MASTERARBEIT | MASTER'S THESIS

Titel | Title

Impact of Weather Conditions on Public Transport Ridership:
A Study of Bratislava, Slovakia

verfasst von | submitted by

Petra Baliová

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 977
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Business Analytics
Betreut von | Supervisor:	Univ.-Prof. Dr. Jan Fabian Ehmke

Abstract

EN This thesis investigates the utilisation and ridership patterns of public transport in Bratislava, with a focus on the relationship between weather conditions and passenger numbers across buses, trolleybuses, and trams. Its methodology is grounded in the CRISP-DM model. The modelling involved training SARIMAX for time series and XGBoost for predictive analysis. The results reveal that temperature has a modest impact on bus ridership on weekends, leading to a slight increase in number of passengers, while precipitation has a significant negative impact on ridership of both buses and trolleybuses. Humidity minimally affects bus usage but significantly impacts trolleybuses on weekdays. Tram ridership declines during weekends with higher temperatures and precipitation. School and public holidays also exhibit a notable decrease in ridership across all modes of transport. Lastly, the study offers actionable strategies for urban transportation planning, emphasising data-driven decision-making such as employing resource allocation based on temporal and weather-related patterns to anticipate ridership shifts. In summary, this research contributes to understanding the impact of weather conditions on public transport ridership, providing valuable insights for the enhancement of Bratislava's transit system.

DE Diese Masterarbeit erforscht die Nutzungs- und Fahrgastzahlmuster des öffentlichen Personenverkehrs in Bratislava mit Fokus auf den Zusammenhang zwischen Wetterbedingungen und Fahrgastzahlen bei Bussen, Oberleitungsbussen und Straßenbahnen. Die Methode basiert auf dem CRISP-DM Modell. Die Modellierung beinhaltet ein SARIMAX-Training für Zeitreihen sowie XGBoost für prädiktive Analyse (predictive analysis). Die Ergebnisse zeigen, dass die Temperatur einen leichten positiven Effekt auf die Busnutzung an Wochenenden hat. Auswirkungen von Niederschlag sind dagegen signifikant negativ bei Bussen mit und ohne Oberleitung. Luftfeuchtigkeit führt zu kaum Unterschieden bei der Busnutzung, allerdings zu starken Effekten bei der Nutzung von Oberleitungsbussen an Wochenenden. Die Straßenbahnnutzung zeigt sich abhängig von hohen Temperaturen und Niederschlag an Wochenenden mit verminderter Nutzung. Schulferien sowie Feiertage führen zu einem allgemeinen Rückgang bei allen öffentlichen Transportmöglichkeiten. Abschließend erschließt diese Studie Strategien für die Planung öffentlicher Verkehrssysteme mit Fokus auf der datenbasierten Entscheidungsfindung, wie z.B. Personalplanung basierend auf Zeit- und Wetter Faktoren. Im Ganzen zeigt diese Arbeit somit Abhängigkeiten der Fahrgastzahlen im öffentlichen Verkehrssystemen von Wetterbedingungen auf, aus denen Ideen für die Verbesserung des öffentlichen Personennahverkehrs in Bratislava abgeleitet werden.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Univ.- Prof. Dr. Jan Fabian Ehmke for his time, guidance and valuable comments. I would also like to extend my gratitude to The Municipality of the City of Bratislava, especially to PhDr. Viktor Robert Bederka and MSc Milan Fedorik, for their trust, guidance and for giving me the opportunity to work with such an interesting dataset.

Contents

1. Introduction

The proposed research entails a comprehensive analysis of the utilisation and ridership patterns of public transportation in Bratislava, specifically examining the relationship between weather conditions and the number of passengers using different vehicle types, these being buses, trolleybuses, or trams. The study was proposed and conducted in close collaboration with The Municipality of the City of Bratislava, the exclusive owner of public transport services in the city. The research will utilise data obtained from the Public Transport Company of Bratislava (DPB) and weather data sourced from the Slovak Hydrometeorological Institute (SHMU). The study aims to validate existing hypotheses regarding the correlation between Bratislava's public transport ridership and various weather patterns. The analysis will focus on the year 2018 due to data unavailability for specific months in 2019, the substantial impact of the COVID-19 pandemic on 2020 and 2021, and the lingering effects in 2022. Despite the temporal distance, municipality representatives express optimism, foreseeing the utility of applying the same analytical framework to study 2018 and 2023, enabling a comprehensive five-year comparison. Furthermore, the study aims to bring a comprehensive examination of the entire public transportation network, therefore, a system-wide analysis approach will be employed.

The main motivation behind conducting the study on the Impact of Weather Conditions on Public Transport Ridership: A Study of Bratislava, Slovakia is that the Municipality of the City of Bratislava, nor the public transportation company of Bratislava (DPB) conducted such an analysis of weather impact on the public transport ridership, so the number of passengers using a particular form of public transport, in the past. Consequently, the absence of prior analyses on the influence of weather conditions on public transport ridership in Bratislava, coupled with the lack of comprehensive assessments by both the city of Bratislava and the public transportation entity DPB, underscores a significant research gap.

This study aims to address this gap, offering valuable insights that can contribute to enhanced public transport planning. Addressing this gap in Bratislava would equip the local public transport authority with essential insights for making better-informed decisions. This includes potential adjustments to the frequency of various public transport modes, such as buses, trolleybuses, or trams, based on the findings of the study, aimed at enhancing passenger comfort. Furthermore, the outcomes could guide informed decisions regarding the optimal specifications of dispatched vehicles, encompassing considerations of size and capacity, thereby facilitating operational cost efficiencies and concurrent improvements in profitability.

The study aims to unravel the interplay between hourly weather changes and ridership across different modes of public transportation in Bratislava, specifically focusing on buses, trolleybuses, and trams. The study will employ time-series modelling methods to investigate the weather-transit relationship at an hourly temporal scale. This study offers an empirical attempt by addressing three key questions:

(1) To what extent do hourly changes in weather conditions affect bus ridership, and how does this effect vary over weekdays and weekends?
(2) To what extent do hourly changes in weather conditions affect trolleybus ridership, and how does this effect vary over weekdays and weekends?
(3) To what extent do hourly changes in weather conditions affect tram ridership, and how does this effect vary over weekdays and weekends?

This comprehensive exploration seeks to contribute valuable insights into the differential responses of various public transport modes to hourly weather dynamics, enabling a more granular understanding that can inform strategic decision-making for optimising service provisions and enhancing the overall commuter experience in Bratislava.

The rest of the paper is structured as follows: The next section will present the study context and data sources. Then, in section 3, already existing literature will be explored to gain more understanding of similar studies conducted in other cities around the world. Section 4 will explain the methodological approach of conducting this study as well as the methodology for modelling the weather-transit ridership. In section 5, the results will be presented in three subsections – the first one will focus on the exploratory data analysis, the second subsection on the results of the descriptive analysis and the third one will be dedicated to the predictive analysis results. Lastly, section 6 will discuss the findings, managerial implications and propose directions for future research before making concluding remarks.

2. Study Context and Data

The section on the study context and data will explore Bratislava's relevant statistics and its public transport network more in detail. First, it will highlight the differences between tram, bus, and trolleybus systems and delve into metrics such as the number of vehicles, kilometres driven, passenger counts, and revenue generated by each mode of transportation in 2018 based on publicly available data. Furthermore, this section will also present the two data sources and reveal features that will be used for further analysis purposes, including auxiliary variables. Additionally, it will provide insights into Bratislava's climate in 2018 based on the data provided by the SHMU.

2.1 Study Context

The public transport network in Bratislava, Slovakia is the study context. Bratislava is the capital of Slovakia and the administrative, economic, political, educational and cultural centre of the state as well as the Bratislava region (*Bratislava - Characteristic of the Region,* 2024). At the beginning of 2018, the city had 429,564 inhabitants with a share of 7.9% of Slovakia's total (*Slovak Republic in Figures*, 2018). The population of the city grew over time reaching 476,922 inhabitants as of December 31st, 2022, presenting 11.02% growth of the city's population (*Slovak Republic in Figures*, 2023).

In Bratislava, buses, trams, and trolleybuses serve as the primary modes of public transportation. Buses are conventional motor vehicles powered by internal combustion engines, operating on roads and able to navigate flexible routes. Trolleybuses, while electrically powered like traditional buses, are constrained by overhead wires for power, limiting their route flexibility compared to conventional buses. However, they still offer a cleaner and quieter alternative to diesel buses, contributing to reduced emissions and noise pollution in urban areas. Trams, on the other hand, are electrically powered rail vehicles running on dedicated tracks, making them the least influenced by traffic congestion, and providing a reliable and efficient mode of transportation.

As Table 1 indicates, at the end of the year 2018, the city transport owned altogether 898 motor vehicles, out of which 60.13% were buses that drove the most kilometres presenting 66.21%

out of a total of 43,150 km driven by the public transport motor vehicles in that year. In regard to seat kilometres, so kilometres made per seat in a vehicle, we observe a similar trend with buses presenting 58%, followed by trams and trolleybuses with 29.18% and 12.93% respectively. These numbers almost copy the number of transported passengers which was 247,114 in total. 58.04% of passengers travelled by bus, followed by 29.19% taking tram and 12.77% opting for trolleybus. Lastly, the total revenue from city transportation was 44,245 € in 2018. It can be observed that the split between different vehicle types perfectly copies the seat kilometres and transported passenger's metrics with buses bringing 58.08% of revenue, trams 29.16% and trolleybuses 12.76% (*Slovak Republic in Figures*, 2023).

Indicator	**Num. of vehicles**	**Driven kilometres**	**Seat kilometres**	**Transported passengers**	**Revenue in €**
***Total** out of which*	898	43,150	5,114,204	247,114	44,245
Bus	540	28,553	2,970,205	143,456	25,697
Trolleybus	162	5,870	661,653	31,527	5,637
Tram	196	8,727	1,492,346	72,130	12,911

Table 1: City transport of Bratislava in 2018 in numbers (Slovak Republic in Figures, 2018)

With a city area of 367.7 km^2, there is 1,824 km of transportation network per 1 km^2 of the area. The public transportation system has a bus stop density of 3.7 stops per 1 km^2, representing an average distance of 268 meters between the stops in a straight line. From the perspective of covering the city area with the public transportation network, the situation can be considered satisfactory (*Concept for the Development of Public Urban Transport in Bratislava for the Years 2013-2025,* 2016). In Figure A (see Appendix), the scheme of public transport is presented, with red lines accounting for trams, green for trolleybuses and blue for buses.

Nonetheless, passenger cars are the primary trip-making mode. In 2018, the total number of registered motor vehicles in the city was 396,451 out of which 77.91% were passenger cars (*Slovak Republic in Figures*, 2023). According to the TomTom Traffic Index 2018, Bratislava's congestion level was at 33%, ranking it as the 69th city among the 403 studied cities with the highest congestion levels (*Traffic Index 2018*). In the Concept of Development of Urban Public

Transport in Bratislava for the years 2013-2025 (2016), the public transportation provider acknowledges the need to take deliberate measures to increase the attractiveness of public transportation. The potential lies in individuals choosing public transport over personal vehicles more frequently, which would enhance overall traffic conditions in the city by reducing congestion and the substantial number of cars (*Concept for the Development of Public Urban Transport in Bratislava for the Years 2013-2025,* 2016). Given this, along with the apparent vulnerability of public transport passengers' travel experience to weather conditions such as accessing and waiting at stops during rain, Bratislava's network offers an interesting context for this study.

Bratislava has a continental climate, with cold winters and warm summers. The city experiences warm summers, typically between June and August, while winters range from December to February and are milder. The rest of the year exhibits moderate temperatures, providing a relatively consistent climate throughout the year. An in-depth examination spanning from 1951 to 2017 revealed that the warm spells exceeding normal temperatures have been steadily increasing, while occurrences of below-normal cold spells have been consistently diminishing (Výberči et al., 2018). Moreover, statistics showed that the city has been experiencing a 1°C temperature increase since 1988. Consequently, higher temperatures have led to an increase in evapotranspiration resulting in occasional but heavy rainfall (Lückerath et al., 2019). Despite this, the total annual precipitation has decreased by nearly 20%, with severe droughts plaguing the region almost every year since the 1990s (Faško et al., 2008).

2.2 Data Sources

To address the three research questions, transit data, and weather measurements are employed as the two principal data sources. Transit data was measured by the public transport provider of the city of Bratislava (DPB). The data covers a twelve-month period from January 1st to December 31st 2018 in the form of two incremental values: the number of boarded passengers and the number of disembarked passengers from each vehicle. From these two values, the number of passengers in each vehicle was calculated and then aggregated per hour. The number of boarded passengers is generated every time a person passes through the door entering a vehicle. Contrarily, the value of disembarked passengers changes every time a person exits the vehicle. These movements are recorded by movement sensors placed by the doors of all vehicles. Each public transport vehicle, whether it is a tram, bus or trolleybus has such a sensor for detecting passenger movement into and from the vehicle.

The data from the public transport provider was quite extensive, provided in twelve files with each file presenting data for one month of the year. Before data cleaning, the twelve files consisted of about 280 million rows altogether. The features that were used for data preparation and analysis include the date, time, vehicle direction, serial number, boarded passengers, disembarked passengers, door action, line, vehicle ID, stop name, and stop ID.

Weather data was acquired from the Slovak Hydrometeorological Institute (SHMU) for the same period as the transit data. It includes measurements of four variables, i.e., temperature, precipitation, humidity, and atmospheric pressure at one-hour intervals for two weather stations. However, as these weather stations are relatively close, only data from one station was used for analysis purposes, this being Weather station Bratislava-Koliba.

The weather conditions captured largely reflect the continental climate, characterised by notable fluctuations in both temperature and humidity throughout both the day and the year. The lowest measured temperature was -14.5 °C and the maximum temperature was 34.5 °C. The lowest humidity was 24% and the highest was 100%. The mean for temperature and humidity was 12.81°C and 69.51% respectively. The variable exhibiting the least variations is atmospheric pressure, which ranged from 959 hPa and 1005 hPa, with a mean of 982 hPa. Lastly, continental climate does not exhibit a lot of precipitation which could be also observed in our data with a mean of 0.45 mm. The highest precipitation level was 28.5 mm measured at the beginning of June.

For analysis purposes, the two principal data sources were merged into a final dataset of about 730,840 rows consisting of the following columns: date, time, line, passenger count, temperature, precipitation, humidity, and atmospheric pressure. Afterwards, auxiliary variables such as day of the week, vehicle type (tram, bus, or trolleybus) as well as binary variables indicating public holidays, school holidays, is weekdays, is peak time, temperature outlier, precipitation outlier, humidity outlier, atmospheric pressure outlier was added into the final dataset to extend the study and explore different patterns of passengers' behaviour.

3. Literature Review

Understanding the dynamics of public transport ridership in urban settings is crucial for efficient transportation planning and policy-making. This literature review section examines various studies that investigate factors influencing public transport usage patterns, ranging from weather conditions to different analytical approaches employed in understanding ridership behaviours.

3.1 Results from Cities Around the World

Numerous studies have investigated the impact of weather conditions, including rain, temperature, wind, humidity, and other factors on public transport ridership. Generally, extreme weather conditions such as very high or low temperatures, strong winds, and heavy precipitation tend to decrease the use of public transport (Stover & McCormack, 2012).

However, it is important to note that the effect of weather on ridership can vary depending on the geographical location. For instance, in Gipuzkoa - a city located in the northeast of Spain, higher temperatures were found to encourage public transport usage due to the pleasant nature of the climate and increased appeal of outdoor activities. Furthermore, multiple linear regression results showed that rain and wind decreased public transport ridership in Gipuzkoa (Arana et al., 2014) as well as in New York City (Singhal et al., 2014). On the other hand, in the Netherlands, it was found to increase ridership as it led to a shift from cycling and walking to public transport (Sabir, 2011). Some studies have explored the impact of other weather-related conditions such as humidity or apparent temperature on public transport ridership. A study on public transport usage in Shenzhen, China demonstrated that humidity negatively affects bus and subway ridership (Zhou et al., 2017). Furthermore, apparent temperature, which presents human-perceived temperature negatively impacted public transport ridership in Brisbane, Australia (Corcoran & Tao, 2017).

To conclude, it is evident that various geographical locations not only exhibit diverse climates but also distinct lifestyles and commuting behaviours. Additionally, the presence of cycling infrastructure and the overall traffic conditions within a city can significantly impact individuals' choices regarding the utilisation of public transportation.

3.2 Different Known Study Approaches

Exploring public transport ridership patterns has so far been approached through various approaches, for example, destination-based analysis, stop-level analysis, and system-level analysis. In a destination-based analysis, the focus is on understanding patterns related to specific destinations or endpoints within a transportation network. This approach involves examining how and why people use public transport to reach particular destinations, such as commercial districts, educational institutions, or residential areas. Factors such as the frequency of trips to specific destinations, the purpose of travel, and the demographic characteristics of commuters heading to these destinations are analysed by using the station-level approach (Tao et al., 2018). This analysis helps planners and policymakers tailor transportation services to meet the demands of specific destinations.

Stop-level analysis involves a more granular examination of ridership patterns at individual stops or stations within the public transportation system. Researchers scrutinise factors such as passenger boarding and alighting patterns, peak hours, and the popularity of specific stops (Tao et al., 2018; Zhou et al., 2017). This approach helps identify high-traffic locations and understand the factors influencing passenger behaviour at each stop. It can inform decisions related to optimising service frequency, improving infrastructure at bus stops, and enhancing the overall efficiency of the transit system.

System-level analysis takes a broader perspective, considering the entire public transportation network as a cohesive system. This approach involves assessing overall ridership trends, route performance, and network connectivity. System-level analysis is crucial for strategic planning, helping policymakers make informed decisions about route expansions, adjustments in service frequency, and improvements in overall network design to enhance the efficiency and attractiveness of the public transport system (Mishra et al., 2012).

A system-level analysis involves a comprehensive examination of the entire public transportation network, taking into account various temporal aspects and usage patterns. This approach looks at the bulk of public transport services, dividing them into different timeframes such as weekdays, weekends, peak hours, and off-peak hours. Researchers analyse overall ridership trends and variations during specific periods, identifying peak demand times and less busy intervals (Nissen et al., 2020; Tao et al., 2018; Zhou et al., 2017). This analysis helps transportation planners optimise service schedules, allocate resources efficiently, and address the specific needs of passengers during different times of the week or day. System-level analysis is instrumental in shaping policies and strategies that enhance the overall functionality and responsiveness of the public transport system. Since the system-wide analysis is perfectly aligned with this study's aims and objectives, it will be further employed to answer this study's research questions.

4. Methodology of the Study

This section outlines the methodology employed in the study, anchored by the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. CRISP-DM provides a systematic approach to the project, guiding the research through its six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Provost & Fawcett, 2013, pp. 26–34). The CRISP-DM framework has been integral throughout the study, ensuring a structured progression from the initial understanding of the business context to the application of findings. By applying the CRISP-DM framework, the study is not only methodologically sound but also aligned with the practical needs and constraints of the domain in question.

The business understanding and data understanding stages of the CRISP-DM process were covered in sections 1 and 2 respectively. In the following part, the approach employed during the more complex stages, these being data preparation, modelling and evaluation will be articulated. Deployment is not a part of this research study and, therefore, will not be covered.

4.1 Employed Technological Frameworks

A combination of different Python libraries to address the specific challenges encountered in this research was carefully selected. The powerful data manipulation capabilities of Pandas and the computational efficiency of NumPy were leveraged in handling and processing datasets of diverse scales and complexities.

The inclusion of XGBoost, a state-of-the-art gradient-boosting regression algorithm, helped to tackle predictive modelling tasks with precision and scalability. Complementing these machine learning components, model validation and hyperparameter optimisation were orchestrated, using scikit-learn's versatile utilities such as ‘train_test_split’ for dataset partitioning and GridSearchCV for parameter tuning. For models’ evaluation purposes, already established metrics such as mean squared error and R-squared were employed to quantify each model’s performance accurately.

For insightful visualisations, I turned to the flexible plotting capabilities of Matplotlib, tailoring plots with fine-grained control facilitated by Matplotlib's dates module, ensuring clarity and interpretability, especially in the context of time series data. Additionally, for Exploratory Data Analysis, I utilised the well-established visualisation tool Tableau.

To delve deeper into temporal patterns and dynamics, statistical models like SARIMAX from Statsmodels were integrated, augmenting traditional machine learning approaches with sophisticated time series analysis techniques. This thoughtfully assembled tech stack not only helped me achieve my research goals but also underscored the importance of adopting a comprehensive approach to addressing complex research questions.

4.2 Data Preparation

The data cleaning process for the dataset from the SHMU began by importing necessary libraries, with pandas facilitating data manipulation and datetime aiding in handling time-related information. As already explained in the Data Source section, unnecessary rows on the Bratislava airport weather station were filtered out, streamlining the dataset for relevant analysis. To better analyse and merge the data in the upcoming steps, the 'datum_cet' column was split into separate 'date' and 'time' columns, utilising pandas' datetime functionalities to extract these elements. Lastly, missing values in the key weather variables such as temperature, humidity, atmospheric pressure, and precipitation were handled through forward filling, where missing data was replaced with the most recent non-null data point from the same day. This method, known as forward imputation, preserves the temporal continuity of the dataset and ensures that missing data do not compromise the integrity of the analysis (Ribeiro & Castro, 2022).

The data sourced from the Bratislava transport provider presented a substantial dataset for analysis, comprising 12 files, each with approximately 24 million rows and 25 distinct attributes. The magnitude and complexity of this data necessitated an extensive data cleaning and preparation process. One notable task involved deriving passenger counts from two specific columns: 'embarked' and 'disembarked'. This step was crucial in accurately assessing the volume of passengers utilising the transport system. The methodology employed in cleaning and preparing this dataset included a series of carefully executed steps designed to ensure the data's accuracy, consistency, and usability for subsequent analyses.

To streamline the dataset for further analysis, a substantial part of the cleaning involved removing several columns considered unnecessary such as driver, planned departure, stop number or untitled. These attributes were found to either not contain reliable information or were not beneficial for the study's purpose. Following the removal of these columns, the attributes were renamed to their English equivalents to standardise the dataset and enhance readability. The cleaning process also involved deleting rows where no change in passenger numbers was observed, specifically where the door action was recorded as 0, indicating that the door remained closed. The 'door_action' column was subsequently dropped as it became redundant after this filtering. Further refinement of the dataset involved dropping rows based on specific conditions such as unknown vehicle direction, zero passenger counts indicating potential incorrect detections or system errors and stops associated with depots likely involving only driver movement.

