**PUBLIC TRANSPORTATION ANALYSIS**

**PHASE 5: PROJECT DOCUMENTATION AND SUBMISSION**

**PROJECT OBJECTIVE:**

Public transportation systems play a pivotal role in urban areas, ensuring the seamless movement of people. It is an efficient mode of travel due to its ability to carry a large number of passengers at once. By utilizing dedicated lanes or routes, it can bypass traffic congestion, ensuring a more reliable and timely journey. Moreover, advancements in technology have enabled real-time tracking and scheduling systems, further enhancing the efficiency of public transit.

However, various challenges affect the efficiency and quality of these services. Timeliness, passenger satisfaction, and operational effectiveness are crucial aspects that demand continuous evaluation and improvement. Delays, overcrowding, and passenger dissatisfaction can lead to decreased ridership and affect the overall urban mobility experience. This project aims to address these challenges by analyzing public transportation data comprehensively. By focusing on on-time performance, passenger feedback, and service efficiency, we intend to identify key bottlenecks, assess customer experience, and propose data driven strategies. Through this analysis, our goal is to enhance the overall quality of public transportation, making it more reliable, convenient, and passenger-friendly.

**DESIGN THINKING PROCESS:**

**ANALYSIS OBJECTIVES:**

**1.On-Time Performance**

Define specific objectives for analyzing public transportation data such as assessing on-time performance. One of the primary objectives is to evaluate the on-time performance of public transportation services. We will measure and report the percentage of services that adhere to their schedules.

**2. Efficiency:**

Identify objectives for analyzing the efficiency of public transportation services. To determine the efficiency of public transportation services, we will assess factors such as route optimization, vehicle utilization, and punctuality.

**3. Passenger Satisfaction:**

Assess passenger satisfaction through data analysis. Another key objective is to gauge passenger satisfaction. This will involve the collection and analysis of passenger feedback through surveys or other available data sources

**DATA COLLECTION PROCESS:**

In order to analyze public transportation data, we need to identify trustworthy sources and methods for collecting transportation data. These sources could include schedules, real-time updates, and passenger feedback.

**1.Schedules Data:**

We will collect schedules data from the provided dataset. This data will include information about planned departure and arrival times, routes, and stops.

**2.Real-time Updates:**

Real-time data will be gathered to track actual departure and arrival times, allowing us to measure on-time performance accurately.

**3.Passenger Feedback:**

Passenger feedback will be collected through surveys or online platforms, if available. This data will provide insights into passenger satisfaction and areas for improvement.

**4.Weather Data:**

Weather data may also be considered to understand its impact on service efficiency and delays.

**DATA VISUALISATION:**

To effectively communicate insights from our analysis, we need a plan for visualizing the data. IBM Cognos is an excellent tool for creating informative dashboards and reports.

**1.IBM Cognos Dashboards:**

We will use IBM Cognos to design informative dashboards and reports. These dashboards will include visualizations such as line charts for tracking on-time performance trends, bar charts for comparing passenger satisfaction across different routes, and geographic maps to visualize service efficiency based on location.

**2.Interactive Reports:**

Interactive reports will allow stakeholders to drill down into specific details, making it easier to identify areas that require improvement.

**3.Key Performance Indicators (KPIs):**

We will present KPIs like on time percentage, passenger satisfaction scores, and service efficiency indices prominently on the dashboards.

**CODE INTEGRATION:**

**1.Data Cleaning:**

Code will be used to clean and preprocess the raw transportation data. This may include handling missing values, standardizing data formats, and removing outliers. Clean the data to ensure accurate, unbiased analysis results.

**2.Data Transformation:**

Code will be employed to transform data into a format suitable for analysis, including merging data from different sources and creating derived variables for deeper insights. Transform the data into a more useful format for further analysis.

**3.Statistical Analysis:**

Advanced statistical analysis, if necessary, will be conducted using code to identify correlations, trends, and potential areas for optimization. Use code to perform statistical analysis and discover meaningful insights.

**INNOVATION:**

**DESIGN AND INNOVATION STRATEGIES:**

Implementing "Gender-Responsive Transportation" could be a creative way to improve public transportation effectiveness while addressing gender-related issues. According to this idea, transportation services would be planned and designed to take differing travel preferences and safety issues for men and women into account. This can entail offering distinct but equal services at particular times to cater to the needs of both genders, ensuring that all passengers travel safely and comfortably. Such a strategy might aid in boosting the number of passengers as well as the general public's opinion of the safety and inclusivity of public transportation networks.

**Data Collection and Feature Engineering**

**Innovation: Comprehensive Data Gathering**

Implement advanced web scraping techniques and leverage real estate APIs to collect diverse datasets encompassing property features, location data, market trends, and historical price data. Apply innovative feature engineering techniques, such as text summarization for property descriptions, to extract meaningful information from both structured and unstructured data. Collect and analyze passenger feedback and sentiment data from sources like social media, surveys, and customer support interactions

**Data Pre-processing**

**Innovation: Natural Language Processing (NLP) for Sentiment Analysis**

Utilize Natural Language Processing (NLP) techniques to pre-process textual data, including passenger feedback and comments. Develop a custom NLP pipeline that includes tokenization, lemmatization, sentiment analysis, and topic modelling to extract valuable insights from passenger comments. Handle missing data with innovative methods, such as imputation based on historical patterns and feedback from similar situations.

**Model Selection and Training**

**Innovation: Machine Learning and Deep Learning Integration**

Employ a combination of machine learning algorithms, such as Random Forests, Support Vector Machines, and XGBoost, for service disruption prediction. Incorporate deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to analyze temporal patterns and passenger sentiment in textual data. Develop an ensemble model that combines the strengths of machine learning and deep learning approaches for more accurate predictions.

**Model Interpretability and Visualization**

**Innovation: Explainable AI (XAI):**

Incorporate Explainable AI techniques such as SHAP values and LIME to provide transparent explanations for model predictions. This helps stakeholders understand the rationale behind efficiency assessments and recommendations. Develop an interactive dashboard with visualizations that showcase key performance indicators, route efficiency scores, and passenger sentiment trends. This user-friendly interface ensures that stakeholders can easily access and interpret the analysis results.

**Data Intergration with IBM cognos**

Use IBM Cognos' data integration capabilities to combine and merge data from different sources into a unified dataset. This often requires using ETL (Extract, Transform, Load) processes.

**Public Transporatation efficiency reports and dashboards**

The development of simple-to-understand visual representations of critical data is required for creating reports and dashboards for public transportation efficiency. These tools offer quick, high-level insights into things like riding patterns, punctuality, maintenance requirements, and fuel usage. They enable transportation authorities to evaluate the system's condition swiftly, pinpoint areas that require repair, and make data-driven choices. Reports and dashboards enable stakeholders to improve service quality, optimize routes, cut costs, and guarantee a more effective and dependable public transportation system for commuters by making data easily understandable and accessible.

**Deploying and monitoring the model**

Deploying and monitoring a Public Transportation Efficiency model involves implementing it within the transit system's infrastructure. This includes integrating data sources, setting up real-time monitoring, and establishing alerts for anomalies. Regularly assessing the model's performance ensures it continues to provide accurate insights for route optimization, cost reduction, and improved service quality. Effective deployment and ongoing monitoring are essential to sustain and enhance public transportation efficiency, benefiting both commuters and the transportation authorities

**Continuous Improvement and Feedback Loops**

**Innovation: Feedback Mechanisms:**

Establish mechanisms for continuous feedback from passengers, transit staff, and city officials. This feedback loop will allow for ongoing adjustments and improvements to the public transport system. By incorporating these design and innovation strategies, Public Transport Efficiency Analysis can become a dynamic and data-driven process that leads to more effective, user-centric, and sustainable public transportation systems.

**Innovation: Model Maintenance and Improvement**

Establish a continuous learning framework that adapts to changing conditions and passenger preferences. Regularly retrain the models using new data to improve prediction accuracy and sentiment analysis. Implement automated data pipelines for seamless data ingestion, model retraining, and feedback incorporation.

**DEVELOPMENT PHASES:**

**Data Exploration and Understanding**

* + Load the dataset using Pandas.
  + Our focus will be on understanding the dataset's structure, consisting of 6 columns: TripID, RouteID, StopID, StopName, WeekBeginning, and NumberOfBoardings and understand the column meanings, and potential relationships between variables.
  + Identify data quality issues, missing values, and outliers.

**Data Preprocessing**

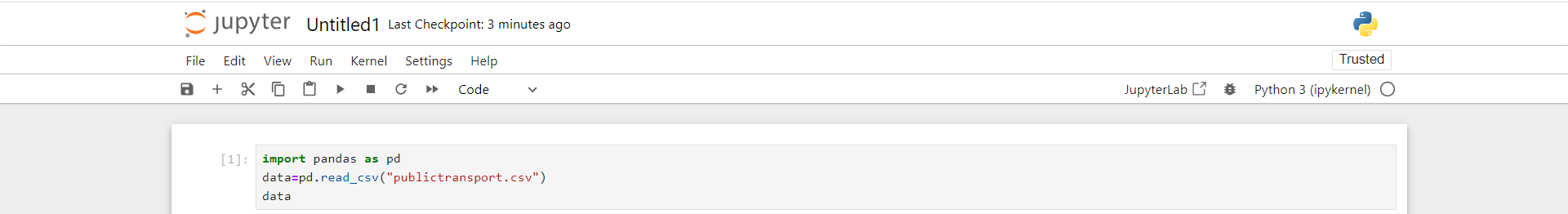
* + Select relevant columns for analysis (e.g., TripID, RouteID, StopName).
  + Handle missing data, duplicates, and irrelevant entries.
  + Convert data types if needed

