**PUBLIC TRANSPORTATION ANALYSIS**

**PHASE 3: DEVELOPMENT PART 1**

**INTRODUCTION:**

This project aims to improve public transportation by using data analysis and machine learning. We will explore the provided dataset, identify issues, and preprocess the data. Our goal is to predict service disruptions and analyze passenger feedback sentiment. With the power of machine learning, we'll uncover insights to enhance transportation services. This document outlines our step-by-step approach.

**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**

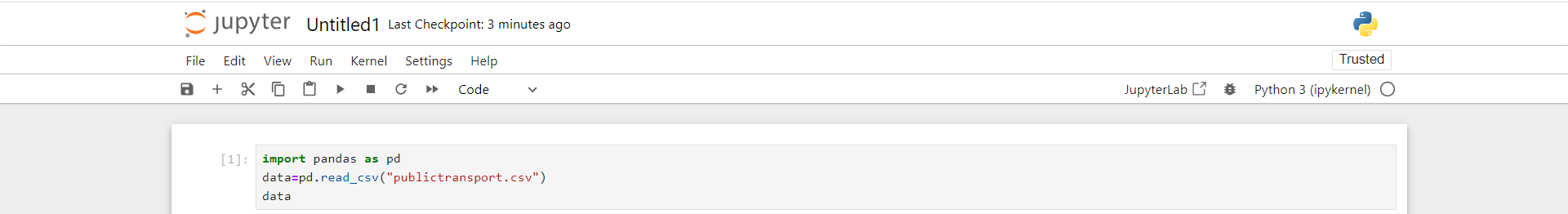
* + Select relevant columns for analysis (e.g., TripID, RouteID, StopName).
  + Handle missing data, duplicates, and irrelevant entries.
  + Convert data types if needed

