Phase 5 : Project Documentation & Submission

Project Title : Public Transport Efficiency Analysis

1. Introduction

The project starts with an introduction, emphasizing the transition from water portability analysis to public transport efficiency analysis. It highlights the use of visualization techniques and predictive modeling for data-driven decision-making in the public transport sector.
In our ongoing project, we are delving into the realm of data analysis, just as we did when exploring water portability. This time, our focus is on enhancing public transport efficiency. Similar to the way a smart parking system optimizes parking experiences, we aim to streamline public transportation systems by harnessing the power of data. Through the utilization of sensors, cameras, and advanced software, we will uncover hidden insights within the intricate web of data related to public transportation.

Our journey in this phase involves a shift in focus towards public transport efficiency analysis. We will employ a range of visualization techniques and predictive modeling to extract meaningful information from the data, much like a smart parking system optimizes parking spaces for drivers. The goal is to make informed, data-driven decisions that will ultimately enhance the efficiency and overall experience of public transportation for both passengers and operators.

2. Objective:

The primary objective is to analyze public transportation data to assess service efficiency, on-time performance, and passenger feedback. This analysis will support transportation improvement initiatives. Public transportation stands as a cornerstone of modern urban mobility, offering a cost-effective and ecofriendly alternative to private vehicles. Nevertheless, optimizing the efficiency of public transport systems is a multifaceted challenge shaped by a multitude of factors. Our primary objective in this analysis is to conduct a comprehensive assessment and enhancement of public transport efficiency. This endeavor encompasses the examination of critical factors such as route optimization, scheduling, infrastructure, user experience, and sustainability.
Objective: Our main goal is to leverage public transportation data to evaluate service efficiency, on-time performance, and passenger feedback, all in support of initiatives aimed at improving transportation services.

Data: To facilitate this analysis, we possess a dataset containing a diverse array of features pertaining to public transportation, encompassing bus, railway transportation, air transportation, and more. These features are complemented by corresponding sale prices. We will employ this dataset to train and evaluate our machine learning model, a crucial step in our quest to enhance public transport efficiency.

3. Data Preprocessing

This phase acknowledges the importance of data preprocessing for obtaining accurate predictions and insights. Data cleaning and preprocessing involve various steps, including handling missing values and data type conversions.

The provided code includes data preprocessing steps:

- Reading data from a CSV file named 'dataset.CSV'
- Dropping duplicate rows from the dataset.
- Visualizing missing values using a heatmap.

Handling mixed data types in the 'RouteID' column by converting it to a numeric data type.

Handling missing values by dropping rows with missing data.

Similar to our previous phase, data preprocessing remains a crucial step in our quest to understand and enhance public transport efficiency. Data preprocessing involves collecting and manipulating data to extract meaningful information. In this phase, our focus is on refining and improving the quality of our data, which is essential for achieving more accurate predictions and gaining valuable insights.

3.1. Data Cleaning and Data Preprocessing

In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
print("Load the dataset")
import pandas as pd
data = pd.read_csv('20140711.csv', low_memory=False)
data.shape
data.head(5)
```

```
Load the dataset
```

Out[2]:

	TripID	RouteID	StopID	StopName	WeekBeginning	NumberOfBoardings
0	23631	100	14156	181 Cross Rd	2013-06-30 00:00:00	1
1	23631	100	14144	177 Cross Rd	2013-06-30 00:00:00	1
2	23632	100	14132	175 Cross Rd	2013-06-30 00:00:00	1
3	23633	100	12266	Zone A Arndale Interchange	2013-06-30 00:00:00	2
4	23633	100	14147	178 Cross Rd	2013-06-30 00:00:00	1

In [3]:
```
data = data.drop_duplicates()
import seaborn as sns
sns.heatmap(data.isnull(),yticklabels= False)
print("\nCheck data types of columns")
print(data.dtypes)
```