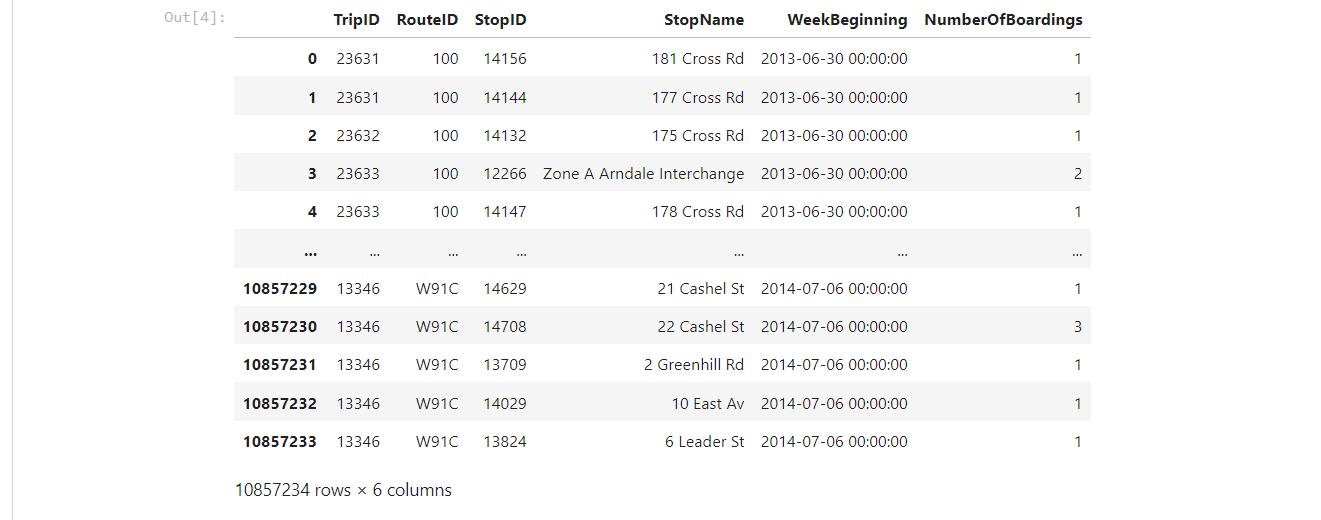
**Predicting Service Disruptions**

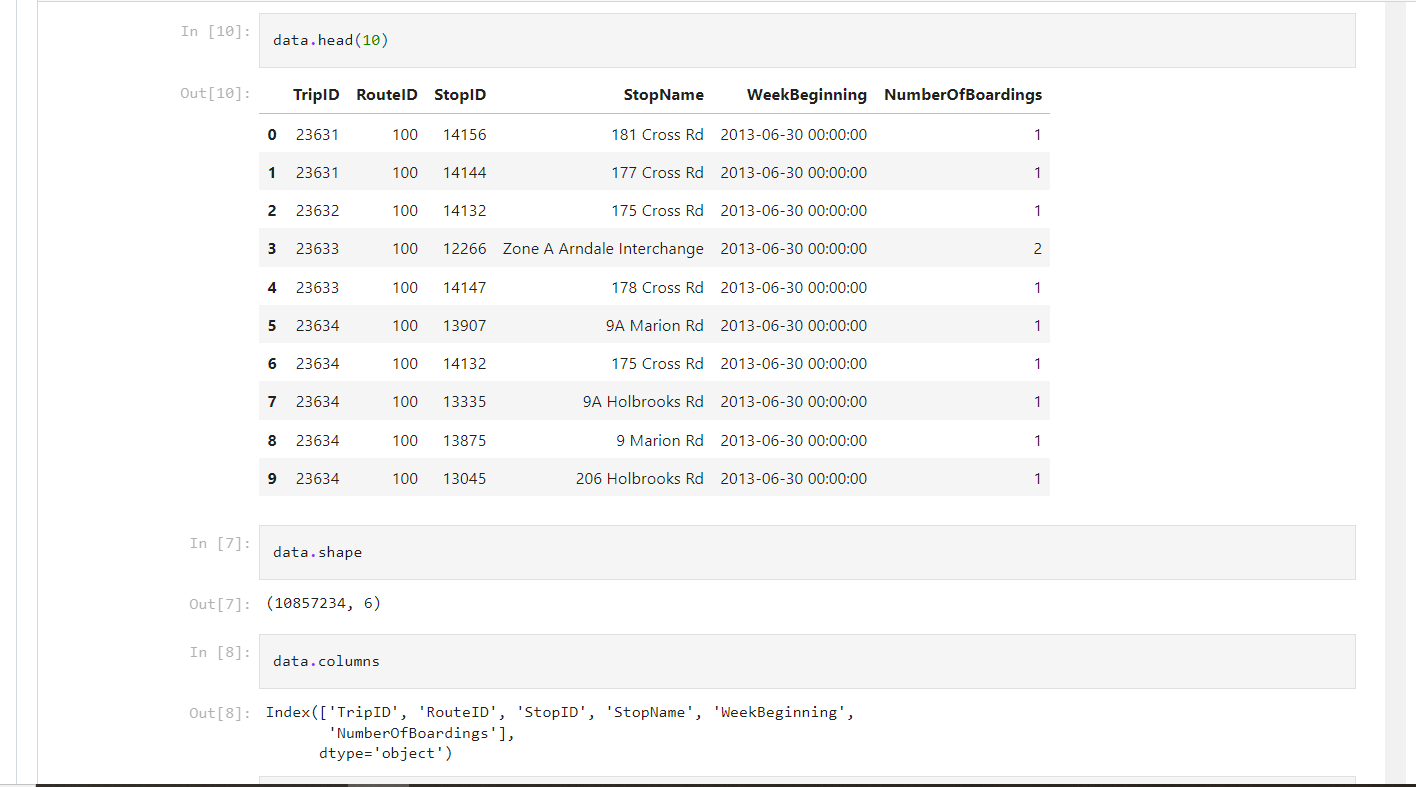
* + Innovation: Define how service disruption is determined from given features
  + Select a set of features and Service Disruption as target feature
  + Create DecisionTreeClassifier and train on 80% of dataset
  + Test the classifier on remaining 20% of dataset

**Sentiment Analysis for Passenger Feedback**

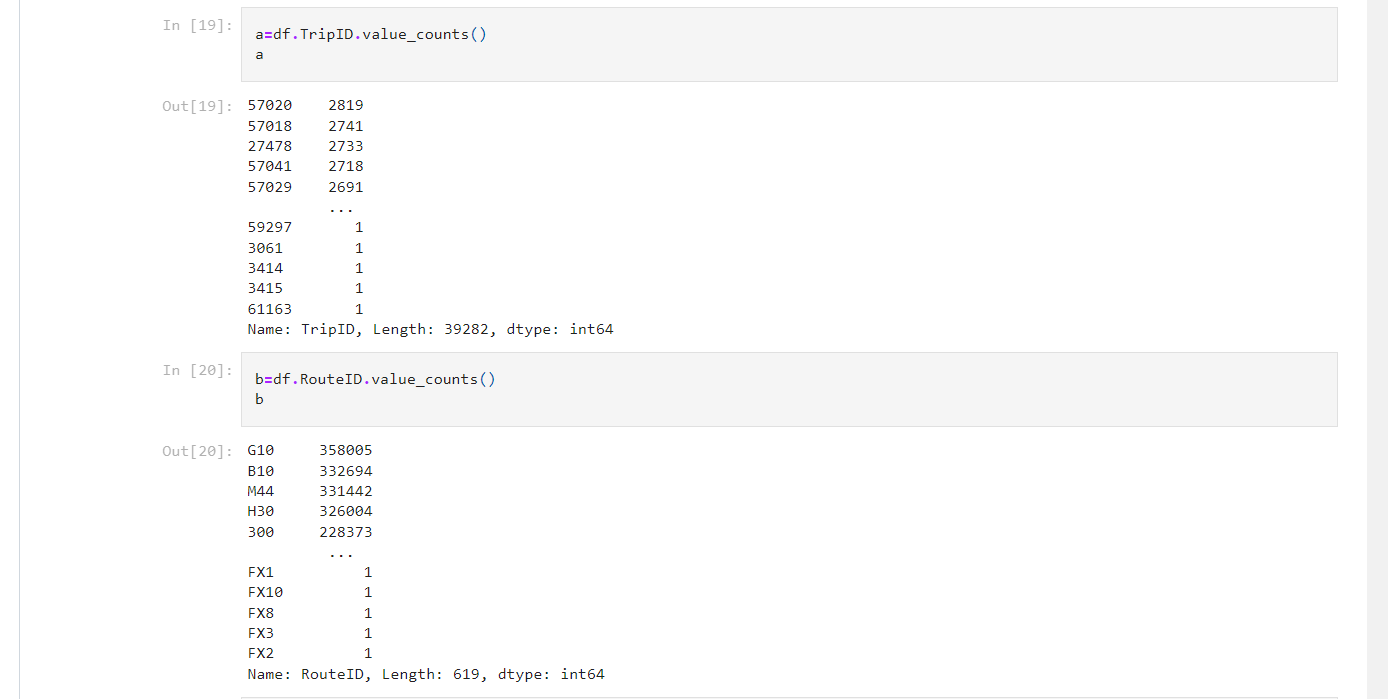
* 1. **Data Preprocessing**
     + For sentiment analysis, we need to extract and clean the text data containing passenger feedback.
     + Load the dataset using Pandas.
     + Select relevant columns for sentiment analysis (e.g., TripID, StopName).
     + Remove duplicates and any irrelevant entries.
     + Handle missing data, if any.
  2. **Text Preprocessing**
     + The text data may contain noise and irrelevant information. Text preprocessing is essential to ensure the accuracy of sentiment analysis.
     + Tokenization: Split text into words.
     + Lowercasing: Convert all text to lowercase.
     + Removing special characters and punctuation.
     + Stopword Removal: Eliminate common words (e.g., "the," "and") that do not carry sentiment.
     + Lemmatization or stemming to reduce words to their base form.
  3. **Model Selection** **VADER Model for Sentiment Analysis:**
     + VADER is a specialized NLP model for sentiment analysis.
     + It provides polarity and intensity scores.
     + Suitable for real-time analysis and informal text.
     + Ideal for public transportation feedback analysis.
  4. **Feature Engineering**
     + Create additional features or transformations that could enhance the analysis, such as time-based aggregations, seasonality, or weather data.
     + Machine Learning Model Development
     + Random Forest is an ensemble learning method that can be used for public transportation analysis as it can handle complex, multifaceted data.
     + It combines multiple decision trees for enhanced accuracy and robustness.
     + The Random Forest model has high accuracy, can handle large datasets, reduces overfitting, is robust to outliers and handles non-linearity.
  5. **Model Training and Validation**
     + Split the dataset into training and testing sets.
     + Train the models for both service disruption prediction and overall analysis.
     + Evaluate the model's performance using relevant metrics.
     + Fine-tune the models if necessary.
  6. **Integration with IBM Cognos**
     + Integrate the machine learning and sentiment analysis results into IBM Cognos for streamlined data analytics and reporting.
  7. **Data Visualization and Reporting**
     + Create dashboards and reports in IBM Cognos to display insights from the analysis.
     + Utilize charts, graphs, and maps to make the results easily interpretable for decision-makers.