To align the dataset with an hourly weather dataset, the 'time' column was rounded to the nearest hour, ensuring each time entry corresponds to the closest hour, with entries at or beyond the half-hour mark rounded up, and those before it rounded down. For example, if a record in the dataset has a timestamp of 05:31:12, it is rounded up to 06:00:00 because it is past the half-hour mark. Conversely, a timestamp of 05:29:59 is rounded down to 05:00:00, as it falls before the half-hour mark. This adjustment facilitates direct comparison with hourly weather data.

The dataset was grouped by date, time, line, serial number, and vehicle ID facilitating more structured analysis moving forward. Then, custom lambda functions were applied to create new columns indicating the start, end, and total passenger count for each journey, refining the dataset further for detailed analysis. The first function, 'find_first_record', identifies the initial passenger count that closely aligns with the journey's median, restricted by a less than 1000 passenger difference, aiming to represent a typical load while excluding outliers. The second function, 'find_last_record', mirrors this approach for the journey's end, selecting a passenger count near the median to avoid skewing data with end-of-journey drops, thus ensuring consistency in passenger load representation. Finally, 'get_total_pax' calculates the passenger change across the journey by subtracting the initial count from the final. This method helped in calculating the total number of passengers for each journey, critical for further analysis.

Furthermore, the data from 12 monthly CSV files was merged into a single DataFrame, creating a comprehensive dataset for the entire year. Then, the integration of DPB (transportation data) and SHMU (meteorological data) took place. It was conducted based on matching timestamps, ensuring that each transportation event was associated with corresponding weather conditions.

To facilitate further analysis, new binary attributes were introduced such as 'is_weekday' to indicate whether a record pertains to a weekday, and 'is_peak_time' to identify records falling within peak hours on weekdays, specifically between 7-9 AM and 5-7 PM. The peak hours were determined based on conventional working hours, assuming these periods witness heightened activity. The dataset was further refined by incorporating additional attributes such as ‘school_holidays’, ‘public_holidays’, ‘day_before_public_holidays’ and ‘day_of_week’, which could play a crucial role in understanding transportation trends. Then, a cleaning operation was performed to remove specific lines from the dataset. These lines, namely 408 and 410, were identified as special buses exclusively operating on New Year's Eve, deemed irrelevant for the analysis. Additionally, several other lines, including 602, 803, 831, 871, 872, 891, 895, and 957, were flagged as being solely used for operational purposes without passengers and were consequently deleted.

Following the cleaning process, the data was structured to facilitate further analysis. Specifically, an effort was made to categorise the transportation data into three distinct clusters representing different vehicle types: buses, trolleybuses, and trams. To achieve this, each line number was mapped to its corresponding vehicle type, and a new column labelled 'vehicle_type' was created in the dataset, integrating this mapping. This categorisation was further used to split the dataset into three files to build both SARIMAX and XGBoost models.

The above steps formed a comprehensive approach to cleaning and organising the dataset, ensuring it was primed for in-depth analysis and interpretation. These preparatory actions set the foundation for robust analysis, aiming to extract meaningful insights from the vast quantities of data.

4.3 Modelling and Evaluation

The methodology employed during the modelling stage aiming to obtain the study results discussed in detail in section 5 varies according to the objectives of each analysis conducted.

Further, the methodology followed in the Exploratory Data Analysis, the Time Series Analysis and last but not least the Predictive Analysis will be presented.

4.3.1 Exploratory Data Analysis

In the Exploratory Data Analysis phase, the analysis utilises the visual analytics platform Tableau as the primary tool. The versatility of Tableau as a tool not only enables the creation of dynamic visualisations but also supports a granular analysis that accommodates the complexities of the dataset and the specific research objectives. This multifaceted approach allows for an in-depth exploration of how different factors influence public transit usage, encompassing a range of analyses tailored to the specific objectives of the research. The Exploratory Data Analysis separately considers the patterns of ridership on weekdays and weekends, allowing for a detailed understanding of how transit behaviour varies between regular workdays and weekends.

The analysis begins by examining the average monthly ridership, identifying seasonal trends and peak periods of demand. This is followed by an investigation into average hourly ridership, revealing critical insights into peak and off-peak hours. Through this, a clear picture of passenger flow throughout the day emerges, highlighting when demand for public transit is highest. Further, the study delves into ridership based on vehicle type, including buses, trolleybuses, and trams. This aspect of the analysis examines the variations in ridership patterns across different modes of transportation, offering insights into passenger preferences and usage intensity for each vehicle type.

The analysis is also dedicated to understanding the impact of extreme weather conditions on transit usage. This includes exploring how very high or low temperatures, heavy precipitation, and extreme atmospheric pressure and humidity events affect ridership. Such an examination provides insights into behavioural shifts during heatwaves, cold spells, and periods of heavy rain or snow, as well as the potential impact of atmospheric pressure on passenger decisions. For the analysis concerning the behaviour of passengers under various extreme weather conditions, outliers were identified using the 95th percentile method. The 95th percentile method is a statistical technique used to identify outliers in a dataset by setting thresholds. This approach involves arranging the data points in ascending order and identifying the value below which 95% of the data points fall. The 95th percentile value is considered a threshold, above which data points are treated as outliers (Aguinis et al., 2013). This method is particularly useful in analyses where the focus is on understanding the extremes of the data distribution, such as extreme weather conditions affecting passenger behaviour. By excluding the top 5% of the data points, which are considered outliers, the analysis can focus on the typical patterns within the majority of the data, providing insights into the standard passenger behaviour under most weather conditions. The use of the 95th percentile method allows for a robust analysis by minimising the influence of extreme values, which might potentially skew the interpretation of the data.

4.3.2 Time Series Analysis

For this purpose, the Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX) model was employed. SARIMAX is an extension of the ARIMA (Autoregressive Integrated Moving Average) model that incorporates both seasonal components and exogenous variables (Karim et al., 2023). This model is particularly well-suited for analysing time series data where seasonal effects are pronounced, and where external factors (in this case, weather conditions) are believed to influence the time series (Tao et al., 2018).

To accommodate SARIMAX modelling, which cannot handle null values and requires continuous data, hours during which trams and trolleybuses were not operational were excluded from the dataset. Trams were found not to run between 1 AM and 4 AM, and thus data corresponding to these hours were filtered out. Similarly, trolleybuses were not operating between 1 AM and 3 AM, and data for these hours was also excluded. The filtering process did not apply to buses, given their continuous operation throughout the night as night buses, thereby obviating the need for any filtering at the bus level. This preprocessing step ensures that the dataset contains only valid, continuous data, aligning with the requirements of the SARIMAX model and reflecting the operational schedules of trams and trolleybuses, allowing for more accurate modelling of public transportation trends.

Achieving stationarity within the time series data is another cornerstone of the SARIMAX model, accomplished through differencing. This process, part of the model's integration component, ensures that the time series being analysed does not exhibit changing statistical properties over time, which is crucial for the reliability of the model's outputs. Moreover, the

SARIMAX model delves into modelling dependencies within the data. It does so by leveraging its autoregressive and moving average components to capture the dependencies that arise both from the series' previous values and from the errors in previous forecasts (Karim et al., 2023). This dual approach enables the model to account for the complex interrelations within the time series data, offering a detailed understanding of how past values and errors influence future ridership numbers.

This comprehensive approach involved constructing a SARIMAX model for each specific weather variable, separately computed for each type of public transportation and differentiated between weekdays and weekends. This detailed breakdown allowed me to dissect and understand the unique temporal patterns and influential factors that affect the ridership of each mode of transportation distinctly. This nuanced analysis was refined by baseline models, which incorporate dummy variables to account for public and school holidays (Tao et al., 2018). Altogether, this extensive modelling effort entails the creation and examination of 30 SARIMAX models, each shedding light on the subtle forces that shape public transit usage.

In my analysis, several key metrics across the three modes of transport were evaluated to assess the efficacy of the SARIMAX models. The coefficients were considered to understand the magnitude and direction of the impact that each variable has on ridership. The p-values were crucial in determining the statistical significance of these coefficients, with values below 0.05 indicating a strong likelihood that the observed effect is not due to weather change (Thiese et al., 2016). To compare the models' fit, the Akaike Information Criterion (AIC) was used, where lower values suggest a more suitable model fit given the number of parameters (Wagenmakers & Farrell, 2004). For accuracy, the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) were calculated, providing insights into the average prediction errors in squared and absolute terms, respectively; lower values here reflect more precise predictions. Lastly, the normalised MAE allowed for comparing the performance of the models across different scales of data by providing a proportionate error metric (Müller-Plath & Lüdecke, 2024).

4.3.3 Predictive Analysis

For the purpose of predictive analysis, XGBoost was chosen to overcome the shortcomings of SARIMAX models in terms of prediction accuracy and to harness its capability to manage the intricacies present in the urban mobility data of Bratislava. XGBoost stands for eXtreme

Gradient Boosting and is an advanced implementation of gradient boosting that constructs a predictive model through an ensemble of weak prediction models, predominantly decision trees. It is well known for its efficiency, scalability, and performance in both machine learning and applied scenarios (Tarwidi et al., 2023).

In the essence of its operation, XGBoost builds upon the gradient boosting framework, where it sequentially introduces new models that aim to correct the residuals or errors of preceding models. These models are then combined to formulate the final predictive output (Asselman et al., 2021). The process is characterised by its stage-wise additive nature, with each new model incrementally reducing the loss function, a quantification of the discrepancy between the model's predictions and the actual outcomes. XGBoost employs decision trees as its foundational learners (Elavarasan & Vincent, 2020). These trees segment the data into branches based on derived decision rules from the input features, culminating in leaves that issue predictions. The collective predictions from all trees are aggregated to deliver the ultimate forecast (Ali et al., 2023).

A pivotal aspect of XGBoost is its utilisation of gradient optimisation. Each iteration sees the introduction of a novel tree designed to amend the inaccuracies of existing trees, employing the gradient descent algorithm to minimise the loss by adjusting trees (Sahin, 2020). This process employs a distributed weighted quantile to determine the optimum number of breakpoints between weighted datasets (Ali et al., 2023). This allows the algorithm to find more effective split points that minimise the loss function, thereby improving the model’s performance. Another distinctive feature of XGBoost is the incorporation of a regularisation term within the loss function, aimed at controlling model complexity (Singh, & Rawat, 2023). This regularisation aids in averting overfitting, thereby enhancing the model's robustness and efficacy on unseen data by favouring simpler, more general models over complex ones (XGBoost, 2022a).

Designed for high efficiency and scalability, XGBoost leverages both hardware optimisations and algorithmic improvements, such as effective tree pruning, to expedite the learning process and adeptly manage large datasets. The model also offers an array of hyperparameters governing the learning process, including tree depth, learning rate, and the number of trees, whose careful adjustment can significantly influence model outcomes (Tarwidi et al., 2023).

The data was thoroughly segmented, training separate models for weekdays and weekends to accurately capture the distinct travel behaviours characteristic of these periods. In total, six XGBoost models were constructed, one for each type of public transport vehicle. The feature selection was critical to the success of these models. In the process of feature selection, a comprehensive set of variables was included: temperature, humidity, precipitation (including binary indicators for the occurrence of precipitation), atmospheric pressure, time of day, whether the day was a weekday or not, and calendar attributes such as year, month, day, and the specific day of the week. This broad spectrum of features was deliberately chosen to encapsulate the dynamic nature of ridership, which is influenced by a confluence of meteorological, temporal, and calendar-specific factors. The significance of the feature selection process cannot be overstated. By identifying and including the most relevant predictors, the models' learning was focused on the most impactful variables, enhancing the accuracy and interpretability of the results (XGBoost, 2022b). This step also helped in avoiding the noise that irrelevant features might introduce, streamlining the models to reflect the true influencers of ridership patterns.

As shown in Figure 1 depicting weekend ridership, the feature importance analysis highlighted temperature, atmospheric pressure (not the binary one), time of day, humidity and day as the most significant predictors. As a result, these features were found to hold the greatest sway over the models' performance. The prominence of these variables underscores the sensitivity of weekend transit use to environmental conditions and the specific hours of operation, which are likely tied to the recreational and flexible nature of weekend travel.

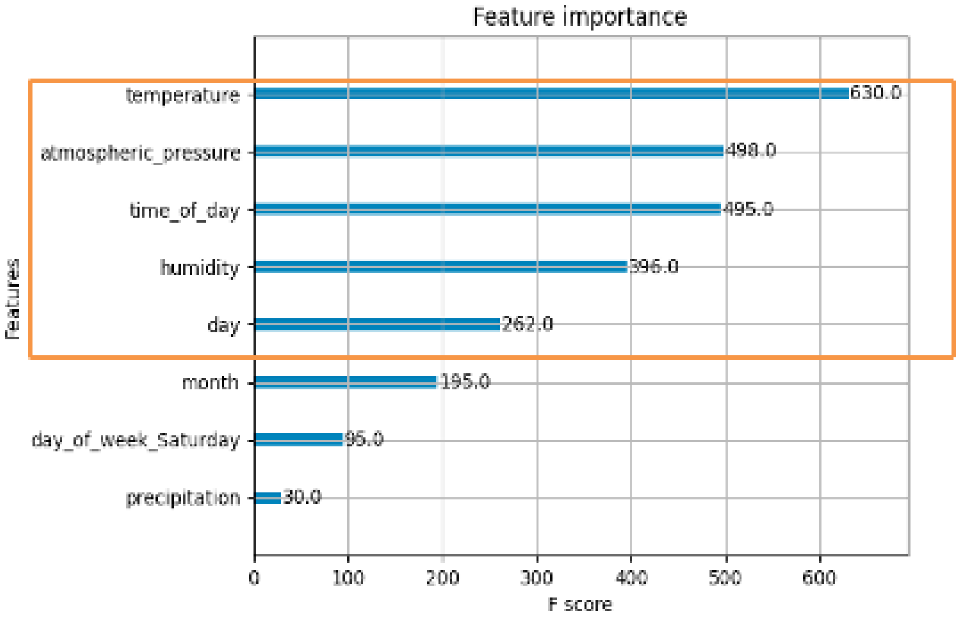


Figure 1: Feature importance for weekend XGBoost models

Interestingly when the 'month' variable was included in the models, there was a noticeable degradation in performance. This suggests that the inclusion of 'month' introduced noise rather than providing clarity, potentially due to overlapping effects with other features or a less direct influence on ridership. Consequently, 'month' was excluded from the final models to ensure a more precise and focused analysis, allowing the models to better capture the true patterns in the data without unnecessary complexity.

In the weekday predictive analysis using XGBoost shown in Figure 2, the feature importance evaluation highlighted six features as the most impactful: time of day, temperature, atmospheric pressure, day, humidity, and month. Each of these top six features played a crucial role in the models' ability to accurately assess ridership patterns. Time of day was the most influential, likely reflecting peak commuting hours, followed by temperature and atmospheric pressure, which may correlate with comfort and travel conditions.

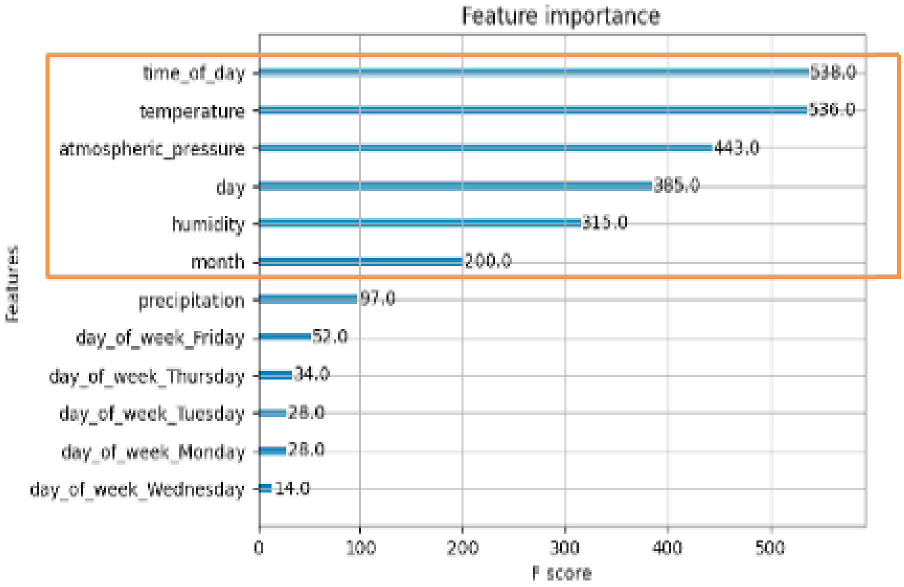


Figure 2: Feature importance for weekday XGBoost models

The inclusion of the day as a variable suggests that daily ridership could vary significantly throughout the week, perhaps due to varying work or school schedules. Humidity was also a key feature, potentially influencing the comfort level of riders or their willingness to use public transport. The month feature provided seasonality insights, capturing broader trends throughout the year.

Experimentation with the models revealed that removing any of these six features resulted in poorer performance, underscoring their collective importance in the model. Conversely, when the feature 'precipitation' was included, which ranked seventh in importance, there was a noticeable decline in the model's predictive accuracy. This observation led to the decision to exclude precipitation from the final model, thereby streamlining the feature set to those most determinative of weekday ridership behaviour. The prioritisation of these features was key to enhancing the model's interpretability and ensuring the reliability of the ridership predictions.

Then, I partitioned the dataset into a 60:20:20 ratio, dedicating 60% to training the models, 20% to the validation set for fine-tuning hyperparameters, and the remaining 20% for testing to evaluate the models on unseen data. This partitioning was a strategy to mitigate the risks of overfitting and underfitting, thus ensuring that the models maintain generalisability when applied to new, unseen data.

During the validation phase, I rigorously engaged in hyperparameter tuning, using the 20% of data reserved for validation to optimise the settings and enhance model performance without biasing the model to the training data. This step was critical as it allowed the adjustment of model parameters to improve prediction without affecting the test set, which remained an independent check on model performance.

The model evaluation metrics I employed included the Root Mean Square Error (RMSE), R-squared, and Mean Absolute Percentage Error (MAPE). RMSE measures the average magnitude of the errors between predicted and observed values, indicating model accuracy; the lower the RMSE, the more accurate the model's predictions. R-squared is a statistical measure that represents the proportion of variance for the dependent variable that's explained by the independent variables in the model, essentially indicating the model's explanatory power; a higher R-squared value suggests a better fit of the model to the data. MAPE expresses average absolute error in percentage, providing insights into the accuracy of the model in terms of the percentage error in predictions; a lower MAPE value indicates higher accuracy (Singh, & Rawat, 2023).

Additionally, the forecast results obtained from the 6 XGBoost models were visualised through two distinct approaches to facilitate a comprehensive evaluation of the models' performance. The first visualisation approach employed a scatter plot, which was constructed using a shuffled dataset. This scatter plot was designed with the actual passenger count on the x-axis and the predicted passenger count on the y-axis, utilising the testing dataset. This method allowed for a direct, visual comparison of the model's predictions against actual ridership figures, highlighting the accuracy and precision of the models in forecasting passenger counts. The second visualisation strategy took a different approach by focusing on the validation dataset to examine the evolution of actual versus predicted passenger counts over a specific timeframe, using a line chart. To achieve this, the dates within the validation set were thoroughly identified and subsequently, plotted using an unshuffled dataset. This enables the reader with a sequential, hour-by-hour comparison of actual and predicted counts. This method provides insights into the models' forecasting performance and offers a detailed view of how well each model’s predictions aligned with actual ridership trends over time.

5. Results

In this section, the modelling results of the impact of weather on system-level ridership will be presented. This section will be divided into three parts, these being exploratory data analysis, time series analysis and predictive analysis.

5.1 Exploratory Data Analysis

In this section, comprehensive insights derived from analysing ridership patterns across various temporal and weather-related dimensions will be explored. First, average monthly ridership trends will be analysed, followed by a detailed exploration of hourly ridership patterns. Subsequently, ridership dynamics based on vehicle type will be investigated, shedding light on commuter preferences and utilisation patterns. Then, the impact of extreme weather conditions on ridership will be explored. Finally, there will be an assessment of the Variance Inflation Factor considering the presence of collinearity between passenger count and weather variables, offering insights into how these factors interact and influence each other.

5.1.1 Average Monthly Ridership

To commence, Figure 3 illustrates the comprehensive monthly ridership patterns across the entire year, followed by Figures 4 and 5, providing a detailed view of public transport usage on weekdays and weekends, respectively. As can be observed in Figure 3, the overall passenger count exhibited a gradual decline from May to September. The most significant percentage decrease occurred from June to July, amounting to 1.47%. September recorded the lowest average passenger count at 6.59%, which could be influenced by the transitional period from summer to fall. During this time, individuals may experience changes in routines, such as the conclusion of vacation periods and the beginning of the school year, potentially impacting public transport demand. When detailly examining a decrease in usage from August to September (please see Figure B in Appendix), it can be observed that the usage of buses and trams experiences a percentual decrease by 15.44% and 51.52% respectively, but the usage of trolleybuses follows an opposite trend and increases by 46.15%.

Conversely, October had the highest average at 9.52%, suggesting a resurgence in public transport use. This increase could be linked to factors such as cooler weather, the resumption of regular work and school or university schedules, or specific events or activities that attract higher ridership during the fall season.

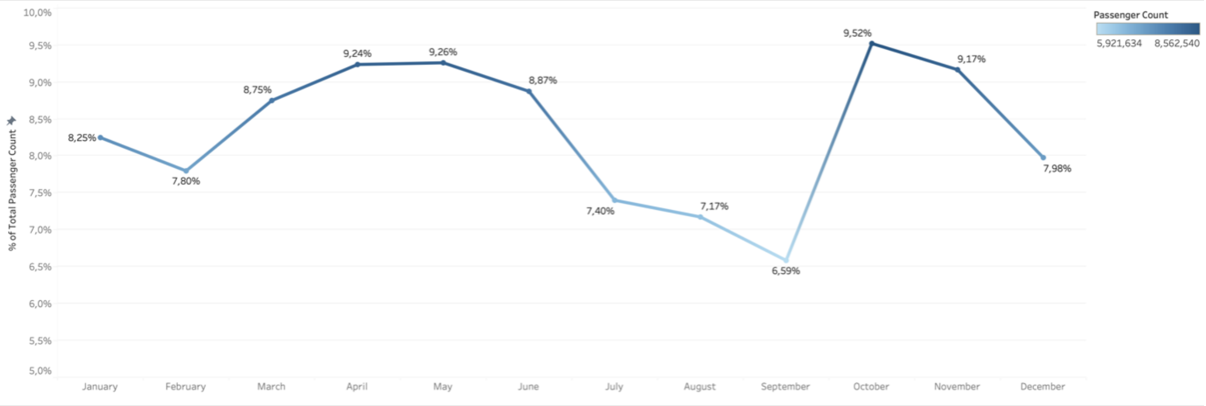


Figure 3: Overall average monthly ridership patterns in percentage

Upon excluding public holidays from the graph, no significant difference was observed (see Figure C in the Appendix). This trend persisted across various analyses, leading to the consideration of public holidays as regular days.

