A Project Report on

MULTIMODAL TRANSPORTATION ANALYSIS

Submitted in partial fulfillment of the

requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by:

Saksham Garg 19116065	**Arush Sharma** 19116011

Under the guidance of:

Prof. Dheeraj Kumar

Department of Electronics and Communication Engineering

Indian Institute of Technology Roorkee

May 2022

DECLARATION

We hereby declare that the work which is being presented in this report entitled **“MULTIMODAL TRANSPORTATION ANALYSIS”** in partial fulfillment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in **Electronics and Communication Engineering** to the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee is an authentic record of our own work carried out under the supervision of Dr. Supervisor and Dr. Co-supervisor, Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee. This work has been done during July 2022 to May 2023. We have not submitted the matter embodied in this report for the award of any other degree or diploma.

Date: 16th May, 2023
Place: Roorkee

Arush Sharma (19116011)	**Saksham Garg** (19116065)

CERTIFICATE

This is to certify that the project report entitled **“MULTIMODAL TRANSPORTATION ANALYSIS”** submitted by ***Arush Sharma*** and ***Saksham Garg*** to the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee toward partial fulfillment of the requirements for the degree of **BACHELOR OF TECHNOLOGY** in **Electronics and Communication Engineering** is a record of bonafide work carried out under my supervision and guidance.

Date: 16th May, 2023
Place: Roorkee

Dheeraj Kumar

Dr. Supervisor

Assistant/Associate/Professor
Department of Electronics and Communication Engineering
Indian Institute of Technology Roorkee

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Dheeraj Kumar for his invaluable guidance and support throughout the duration of our B.Tech Project. His expertise, mentorship, and encouragement have been instrumental in shaping our research and enabling us to accomplish our project goals. We extend our heartfelt appreciation for entrusting us with this project and providing us with the opportunity to explore the fascinating domain of Data Mining and Data Analysis. The knowledge and skills we have acquired during this journey have greatly enriched our understanding of the subject matter. We would also like to acknowledge the contributions of our fellow classmates and friends who have provided assistance and insightful discussions, contributing to the overall success of our project. Their support and collaboration have been invaluable in overcoming challenges and refining our ideas. Lastly, we would like to express our gratitude to our families for their unwavering support and encouragement throughout this endeavor. Their belief in our abilities has been a constant source of motivation. This project has been an incredible learning experience, and we are grateful to all those who have played a part in making it a success.

ABSTRACT

This project presents an analysis of Delhi's multimodal public transportation system, focusing on the Metro and Bus networks. Utilizing passenger e-ticket data and static bus information, we examine station and route loads to uncover insights into peak periods and heavily utilized segments. Additionally, through a comprehensive literature review, we enhance our understanding of the system's operational characteristics and benchmark our findings against established knowledge. The value of this analysis lies in its ability to inform resource allocation, optimize route planning, and facilitate evidence-based decision-making for both transportation authorities and commuters. However, to further enhance the efficiency and effectiveness of Delhi's public transportation, we emphasize the need for increased availability of public transportation data to support data-driven planning, foster innovation, and promote transparency. This project contributes to the ongoing efforts of improving the city's transportation infrastructure and services.

CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

1.1. Introduction To The Problem Statement

The objective of this project is to perform Multimodal Transportation Analysis for the Metro and Bus systems in Delhi. Building upon our previous semester's work, our focus has expanded to include the analysis of both Metro and Bus data to gain insights into passenger mobility patterns and enhance the efficiency of public transportation.

1.1.1 Problem Description

The objective of this research is to analyze the passenger mobility patterns in Delhi, specifically focusing on the Delhi Metro and bus services. The analysis involves examining the station and route loads at various times of the day to gain insights into the utilization and distribution of public transportation.For the Delhi Metro, the study aims to analyze the station loads and understand the demand patterns throughout the day. By evaluating the passenger volumes at different metro stations, we can identify areas of high traffic and assess the efficiency of the metro system.Additionally, the research includes an evaluation of the station and route loads for bus services. By analyzing the passenger volumes at bus stops and studying the efficiency of bus routes, we aim to improve the bus transportation network and enhance the overall commuting experience.Through this comprehensive analysis of passenger mobility in the Delhi Metro and bus services, this research aims to provide valuable insights for optimizing public transportation systems and facilitating better transportation planning in the city.

1.2 Motivation

The analysis of passenger mobility in the Delhi Metro and bus services holds significant importance for various stakeholders, including the government, transportation authorities, and the general public. This section discusses the motivations behind this research and the benefits it offers.

1.2.1 Benefits for Government

Accurate and insightful analysis of passenger mobility data provides valuable information to government agencies and transportation authorities. By understanding the utilization patterns of the Delhi Metro and bus services, policymakers can make informed decisions regarding infrastructure development, route planning, and resource allocation. This analysis aids in optimizing the efficiency of public transportation, reducing congestion, and enhancing the overall commuting experience for citizens.

Furthermore, the data-driven insights obtained from this research can guide the government in making informed decisions about future expansion plans, such as the construction of new metro lines, bus routes, and associated infrastructure investments. By strategically leveraging this information, the government can ensure sustainable and efficient transportation systems that cater to the growing demands of the city's population.

1.2.2 Benefits for the Public

The analysis of passenger mobility patterns in the Delhi Metro and bus services directly benefits the general public. By making this data accessible and understandable, commuters can make informed decisions regarding their travel plans, choose optimal routes, and avoid peak-hour congestion. This empowers individuals to plan their journeys more efficiently, reduce travel time, and enhance their overall commuting experience.

Moreover, by analyzing station and route loads, this research provides insights into areas of high passenger traffic, allowing commuters to avoid crowded stations and buses, and potentially find alternatives that offer a more comfortable and convenient travel experience. By empowering the public with this knowledge, the research contributes to the enhancement of commuter satisfaction and promotes the usage of public transportation as a viable and reliable mode of travel.

1.2.3 Importance of Making Data Public

Transparency and accessibility of data are fundamental to the success of any data-driven analysis. By advocating for the publication of relevant transportation data, including metro and bus load data, this research emphasizes the significance of open data initiatives. Public availability of this data encourages collaborative efforts, fosters innovation, and enables independent researchers and developers to contribute to the improvement of transportation systems. Moreover, it facilitates accountability and public scrutiny, ensuring that decisions related to transportation planning and resource allocation are made in the best interest of the citizens.

By highlighting the importance of making data public, this research aims to encourage the government and relevant authorities to adopt policies and practices that promote transparency, openness, and data-driven decision-making in the realm of public transportation.

1.3 Related Literature

Multimodal transportation analysis is a vast field that encompasses extensive research and data mining efforts. Several studies have contributed to the understanding of various aspects related to this domain. Hussain, Bhaskar, and Chung (2021) [4] analyze and estimate a transit origin-destination (OD) matrix using smartcard data, considering recent developments and future research challenges. Their work focuses on utilizing smartcard data to understand passenger mobility patterns and improve transportation planning. Munizaga and Palma (2012) [5] present an estimation approach for a disaggregate multimodal public transport OD matrix using passive smartcard data from Santiago, Chile. Their study showcases the application of smartcard data in understanding passenger travel patterns and optimizing public transportation systems. Huang et al. (2020) [6] propose a method for estimating bus

OD matrices using multisource data. Their approach integrates data from various sources to infer passenger travel patterns and optimize bus service planning. Gao et al. (2015) [7] discuss a calculation method for OD matrices in a multi-modal transit network based on traffic big data. Their study highlights the significance of utilizing comprehensive traffic data to analyze passenger flows and enhance transportation network efficiency. Afandizadeh Zargari et al. (2021) [8] explore the estimation of hourly origin-destination matrices using intelligent transportation systems data and deep learning techniques. Their research emphasizes the application of advanced technologies to analyze travel patterns and improve transportation system performance.

These studies collectively contribute to the body of knowledge in multimodal transportation analysis by employing various data sources, analytical techniques, and optimization approaches. By leveraging the insights gained from these research endeavors, we aim to enhance our understanding of passenger mobility in the context of the Delhi Metro and bus transportation system.

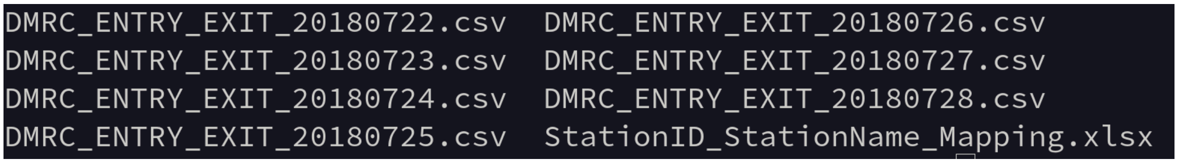
Chapter 2: Data: Sourcing and Analysis

2.1 Metro Data

In our project, we began by acquiring the data on Delhi Metro Passenger Mobility. This dataset was provided to us by our project supervisor and was originally released by the Delhi Metro Rail Corporation (DMRC) during a hackathon they conducted in 2018. The dataset serves as a valuable resource for studying the passenger mobility patterns within the Delhi Metro network.