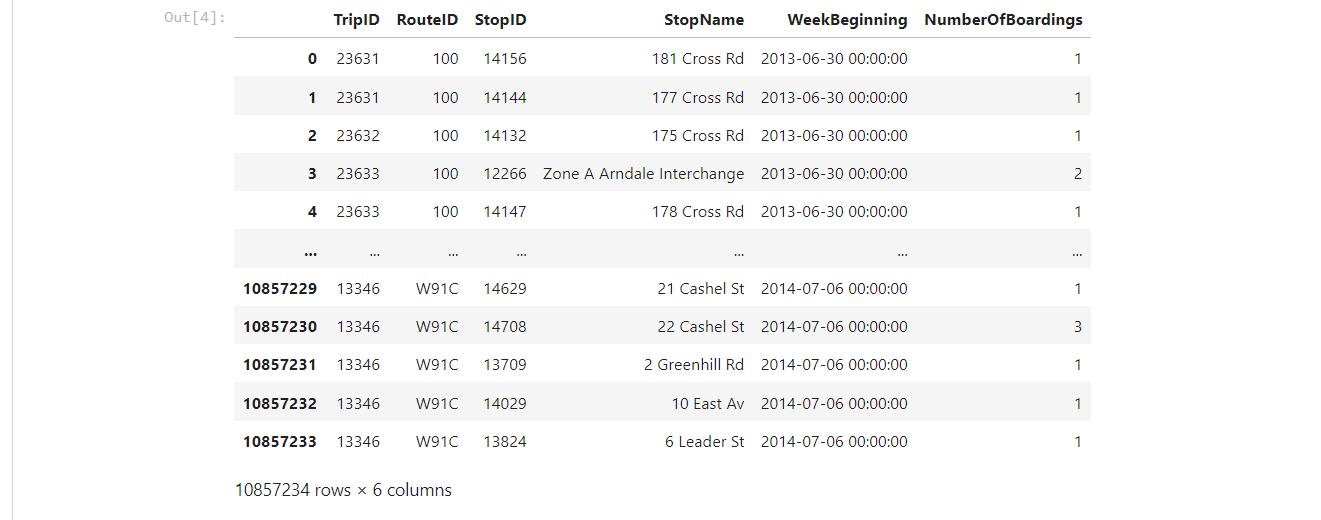
**Predicting Service Disruptions**

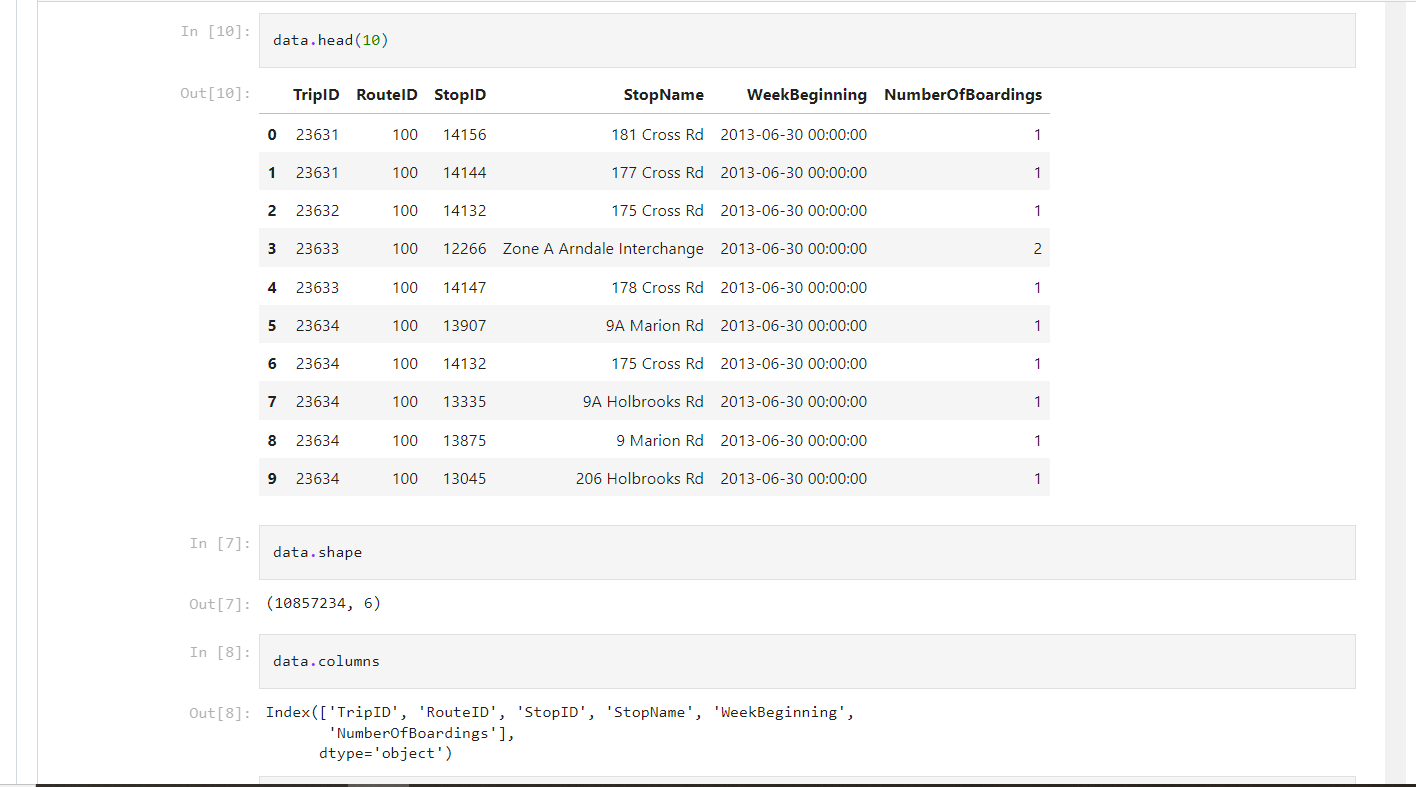
* + Innovation: Define how service disruption is determined from given features
  + Select a set of features and Service Disruption as target feature
  + Create DecisionTreeClassifier and train on 80% of dataset
  + Test the classifier on remaining 20% of dataset