Weekday usage of public transport can be observed in Figure 4. It can be seen that its usage peaked in October at 9.79% with trams being used the most (38.18% out of total usage in September), closely followed by May at 9.41%. Conversely, the least utilisation occurred in September (6.36%), with August (7.20%) and July (7.27%) also experiencing comparatively lower ridership. Please refer to Figure D to see the monthly usage on weekdays broken down by vehicle type.

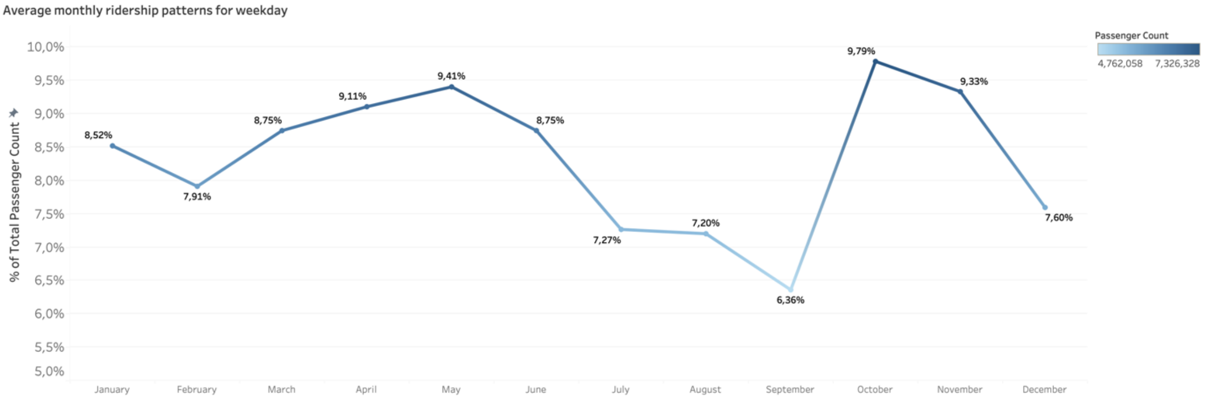


Figure 4: Average monthly ridership patterns on weekdays in percentage

This variance in weekday ridership could be influenced by factors such as seasonal changes, school and work schedules, and weather conditions. For instance, October and May might see increased public transport usage due to favourable weather and regular work or school routines, while the decline in September usage can be attributed to transitional periods marked by shifts from one routine or schedule to another. September is recognised as a transitional month between seasons, signifying the conclusion of the vacation period and the commencement of a new school year for most scholars.

Figure 5 presents usage during weekends when public transport experienced its highest utilisation in April and December, registering 9.88% and 9.87%, respectively. This elevated usage may be influenced by several factors. In April, milder weather conditions could encourage individuals to engage in outdoor activities or events, leading to increased reliance on public transportation. Additionally, December often coincides with the holiday season, and more frequent social and festive activities may contribute to a surge in public transport use as people travel for gatherings, shopping, and celebrations.

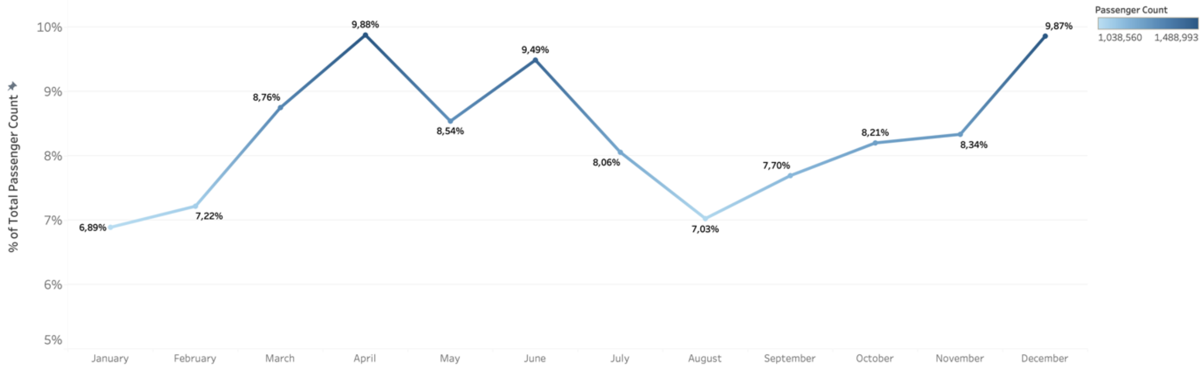


Figure 5: Average monthly ridership patterns on weekends in percentage

Conversely, the lowest usage was observed in January (6.89%), August (7.03%) and February (7.22%). August may witness decreased public transport utilisation due to vacation periods, as individuals tend to spend longer periods outside of the city. January and February, being a winter month, might experience lower public transport demand due to potential weather-related challenges and a general tendency for reduced outdoor activities during colder periods. Please see Figure E in the Appendix for a breakdown of the monthly usage of public transport on weekends based on vehicle type.

5.1.2 Average Hourly Ridership

The following examination of average hourly ridership offers valuable insights into the consistent and variable demand for public transportation services. This analysis aids in identifying peak periods, distinguishing between weekdays and weekends, and understanding the specific demands for different vehicle types throughout the day. In general, public transport experiences the highest amount of passengers at 4 PM (7.96%), followed by 5 PM (7.87%) and 8 AM (7.55%). For a more detailed examination, please refer to Figure F in the Appendix. When public holidays were excluded from the analysis, there was no visible change (Figure G).

As shown in Figure 6, there is a substantial variation in passenger behaviour between weekdays and weekends, particularly in relation to different times of the day. This difference can be attributed to the distinct routines and activities that individuals engage in during weekdays, such as commuting to work or school, compared to the more leisure-oriented patterns observed on weekends. On weekdays, public transport experiences the highest amount of passengers at 4 PM (7.96%), followed by 5 PM (7.87%) and 8 AM (7.55%). The least busy times during the usual work and school days are between 10 AM and 12 PM. On the other hand, weekends showcase a distinct ridership pattern, where the number of passengers gradually increases with each passing hour, reaching its peak at 6 PM (8.01%), followed by a continuous decline towards the evening hours.

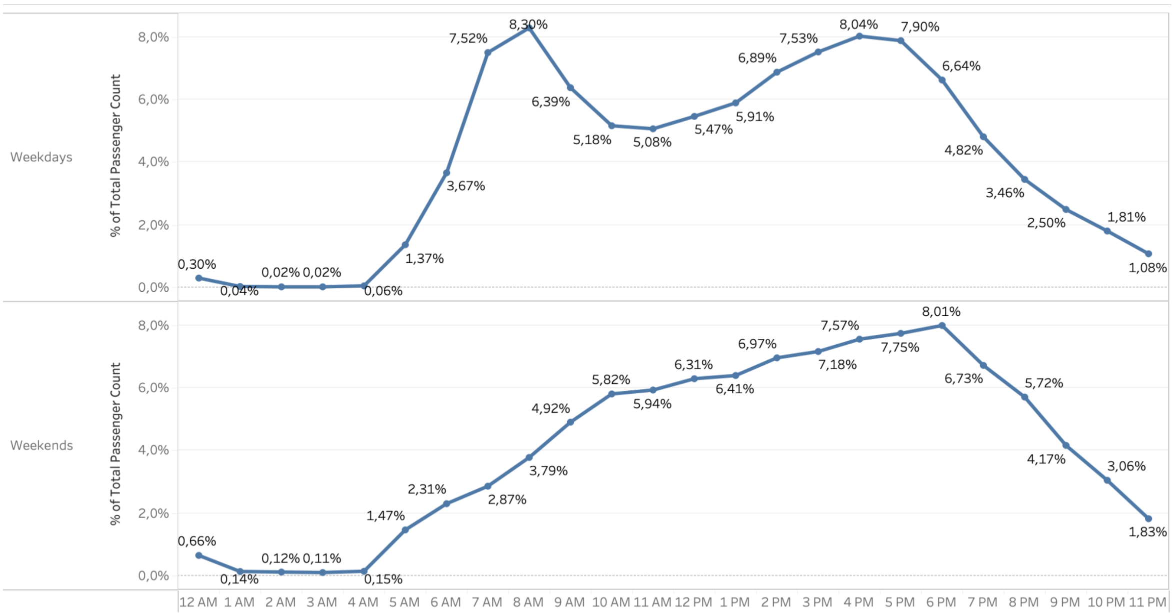


Figure 6: Average hourly ridership patterns on weekdays and weekends

Figure 7 also presents a comparison between the average hourly ridership on weekdays and weekends but additionally broken down by vehicle type. The percentage values represent the proportion of usage for specific modes of public transport out of the total public transport usage throughout the year. This implies that the cumulative sum of all these values equals to 100%.

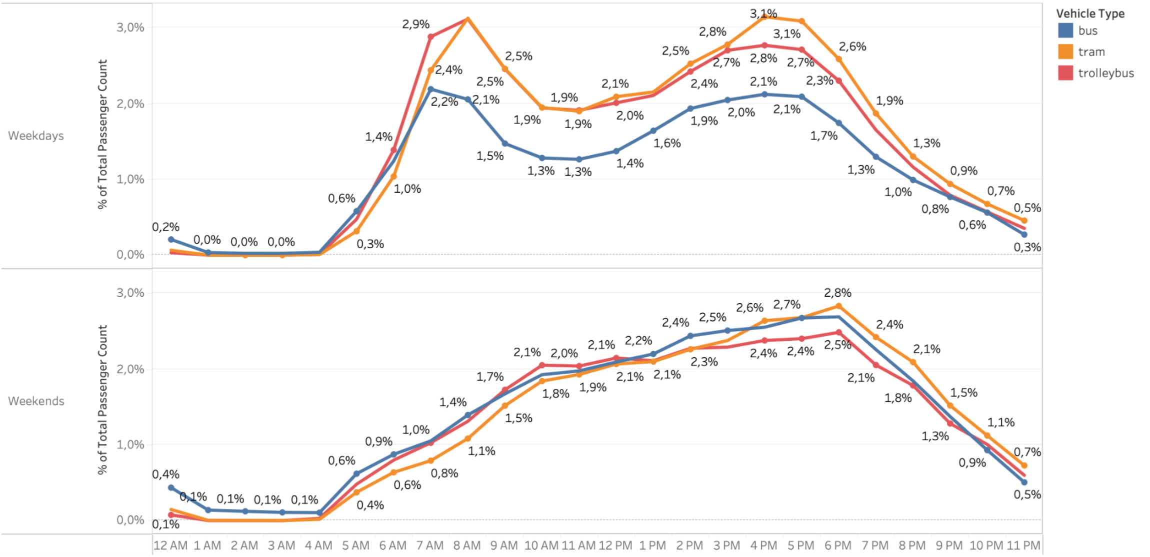


Figure 7: Average hourly ridership patterns by vehicle type on weekdays and weekends

Interestingly, the percentage usage of all three vehicle types is almost the same on weekends, suggesting a more balanced distribution of public transport modes during leisure days. On the other hand, on weekdays, the lower usage of buses between 7 AM and 6 PM could be attributed to the prevalence of alternative transportation means during typical commuting hours, such as personal vehicles or other modes of transit better suited for daily work and school routines.

Lastly, the general trend of the overall average hourly ridership throughout the year closely aligns with the pattern observed in the upper graph depicting average hourly ridership on weekdays. For a more detailed examination, please refer to Figure H in the Appendix.

5.1.3 Ridership Based on Vehicle Type

In this section, a comprehensive analysis of ridership patterns based on distinct vehicle types - namely buses (depicted in blue), trolleybuses (represented in red), and trams (highlighted in orange) will be presented. It aims to explore the usage dynamics of each vehicle type and uncover insights into commuter preferences, peak and off-peak hours, the difference between passengers' behaviour during weekdays and weekends, behaviour a day before and during public holidays, and the overall impact of these modes on the public transportation system. Through detailed examination and visualisation of ridership trends, the distinctive roles and contributions of buses, trolleybuses, and trams within the urban transit landscape will be explored.

Overall, trams exhibit the highest utilisation at 36.3%, indicating a notable preference for this mode of transportation. Trolleybuses closely follow with a usage rate of 35.2%, showcasing their significant role in the overall ridership. Buses, while still substantial, hold a slightly lower share at 28.5% (see Figure I in the Appendix). This distribution underscores the diverse preferences and efficiencies associated with each vehicle type within the public transportation system. After excluding public holidays from this analysis, no significant changes were revealed. For a more detailed examination, please refer to Figure J in the Appendix.

In Figure 8, passenger behaviour variation between weekdays and weekends can be observed. On weekdays, the preference for trams remains high at 36.9%, closely followed by trolleybuses at 35.8%, and buses at 27.3%. However, on weekends, the preferences are more balanced among all three modes of transport, with trams at 33.2%, trolleybuses at 32.3%, and buses at 34.5%. This distinction highlights a shift in commuter choices based on the day of the week, providing valuable insights into the dynamics of public transportation usage.

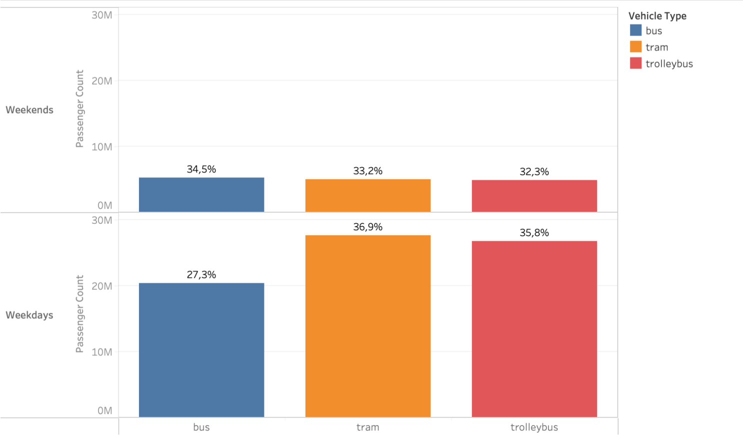


Figure 8: Usage of different modes of transport on weekdays and weekends

The weekday preferences for trams, trolleybuses, and buses reflect commuter choices during typical workdays. The sustained high preference for trams suggests their efficiency or popularity for daily commuting needs. Trolleybuses closely follow, indicating their significant role in the weekday transportation mix, while buses, though still substantial, exhibit a slightly lower preference, possibly influenced by factors such as route coverage or scheduling.

Upon examining the disparity between peak and off-peak hours in Figure 9, it becomes evident that buses are slightly less favoured during peak hours compared to off-peak hours. In contrast, the usage of trams and trolleybuses sees an increase of 2.79% and 4.31%, respectively, during peak hours. This pattern may be influenced by various factors, such as commuters opting for more efficient or faster modes of transportation like trams and trolleybuses during busy periods, while buses may experience a marginal decrease in preference due to factors like increased congestion or longer travel times.

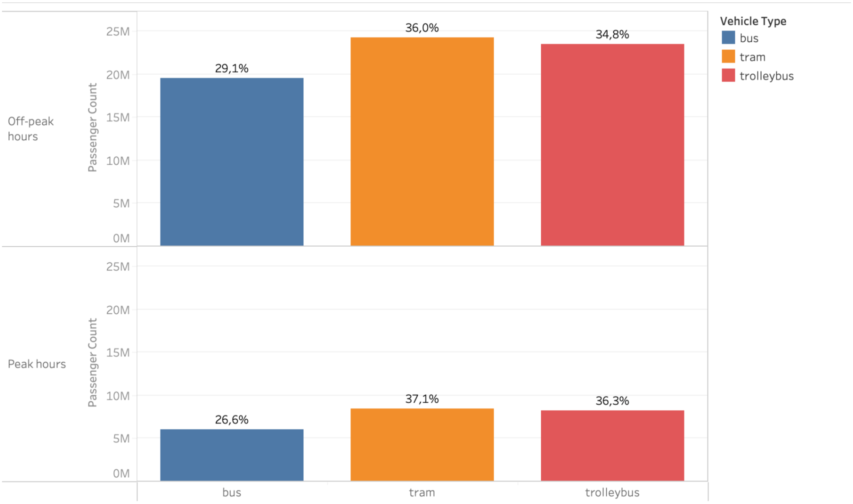


Figure 9: Usage of different modes of transport during off-peak and peak hours

Examining passengers' behaviour on public holidays reveals a pattern closely mirroring weekend trends, where buses are the most utilised at 34.2%, followed by trolleybuses at 33.1% and trams at 32.7% (refer to Figure K in the Appendix). Additionally, analysing passenger behaviour on the day before public holidays (Figure L) indicates that buses are the least favoured, resembling weekday trends, with a usage rate approximately 6.4% lower than trams and 5.7% lower than trolleybuses. This distinction may be influenced by altered travel patterns and reduced demand for buses during the transitional period leading to public holidays.

5.1.4 Ridership in Extreme Weather Conditions

In this section, the impact of individual weather variables - specifically, temperature, precipitation, and atmospheric pressure - on the utilisation of various transportation modes by comparing patterns on weekdays and weekends will be explored. Interestingly, humidity did not present a similar pattern of extreme values. The data within the 'humidity' column did not exceed the threshold established by the 95th percentile, suggesting that humidity levels are relatively stable within the dataset and do not exhibit extreme variances that could significantly influence transit usage. As a result, humidity will not be under examination, as no outliers were detected in this variable.

5.1.4.1 Ridership in Extreme Temperature Conditions

For temperature, a total of 140,078 data points were identified as outliers, constituting approximately 4.98% of the dataset. This indicates that nearly 5% of temperature records were significantly higher or lower than the norm, which could influence passenger behaviour as people may seek alternative means of transportation or change their travel habits during temperature extremes. In Figure 10, the impact of extreme temperatures on weekday ridership is evident. During weekdays, the extreme temperature results in a slight increase in bus usage by 1% and tram usage by 2.2%. However, trolleybuses experience a decrease of 3.2% in usage during extreme temperature conditions. This variation in ridership may be influenced by factors such as the efficiency or comfort of each mode of transport during extreme temperatures, as well as individual preferences or alternative transportation options.

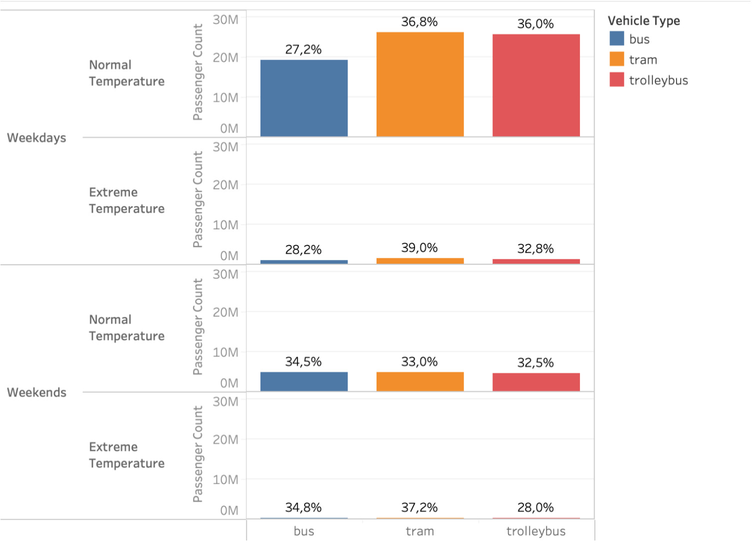


Figure 10: Impact of extreme temperature conditions on modes of transport

On weekends, the influence of extreme temperatures is less pronounced for buses, which are used similarly regardless of the temperature. Trams, on the other hand, see a notable increase in usage by 4.2% during extreme temperature conditions. Conversely, trolleybuses experience a decrease in usage by 4.5% on weekends when faced with extreme temperatures. These patterns suggest that, during weekends, passengers may adapt their choices based on temperature conditions, with trams being favoured in extreme heat, while trolleybuses are slightly less utilised.

5.1.4.2 Ridership in Extreme Precipitation Conditions

In the case of precipitation, 131,081 data points were marked as outliers, amounting to 4.66% of the observations. Such findings highlight instances of significant rainfall or snowfall that could lead to disruptions in transit services or deter passengers from using public transit due to inclement weather conditions. However, upon examining the difference in passenger behaviour when opting for a different mode of transport during extreme precipitation on weekdays and weekends, minimal changes are observed. The most significant percentage difference is noted in the usage of buses on weekends during extreme precipitation, with a modest decrease of 1%. These subtle variations observed in Figure 11 suggest that, overall, commuters maintain relatively consistent preferences for modes of transport regardless of the extreme precipitation events.

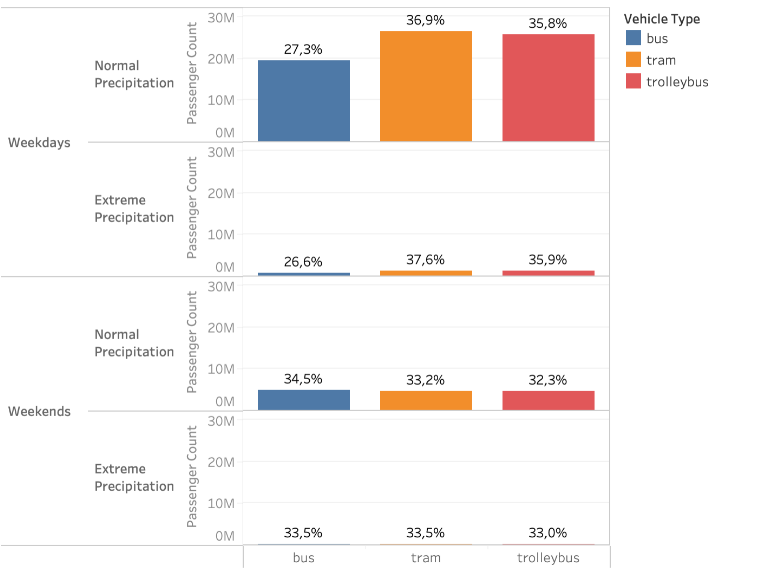


Figure 11: Impact of extreme precipitation conditions on modes of transport

5.1.4.3 Ridership in Extreme Atmospheric Pressure Conditions

For atmospheric pressure, 136,528 data points fell into the outlier category, which is 4.85% of the measurements. Variations in atmospheric pressure could be associated with a range of weather conditions that might affect transit ridership, from clear, calm days to stormy periods that could impact service reliability and passenger comfort. High atmospheric pressure is generally associated with fair weather, while low pressure may indicate the possibility of rain or storms (Seco et al., 2012). As a result, commuters may adjust their transportation choices based on weather conditions. Figure 12 shows these changes in passengers' behaviour.