2.1.1 About the Dataset

We started with the following files for Delhi Metro :

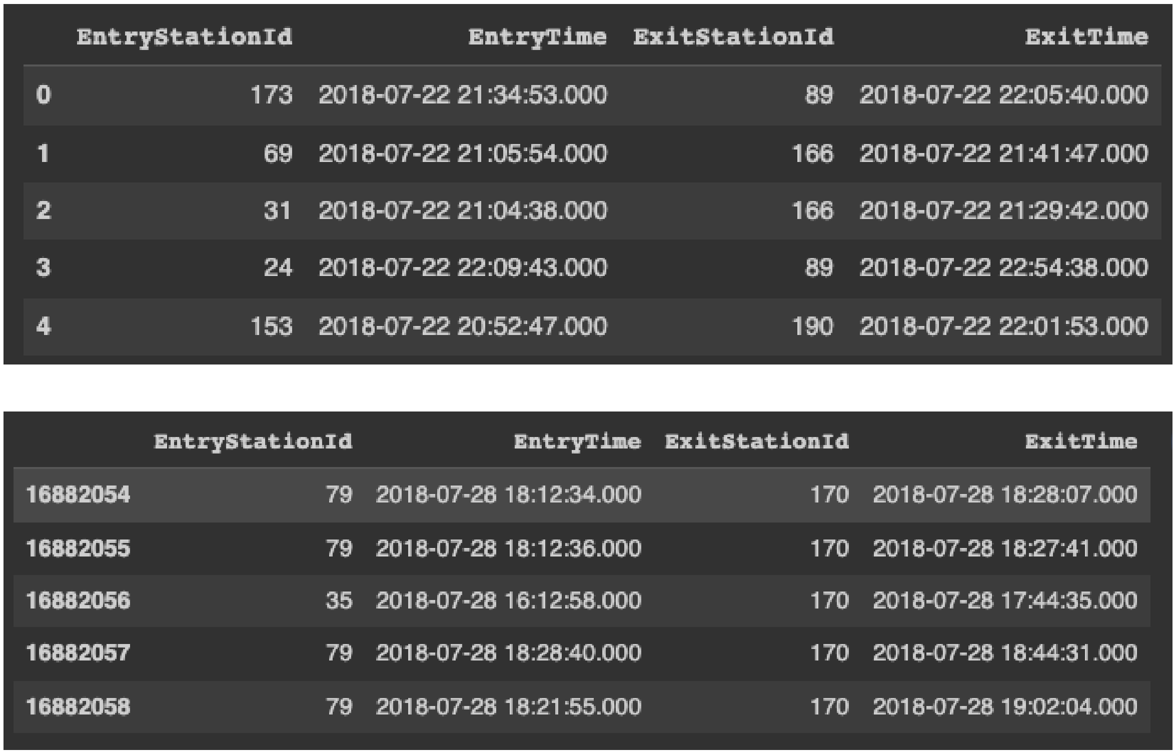


```
DMRC_ENTRY_EXIT_20180722.csv  DMRC_ENTRY_EXIT_20180726.csv
DMRC_ENTRY_EXIT_20180723.csv  DMRC_ENTRY_EXIT_20180727.csv
DMRC_ENTRY_EXIT_20180724.csv  DMRC_ENTRY_EXIT_20180728.csv
DMRC_ENTRY_EXIT_20180725.csv  StationID_StationName_Mapping.xlsx
```

Fig 1: List of given files in Dataset

The core component of the metro data we worked with is the Passenger Entry-Exit Data. This dataset provides crucial information about the entry and exit of passengers at each metro station. It includes attributes such as the entry station ID, entry time, exit station ID, and exit time. This data allows us to analyze the flow of passengers throughout the Delhi Metro network and understand the patterns of passenger mobility.

The dataset contains 7 CSV files and one xlsx file. The CSV files contains data in the following format:-

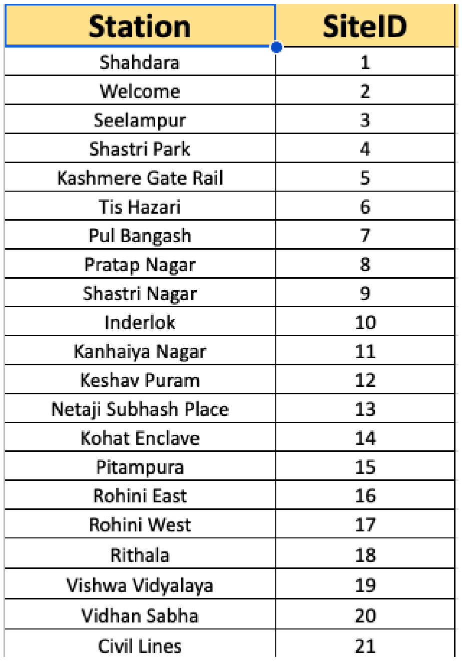


	EntryStationId	EntryTime	ExitStationId	ExitTime
0	173	2018-07-22 21:34:53.000	89	2018-07-22 22:05:40.000
1	69	2018-07-22 21:05:54.000	166	2018-07-22 21:41:47.000
2	31	2018-07-22 21:04:38.000	166	2018-07-22 21:29:42.000
3	24	2018-07-22 22:09:43.000	89	2018-07-22 22:54:38.000
4	153	2018-07-22 20:52:47.000	190	2018-07-22 22:01:53.000

	EntryStationId	EntryTime	ExitStationId	ExitTime
16882054	79	2018-07-28 18:12:34.000	170	2018-07-28 18:28:07.000
16882055	79	2018-07-28 18:12:36.000	170	2018-07-28 18:27:41.000
16882056	35	2018-07-28 16:12:58.000	170	2018-07-28 17:44:35.000
16882057	79	2018-07-28 18:28:40.000	170	2018-07-28 18:44:31.000
16882058	79	2018-07-28 18:21:55.000	170	2018-07-28 19:02:04.000

Fig 2: Passenger Data in a Pandas Dataframe

Along with this, we also have a StationID_StationName_Mapping file that actually contains the mapping of the Entry and Exit Station IDs as can be seen in the CSV files. The ID numbers are mapped to the Station names.



Station	SiteID
Shahdara	1
Welcome	2
Seelampur	3
Shastri Park	4
Kashmere Gate Rail	5
Tis Hazari	6
Pul Bangash	7
Pratap Nagar	8
Shastri Nagar	9
Inderlok	10
Kanhaiya Nagar	11
Keshav Puram	12
Netaji Subhash Place	13
Kohat Enclave	14
Pitampura	15
Rohini East	16
Rohini West	17
Rithala	18
Vishwa Vidyalaya	19
Vidhan Sabha	20
Civil Lines	21

Fig 3: Station ID to Station Name Mapping

2.1.1 Formatting

Before delving into the analysis of the Metro Data, it was essential to address the issue of data formatting. The dataset provided to us initially was highly unformatted, with irregular whitespaces and inconsistent naming conventions. To ensure the dataset's usability and compatibility with our analysis code, we embarked on a process of standardizing the data. One of the primary challenges we encountered was the presence of irregular whitespaces and formatting inconsistencies within the dataset. This hindered our ability to perform seamless data processing and analysis. To overcome this hurdle, we undertook the task of standardizing the dataset by eliminating unnecessary whitespaces, ensuring uniform spacing, and aligning the data structure. Moreover, we discovered inconsistencies in the naming of metro stations within the dataset. These discrepancies posed a significant obstacle when integrating external data sources or conducting comparative analyses. To address this issue, we implemented a methodology that utilized fuzzy search techniques. This approach allowed us to resolve inconsistencies and establish accurate correspondences between the station names in our dataset and those obtained from other online sources.

2.1.2 Shortcomings

Despite the valuable insights the Metro Data provided, it had certain limitations and shortcomings that needed to be addressed. These shortcomings are as follows:

Limited Timeframe: The dataset covered a specific timeframe that is, it only has data on stations that were constructed till 2018, and thus, the analysis was limited to that particular period.

Absence of Real-Time Data: The dataset consisted of historical data and did not include real-time updates. This limited our ability to capture dynamic changes and fluctuations in passenger mobility patterns.

Incomplete Information: The dataset primarily focused on passenger entry and exit information, lacking additional details such as demographic data, trip purposes, or fare information. This limited the depth of analysis we could perform.Missing Data: Some entries in the dataset had missing or incomplete information, which required careful handling during the analysis to ensure accurate results.

Despite these shortcomings, the Metro Data provided a valuable foundation for our analysis, allowing us to gain insights into passenger mobility patterns in the Delhi Metro system.

2.1.3 Additional Data

In addition to the primary dataset, we acquired additional data from various sources to enrich our analysis and provide a more comprehensive understanding of the Delhi Metro system. The additional data included:

Line Information: To incorporate information about the lines to which each metro station belongs, we sourced a dataset from Kaggle. This dataset provided details on the lines, including line codes and names. By leveraging the VLOOKUP function, we seamlessly integrated this data with our existing dataset, allowing us to analyze the metro system at a line level.

Geographical Coordinates: Accurate geographical coordinates of each metro station are crucial for spatial analysis and visualization. To obtain this information, we utilized web scraping techniques with the help of the Selenium library in Python. By automating the process, we collected the latitude and longitude coordinates of each station, ensuring precise geospatial representation in our analysis.

2.1.4 Data Preprocessing

To enhance performance and reduce processing time for our large dataset (16.8 million+ entries), we implemented data caching using local storage. This approach involved storing the processed data locally, allowing us to quickly access and reuse it without repeated computations.

By following these steps, we optimized our workflow:

Data Processing: We performed data cleaning, transformation, and analysis on the complete dataset.

Local Storage Storage: After processing, we stored the data in the local storage of our system.

Retrieval and Reuse: When needed for subsequent analyses or computations, we directly loaded the processed data from local storage, eliminating the need to rerun intensive operations on the entire dataset.

