. To address this, parallel processing was employed, using a ThreadPoolExecutor. The data was divided into manageable chunks which were processed concurrently. This approach helped to optimize performance and reduce processing time.

Chapter 4: Identification of Congested Routes

Identifying congested routes is a critical step in understanding traffic flow and managing urban congestion. This chapter outlines the methodology used to pinpoint the most congested routes within Nottingham. The primary goal is to assess the average speed and bus density across different routes, identify congestion patterns, and classify congested route segments based on these patterns.

4.1 Calculation of Average Speed and Bus Density

The key metrics of interest in this analysis are average speed and bus density. The calculations were made for the specific time interval under consideration. These calculations provided insights into the traffic situation on each route link during the specified time interval. The average speed and bus density were calculated as follows:

4.1.1 Average Speed Calculation

The run time for a bus on each route link was calculated by measuring the time taken by the bus to travel between the two stops of that route link. Speed was then calculated by dividing the track distance by the run time. To obtain the average speed for each route link, the speeds of all the buses travelling on that route link were averaged.

4.1.2 Bus Density Calculation

Bus density is defined as the number of buses per unit distance on each route link. It was calculated using the following formula:

$$Bus\ Density = \frac{Number\ of\ buses}{Track\ Distance}$$

The data was grouped by route link and the number of buses on each route link were counted. The total number of buses passing through each route link was divided by the length of that route link to get the bus density for that route link. Refer to appendix E for sample data.

4.2 Exploratory Data Analysis

4.2.1 Distribution of Average Speed and Bus Density

The distribution of average speed and bus density was explored using histograms and kernel density estimation (KDE). Figure 4.1 provides a visual representation of the distribution of the average speed and bus density across various route links.

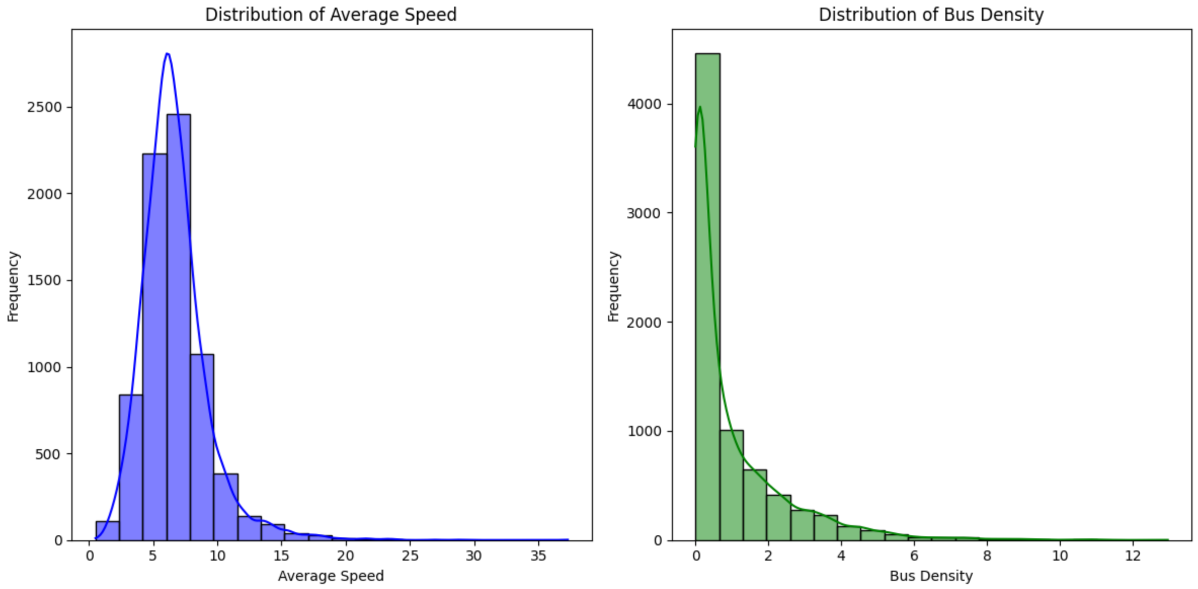


Figure 4.1: Distribution of Average Bus Speed and Bus Density

The histogram on the left side of Figure 4.1 represents the distribution of average speed. It indicates a right-skewed distribution. The majority of the routes have average speeds clustered between 5 m/s and 10 m/s, with a significant drop-off as the speed increases. This suggests that most buses on the analyzed routes travel at relatively low speeds, which could indicate congestion or frequent stops. The KDE curve confirms this with a sharp peak at around 7-8 m/s mark, highlighting that this is the most common speed range across the dataset.

The right side of Figure 4.1 displays the distribution of bus density, which also exhibits a right-skewed pattern. Most route links have a low bus density, typically around 0.5 bus per meter. This is evident from peak in the histogram and the corresponding KDE curve, which shows a sharp decline as the bus density increases. Only a small number of route links have a high bus density, which may suggest that these links are more congested.

The skewed distributions in both histograms suggest that traffic conditions are not uniform across the routes. The concentration of average speeds in the lower range and the majority of route links having low bus densities might indicate that only specific areas experience high traffic volumes leading to reduced speeds and potential congestion. The analysis implies that most congested route links are likely those with low average speeds and high bus densities.

4.2.2 Box Plot Analysis

Box plots were used to visualize the spread and identify potential outliers in the average speed and bus density data. These visualizations helped in understanding the central tendency and variability within the data. Figure 4.2 presents the box plots for average speed and bus density across various route links.

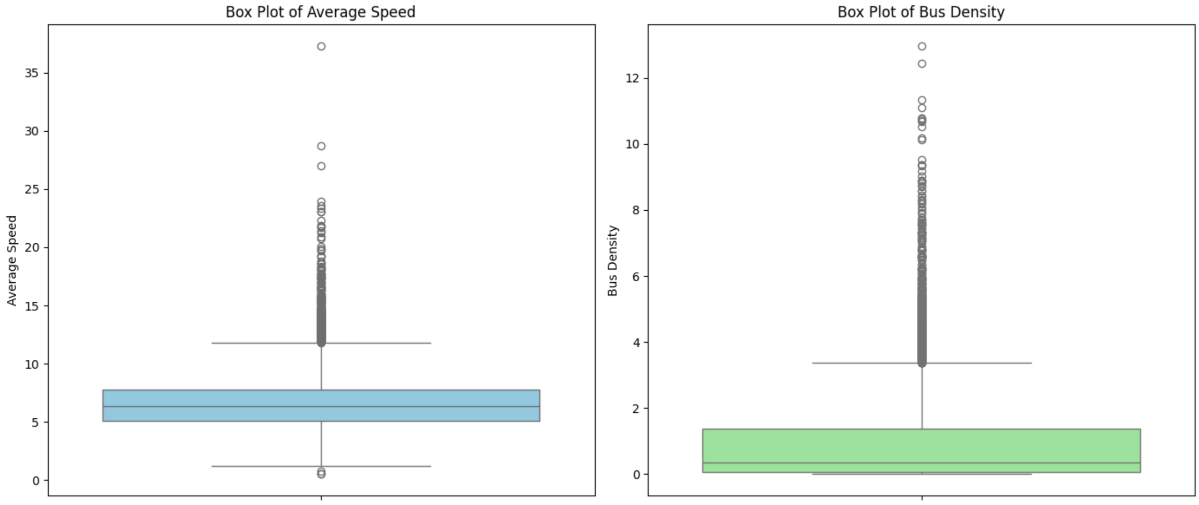


Figure 4.2: Box plots for Average Bus Speed and Bus Density

The box plot on the left side of Figure 4.2 shows the boxplot for average speed. The median average speed is observed to be around 7 m/s, with the interquartile range (IQR) spanning from approximately 5 to 8 m/s. This confirms that the majority of route links have average speeds within this range, consistent with the findings from the histogram in Figure 4.1. However the boxplot also reveals a significant number of outliers, with some route links having very high average speeds, some of them being even more than 20 m/s.

The box plot on the right side of Figure 4.2 shows the boxplot for bus density. The median bus density is very low, close to 0.25 buses per meter, with the IQR ranging from 0.1 to 1.5. This indicates that on most route links, bus density remains low, which could imply less congestion. This confirms the findings from Figure 4.1. Similar to the average speed box plot, the bus density box plot also highlights several outliers.

4.2.3 Outlier Removal

To ensure that the analysis is not skewed by the extreme values, outliers in the average speed were identified and removed based on z-scores. The threshold for identifying outliers was set to a z-score of 5, ensuring that only significant deviations from the mean were considered outliers. Figure 4.3 (a) and Figure 4.3 (b) show the histogram and box plot of the average speeds after outlier removal.

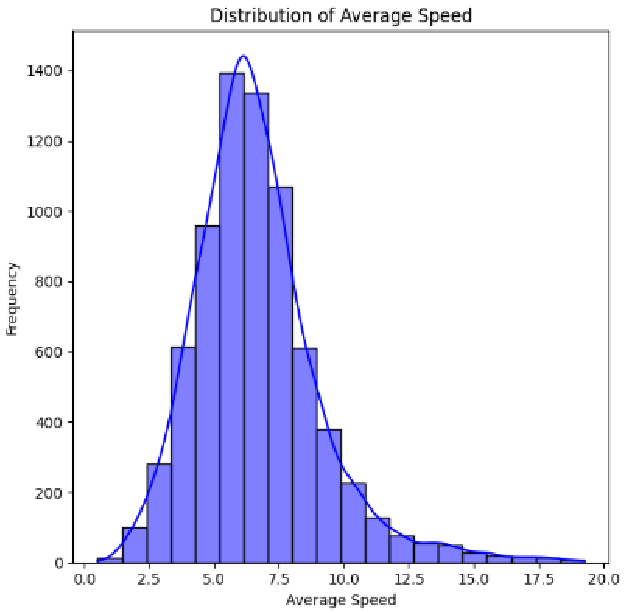


Figure 4.3 (a): Histogram for Average Speed after Outlier Removal

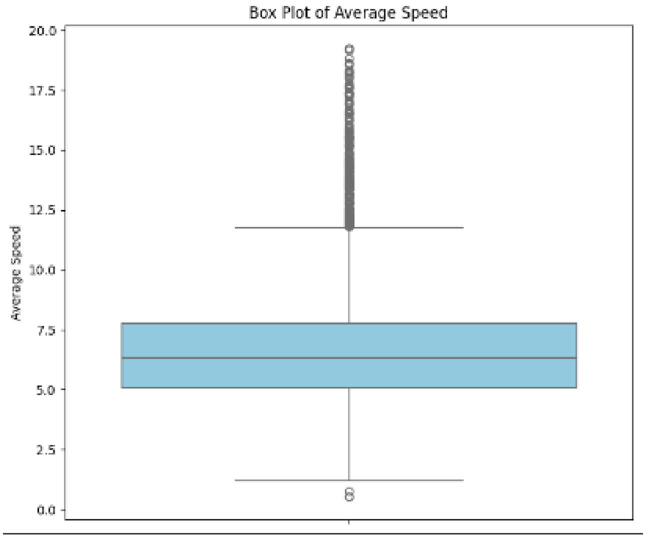


Figure 4.3 (b): Box plot for Average Speed after Outlier Removal

4.3 Cluster Analysis

Clustering was performed to classify route segments based on average speed and bus density, identifying areas with varying levels of congestion.

4.3.1 k-Means Clustering

The k-Means algorithm was employed to classify route segments into clusters representing different congestion levels. Different cluster configurations were tested to determine the most suitable number of clusters that effectively differentiated the route links on the basis of average speed and bus density. The features were normalized before applying the clustering algorithms to ensure that both the features contributed equally to the clustering process.

k = 2

Initially, the clustering was performed with two clusters, where the algorithm divided the route links into two clusters. Figure 4.4 shows that the clusters are clearly distinguishable. The yellow cluster primarily represents route links with lower average speeds and higher bus densities, indicating congested routes. The purple cluster represents route links with higher average speeds and lower bus densities, which can be considered less congested.

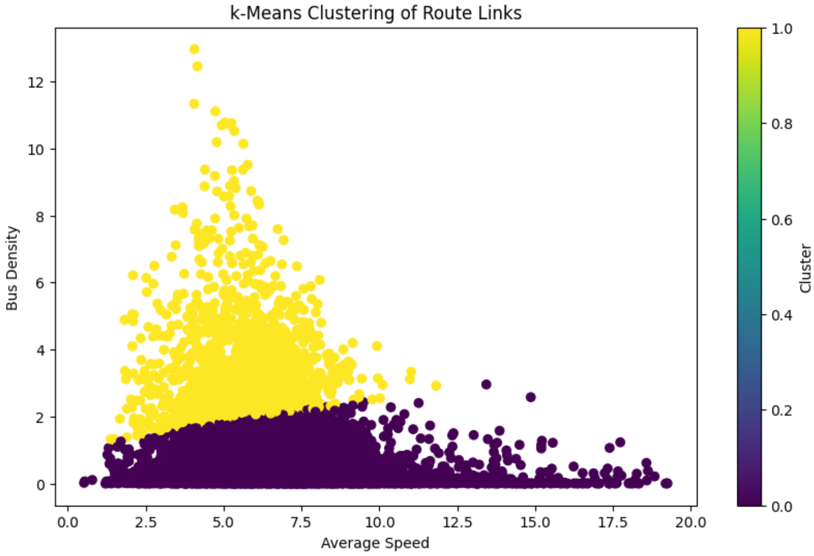


Figure 4.4: Scatter Plot for k-means with k = 2

k = 3

Increasing the number of clusters to three allowed a more granular classification of the route links. Figure 4.5 shows that the data points are divided into three distinct clusters. The yellow cluster represents highly congested route links, with low speeds and high bus densities. The purple cluster includes route links with slightly better traffic conditions. The teal cluster captures the least congested route links with relatively high average speeds and low bus density.

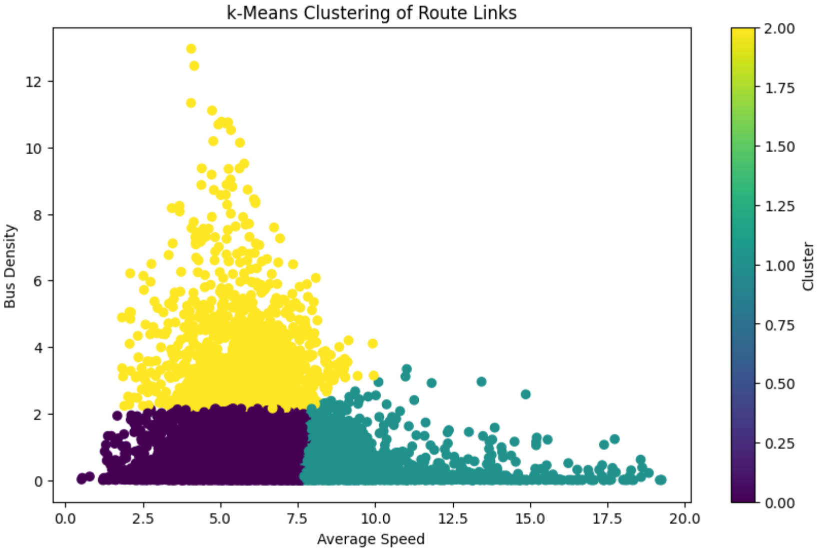


Figure 4.5: Scatter Plot for k-means with k = 3

k = 4

Finally, the k-means algorithm was tested with four clusters to see if a more detailed segmentation could yield insights. However, this increased the complexity without providing significant additional value over the k = 3 model. The four clusters can be seen in the Figure 4.6.