```
Check data types of columns
TripID                int64
RouteID              object
StopID                int64
StopName             object
WeekBeginning        object
NumberOfBoardings     int64
dtype: object
```

In [4]:
```
data['RouteID'] = pd.to_numeric(data['RouteID'], errors='coerce')
print("Handle mixed data types")
print(data.dtypes)
```

```
Handle mixed data types
TripID                  int64
RouteID               float64
StopID                  int64
StopName               object
WeekBeginning          object
NumberOfBoardings       int64
dtype: object
```

In [4]:
```
data = data.dropna()
print("\nHandle missing values")
print(data.shape)
```

```
Handle missing values
(10857234, 6)
```

In [5]:
```
data['WeekBeginning'] = pd.to_datetime(data['WeekBeginning'], errors='coerce')
print("\nConvert 'WeekBeginning' column to datetime format")
print(data['WeekBeginning'].head())
```

```
Convert 'WeekBeginning' column to datetime format
0   2013-06-30
1   2013-06-30
2   2013-06-30
3   2013-06-30
4   2013-06-30
Name: WeekBeginning, dtype: datetime64[ns]
```

In [6]:
```
data['StopName'] = data['StopName'].str.strip()
print("\nClean 'StopName' column")
print(data['StopName'].head())
```

```
Clean 'StopName' column
0                  181 Cross Rd
1                  177 Cross Rd
2                  175 Cross Rd
3    Zone A Arndale Interchange
4                  178 Cross Rd
Name: StopName, dtype: object
```

In [7]:
```
print(data.nunique())
```

