PUBLIC TRANSPORTATION ANALYSIS USING MACHINE LEARNING

Phase-3 submission document

Project Title: Public transportation analysis

Phase 3: Development part 1

Topic: Start building the public transportation analysis by loading and preprocessing the dataset.

PUBLIC TRANSPORTATION ANALYSIS

Introduction:

- Public transportation is an indispensable element of modern urban living, offering a lifeline for millions of people who rely on buses, trains, subways, trams, and other transit modes to commute, access essential services, and reduce the environmental impact of personal vehicle use.

- However, despite its undeniable importance, public transportation systems often face a multitude of challenges that necessitate rigorous analysis and problem-solving.

- This analysis seeks to address the complex and pressing issues that plague public transportation networks, which impact the quality of service, ridership numbers, environmental sustainability, and the economic viability of these systems.

- This journey begins with the fundamental steps of data loading and preprocessing.

- This introduction will guide you through the initial steps of the process. We'll explore how to import essential libraries, load the dataset and perform critical preprocessing steps. Data preprocessing is crucial as it helps clean, format and prepare the data for further analysis. This includes its quality and accuracy.

Given Dataset:

	A	B	C	D	E	F	G
1	TripID	RouteID	StopID	StopName	WeekBegi	NumberOfBoardings	
2	23631	100	14156	181 Cross Rd	########	1	
3	23631	100	14144	177 Cross Rd	########	1	
4	23632	100	14132	175 Cross Rd	########	1	
5	23633	100	12266	Zone A Arndale Interchange	########	2	

1048571	45682	171	14325	16 Fullarton Rd	########	1
1048572	45682	171	13929	8 Fullarton Rd	########	2
1048573	45682	171	13758	3 Glen Osmond Rd	########	3
1048574	45682	171	13967	9 Fullarton Rd	########	1
1048575	45682	171	13808	5 Fullarton Rd	########	1
1048576	45682	171	13845	6 Fullarton Rd	########	3

Importance of loading and processing dataset:

Loading and preprocessing datasets are critical steps in the data analytics process, and they carry significant importance for several reasons:

- **Data Quality Assurance**
- **Data Understanding**
- **Data Consistency**
- **Missing Data Handling**
- **Feature Engineering**

Challenges involved in loading and pre processing public transportation analysis:

Loading and preprocessing data for public transportation analysis can be particularly challenging due to the nature of transportation data, which often includes vast and complex datasets from various sources. Here are some of the common challenges involved:

- **Real-Time Data**
- **Data Volume**
- **Data Variety**
- **Data Quality**

How to overcome the challenges of loading and preprocessing a public transportation analysis dataset:

Overcoming the challenges involved in loading and preprocessing data for public transportation analysis requires a combination of technical solutions, data governance, and effective strategies. Here are some ways to address these challenges:

- **Data Quality Assurance**
- **Data Integration and Standardization**
- **Real-Time Data Processing**
- **Data Privacy and Security**
- **Geospatial Data Handling**
- **Data Security**

1.Loading the dataset:

Loading a dataset is a fundamental step in data analytics, where you import your data into the chosen data analysis tool or software for further exploration and analysis. The specific steps for loading a dataset can vary depending on the software or programming language you are using. Here's a general overview of how to load a dataset:

- **Choose Your Data Analysis Tool**
- **Prepare Your Data File**
- **Import Your Data**
  a. **Python (using Pandas)**
  b. **R**

Some common data preprocessing tasks include:

Data cleaning:

- This involves identifying and correcting errors and inconsistencies in the data.
- For example, this may involve removing duplicate records, correcting typos, and filling in missing values.