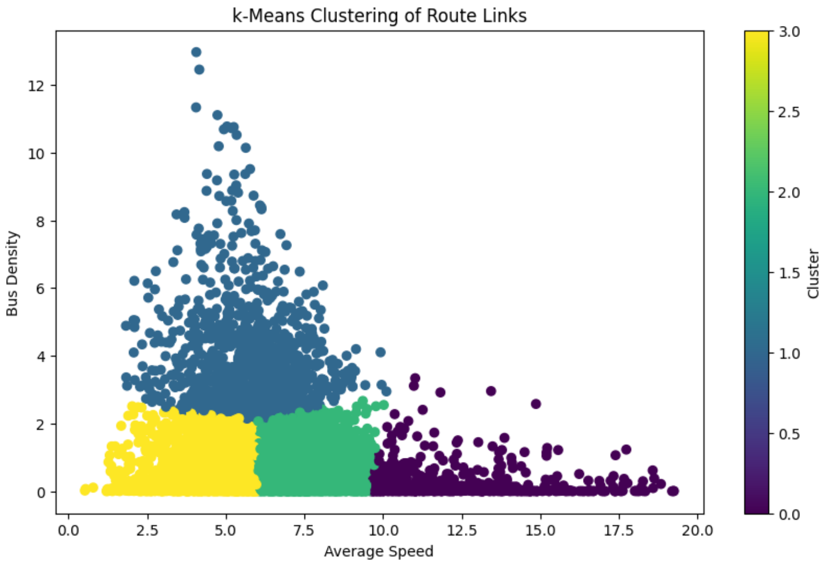


Figure 4.6: Scatter Plot for k-means with k = 4

4.3.2 DBSCAN Clustering

DBSCAN was also employed to classify route links based on the same features i.e., average speed and bus density. Unlike k-means, which requires specification of the number of clusters in advance, DBSCAN identifies the clusters based on the density of the data points, which makes it particularly effective for classifying clusters of varying shapes and sizes and identifying noise points.

Two key parameters were set for the DBSCAN algorithm:

- **Epsilon (ϵ):** This parameter defines the maximum distance between two points to be considered as part of the same neighbourhood. The value of epsilon was set to 0.5

- **Minimum Samples:** This specifies the minimum number of points required to form a cluster. Points that do not meet this criterion are classified as noise. The value of minimum samples was set to 2.

After applying DBSCAN, the route links were divided into three clusters. Figure 4.7 shows that the yellow cluster represents congested route links with a low average speed and high bus density. The purple cluster represents route links that are not congested and have high average speed and low bus density, while the teal cluster represents route links with medium level of congestion.

However, it can be observed that most of the route links were classified as medium congestion (teal), while the low (purple) and high (yellow) congestion clusters contained only two route links each. This shows that while DBSCAN offers the advantage of detecting clusters of various shapes and sizes, its performance in this analysis has potential limitations. Therefore, k-means may be more suitable for this congestion analysis, where a more balanced distribution of clusters is desired.

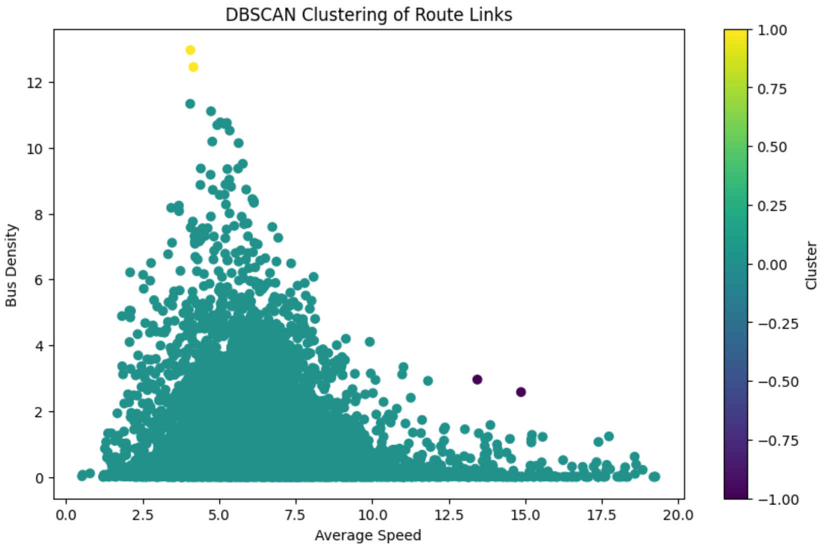


Figure 4.7: Scatter Plot for DBSCAN Clustering

Chapter 5: Analysis of Congested Routes

In the previous chapter, the k-means algorithm with three clusters was identified as the most suitable approach for this study. This chapter provides the analysis of congestion patterns observed across various bus routes. The findings and visualizations produced during this analysis are presented below.

5.1 Distribution of Congestion Levels

The k-means clustering algorithm, with k = 3 was used to categorize the route links into three distinct congestion levels: low, medium and high. The results of this classification were visualized using a pie chart, which effectively illustrates the proportion of route links within each congestion category.

The pie chart provides a comprehensive overview of the distribution of congestion levels across the route network, highlighting the relative prevalence of each category. This visualization serves as a foundation for a more in-depth analysis in subsequent sections.

As shown in Figure 5.1, 24.5% route links were classified as experiencing low congestion, 60.3% as medium congestion and 15.3% as highly congested. This distribution highlights the varying degrees of congestion present throughout the network and helps identify areas that may require targeted interventions. The exact number of route links in each category is detailed in Table 5.1.

Table 5.1: Number of route links in each congestion category

	Number of Route Links
High Congestion	1128
Medium Congestion	4454
Low Congestion	1810

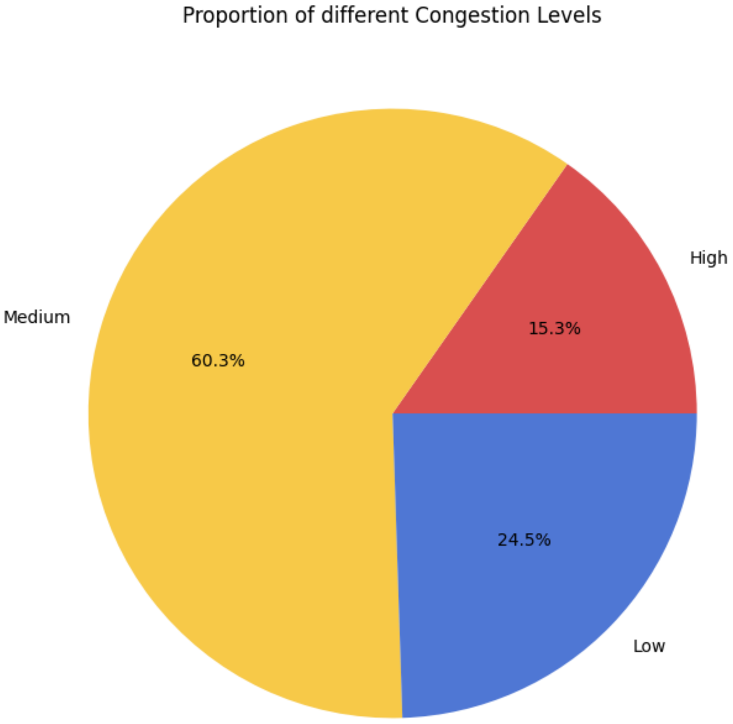


Figure 5.1: Pie Chart to illustrate the proportion of route links under each congestion level

5.2 Visualization of Congestion Levels on the Map

To gain a deeper understanding of the spatial distribution of congestion, the route links were mapped using the python library ‘folium’. Each route segment was color-coded based on its congestion level: red for high congestion, yellow for medium congestion, and blue for low congestion.

It is important to note that some route links may experience congestion in one direction but not in the opposite direction. These route links are overlapped to reflect this difference. For instance, in Figure 5.2, some route links are depicted in orange. This indicates that the link is categorized as red (high congestion) in one direction and yellow (medium congestion) in the other.

This map serves as a powerful visual tool to identify congestion hotspots and areas with severe congestion. It not only shows the geographical spread of congestion but also enables stakeholders to pinpoint specific routes or areas that may require intervention or further investigation.

While the map might not confirm the direction of traffic flow experiencing congestion, it provides valuable insights into the general areas that are affected. This helps in understanding the broader patterns of congestion.

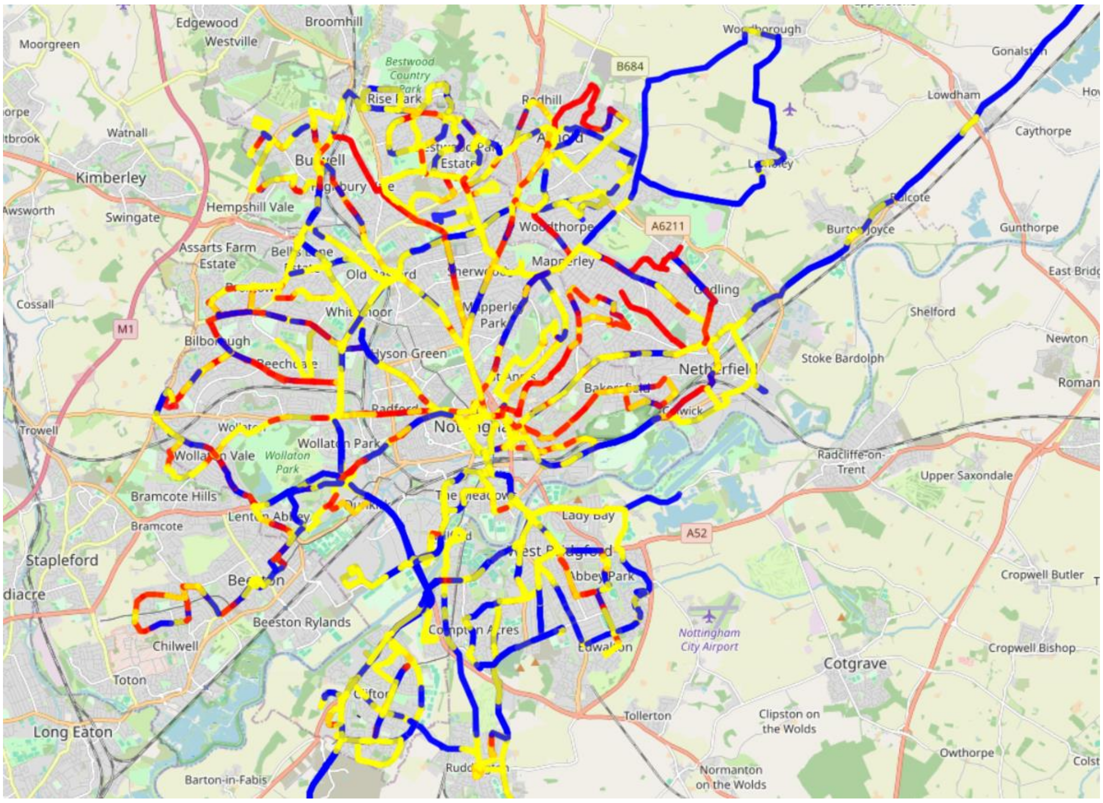


Figure 5.2: Visualization of the route links with different congestion levels

5.3 Analysis of Congestion by Route Section

Each route section comprises multiple route links. After categorizing these links into different congestion levels, a more detailed analysis was conducted by calculating the percentage of congestion for each route section. This was done by comparing the number of congested route links to the total number of route links within each section. A sample can be found in appendix F.

$$Congestion\ Percentage = \frac{Number\ of\ Congested\ Route}{Total\ Number\ of\ Route\ Links}$$

The bar chart in Figure 5.3 shows the percentage of congestion for each route section. Only sections with more than 20% congestion were included in the chart, allowing for a focused analysis on the most significantly affected areas.

As seen in Figure 5.3, several route sections show very high congestion levels, with some even approaching 100%. This indicates that nearly all route links in these route sections are experiencing congestion. Identifying these heavily congested route sections is crucial for prioritizing intervention and implementing measures to alleviate traffic congestion.

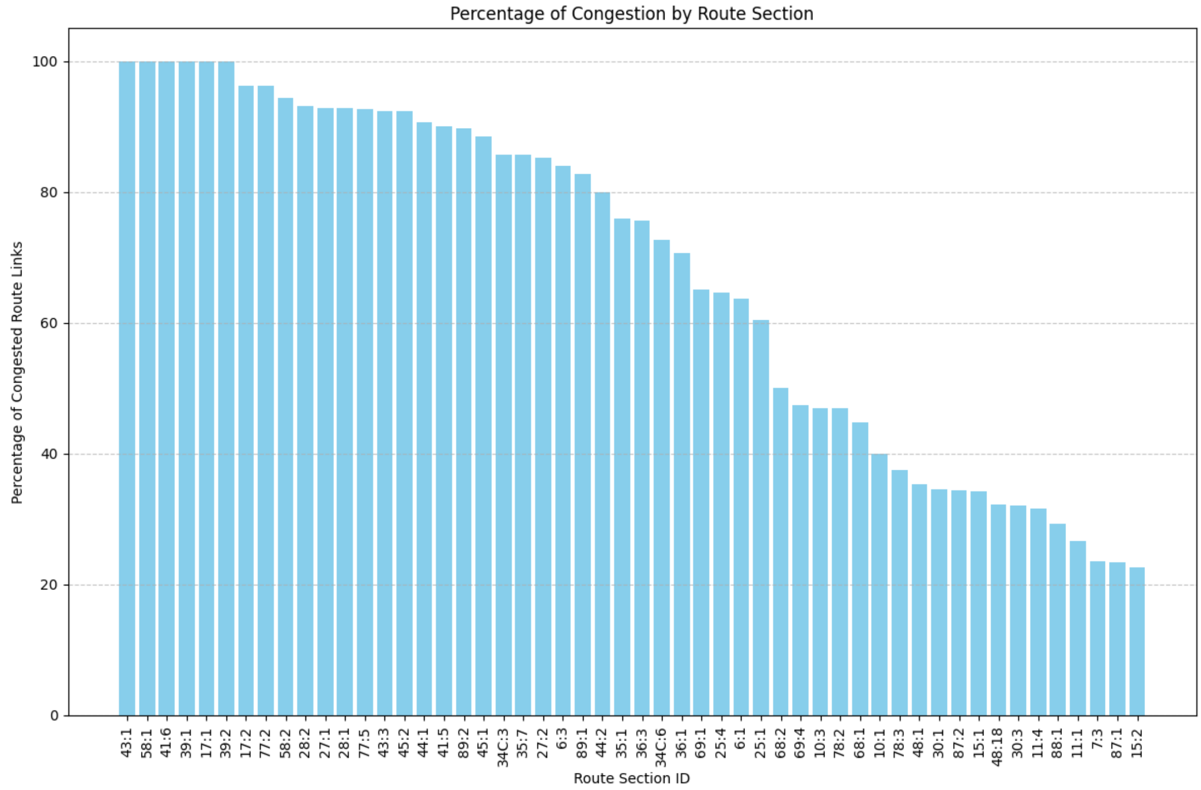


Figure 5.3: Bar Chart to illustrate the congestion percentage of each route

To refine the analysis further, a filter was applied to identify the route sections with more than 50% congestion. This revealed 33 such route sections that had significantly high congestion levels exceeding 50%. By targeting these routes, more optimized pathways with reduced congestion can be explored.