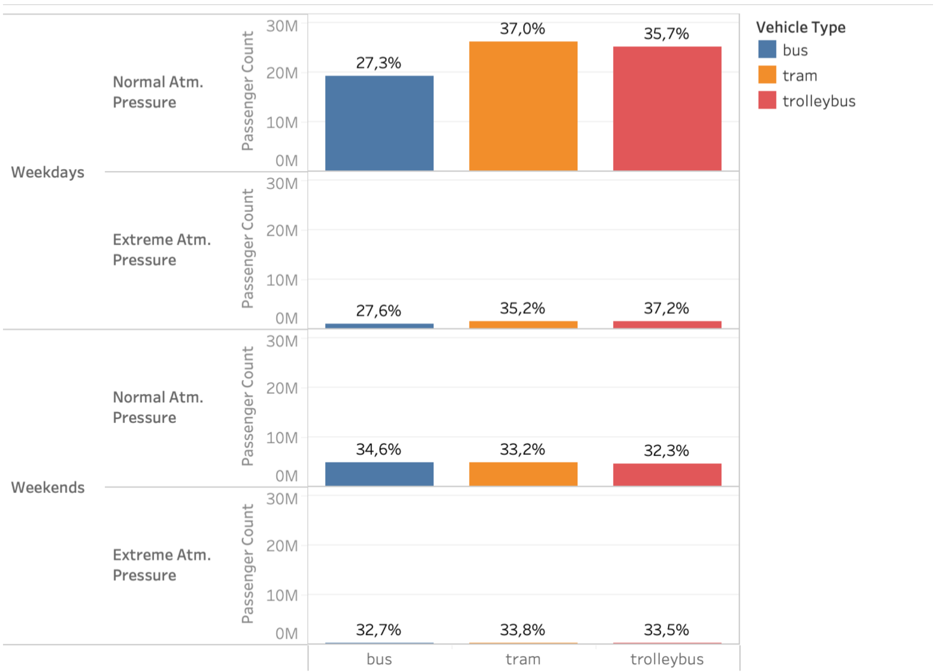


Figure 12: Impact of extreme atmospheric pressure conditions on modes of transport

On weekdays, the usage of trams decreases by 1.5%, and the usage of trolleybuses decreases by 1.8% during extreme atmospheric pressure conditions. However, the usage of buses remains relatively stable. Conversely, on weekends, the usage of buses decreases by 1.9% during extreme atmospheric pressure conditions, while the usage of trams and trolleybuses remains unchanged.

Since atmospheric pressure is not directly visible or perceptible to passengers, it makes it challenging to explain shifts in their behaviour solely based on this factor. However, passengers may indirectly respond to atmospheric pressure through its correlation with certain weather conditions. For instance, changes in atmospheric pressure are often associated with shifts in temperature, humidity, and the likelihood of precipitation.

5.1.5 Variance Inflation Factor

Variance Inflation Factor (VIF) assesses the level of multicollinearity among predictor variables in a regression model. Multicollinearity refers to the correlation between independent variables, and high levels of multicollinearity can affect the reliability and interpretability of regression results (Senthilnathan, 2019). VIF values range from 1 to positive infinity (Shah et al., 2023), with a VIF of 1 signifying no correlation implying perfect independence. Ideally, lower VIF values, closer to 1, are desirable, while higher values suggest increasing levels of multicollinearity (Senthilnathan, 2019).

In this study, the Variance Inflation Factor was employed to assess the degree of multicollinearity between 'passenger count' and the weather variables. The calculated VIF for 'passenger count' in the context of the selected predictor variables - 'temperature', 'humidity', 'atmospheric pressure', and 'precipitation' - is 1.588. This value indicates relatively low multicollinearity, providing a positive outlook for regression analysis (Kumari, 2008). Specifically, it suggests that 'passenger count' exhibits a reasonable level of independence from the chosen weather-related predictor variables.

In practical terms, this outcome is favourable for the regression analysis, implying that 'passenger count' is not heavily influenced by collinearity with the selected weather variables. This independence enhances the reliability and interpretability of the regression model, as the predictor variables can be considered relatively distinct in their influence on passenger count.

5.1.6 Collinearity between Passenger Count and Weather

In examining the relationship between passenger count and weather conditions, an exploration of collinearity becomes crucial. The interplay between passenger count - the dependent variable - and various weather-related factors serves as a fundamental aspect of understanding how external elements contribute to patterns in public transportation utilisation. For such purposes, a correlation coefficient, denoted as 'r', was calculated.

The correlation coefficient is a statistical measure that evaluates the strength and direction of a linear relationship between two variables (Onur et al., 2020). Its range spans from -1 to 1, where a correlation coefficient of 1 signifies a perfect positive correlation—meaning that as one variable increases, the other increases proportionally. Conversely, a correlation coefficient of -1 indicates a perfect negative correlation - when one variable increases, the other decreases proportionally. A correlation coefficient of 0 suggests no linear correlation between the variables; they are not related linearly. Furthermore, the magnitude of the correlation coefficient indicates the strength of the relationship. Values closer to 1 or -1 represent stronger correlations, while values closer to 0 indicate weaker or negligible correlations (Schober et al., 2018). It is important to note that correlation does not imply causation, emphasising the need for careful interpretation and consideration of additional factors in understanding the relationship between variables.

Collinearity, especially in the context of multiple regression analysis, unveils the degree to which independent variables such as temperature, humidity, atmospheric pressure, and precipitation may exhibit correlations across various conditions. These conditions include weekdays, weekends, public holidays, school holidays, peak time, off-peak time, and more. This exploration seeks to reveal the interrelationships among these variables, providing insights into how they may be correlated under different circumstances.

	Temperature	**Precipitation**	**Atmospheric Pressure**	**Humidity**
Any Day	0.004138	-0.006800	-0.000410	-0.041595
Weekday	-0.002845	0.004625	-0.033539	-0.033539
Weekend	0.033504	0.000591	0.003518	-0.079633
Public Holidays	**0.167211**	-0.021324	-0.037021	**-0.187501**
Peak-Time (non-pub. holidays)	-0.049025	-0.006902	-0.000443	0.02201
Off Peak-Time (non-pub. holidays)	0.020572	-0.006270	0.002905	-0.066397
School Holidays	0.017523	-0.010363	0.000073	-0.045423
School Holidays, No Public Holidays and Weekdays	0.002770	-0.008881	0.002478	-0.034485
School Holidays, No Public Holidays and Weekends	0.054788	-0.001433	-0.002565	-0.086147

Table 2: Collinearity among passenger count and weather variables under different conditions

The correlation values presented in Table 2 offer insights into the interrelationships among variables within the dataset. In general, it can be observed that there is no strong correlation between any of the weather variables and different conditions. The most notable correlation is observed at 0.167211 between temperature and public holidays. However, even this correlation signifies only a weak positive association, suggesting a subtle tendency for public holiday occurrences to increase slightly as temperature rises. On the other hand, the correlation of -0.187501 between humidity and public holidays indicates a weak negative correlation.

It is essential to recognise that these correlation values fall on the lower end of the scale, suggesting that the linear relationships between temperature, humidity, and public holidays are not strongly pronounced. While correlation values near 0 indicate weak relationships, it is crucial to consider the broader context, potential nonlinear associations, and the influence of other factors that may contribute to the observed patterns. In summary, the correlation coefficients highlight subtle tendencies between temperature, humidity, and public holidays, emphasising the importance of further exploration and analysis to uncover more nuanced relationships within the dataset.

The following analysis is dedicated to an examination of collinearity between passenger count in buses, trams, and trolleybuses and independent weather variables, these being temperature, precipitation, atmospheric pressure, and humidity. Similarly to the previous section, the correlation coefficients between these variables are also notably low (see Table 3). This suggests that there is a lack of collinearity, emphasising that the changes in passenger count are not associated with the variations in the usage of different vehicle types in specific weather conditions. Essentially, the independence of these variables implies that each vehicle is unaffected by variations in each weather variable.

	Temperature	**Precipitation**	**Atmospheric Pressure**	**Humidity**
Bus	-0.001149	-0.004697	-0.001291	-0.040997
Tram	0.054022	-0.006950	0.011569	-0.087894
Trolleybus	0.000979	-0.012448	0.006189	-0.038882

Table 3: Collinearity among passenger count and weather variables based on vehicle type

In the exploration of co-linearity between passenger count and weather variables depicted in Table 4, it can be seen that negligible observations emerge during both extreme temperature events and typical weather conditions. There is a very slight tendency for the passenger count to decrease as temperatures rise during extreme heat, but it is crucial to emphasise that this relationship is exceptionally weak. Similarly, under typical weather conditions, the impact on passenger count is minimal, with an almost imperceptible decrease. These nuanced findings underscore the complexity of the interplay between weather variables and ridership, with the identified trends being subtle and requiring careful consideration in the broader context of factors influencing public transport usage.

	Normal Weather	Extreme Weather
Temperature	0.000543	-0.018549
Precipitation	-0.001174	-0.012692
Atmospheric Pressure	-0.005572	-0.015622

Table 4: Collinearity among passenger count and weather variables based on weather conditions

5.1.7 Summary of Exploratory Data Analysis

In this extensive exploration of the impact of various factors on public transportation ridership, several key findings emerged. Firstly, when examining average monthly ridership patterns, it was observed that ridership exhibited a gradual decline from May to September, with the most significant decrease occurring from June to July. September recorded the lowest average passenger count, influenced by the transitional period from summer to fall. October, however, experienced the highest average ridership, indicating a resurgence, potentially influenced by cooler weather and the resumption of regular work and school schedules.

Moreover, when analysing average hourly ridership patterns, it became apparent that significant variations exist between weekdays and weekends. Weekdays follow a more predictable pattern with peak hours during typical commuting times, while weekends show a gradual increase in ridership throughout the day, peaking in the evening. The preference for different vehicle types also showcased interesting dynamics, with trams being the most utilised mode overall, closely followed by trolleybuses and buses. Weekday preferences indicated a sustained high preference for trams, reflecting their efficiency for daily commuting, while weekends displayed a more balanced distribution among all three modes.

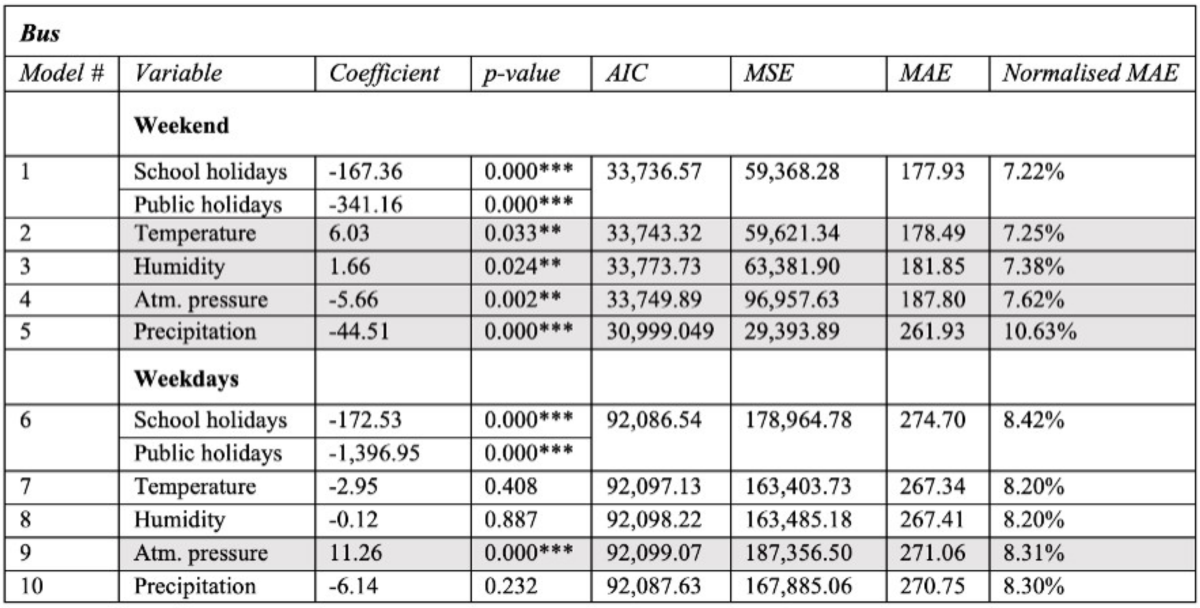
Collinearity analysis between passenger count and weather variables demonstrated relatively low multicollinearity, enhancing the reliability of regression analysis. Additionally, exploring collinearity between passenger count and vehicle types revealed a lack of correlation, suggesting that changes in passenger count are not significantly associated with variations in the usage of different vehicle types under specific weather conditions. Similarly, the co-linearity analysis comparing extreme and normal weather conditions indicated a negligible impact on passenger count in both scenarios. Overall, these findings lay the foundation for deeper insights into the complex dynamics of public transportation ridership. Having established a strong foundation and a thorough understanding of the dataset, the subsequent section will dive into predictive analysis.

5.2 Time Series Analysis

In this time series analysis, a rigorous exploration of how various factors influence ridership patterns across buses, trolleybuses, and trams will be conducted.

5.2.1 SARIMAX: Buses

The time series analysis of bus ridership shown in Table 5 yields informative insights into how various factors affect usage on weekends and weekdays. On weekends, the presence of school holidays results in a decrease of 167.36 in ridership, while public holidays lead to a more substantial decrease of 341.16, as reflected by their negative coefficients and highly significant p-values. Weather variables also play a discernible role: temperature increases ridership modestly, with a coefficient of 6.03, and humidity slightly increases it, with a coefficient of 1.66, both statistically significant. Atmospheric pressure shows a small negative effect with a coefficient of -5.66, and precipitation notably decreases ridership by 44.51.



Bus							
Model #	*Variable*	*Coefficient*	*p-value*	*AIC*	*MSE*	*MAE*	*Normalised MAE*
	Weekend						
1	School holidays	-167.36	0.000***	33,736.57	59,368.28	177.93	7.22%
	Public holidays	-341.16	0.000***				
2	Temperature	6.03	0.033**	33,743.32	59,621.34	178.49	7.25%
3	Humidity	1.66	0.024**	33,773.73	63,381.90	181.85	7.38%
4	Atm. pressure	-5.66	0.002**	33,749.89	96,957.63	187.80	7.62%
5	Precipitation	-44.51	0.000***	30,999.049	29,393.89	261.93	10.63%
	Weekdays						
6	School holidays	-172.53	0.000***	92,086.54	178,964.78	274.70	8.42%
	Public holidays	-1,396.95	0.000***				
7	Temperature	-2.95	0.408	92,097.13	163,403.73	267.34	8.20%
8	Humidity	-0.12	0.887	92,098.22	163,485.18	267.41	8.20%
9	Atm. pressure	11.26	0.000***	92,099.07	187,356.50	271.06	8.31%
10	Precipitation	-6.14	0.232	92,087.63	167,885.06	270.75	8.30%

Table 5: Time series analysis of bus ridership using the SARIMAX model (*** p<0.001, **p<0.05)

For weekdays, the trend is similar for school and public holidays, showing a significant reduction in bus use, but with school holidays showing a smaller decrease of 172.53 compared to weekends. Public holidays have a pronounced negative effect, with a decrease of 1,396.95 in ridership. Temperature changes do not significantly influence weekday ridership, nor does humidity, which has an insignificant and minuscule coefficient. However, atmospheric pressure contributes positively to ridership, with a significant coefficient of 11.26. Precipitation on weekdays does not have a statistically significant effect, with a p-value of 0.232.

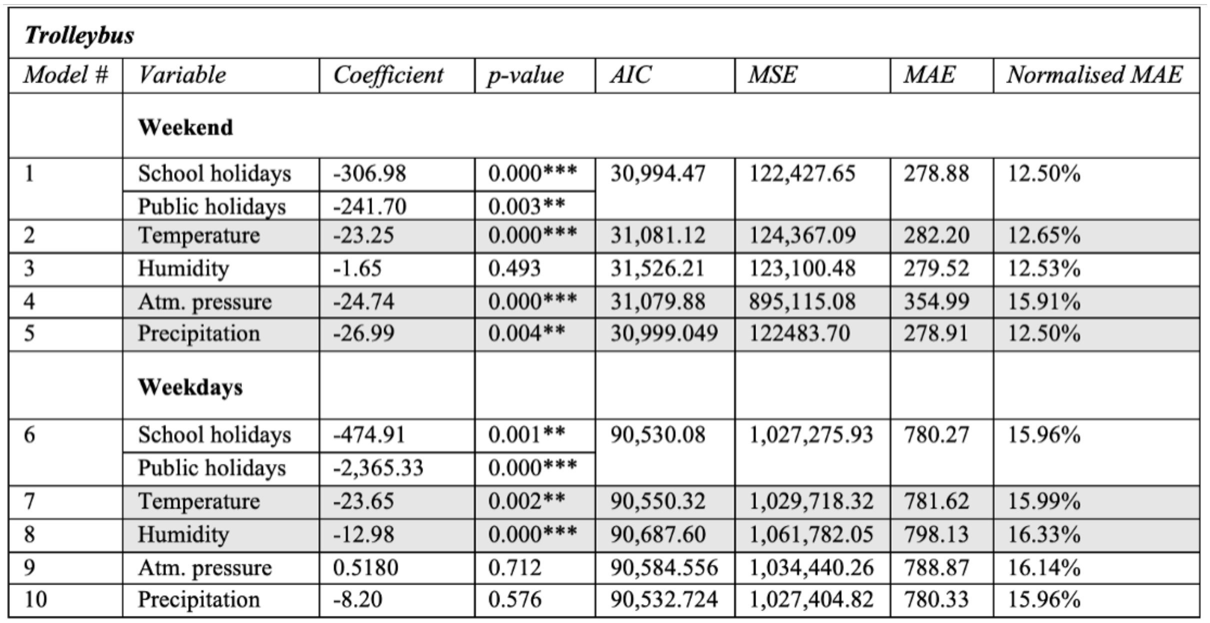
The AIC values in the bus ridership models reveal a more satisfactory statistical fit for the weekend models, with AIC values ranging from 30,994.049 to 33,773.73, as opposed to the higher AIC values ranging from 92,086.54 to 92,099.07 for the weekday models. This distinction suggests that the weekend models might capture the relationship between ridership and the influencing factors with more accuracy, whereas the weekday models could be contending with additional complexities or variables not accounted for within the model framework.

While the AIC values give us an insight into the model fit, the MSE and MAE are indicative of the variability in the data that the models are attempting to capture, with higher values for the weekday models signalling larger deviations from observed values. These deviations are quantified by MSE values from 178,964.78 to 187,856.50 and MAE values from 270.75 to

274.70. The Normalised MAE, providing a proportionate measure of error, shows a consistent pattern across the models, but highlights the significant influence of precipitation on weekends, where it leads to higher normalised errors of up to 10.63%. These metrics together not only affirm the differences in weekend and weekday ridership patterns but also reveal the challenges in predicting weekday ridership of buses with the current model parameters.

5.2.2 SARIMAX: Trolleybuses

For trolleybuses, the time series analysis reveals clear patterns regarding ridership influences during both weekends and weekdays. As shown in Table 6, on weekends, school holidays contribute to a decrease in ridership by 306.98, while public holidays show a reduction of 241.70, both with highly significant p-values, indicating strong confidence in these results. Temperature shows a considerable negative impact with a coefficient of -23.25, again statistically significant. Humidity does not significantly affect weekend ridership, which is supported by a p-value of 0.493. Atmospheric pressure and precipitation both show a negative influence on ridership, with coefficients of -24.74 and -26.99 respectively, and both variables are statistically significant.



Trolleybus							
Model #	*Variable*	*Coefficient*	*p-value*	*AIC*	*MSE*	*MAE*	*Normalised MAE*
	Weekend						
1	School holidays	-306.98	0.000***	30,994.47	122,427.65	278.88	12.50%
	Public holidays	-241.70	0.003**				
2	Temperature	-23.25	0.000***	31,081.12	124,367.09	282.20	12.65%
3	Humidity	-1.65	0.493	31,526.21	123,100.48	279.52	12.53%
4	Atm. pressure	-24.74	0.000***	31,079.88	895,115.08	354.99	15.91%
5	Precipitation	-26.99	0.004**	30,999.049	122483.70	278.91	12.50%
	Weekdays						
6	School holidays	-474.91	0.001**	90,530.08	1,027,275.93	780.27	15.96%
	Public holidays	-2,365.33	0.000***				
7	Temperature	-23.65	0.002**	90,550.32	1,029,718.32	781.62	15.99%
8	Humidity	-12.98	0.000***	90,687.60	1,061,782.05	798.13	16.33%
9	Atm. pressure	0.5180	0.712	90,584.556	1,034,440.26	788.87	16.14%
10	Precipitation	-8.20	0.576	90,532.724	1,027,404.82	780.33	15.96%

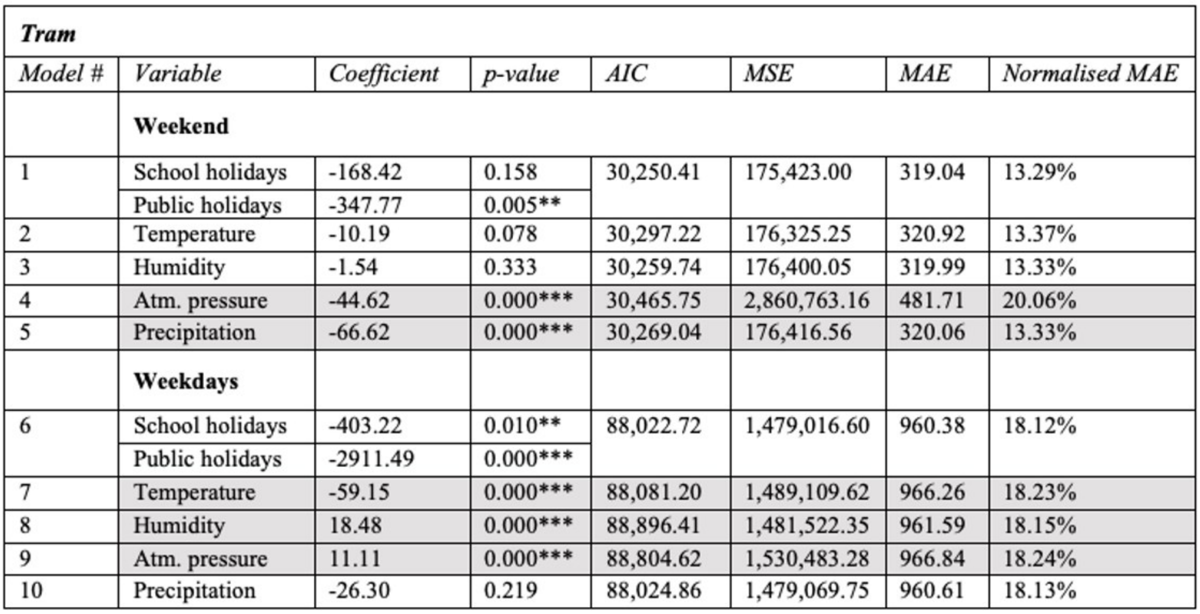
Table 6: Time series analysis of trolleybus ridership using the SARIMAX model (*** p<0.001, **p<0.05)

Weekdays exhibit a more pronounced effect; school holidays decrease ridership by 474.91, and public holidays show a drastic decrease with a coefficient of -2,365.33, highlighting a much stronger impact compared to weekends. Temperature slightly decreases ridership with a coefficient of -23.65, and humidity shows a significant negative effect on ridership with a coefficient of -12.98. Atmospheric pressure has an insignificant positive effect, with a coefficient close to zero and a p-value of 0.712, indicating no clear influence. Precipitation also does not significantly affect ridership on weekdays, as suggested by a p-value of 0.576.