**Sentiment Analysis for Passenger Feedback**

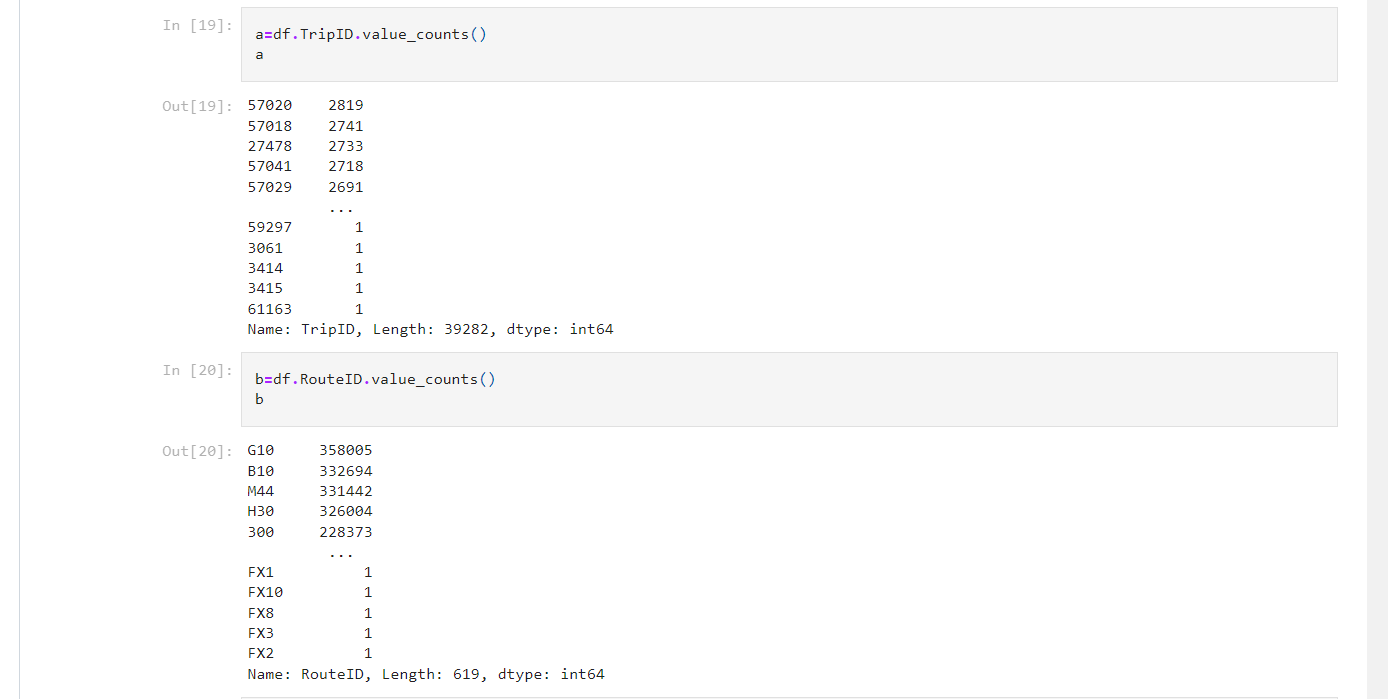
* 1. **Data Preprocessing**
     + For sentiment analysis, we need to extract and clean the text data containing passenger feedback.
     + Load the dataset using Pandas.
     + Select relevant columns for sentiment analysis (e.g., TripID, StopName).
     + Remove duplicates and any irrelevant entries.
     + 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.