Figure 5.4 provides a closer look at the congestion levels on route section 30:3, which was found to be 32.1% congested. The visualization shows that approximately 32% of this route section is marked in red. This validates the calculated congestion percentage.

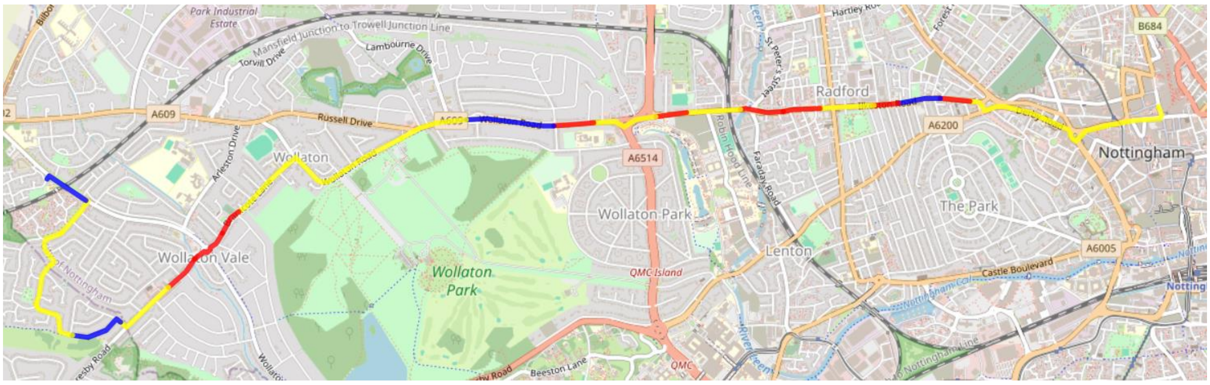


Figure 5.4: Visualization of congestion levels across Route 30:3

5.4 Congestion Analysis by Day of the Week

The congestion levels of route links were analysed separately for each day of the week and visualized using a heatmap. This approach enabled us to detect variations in congestion patterns throughout the week, revealing days with consistently high congestion or those with significant variability.

Figure 5.5 illustrates the congestion intensity across the week, highlighting patterns that might not be apparent from daily analysis alone. This temporal perspective is essential for understanding how congestion fluctuates and for planning interventions that account for these variations.

The heatmap in Figure 5.5 reveals that the congestion levels vary by day, with notable differences between weekdays and weekends. For some routes, congestion is higher on weekdays, likely due to commuter traffic related to jobs and offices, while weekends show reduced congestion. Conversely, some routes maintain consistent congestion levels throughout the week, indicating uniform traffic patterns irrespective days.

Additionally, the heatmap shows that certain routes are inactive on specific days. For example, Figure 5.6 shows that Route 49:4 operates only on Saturdays, whereas Routes 49:1, 49:2, 49:3, 49A:1, 49A:2, 49A:3, and 49A:4 are only operational during weekdays. Furthermore, Routes 48:1 and 49:18 experience higher congestion from Monday to Saturday but have reduced congestion on Sundays. This insight highlights the varying demand for transportation services throughout the week and can inform targeted management strategies.

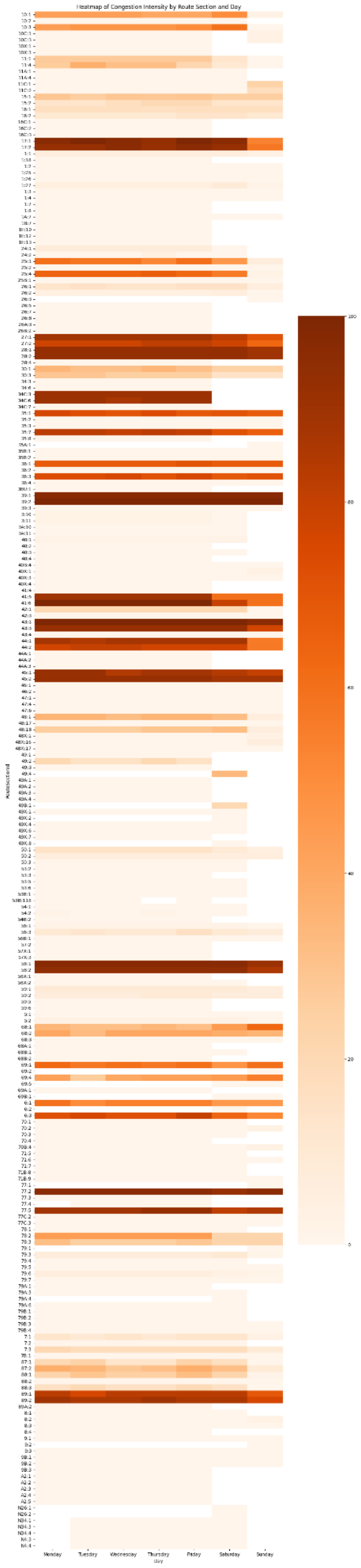


Figure 5.5: Heatmap for congestion intensity for each route across the week

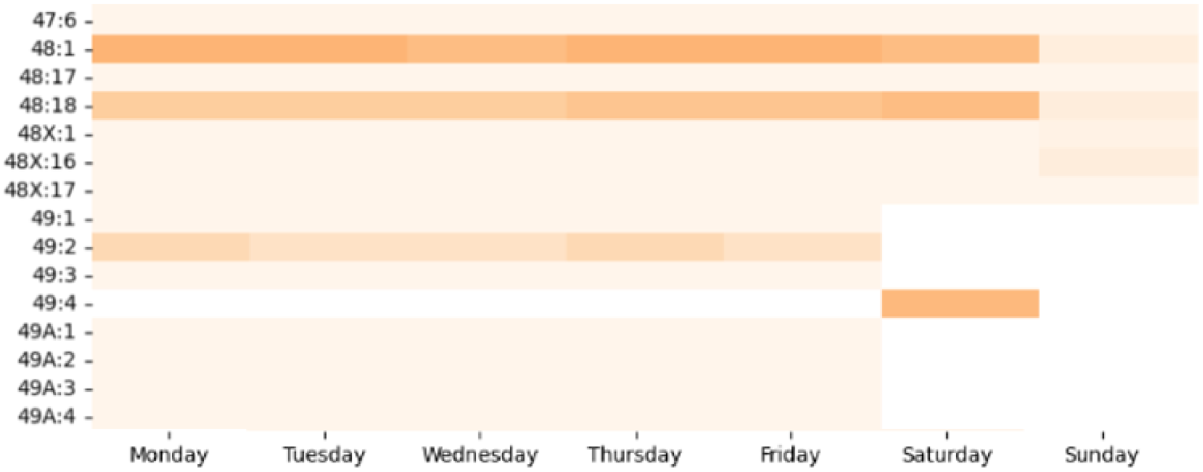


Figure 5.6: Heatmap by weekdays for specific routes

5.5 Congestion Analysis by Time Interval

After analysing the congestion levels for each day of the week, the congestion was examined during different intervals throughout the day. A heatmap by time interval was generated to get a detailed visualization of congestion patterns across various times of the day. The objective was to identify specific times of the day when congestion is most severe and to understand how it varies across different routes throughout the day.

The day was divided into 12 two-hour intervals, ranging from ’00:00 – 02:00’ to ’22:00 – 00:00’. These intervals were chosen to capture both peak and off-peak hours, providing a comprehensive view of the traffic flow over the entire day.

Figure 5.7 shows that most routes experience peak congestion between 8:00 and 18:00, aligning with typical working hours when traffic volume is at its highest. In contrast, off-peak hours particularly after 18:00 show significantly lower congestion levels.

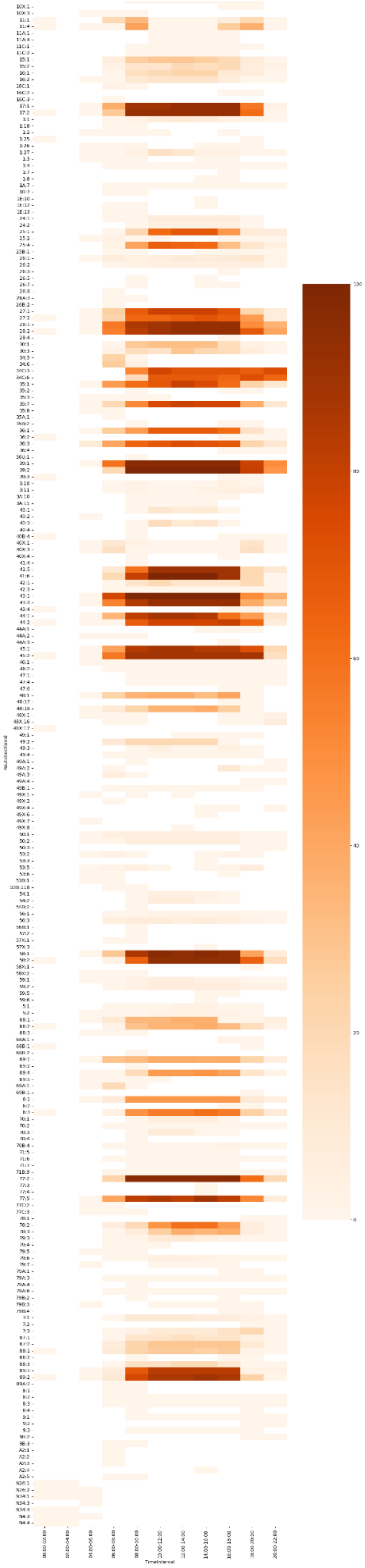


Figure 5.7: Heatmap for congestion intensity across different intervals of the day

A similar pattern can be observed in the line graph in Figure 5.8 which shows the trend in the congestion levels for different routes. Only the routes with relatively higher congestion levels are plotted to reduce overlapping and improve interpretability. It can clearly be seen that most of the routes have a higher congestion level during the 8:00 – 18:00 interval.

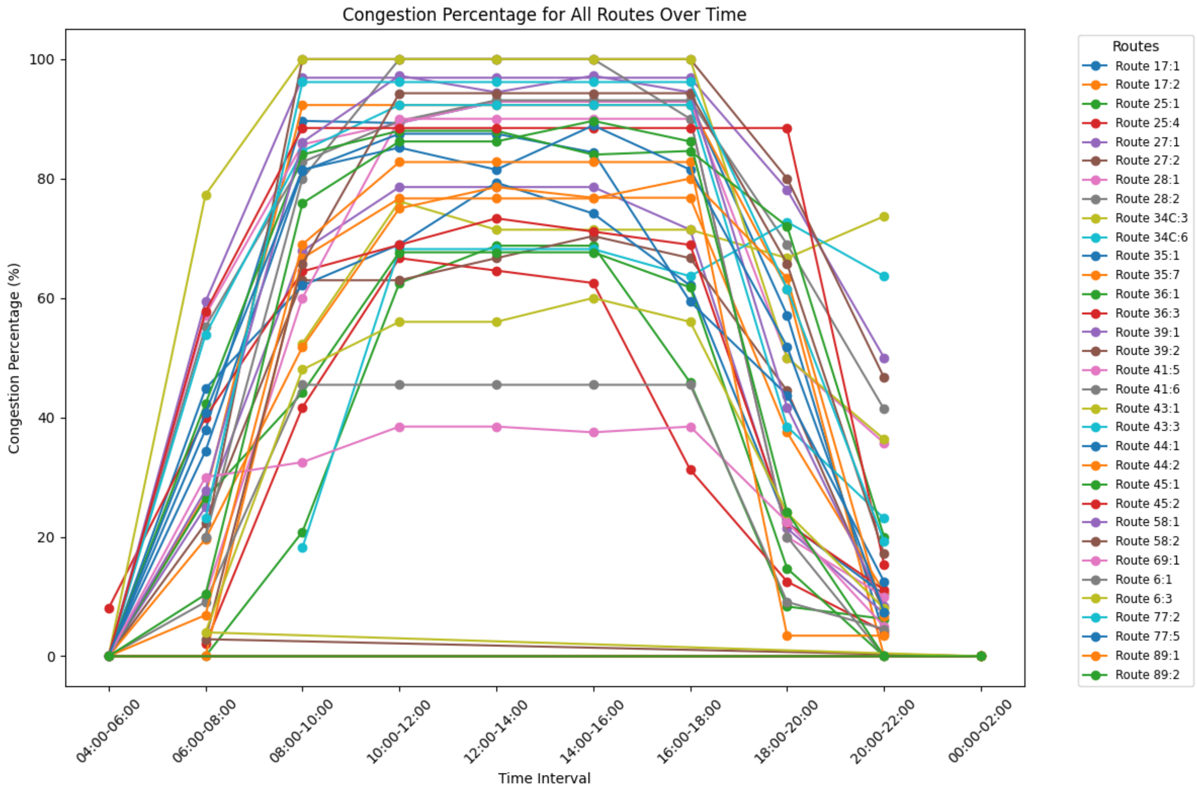


Figure 5.8: Line Graph for Congestion Percentage throughout the day

While most of the routes display similar congestion patterns throughout the day, some routes exhibit specific congestion peaks during certain time intervals. For instance, Figure 5.8 highlights that Route 11:4 experiences relatively higher congestion levels during the intervals 08:00 – 10:00 and 16:00 – 20:00, which are the typical commuting hours.