By leveraging local storage for data caching, we achieved faster access to preprocessed data, improving overall efficiency and reducing processing time. It enabled us to focus on data analysis and interpretation, rather than repetitive computations.

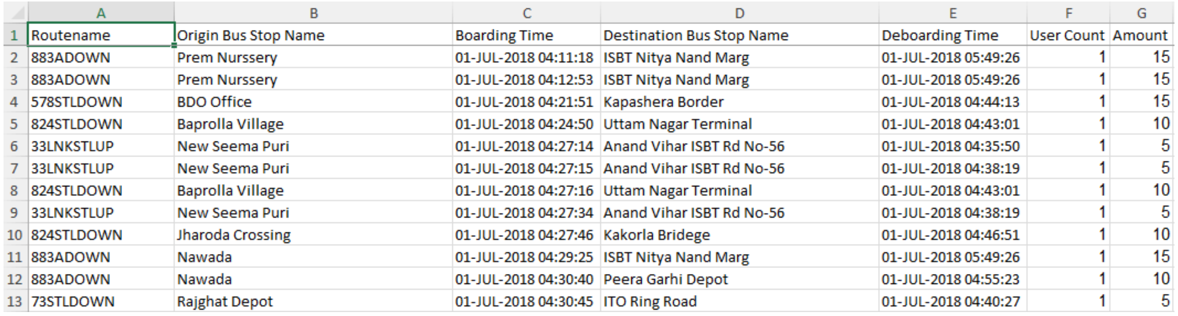
Implementing data caching with local storage provided us with an efficient way to manage and reuse processed data, streamlining our project's workflow and enhancing performance.

2.2 Bus Data

2.2.1 About the Dataset

The bus data that we began with in this project consists of passenger e-ticket data and DTC (Delhi Transport Corporation) bus route trip data.

The passenger e-ticket data includes the following information:

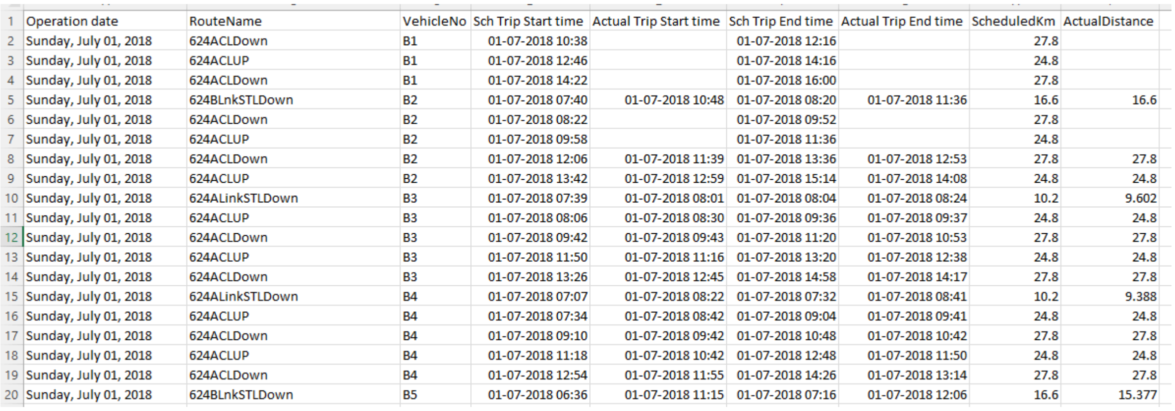


	A	B	C	D	E	F	G
1	Routename	Origin Bus Stop Name	Boarding Time	Destination Bus Stop Name	Deboarding Time	User Count	Amount
2	883ADOWN	Prem Nurssery	01-JUL-2018 04:11:18	ISBT Nitya Nand Marg	01-JUL-2018 05:49:26	1	15
3	883ADOWN	Prem Nurssery	01-JUL-2018 04:12:53	ISBT Nitya Nand Marg	01-JUL-2018 05:49:26	1	15
4	578STLDOWN	BDO Office	01-JUL-2018 04:21:51	Kapashera Border	01-JUL-2018 04:44:13	1	15
5	824STLDOWN	Baprolla Village	01-JUL-2018 04:24:50	Uttam Nagar Terminal	01-JUL-2018 04:43:01	1	10
6	33LNKSTLUP	New Seema Puri	01-JUL-2018 04:27:14	Anand Vihar ISBT Rd No-56	01-JUL-2018 04:35:50	1	5
7	33LNKSTLUP	New Seema Puri	01-JUL-2018 04:27:15	Anand Vihar ISBT Rd No-56	01-JUL-2018 04:38:19	1	5
8	824STLDOWN	Baprolla Village	01-JUL-2018 04:27:16	Uttam Nagar Terminal	01-JUL-2018 04:43:01	1	10
9	33LNKSTLUP	New Seema Puri	01-JUL-2018 04:27:34	Anand Vihar ISBT Rd No-56	01-JUL-2018 04:38:19	1	5
10	824STLDOWN	Jharoda Crossing	01-JUL-2018 04:27:46	Kakorla Bridege	01-JUL-2018 04:46:51	1	10
11	883ADOWN	Nawada	01-JUL-2018 04:29:25	ISBT Nitya Nand Marg	01-JUL-2018 05:49:26	1	15
12	883ADOWN	Nawada	01-JUL-2018 04:30:40	Peera Garhi Depot	01-JUL-2018 04:55:23	1	10
13	73STLDOWN	Rajghat Depot	01-JUL-2018 04:30:45	ITO Ring Road	01-JUL-2018 04:40:27	1	5

Fig 4. Passenger e-ticket data

Routename: The name of the bus route.
Origin Bus Stop Name: The name of the bus stop where passengers boarded the bus.
Boarding Time: The time at which passengers boarded the bus.
Destination Bus Stop Name: The name of the bus stop where passengers disembarked.
Deboarding Time: The time at which passengers disembarked from the bus.
User Count: The number of passengers on the bus.
Amount: The fare amount for the journey.

Additionally, the dataset also includes DTC bus route trip data, which provides information about the routes and schedules of DTC buses.



1	Operation date	RouteName	VehicleNo	Sch Trip Start time	Actual Trip Start time	Sch Trip End time	Actual Trip End time	ScheduledKm	ActualDistance
2	Sunday, July 01, 2018	624ACLDown	B1	01-07-2018 10:38		01-07-2018 12:16		27.8	
3	Sunday, July 01, 2018	624ACLUP	B1	01-07-2018 12:46		01-07-2018 14:16		24.8	
4	Sunday, July 01, 2018	624ACLDown	B1	01-07-2018 14:22		01-07-2018 16:00		27.8	
5	Sunday, July 01, 2018	624BLnkSTLDown	B2	01-07-2018 07:40	01-07-2018 10:48	01-07-2018 08:20	01-07-2018 11:36	16.6	16.6
6	Sunday, July 01, 2018	624ACLDown	B2	01-07-2018 08:22		01-07-2018 09:52		27.8	
7	Sunday, July 01, 2018	624ACLUP	B2	01-07-2018 09:58		01-07-2018 11:36		24.8	
8	Sunday, July 01, 2018	624ACLDown	B2	01-07-2018 12:06	01-07-2018 11:39	01-07-2018 13:36	01-07-2018 12:53	27.8	27.8
9	Sunday, July 01, 2018	624ACLUP	B2	01-07-2018 13:42	01-07-2018 12:59	01-07-2018 15:14	01-07-2018 14:08	24.8	24.8
10	Sunday, July 01, 2018	624ALinkSTLDown	B3	01-07-2018 07:39	01-07-2018 08:01	01-07-2018 08:04	01-07-2018 08:24	10.2	9.602
11	Sunday, July 01, 2018	624ACLUP	B3	01-07-2018 08:06	01-07-2018 08:30	01-07-2018 09:36	01-07-2018 09:37	24.8	24.8
12	Sunday, July 01, 2018	624ACLDown	B3	01-07-2018 09:42	01-07-2018 09:43	01-07-2018 11:20	01-07-2018 10:53	27.8	27.8
13	Sunday, July 01, 2018	624ACLUP	B3	01-07-2018 11:50	01-07-2018 11:16	01-07-2018 13:20	01-07-2018 12:38	24.8	24.8
14	Sunday, July 01, 2018	624ACLDown	B3	01-07-2018 13:26	01-07-2018 12:45	01-07-2018 14:58	01-07-2018 14:17	27.8	27.8
15	Sunday, July 01, 2018	624ALinkSTLDown	B4	01-07-2018 07:07	01-07-2018 08:22	01-07-2018 07:32	01-07-2018 08:41	10.2	9.388
16	Sunday, July 01, 2018	624ACLUP	B4	01-07-2018 07:34	01-07-2018 08:42	01-07-2018 09:04	01-07-2018 09:41	24.8	24.8
17	Sunday, July 01, 2018	624ACLDown	B4	01-07-2018 09:10	01-07-2018 09:42	01-07-2018 10:48	01-07-2018 10:42	27.8	27.8
18	Sunday, July 01, 2018	624ACLUP	B4	01-07-2018 11:18	01-07-2018 10:42	01-07-2018 12:48	01-07-2018 11:50	24.8	24.8
19	Sunday, July 01, 2018	624ACLDown	B4	01-07-2018 12:54	01-07-2018 11:55	01-07-2018 14:26	01-07-2018 13:14	27.8	27.8
20	Sunday, July 01, 2018	624BLnkSTLDown	B5	01-07-2018 06:36	01-07-2018 11:15	01-07-2018 07:16	01-07-2018 12:06	16.6	15.377

Fig 5: Bus Trip Information

2.2.2 Shortcomings

There were a lot of shortcomings in the bus data. As is apparent from the screenshot of the data placed above as well, there are cells with empty values which would introduce inconsistencies in the results.