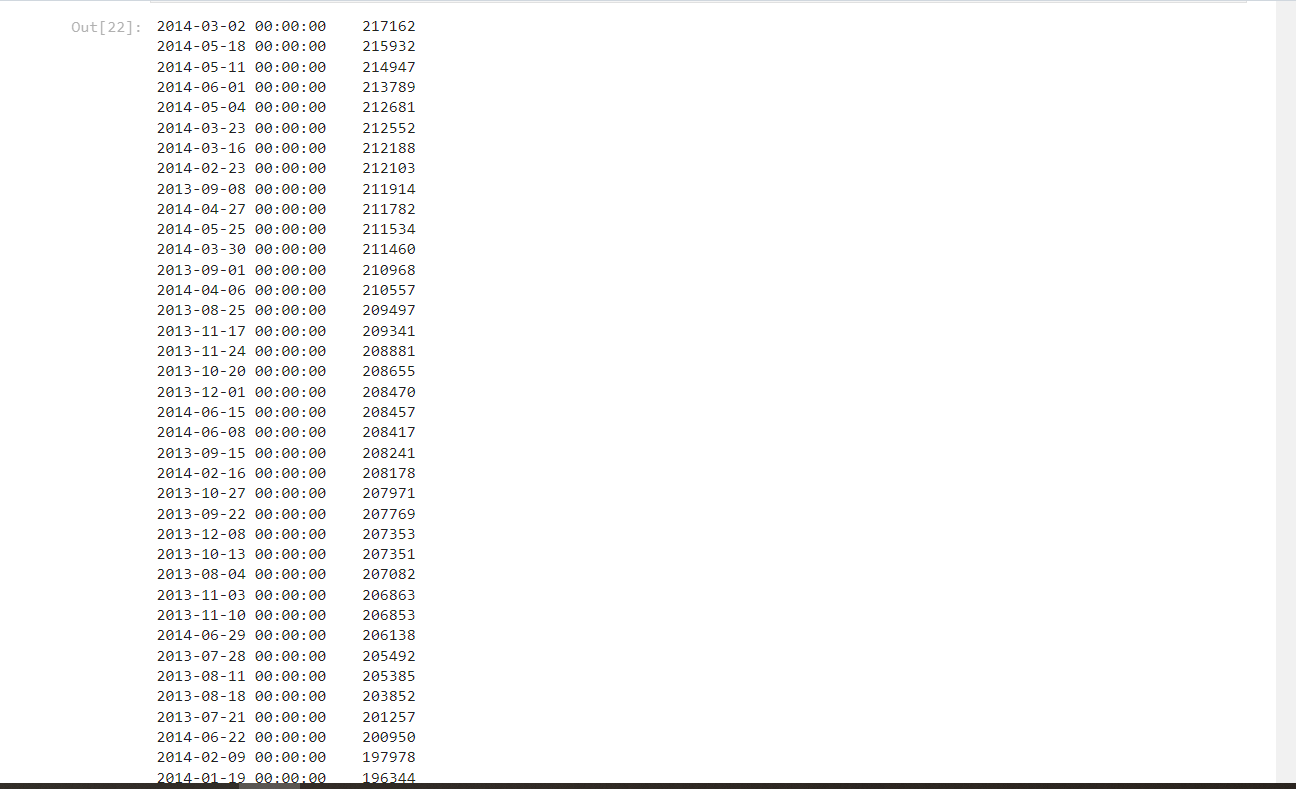


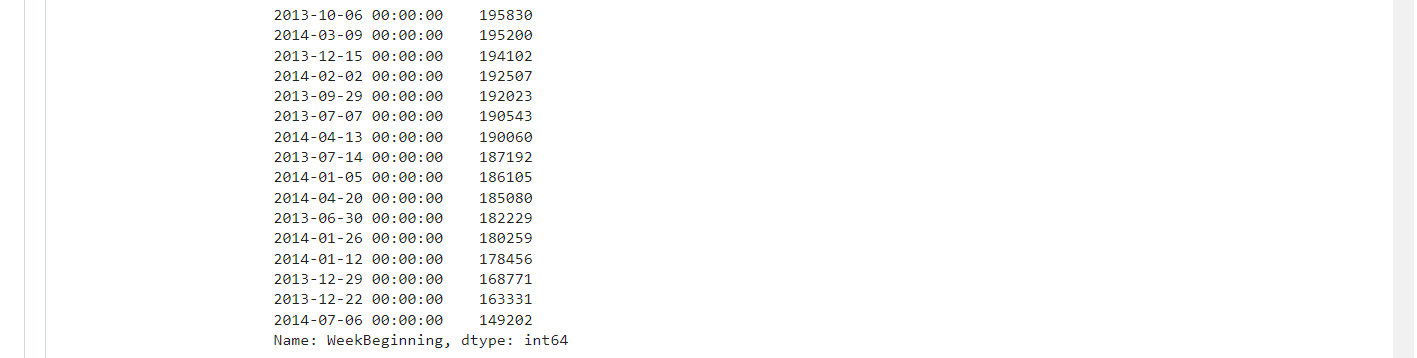




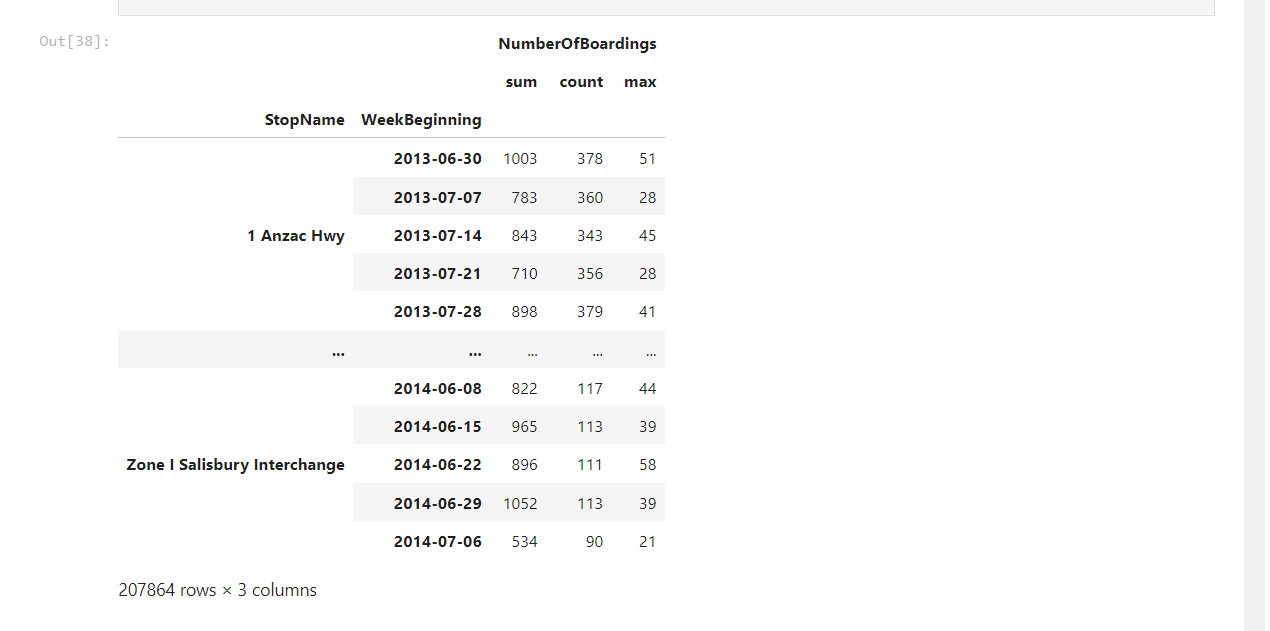




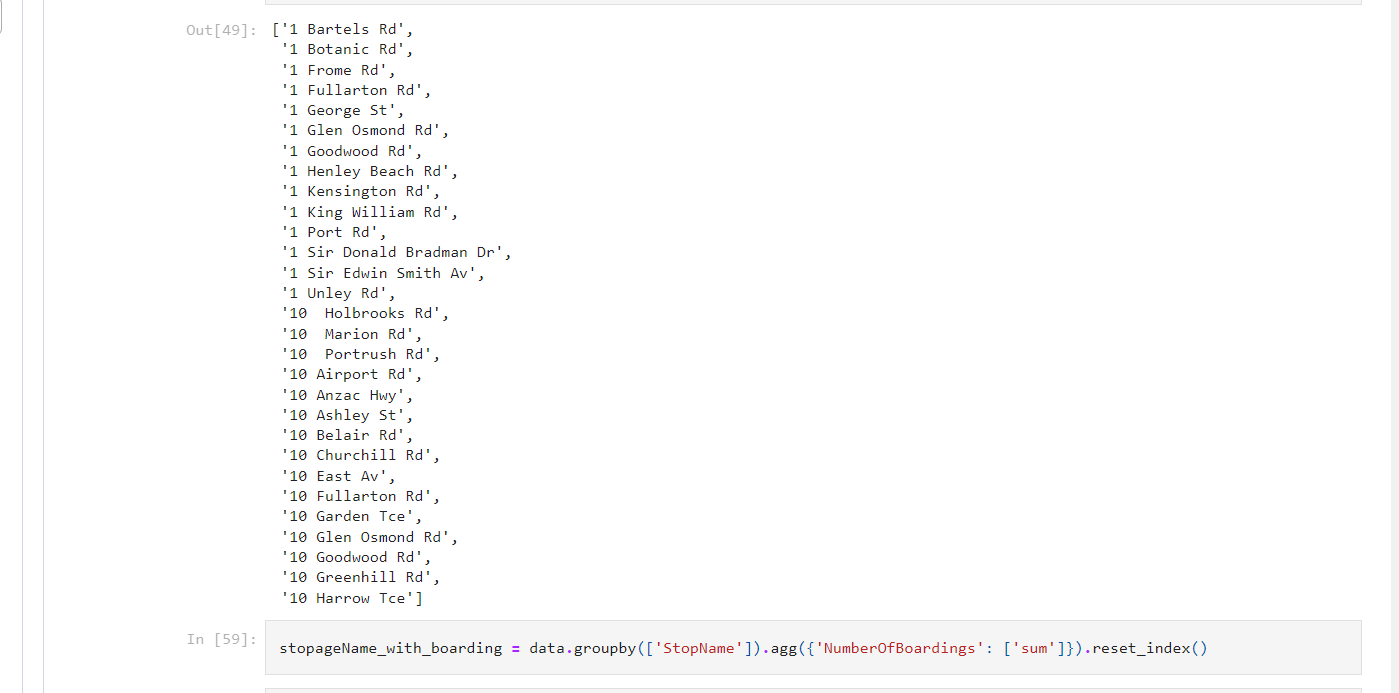


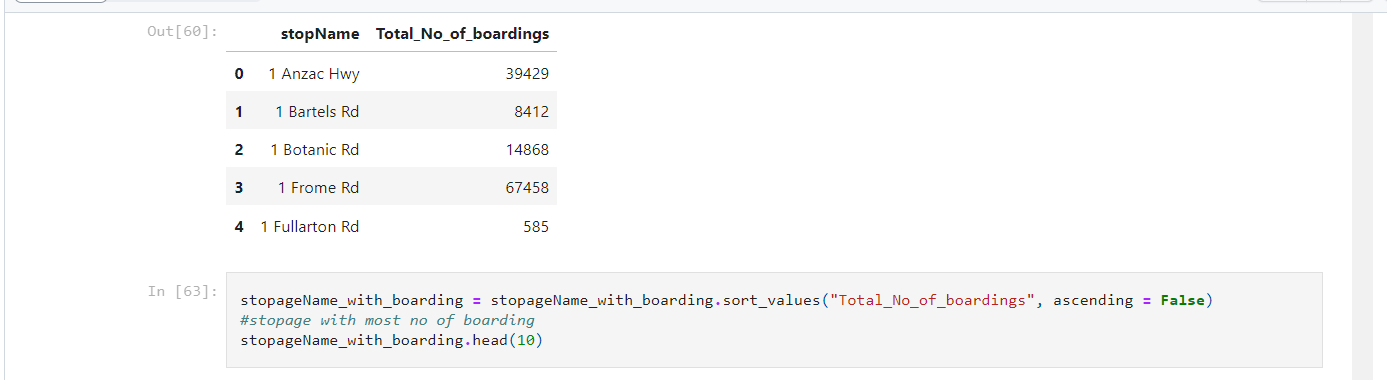




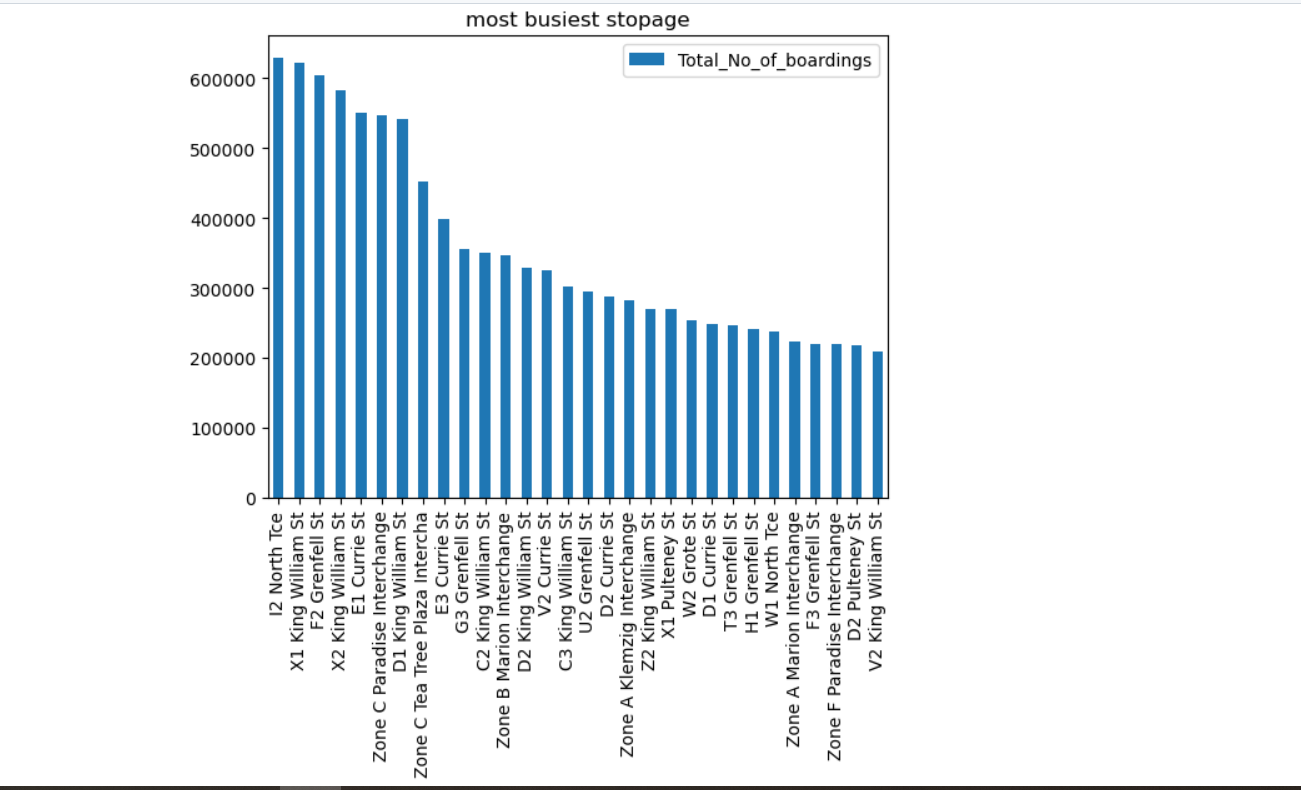


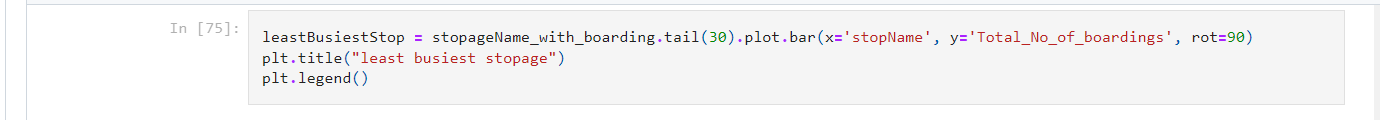