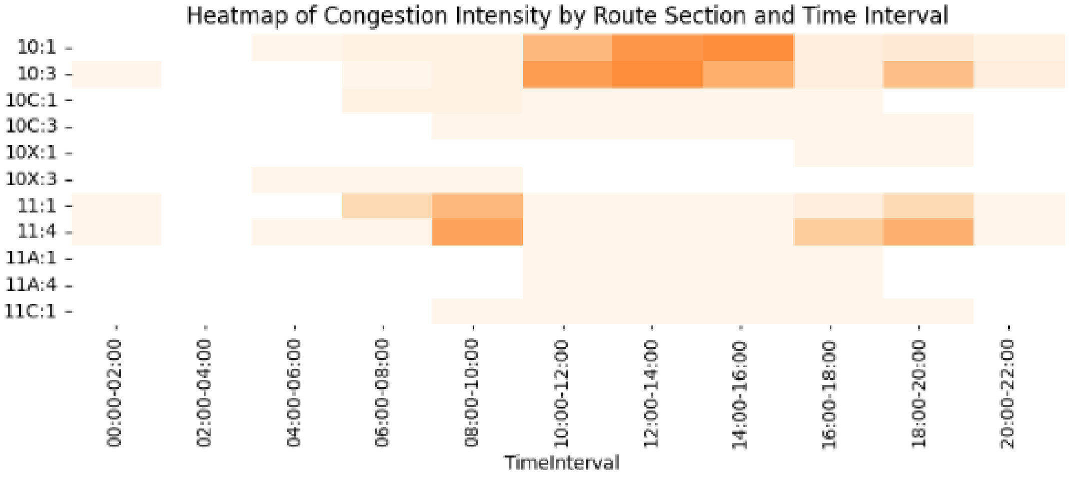


Figure 5.9: Heatmap by time interval for specific routes

The congestion pattern of Route 11:4 can be observed in the line plot shown in Figure 5.10, which illustrates the congestion peaks of the route. We can see that the route is not congested for most part of the day, but experiences peak congestion during the commuting hours.

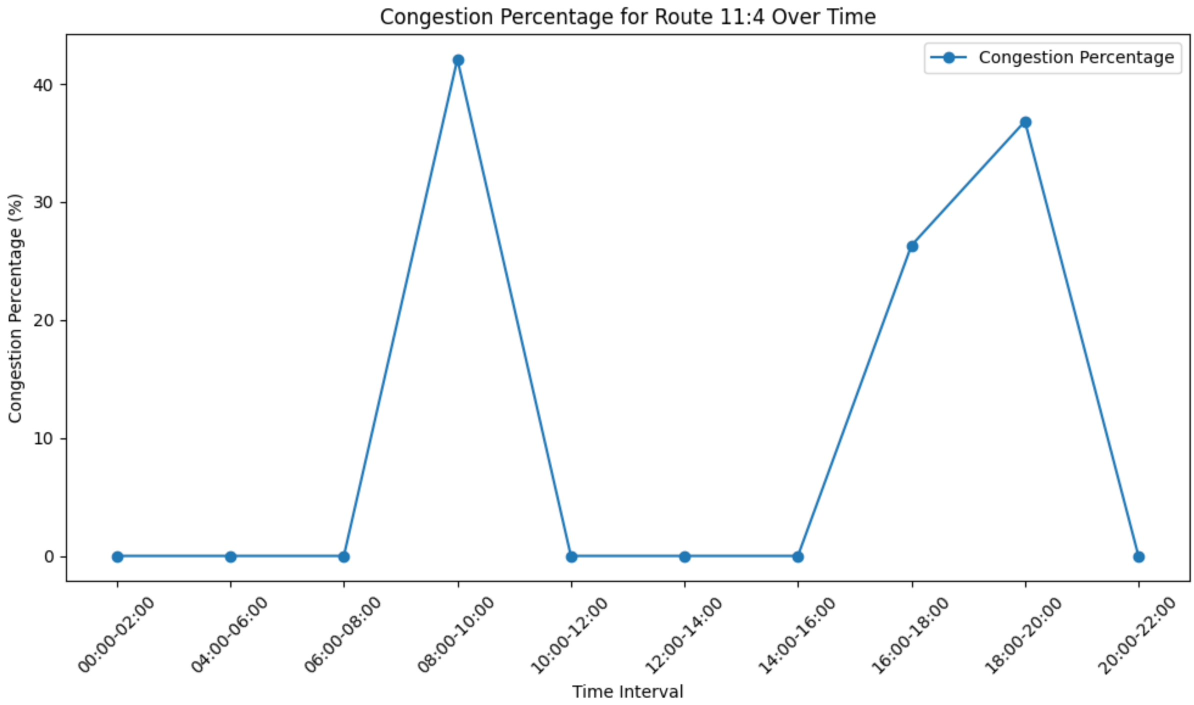


Figure 5.10: Line plot for congestion patterns of Route 11:4 during different time intervals

Chapter 6: Route Optimization and Evaluation

This chapter explores the methodologies and techniques used to optimize bus routes by proposing and evaluating alternative paths. By leveraging graph-based algorithms, the aim is to mitigate congestion on existing routes by suggesting more efficient alternatives. The chapter will detail the process of generating these alternative routes, the criteria and metrics used for their evaluation, and the comprehensive analysis between the proposed alternative routes and the current main routes.

6.1 Proposing Alternatives

The primary objective of route optimization is to find alternative routes that potentially improve travel efficiency. This may involve suggesting new bus stops to divert the existing paths. However, for this study, the focus is solely on proposing alternative routes that utilize the existing bus stops, ensuring minimal disruption to the current infrastructure.

6.1.1 Creating a Graph for Route Analysis

A directed graph was constructed using the python library ‘NetworkX’, to model the bus routes. In this graph, each node represents a bus stop, while each directed edge represents a route link between two stops. The weight of each edge is determined by the average run time between the respective stops, providing a measure of travel time. This graph-based representation allows for the application of efficient graph algorithms to analyse and optimize the network and find alternative routes.

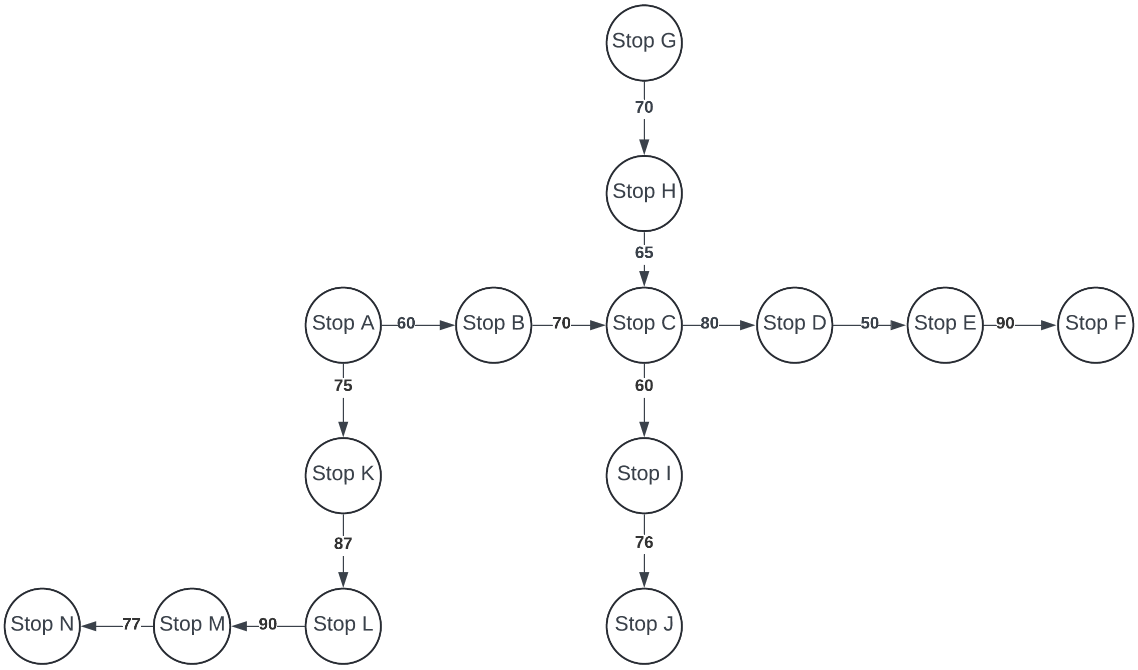


Figure 6.1: Directed graph representation of bus routes, where node denotes a bus stop and each edge represents the travel time between stops

6.1.2 Finding Alternative Routes

Alternative routes with reduced travel time were proposed for all the routes having congestion percentage greater than 70%. In cases where a suitable alternative route could not be found using the existing bus stops, the original route was retained.

To identify an alternative route for a given route section, the start and end points of the route were first determined. Dijkstra’s algorithm was then applied to find the shortest path, in terms of runtime, between these two points. The result was an optimized route represented as a sequence of bus stops.

For example, an alternative route was proposed for Route 36:1, which had a congestion percentage of 70.59. The alternative route reduced the travel time significantly, from 37.2633 minutes to 20.408 minutes, offering a more efficient path. The specific paths for the actual and alternative routes are detailed below.

Table 6.1: Comparison of Main and Alternative Route for the Route 36:1.

	Run Time (minutes)	**Sequence of Stops**
Main Route	37.2633	3390J4 → 3390E2 → 3390A4 → 3390Y4 → 3390CC04 → 3390LE07 → 3390LE08 → 3390LE09 → 3390LE10 → 3390LE11 → 3390LE12 → 3390QM04 → 3390UN15 → 3390UN16 → 3390UN17 → 3390UN18 → 3390UN19 → 3390UN20 → 3390UN21 → 3390UN22 → 3390UN23 → 3300BR0524 → 3300BR0523 → 3300BR0241 → 3300BR0213 → 3300BR0613 → 3300BR0095 → 3300BR0097 → 3300BR0080 → 3300BR0075 → 3300BR0503 → 3300BR0026 → 3300BR0028 → 3300BR0626 → 3300BR0197
Alternative Route	20.408	3390J4 → 3390E2 → 3390A3 → 3390Y4 → 3390CC04 → 3390LE07 → 3390LE08 → 3390LE09 → 3390LE10 → 3390LE11 → 3390LE12 → 3390QM04 → 3390UN43 → 3390UN46 → 3390UN44 → 3390UN45 → 3390UN33 → 3300BR0524 → 3300BR0523 → 3300BR0241 → 3300BR0213 → 3300BR0613 → 3300BR0095 → 3300BR0097 → 3300BR0080 → 3300BR0075 → 3300BR0503 → 3300BR0026 → 3300BR0028 → 3300BR0626 → 3300BR0197

Figure 6.2 provides a visual representation of both the actual and alternative routes, highlighting the differences and improvements. The congested route links are represented by red colour, while blue and yellow represents low and moderate congestion respectively. The alternative route is represented by green colour. It can be seen that the alternative route overlaps with the original route for most of the part. However, it avoids a congested route link to reduce the travel time and make the journey more efficient.

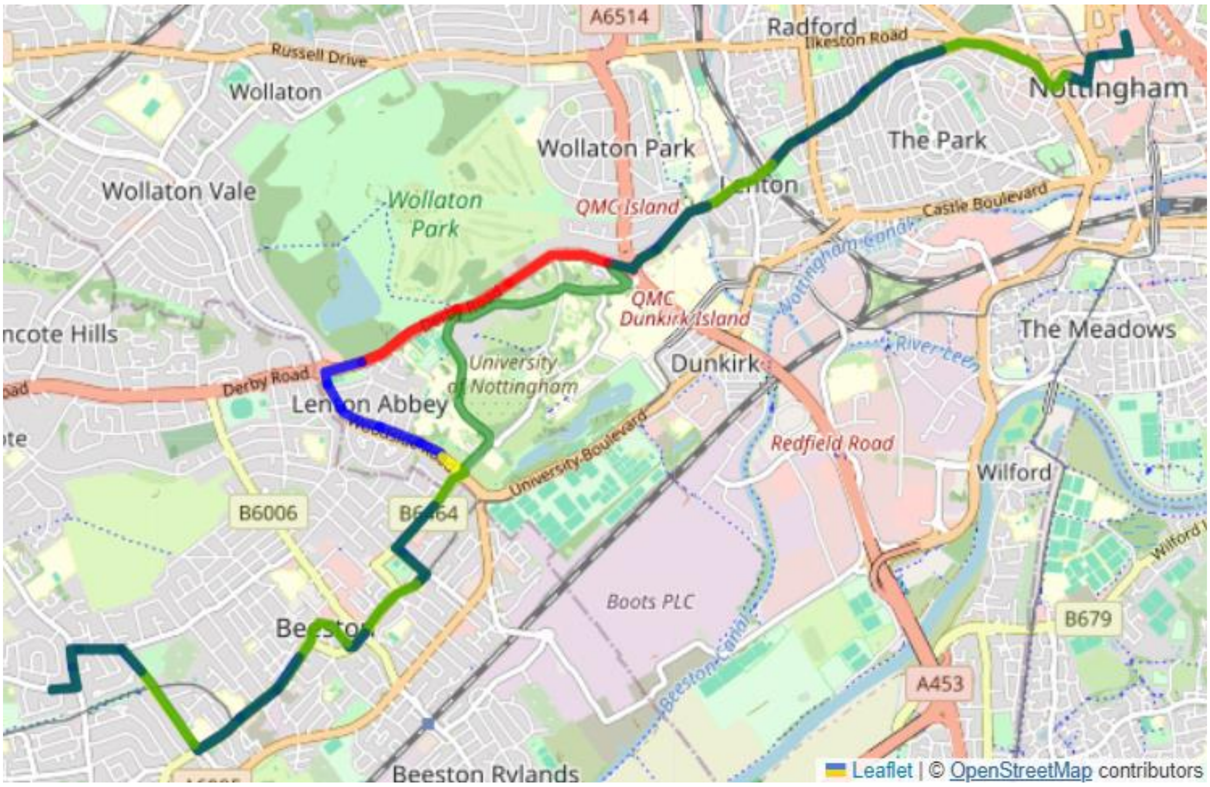


Figure 6.2: Visual representation of the route 36:1 and the proposed alternative route

6.2 Evaluating Alternative Routes

Alternative routes were proposed for all the routes with a congestion percentage greater than 70%. To assess the effectiveness of these proposed routes, various evaluation metrics were employed. The primary focus was on comparing the run times for the main and alternative routes.

6.2.1 Run Time Comparison

The difference in run times between the main and alternative routes serves as a critical metric for assessing the efficiency of the alternative route. This was assessed by calculating the mean run time for both the main and alternative routes, followed by evaluating the percentage decrease in run time. The analysis revealed a significant improvement, with a 12.89% reduction in run time for the alternative routes.