Also, a major issue is that the information about the detailed route is not present in the given dataset, for which we had to search elsewhere to be able to make proper use of this data. Also, similar to how it was in the case of the Metro Station names, the bus stop names are also very inconsistent across datasets that we found from various sources. To incorporate them together, we again had to implement fuzzy search based implementation to approximate the names and identify similarities. With the help of that we were often able to increase the amount of usable data by more than 50%.

2.2.3 Additional Data

We expanded our dataset by including crucial information such as routes, stop times, stops, trips, and a calendar of the bus schedule. To acquire this valuable data, we leveraged the General Transit Feed Specification (GTFS), a standardized format for public transportation data. The GTFS data provided us with a comprehensive and structured collection of data elements, enabling us to establish a robust foundation for our analysis.

However, incorporating the GTFS data into our existing dataset posed a significant challenge. The data originated from a different source and required careful integration to ensure consistency and compatibility. We devoted considerable effort to harmonizing the datasets, aligning the relevant fields, and establishing meaningful relationships between the data elements. This meticulous process allowed us to create a unified and comprehensive dataset that encompassed both the passenger e-ticket data and the enriched GTFS data.

2.2.4 Data preprocessing

Data preprocessing for the bus data involved essential steps to enhance data quality, consistency, and compatibility with our analytical framework. Initially, we focused on removing duplicate and erroneous entries from the dataset, ensuring accurate and reliable data for analysis. We standardized the formatting of bus-related information, including bus stops and routes, to achieve consistency across different variables. Timestamp data was transformed to capture boarding and deboarding times, enabling the analysis of temporal patterns. Validation and verification were conducted by cross-referencing the bus data with external sources, resolving discrepancies using fuzzy search methods. These preprocessing efforts ensured that the bus data was cleansed, formatted, and ready for analysis, maintaining data quality and integrity throughout the process.

Chapter 3 : Research and Methodology

3.1 Modeling the data as a graph

The data we are dealing with involves metro stations, out of which some are connected to each other, in the form of lines (Red Line, Pink Line, etc). Now, if we observe carefully, this can be modeled as an undirected and unweighted graph where the edges are consecutive stations on a line, and the stations are nodes.

3.2 Understanding the problem

We have been given the passenger data which only contains the entry exit, but does not tell us about the route that the passenger has taken. However, to truly analyze the data, we need to find it out so that we can analyze overall route loads.

3.2.1 Path finding

To determine the route load based on the given starting and ending stations, we approach it as a pathfinding problem within a graph representation. By modeling the transportation data as a graph, we can leverage algorithms to identify the shortest path between two points.

In this analysis, we make the assumption that the passengers choose the shortest path when traveling from point A to point B. By considering this assumption, we can estimate the route load by aggregating the number of passengers between each pair of stations along the shortest path.

3.2.2 Shortest path algorithms

Now, between two nodes, a large number of paths exist. However, a metro does not take any path, it only takes the shortest path. To do this, multiple algorithms have been developed, such as Djikstra’s, Breadth First Search, Depth First Search, etc.

3.2.3 Dijkstra's Algorithm

For our problem, we assume that the edges are weighted, as the time to cover each segment between stations may vary depending on factors such as distance or traffic. This assumption was made after a detailed discussion with our supervisor. As a result, we decided to use Dijkstra's Algorithm (for each node separately) to perform this task.

Dijkstra's Algorithm is a graph traversal algorithm used to find the shortest path from a source node to all other nodes in a graph. It is particularly well-suited for weighted graphs, where each edge has an associated cost (such as distance or time). The algorithm maintains a priority queue to track the next most promising node to explore based on the cumulative cost of reaching it. By systematically exploring nodes with the lowest cost first, Dijkstra's algorithm ensures that the shortest path to each node is determined efficiently.

Time Complexity:

$O(|E| \log |V|)$

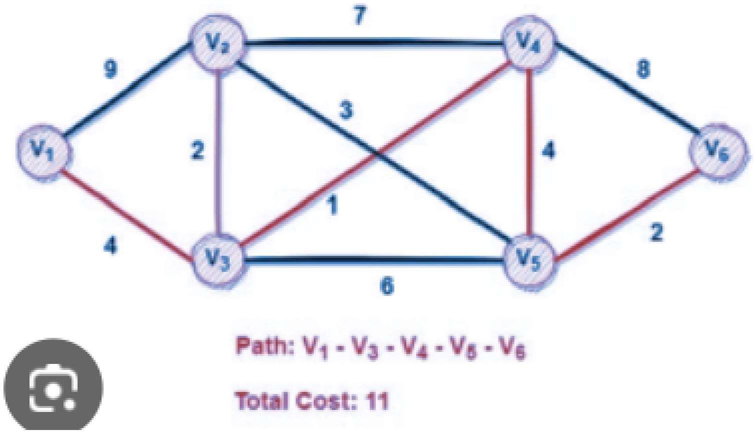


Fig 6: Visualization of Dijkstra’s Algorithm

3.2.4 Fuzzy Search

In our analysis of bus and metro stations, we encountered inconsistencies between the names provided in the dataset and the names found on external sources such as Wikipedia and the internet. These inconsistencies primarily arose due to the presence of names in a mix of Hindi and English, commonly referred to as "Hinglish." For instance, the bus station "Prem

Nursery" was recorded as "Prem Nurssary" in the bus passenger dataset, and the metro station "Ghitorni" appeared as "Ghitorny" in our metro smart card data.

These variations in station names posed significant challenges during the data scraping process, specifically when retrieving geodata from Wikipedia and using VLOOKUP to match entries. To address this issue, we employed fuzzy search techniques, which involve calculating the Levenshtein distance to determine the best possible match for a given string value.

By utilizing fuzzy search, we were able to overcome the discrepancies caused by the differences in station names. This allowed us to effectively match and integrate the data from various sources, ensuring the accuracy and consistency of our analysis.

3.2.5 Handling Shapefiles

Shapefiles were employed for zone-wise analysis of bus stops. Shapefiles are geospatial vector data formats used in GIS applications. By integrating shapefiles delineating the zones in Delhi, bus stops were attributed to their respective zones. This facilitated analysis of bus stop characteristics and performance on a zone-by-zone basis. The approach provided insights into bus stop distributions, passenger load, and route connectivity within each zone. This spatially informed analysis informed recommendations for optimizing the bus system in different zones. Overall, shapefiles enabled a condensed and precise examination of bus stop dynamics and their relationships to specific zones in Delhi.

3.3 Workshop at IIIT-D CSM

A workshop organized by CSM (Centre of Sustainable Mobility), IIIT Delhi proved instrumental in enhancing our understanding of working with bus data. The workshop provided valuable insights and techniques for efficiently handling bus data, such as identifying a set of stops along a given route and plotting stations within a specific radius. These tasks proved to be subsets of our overall analysis objectives. The knowledge gained from the workshop equipped us with practical skills and methodologies, enabling us to effectively tackle the challenges and complexities involved in analyzing the multimodal public transportation system in Delhi

Chapter 4: Metro Analysis

4.1 Station Load Analysis

4.1.1 Calculating Station Load:

For calculating the station load, we first discretized time domain and **bucketed** it into 10 minute buckets (total 6*24=144 time buckets)

We then incremented the numpy array indices corresponding to a person's in-station id, out-station id, time bucket and the day at which he/she was traveling.