AIC values, which serve as an indicator of the trade-off between model fit and complexity, are notably lower for weekend models, with values ranging from 30,994.47 to 31,526.21, compared to the weekday models which present AIC values between 90,530.08 and 90,687.60. This suggests that the weekend models are more effectively capturing the underlying patterns in ridership data, indicating a better fit and parsimonious explanation for weekend trends. On the other hand, the higher AIC values for weekdays imply potential complexities or additional variables that may be affecting ridership but are not included in the current models. Furthermore, the higher MSE and MAE values observed in weekday models underscore a greater prediction error and thus greater variance in ridership. This increased variance is echoed in the Normalised MAE percentages, which are higher on weekdays.

5.2.3 SARIMAX: Trams

The time series models for tram ridership reveal how different variables impact usage during weekends and weekdays. On weekends, the coefficient for school holidays is not significant (p-value of 0.158), suggesting no clear evidence that school holidays impact ridership, whereas public holidays show a significant decrease in ridership by 347.77. Temperature and humidity changes have non-significant coefficients, indicating no substantial effect on weekend tram use. However, atmospheric pressure and precipitation both significantly decrease ridership, with coefficients of -44.62 and -66.62, respectively.



Tram							
Model #	*Variable*	*Coefficient*	*p-value*	*AIC*	*MSE*	*MAE*	*Normalised MAE*
	Weekend						
1	School holidays	-168.42	0.158	30,250.41	175,423.00	319.04	13.29%
	Public holidays	-347.77	0.005**				
2	Temperature	-10.19	0.078	30,297.22	176,325.25	320.92	13.37%
3	Humidity	-1.54	0.333	30,259.74	176,400.05	319.99	13.33%
4	Atm. pressure	-44.62	0.000***	30,465.75	2,860,763.16	481.71	20.06%
5	Precipitation	-66.62	0.000***	30,269.04	176,416.56	320.06	13.33%
	Weekdays						
6	School holidays	-403.22	0.010**	88,022.72	1,479,016.60	960.38	18.12%
	Public holidays	-2911.49	0.000***				
7	Temperature	-59.15	0.000***	88,081.20	1,489,109.62	966.26	18.23%
8	Humidity	18.48	0.000***	88,896.41	1,481,522.35	961.59	18.15%
9	Atm. pressure	11.11	0.000***	88,804.62	1,530,483.28	966.84	18.24%
10	Precipitation	-26.30	0.219	88,024.86	1,479,069.75	960.61	18.13%

Table 7: Time series analysis of tram ridership using the SARIMAX model (*** p<0.001, **p<0.05)

In Table 7, the models for weekdays reveal that several variables significantly influence tram usage. School holidays show a substantial decrease in ridership, with a reduction of 403.22, while public holidays also contribute to a decrease with a coefficient of -291.49. Temperature plays a notable role; as it drops, tram ridership follows, indicated by a coefficient of -59.15. On the other hand, increased humidity correlates with a rise in ridership, marked by a coefficient of 18.48, and atmospheric pressure has a similar positive association, with a coefficient of 11.11. Precipitation's impact on weekdays, however, is not statistically significant, with a p-value of 0.219, suggesting that it does not have a clear effect on ridership during these days.

When considering model performance, the AIC values show a contrast between weekends and weekdays; the much lower AICs for weekend models suggest that they are capturing the fluctuations in ridership with more fidelity. The elevated MSE for the weekend model that includes atmospheric pressure, soaring to 2,860,763.16, may point to potential anomalies or issues with model fit. In tandem, the higher Normalised MAE on weekdays, especially pronounced in the school holidays model at 18.12%, indicates greater discrepancies between observed and estimated ridership, highlighting the complexity and unpredictability of modelling weekday tram usage. These metrics, particularly the higher weekday Normalised MAE, underscore the challenges in accounting for the varied influences on tram ridership, suggesting that weekday patterns may be affected by additional, unmodeled factors.

5.2.4 Summary of Time Series Analysis

In the time series analysis of Bratislava's urban mobility, SARIMAX models were trained to investigate the effects of seasonal variations and exogenous factors, such as weather conditions and holidays, on the ridership of buses, trolleybuses, and trams. The most significant findings emerge from the variations in ridership due to school and public holidays. For buses, ridership drops by 167.36 during weekend school holidays and even more significantly, by 341.16, during public holidays. Interestingly, atmospheric pressure has a small yet consistently negative effect across all transport types, while precipitation impacts are most substantial on tram ridership during weekends, with a decrease of 66.62.

A critical aspect of this analysis is the performance indicators for the models. Weekend models showed lower AIC values, indicating a more robust fit for the data, with bus models scoring between 30,994.049 to 33,773.73 and tram models around 30,250.41. This pattern suggests a more consistent and predictable ridership behaviour during weekends. However, weekday models, particularly for trolleybuses and trams, exhibit higher AIC values - up to 90,687.60 for trolleybuses - pointing to potentially complex dynamics not fully explained by the model. The precision of the models, measured by MSE and MAE, reinforces this notion, with weekday models displaying higher values and indicating larger deviations from observed values, such as the tram model for atmospheric pressure, which sees an MSE as high as 2,860,763.16.

Overall, these findings underscore the complex interplay between various factors and ridership. The higher normalised MAE values across all transport modes during weekdays highlight the unpredictability and complexity inherent in modelling weekday ridership. This analysis illuminates the challenges in capturing the full scope of influences affecting urban mobility, providing a foundation for further research and policy-making to enhance public transport services.

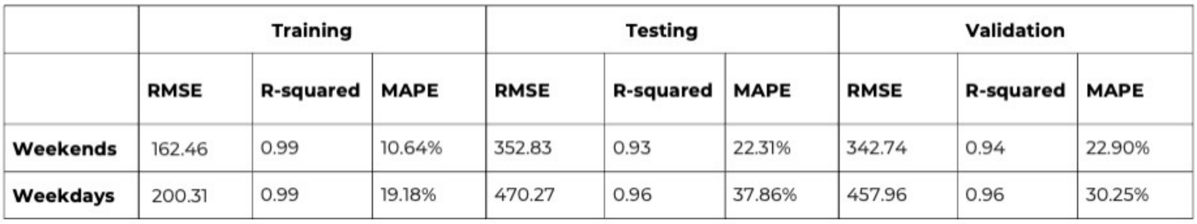
5.3 Predictive Analysis

Although SARIMAX models are traditionally leveraged for predictive purposes, within the scope of this study, their application was directed toward understanding the time series characteristics and the impact of external variables on public transport ridership. The decision to not use SARIMAX models for forecasting resulted from a high variability and the associated prediction errors indicated by the elevated MSE and MAE values in the weekday models. These metrics highlighted the models' limitations in reliably capturing the complex dynamics of weekday ridership, as evidenced by the inconsistencies and large deviations from observed data. Given the limited reliability of SARIMAX models for accurate predictions in this specific context, the potential of XGBoost for predictive analysis given its high performance and efficiency will be explored.

5.3.1 XGBoost: Buses

The results from the XGBoost models for the ridership of buses on weekends and weekdays offer interesting insights into the model's performance across different data partitioning phases. Please see Table 8 for more details.

During the training phase for weekends, the model demonstrated an excellent fit with an R-squared of 0.99, indicating that the model could explain 99% of the variance within the training data, a very high value signifying a strong model. The RMSE stood at 162.46, signifying the model's predictions were, on average, within this range from the actual values. The MAPE of 10.64% reflects that the model's predictions were, on average, off by this percentage, which is an acceptable error rate for real-world predictions. In the testing phase, however, there was a decline in R-squared to 0.93, indicating a slightly less accurate fit to the testing data, though still quite robust. The increase in RMSE to 352.83 and MAPE to 22.31% during testing suggests that while the model remains quite predictive, its performance is not as strong as with the training data, a common occurrence in model deployment.



	Training			Testing			Validation		
	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE
Weekends	162.46	0.99	10.64%	352.83	0.93	22.31%	342.74	0.94	22.90%
Weekdays	200.31	0.99	19.18%	470.27	0.96	37.86%	457.96	0.96	30.25%

Table 8: Results of predictive analysis of bus ridership using the XGBoost model

For weekdays, the training phase results were also strong, with an R-squared of 0.99, but the RMSE was higher at 200.31, indicating larger errors in the model's predictions as compared to the weekend model. The MAPE of 19.18% was significantly higher, showing that the model's predictions were less accurate for the more variable weekday data. In the testing phase, the R-squared remained high at 0.96, indicating good model performance. However, the increase in RMSE to 470.27 and MAPE to 37.86% underscores the challenges of accurately predicting weekday ridership, where the errors and variance are noticeably larger.

These models were further validated, as evidenced by the validation phase results, where the RMSE and MAPE values show how well the model could generalise to another unseen subset of the data. With an RMSE of 342.74 and MAPE of 22.90% for weekends, and an RMSE of 457.96 and MAPE of 30.25% for weekdays, the models prove to be reliable, though with understandable deviations in a complex real-world scenario.

5.3.1.1 Visual Predictions for Buses

The scatter plots in Figure 15, illustrating the predicted passenger count on the y-axis against the actual passenger count on the x-axis, were generated using the shuffled testing dataset to ensure an unbiased evaluation of the model's predictive accuracy. In these plots, points clustered closely around the line of perfect prediction indicate a high degree of model accuracy. The plot for weekends displays a wider dispersion of points from the perfect prediction line, suggesting either a greater inherent variability in weekend ridership or potential overfitting of the model to the training data.

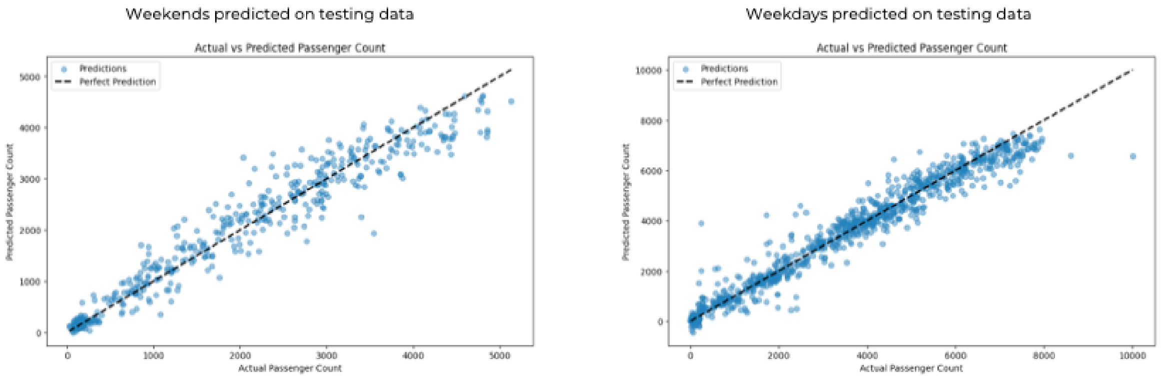


Figure 15: Actual versus predicted passenger count of buses on weekends and weekdays

In contrast, the weekday plot shows a tighter grouping of points, indicating more consistent predictions, although the elevated RMSE and MAPE values highlight complex patterns in weekday ridership that the model may not fully encapsulate. These visual and statistical analyses conducted on shuffled testing data provide a robust framework for assessing model performance, identifying areas for improvement, and enhancing the overall reliability of predictions.

The graph in Figure 16 depicts the actual versus predicted ridership of buses on weekdays, given the dates from October 15th to 17th (Monday to Wednesday) based on the XGBoost model's performance on the validation dataset. The data was not shuffled, indicating that the time series aspect of the data was preserved, which is essential for capturing and predicting patterns over time. The blue line represents the actual number of passengers counted, while the red line represents the predicted count from the XGBoost model.

The model appears to capture the general trend in ridership quite well, closely tracking the actual counts, particularly capturing the peaks and troughs throughout the days. There are periods, especially during the early morning and late evening hours, where the predicted and actual counts align very closely, suggesting the model effectively captures the patterns during these times.

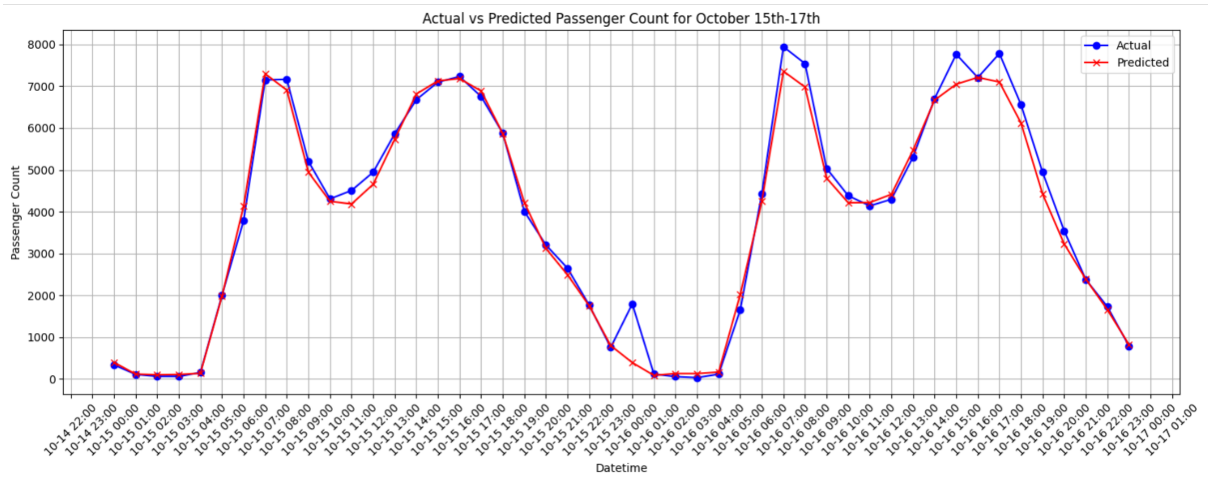


Figure 16: Actual versus predicted passenger count of buses for October 15th-17th, 2018 (weekday)

However, there are a few points, particularly in the midday peaks, where the model does not precisely match the actual counts, other than it it follows the overall trend accurately. This might be due to sudden changes in ridership that the model, despite its accuracy, is not nimble enough to predict, or it could be influenced by factors not included in the model. The precision with which the model forecasts the less dramatic increases and decreases in ridership overnight and during off-peak hours is noteworthy and suggests it has learned the quieter patterns of bus use effectively. Overall, the model's predictions are quite impressive, but the deviations during peak hours could be a focus for further model refinement.

Moreover, Figure 17 presents a comparison between the actual and predicted ridership of buses on weekends, based on the XGBoost model's performance on the validation dataset. Similarly, as for the weekday, the data was not shuffled for purposes of this time series visualisation. The plot shows that the model has done somewhat well in capturing the overall trend of bus ridership across the specified weekend of August 11th to 12th. The predictions in general follow the actual counts, however, it struggles with morning Sunday hours. This alignment indicates that generally, the model understands the temporal dynamics of bus ridership for a weekend, however, there is still room for improvement.

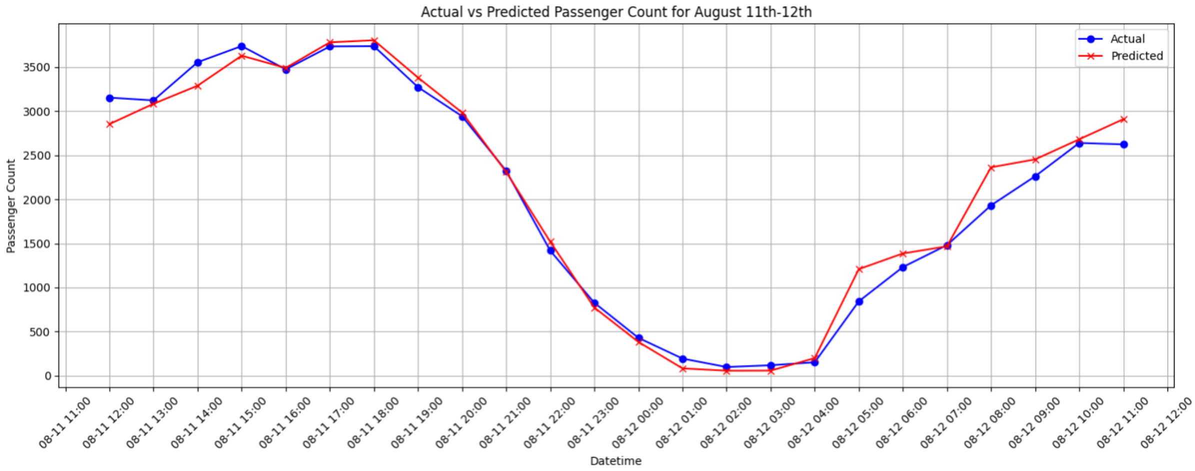
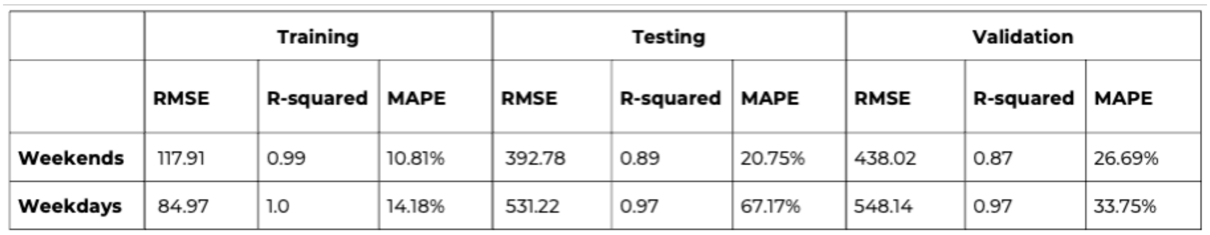


Figure 17: Actual versus predicted passenger count of buses for August 11th-12th, 2018 (weekend)

There is a notable synchronisation in the rise and fall of passenger counts, suggesting that the model has successfully learned the typical weekend patterns, such as increased travel during certain hours of the day. The slight deviations between the actual and predicted values are minimal, implying that the model's parameters are well-tuned to capture the underlying weekend ridership patterns without being misled by the order of the data. Overall, the graph demonstrates the model's robust predictive power and offers confidence in its potential applicability for planning and optimising bus services during weekends.

5.3.2 XGBoost: Trolleybuses

The results for the XGBoost models predicting trolleybus ridership provide distinct performance metrics for weekends and weekdays as can be seen in Table 9. For weekends, the training phase indicates a high level of accuracy with an RMSE of 117.91 and an R-squared of 0.99, suggesting the model predictions were very close to the actual data. The MAPE at 10.81% indicates a reasonable average percentage error. However, in the testing phase, there is a notable increase in RMSE to 392.78 and a rise in MAPE to 20.75%, indicating that the model's predictions deviated more from the actual counts when faced with unseen data. The validation phase showed a further increase in RMSE to 438.02 and a decrease in R-squared to 0.87, coupled with a MAPE of 26.69%, which suggests that the model’s ability to generalise to new data is less accurate than in the training phase.



	Training			Testing			Validation		
	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE
Weekends	117.91	0.99	10.81%	392.78	0.89	20.75%	438.02	0.87	26.69%
Weekdays	84.97	1.0	14.18%	531.22	0.97	67.17%	548.14	0.97	33.75%

Table 9: Results of predictive analysis of trolleybus ridership using the XGBoost model

For weekdays, the model achieved perfect training performance with an R-squared of 1.0, but this is likely a sign of overfitting, especially considering the high MAPE of 14.18% even during training. The overfitting is substantiated in the testing phase, where RMSE jumps to 531.22, and the MAPE escalates to 67.17%, highlighting significant prediction errors. The validation metrics confirm this with an RMSE of 548.14 and a MAPE of 33.75%, although the R-squared remains high at 0.97. These figures reflect challenges in the model's predictive capacity, particularly in capturing the more complex weekday patterns.

5.3.2.1 Visual Predictions for Trolleybuses

The scatter plots in Figure 18 provide a visual context for the above-explained metrics. The plot for weekends shows data points that are more scattered around the line of perfect prediction, indicating variability in accuracy.

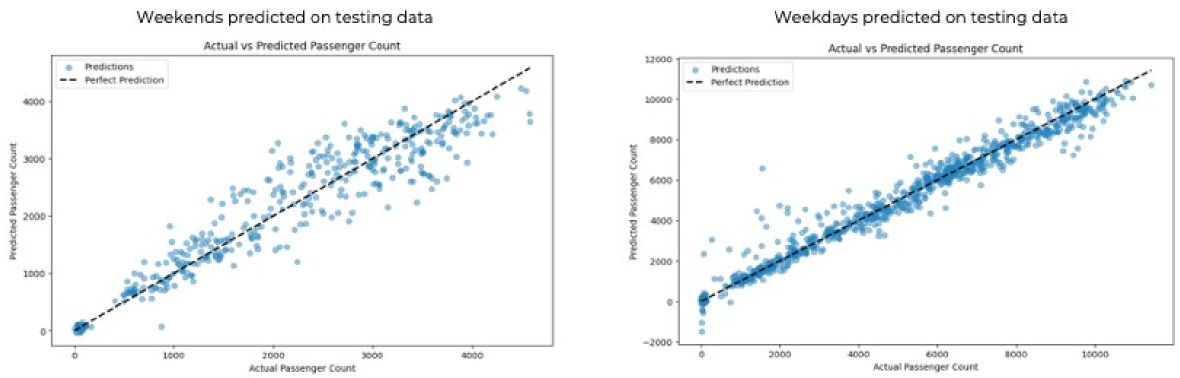


Figure 18: Actual versus predicted passenger count of trolleybuses on weekends and weekdays

The weekday plot, while showing a tighter cluster around the line, still exhibits significant deviation for higher counts, underscoring the discrepancies between predicted and actual ridership. These visual and numerical results suggest that while the models capture overall trends, there is room for improvement in accuracy, particularly in managing the high variability and complexity of weekday ridership patterns. Figure 19 presents the actual versus predicted passenger count for trolleybuses during a typical weekday period from October 15th to 17th (Monday to Wednesday).