Table 6.2: Run Time Comparison for Main and Alternative Routes

Main Route Run Time (minutes)	**Alternative Route Run Time (minutes)**	**Percentage Decrease**
1671.57	1456.06	12.89%

6.2.2 Statistical Evaluation

A paired t-test was used to determine if the differences in the run times were statistically significant. This approach takes into account the natural pairing of main and alternative routes for each route section. The t-test was applied to the run times after identifying alternative routes for all the routes with a congestion percentage greater than 70%.

The paired t-test results indicated strong statistical significance, with a t-statistic of 3.463 and a p-value of 0.00174, which is well below the conventional threshold of 0.05. This suggests that the observed difference in run times between the main and alternative routes is unlikely to be due to random chance.

These findings confirm that the alternative routes offer a meaningful reduction in run times compared to the main routes. The significant –value from the paired t-test strongly supports the efficacy of the route optimization algorithm, demonstrating that the proposed alternative routes effectively reduce congestion and improve travel efficiency.

Chapter 7: Conclusion and Future Work

This study demonstrates the potential of using clustering and graph-based algorithms to analyse and optimize bus routes in response to congestion. By leveraging these methodologies, bus operators can make informed decisions to improve the efficiency and effectiveness of their bus networks. The insights gained and the proposed alternative routes offer a valuable contribution to enhancing urban transportation systems.

The findings of this research underscore the importance of ongoing research and adaptation in transportation planning. As cities evolve and traffic patterns change, continuous analysis and optimization will be crucial for maintaining an efficient and responsive public transportation system. Through the application of innovative analytical techniques and a commitment to addressing congestion, we can work towards creating more efficient, reliable and passenger friendly transportation networks for urban environments.

This chapter summarizes the key findings of the study, discusses the implications of these findings, and outlines the potential avenues for future research and improvements in bus route optimization.

7.1 Summary of Findings

This study has successfully employed k-means clustering and graph-based algorithms to analyse and optimize the Nottingham City Transport (NCT) bus routes, with a focus on traffic congestion. The key findings are summarized as follows:

7.1.1 Congestion Analysis

The k-means clustering algorithm with three clusters (low, medium, and high) effectively categorized route links into distinct congestion levels. The analysis revealed that 24.5% of the route links experienced low congestion, 60.3% had medium congestion, and 15.3% were highly congested.

The spatial distribution of congestion, visualized using a map, highlighted specific hotspots and areas with severe congestion. The congestion patterns varied across different days of the week and time intervals, with peak congestion typically occurring during weekday working hours.

7.1.2 Route Optimization

Alternative routes were proposed for routes with congestion levels greater than 70% using Dijkstra’s algorithm. These alternative routes showed significant improvements in travel time, with an average reduction of 12.89%.

The effectiveness of the proposed routes was confirmed through a paired t-test, which demonstrated a statistically significant reduction in run times.

7.2 Implications

The findings of this study provide valuable insights into the congestion patterns within the bus route network and offer practical solutions for improving travel efficiency. The analysis and proposed alternative route can assist transit authorities in:

Targeted Interventions: Identifying specific congested routes and areas where targeted interventions could alleviate congestion and improve service quality.

Operational Planning: Adjusting schedules and resource allocation based on congestion patterns observed during different times of the day and days of the week.

Route Optimization: Implementing alternative routes that reduce travel time and enhance the efficiency of the bus network, potentially leading to better overall service and increased passenger satisfaction.

7.3 Future Work

While this study has provided a solid foundation for understanding and optimizing bus routes, several areas warrant further investigation:

Dynamic Traffic Conditions: Future research could incorporate real-time traffic data to account for dynamic changes in congestion. This would enable more responsive and adaptive route optimization strategies that reflect current traffic conditions.

Expanded Data Analysis: Analysing additional factors such as weather conditions, special events, and roadworks could provide a more comprehensive view of congestion patterns and further refine route optimization.

Passenger Behaviour Analysis: Studying passenger behaviour, including the patterns for boarding and getting off the bus, could provide insights into how congestion impacts passenger experience and help in designing more efficient routes and schedules.

Introducing New Bus Stops: This study explores the possibility of optimizing current bus routes by utilizing existing bus stops and route links. Additionally, introducing new bus stops or route links could offer even more efficient route alternatives, further enhancing the effectiveness of the proposed solutions.

References

[1] Almeida, A., Brás, S., Sargento, S., and Oliveira, I. (2023) ‘Exploring bus tracking data to characterize urban traffic congestion’, Journal of Urban Mobility, 4, Art. no. 100065. Available at: https://doi.org/10.1016/j.urbmob.2023.100065.

[2] Ma, J., Chan, J., Ristanoski, G., Rajasegarar, S., and Leckie, C. (2019) ‘Bus travel time prediction with real-time traffic information’, *Transportation Research Part C: Emerging Technologies*, 105, pp. 536-549. Available at: https://doi.org/10.1016/j.trc.2019.06.008.

[3] Chien, S. I.-J., Dimitrijevic, B. V., and Spasovic, L. N. (2003) ‘Optimization of Bus Route Planning in Urban Commuter Networks’, *Journal of Public Transportation*, 6(1), pp. 53-79. Available at: https://doi.org/10.5038/2375-0901.6.1.4.

[4] Huang, K., Xu, L., Chen, Y., Cheng, Q., and An, K. (2020) ‘Customized Bus Route Optimization with Real-Time Data’, *Journal of Advanced Transportation*. Available at: https://www.researchgate.net/publication/343967029

[5] Shi, Q., Zhang, K., Weng, J., Dong, Y., Ma, S., and Zhang, M. (2021) ‘Evaluation model of bus routes optimization scheme based on multi-source bus data’, *Transportation Research Interdisciplinary Perspectives*, 10, p. 100342. Available at: https://doi.org/10.1016/j.trip.2021.100342.

[6] Afrin, T. and Yodo, N. (2020) ‘A Survey of Road Traffic Congestion Measures towards a Sustainable and Resilient Transportation System’, *Sustainability*, 12, p. 4660. Available at: https://doi.org/10.3390/su12114660.

[7] Lomax, T., Turner, S., Shunk, G., Levinson, H. S., Pratt, R. H., Bay, P. N., and Douglas, G. B. (1997) ‘Quantifying Congestion. Volume 1: Final Report’, NCHRP Report 398, Transportation Research Board, Washington, DC. Available at: http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_398.pdf.

[8] Gao, P., Liu, Z., Tian, K., and Liu, G. (2016) ‘Characterizing Traffic Conditions from the Perspective of Spatial-Temporal Heterogeneity’, *ISPRS International Journal of Geo-Information*, 5, p. 34. Available at: https://doi.org/10.3390/ijgi5030034.

[9] Chang, M.-F. and Herman, R. (1978) ‘An Attempt to Characterize Traffic in Metropolitan Areas’, *Transportation Science*, 12(1), pp. 58-79. Available at: https://doi.org/10.1287/trsc.12.1.58.

[10] Li, Q., Tan, H., Jiang, Z., Wu, Y., and Ye, L. (2021) ‘Nonrecurrent traffic congestion detection with a coupled scalable Bayesian robust tensor factorization model’, *Neurocomputing*, 430, pp. 138-149. Available at: https://doi.org/10.1016/j.neucom.2020.10.091.

[11] Gmira, M., Gendreau, M., Lodi, A., and Potvin, J.-Y. (2020) ‘Travel speed prediction based on learning methods for home delivery’, *EURO Journal on Transportation and Logistics*, 9(4), p. 100006. Available at: https://doi.org/10.1016/j.ejtl.2020.100006.

[12] Aftabuzzaman, M. (2007) ‘Measuring traffic congestion-a critical review’, in *Proceedings of the 30th Australasian Transport Research Forum*, vol. 1, ETM Group, London, UK, Sep. 2007. Available at: https://australasiantransportresearchforum.org.au/wp-content/uploads/2022/03/2007_Aftabuzzaman.pdf.

[13] Fei, X. and Gkountouna, O. (2018) ‘Spatiotemporal clustering in urban transportation: a bus route case study in Washington D.C.’, *SIGSPATIAL Special*, 10(2), pp. 26–33. Available at: https://doi.org/10.1145/3292390.3292396.

[14] Diker, A. C. and Nasibov, E. (2012) ‘Estimation of traffic congestion level via FN-DBSCAN algorithm by using GPS data’, in *2012 IV International Conference "Problems of Cybernetics and Informatics" (PCI)*, Baku, Azerbaijan, 2012, pp. 1-4. Available at: https://doi.org/10.1109/ICPCI.2012.6486279.

[15] Yang, Q., Yue, Z., Chen, R., Zhang, J., Hu, X., and Zhou, Y. (2019) ‘Real-time detection of traffic congestion based on trajectory data’, *The Journal of Engineering*, vol. 2019, pp. 8251-8256. Available at: https://doi.org/10.1049/joe.2019.0872.

[16] Tong, H. Y. and Ng, K. W. (2021) ‘A bottom-up clustering approach to identify bus driving patterns and to develop bus driving cycles for Hong Kong’, *Environmental Science and Pollution Research*, 28, pp. 14343–14357. Available at: https://doi.org/10.1007/s11356-020-11554-w.

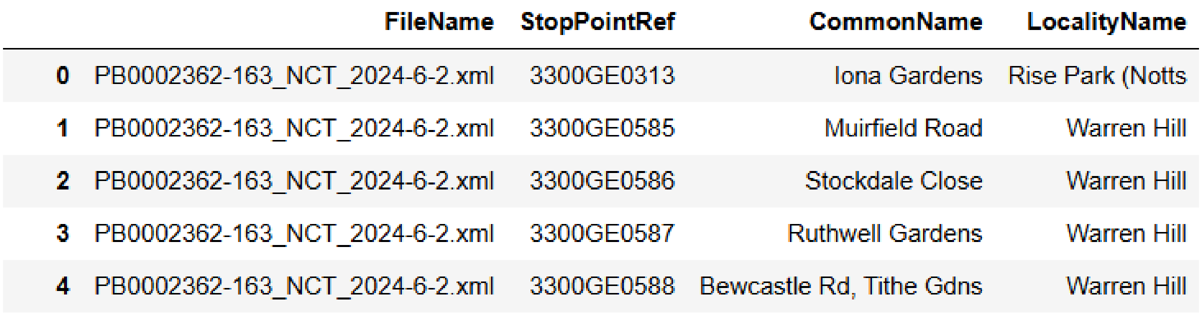
[17] Anbaroglu, B., Heydecker, B., and Cheng, T. (2014) ‘Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks’, *Transportation Research Part C: Emerging Technologies*, 48, pp. 47-65. Available at: https://doi.org/10.1016/j.trc.2014.08.002.

Appendix A: Raw Data Samples for Timetable Data

The timetable data was stored in the form of multiple pandas data frames, each containing information about a different aspect of the bus service. Here are samples from each data frame.

Stop Points

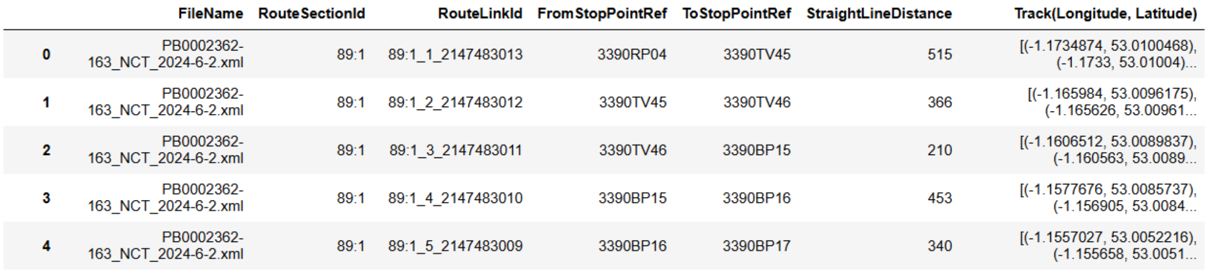
Table A.1: Sample data for Stop Points



	FileName	StopPointRef	CommonName	LocalityName
0	PB0002362-163_NCT_2024-6-2.xml	3300GE0313	Iona Gardens	Rise Park (Notts
1	PB0002362-163_NCT_2024-6-2.xml	3300GE0585	Muirfield Road	Warren Hill
2	PB0002362-163_NCT_2024-6-2.xml	3300GE0586	Stockdale Close	Warren Hill
3	PB0002362-163_NCT_2024-6-2.xml	3300GE0587	Ruthwell Gardens	Warren Hill
4	PB0002362-163_NCT_2024-6-2.xml	3300GE0588	Bewcastle Rd, Tithe Gdns	Warren Hill

Route Links

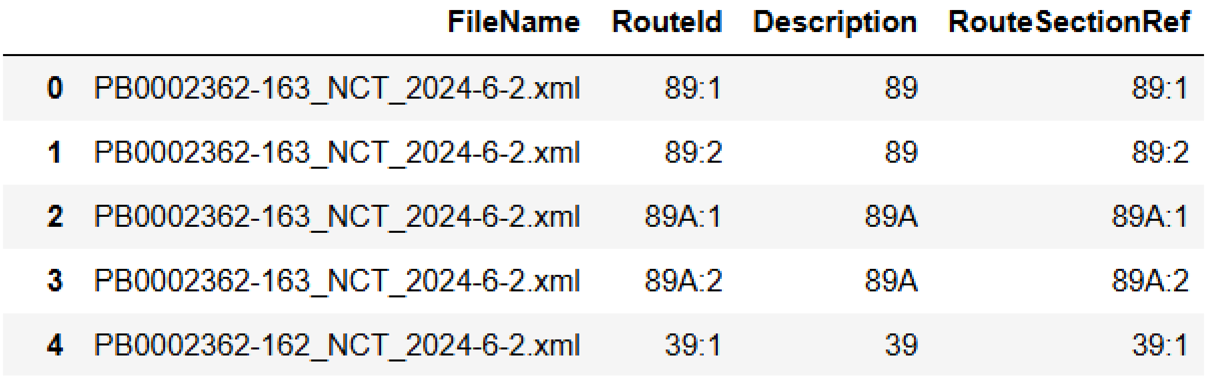
Table A.2: Sample data for Route Links



	FileName	RouteSectionId	RouteLinkId	FromStopPointRef	ToStopPointRef	StraightLineDistance	Track(Longitude, Latitude)
0	PB0002362-163_NCT_2024-6-2.xml	89:1	89:1_1_2147483013	3390RP04	3390TV45	515	[(-1.1734874, 53.0100468), (-1.1733, 53.01004)...
1	PB0002362-163_NCT_2024-6-2.xml	89:1	89:1_2_2147483012	3390TV45	3390TV46	366	[(-1.165984, 53.0096175), (-1.165626, 53.00961...
2	PB0002362-163_NCT_2024-6-2.xml	89:1	89:1_3_2147483011	3390TV46	3390BP15	210	[(-1.1606512, 53.0089837), (-1.160563, 53.0089...
3	PB0002362-163_NCT_2024-6-2.xml	89:1	89:1_4_2147483010	3390BP15	3390BP16	453	[(-1.1577676, 53.0085737), (-1.156905, 53.0084...
4	PB0002362-163_NCT_2024-6-2.xml	89:1	89:1_5_2147483009	3390BP16	3390BP17	340	[(-1.1557027, 53.0052216), (-1.155658, 53.0051...