Data transformation:

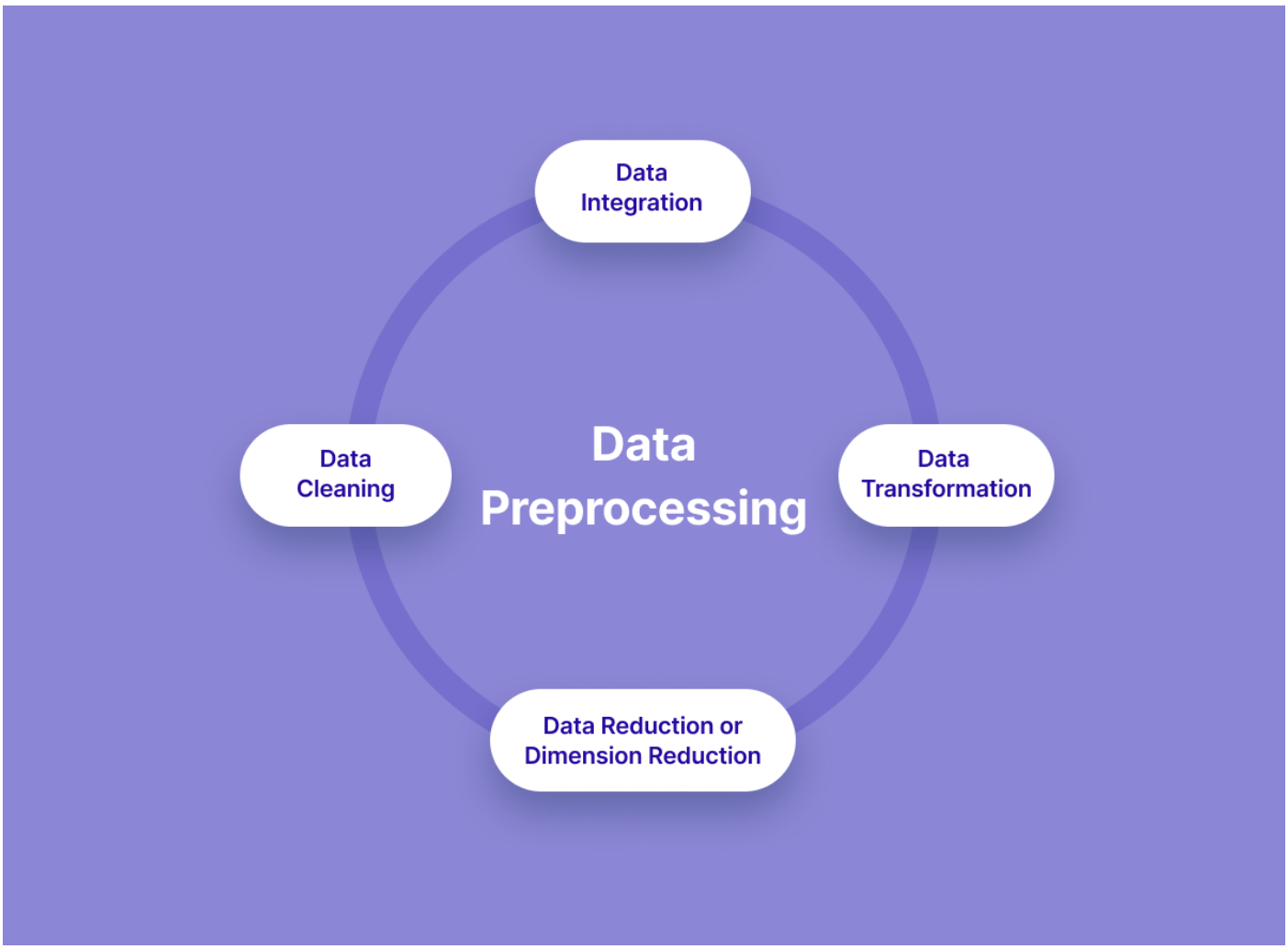
- This involves converting the data into a format that is suitable for the analysis task.
- For example, this may involve converting categorical data to numerical data, or scaling the data to a suitable range.

Handling Duplicate Data:

- Identify and remove duplicate records to avoid skewing analysis results.