4.1.2 Station Load Results:

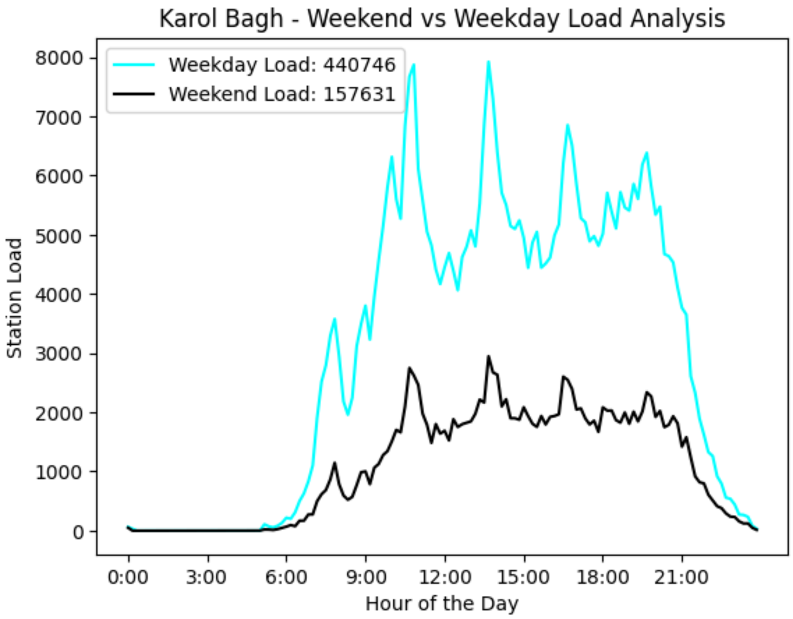


Fig 7: Karol Bagh - Weekend vs Weekday Load Analysis

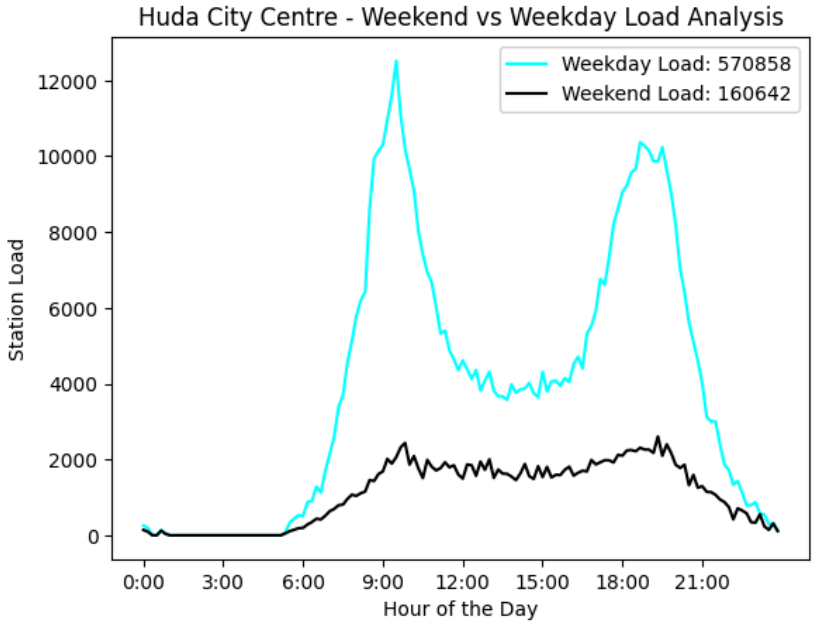


Fig 8: Huda City Centre - Weekend vs Weekday Load Analysis

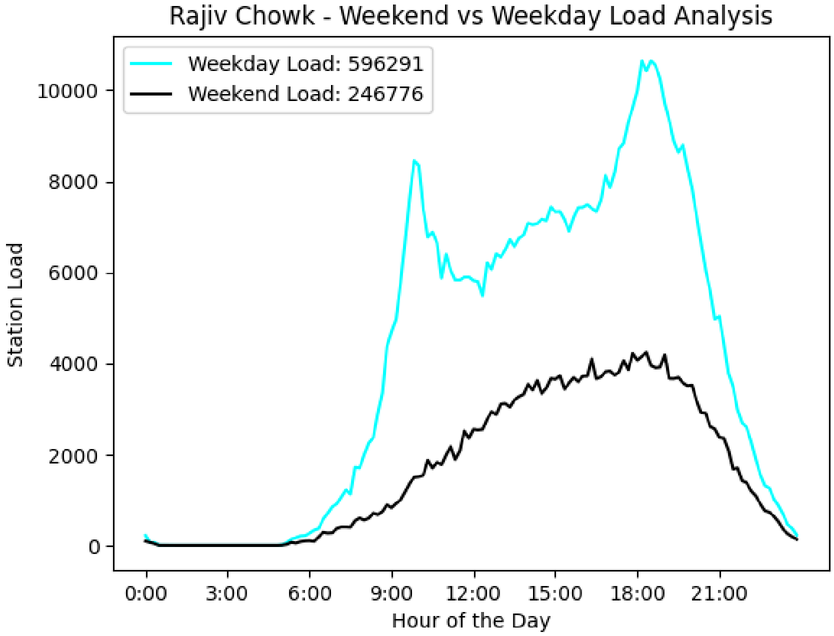


Fig 9: Rajiv Chowk - Weekend vs Weekday Load Analysis

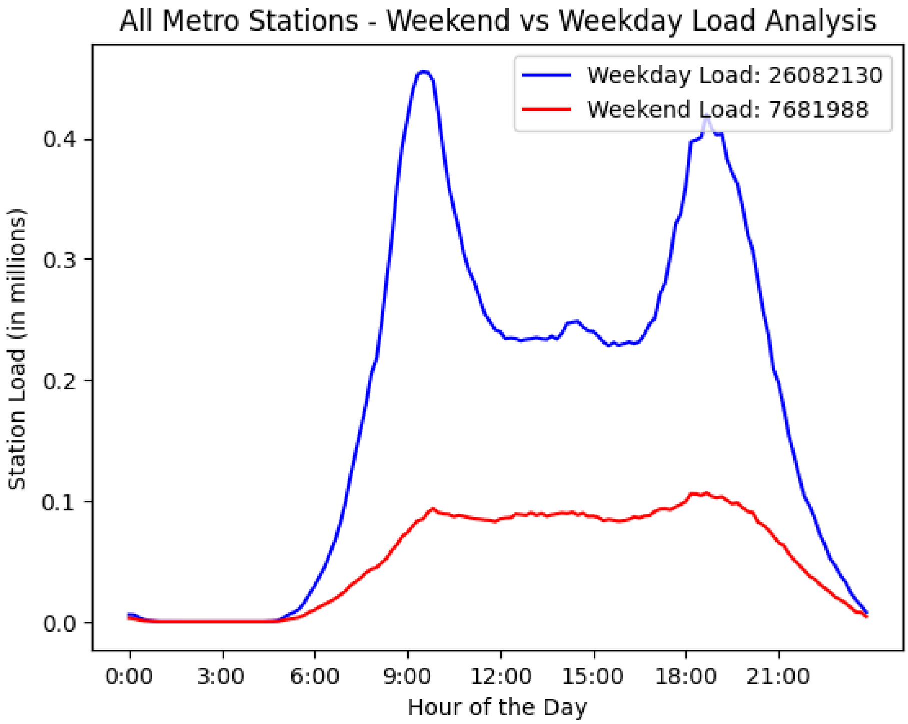


Fig 10: Weekend vs Weekday Load Analysis for all Metro Stations

- It is quite apparent from the plots that Delhi metro is quite busy in general, during the weekdays compared to weekends.

- 9:00 AM and 6:00 PM are found to be the busiest hours as clear peaks can be observed around this time, on both weekdays as well as weekends.. This is also understandable since most people are commuting from their workplaces to their homes (or vice versa) during these hours.

4.1 Route Load Analysis

4.2.1 Calculating Route Load:

First we pulled Delhi Metro Geodata from Wikipedia and mapped it to all the station ids present in our dataset.Then, an undirected unweighted graph was constructed using the dataset, drawing an edge between consecutive stations on a given line. A shortest path algorithm was implemented from scratch to find the shortest path between any 2 given stations. For any 2 starting and ending stations, all the stations lying in the optimal path between them are incremented for journey in the corresponding numpy array indices.

Finally, a plot indicating route load is plotted. For that, we first calculated the z score for a particular route load using the following formula:

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

Edges/Routes having a Z-Score > 1 (lying in the top 84 percentile) are assigned highest edge weight while those having a Z-Score of < -1 (lying in the below 16 percentile), least edge weight.

4.2.2 Route Load Results:

We plot the Delhi Metro route load for 3 randomly selected time slots.

(**Note**: Thickness in the plots is indicative of how busy a particular route is. Thicker the edge, heavier the load.)

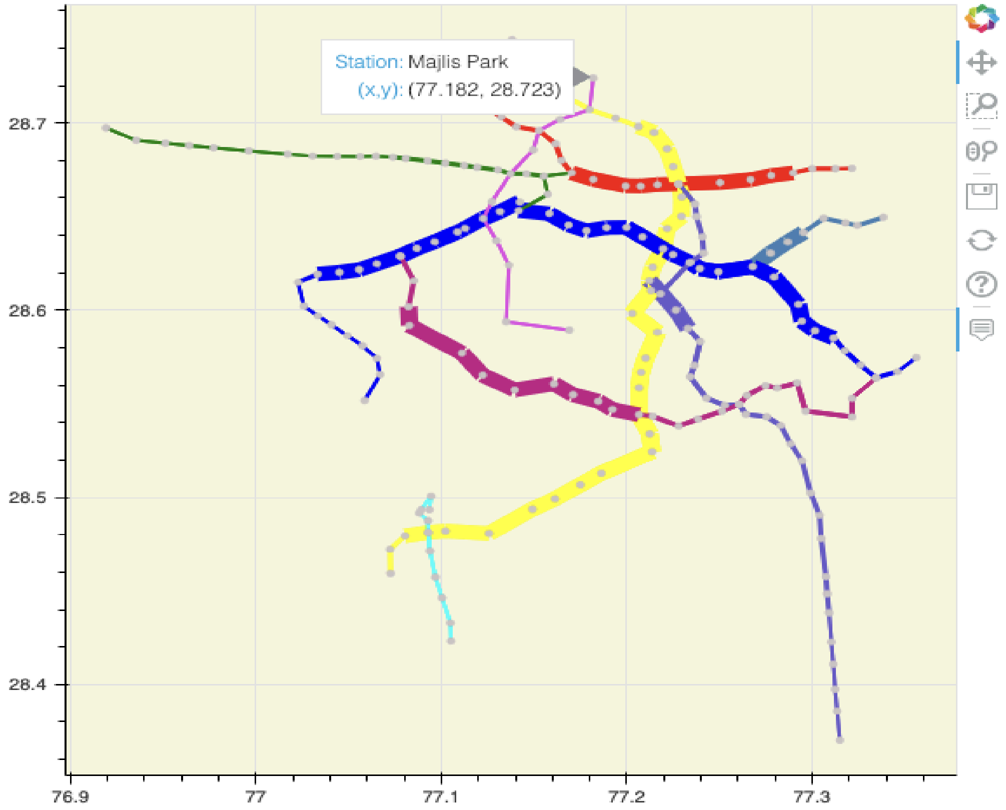


Fig 11: Delhi Metro Routes at 8:45 AM on Weekday

8:45 AM on a weekday: At this time, most of the lines (Blue, Red, Magenta, Yellow) and some portions of the Violet Line are busy. This is obviously what one might expect since 8:45 AM on a weekday is when most people are commuting to their workplaces from their homes.

