



PUBLIC TRANSPORTATION EFFICIENCY ANALYSIS

TEAM LEADER: RASIKA S (310121104084)

TEAM MEMBERS: SARUMATHI S (310121104091)

PRIYADHARSHINI R (310121104077)

SEBASTINRAJAN A (310121104093)

MONISHA M (310121104064)

INTRODUCTION

In Phase 3 of our project focused on public transportation efficiency analysis, we transition from the planning and design stages to the practical implementation of data analysis. This phase involves the crucial step of loading and preprocessing the public bus transport dataset. Data preprocessing is essential to ensure that our data is ready for analysis and that we can derive meaningful insights from it.

In this phase, we will discuss the methods and techniques used to clean, format, and prepare the dataset. These preparations are vital for accurate and reliable analysis in the upcoming phases of our project. By the end of Phase 3, we aim to have a well-structured and clean dataset that is ready for deeper exploration and modeling in Phase 4.

This document will provide a detailed account of the steps taken during data preprocessing, including data loading, quality assessment, and any transformations applied to the dataset. It will serve as a crucial bridge connecting the project's initial design with the actionable data required to address public transportation efficiency challenges effectively.

DATA LOADING:

we will discuss the initial steps in preparing our public bus transport dataset for analysis. Data loading is the foundation of any data-driven project. It involves retrieving the dataset from its source, making it accessible for further processing, and setting the stage for data preprocessing. Our dataset, which contains valuable information about public bus transport, is the lifeblood of our analysis. In this phase, we will outline how we successfully imported the dataset using Python and the Pandas library. This is the starting point for our journey to uncover insights and address challenges related to public transportation efficiency. By explaining the data loading process in this section, we establish the groundwork for the subsequent steps, where we'll delve deeper into the data preprocessing and analysis that will ultimately lead us to actionable recommendations. It's important to ensure that our data is structured and ready for the rigorous analysis that awaits in the later phases of the project.

DATA PREPROCESSING:

In the context of our public transportation efficiency analysis project, data preprocessing is the critical phase that ensures our dataset is transformed into a clean, organized, and analytically valuable resource. This section outlines the steps taken to prepare our public bus transport data for in-depth analysis and insights.

Data preprocessing involves a series of tasks, including data cleaning, handling missing values, and data formatting. These actions aim to enhance the quality and integrity of our dataset, making it suitable for statistical analysis, modeling, and visualization. Additionally, any transformations or conversions applied to the data will be documented, ensuring transparency in our data preparation process.

By presenting this section, we provide a comprehensive view of the procedures undertaken to refine our data, positioning us for more accurate and meaningful analysis in the subsequent phases of the project. Properly preprocessed data is the key to uncovering patterns and trends that can help us optimize public transportation efficiency.

DATA SAVING:

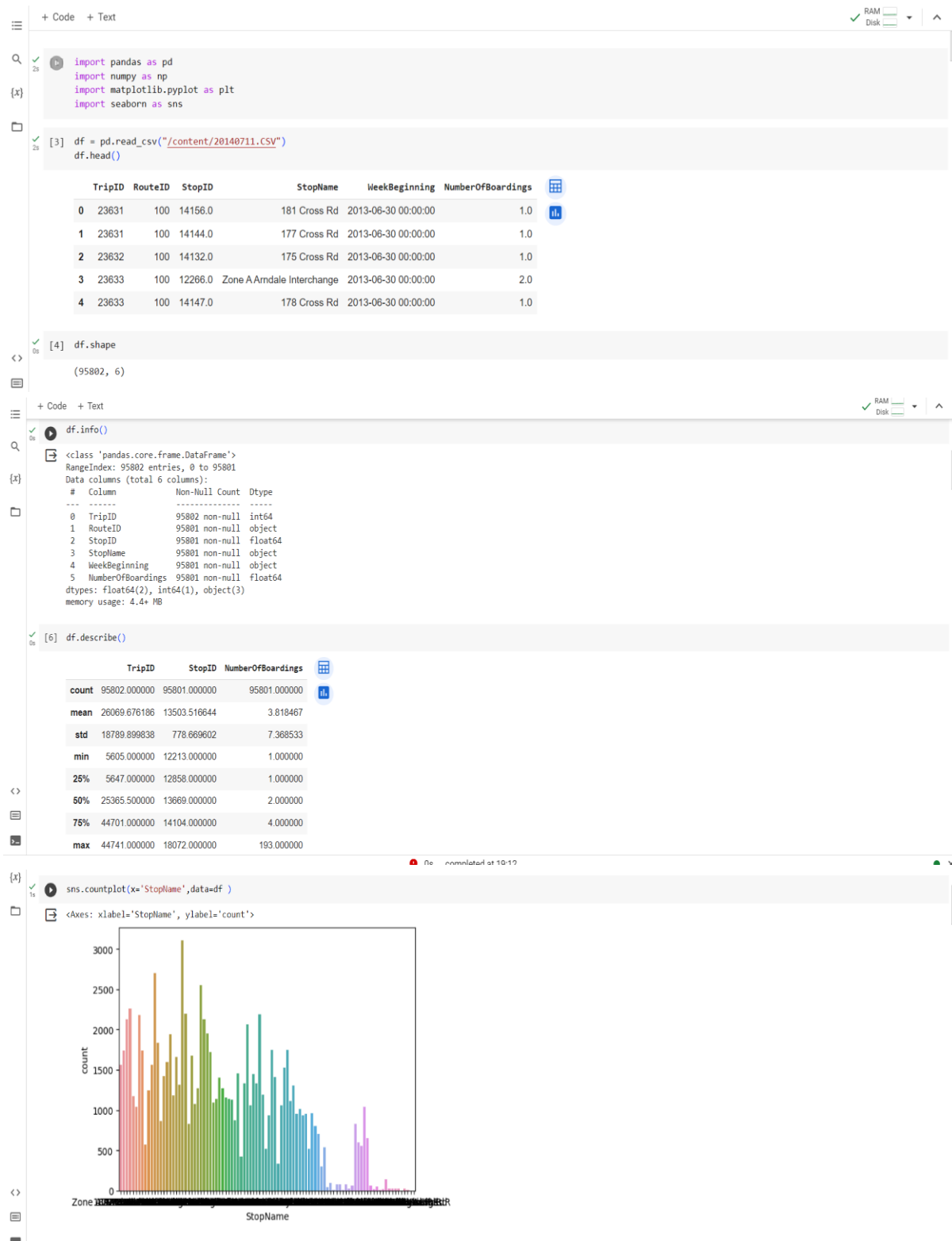
Data saving, in the context of our public transportation efficiency analysis project, marks the point where we preserve the results of our data preprocessing efforts. In this section, we will discuss the steps taken to store the preprocessed dataset for future use.