Data Visualization:

- Use data visualization techniques to explore the dataset, identify trends, and gain insights before analysis begins.



1.Load the Dataset:

Load your dataset into a Pandas Data Frame. You can typically find passenger datasets in CSV format, but you can adapt this code to other formats as needed.

Program:

```
df=pd.read_csv("C:\\Users\\SWETHA.U\\Downloads\\archive\\20140711.CSV")
print(df)
```

Output:

```
          TripID RouteID  StopID                    StopName  \
0          23631     100   14156                181 Cross Rd
1          23631     100   14144                177 Cross Rd
2          23632     100   14132                175 Cross Rd
3          23633     100   12266  Zone A Arndale Interchange
4          23633     100   14147                178 Cross Rd
...          ...     ...     ...                         ...
10857229   13346    W91C   14629                21 Cashel St
10857230   13346    W91C   14708                22 Cashel St
10857231   13346    W91C   13709              2 Greenhill Rd
10857232   13346    W91C   14029                  10 East Av
10857233   13346    W91C   13824                 6 Leader St

                WeekBeginning  NumberOfBoardings
0         2013-06-30 00:00:00                  1
1         2013-06-30 00:00:00                  1
2         2013-06-30 00:00:00                  1
3         2013-06-30 00:00:00                  2
4         2013-06-30 00:00:00                  1
...                       ...                ...
10857229  2014-07-06 00:00:00                  1
10857230  2014-07-06 00:00:00                  3
10857231  2014-07-06 00:00:00                  1
10857232  2014-07-06 00:00:00                  1
10857233  2014-07-06 00:00:00                  1

[10857234 rows x 6 columns]
```

2. Preprocessing:

Program:

```
import pandas as pd


# Load the dataset into a pandas DataFrame
df = pd.read_csv('/content/Dataset.csv')


# Remove rows with any missing values in any column
```