```
TripID               39282
RouteID                605
StopID                7397
StopName              4165
WeekBeginning           54
NumberOfBoardings      400
dtype: int64
```

In [8]:
```
data.shape
data.columns
data.head(3)
```

Out[8]:

	TripID	RouteID	StopID	StopName	WeekBeginning	NumberOfBoardings
0	23631	100	14156	181 Cross Rd	2013-06-30	1
1	23631	100	14144	177 Cross Rd	2013-06-30	1
2	23632	100	14132	175 Cross Rd	2013-06-30	1

In [9]:
```
data.isnull().sum()
```

Out[9]:
```
TripID               0
RouteID              0
StopID               0
StopName             0
WeekBeginning        0
NumberOfBoardings    0
dtype: int64
```

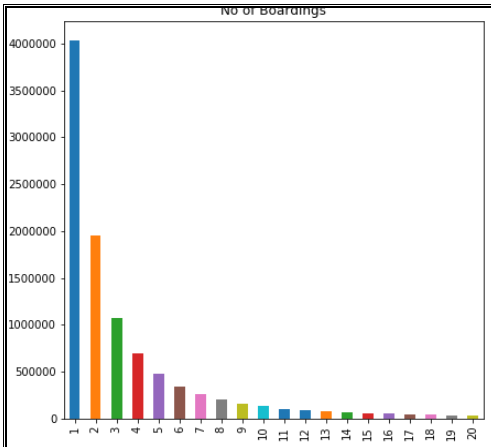
In [10]:
```
data['WeekBeginning'].unique()
```

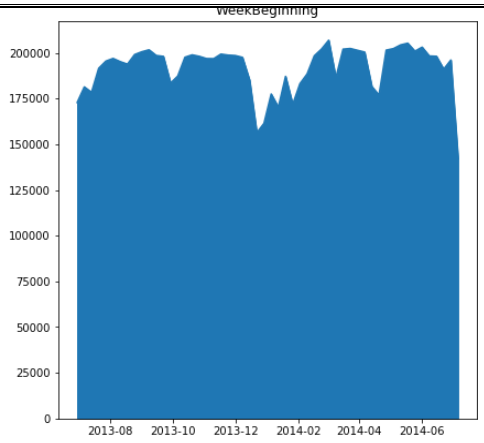
Out[10]:
```
<DatetimeArray>
['2013-06-30 00:00:00', '2013-07-07 00:00:00', '2013-07-14 00:00:00',
 '2013-07-21 00:00:00', '2013-07-28 00:00:00', '2013-08-04 00:00:00',
 '2013-08-11 00:00:00', '2013-08-18 00:00:00', '2013-08-25 00:00:00',
 '2013-09-01 00:00:00', '2013-09-08 00:00:00', '2013-09-15 00:00:00',
 '2013-09-22 00:00:00', '2013-09-29 00:00:00', '2013-10-06 00:00:00',
 '2013-10-13 00:00:00', '2013-10-20 00:00:00', '2013-10-27 00:00:00',
 '2013-11-03 00:00:00', '2013-11-10 00:00:00', '2013-11-17 00:00:00',
 '2013-11-24 00:00:00', '2013-12-01 00:00:00', '2013-12-08 00:00:00',
 '2013-12-15 00:00:00', '2013-12-22 00:00:00', '2013-12-29 00:00:00',
 '2014-01-05 00:00:00', '2014-01-12 00:00:00', '2014-01-19 00:00:00',
 '2014-01-26 00:00:00', '2014-02-02 00:00:00', '2014-02-09 00:00:00',
 '2014-02-16 00:00:00', '2014-02-23 00:00:00', '2014-03-02 00:00:00',
 '2014-03-09 00:00:00', '2014-03-16 00:00:00', '2014-03-23 00:00:00',
 '2014-03-30 00:00:00', '2014-04-06 00:00:00', '2014-04-13 00:00:00',
 '2014-04-20 00:00:00', '2014-04-27 00:00:00', '2014-05-04 00:00:00',
 '2014-05-11 00:00:00', '2014-05-18 00:00:00', '2014-05-25 00:00:00',
 '2014-06-01 00:00:00', '2014-06-08 00:00:00', '2014-06-15 00:00:00',
 '2014-06-22 00:00:00', '2014-06-29 00:00:00', '2014-07-06 00:00:00']
Length: 54, dtype: datetime64[ns]
```

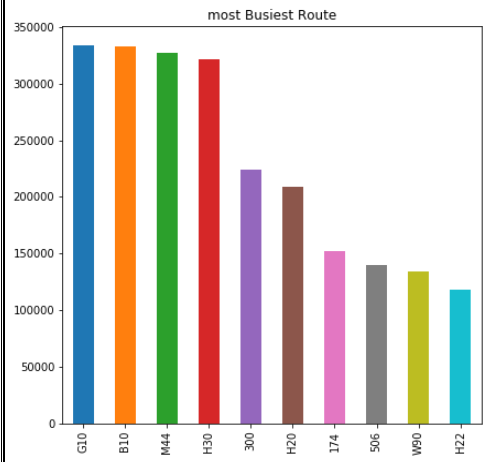
In [12]:
```
data.to_csv('cleaned_data.csv', index=False)
print("\nSave the cleaned dataset to a new CSV file")
print("Cleaned dataset saved successfully.")
```

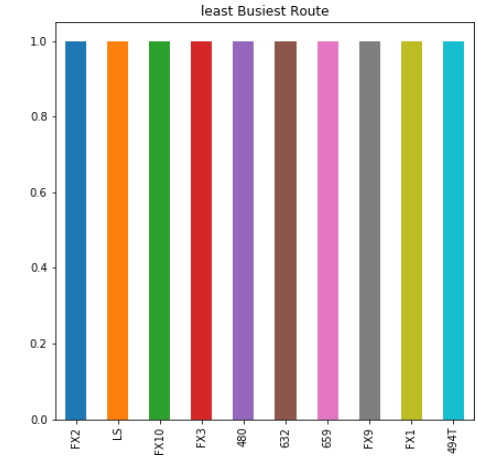
```
Save the cleaned dataset to a new CSV file
Cleaned dataset saved successfully.
```