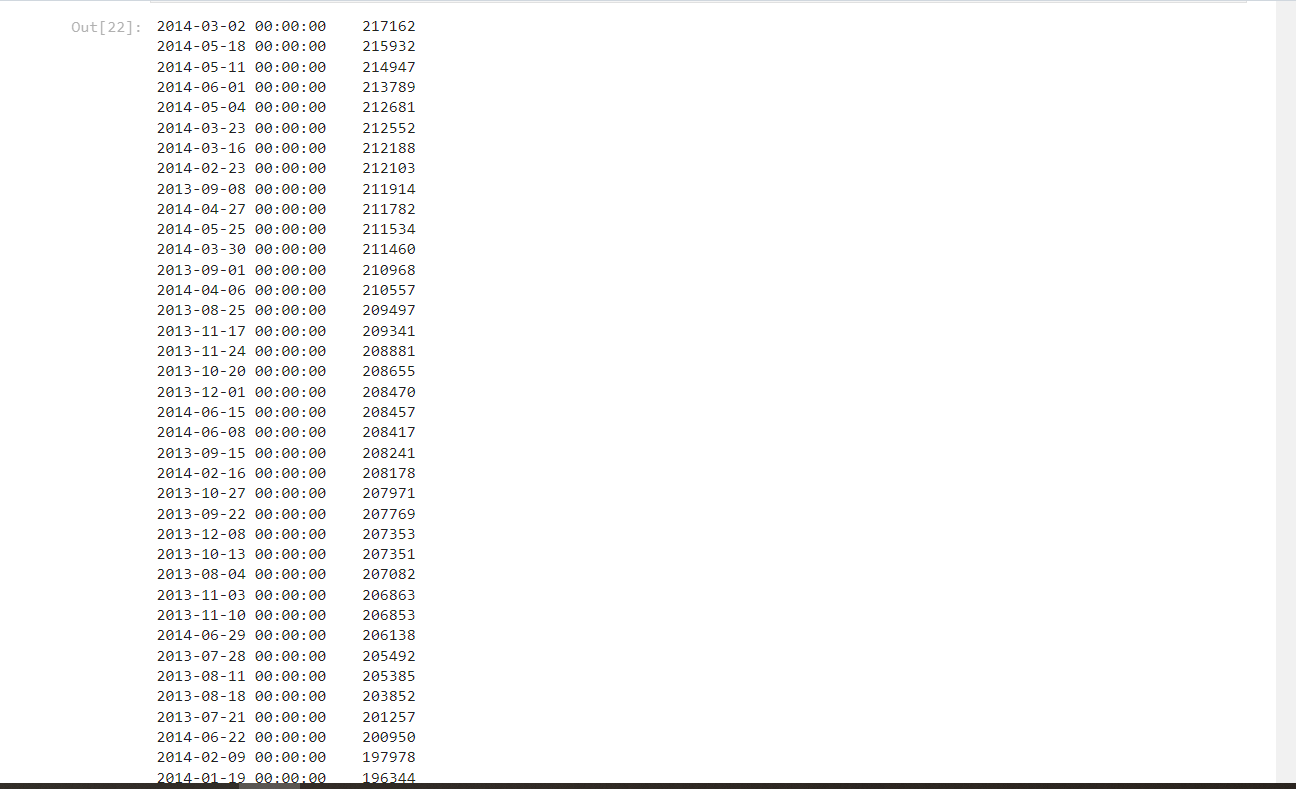


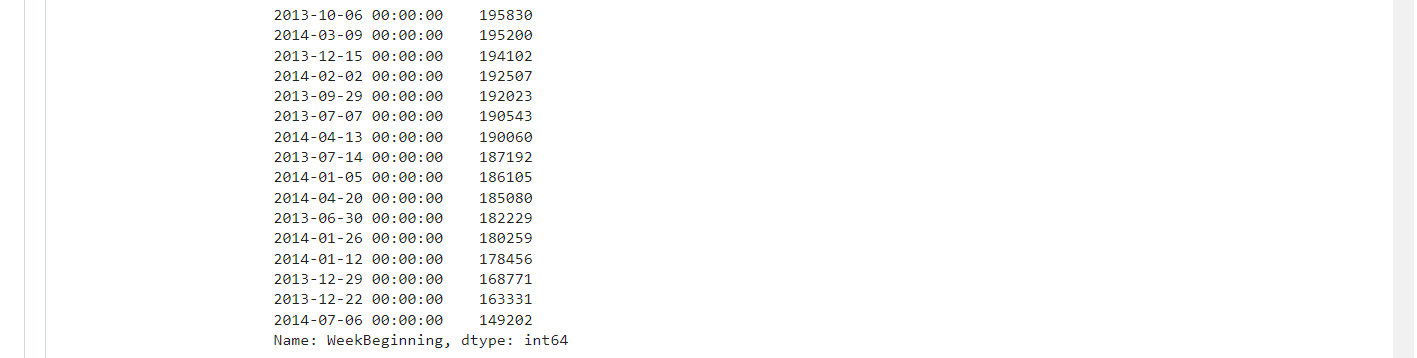




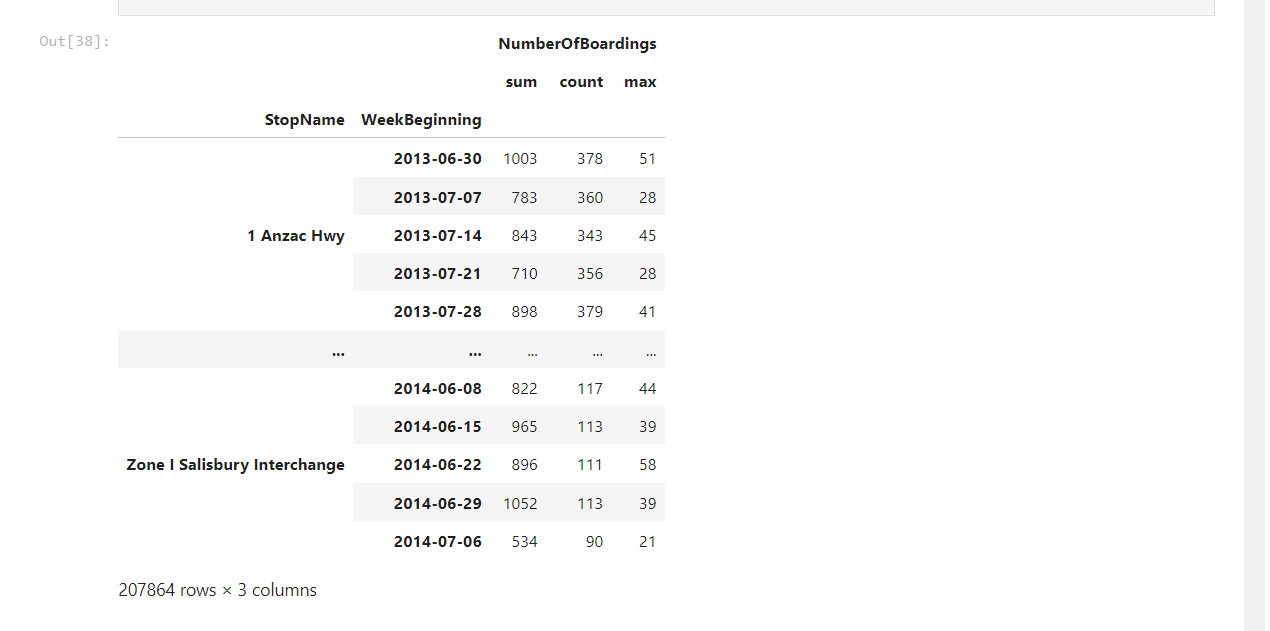