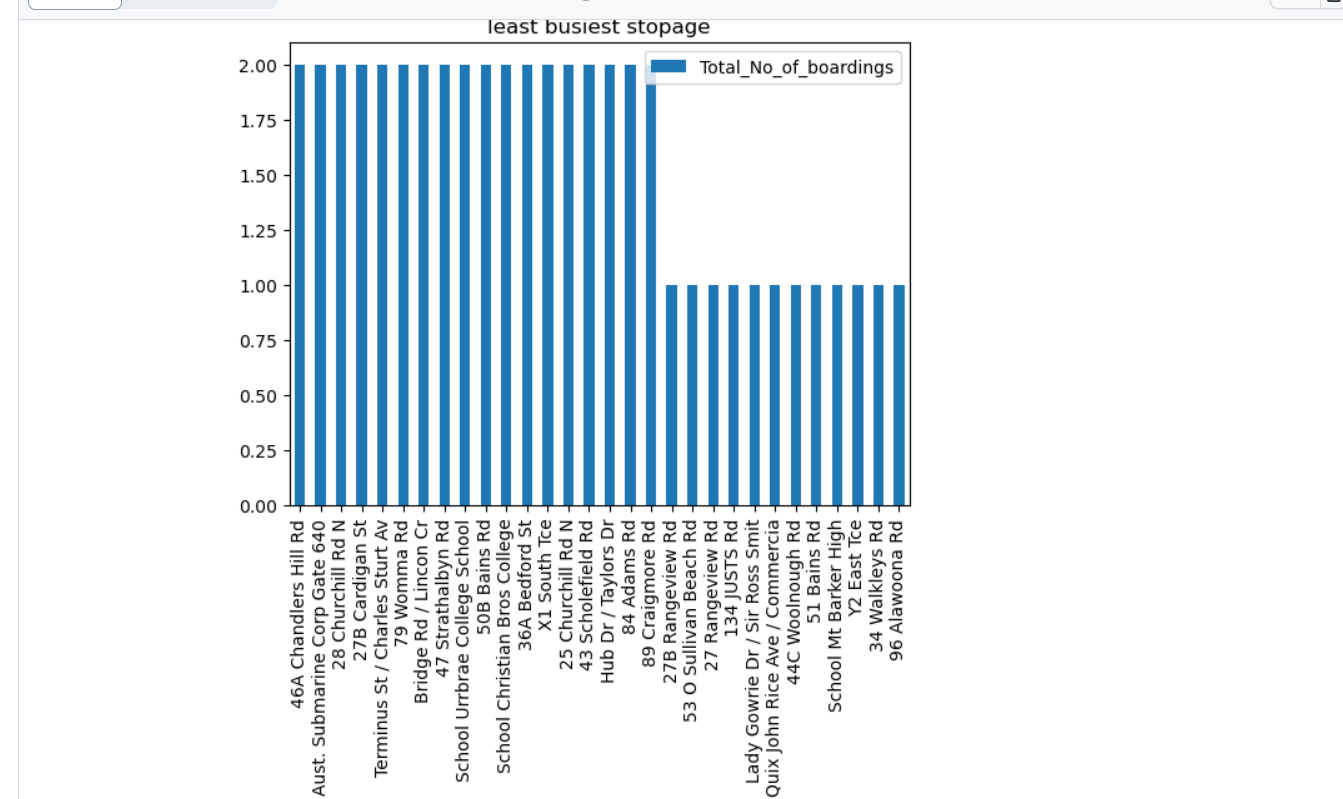


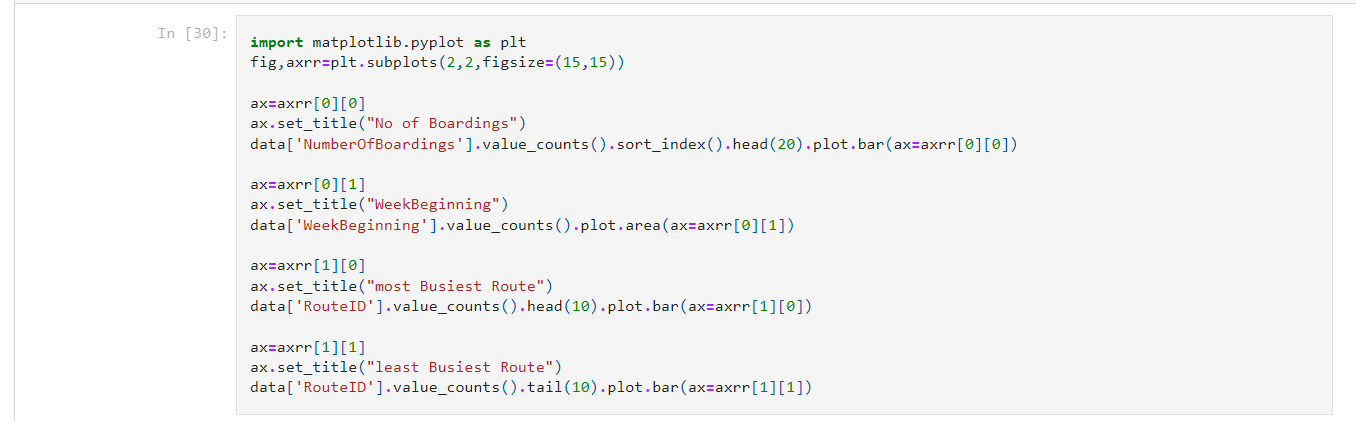


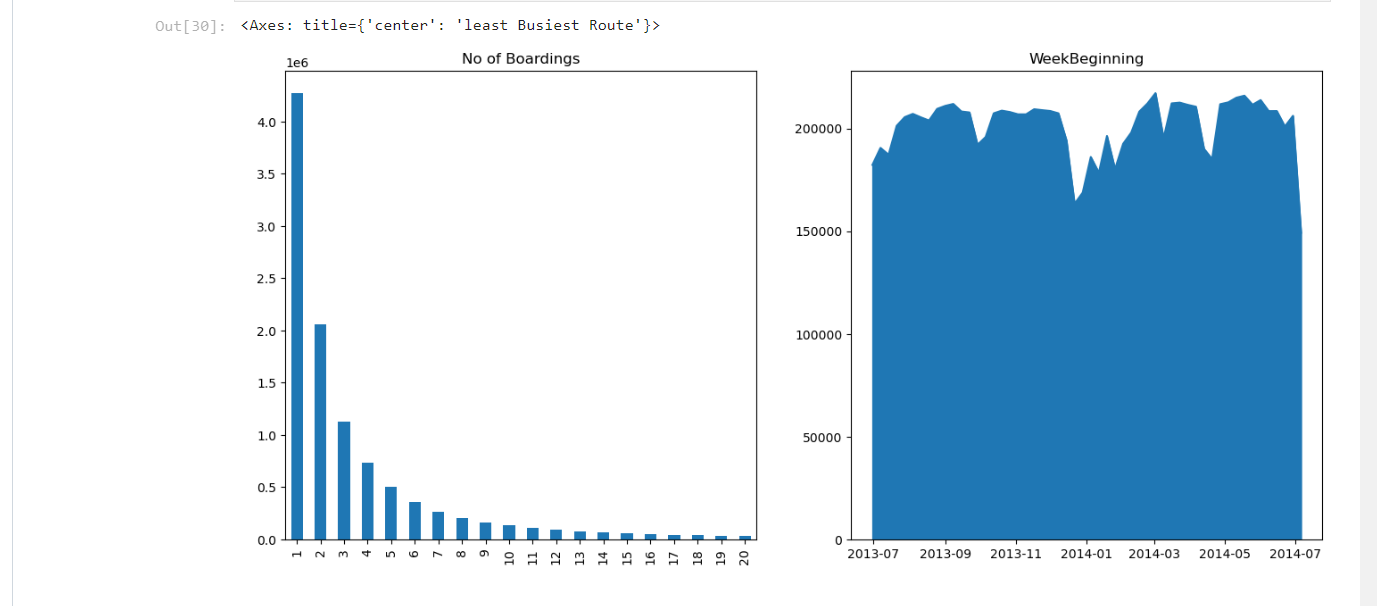


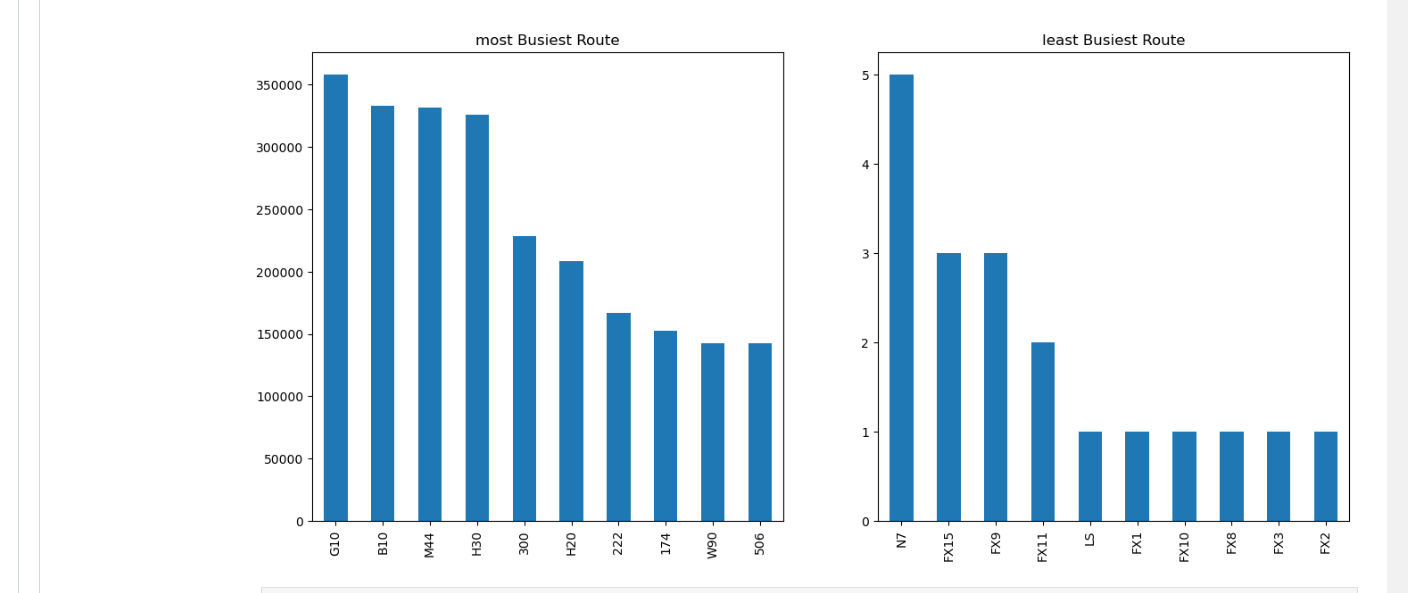








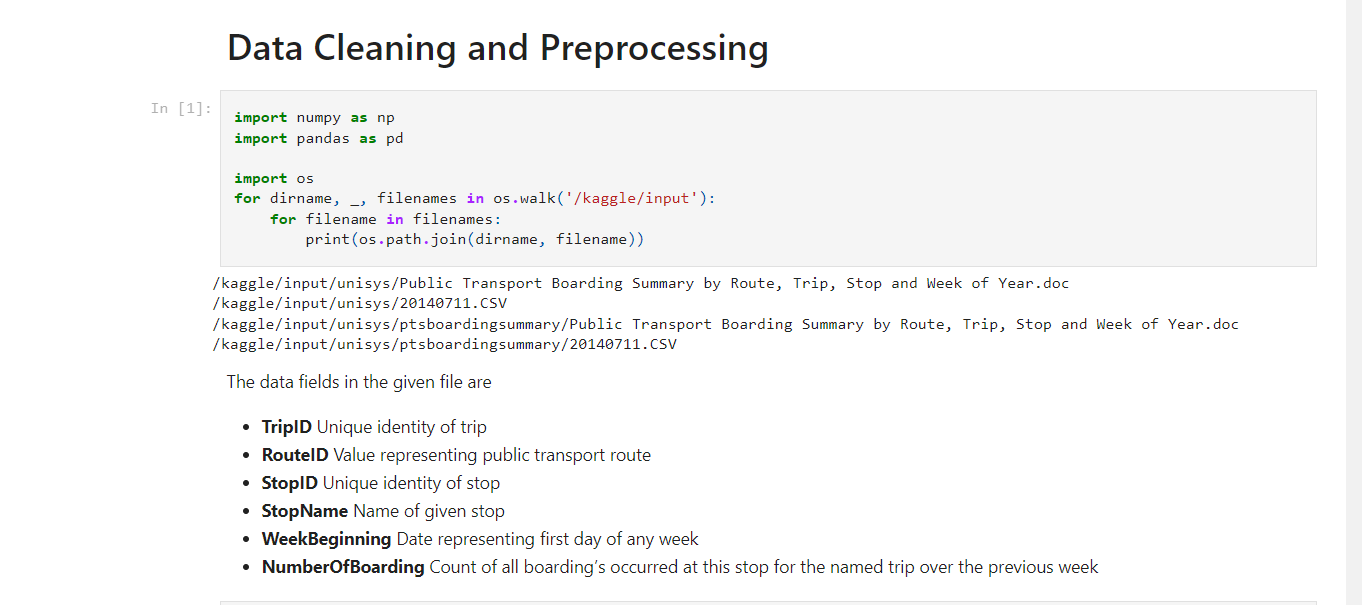


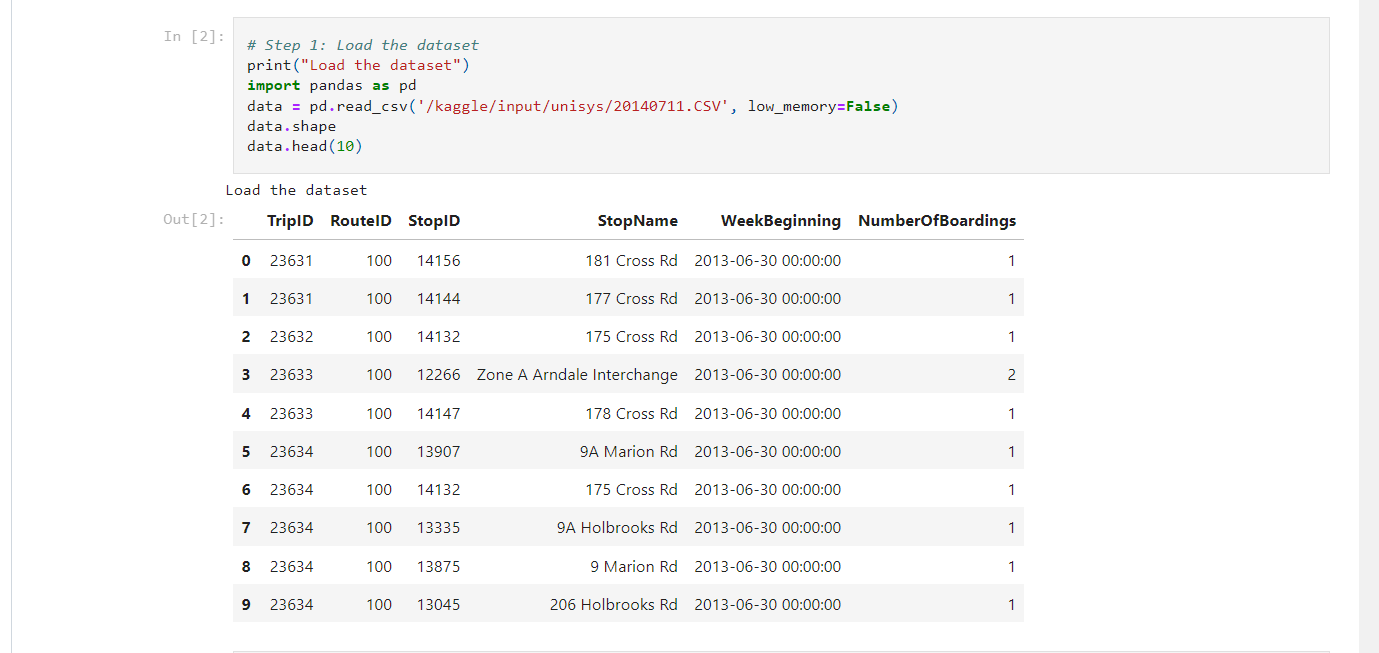


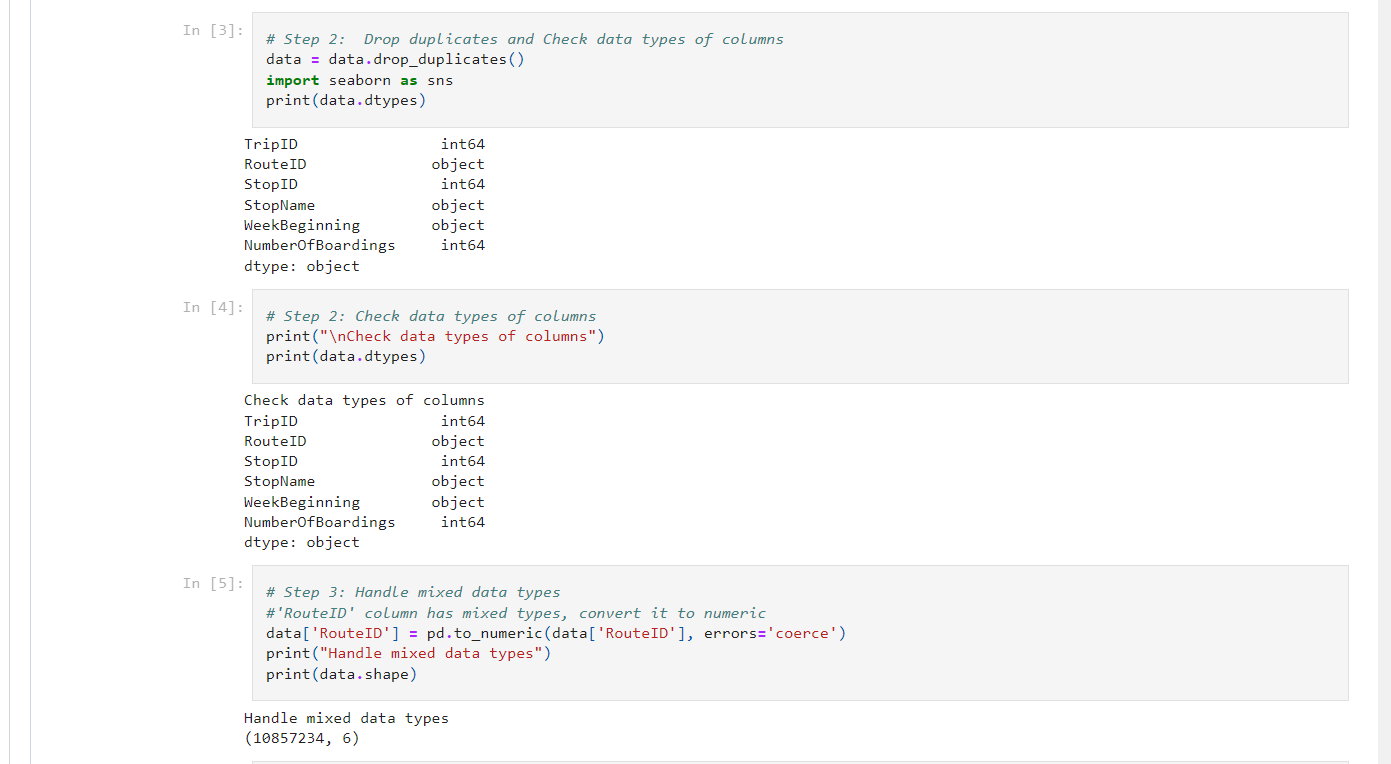