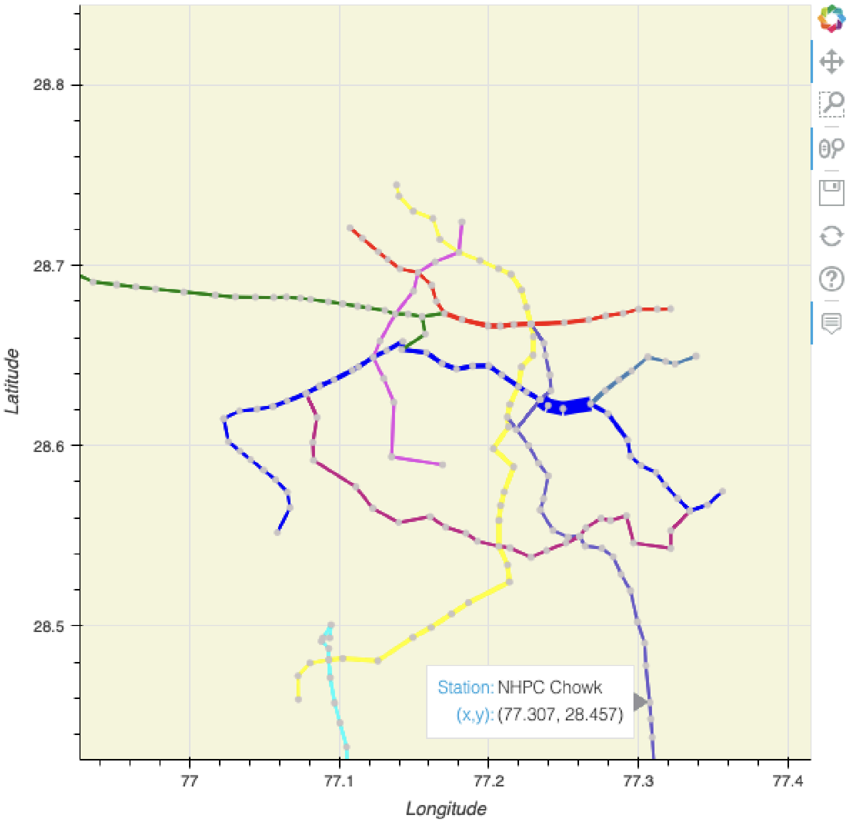


Fig 12: Delhi Metro Routes at 12:34 PM on a Weekday

- **12:34 PM on a weekday**: Only 4 stations (Mandi House,Pragati Maidan, Indraprastha and Yamuna Bank) lying on Blue Line experience a considerable amount of traffic at this point of time.

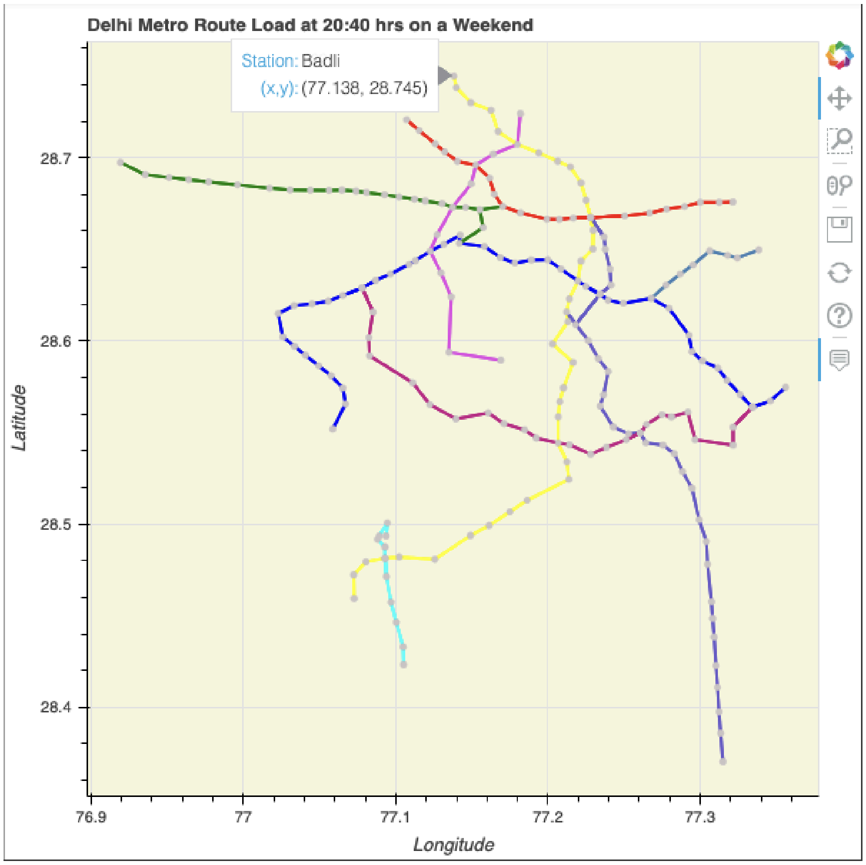


Fig 13: Delhi Metro Routes at 8:40 PM on a Weekend

- **8:40 PM on a weekend**: At this time, Delhi Metro seems to be relatively low on traffic. This also seems logical since one would not expect a lot of people using the metro service at this late hour and especially on a weekend.

Chapter 5: Bus Analysis

5.1 Station Load Analysis

A similar technique was applied for analyzing the station load in the bus system. However, the data for buses was even more unstructured due to inconsistencies in station names. These inconsistencies posed challenges as there were variations in naming conventions. Despite this hurdle, we employed strategies to handle the discrepancies and ensure accurate analysis of bus station loads.

5.1.1 Calculating Station Load

To calculate the station load, we initially divided the time domain into discrete 10-minute intervals, resulting in 144 time buckets covering a span of 24 hours. Next, we incremented the corresponding indices in a numpy array based on the where the passenger boarded/deboarded the bus stop,their day of travel and the time of the day during boarding/deboarding.