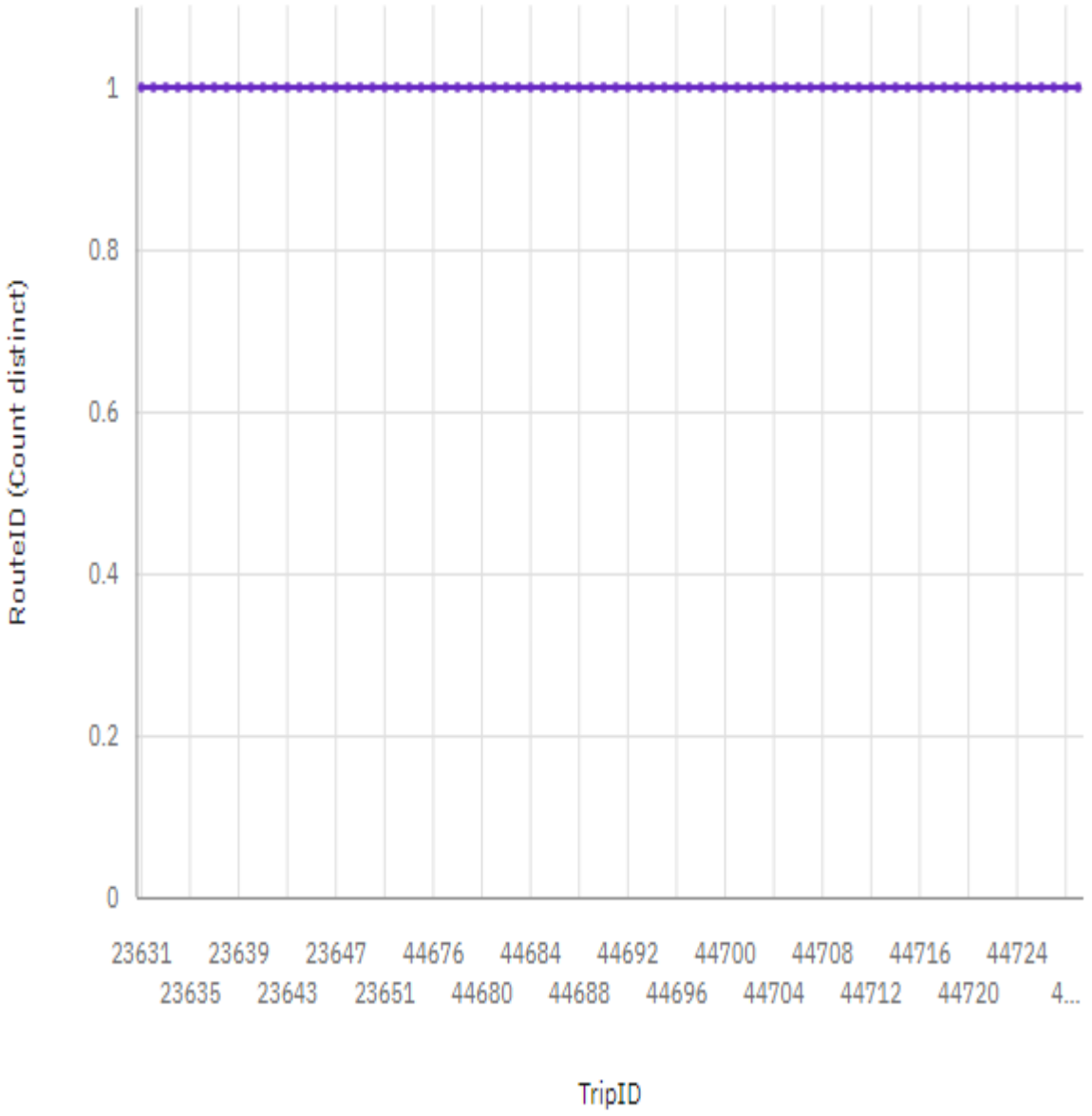
```
df_cleaned = df.dropna()

# Or, fill missing values in specific columns (if applicable)
# df_cleaned['NumberOfBoardings'].fillna(0, inplace=True)

# Remove duplicate rows based on all columns
df_cleaned = df_cleaned.drop_duplicates()

# Convert specific columns to appropriate data types
df_cleaned['TripID'] = df_cleaned['TripID'].astype(int)
df_cleaned['NumberOfBoardings'] = df_cleaned['NumberOfBoardings'].astype(int)
# Convert other columns as needed

# Extract day, month, year, etc. from 'WeekBeginning' column
df_cleaned['WeekBeginning'] = pd.to_datetime(df_cleaned['WeekBeginning'])
df_cleaned['Year'] = df_cleaned['WeekBeginning'].dt.year
df_cleaned['Month'] = df_cleaned['WeekBeginning'].dt.month
df_cleaned['Day'] = df_cleaned['WeekBeginning'].dt.day
# Extract other features as needed

# Save the cleaned dataset to a new CSV file
df_cleaned.to_csv('project_dataset.csv', index=False)
```

Visualization:

Visualization in data analytics is the process of representing data graphically to gain insights, identify patterns, and communicate findings effectively. Data visualization plays a crucial role in the data analysis process as it helps analysts and stakeholders better understand complex datasets, make data-driven decisions, and communicate their findings to others.

Using IBM Cognos for Visualization:

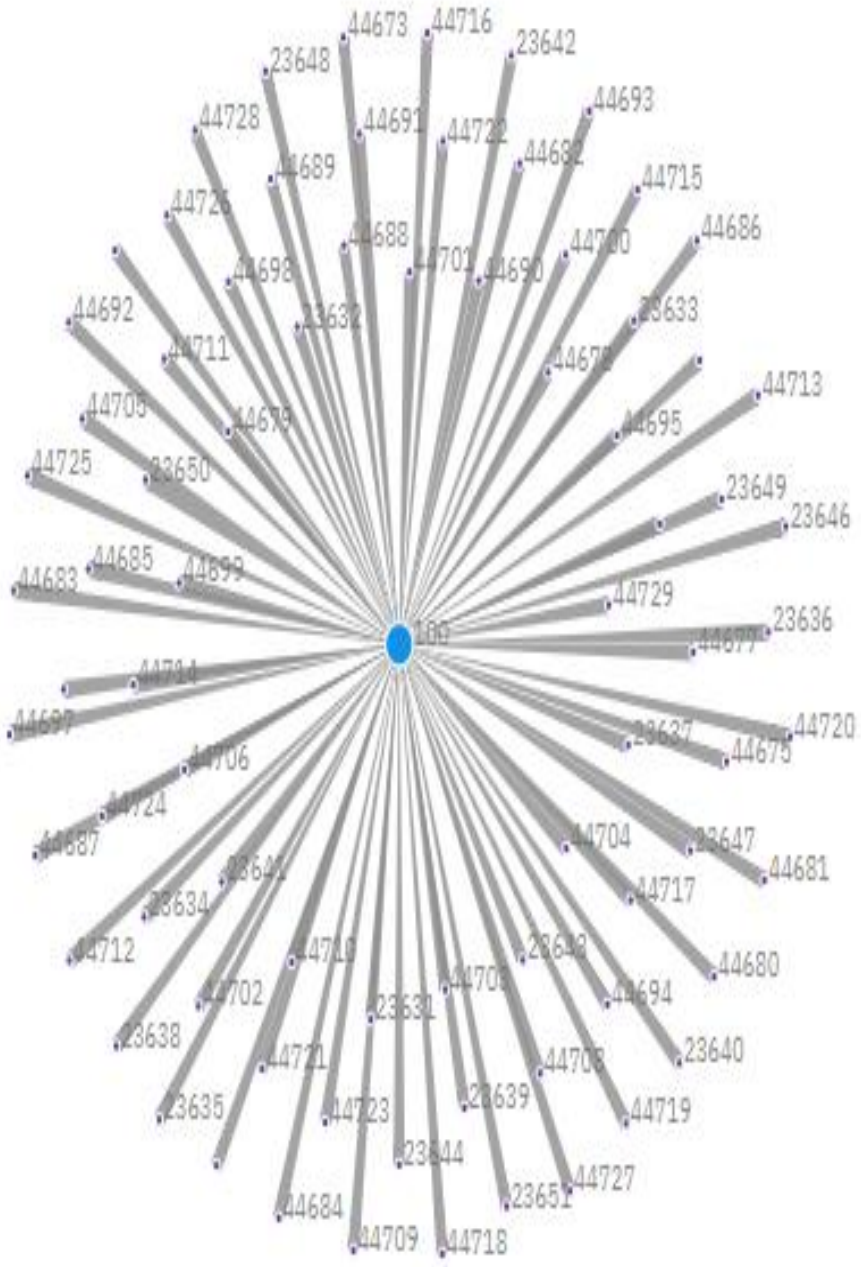
RouteID by TripID



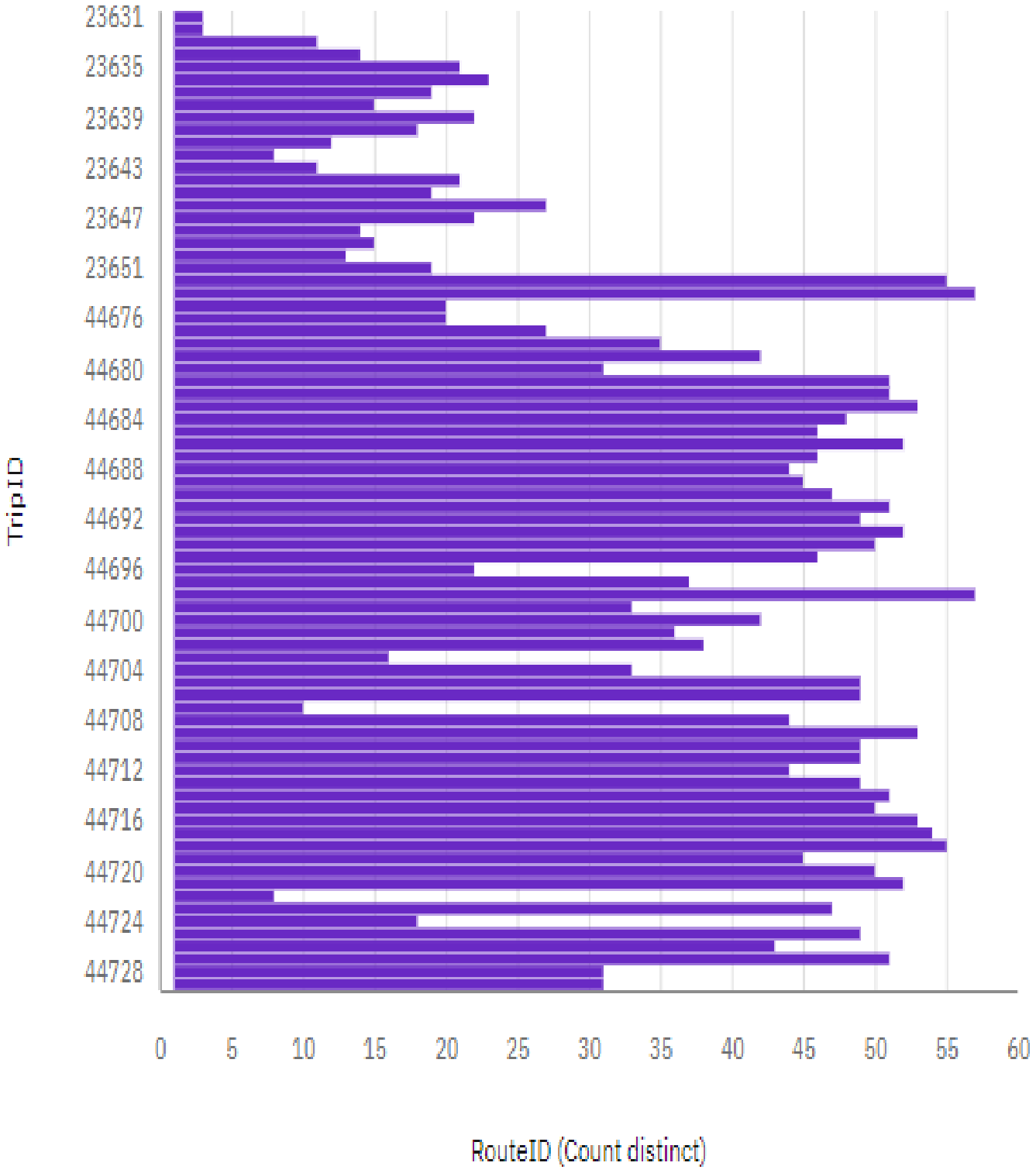
TripID to RouteID

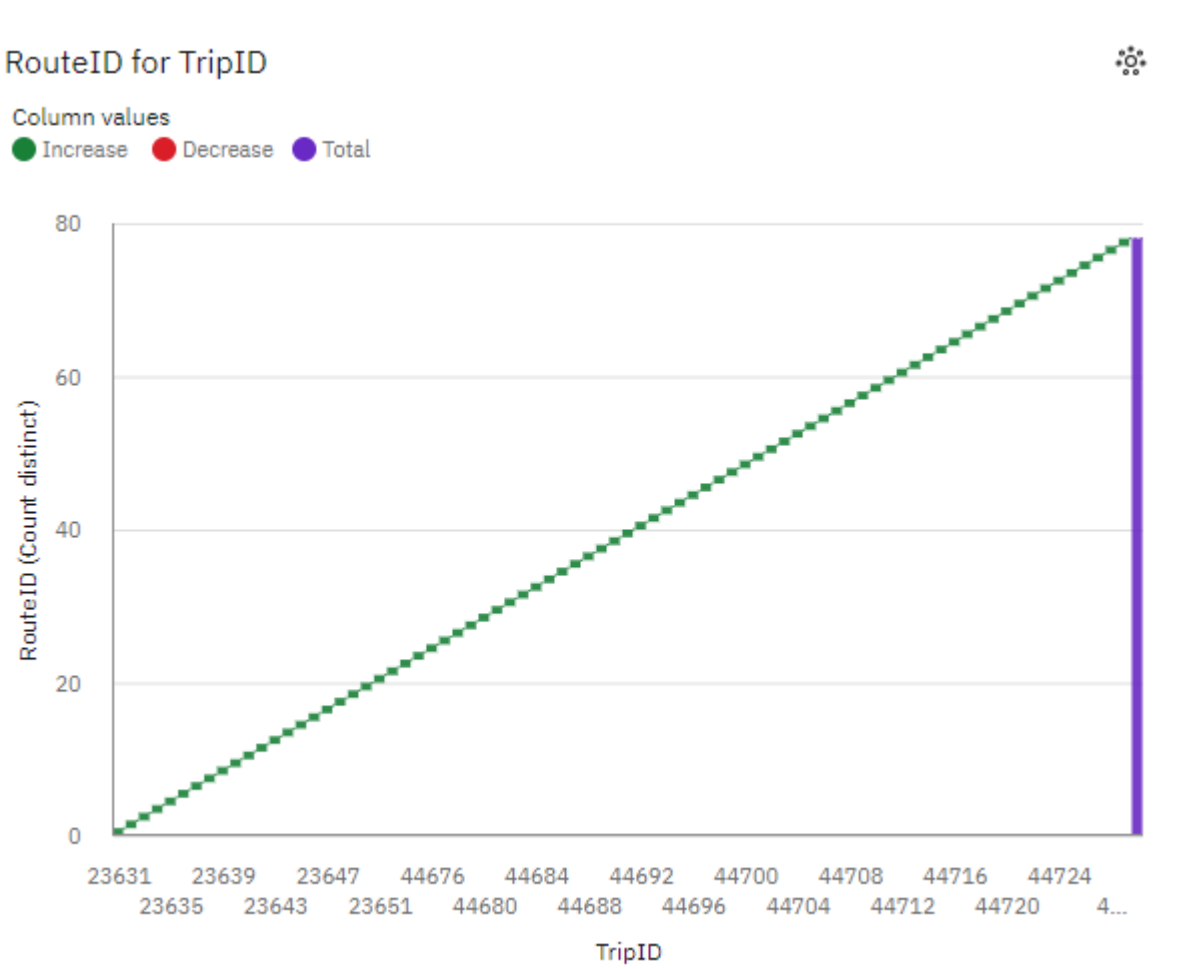
From
TripID

To
RouteID



RouteID by TripID





Conclusion:

In our journey to analyse public transportation, we have undertaken a comprehensive exploration of a crucial aspect of urban infrastructure. This analysis began with data collection, organization, and preprocessing, similar to our house price prediction model project. These steps are vital for effective analysis of public transportation systems.

Understanding the structure and characteristics of public transportation data, along with identifying any potential issues, has been pivotal. Exploratory data analysis (EDA) has provided valuable insights into the efficiency, accessibility, and reliability of public transportation services within a given area.

Data preprocessing, just like in the house price prediction model project, has played a central role in this analysis. Cleaning, transforming, and refining the transportation dataset have been necessary to ensure that it aligns with the requirements for further analysis and decision-making regarding improvements or policy changes.

With these foundational steps completed, our dataset is now well-prepared for more advanced analyses, such as route optimization, demand forecasting, or identifying areas for investment and enhancement in public transportation services. This comprehensive analysis will ultimately contribute to informed decision-making and

potentially lead to improvements in the public transportation system, benefiting the community as a whole.