**Analysis Objectives:**

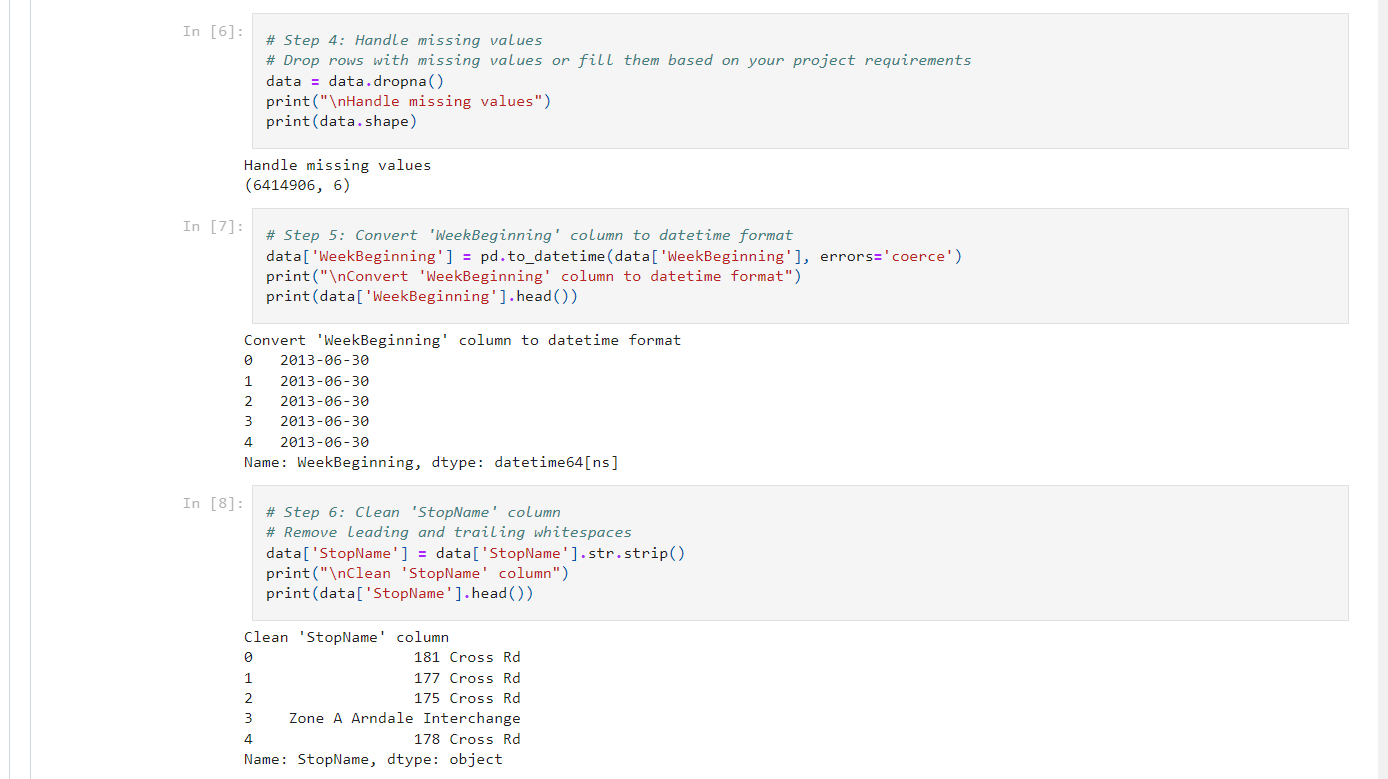
The primary objectives of this project are to assess and improve public transportation efficiency. This involves evaluating factors such as ridership trends, route optimization, on-time performance, and environmental impact. We seek to leverage IBM Cognos for data visualization