5.1.2 Station Load Results

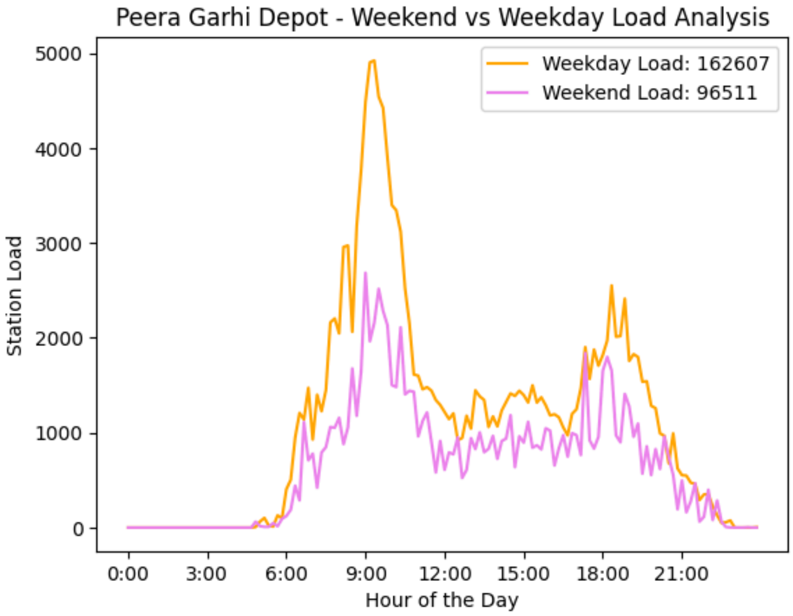


Fig 14: Peera Garhi Depot - Weekend vs Weekday Load Analysis

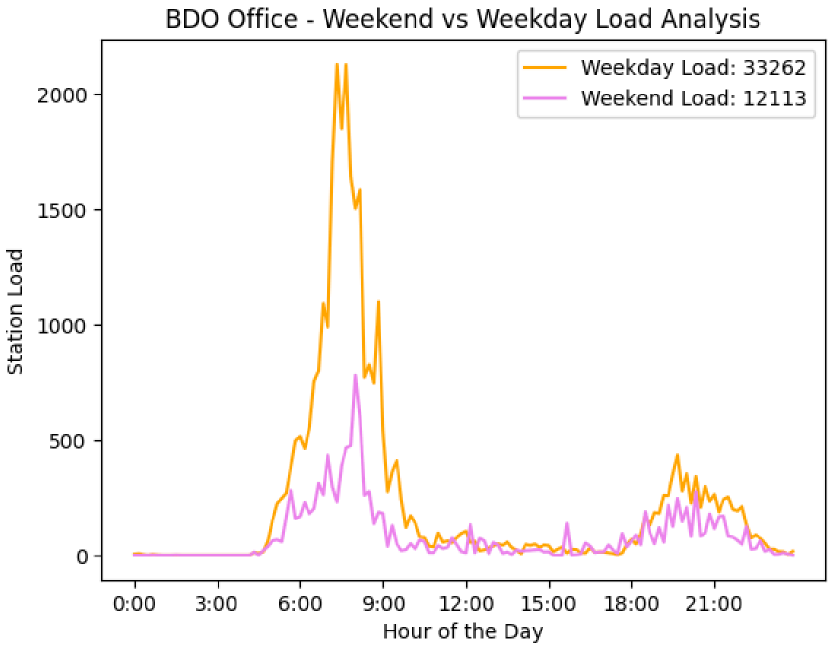


Fig 15: BDO Office - Weekend vs Weekday Load Analysis

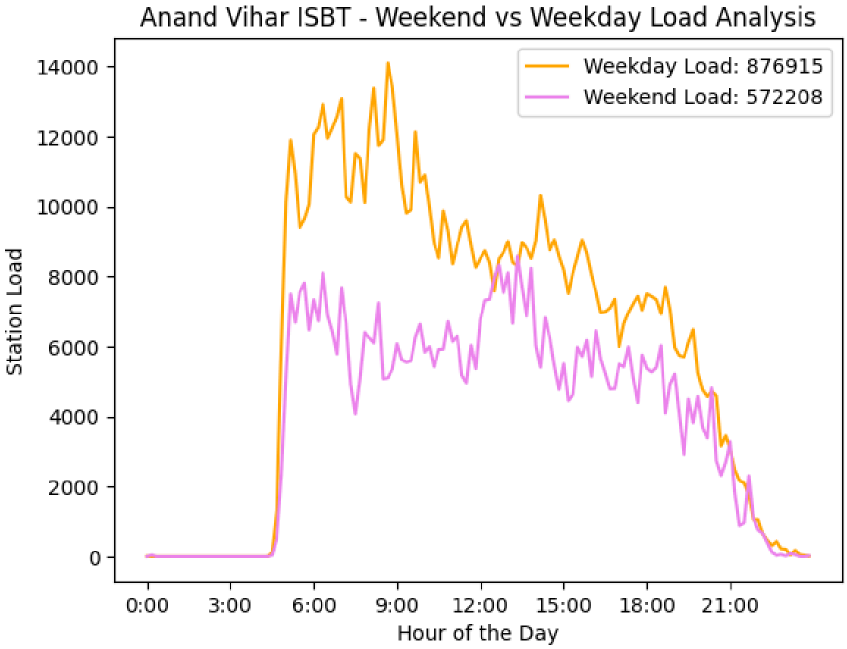


Fig 16: Anand Vihar ISBT - Weekend vs Weekday Load Analysis

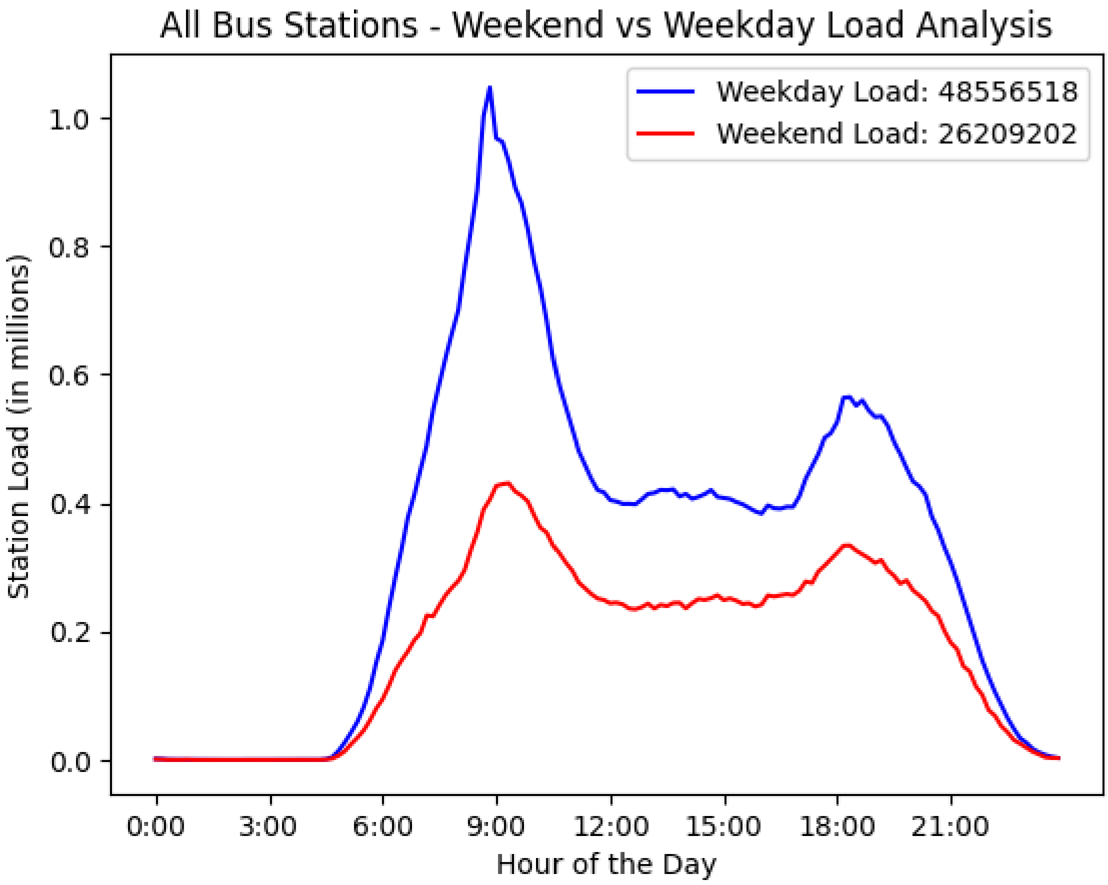


Fig 17: Weekend vs Weekday Load Analysis for all Bus Stations

- It is evident from the analysis that the bus system in Delhi experiences higher activity levels on weekdays compared to weekends, similar to the metro system.
- During the weekdays and weekends, the busiest hours for buses are observed to be between 9:00 AM and 6:00 PM, coinciding with peak commuting times when people travel between their workplaces and homes.
- It is noteworthy that there is relatively less congestion in buses during the evenings compared to the mornings. This could be attributed to safety concerns associated with bus travel after dark and the preference for metro transportation due to traffic conditions.

5.2 Bus Stations: Zonewise Analysis

A zone-wise analysis of bus stops was conducted to gain insights into the congestion levels and connectivity within New Delhi. The city was segmented into nine distinct zones, allowing us to examine the number of bus stations and station density within each zone. This approach aimed to provide a comprehensive understanding of the crowd density and connectivity quality within specific zones. By analyzing the distribution and density of bus stops, we obtained valuable insights into the transportation infrastructure's effectiveness and identified areas that require improved connectivity.