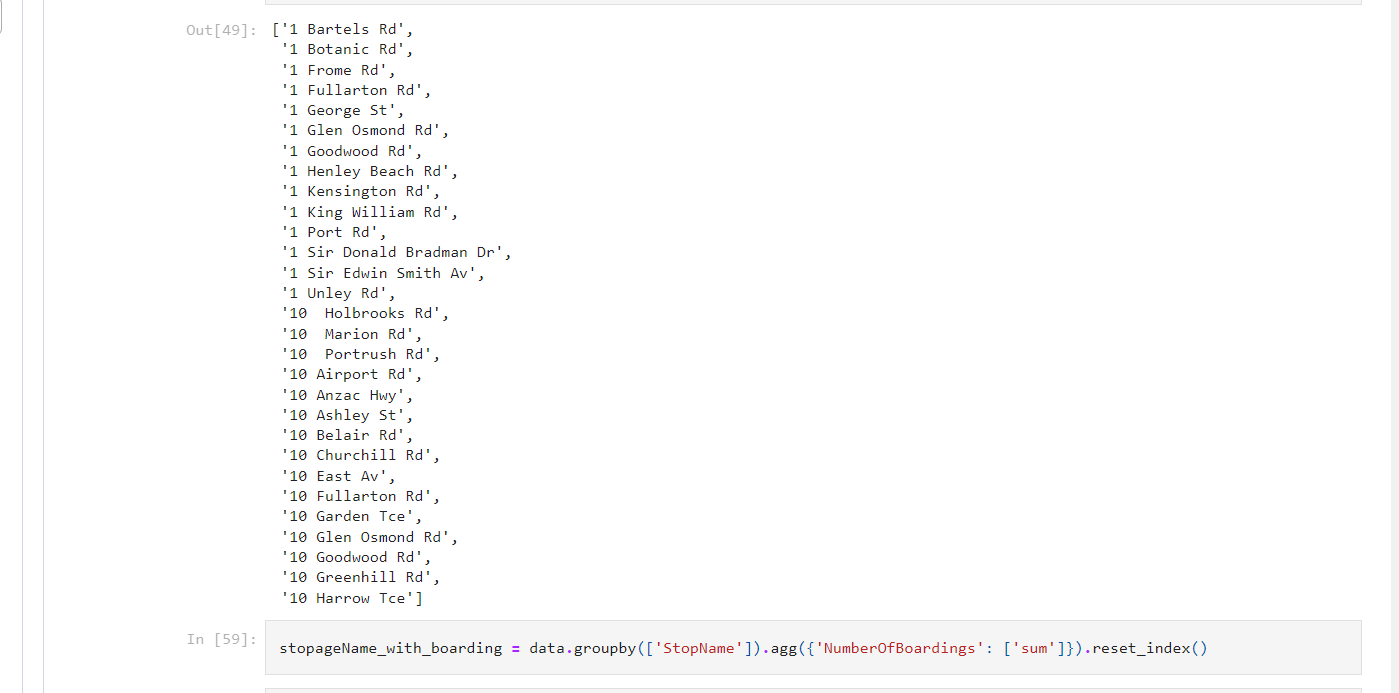


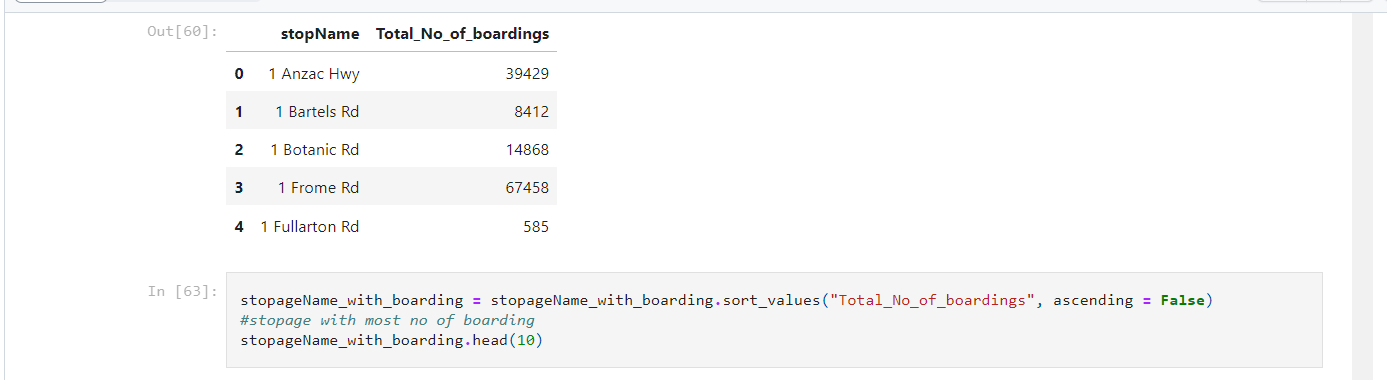




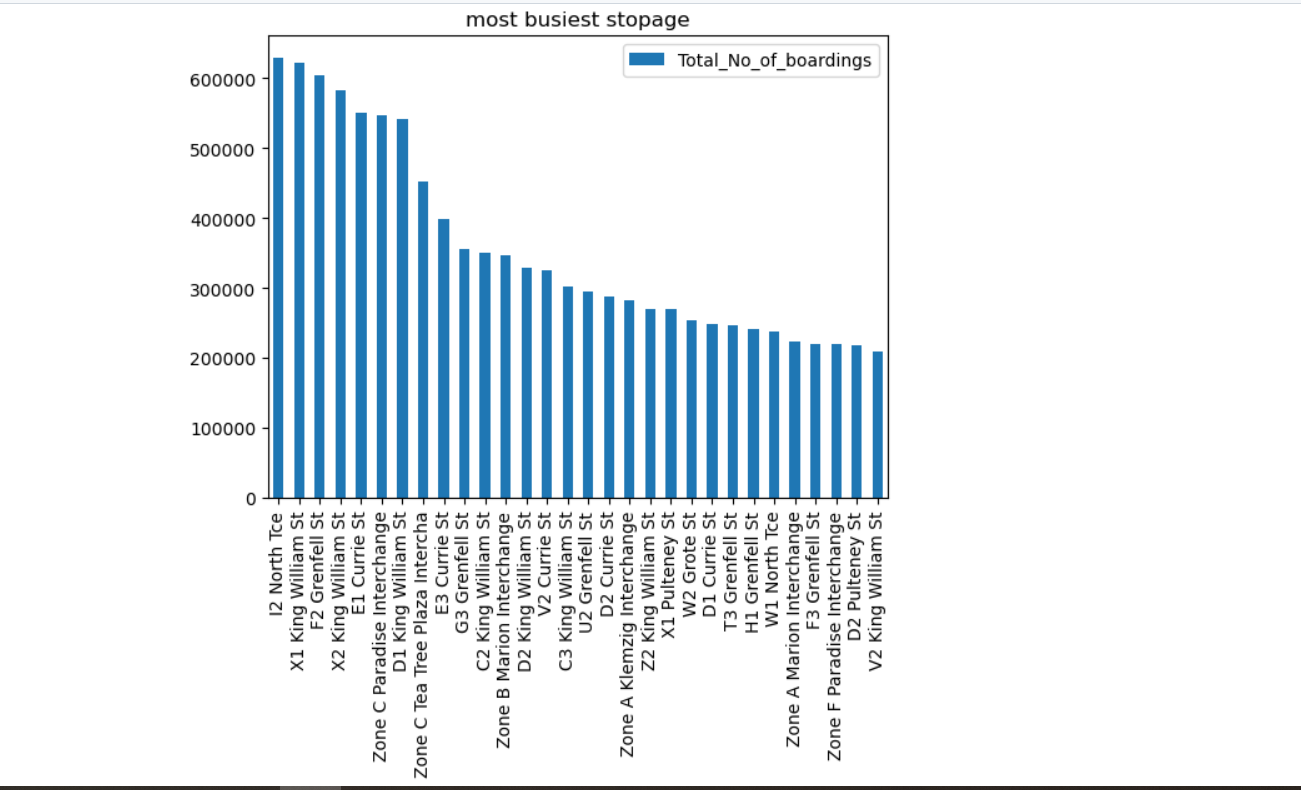


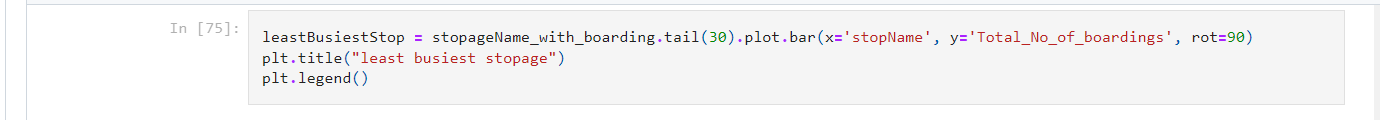