to gain actionable insights, enhance decision-making for transportation authorities, and contribute to more sustainable and effective urban mobility systems.

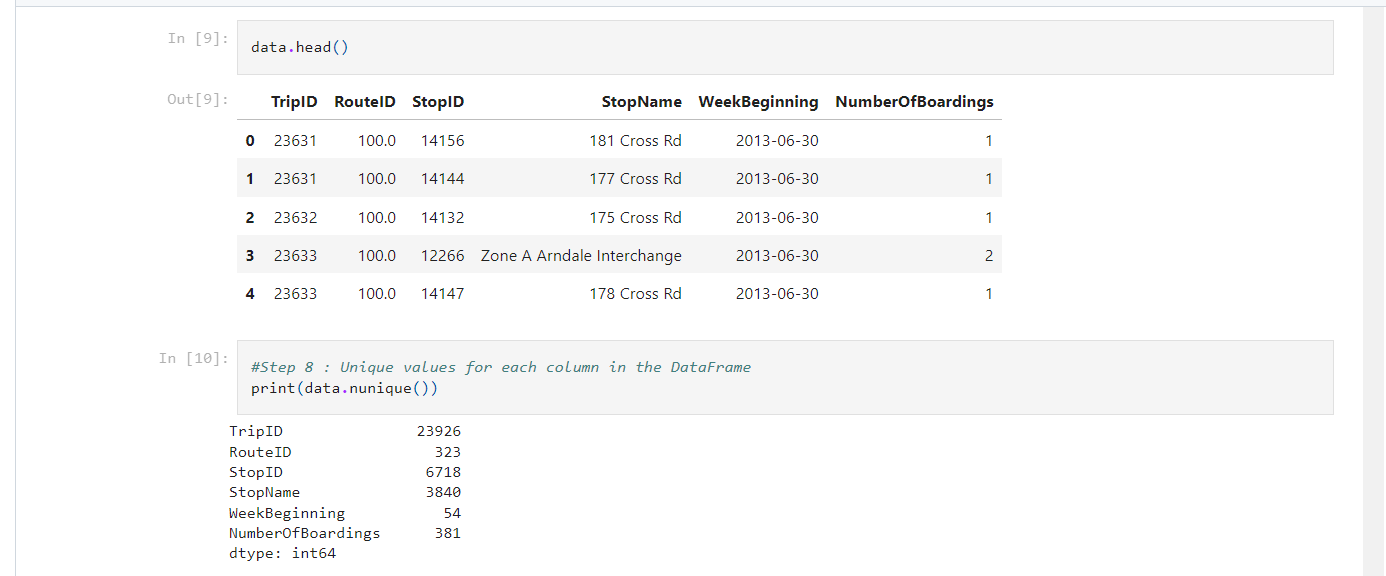
At present we tried visualisations that show how NumberOfBoardings is distributed across routes, stops and a week.