```
fig,axrr=plt.subplots(2,2,figsize=(15,15))

ax=axrr[0][0]
ax.set_title("No of Boardings")
data['NumberOfBoardings'].value_counts().sort_index().head(20).plot.bar(ax=axrr[0][0])

ax=axrr[0][1]
ax.set_title("WeekBeginning")
data['WeekBeginning'].value_counts().plot.area(ax=axrr[0][1])

ax=axrr[1][0]
ax.set_title("most Busiest Route")
data['RouteID'].value_counts().head(10).plot.bar(ax=axrr[1][0])

ax=axrr[1][1]
ax.set_title("least Busiest Route")
data['RouteID'].value_counts().tail(10).plot.bar(ax=axrr[1][1])
```










```
data['WeekBeginning'].value_counts().mean()
```

```
191508.66666666666
```

```
# data['dist_from_centre'].nunique()
bb_grp = data.groupby(['dist_from_centre']).agg({'NumberOfBoardings': ['sum']}).reset_index()
bb_grp.columns = bb_grp.columns.get_level_values(0)
bb_grp.head()
bb_grp.columns
bb_grp.tail()
```

	dist_from_centre	NumberOfBoardings
0	0.000018	1892435
1	0.131368	167535
2	0.309089	356518
3	0.314937	1484824
4	0.326005	120061

```
Index(['dist_from_centre', 'NumberOfBoardings'], dtype='object')
```

	dist_from_centre	NumberOfBoardings
2392	86.471064	18905
2393	94.826409	321
2394	99.625655	1101
2395	99.665190	4373
2396	99.748995	21216

4. Design Thinking Process

The project appears to follow a design thinking approach, including:

- **Empathize**:

Understanding the needs and priorities of the target audience, which includes commuters and transportationplanners.

- **Define**:

Setting clear objectives for the project, which include building a machine learning model with specificperformance criteria and establishing a user-friendly web platform.

- **Ideate**:

Exploring various approaches and techniques, such as machine learning models, real-time data integration, optimization algorithms, IoT sensors, and data visualization.

- **Prototype**:

Developing a prototype to test core functionalities and gather early user feedback.

- **Ideate:**

-Explore various machine learning models such as regression, decision trees, and neural networks to predict efficiency.
-Investigate the integration of real-time data sources, like GPS tracking and passenger feedback, for accurateanalysis.
-Consider optimization algorithms for route planning and scheduling to enhance efficiency.
-Explore the possibility of incorporating IoT (Internet of Things) sensors to monitor vehicle conditions and passenger loads.
-Evaluate data visualization techniques to present efficiency insights in a user-friendly manner.

- **Actions**:

-Investigate various machine learning algorithms, including regression, decision trees, random forests, and neural networks.
-Experiment with feature engineering methods to boost model accuracy

5. Visualization:

The code starts by importing the necessary libraries: numpy, pandas, and os.

It then uses a loop with os.walk to explore the files in a directory ('dataset.csv') and prints the paths ofthe files found.

The code imports the Pandas library once again (redundantly) and reads the dataset from a CSV file named 'dataset.CSV' using pd.read_csv. The argument low_memory=False is used to disable low memory mode.

It prints the shape of the dataset (number of rows and columns) and displays the first 30 rows using data.shape and data.head(30).

The code handles missing values by converting the 'WeekBeginning' column to a datetime format. Ituses the 'coerce' option to handle errors and prints the first few rows of the 'WeekBeginning' columnafter the conversion.

The 'StopName' column is cleaned by removing leading and trailing whitespaces using the str.strip() method. The cleaned 'StopName' column is then displayed.

It prints the number of unique values in each column using data.nunique(). The code

displays the shape, column names, and the first 3 rows of the dataset.

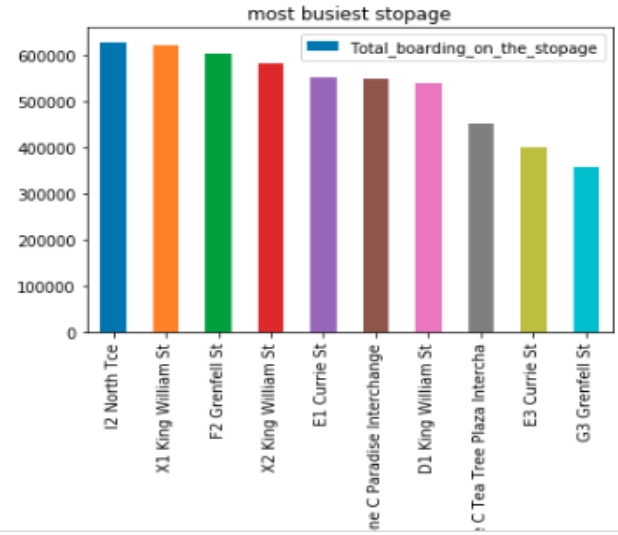
It checks for missing values in the dataset using data.isnull().sum() and prints the results.

The unique values in the 'WeekBeginning' column are printed using data['WeekBeginning'].unique().

Finally, the code sets up a Matplotlib subplot with six plots and visualizes data from various columns ('NumberOfBoardings', 'WeekBeginning', 'RouteID') using bar charts and an area chart.