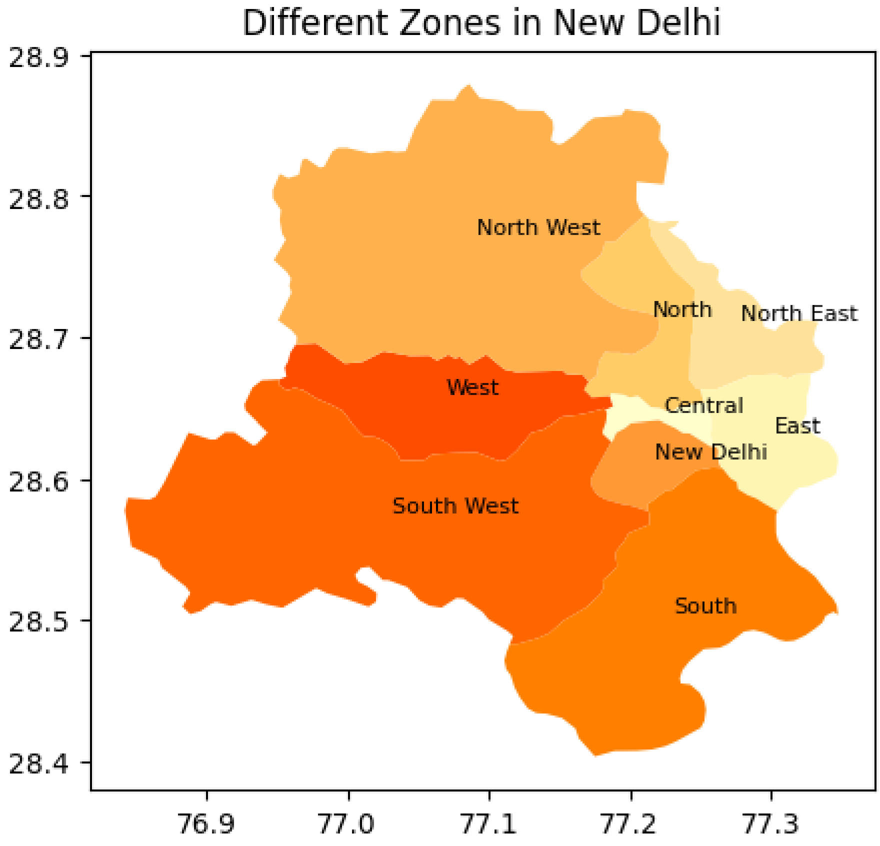


Fig 18: Zonewise classification of Delhi bus network

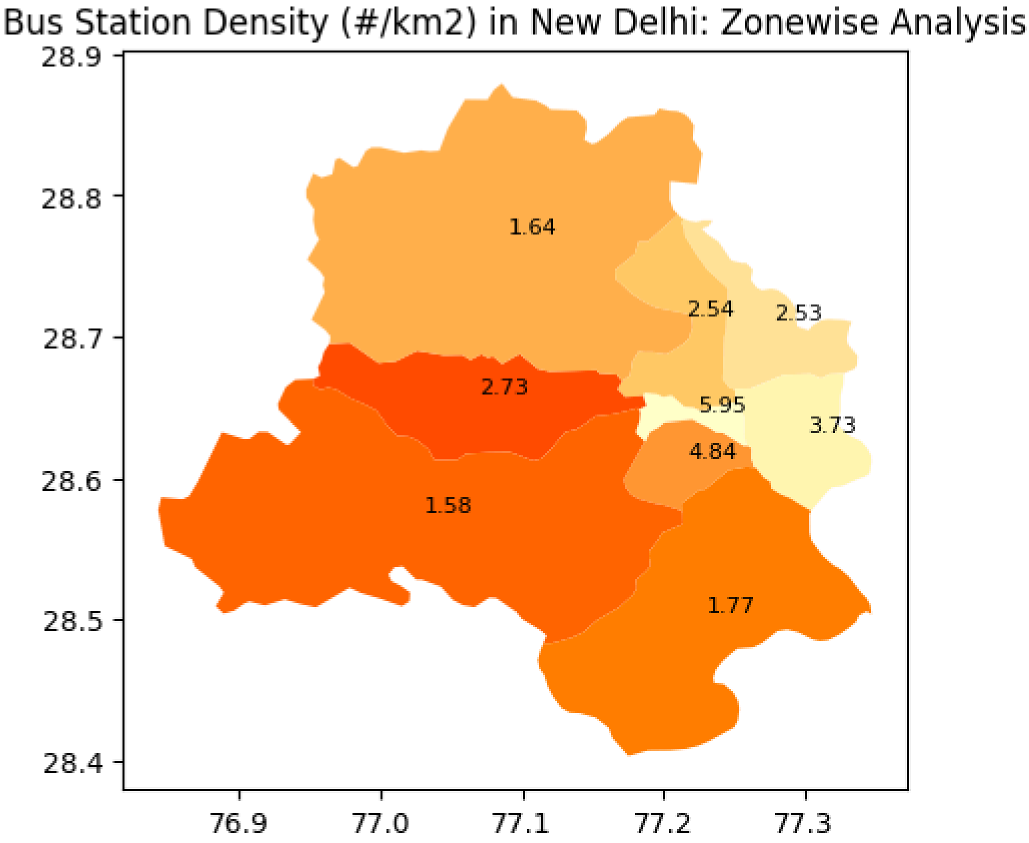


Fig 19: Bus station density : Zonewise analysis

- As we can see, North West Delhi region has the least Bus Stop Density of 1.64 Bus stops per km^2 across all zones while Central Delhi seems to have the highest Bus Stop Density of 5.95 Bus stops per km^2.
- This zone-wise analysis contributes to a better understanding of the overall performance and efficiency of the bus system in Delhi.

Chapter 6

6. Recommendations

6.1 Metro

To enhance the efficiency of the public transportation system and improve passenger experience, several strategies and measures can be implemented based on the analysis of the multimodal transportation data. Firstly, increasing the frequency of metro services on congested routes can help alleviate overcrowding and reduce delays. By identifying off-peak hours specific to each station, incentives such as discounts can be provided to encourage travel during less crowded periods.

Optimizing route planning is another crucial aspect that can be derived from the station load data. By identifying high-demand routes, transportation authorities can schedule more frequent services during peak hours, ensuring smoother travel and reduced congestion. Additionally, leveraging the station load data can enhance passenger information systems. Real-time information on crowded stations, expected wait times, and alternative routes can be provided through digital displays, mobile apps, and public address systems.

To address the growing demand, capacity expansion measures should be considered. Stations and routes that consistently experience high congestion can be identified, and infrastructure improvements such as increasing the number of coaches or buses can be implemented. Furthermore, there is an opportunity to generate revenue and provide relevant information by partnering with advertising agencies to strategically place advertisements inside metros and buses.

Integration with other modes of transport is crucial for seamless travel experiences. Analyzing the data to identify major interchanges between metro and bus stations can lead to improved connectivity and the implementation of integrated ticketing systems or dedicated shuttle services. Crowd management strategies can also be developed based on the station load data to ensure passenger safety and a smooth flow of commuters during peak hours.

Collecting feedback from passengers is vital in understanding their concerns, suggestions, and priorities. Establishing feedback mechanisms can help transportation authorities make informed decisions and prioritize improvements that address the specific needs of the commuters. By incorporating these strategies and measures, the multimodal transportation system can be optimized, resulting in enhanced efficiency, reduced overcrowding, and improved passenger satisfaction.

6.2 Buses

To enhance bus service reliability, it is crucial to implement various measures aimed at improving the overall efficiency and performance of the system. These measures include the establishment of dedicated bus lanes, prioritizing buses at traffic signals, and implementing better traffic management strategies, especially during peak hours. By ensuring smoother bus operations, travel time can be significantly reduced, making buses a more attractive option even during congested evening hours.

Additionally, the implementation of integrated ticketing and fare systems is essential to facilitate seamless transfers between buses and metros with a single ticket or smart card. This integrated approach encourages intermodal travel, providing passengers with greater convenience and eliminating the need for separate ticket purchases.

To further enhance the bus system, exploring public-private partnerships can bring innovative solutions and technologies on board. Collaborating with private companies can introduce features like mobile ticketing apps, on-board infotainment systems, and digital advertising platforms. These partnerships not only enhance the passenger experience but also generate additional revenue for the government.

Furthermore, targeted marketing campaigns should be developed to raise awareness among commuters about the advantages of bus travel during congested hours. By emphasizing the reliability, comfort, and cost-effectiveness of buses compared to other modes of transport, more individuals can be encouraged to choose buses as their preferred means of transportation. Offering incentives such as discounts, loyalty programs, or special passes can further attract commuters to the bus system.

Finally, data-driven planning and optimization play a crucial role in ensuring the efficiency of bus services. Analyzing passenger data, including travel patterns and preferences, allows for continuous optimization of bus routes, frequencies, and capacity. Utilizing predictive analytics enables the anticipation of demand fluctuations, ensuring that bus services are adjusted accordingly to effectively utilize available resources.

Incorporating these strategies and initiatives into the bus system can significantly enhance its reliability, efficiency, and attractiveness to commuters, ultimately improving the overall public transportation experience.

Chapter 7

7. Conclusion and Future Work

7.1 Summary and conclusion

In this study, we explored the passenger mobility analysis of the Delhi Metro and bus systems. Through comprehensive data sourcing, analysis, and preprocessing, we gained valuable insights into the station and route load patterns, as well as overall passenger flow. By leveraging datasets from various sources, including the Delhi Metro passenger entry-exit data, additional data on metro lines and bus routes, and GTFS data, we were able to create a holistic view of the transportation system.

Our analysis revealed the significant benefits of such studies for both the government and the public. The findings can assist transportation authorities in optimizing resource allocation, improving operational efficiency, and enhancing the overall passenger experience. For the public, the insights gained can aid in making informed decisions regarding travel routes, timings, and mode of transportation.

7.2 Future work

While this study provides valuable insights into the passenger mobility of the Delhi Metro and bus systems, there are several avenues for future research and improvement. First, integrating real-time data from both the metro and bus systems can provide more accurate and up-to-date information on passenger flow and congestion. This can be achieved by incorporating live tracking systems and passenger count data into the analysis framework.

Additionally, the inclusion of demographic data, such as age groups, occupation, and travel purposes, can offer a deeper understanding of passenger preferences and behavior. This can enable the development of targeted strategies and services to cater to specific passenger segments.

Moreover, expanding the analysis to encompass multi-modal transportation, including other modes like taxis and shared mobility services, can provide a more comprehensive understanding of the overall transportation ecosystem. By incorporating data from diverse sources, a holistic perspective on urban mobility can be obtained.

In conclusion, this study serves as a foundation for further research in the field of passenger mobility analysis and offers valuable insights for transportation planning and management. By leveraging data-driven approaches, policymakers and transportation authorities can make informed decisions to improve the efficiency, accessibility, and sustainability of public transportation systems.

References

[1] https://en.wikipedia.org/wiki/Delhi_Metro

[2] https://selenium-python.readthedocs.io/

[3] https://matplotlib.org/

[4] Etikaf Hussain a, Ashish Bhaskar a, *, Edward Chung b, “Transit OD matrix estimation using smartcard data: Recent developments and future research challenges”, Transportation Research Part C 125 (2021) 103044

[5] Marcela A. Munizaga a,⇑, Carolina Palma b,1, “Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile”, Transportation Research Part C 24 (2012) 9–18 (http://dx.doi.org/10.1016/j.trc.2012.01.007)

[6] Di Huang ,1 Jun Yu,1 Shiyu Shen,2 Zhekang Li,1 Luyun Zhao,2 and Cheng Gong2, “A Method for Bus OD Matrix Estimation Using Multisource Data”, Journal of Advanced Transportation (https://doi.org/10.1155/2020/5740521)

[7] Li -Xiao Gao, Ji-hua Hu, Guo-yuan Li, Jia-xian Liang, “A calculation method of OD matrix in multi-modal transit network based on traffic big data”, The 3rd International Conference on Transportation Information and Safety, June 25 – June 28, 2015, Wuhan, P. R. China

[8] Shahriar Afandizadeh Zargari 1, * , Amirmasoud Memarnejad 1, Hamid Mirzahossein, “Hourly Origin–Destination Matrix Estimation Using Intelligent Transportation Systems Data and Deep Learning”, Sensors 2021, 21, 7080

[9] https://python-visualization.github.io/folium/