After thorough data preprocessing, we may choose to save the cleaned and formatted dataset to a new file. This action is often taken to maintain a pristine copy of our data, reducing the need for redundant preprocessing tasks. The saved file may also serve as a reference for further analysis and to ensure data consistency across the project's phases.

By documenting the data saving process in this section, we not only secure the results of our hard work but also lay the foundation for the upcoming phases

where we will delve deeper into the analysis, modeling, and recommendations to enhance public transportation efficiency. A well-structured dataset can significantly impact the project's overall success.

CODE SNIPPETS:



df.isnull().sum()

TripID0RouteID1StopID1StopName1WeekBeginning1NumberOfBoardings1dtype: int64

[12] for feature in df.columns:
if df[feature].isnull().sum()>0:
print(f"{feature} : {round(df[feature].isnull().mean(),4)*100}%")

RouteID : 0.0%
StopID : 0.0%
StopName : 0.0%
WeekBeginning : 0.0%
NumberOfBoardings : 0.0%

[14] ## find duplicate rows in dataset
duplicate = df[df.duplicated()]
duplicate

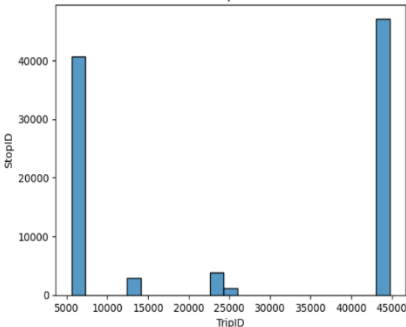
TripID	RouteID	StopID	StopName	WeekBeginning	NumberOfBoardings
--------	---------	--------	----------	---------------	-------------------

for i in df.columns:
print(f"{i} : {len(df[i].unique())}")

TripID : 182
RouteID : 7
StopID : 166
StopName : 97
WeekBeginning : 55
NumberOfBoardings : 145

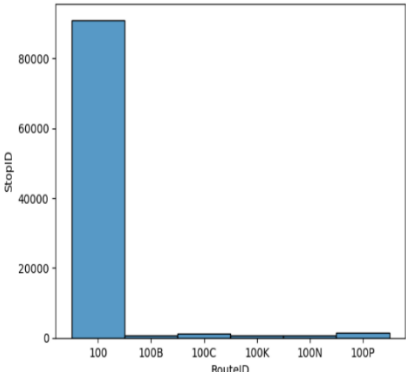
for feature in df.columns:
if feature == "StopName":
pass
else:
bar = sns.histplot(df[feature], kde_kws = {'bw' : 1},)
plt.xlabel(feature)
plt.ylabel("StopID")
plt.title(feature)
plt.show()