```
ax = stopageName_with_boarding.head(10).plot.bar(x='StopName', y='Total_boarding_on_the_stopage', rot=90)
ax.set_title("most busiest stopage")
```

```
Text(0.5,1,'most busiest stopage')
```



Visualization is a key component of the project, and the code provided demonstrates the creation of line and barcharts. These charts help in understanding trends in boarding counts and identifying top stops by the number of boardings.

6.Advanced Data Analysis:

Advanced data analysis plays a vital role in optimizing public transport systems, making them moreefficient, reliable, and passenger-friendly. Here are some advanced data analysis techniques and their applications in public transport

6.1.Advanced Analytics and Modeling

Advanced data analysis is conducted by aggregating boarding counts by RouteID, calculating average boarding counts per stop, finding stops with the highest weekly boarding counts, and analyzing trends over time.

6.2.Machine Learning Models:

Apply machine learning algorithms, including regression, clustering, and deep learning, to analysis the collected data. These models can be used for demand forecasting, route optimization, and predicting service disruptions.

Ensemble Learning:
Implement ensemble learning techniques to combine the predictions of multiple models, enhancing the accuracy and robustness of our analysis. Ensemble methods like Random Forests or Gradient Boosting can be particularly effective.

6.3.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.

7. Supporting Transportation Improvement Initiatives:

The insights derived from this analysis can support transportation improvement initiatives by providing data-driven information on various aspects of public transport efficiency.

These insights may help in making decisions related to route planning, scheduling, and resource allocation. For example, understanding passenger boardings and on-time performance can lead to optimized transportation services, reduced congestion, and improved overall quality of transportation services.

The information can be valuable for transportation planners and decision-makers to enhance the efficiency of public transport systems.

- **Route Optimization**:

By analyzing data on passenger boardings and ridership patterns, transportation authorities can identify highdemand routes and underutilized ones. This information can help them optimize routes, add more services to popular routes, and reallocate resources to better serve passengers.

- **Scheduling Improvements**:

Data on on-time performance and delays can be used to refine and improve transportation schedules. Timely arrivals and departures are critical for public transport systems, and by identifying the causes of delays, transportation authorities can work to minimize them.

- **Resource Allocation**:

With insights into passenger demographics and travel patterns, authorities can allocate resources more effectively. This might involve deploying more buses or trains during peak hours, increasing the frequency of service on specific routes, or adjusting staffing levels based on demand.

- **Cost Efficiency**:

Data-driven decision-making can also lead to cost savings for transportation agencies. By eliminating underperforming routes or reallocating resources more efficiently, agencies can operate with a reduced budget while maintaining or even improving service quality.

- **Environmental Benefits**:

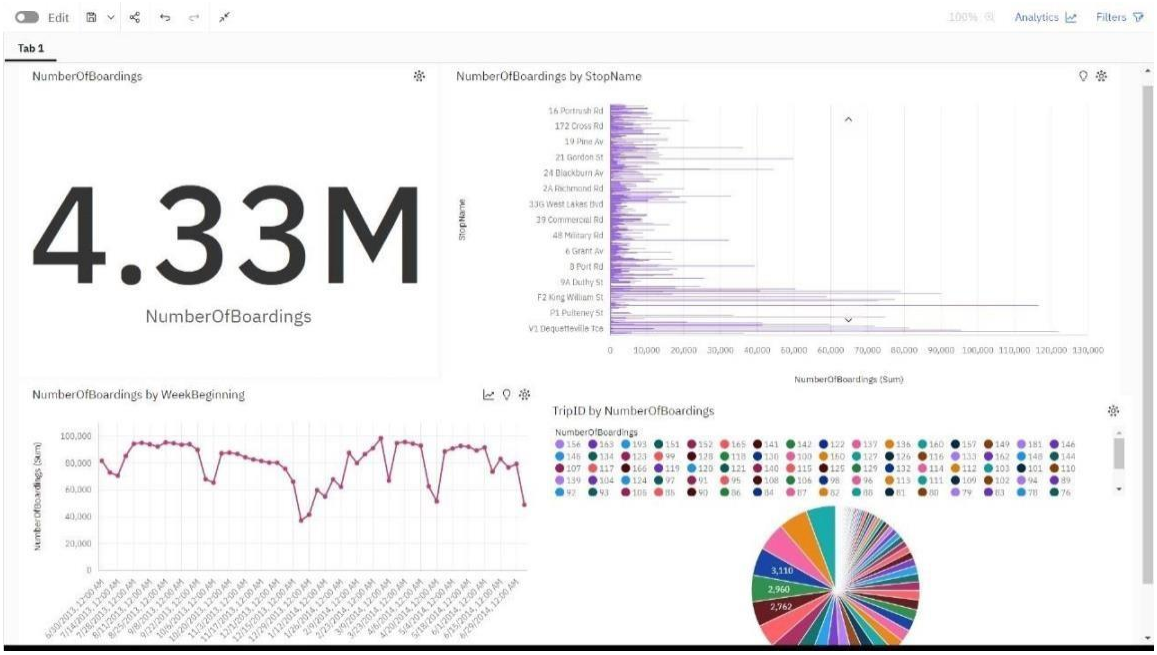
A more efficient public transportation system can have a positive impact on the environment. It can reduce the number of individual vehicles on the road, leading to lower greenhouse gas emissions and improved air quality in urban areas.

- **Safety Enhancements**:

Analyzing data can help identify potential safety issues in the public transport system. For example, if there are areas with a high incidence of accidents or security concerns, authorities can take measures to improve safety for passengers and employees.

In summary, data-driven insights derived from the analysis of public transportation data can play a crucial role in enhancing the efficiency, quality, and sustainability of public transport systems. This, in turn, can lead to better mobility options, reduced congestion, and a more pleasant and environmentally friendly urban environment.

8.IBM cognos final report:



9.Conclusion:

The project concludes by summarizing the data analysis work, emphasizing the use of visualization libraries like Matplotlib and Seaborn, and the application of data-driven techniques for understanding public transport efficiency

In this project, we embarked on a comprehensive journey to understand and optimize public transport efficiency through data analysis. By employing a structured approach and leveraging powerful data analysis tools, we've unveiled insights and established a foundation for data-driven decision-making within the public transport sector.

Throughout the project, we've emphasized the importance of data preprocessing as a critical step. It is essential for refining and enhancing the quality of the data, which, in turn, paves the way for more accurate predictions and insights. These insights have the potential to support a wide range of transportation improvement initiatives, ultimately benefiting commuters and urban development.

This framework will not only be a valuable resource for urban planners and transit agencies but will also contribute to the advancement of data-driven decision-making in public transportation.

Through the fusion of cutting-edge technologies and methodologies, our ultimate goal is to provide a comprehensive and insightful solution for evaluating and enhancing public transport efficiency.