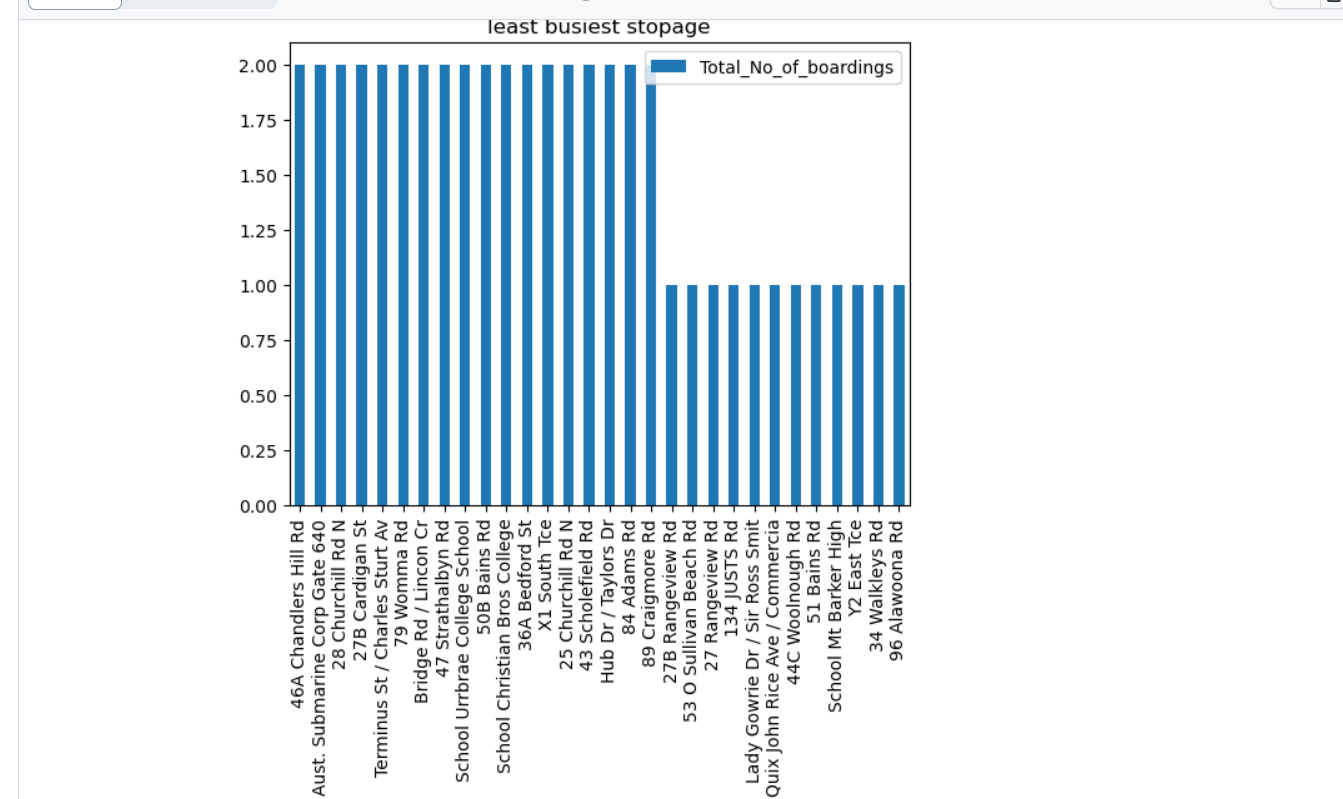


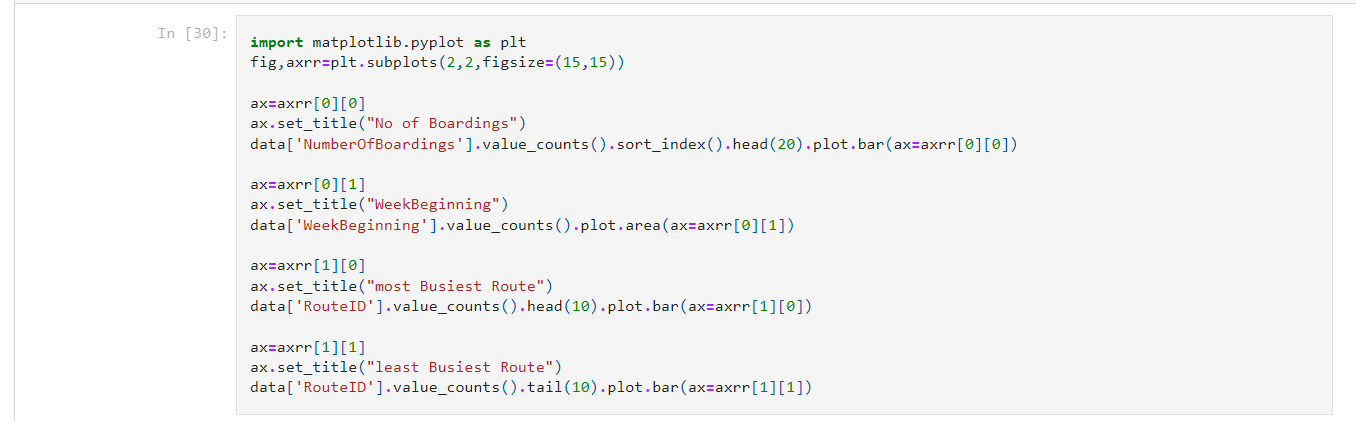


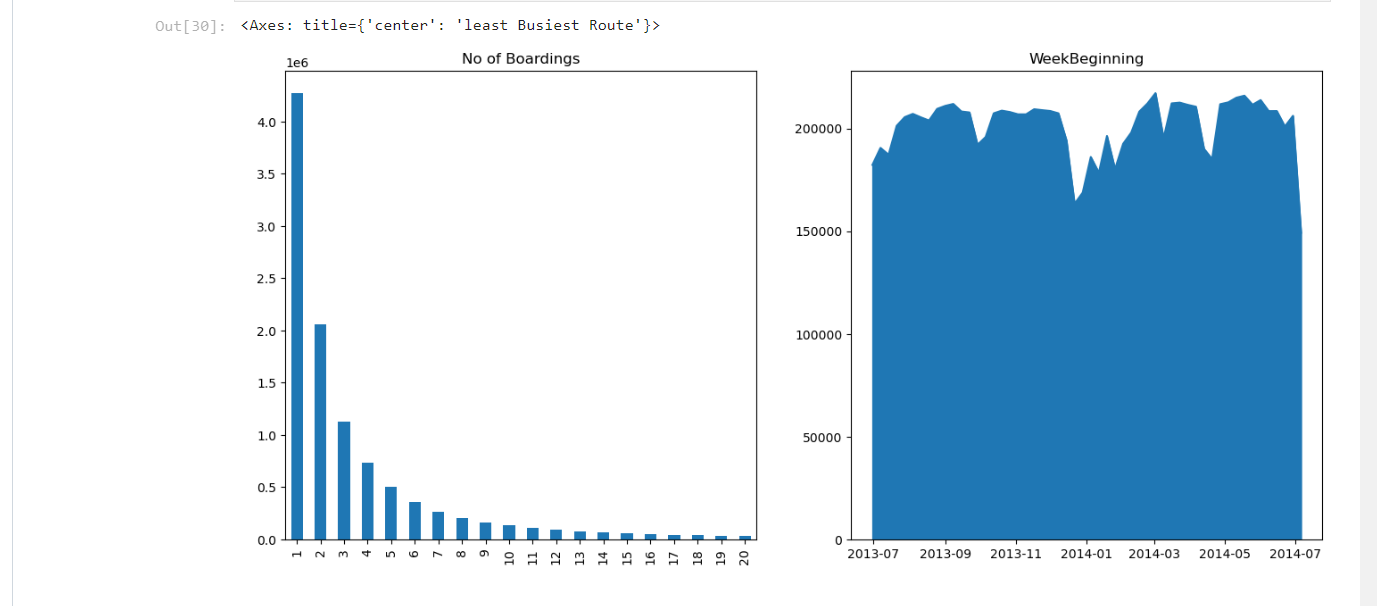