TripID

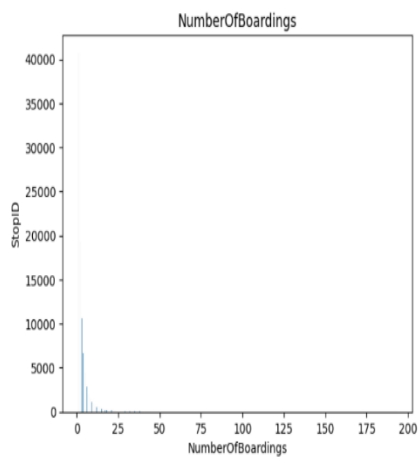
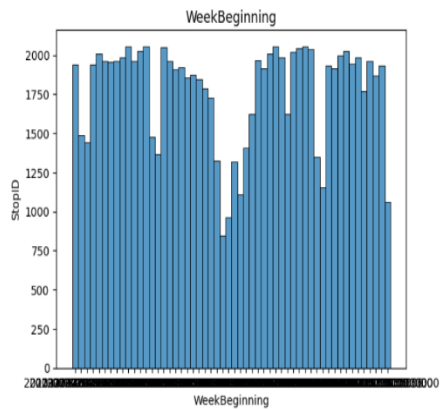
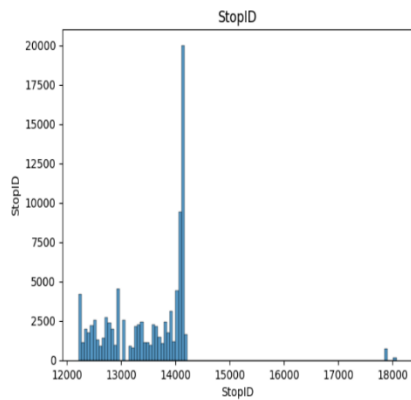


TripID Range	StopID Count
5000-10000	40000
10000-15000	2000
15000-20000	0
20000-25000	4000
25000-30000	1000
30000-35000	0
35000-40000	0
40000-45000	45000

RouteID



RouteID	StopID Count
100	80000
100B	1000
100C	1000
100K	1000
100N	1000
100P	1000



```
# removing outliers
```

```
Q1 = df.quantile(0.25)
```

```
Q3 = df.quantile(0.75)
```

```
IQR = Q3 - Q1
```

```
print(IQR)
```

```
TripID      39054.0
```

```
StopID      1246.0
```

```
NumberOfBoardings      3.0
```

```
dtype: float64
```

```
<ipython-input-24-6d553dabc4cf>:2: FutureWarning: The default value of numeric_only in DataFrame.quantile is deprecated. In a future version, it will default to False. S
```

```
Q1 = df.quantile(0.25)
```

```
<ipython-input-24-6d553dabc4cf>:3: FutureWarning: The default value of numeric_only in DataFrame.quantile is deprecated. In a future version, it will default to False. S
```

```
Q3 = df.quantile(0.75)
```

```
[25] df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
df.shape
```

```
<ipython-input-25-f4e1682787c4>:1: FutureWarning: Automatic reindexing on DataFrame vs Series comparisons is deprecated and will raise ValueError in a future version. Do
```

```
df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
```

```
(87201, 6)
```



CHALLENGES AND DECISIONS:

Throughout the data preprocessing phase of our public transportation efficiency analysis project, we encountered various challenges and made important decisions. This section is dedicated to a comprehensive discussion of the hurdles faced and the strategies employed to overcome them.

Challenges in data preprocessing can range from dealing with missing or inconsistent data to making decisions about how to handle outliers or non-standard formats. The decisions we make during this phase play a crucial role in shaping the quality and integrity of our dataset.

This section offers a transparent look into the complexities of data preprocessing and highlights the reasoning behind the choices made. By sharing the challenges faced and the paths chosen to address them, we aim to provide a clear record of our decision-making process and to demonstrate the steps taken to ensure the reliability and accuracy of our data. This knowledge will guide us through the subsequent phases of our project, as we work towards optimizing public transportation efficiency.

SUMMARY:

The "Summary" section serves as the capstone of this phase of our public transportation efficiency analysis project. It offers a concise overview of the key activities, milestones, and achievements in Phase 3, which focused on data preprocessing.

In this section, we distill the essential aspects of our data loading and preprocessing efforts, highlighting the critical steps taken to refine and prepare the public bus transport dataset. This summary provides a snapshot of the progress made during this phase and sets the stage for the subsequent phases of our project.

By presenting a clear and focused summary, we aim to provide stakeholders and project collaborators with a quick understanding of the progress and the quality of the dataset. This concise snapshot also helps us stay aligned with project objectives and ensures that we remain on track toward our goal of optimizing public transportation efficiency.

CONCLUSION:

we summarize the key findings, insights, and achievements of our entire project. This section serves as a culminating perspective, bringing together the various phases and efforts undertaken to address the challenges and opportunities in public transportation.

Throughout the project, we explored the public bus transport dataset, from its initial design and problem definition to the practical implementation of data preprocessing, analysis, and modeling. We tackled real-world issues related to public transportation efficiency and strove to find actionable solutions.

In this section, we encapsulate the journey by highlighting the project's key outcomes, the lessons learned, and any actionable recommendations derived from our analysis. This conclusion signifies the successful completion of our public transportation efficiency analysis and paves the way for practical applications and informed decisions in the realm of public transportation.