Routes

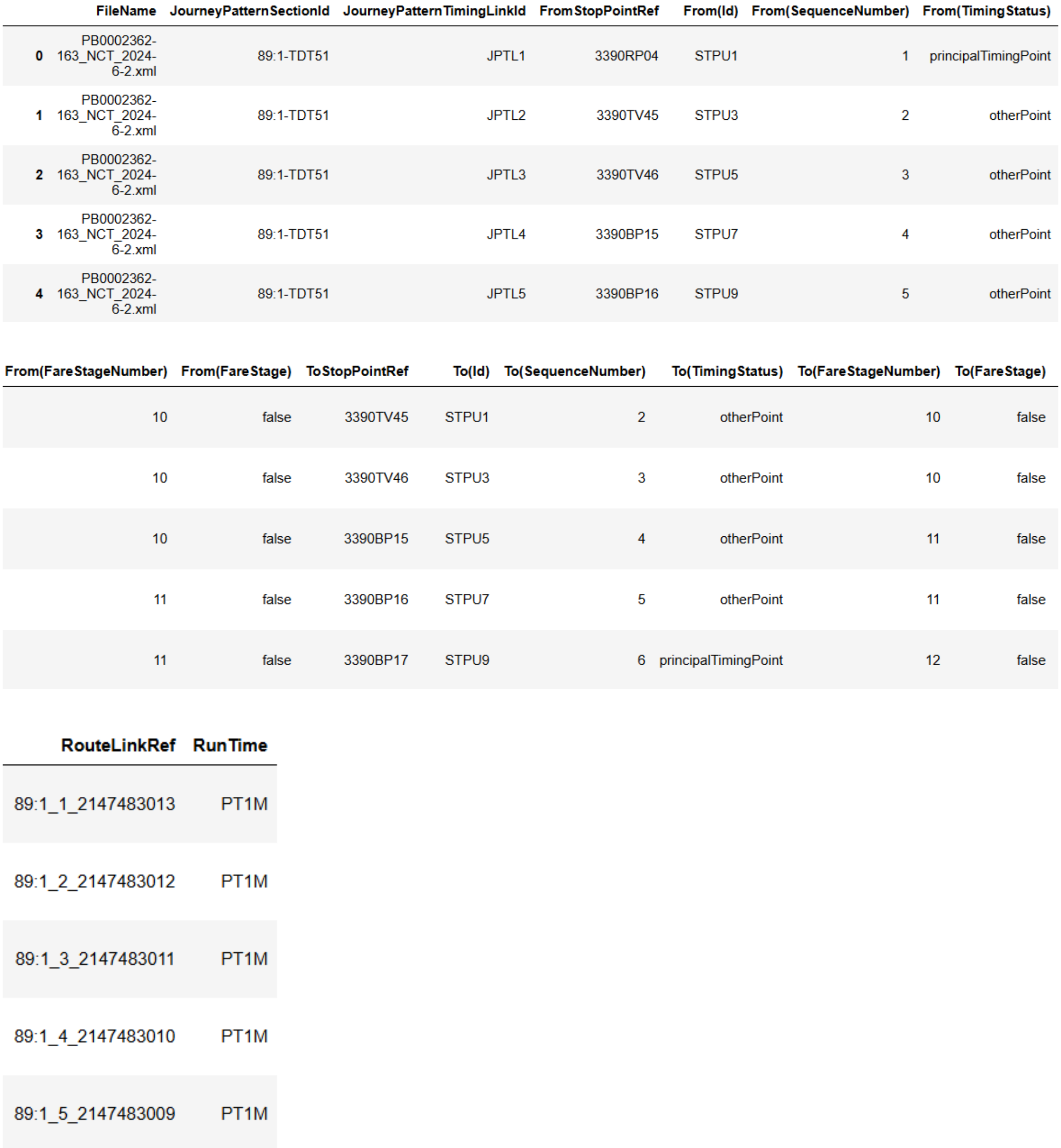
Table A.3: Sample data for Routes



	FileName	RouteId	Description	RouteSectionRef
0	PB0002362-163_NCT_2024-6-2.xml	89:1	89	89:1
1	PB0002362-163_NCT_2024-6-2.xml	89:2	89	89:2
2	PB0002362-163_NCT_2024-6-2.xml	89A:1	89A	89A:1
3	PB0002362-163_NCT_2024-6-2.xml	89A:2	89A	89A:2
4	PB0002362-162_NCT_2024-6-2.xml	39:1	39	39:1

Journey Pattern Timing Links

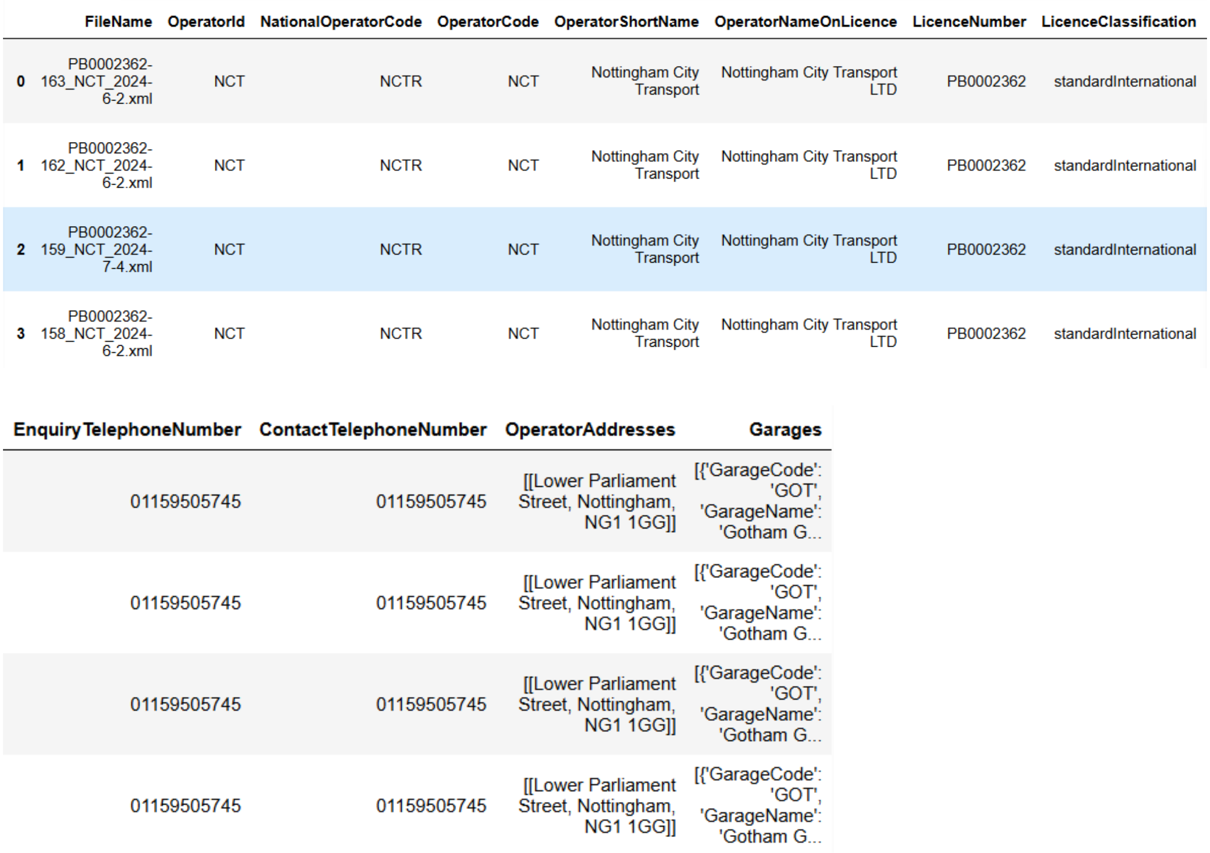
Table A.4: Sample data for Journey Pattern Timing Links



	FileName	JourneyPatternSectionId	JourneyPatternTimingLinkId	FromStopPointRef	From(Id)	From(SequenceNumber)	From(TimingStatus)	From(FareStageNumber)	From(FareStage)	ToStopPointRef	To(Id)	To(SequenceNumber)	To(TimingStatus)	To(FareStageNumber)	To(FareStage)	RouteLinkRef	RunTime
0	PB0002362-163_NCT_2024-6-2.xml	89:1-TDT51	JPTL1	3390RP04	STPU1	1	principalTimingPoint	10	false	3390TV45	STPU1	2	otherPoint	10	false	89:1_1_2147483013	PT1M
1	PB0002362-163_NCT_2024-6-2.xml	89:1-TDT51	JPTL2	3390TV45	STPU3	2	otherPoint	10	false	3390TV46	STPU3	3	otherPoint	10	false	89:1_2_2147483012	PT1M
2	PB0002362-163_NCT_2024-6-2.xml	89:1-TDT51	JPTL3	3390TV46	STPU5	3	otherPoint	10	false	3390BP15	STPU5	4	otherPoint	11	false	89:1_3_2147483011	PT1M
3	PB0002362-163_NCT_2024-6-2.xml	89:1-TDT51	JPTL4	3390BP15	STPU7	4	otherPoint	11	false	3390BP16	STPU7	5	otherPoint	11	false	89:1_4_2147483010	PT1M
4	PB0002362-163_NCT_2024-6-2.xml	89:1-TDT51	JPTL5	3390BP16	STPU9	5	otherPoint	11	false	3390BP17	STPU9	6	principalTimingPoint	12	false	89:1_5_2147483009	PT1M

Operators

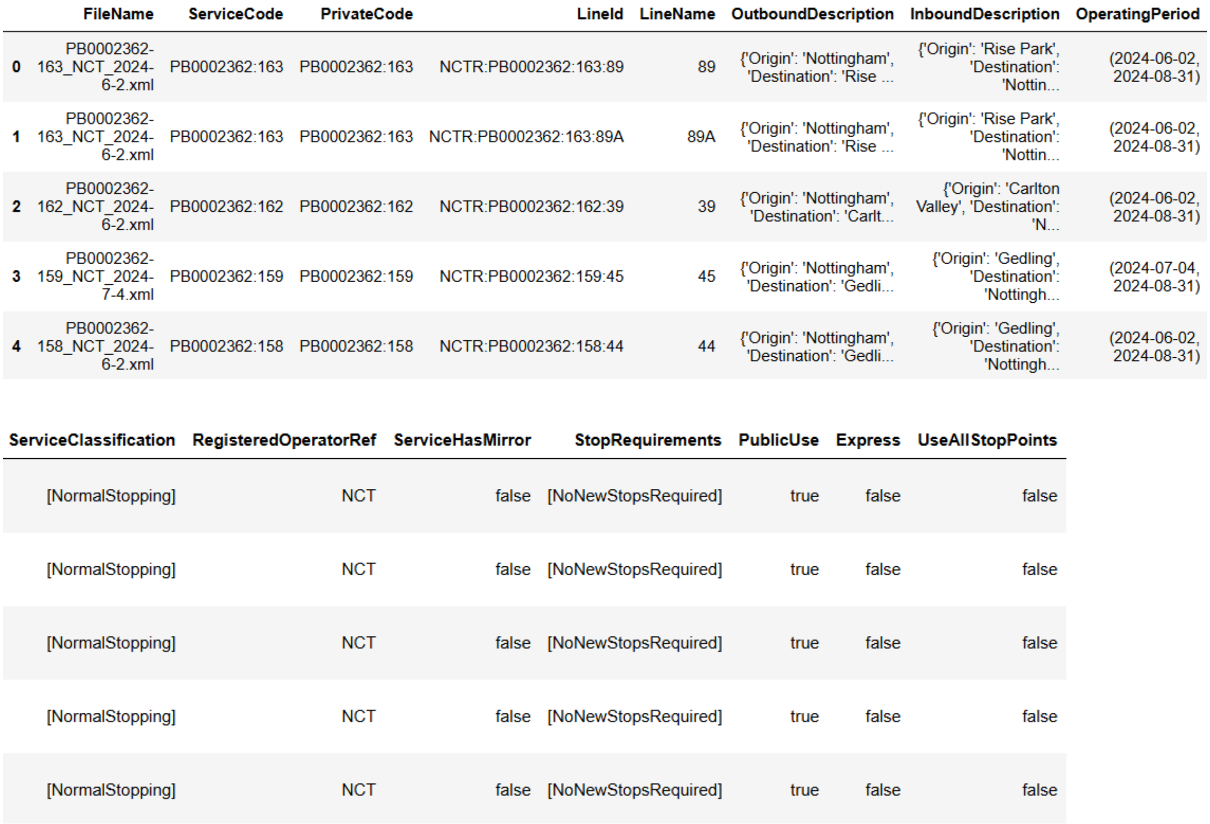
Table A.5: Sample data for Operators



	FileName	OperatorId	NationalOperatorCode	OperatorCode	OperatorShortName	OperatorNameOnLicence	LicenceNumber	LicenceClassification	EnquiryTelephoneNumber	ContactTelephoneNumber	OperatorAddresses	Garages
0	PB0002362-163_NCT_2024-6-2.xml	NCT	NCTR	NCT	Nottingham City Transport	Nottingham City Transport LTD	PB0002362	standardInternational	01159505745	01159505745	[[Lower Parliament Street, Nottingham, NG1 1GG]]	[{'GarageCode': 'GOT', 'GarageName': 'Gotham G...
1	PB0002362-162_NCT_2024-6-2.xml	NCT	NCTR	NCT	Nottingham City Transport	Nottingham City Transport LTD	PB0002362	standardInternational	01159505745	01159505745	[[Lower Parliament Street, Nottingham, NG1 1GG]]	[{'GarageCode': 'GOT', 'GarageName': 'Gotham G...
2	PB0002362-159_NCT_2024-7-4.xml	NCT	NCTR	NCT	Nottingham City Transport	Nottingham City Transport LTD	PB0002362	standardInternational	01159505745	01159505745	[[Lower Parliament Street, Nottingham, NG1 1GG]]	[{'GarageCode': 'GOT', 'GarageName': 'Gotham G...
3	PB0002362-158_NCT_2024-6-2.xml	NCT	NCTR	NCT	Nottingham City Transport	Nottingham City Transport LTD	PB0002362	standardInternational	01159505745	01159505745	[[Lower Parliament Street, Nottingham, NG1 1GG]]	[{'GarageCode': 'GOT', 'GarageName': 'Gotham G...