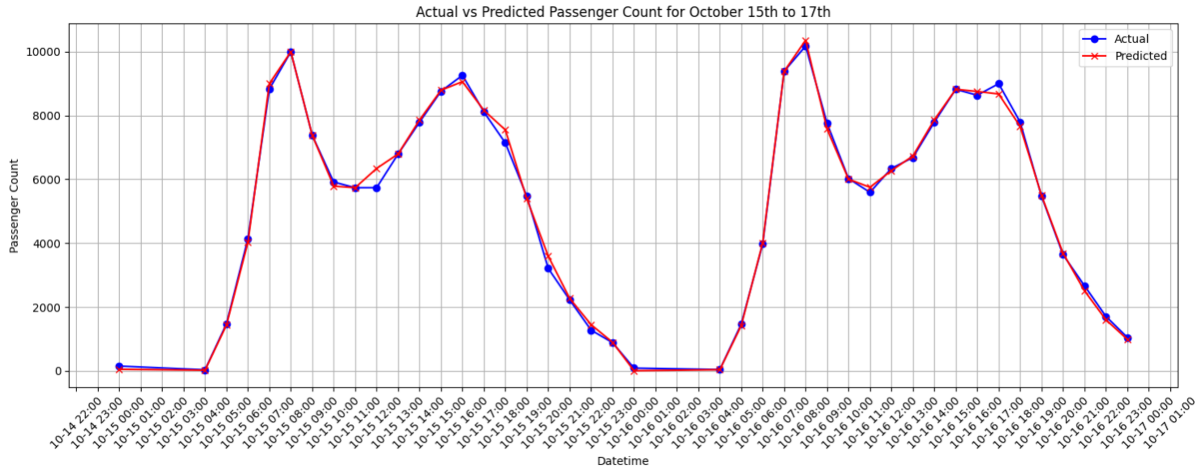


Figure 19: Actual versus predicted passenger count of trolleybuses for October 15th-17th, 2018 (weekday)

The model appears to follow the actual ridership trends closely, with the predicted values aligning well with the actual data. This suggests that the model has effectively learned the daily patterns of trolleybus usage throughout the weekdays. The predictions accurately reflect the sharp increases and decreases in passenger counts, which are characteristic of rush hours in the morning and evening when people are likely commuting to and from work or school. However, there are points, particularly during the peak hours, where the model either overestimates or underestimates the actual count. The overestimations and underestimations may be due to sudden, unpredictable spikes or drops in ridership that the model parameters are not adjusted to handle, or they could be related to external factors not included in the model.

Despite these discrepancies, the overall accuracy of the model is notable, especially in non-peak periods, where the predicted and actual counts are very close. This high level of accuracy during off-peak times might be attributed to more stable and predictable ridership patterns, as opposed to the peak times which can be influenced by a greater number of variables. Overall, the model demonstrates strong predictive capabilities, with potential areas for refinement identified in the peak periods where the complexities of ridership patterns are most pronounced. The following graph in Figure 20 shows the actual versus predicted passenger count for trolleybus ridership during a weekend, from October 20th to 21st.

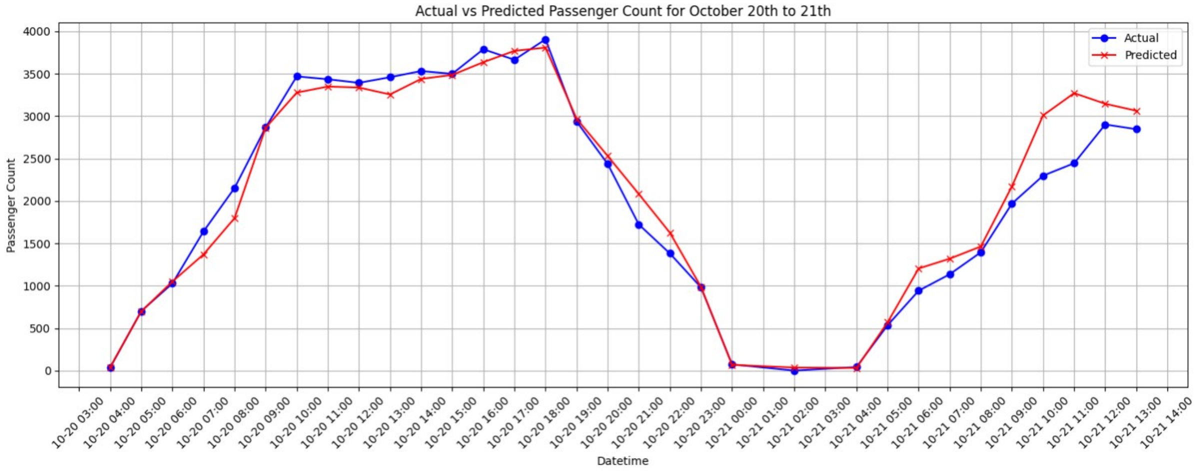
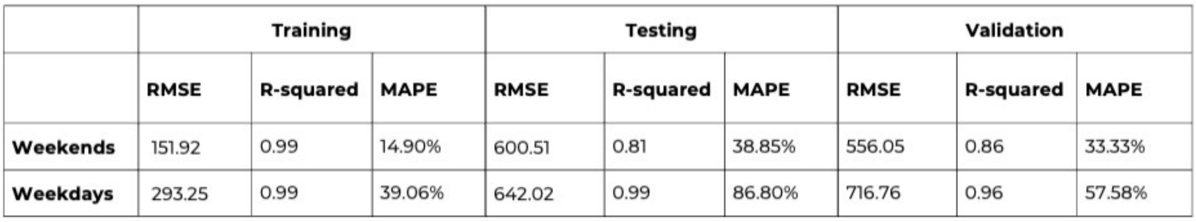


Figure 20: Actual versus predicted passenger count of trolleybuses for October 20th-21st, 2018 (weekend)

The model appears to predict the ridership trend with a high degree of accuracy, closely following the actual counts. There is a clear correlation in the rise and fall of passenger numbers, with the predictions mirroring the actual data points throughout most of the time period. This suggests the model has successfully captured the key factors influencing trolleybus usage over the weekend. Notably, the predictions and actual counts are almost indistinguishable during the early hours and late evenings, but some discrepancies are visible during peak times, particularly in the midday and early evening. These discrepancies might be due to unmodeled factors that specifically affect ridership at these times, or they could stem from random variations that are difficult to predict. The consistency in the model's performance over two days indicates its potential reliability for planning and operational adjustments during weekend periods.

5.3.3 XGBoost: Trams

As summarised in Table 10, The XGBoost model's performance for predicting tram ridership shows different results for training, testing, and validation phases for both weekends and weekdays. During the training phase, both weekend and weekday models have high R-squared values (0.99), indicating that the models can explain 99% of the variance in the data. The RMSE is lower for weekends (151.92) than weekdays (293.25), suggesting that the predictions for weekend ridership are closer to the actual values. The MAPE values, which represent the average percentage error between the predicted and actual values, are 14.90% for weekends and 39.06% for weekdays, showing that the weekend predictions are, on average, more accurate.



	Training			Testing			Validation		
	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE	RMSE	R-squared	MAPE
Weekends	151.92	0.99	14.90%	600.51	0.81	38.85%	556.05	0.86	33.33%
Weekdays	293.25	0.99	39.06%	642.02	0.99	86.80%	716.76	0.96	57.58%

Table 10: Results of predictive analysis of tram ridership using the XGBoost model

In the testing phase, both models experience a drop in R-squared and an increase in MAPE. The R-squared for weekends falls to 0.81, and for weekdays to 0.99, indicating a good fit to the unseen data despite being lower than the training phase. The MAPE increases substantially, to 38.85% for weekends and 86.80% for weekdays, showing that the model's predictions are less accurate when applied to new data.

The validation phase results are similar to the testing phase, with the weekend R-squared at 0.86 and weekday R-squared at 0.96. However, while the MAPE decreases to 33.33% for weekends and 57.58% for weekdays compared to the testing phase, suggesting improved accuracy on the validation data, these values, particularly for weekends, remain notably high. This implies that the prediction error could still be significant.

5.3.3.1 Visual Predictions for Trams

The scatter plots in Figure 21 illustrate the predicted versus actual passenger counts for weekends and weekdays using shuffled testing datasets. Both plots show that the predictions are in line with the actual counts, but with some variability as indicated by the spread of the points around the line of perfect prediction. The weekend model appears to have a wider spread, particularly for higher passenger counts, which might reflect greater difficulty in predicting peak ridership times. The weekday model shows a tighter cluster of points, indicating more consistent predictions, though there is still noticeable variation.

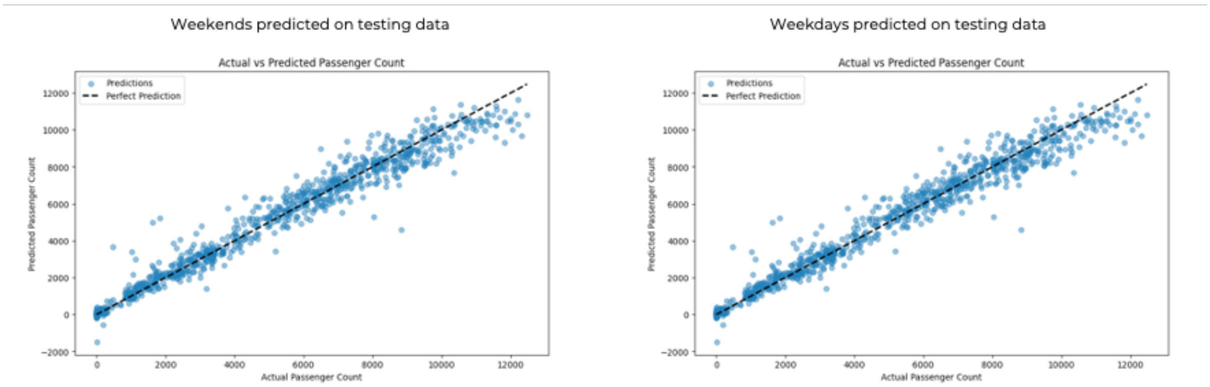


Figure 21: Actual versus predicted passenger count of trams on weekends and weekdays

Overall, the model performs well, but the increased MAPE during the testing phase and the spread of points in the scatter plots suggest there is room for improvement. The results demonstrate the challenges in predicting public transit ridership but also confirm the potential of machine learning models to provide valuable insights into transit patterns.

Further, the actual versus predicted values using the validation dataset were plotted to assess the performance of the tram ridership model for weekdays, from October 15th to 17th (Monday to Wednesday). The graph in Figure 22 illustrates that the predicted values follow the same general pattern as the actual counts, suggesting that the model is capturing the overall trend in tram usage during weekdays. The predictions closely match the actual data at several points, particularly during peak hours, indicative of the model's ability to understand and replicate the daily commuting patterns.

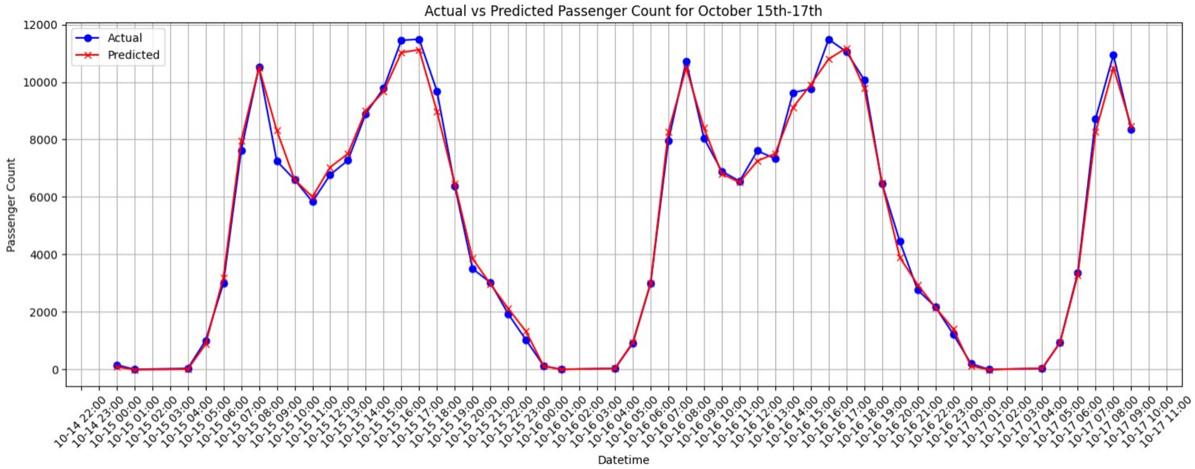


Figure 22: Actual versus predicted passenger count of trams for October 15th-17th, 2018 (weekday)

Despite this, there are moments, especially during peak periods, when the model's predictions deviate from the actual counts. These discrepancies could be due to a variety of factors not accounted for in the model, such as irregular fluctuations in ridership or events that cause sudden changes in the number of passengers. The model's performance on the validation dataset, which it was not trained on, suggests a good level of generalisability. However, the presence of some errors, particularly during peak hours, indicates opportunities for further refining the model to enhance its accuracy across all times of the day.

Furthermore, the graph in Figure 23 depicts the actual versus predicted passenger count for trams during a weekend from October 13th to 14th. Observing the trends, the model does not seem to have a strong handle on the overall ridership patterns, quite deviating from the actual data with its predictions, especially during lunch and early afternoon hours. The lines clearly do not ascend and descend in sync, indicating that the model predictions are not capturing the passenger behaviour well enough. This observation aligns with the high MAPE on the validation dataset, suggesting that the model's performance is suboptimal. Such discrepancies could be due to various factors that affect ridership more strongly at these times and that the model may not fully account for, such as weekend events or changes in the weather. On the other hand, there are also some periods where the model follows the actual passenger count quite well, for example, during early morning and evening hours.

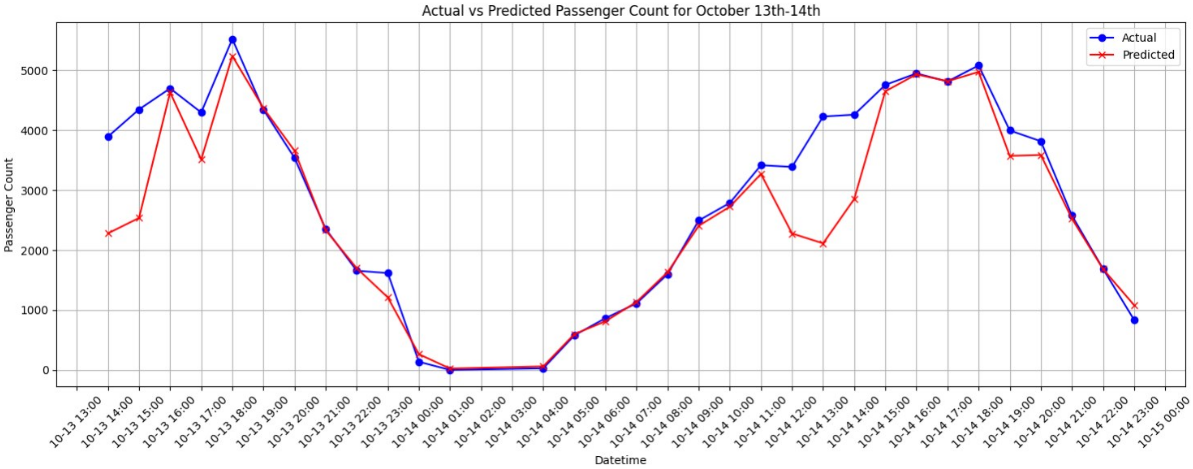


Figure 23: Actual versus predicted passenger count of trams for October 13th-14th, 2018 (weekend)

The close prediction and at the same time distinction between the predicted and actual counts throughout depending on the hour of the day suggests that the model can be a useful tool for predicting weekend tram ridership, however, it needs significant refinement.

5.3.4 Summary of Predictive Analysis

In the predictive analysis, the XGBoost models were deployed to discern the patterns in public transport ridership for buses, trolleybuses, and trams, taking into account the variations between weekends and weekdays. The models demonstrate high predictive accuracy during the training phase, indicating a strong grasp of the historical data patterns for all three transport modes. The testing phase results were quite promising for weekend ridership, with the models demonstrating a high level of accuracy. Conversely, the predictions for weekdays indicate that the models could benefit from additional refinement to address the increased complexity observed. Throughout the validation phase, the consistently strong performance on weekends reiterate the models' efficacy in capturing weekend patterns. However, it also underscores the necessity for enhanced model refinement specifically for tram ridership predictions on weekends. This need arises from the model's inaccuracies not only in predicting individual data points but also in forecasting the overall ridership trend during weekends. Overall. this pattern of stronger performance during weekends and more challenging weekday predictions echoes the results seen with the SARIMAX model in the time series analysis.

Visualisations play a pivotal role in interpreting the models' performance, displaying the predicted versus actual ridership counts. These graphs offer a straightforward comparison, emphasising where the models succeeded in mirroring actual trends and where discrepancies lay. Notably, the visual data often depicts a close alignment during off-peak hours but reveals deviations during peak periods, underscoring the potential need for model enhancements.

Overall, the predictive analysis using XGBoost provide valuable insights into the dynamics of urban mobility. While the models show potential utility for operational planning, the results indicate that further refinements could improve their applicability, especially for handling the intricate patterns of weekday ridership and also trams' weekend ridership. The combined quantitative and visual analyses illustrate the strengths and limitations of the models, forming a solid foundation for advancing public transport analytics.

5.4 Discussion

The comprehensive analysis undertaken in this study reveals interesting patterns in public transportation ridership in relation to various factors, including weather conditions, time of day, and types of days (weekdays, weekends, and holidays). Utilising both time series and predictive analysis models, such as SARIMAX and XGBoost, this research offers profound insights into the multifaceted nature of urban mobility.

The time series analysis shows how weather conditions distinctly influence ridership for buses, trolleybuses, and trams, addressing the hypotheses proposed. The hypothesis number 1 suggested investigating how hourly changes in weather conditions affect bus ridership, and how this effect varies between weekdays and weekends. For buses, the analysis reveals that temperature has a modest impact, leading to a slight increase in ridership on weekends. However, temperature changes do not significantly influence weekday ridership. Precipitation emerges as a significant factor, significantly decreasing ridership on both weekends and weekdays, indicating a strong deterrent effect of rain on bus usage during rainy weather. Atmospheric pressure shows a small negative effect on weekends but contributes positively to ridership on weekdays, suggesting nuanced variations in bus usage patterns depending on atmospheric conditions. Humidity appears to have a minor role, slightly boosting weekend ridership but lacking a significant effect on weekdays. Overall, these findings suggest that weather conditions, particularly precipitation, play a crucial role in influencing bus ridership, with variations observed between weekdays and weekends.

For hypothesis number 2, exploring the impact of hourly changes in weather conditions on trolleybus ridership also reveals significant trends. Temperature exhibits a considerable negative effect on both weekends and weekdays, indicating decrease in ridership during colder weather. Precipitation emerges as a significant deterrent, decreasing ridership notably on both weekends and weekdays. Atmospheric pressure influences weekend ridership negatively and has an insignificant positive effect on weekdays, suggesting potential weekday variations in trolleybus usage patterns in response to atmospheric conditions. While humidity does not significantly affect weekend ridership, it shows a significant negative impact on weekday ridership, implying that higher humidity levels deter people from using trolleybuses during weekdays.

For hypothesis number 3, examining the impact of hourly changes in weather conditions on tram ridership, temperature is found to decrease ridership on both weekends and weekdays, suggesting a preference for trams in cooler weather conditions. While precipitation significantly decreases weekend ridership, its effect on weekdays is not statistically significant, indicating that rain may primarily discourage tram usage on weekends. Atmospheric pressure negatively impacts weekend ridership but is positively associated with weekday ridership, suggesting potential differences in tram usage patterns depending on atmospheric conditions. Additionally, humidity increases ridership on weekdays but does not have a clear effect on weekend ridership, implying that humidity may influence tram usage differently throughout the week. These findings underscore the complex interplay between weather conditions and public transport ridership, highlighting distinct patterns between weekdays and weekends ridership of tram.

Overall, the weekend models for all three modes of transport demonstrate a more consistent pattern in relation to weather conditions than the weekday models, as indicated by the lower AIC values for weekends. This predictability on weekends stands in contrast to the complexity and unpredictability of weekday ridership, underscoring the need for a more detailed understanding of the interplay between weather conditions, holidays, and urban mobility. Furthermore, the differences in weather impacts between transport modes may reflect the unique ways in which riders use these services.

Additionally, one key finding is the significant impact of school and public holidays on ridership across all modes of transport, with a notable decrease in usage during these periods.

This outcome underscores the importance of academic and work schedules in shaping daily transportation patterns. The analysis also highlighted the variable impact of different factors on weekends versus weekdays. For instance, the predictive accuracy and model performance differed between these two time segments, suggesting varying degrees of predictability and underlying dynamics.

Nonetheless, the XGBoost models provide valuable insights into the ridership patterns of buses, trolleybuses, and trams in Bratislava. For buses, the model exhibits high accuracy in predicting weekend ridership but faces challenges with weekdays, where higher variability in data leads to less accurate predictions. Similarly, for trolleybuses, the model performs well in training but struggles with unseen data, particularly on weekdays. Tram ridership predictions show strong correlations with actual data on weekdays except for a few discrepancies during peak hours, however, struggle notably with weekend predictions. Additionally, the exploration of feature importance within the XGBoost models further elucidated the critical role of specific variables, such as temperature, atmospheric pressure, and time of day, in influencing ridership, while revealing that certain features like precipitation may not significantly impact the XGBoost models' outcomes. These findings highlight the complexities of predicting public transit ridership and underscore the need for continued refinement to enhance model accuracy and reliability across different days and modes of transportation.

