**Public Transportation Efficiency Analysis**

**Introduction**

Public transportation systems play a crucial role in the mobility and sustainability of urban environments. They are the lifeblood of cities, connecting people to employment, education, and essential services while also mitigating traffic congestion and reducing carbon emissions. In an era of growing urbanization and environmental concerns, the efficiency of public transportation systems has never been more critical.

**1. Objectives**

In Phase 4, we aim to achieve the following objectives:

1. **Data Exploration and Visualization:** We will thoroughly explore our dataset, uncover patterns and trends, and create informative visualizations that shed light on the current state of public transportation services. Our visualizations will focus on key metrics related to on-time performance, passenger feedback, and service efficiency.
2. **Dashboard Design:** Using IBM Cognos, we will design interactive dashboards that condense complex data into easily digestible formats. These dashboards will serve as a valuable resource for stakeholders and decision-makers to monitor and assess performance.
3. **Supervised Learning - Regression Analysis:** We will implement regression analysis to predict and model critical aspects of public transportation. Regression models will enable us to understand the relationship between independent variables and our target metrics, paving the way for predictive analysis.

**2. Data Preparation and Preprocessing**

In this section, we address the critical aspects of data preparation and preprocessing, a fundamental step to ensure that our analysis is based on clean, accurate, and reliable data. The quality of our insights and models hinges on the integrity of the dataset, and therefore, meticulous attention to data preparation is paramount.

**2.1 Data Sources**

Our dataset is a comprehensive repository of public transportation information, including details on routes, stops, schedules, passenger feedback, and performance metrics. It comprises data from various sources, all aimed at providing a holistic view of the public transportation system. The data sources include:

* Route Information: This dataset provides details about different routes, their names, and other route-specific information.
* Stop Information: Stop data contains valuable information about stops, such as their names, locations, and unique identifiers.
* Schedules: Schedules data is crucial for understanding the timing and frequency of transportation services.
* Passenger Feedback: Passenger feedback records include feedback from passengers who have used the public transportation services, helping us gauge satisfaction levels and areas for improvement.
* Performance Metrics: Performance data records metrics such as on-time performance and efficiency indicators, providing valuable insights into the quality of service.

**2.2 Data Cleaning**

To ensure the reliability of our dataset, we executed a series of data cleaning steps:

* Handling Missing Values: We identified and addressed missing values in our dataset. This involved examining each column for missing data and applying suitable strategies, such as imputation or removal, to maintain data completeness and accuracy.
* Data Deduplication: Duplicates can distort the accuracy of our analysis. We conducted data deduplication to eliminate redundant records and maintain a unique and clean dataset.
* Data Validation: Rigorous validation checks were performed to identify and rectify any inconsistencies, anomalies, or outliers. Data validation helped us ensure that the information was internally coherent and adhered to predefined standards.
* Data Integration: In some instances, data from multiple sources was integrated to form a comprehensive dataset. The integration process involved aligning data with common identifiers to establish relationships and dependencies.

**2.3 Data Structuring**

For efficient analysis, the dataset was structured into a coherent and organized format. We implemented the following steps:

* Column Standardization: To facilitate consistent analysis, we standardized column names, ensuring clarity and uniformity in data representation.
* Date and Time Formatting: Time-related data, particularly in schedules and performance metrics, was formatted consistently to simplify temporal analysis and visualization.
* Categorization: Variables that represented categorical data, such as stop names or route IDs, were categorized for easy interpretation and visualization.

**2.4 Data Documentation**

Throughout the data preparation and preprocessing phase, we maintained comprehensive documentation. This documentation provides transparency and traceability, ensuring that the processes applied to the dataset are well-documented and reproducible.

The result of our data preparation and preprocessing efforts is a structured, clean, and reliable dataset ready for in-depth exploration, visualization, and regression analysis. This robust dataset will underpin our subsequent analyses, providing a solid foundation for meaningful insights and predictions.

**3. Supervised Learning - Regression**

We embark on a journey into the realm of supervised learning, with a specific focus on regression analysis. Supervised learning is a subfield of machine learning that enables us to build predictive models based on labeled data, where the outcome variable (dependent variable) is known. In our case, this approach allows us to predict and model essential aspects of public transportation with the aim of making informed decisions and driving improvements.