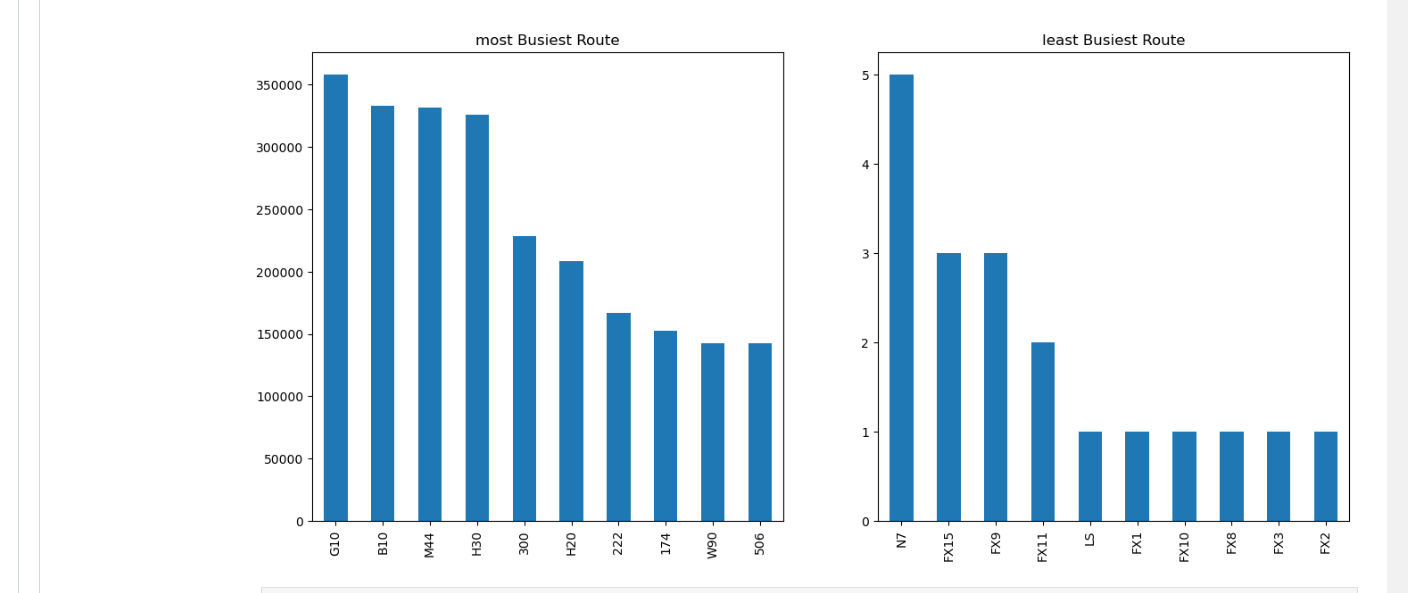












**CONCLUSION:**

By following these steps, we can effectively enhance public transportation analysis by predicting service disruptions and analyzing passenger sentiment, ultimately leading to improvements in the transportation system.