In the realm of urban mobility, the influence of weather conditions on public transport ridership reveale intriguing and comparable results with findings from cities presented in the literature review. In Gipuzkoa, Spain, higher temperatures are associated with increased public transport usage, contrasting with Bratislava's observation of temperature changes showing no significant influence on bus ridership. Meanwhile, in Brisbane, Australia, apparent temperature negatively affects ridership, suggesting diverse responses to temperature across regions. Notably, while temperature changes in Bratislava did not significantly affect bus ridership, they exhibited a considerable negative effect on tram and trolleybus ridership. This underscores the importance of considering specific modes of public transportation and local climate nuances when analysing the impact of weather conditions on ridership patterns. These discrepancies highlight the need to account for local climate variations and infrastructure differences to better understand the relationship between weather conditions and public transportation usage.

Moreover, in Gipuzkoa, Spain, and New York City, rain decreases ridership, consistent with Bratislava's findings for buses. Conversely, in the Netherlands, rain increases ridership, contrasting with the observed deterrent effect in Bratislava. Bratislava's tram ridership is similarly affected by precipitation, with a significant decrease in weekend ridership but no statistically significant effect on weekdays. For trolleybuses, Bratislava's findings mirror those of buses, with precipitation significantly decreasing ridership on both weekends and weekdays. This is consistent with the observed decrease in ridership in Gipuzkoa and New York City due to rain. Overall, while Bratislava's findings align with those of Gipuzkoa and New York City regarding the deterrent effect of rain on bus and trolleybus ridership, the impact on tram ridership appears to be more pronounced on weekends. In contrast, the Netherlands demonstrates a unique response to rain, with increased ridership across all modes of public transportation. These comparisons highlight the complex and varied influences of precipitation on public transportation usage across different cities.

Humidity also has a nuanced influence on different modes of transportation. In Bratislava in the case of buses, humidity appears to have a minor role, slightly boosting weekend ridership but lacking a significant effect on weekdays. However, for trolleybuses, higher humidity levels show a significant negative impact on weekday ridership, suggesting that people may be deterred from using trolleybuses during humid weekdays. Similarly, in Shanghai, the findings mirror those of Bratislava for trolleybuses, where humidity shows a significant negative impact on ridership. Tram ridership, on the other hand, increases on weekdays with higher humidity levels, indicating a potentially positive effect.

6. Managerial Implications, Limitations and Further Research

This section presents managerial implications and actionable strategies for urban transportation planning and management. By embracing data-driven decision-making, transportation authorities can anticipate changes in ridership patterns, leading to more responsive and efficient public transit services. Despite the valuable insights provided, the study has limitations that warrant consideration and, therefore, will also be discussed further. Lastly, avenues for further research stemming from the findings of the study on public transportation ridership in Bratislava will be proposed.

6.1 Managerial Implications

The findings of this study offer valuable insights for urban transportation planning and management, highlighting the importance of adaptive and responsive strategies to meet the dynamic demands of public transit users. One practical application of these insights is in adaptive scheduling, where transportation authorities can optimise scheduling and fleet management to accommodate significant shifts in ridership identified during school, public holidays and different weather conditions on weekdays or weekends. By tailoring service levels to anticipated demand, operational efficiency and passenger satisfaction can be significantly enhanced.

Moreover, the influence of weather conditions on ridership suggests that transit systems could benefit from developing weather-responsive operational strategies. This approach could encompass flexible scheduling, enhanced communication with riders regarding weather-related service adjustments, and the deployment of additional resources during adverse weather conditions to ensure service reliability.

An understanding of temporal and weather-related ridership patterns also facilitates more effective resource allocation. This could involve maintaining or enhancing service by deploying longer or additional vehicles during peak times, as identified through hourly ridership analysis, while also considering the reallocation of resources during off-peak hours or less busy periods.

Furthermore, the study’s insights into preferences for different modes of transport and the impact of specific variables on ridership can guide infrastructure development and service improvements. For example, improving the attractiveness and comfort of modes that show higher usage potential under certain conditions, like trams during extreme temperatures, could encourage more people to use specific public transport types.

Finally, the methodologies and findings from this study underscore the importance of data-driven decision-making in public transport management. By incorporating predictive analytics into planning and operational decisions, transportation authorities can better anticipate changes in ridership patterns, leading to improved service responsiveness and efficiency. This comprehensive approach to transportation planning and management not only addresses current challenges but also sets the stage for more sustainable and user-friendly public transit systems in the future.

In summary, this study not only enriches the academic understanding of factors influencing public transportation usage but also provides actionable insights for transportation authorities and city planners of Bratislava. By incorporating these findings into strategic planning and operational adjustments, the city can better meet the dynamic needs of urban populations, fostering more efficient, reliable, and user-friendly transportation systems.

6.2 Limitations and Further Analysis

In section 5, XGBoost models’ predictions of passenger count for buses, trolleybuses and trams separately for weekdays and weekends were analysed. It could be observed that some data points were not predicted very accurately and deviated from the line of perfect prediction, specifically in scatterplots in Figures 15, 18 and 21. Therefore, the study was extended by an example analysis in which the top five instances with the largest discrepancies between actual and predicted counts were extracted. This was done separately for each trained XGBoost model. Then, to further explore potential causes for these inaccuracies, any major events in Bratislava on the corresponding days and times that might have affected bus ridership and hence the model's predictions were searched for online. However, no documented events that could account for the discrepancies were discovered. Information about specific events that might explain the model's poor performance could have been lost or might no longer be available online. For a more accurate and timely analysis, such an investigation should ideally be conducted with more structured and recent data, ensuring that the potential impact of such events on ridership predictions can be accurately assessed.

Consequently, one constraint of this research is the potential influence of external events that remain unexplored, particularly because the dataset pertains to the year 2018, rendering such events increasingly difficult to investigate retrospectively. Social happenings like concerts, sports fixtures including football and hockey matches, and various other events could substantially affect public transport ridership at specific times. The absence of this data introduces a notable limitation, as these occurrences can significantly alter ridership patterns and, therefore, the predictive accuracy of the model under study.

Another limitation to consider within the scope of this study is the range of weather data at disposal. The dataset incorporated variables such as temperature, precipitation, humidity, and atmospheric pressure. Nevertheless, these parameters represent only a subset of the myriad weather conditions that could influence public transport ridership. It stands to reason that additional meteorological factors if included, might significantly refine the predictive capabilities of the model. Factors such as wind speed, visibility or cloudiness could potentially have a pronounced effect on transit usage. Incorporating a more comprehensive suite of weather data could therefore be an invaluable extension to the study, potentially yielding a model that more accurately mirrors the complexities of public transport dynamics.

Unfortunately, detailed information regarding the presence or absence of air conditioning in each vehicle type (bus, tram, or trolleybus) was not available for the year 2018, and even the city representatives lacked this level of granularity. Consequently, the study did not account for the potential impact of air conditioning on public transport ridership in the analysis. However, it is worth noting that considering this factor could be beneficial for future research, particularly in more recent years, to ascertain whether it indeed influences ridership patterns. Integrating such data could provide valuable insights into the interplay between comfort amenities and public transport usage, thereby enriching our understanding of factors shaping transportation preferences and behaviours.

Furthermore, the study presents an opportunity for expansion by examining the transportation choices on days and times when public transport ridership is lower. It would be instructive to determine whether individuals shift to alternative modes of transportation such as cars, bicycles, electric scooters, or even walking. Additionally, it is conceivable that these preferences may vary according to different areas of the city or between various types of transportation. An in-depth analysis of these patterns could reveal nuanced behaviours and preferences, further enriching our understanding of urban mobility and the factors influencing the selection of transport modes.

Additionally, the study was limited by the scope of weather data, which was sourced from just two SHMU weather stations located nearby, yielding highly similar datasets. Consequently, data exclusively from the station nearer to the city centre was used only. However, should data become available from a diverse array of weather stations dispersed throughout the city, there would be an opportunity for a more granular analysis. With a broader geographical spread of data points, it would be possible to conduct a spatial analysis, potentially unveiling more detailed correlations between weather conditions and public transport ridership across different city sectors. This could significantly enhance the predictive model by accounting for microclimatic variations within the urban environment.

Last but not least, there exists a promising research potential for a longitudinal study through a five-year comparison. By applying the data preparation and cleaning methodologies that I have already established, it would be feasible to adapt the existing script to process the 2023 dataset in a similar manner. This would set the stage for a comparative analysis between the two years, offering insights into the evolution of public transport ridership behaviours. Such a comparison could reveal significant shifts over time, potentially influenced by a variety of factors, including the long-term impacts of the COVID-19 pandemic, changes in workplace policies promoting remote work, and other societal transformations. Understanding these trends could prove invaluable in forecasting future public transport needs and in planning for sustainable urban transit systems.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. doi:10.1177/1094428112470848

Ali, Z. A., Abduljabbar, Z. H., Tahir, H. A., Sallow, A. B., & Almufti, S. M. (2023). Extreme gradient boosting algorithm with Machine Learning: A Review. *Academic Journal of Nawroz University*, *12*(2), 320–334. doi:10.25007/ajnu.v12n2a1612

Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. *Transportation Research Part A: Policy and Practice, 59,* 1–12. https://doi.org/10.1016/j.tra.2013.10.019

Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of Student Performance Based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. doi:10.1080/10494820.2021.1928235

Bratislava - Characteristic of the Region. (2024). Retrieved from https://shorturl.at/cdL68

Concept for the Development of Public Urban Transport in Bratislava for the Years 2013-2025. (2016). Retrieved from https://cdn-api.bratislava.sk/strapi-homepage/upload/Koncepcia_rozvoja_MHD_2013_2025_b46d95b779.pdf

Corcoran, J., & Tao, S. (2017). Mapping spatial patterns of bus usage under varying local temperature conditions. *Journal of Maps, 13(1),* 74–81. https://doi.org/10.1080/17445647.2017.1378933

Elavarasan, D., & Vincent, D. R. (2020). Reinforced XGBoost machine learning model for Sustainable Intelligent Agrarian Applications. *Journal of Intelligent &amp; Fuzzy Systems*, *39*(5), 7605–7620. doi:10.3233/jifs-200862

Faško, P., Lapin, M., & Pecho, J. (2008). 20-year extraordinary climatic period in Slovakia. *Meteorol. Časopis*, *11*, 99-105.

Historical Maps of Public Transport Lines. (2018). Retrieved from https://imhd.sk/ba/mapa-schema/1907/Historick%C3%A9-mapy-liniek-MHD-a-IDS-10-2-2018

Karim, A. A., Pardede, E., & Mann, S. (2023). A model selection approach for time series forecasting: Incorporating Google Trends data in Australian Macro Indicators. *Entropy*, *25*(8), 1144. doi:10.3390/e25081144

Kumari, S. S. (2008). Multicollinearity: Estimation and elimination. *Journal of Contemporary research in Management*, *3*(1), 87-95.

Lückerath, D., Streberová, E., Bogen, M., Rome, E., Ullrich, O., & Pauditsová, E. (2019). Climate change impact and vulnerability analysis in the city of bratislava: Application and lessons learned. *Critical Information Infrastructures Security*, 83–94. doi:10.1007/978-3-030-37670-3_7

Mishra, S., Welch, T. F., & Jha, M. K. (2012). Performance indicators for public transit connectivity in multi-modal Transportation Networks. *Transportation Research Part A: Policy and Practice*, *46*(7), 1066–1085. doi:10.1016/j.tra.2012.04.006

Müller-Plath, G., & Lüdecke, H.-J. (2024). Normalized coefficients of prediction accuracy for Comparative Forecast Verification and modeling. *Research in Statistics*, *2*(1). doi:10.1080/27684520.2024.2317172

Nissen, K. M., Becker, N., Dähne, O., Rabe, M., Scheffler, J., Solle, M., & Ulbrich, U. (2020). How does weather affect the use of public transport in Berlin? *Environmental Research Letters*, *15*(8). doi:10.1088/1748-9326/ab8ec3

Onur, S. G., Altin, K. T., Yurtseven, B. D., Haznedaroglu, E., & Sandalli, N. (2020). Children's drawing as a measurement of dental anxiety in Paediatric Dentistry. *International Journal of Paediatric Dentistry*, *30*(6), 666–675. doi:10.1111/ipd.12657

Provost, F., & Fawcett, T. (2013). Chapter 2: Business Problems and Data Science Solutions. In *Data Science for Business* (1st ed., pp. 26–34). essay, Sebastopol, California: O'Reilly Media.

Ribeiro, S. M., & Castro, C. L. (2022). Missing data in time series: A review of imputation methods and case study. *Learning and Nonlinear Models*, *20*(1), 31–46. doi:10.21528/lnlm-vol20-no1-art3

Sabir, M. (2011). Weather and travel behaviour. [PhD thesis, VU University].

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and Random Forest. *SN Applied Sciences*, *2*(7). doi:10.1007/s42452-020-3060-1

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia &amp; Analgesia*, *126*(5), 1763–1768. doi:10.1213/ane.0000000000002864

Senthilnathan, S. (2019). Usefulness of correlation analysis. *SSRN Electronic Journal*. doi:10.2139/ssrn.3416918

Shah, S., Houda, M., Khan, S., Althoey, F., Abuhussain, M., Abuhussain, M. A., … Javed, M. F. (2023). Mechanical behaviour of e-waste aggregate concrete using a novel machine learning algorithm: Multi Expression Programming (MEP). *Journal of Materials Research and Technology*, *25*, 5720–5740. doi:10.1016/j.jmrt.2023.07.041

Singh, D. K., & Rawat, N. (2023). Machine learning for weather forecasting: XGBoost vs SVM VS random forest in predicting temperature for Visakhapatnam. *International Journal of Intelligent Systems and Applications*, *15*(5), 57–69. doi:10.5815/ijisa.2023.05.05

Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice, 69,* 379–391. https://doi.org/10.1016/j.tra.2014.09.008

Slovak Republic in Figures 2018. (2018). Retrieved from https://shorturl.at/R0269

Slovak Republic in Figures 2023. (2023). Retrieved from https://shorturl.at/wzFMW

Stover, V., & McCormack, E. (2012). The impact of weather on bus ridership in Pierce County, Washington. *Journal of Public Transportation, 15(1),* 95–110. https://doi.org/10.5038/2375-0901.15.1.6

Tao, S., Corcoran, J., Rowe, F., & Hickman, M. (2018). To travel or not to travel: 'weather' is the question. modelling the effect of local weather conditions on bus ridership. *Transportation Research Part C: Emerging Technologies*, *86*, 147–167. doi:10.1016/j.trc.2017.11.005

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, *10*, 102119. doi:10.1016/j.mex.2023.102119

Thiese, M. S., Ronna, B., & Ott, U. (2016). P value interpretations and considerations. *Journal of Thoracic Disease*, *8*(9). doi:10.21037/jtd.2016.08.16

Traffic Index 2018. (2018). Retrieved from https://traffic-index-docs.s3-eu-west-1.amazonaws.com/TomTomTrafficIndex-Ranking-2018-full.pdf

Výberči, J., Pecho, J., Faško, P., & Bochníček, J. (2018). Warm and cool spells in Slovakia (1951–2017) in the context of climate change. *Meteorological journal*, *21*(2018), 101-108.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin &amp; Review*, *11*(1), 192–196. doi:10.3758/bf03206482

XGBoost. (2022a). *Introduction to Boosted Trees*. Retrieved January 13, 2024, from https://xgboost.readthedocs.io/en/stable/tutorials/model.html

XGBoost. (2022b). *XGBoost Parameters*. Retrieved January 13, 2024, from https://xgboost.readthedocs.io/en/stable/parameter.html

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies*, *75*, 17–29. doi:10.1016/j.trc.2016.12.001

Appendix

Figure A: Public transport scheme effective from February 10th, 2018 (*Historical Maps of Public Transport Lines,* 2018)

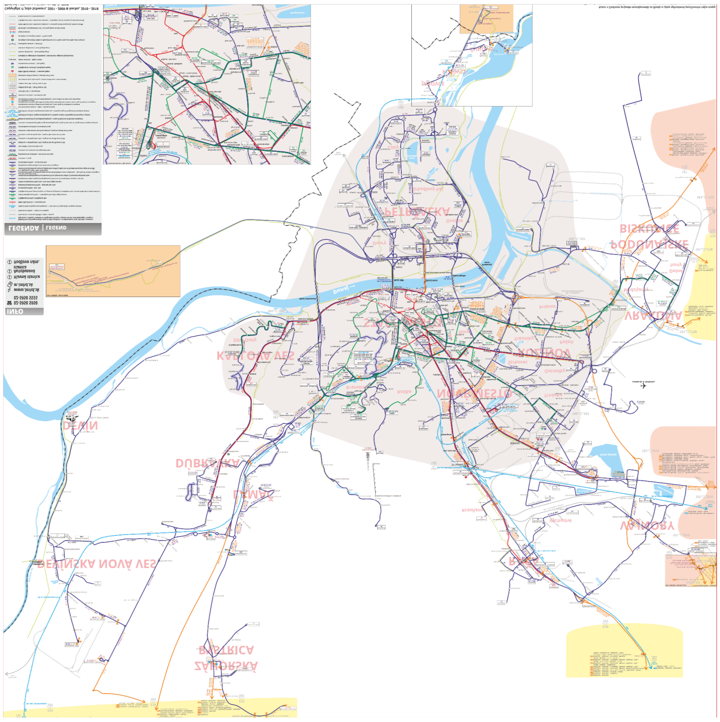


Figure B: Overall monthly ridership patterns drilled down by vehicle type

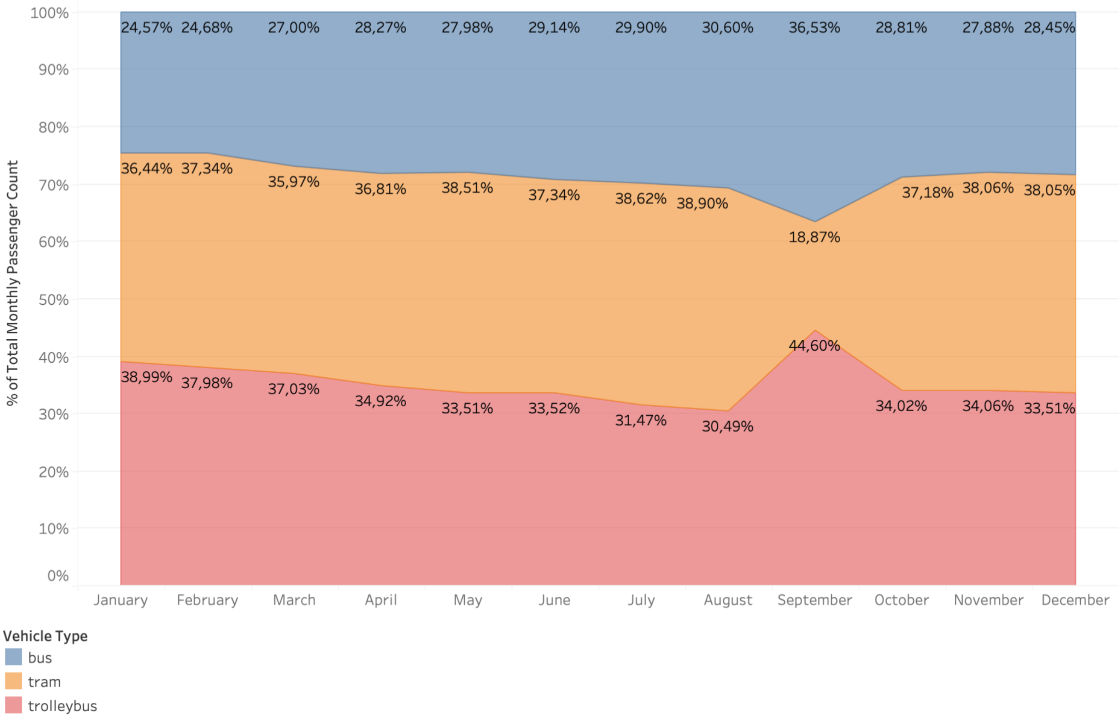


Figure C: Overall average monthly ridership patterns (excluding public holidays)