**3.1 Regression Tasks**

Our regression analysis focuses on specific tasks that align with our project objectives. These tasks involve predicting and modeling key metrics and performance indicators, including:

1. On-Time Performance Prediction: We aim to predict on-time performance, a critical measure of the reliability and punctuality of public transportation services. Understanding the factors influencing on-time performance is essential for improving service quality.
2. Passenger Satisfaction Modeling: Passenger satisfaction is a fundamental aspect of public transportation. We intend to model passenger feedback and satisfaction scores to gain insights into the aspects of service that matter most to passengers.
3. Service Efficiency Optimization: Modeling service efficiency metrics allows us to identify areas where public transportation can be optimized. Efficiency models assist in resource allocation and improving overall service quality.

**3.2 Feature Selection**

A fundamental step in regression analysis is the selection of relevant features (independent variables) that have an impact on the target variable. Our feature selection process involved evaluating the significance of various factors, including:

* Route Information: Factors related to specific routes, such as route IDs, stops, and schedules, can influence performance and efficiency.
* Temporal Factors: Variables related to time and schedules, such as day of the week and time of day, play a pivotal role in on-time performance analysis.
* Passenger Feedback: Feedback from passengers, encompassing aspects like cleanliness, safety, and accessibility, can be valuable for modeling passenger satisfaction.
* Service Metrics: Metrics related to service operations, such as frequency, capacity, and maintenance, may contribute to efficiency modeling.

**3.3 Model Selection and Training**

In Phase 4, we explored and employed various regression models to fulfill our tasks. Our model selection included traditional regression models as well as more advanced techniques. The chosen models include:

* Linear Regression: A foundational regression model that establishes linear relationships between independent and dependent variables.
* Decision Trees: Decision trees provide a clear visual representation of how decisions lead to outcomes. They are particularly useful for modeling passenger satisfaction.
* Random Forests: A robust ensemble technique that combines multiple decision trees for increased predictive power.

These models were trained using our meticulously prepared dataset, with appropriate data splitting into training and testing sets to ensure model generalization.

**3.4 Model Evaluation and Interpretation**

The performance of our regression models was evaluated using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2). These metrics helped us assess the accuracy and suitability of our models for the tasks at hand. Furthermore, the interpretation of model results allowed us to gain insights into the relationships between independent variables and the dependent variables, shedding light on the factors influencing on-time performance, passenger satisfaction, and service efficiency.

**3.5 Predictive Analysis**

The regression models developed in this phase serve as powerful tools for predictive analysis. We can now make informed predictions regarding on-time performance, passenger satisfaction, and service efficiency based on real-time or hypothetical scenarios. These predictions empower decision-makers with the ability to assess the impact of potential changes and interventions.

The results of our regression analysis are poised to guide transportation improvement initiatives and foster data-driven decision-making. Our journey into the world of supervised learning and regression equips us with the tools needed to enhance the public transportation experience.

**SOURCE CODE**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load your dataset (replace 'your\_dataset.csv' with the actual file path)

data = pd.read\_csv('/content/20140711 (1).csv')

# Basic data exploration

print(data.head()) # Display the first few rows of the dataset

print(data.info()) # Get information about the dataset, e.g., data types

# Visualization: On-Time Performance Over Time

plt.figure(figsize=(12, 6))

sns.lineplot(x='WeekBeginning', y='NumberOfBoardings', data=data)

plt.xlabel('Week Beginning')

plt.ylabel('Number of Boardings')

plt.title('On-Time Performance Over Time')

plt.xticks(rotation=45)

plt.grid(True)

plt.show()

# Visualization: Distribution of Number of Boardings

plt.figure(figsize=(10, 6))

sns.histplot(data['NumberOfBoardings'], bins=30, kde=True)

plt.xlabel('Number of Boardings')

plt.ylabel('Frequency')

plt.title('Distribution of Number of Boardings')

plt.grid(True)

plt.show()