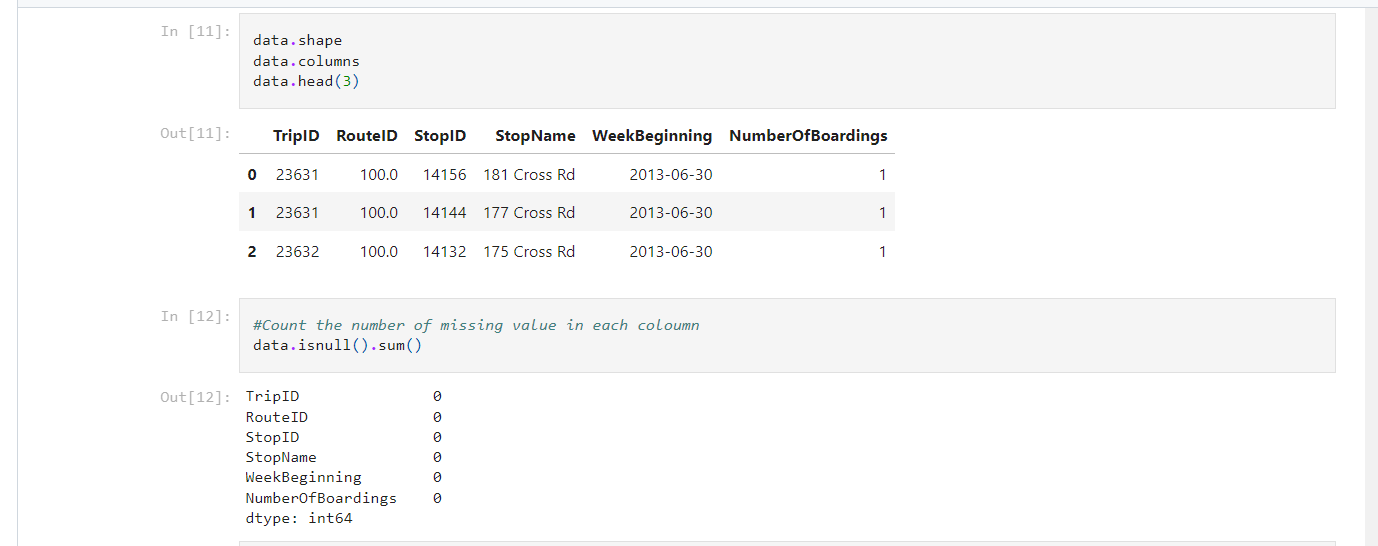


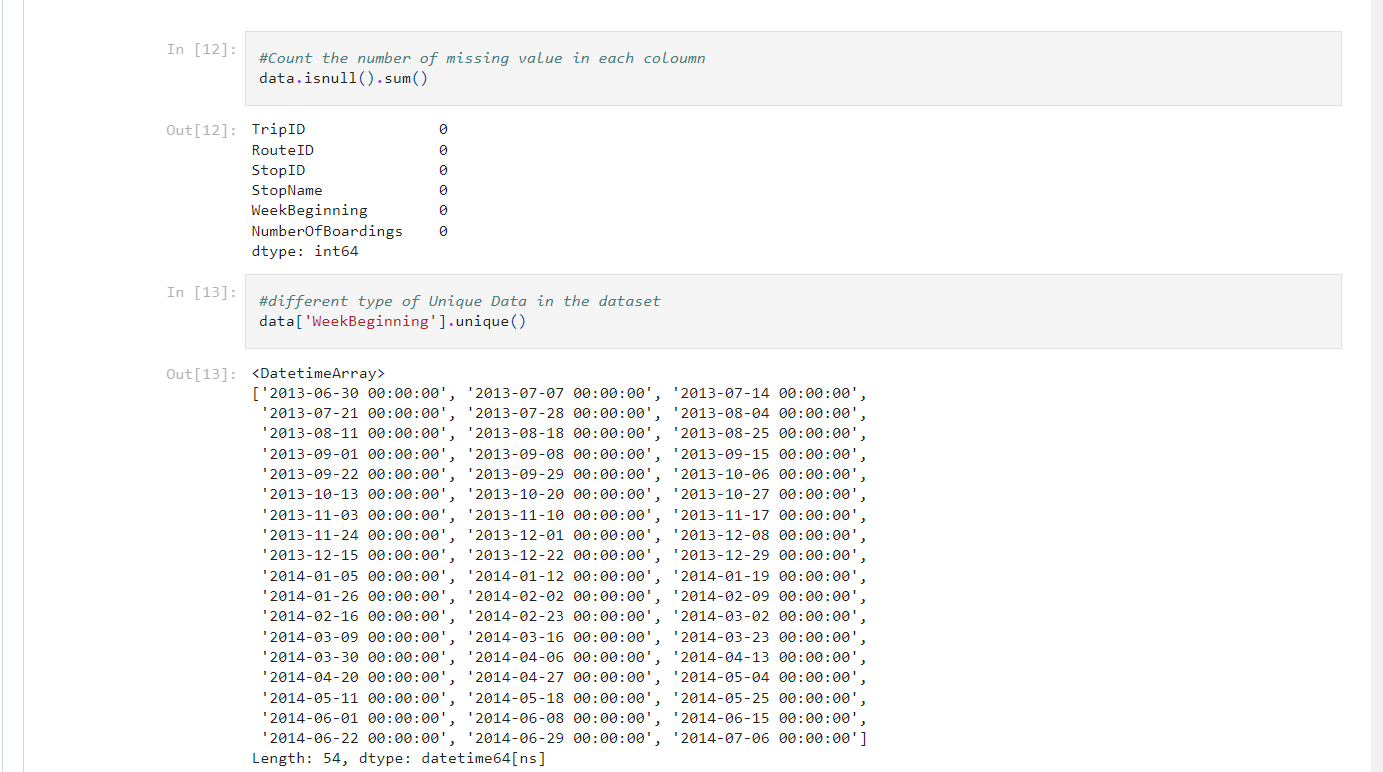


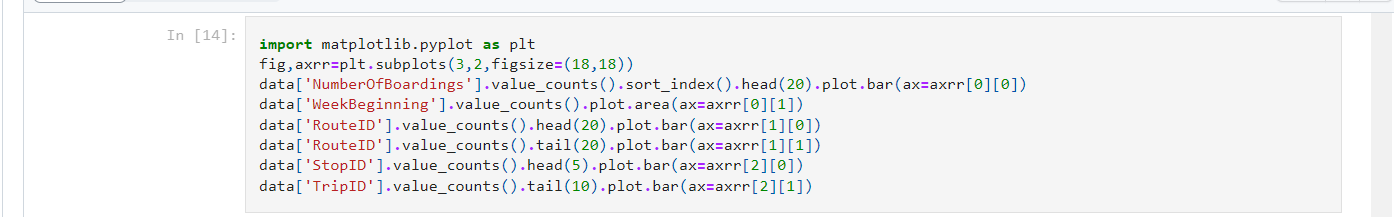


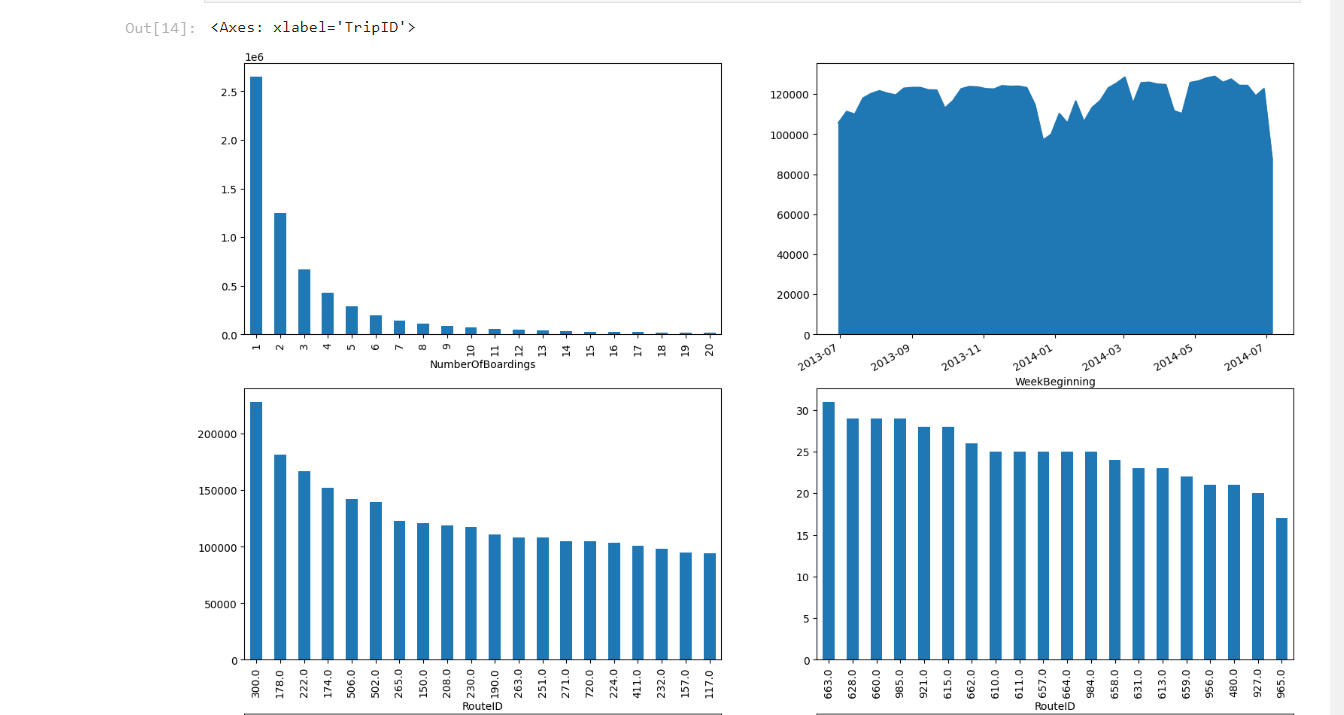


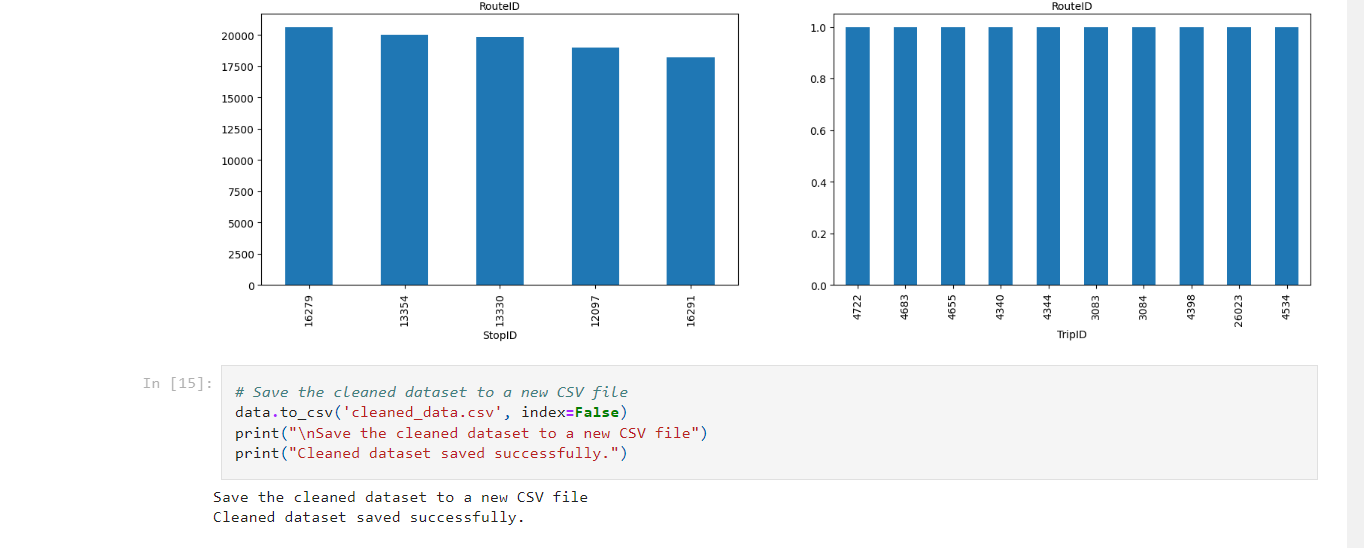


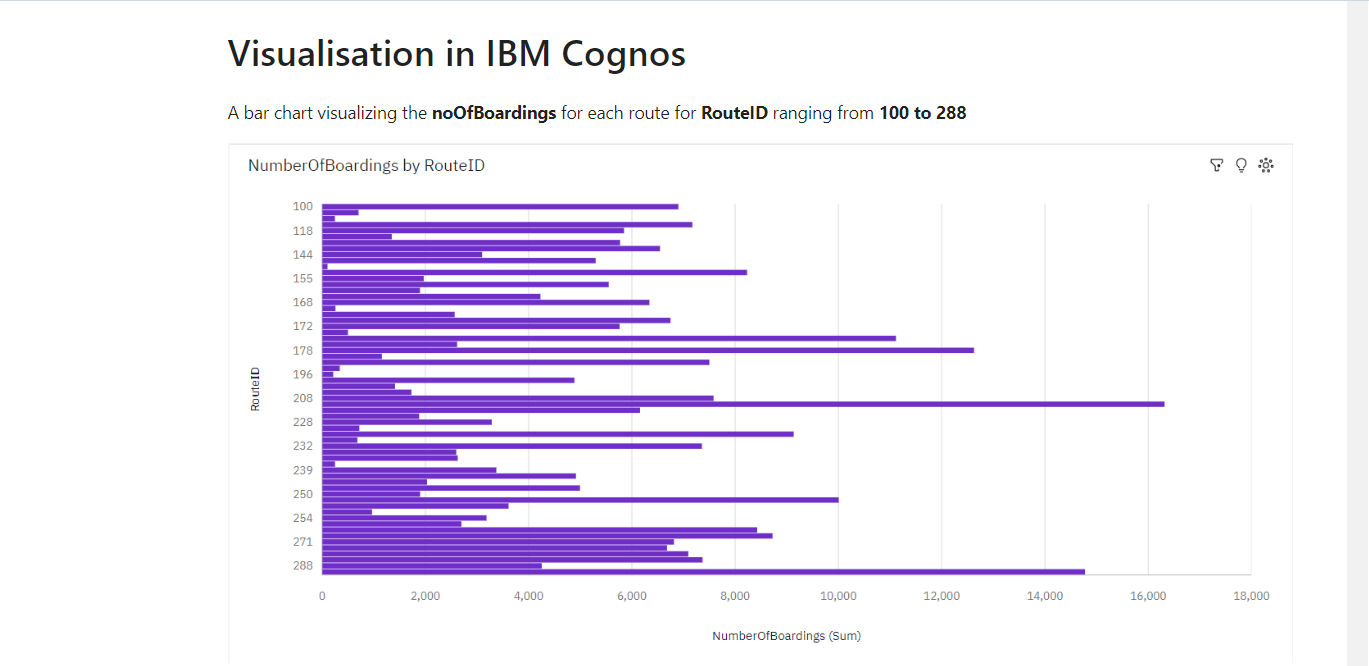


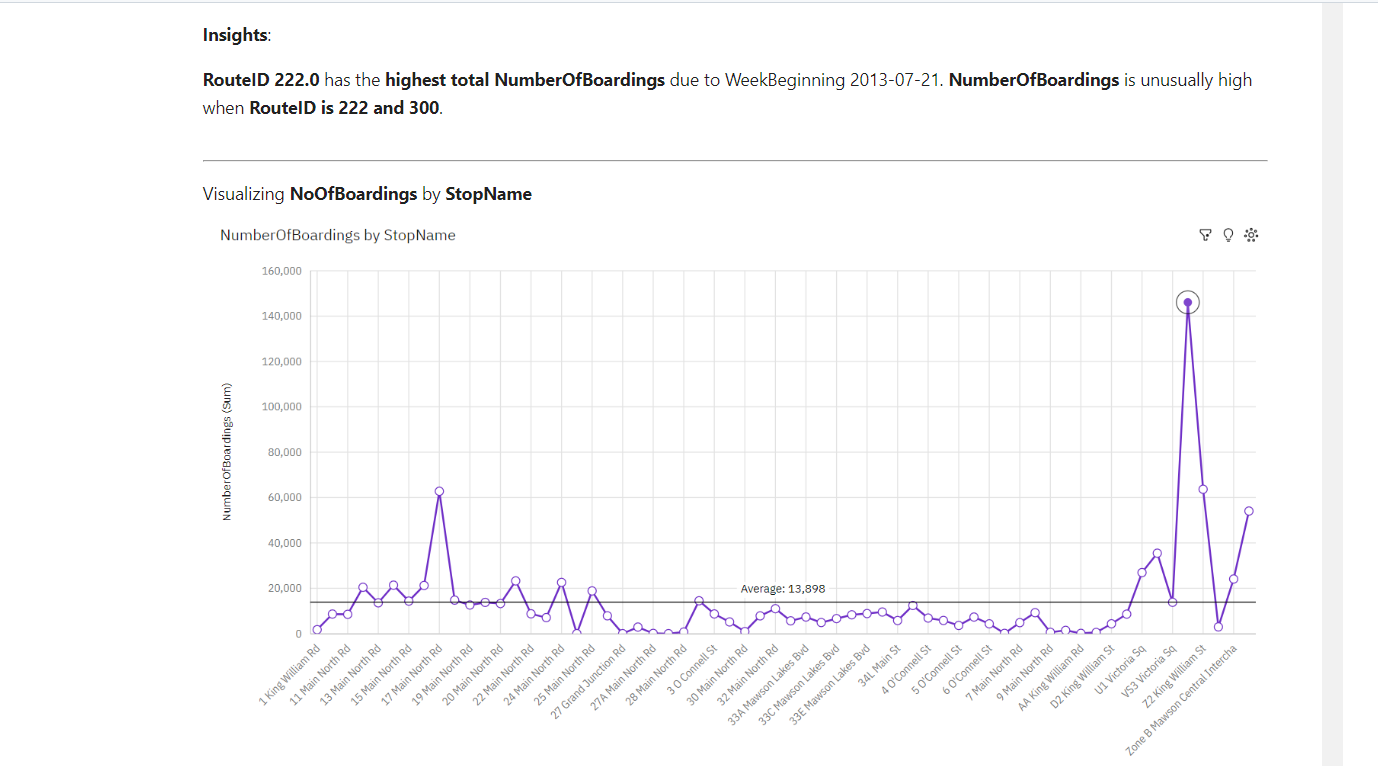


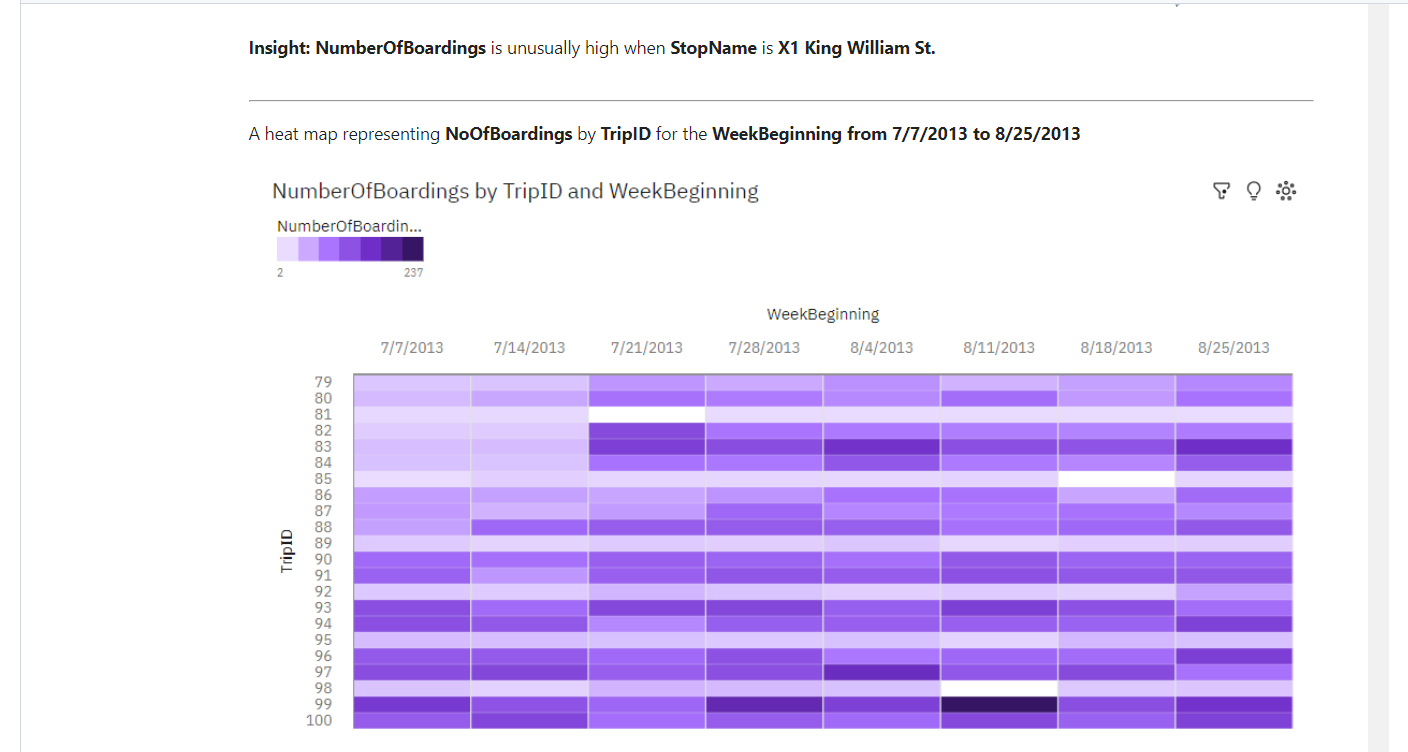












**CONCLUSION:**

This document outlines the project's objectives, design thinking process, data collection process, data visualisation, analysis objectives, innovation, development phases and the role of code in analyzing public transportation data to improve service efficiency. The defined timeline provides a structured approach to project execution.

Through the analysis of public transportation data, we have identified areas that require improvement and support transport improvement initiatives. Effective data visualization strategies and code integration will simplify complex transportation data analysis and provide actionable insights for public transportation improvement.

In conclusion, the use of IBM Cognos for visualization in the public transportation efficiency analysis project has brought about positive changes, leading to more efficient and user-friendly services, better decision-making, and improved sustainability