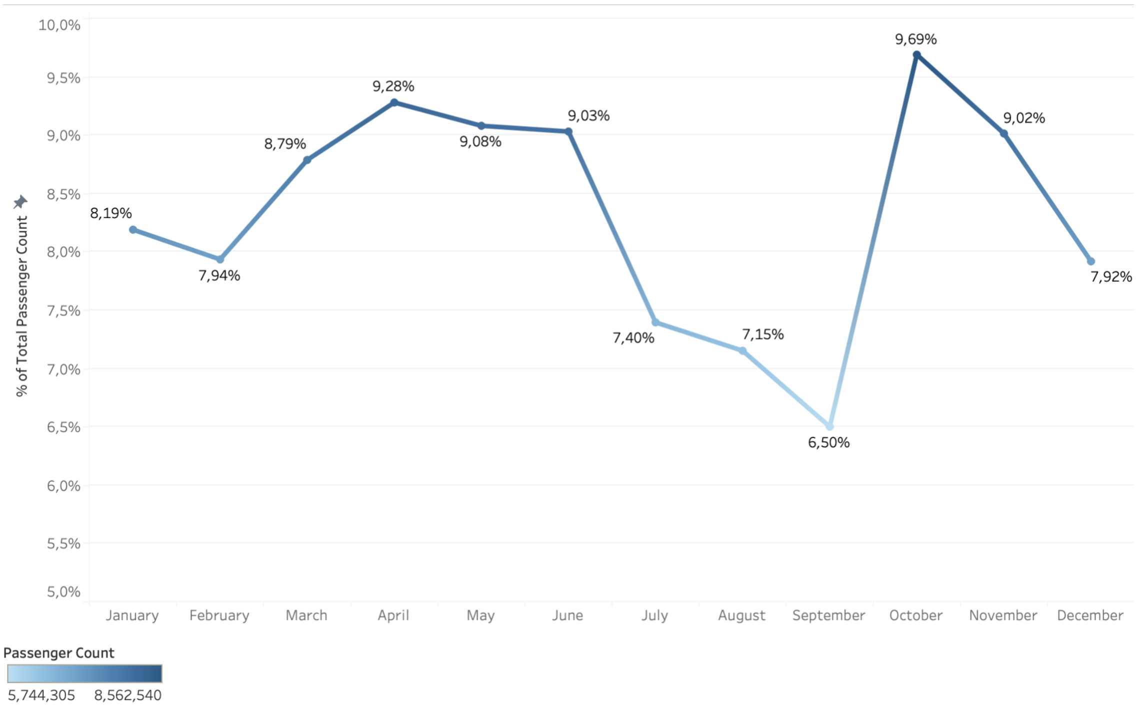


Figure D: Monthly usage on weekdays broken down by vehicle type

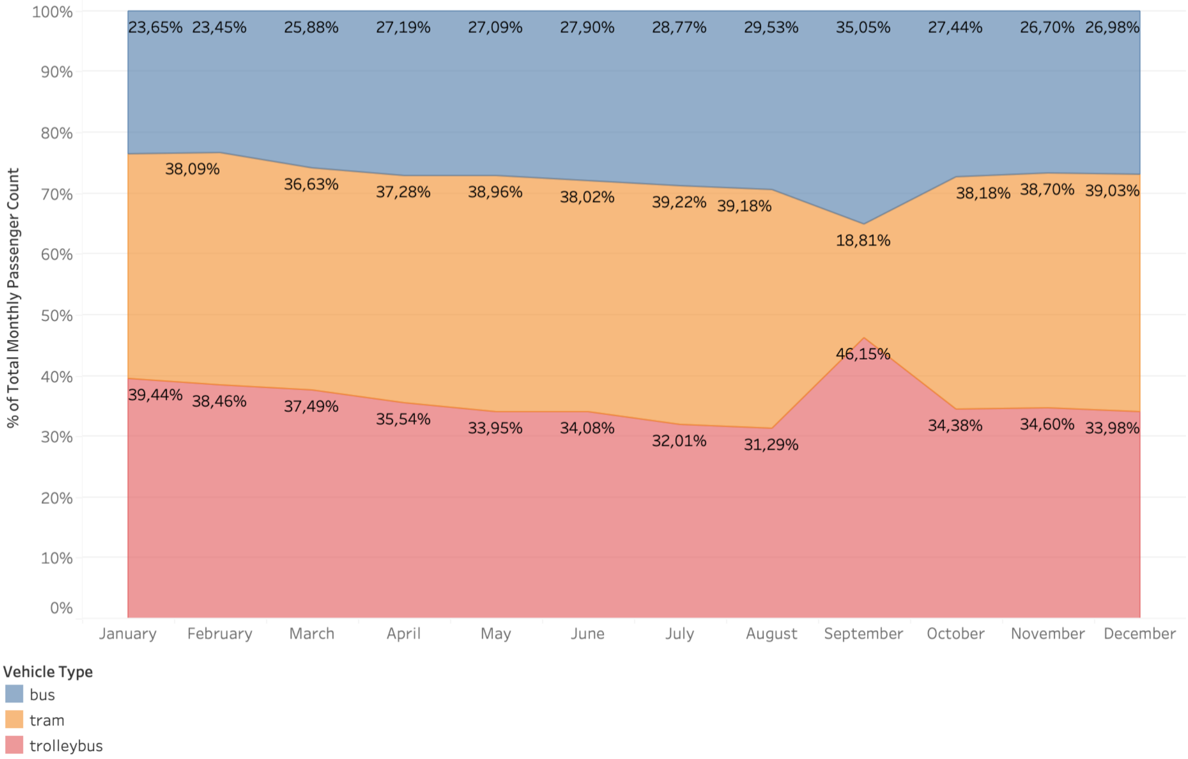


Figure E: Monthly ridership patterns on weekends drilled down by vehicle type

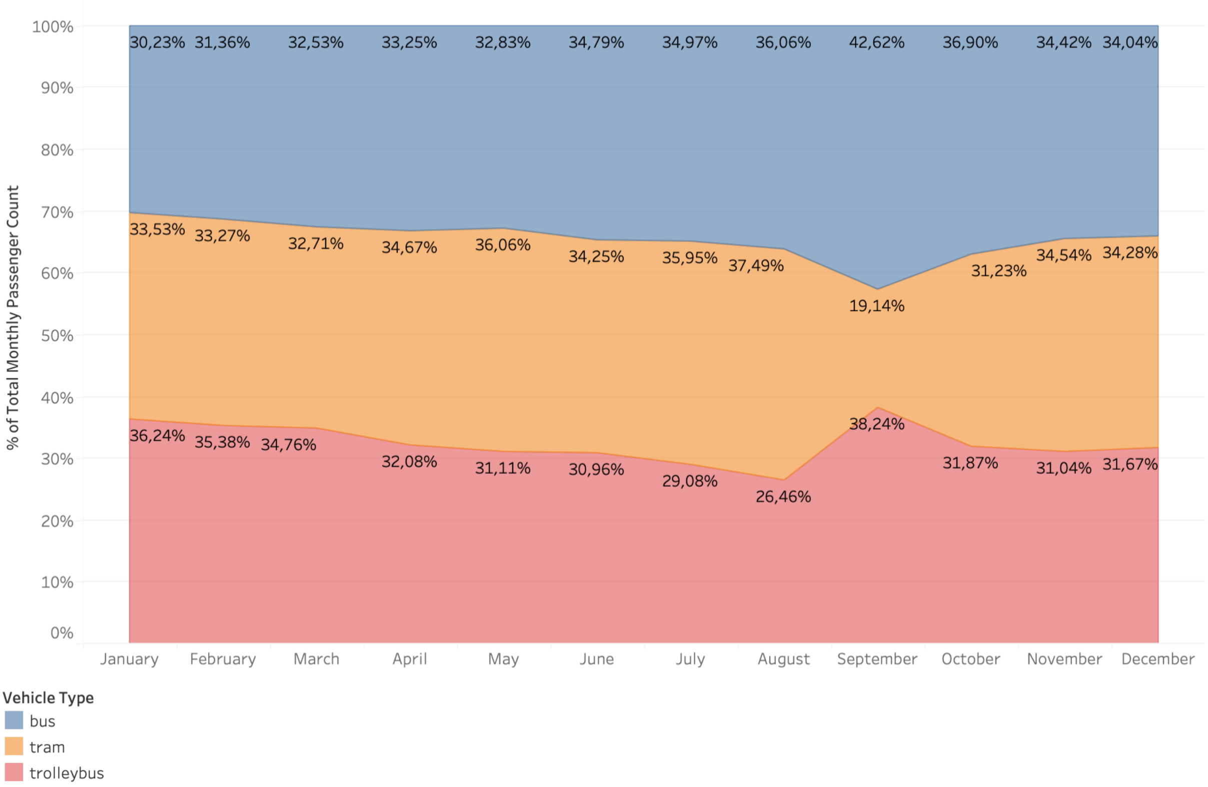


Figure F: Average hourly ridership patterns of all vehicle types

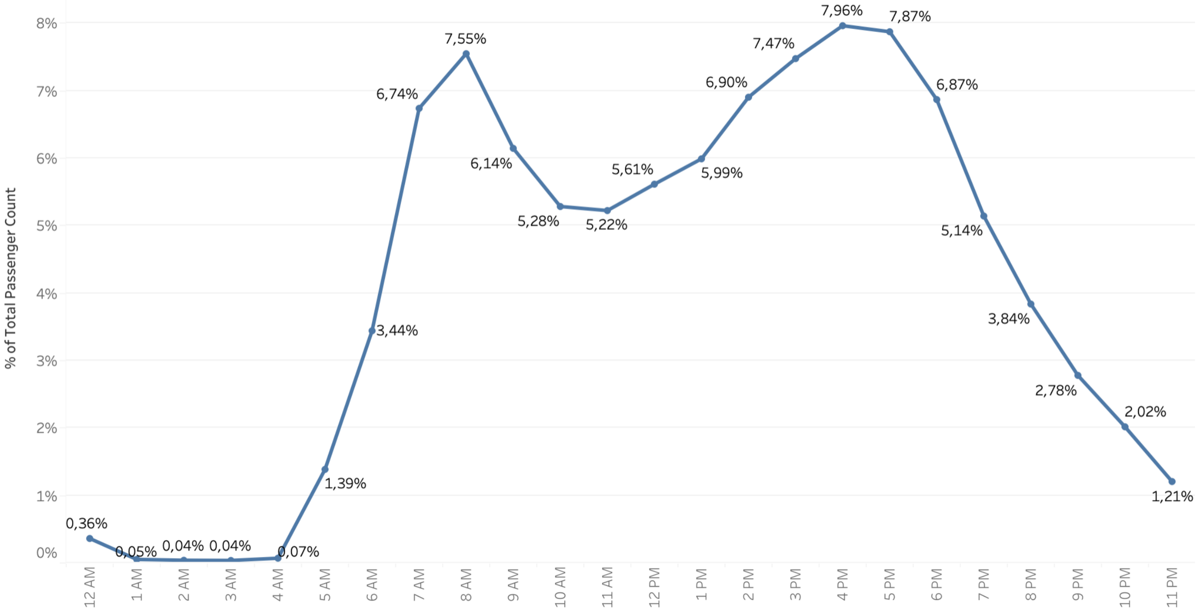


Figure G: Average hourly ridership patterns of all vehicle types (excluding public holidays)

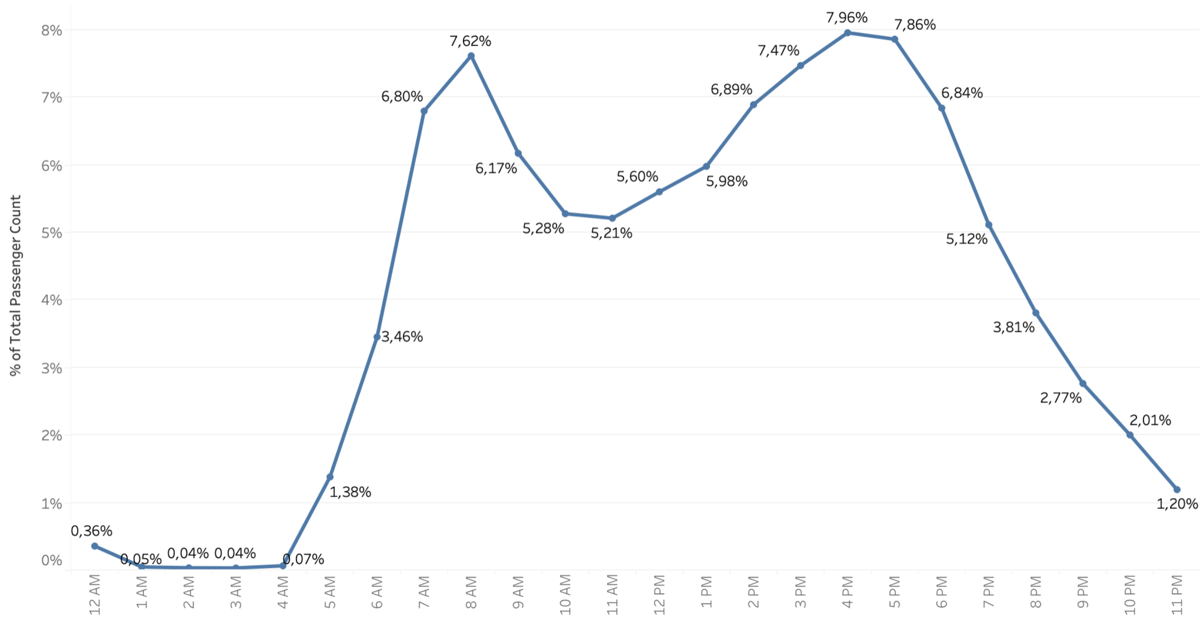


Figure H: Average hourly ridership patterns drilled down by vehicle type

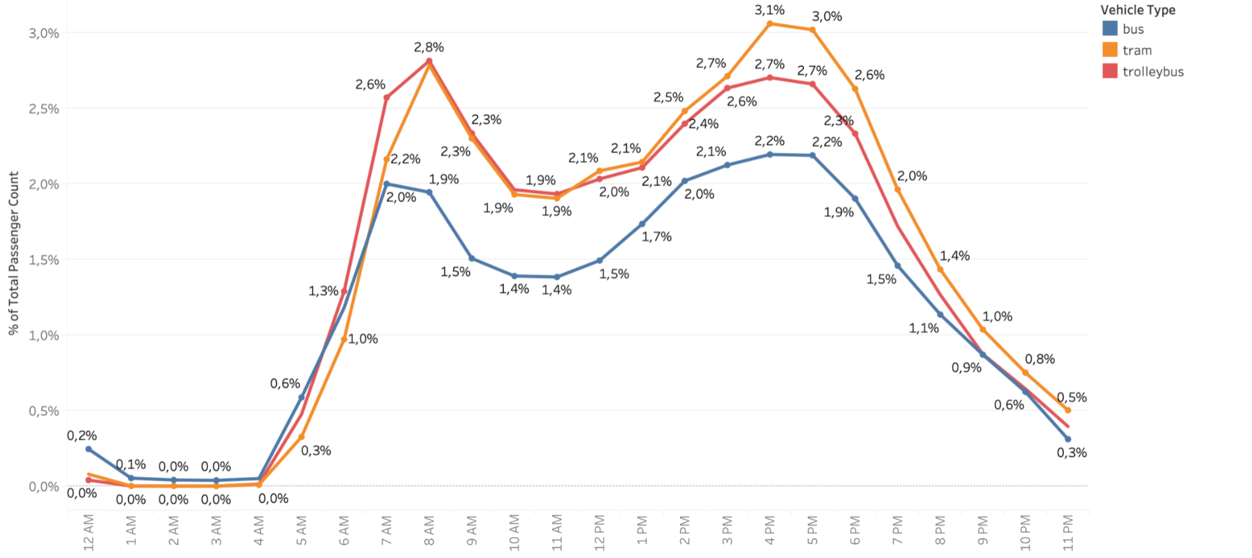


Figure I: Different vehicle type usage

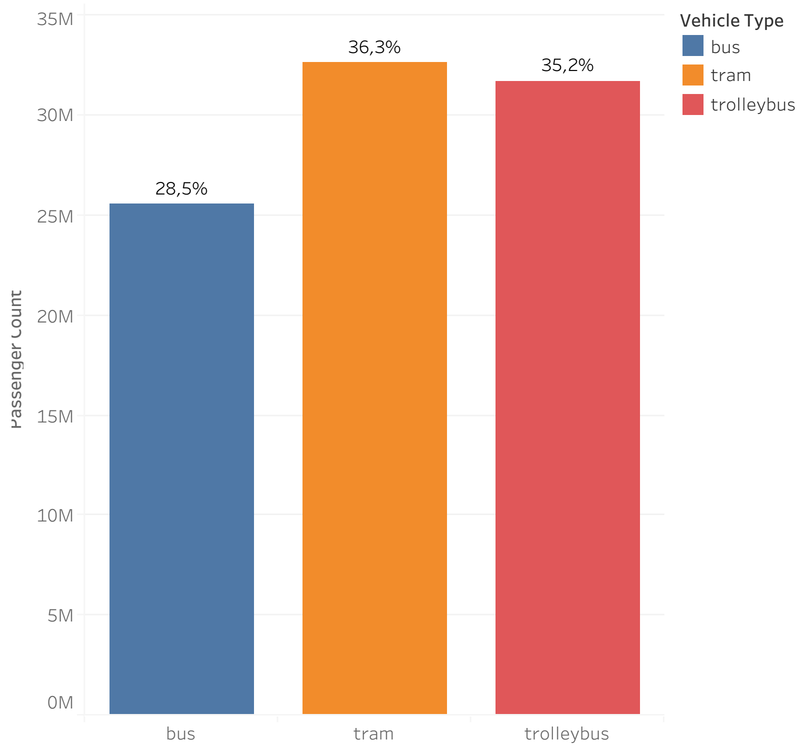


Figure J: Different vehicle type usage (excluding public holidays)

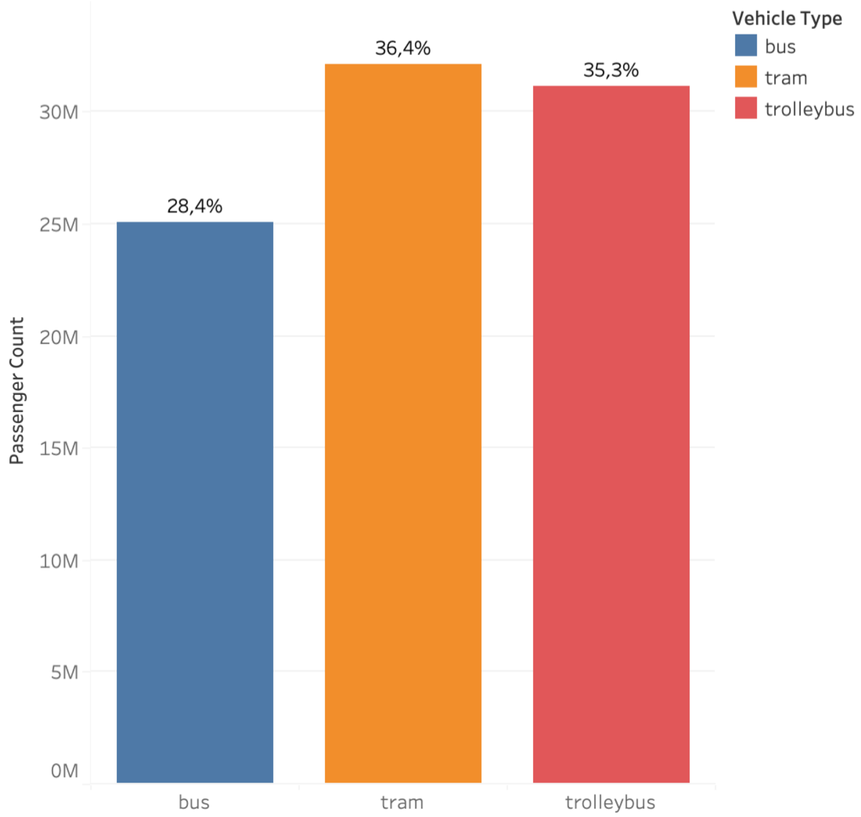


Figure K: Different vehicle type usage during public holidays

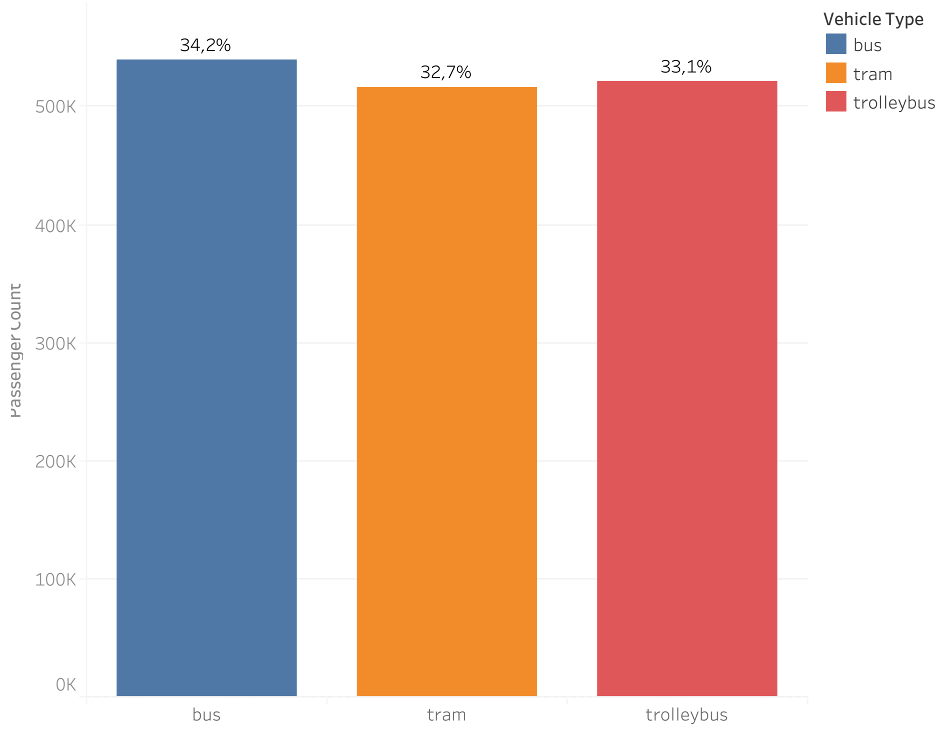
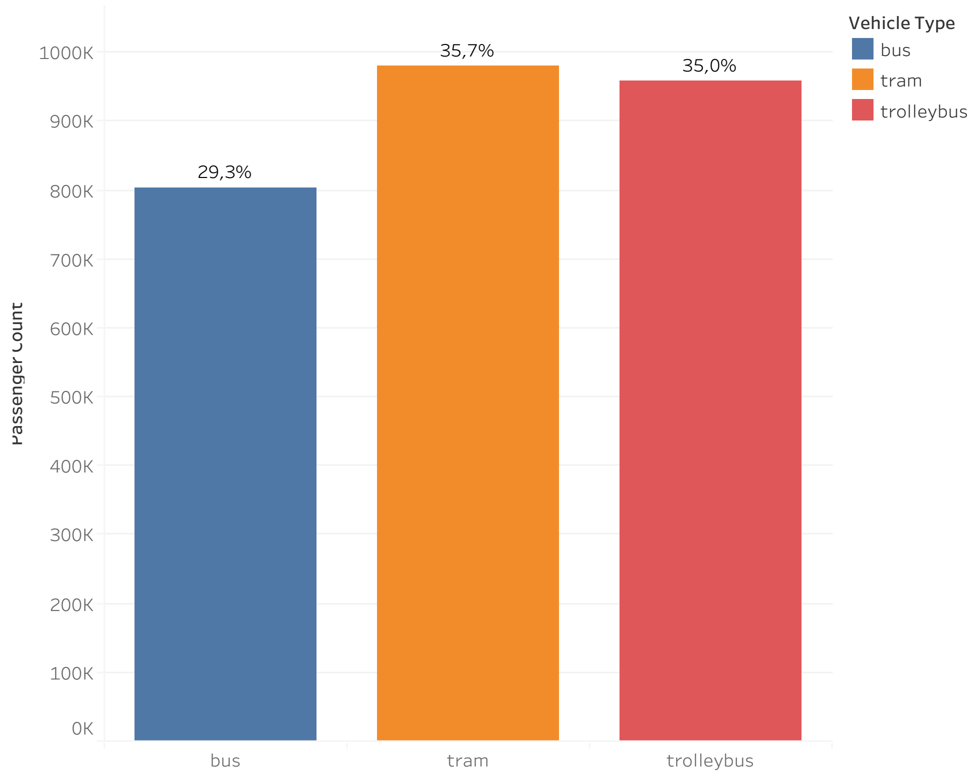


Figure L: Different vehicle type usage one day before public holidays


Vehicle Type
🟦 bus
🟧 tram
🟥 trolleybus
Passenger Count
1000K
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
bus
tram
trolleybus