**OUTPUT**

TripID RouteID StopID StopName WeekBeginning \

0 23631 100 14156 181 Cross Rd 6/30/2013 0:00

1 23631 100 14144 177 Cross Rd 6/30/2013 0:00

2 23632 100 14132 175 Cross Rd 6/30/2013 0:00

3 23633 100 12266 Zone A Arndale Interchange 6/30/2013 0:00

4 23633 100 14147 178 Cross Rd 6/30/2013 0:00

NumberOfBoardings

0 1

1 1

2 1

3 2

4 1

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1048575 entries, 0 to 1048574

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 TripID 1048575 non-null int64

1 RouteID 1048575 non-null object

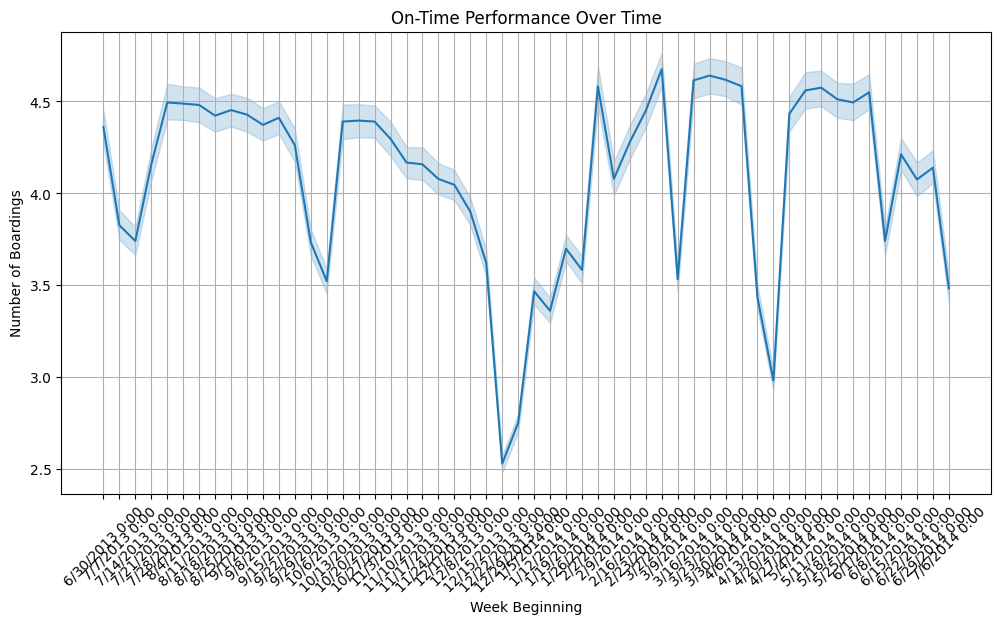
2 StopID 1048575 non-null int64

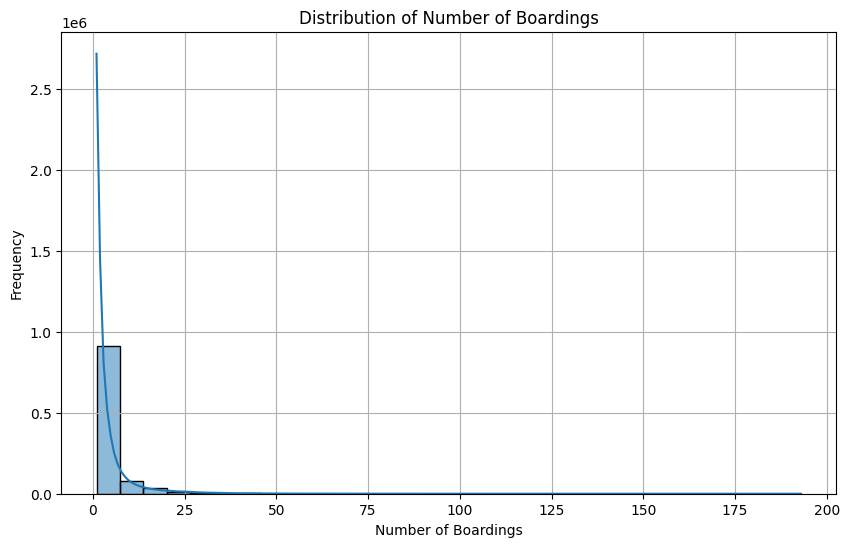
3 StopName 1048575 non-null object

4 WeekBeginning 1048575 non-null object

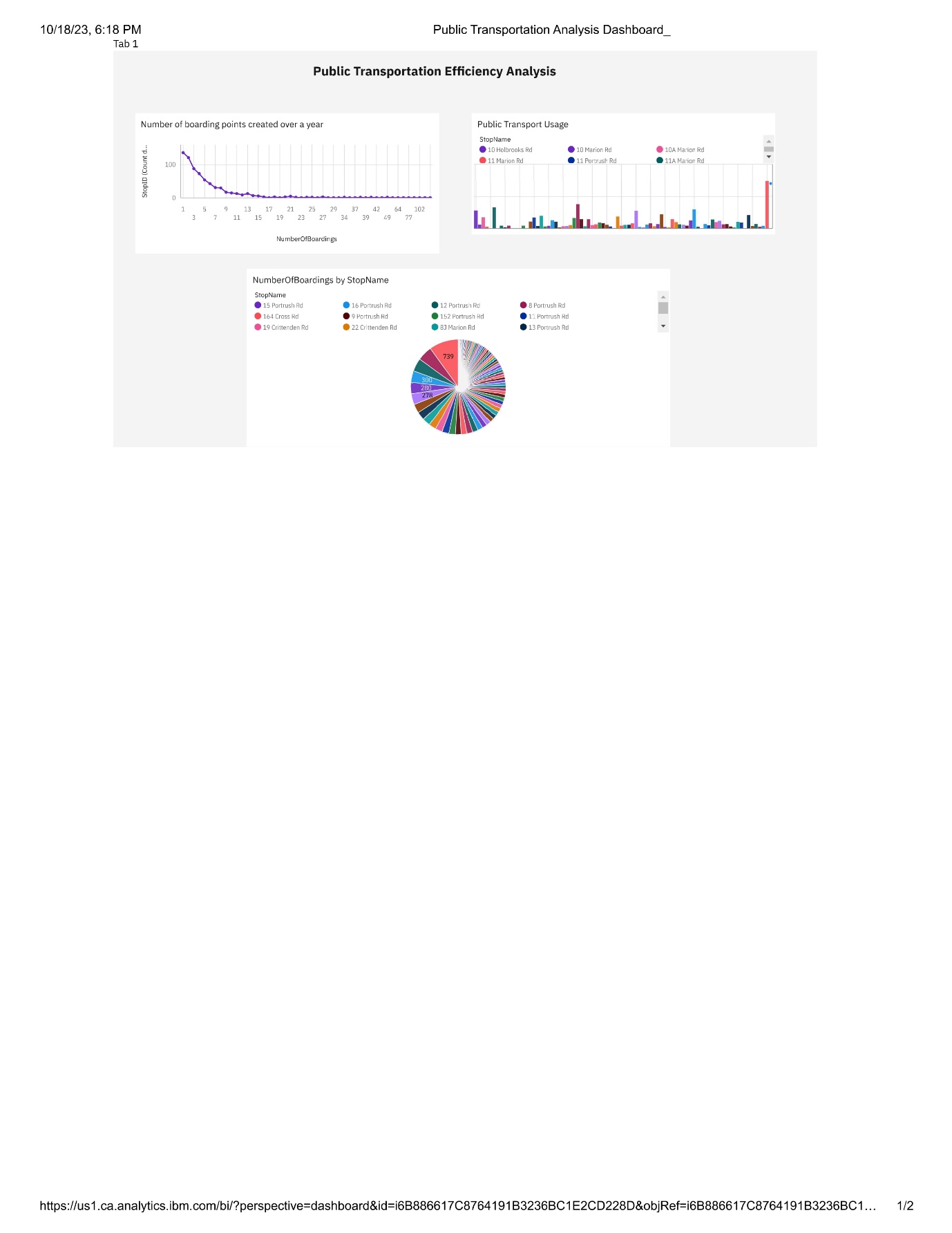
5 NumberOfBoardings 1048575 non-null int64

dtypes: int64(3), object(3)

****memory usage: 48.0+ MB

****

**IBM Cognos**

****