Service Lines

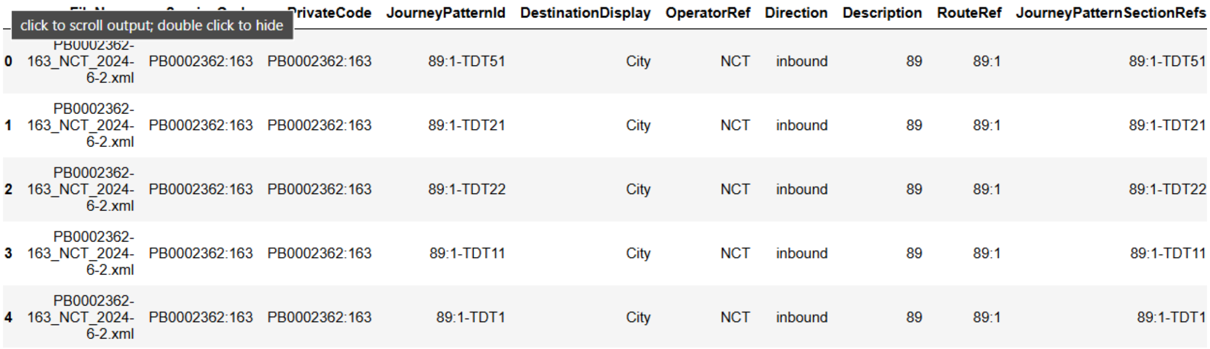
Table A.6: Sample data for Service Lines



	FileName	ServiceCode	PrivateCode	LineId	LineName	OutboundDescription	InboundDescription	OperatingPeriod	ServiceClassification	RegisteredOperatorRef	ServiceHasMirror	StopRequirements	PublicUse	Express	UseAllStopPoints
0	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	NCTR:PB0002362:163:89	89	{'Origin': 'Nottingham', 'Destination': 'Rise ...	{'Origin': 'Rise Park', 'Destination': 'Nottin...	(2024-06-02, 2024-08-31)	[NormalStopping]	NCT	false	[NoNewStopsRequired]	true	false	false
1	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	NCTR:PB0002362:163:89A	89A	{'Origin': 'Nottingham', 'Destination': 'Rise ...	{'Origin': 'Rise Park', 'Destination': 'Nottin...	(2024-06-02, 2024-08-31)	[NormalStopping]	NCT	false	[NoNewStopsRequired]	true	false	false
2	PB0002362-162_NCT_2024-6-2.xml	PB0002362:162	PB0002362:162	NCTR:PB0002362:162:39	39	{'Origin': 'Nottingham', 'Destination': 'Carlt...	{'Origin': 'Carlton Valley', 'Destination': 'N...	(2024-06-02, 2024-08-31)	[NormalStopping]	NCT	false	[NoNewStopsRequired]	true	false	false
3	PB0002362-159_NCT_2024-7-4.xml	PB0002362:159	PB0002362:159	NCTR:PB0002362:159:45	45	{'Origin': 'Nottingham', 'Destination': 'Gedli...	{'Origin': 'Gedling', 'Destination': 'Nottingh...	(2024-07-04, 2024-08-31)	[NormalStopping]	NCT	false	[NoNewStopsRequired]	true	false	false
4	PB0002362-158_NCT_2024-6-2.xml	PB0002362:158	PB0002362:158	NCTR:PB0002362:158:44	44	{'Origin': 'Nottingham', 'Destination': 'Gedli...	{'Origin': 'Gedling', 'Destination': 'Nottingh...	(2024-06-02, 2024-08-31)	[NormalStopping]	NCT	false	[NoNewStopsRequired]	true	false	false

Service Journey Patterns

Table A.7: Sample data for Service Journey Patterns

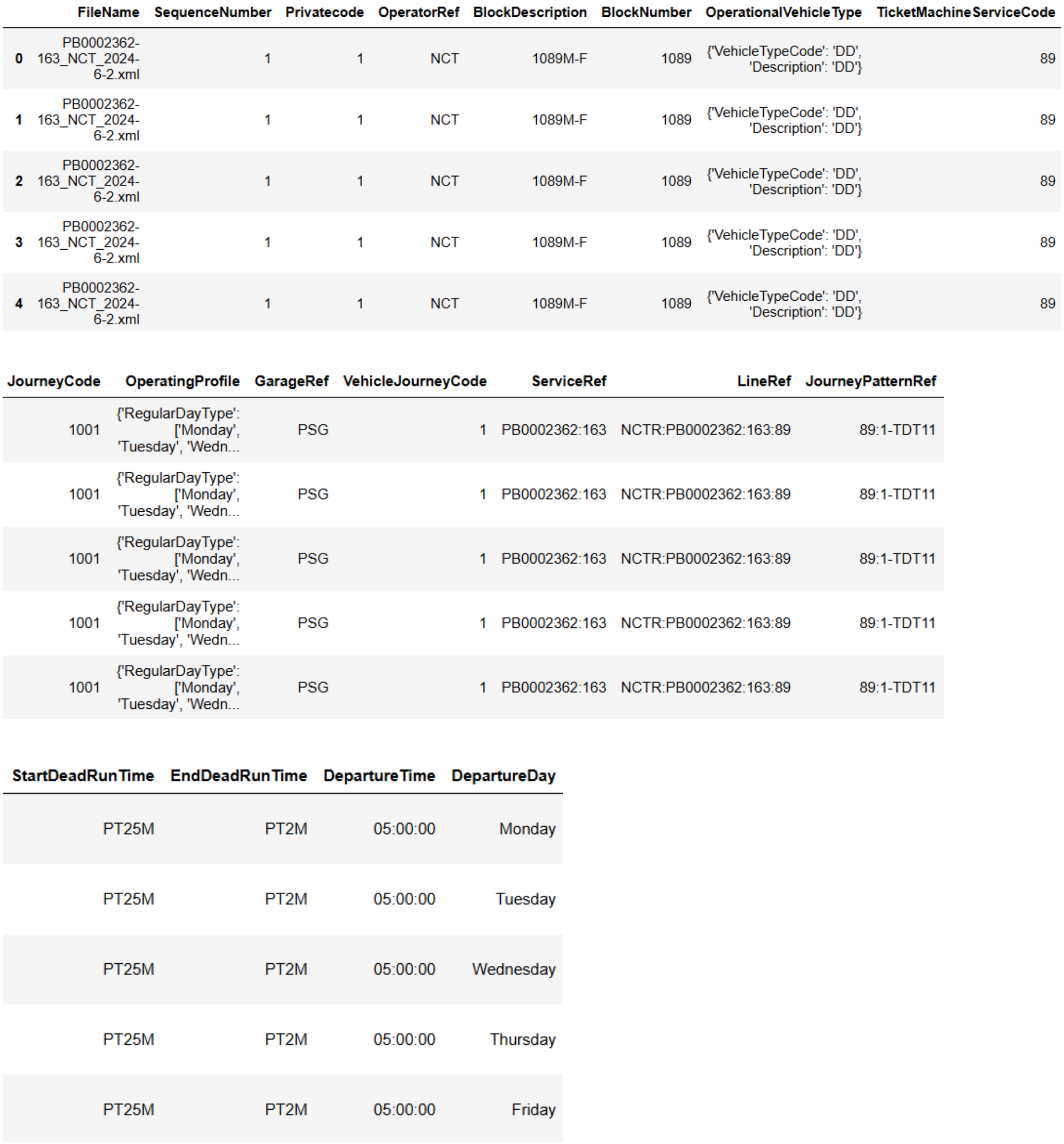


click to scroll output; double click to hide

	FileName	ServiceCode	PrivateCode	JourneyPatternId	DestinationDisplay	OperatorRef	Direction	Description	RouteRef	JourneyPatternSectionRefs
0	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	89:1-TDT51	City	NCT	inbound	89	89:1	89:1-TDT51
1	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	89:1-TDT21	City	NCT	inbound	89	89:1	89:1-TDT21
2	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	89:1-TDT22	City	NCT	inbound	89	89:1	89:1-TDT22
3	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	89:1-TDT11	City	NCT	inbound	89	89:1	89:1-TDT11
4	PB0002362-163_NCT_2024-6-2.xml	PB0002362:163	PB0002362:163	89:1-TDT1	City	NCT	inbound	89	89:1	89:1-TDT1

Vehicle Journeys

Table A.8: Sample data for Vehicle Journeys



	FileName	SequenceNumber	Privatecode	OperatorRef	BlockDescription	BlockNumber	OperationalVehicleType	TicketMachineServiceCode
0	PB0002362-163_NCT_2024-6-2.xml	1	1	NCT	1089M-F	1089	{'VehicleTypeCode': 'DD', 'Description': 'DD'}	89
1	PB0002362-163_NCT_2024-6-2.xml	1	1	NCT	1089M-F	1089	{'VehicleTypeCode': 'DD', 'Description': 'DD'}	89
2	PB0002362-163_NCT_2024-6-2.xml	1	1	NCT	1089M-F	1089	{'VehicleTypeCode': 'DD', 'Description': 'DD'}	89
3	PB0002362-163_NCT_2024-6-2.xml	1	1	NCT	1089M-F	1089	{'VehicleTypeCode': 'DD', 'Description': 'DD'}	89
4	PB0002362-163_NCT_2024-6-2.xml	1	1	NCT	1089M-F	1089	{'VehicleTypeCode': 'DD', 'Description': 'DD'}	89

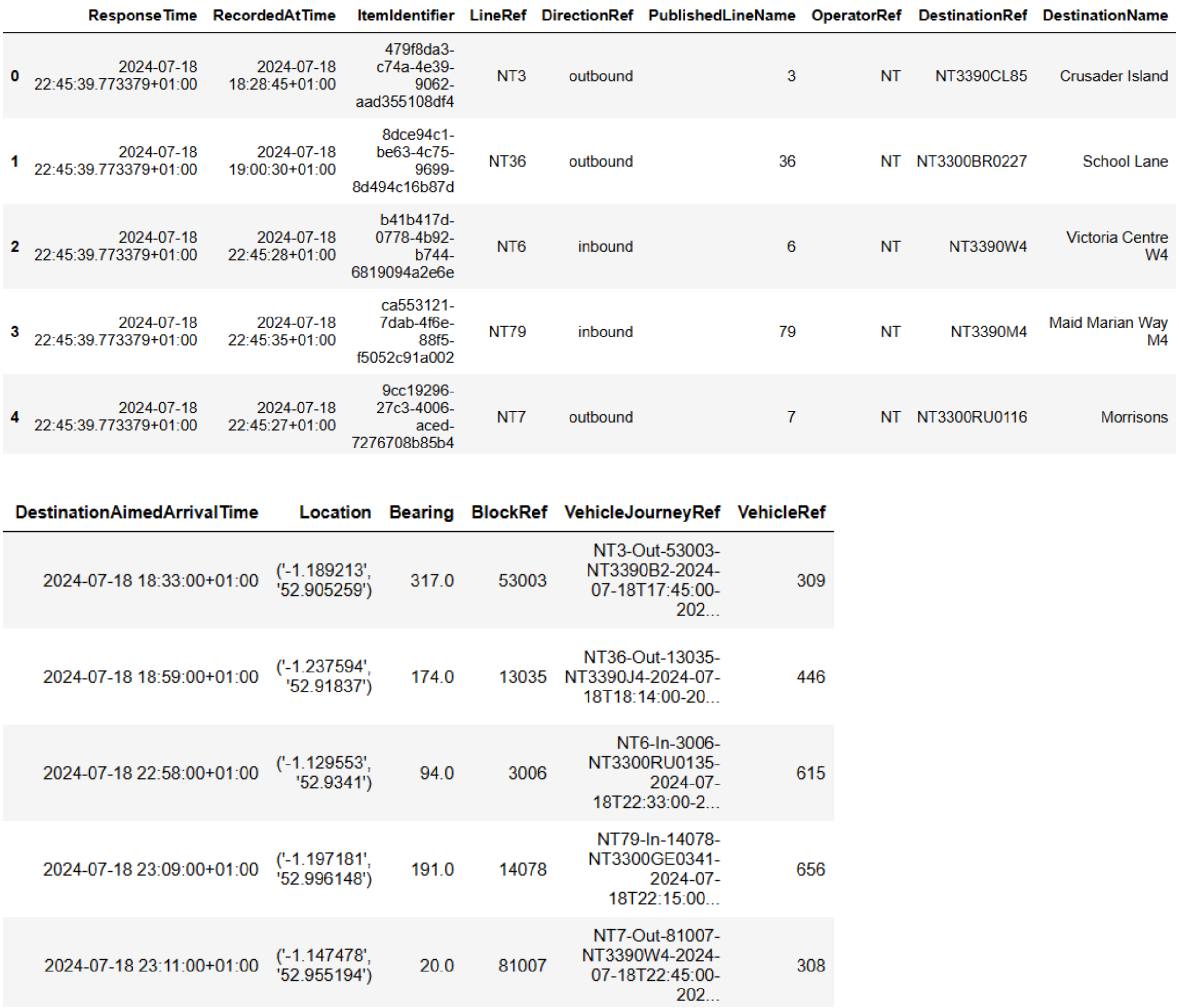
JourneyCode	OperatingProfile	GarageRef	VehicleJourneyCode	ServiceRef	LineRef	JourneyPatternRef
1001	{'RegularDayType': ['Monday', 'Tuesday', 'Wedn...	PSG	1	PB0002362:163	NCTR:PB0002362:163:89	89:1-TDT11
1001	{'RegularDayType': ['Monday', 'Tuesday', 'Wedn...	PSG	1	PB0002362:163	NCTR:PB0002362:163:89	89:1-TDT11
1001	{'RegularDayType': ['Monday', 'Tuesday', 'Wedn...	PSG	1	PB0002362:163	NCTR:PB0002362:163:89	89:1-TDT11
1001	{'RegularDayType': ['Monday', 'Tuesday', 'Wedn...	PSG	1	PB0002362:163	NCTR:PB0002362:163:89	89:1-TDT11
1001	{'RegularDayType': ['Monday', 'Tuesday', 'Wedn...	PSG	1	PB0002362:163	NCTR:PB0002362:163:89	89:1-TDT11

StartDeadRunTime	EndDeadRunTime	DepartureTime	DepartureDay
PT25M	PT2M	05:00:00	Monday
PT25M	PT2M	05:00:00	Tuesday
PT25M	PT2M	05:00:00	Wednesday
PT25M	PT2M	05:00:00	Thursday
PT25M	PT2M	05:00:00	Friday

Appendix B: Raw Data Samples for Location Data

The location data was recorded over a period of 2 weeks. It stores the location of each bus after every 15 seconds.

Table B.1: Sample Location Data



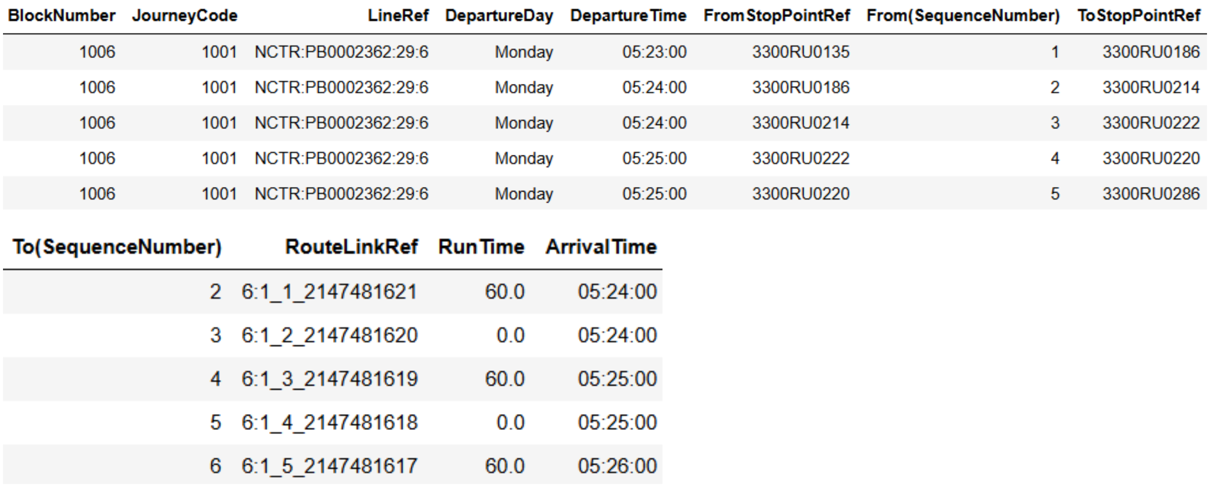
	ResponseTime	RecordedAtTime	ItemIdentifier	LineRef	DirectionRef	PublishedLineName	OperatorRef	DestinationRef	DestinationName
0	2024-07-18 22:45:39.773379+01:00	2024-07-18 18:28:45+01:00	479f8da3-c74a-4e39-9062-aad355108df4	NT3	outbound	3	NT	NT3390CL85	Crusader Island
1	2024-07-18 22:45:39.773379+01:00	2024-07-18 19:00:30+01:00	8dce94c1-be63-4c75-9699-8d494c16b87d	NT36	outbound	36	NT	NT3300BR0227	School Lane
2	2024-07-18 22:45:39.773379+01:00	2024-07-18 22:45:28+01:00	b41b417d-0778-4b92-b744-6819094a2e6e	NT6	inbound	6	NT	NT3390W4	Victoria Centre W4
3	2024-07-18 22:45:39.773379+01:00	2024-07-18 22:45:35+01:00	ca553121-7dab-4f6e-88f5-f5052c91a002	NT79	inbound	79	NT	NT3390M4	Maid Marian Way M4
4	2024-07-18 22:45:39.773379+01:00	2024-07-18 22:45:27+01:00	9cc19296-27c3-4006-aced-7276708b85b4	NT7	outbound	7	NT	NT3300RU0116	Morrisons

DestinationAimedArrivalTime	Location	Bearing	BlockRef	VehicleJourneyRef	VehicleRef
2024-07-18 18:33:00+01:00	('-1.189213', '52.905259')	317.0	53003	NT3-Out-53003-NT3390B2-2024-07-18T17:45:00-202...	309
2024-07-18 18:59:00+01:00	('-1.237594', '52.91837')	174.0	13035	NT36-Out-13035-NT3390J4-2024-07-18T18:14:00-20...	446
2024-07-18 22:58:00+01:00	('-1.129553', '52.9341')	94.0	3006	NT6-In-3006-NT3300RU0135-2024-07-18T22:33:00-2...	615
2024-07-18 23:09:00+01:00	('-1.197181', '52.996148')	191.0	14078	NT79-In-14078-NT3300GE0341-2024-07-18T22:15:00...	656
2024-07-18 23:11:00+01:00	('-1.147478', '52.955194')	20.0	81007	NT7-Out-81007-NT3390W4-2024-07-18T22:45:00-202...	308

Appendix C: Sample of Pre-processed Timetable Data

The timetable data was pre-processed and all the data frames were integrated to get a schedule for each bus journey.

Table C.1: Pre-processed Timetable data sample

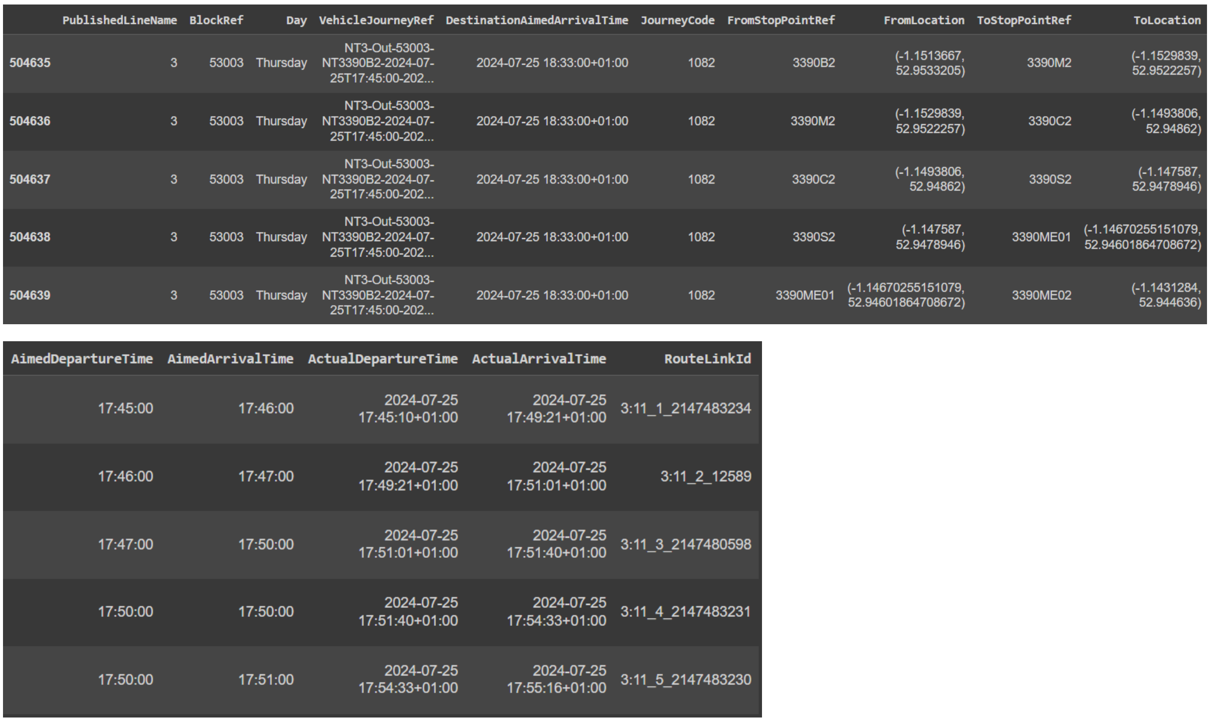


BlockNumber	JourneyCode	LineRef	DepartureDay	DepartureTime	FromStopPointRef	From(SequenceNumber)	ToStopPointRef	To(SequenceNumber)	RouteLinkRef	RunTime	ArrivalTime
1006	1001	NCTR:PB0002362:29:6	Monday	05:23:00	3300RU0135	1	3300RU0186	2	6:1_1_2147481621	60.0	05:24:00
1006	1001	NCTR:PB0002362:29:6	Monday	05:24:00	3300RU0186	2	3300RU0214	3	6:1_2_2147481620	0.0	05:24:00
1006	1001	NCTR:PB0002362:29:6	Monday	05:24:00	3300RU0214	3	3300RU0222	4	6:1_3_2147481619	60.0	05:25:00
1006	1001	NCTR:PB0002362:29:6	Monday	05:25:00	3300RU0222	4	3300RU0220	5	6:1_4_2147481618	0.0	05:25:00
1006	1001	NCTR:PB0002362:29:6	Monday	05:25:00	3300RU0220	5	3300RU0286	6	6:1_5_2147481617	60.0	05:26:00

Appendix D: Integrating Location data with Timetable data

The location data was mapped to the timetable data to get the actual journey of each bus and the aimed departure and arrival times between stops.

Table D.1: Integrated Location and Timetable data

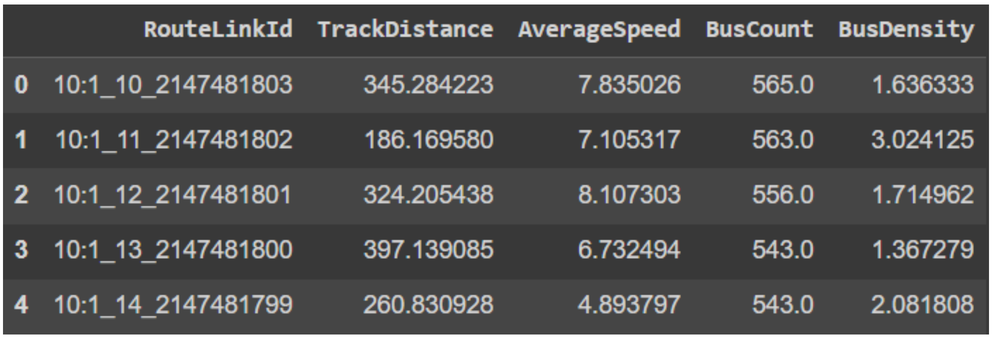


	PublishedLineName	BlockRef	Day	VehicleJourneyRef	DestinationAimedArrivalTime	JourneyCode	FromStopPointRef	FromLocation	ToStopPointRef	ToLocation
504635	3	53003	Thursday	NT3-Out-53003-NT3390B2-2024-07-25T17:45:00-202...	2024-07-25 18:33:00+01:00	1082	3390B2	(-1.1513667, 52.9533205)	3390M2	(-1.1529839, 52.9522257)
504636	3	53003	Thursday	NT3-Out-53003-NT3390B2-2024-07-25T17:45:00-202...	2024-07-25 18:33:00+01:00	1082	3390M2	(-1.1529839, 52.9522257)	3390C2	(-1.1493806, 52.94862)
504637	3	53003	Thursday	NT3-Out-53003-NT3390B2-2024-07-25T17:45:00-202...	2024-07-25 18:33:00+01:00	1082	3390C2	(-1.1493806, 52.94862)	3390S2	(-1.147587, 52.9478946)
504638	3	53003	Thursday	NT3-Out-53003-NT3390B2-2024-07-25T17:45:00-202...	2024-07-25 18:33:00+01:00	1082	3390S2	(-1.147587, 52.9478946)	3390ME01	(-1.14670255151079, 52.94601864708672)
504639	3	53003	Thursday	NT3-Out-53003-NT3390B2-2024-07-25T17:45:00-202...	2024-07-25 18:33:00+01:00	1082	3390ME01	(-1.14670255151079, 52.94601864708672)	3390ME02	(-1.1431284, 52.944636)

AimedDepartureTime	AimedArrivalTime	ActualDepartureTime	ActualArrivalTime	RouteLinkId
17:45:00	17:46:00	2024-07-25 17:45:10+01:00	2024-07-25 17:49:21+01:00	3:11_1_2147483234
17:46:00	17:47:00	2024-07-25 17:49:21+01:00	2024-07-25 17:51:01+01:00	3:11_2_12589
17:47:00	17:50:00	2024-07-25 17:51:01+01:00	2024-07-25 17:51:40+01:00	3:11_3_2147480598
17:50:00	17:50:00	2024-07-25 17:51:40+01:00	2024-07-25 17:54:33+01:00	3:11_4_2147483231
17:50:00	17:51:00	2024-07-25 17:54:33+01:00	2024-07-25 17:55:16+01:00	3:11_5_2147483230

Appendix E: Average Speed and Bus Density

Table E.1: Sample of data after calculating speed and bus density



	RouteLinkId	TrackDistance	AverageSpeed	BusCount	BusDensity
0	10:1_10_2147481803	345.284223	7.835026	565.0	1.636333
1	10:1_11_2147481802	186.169580	7.105317	563.0	3.024125
2	10:1_12_2147481801	324.205438	8.107303	556.0	1.714962
3	10:1_13_2147481800	397.139085	6.732494	543.0	1.367279
4	10:1_14_2147481799	260.830928	4.893797	543.0	2.081808

Appendix F: Congestion Summary

The following congestion summary shows the percentage congestion of each route section. Each route section consists of multiple route links. The ratio of the congested links and total links was calculated to get the congestion percentage.

Table F.2: Congestion Summary of the Route Sections

Route Section Id	**Total Links**	**Congested Links**	**Percentage Congested**
39:1	32	32	100.000000
43:1	22	22	100.000000
39:2	15	15	100.000000
58:1	36	36	100.000000
17:1	28	28	100.000000
41:6	10	10	100.000000
17:2	26	25	96.153846
77:2	26	25	96.153846
58:2	35	33	94.285714
28:2	29	27	93.103448
28:1	28	26	92.857143
27:1	28	26	92.857143
77:5	27	25	92.592593
43:3	13	12	92.307692
45:2	26	24	92.307692
44:1	32	29	90.625000
41:5	10	9	90.000000
89:2	29	26	89.655172
45:1	26	23	88.461538
35:7	56	48	85.714286
34C:3	21	18	85.714286
27:2	27	23	85.185185
6:3	25	21	84.000000
89:1	29	24	82.758621
44:2	30	24	80.000000
35:1	58	44	75.862069
36:3	45	34	75.555556

34C:6	22	16	72.727273
36:1	34	24	70.588235
69:1	40	26	65.000000
25:4	48	31	64.583333
6:1	22	14	63.636364
25:1	48	29	60.416667
68:2	40	20	50.000000
69:4	38	18	47.368421
78:2	32	15	46.875000
10:3	32	15	46.875000
68:1	38	17	44.736842
10:1	30	12	40.000000
78:3	32	12	37.500000
48:1	34	12	35.294118
30:1	29	10	34.482759
87:2	32	11	34.375000
15:1	35	12	34.285714
48:18	31	10	32.258065
30:3	28	9	32.142857
11:4	19	6	31.578947
88:1	41	12	29.268293
11:1	15	4	26.666667
7:3	34	8	23.529412