

PHASE:3

Public Transportation Efficiency Analysis

TEAM MEMBERS:

NAME: Murali V

REG NO:721221104040

NAME: Ragul M

REG NO:721221104044

NAME: Aswin S

REG NO:721221104009

NAME: Arun S

REG NO:72122110400

NAME: Abishek S

REG NO:721221104004

Introduction:

Public transportation plays a critical role in urban planning and sustainability. To ensure its optimal functioning, it's essential to analyse and visualize its efficiency. In this project, we will use IBM Cognos for data visualization to gain insights into public transportation efficiency. The project aims to answer questions such as:

Data Collection:

- Collect transportation data from the provided source. Ensure that the data includes information about routes, schedules, delays, ridership, and any other relevant metrics.

DatasetLink: <https://www.kaggle.com/datasets/rednivrug/unisys?select=20140711.CSV>

Data Preprocessing and Cleaning:

- Clean the collected data to ensure its quality and accuracy.

#importing data set

```
import pandas as pd
```

```
import numpy as np
```

```
import sklearn
```

```
from sklearn.preprocessing import StandardScaler
```

```
data = pd.read_csv("public transport.CSV",dtype={'TripID': int, 'RouteID': str, 'StopID': int,  
'WeekBeginning': str})
```

```
data.info()
```

```
data
```

output:

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10857033 entries, 0 to 10857032  
Data columns (total 6 columns):  
#   Column              Dtype  
---  ---  
0   TripID              int32  
1   RouteID             object  
2   StopID              int32  
3   StopName            object  
4   WeekBeginning       object  
5   NumberOfBoardings   int64  
dtypes: int32(2), int64(1), object(3)  
memory usage: 414.2+ MB
```

```
[2]:
```

	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
3	23633	100	12266	Zone A Arndale Interchange	2013-06-30	2
4	23633	100	14147	178 Cross Rd	2013-06-30	1
...
10857028	13346	W91C	14629	21 Cashel St	2014-07-06	1
10857029	13346	W91C	14708	22 Cashel St	2014-07-06	3
10857030	13346	W91C	13709	2 Greenhill Rd	2014-07-06	1
10857031	13346	W91C	14029	10 East Av	2014-07-06	1
10857032	13346	W91C	13824	6 Leader St	2014-07-06	1

#cleansing the data set

```
data['StopName'].fillna('Unknown', inplace=True)
```

```
data.drop_duplicates(subset=['TripID', 'StopID', 'WeekBeginning'], keep='first', inplace=True)
```

```
data['TripID'] = data['TripID'].astype(int)
```

```
data['RouteID'] = data['RouteID'].astype(str)
```

```
data['StopID'] = data['StopID'].astype(int)
```

```
data['WeekBeginning'] = pd.to_datetime(data['WeekBeginning'])
```

#stored the cleaned data into another file

```
data.to_csv('cleanddataset.csv', index=False)
```

Exploratory Data Analysis (EDA):

```
data = pd.read_csv('cleanddataset.csv')
```

```
data.shape
```

```
data.head(10)
```

```
data.sample(5)
```

output:

```
[38]: (10857033, 6)
```

```
[38]:
```

	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
3	23633	100	12266	Zone A Arndale Interchange	2013-06-30	2
4	23633	100	14147	178 Cross Rd	2013-06-30	1
5	23634	100	13907	9A Marion Rd	2013-06-30	1
6	23634	100	14132	175 Cross Rd	2013-06-30	1
7	23634	100	13335	9A Holbrooks Rd	2013-06-30	1
8	23634	100	13875	9 Marion Rd	2013-06-30	1
9	23634	100	13045	206 Holbrooks Rd	2013-06-30	1

```
data.shape
```

output:

```
[49]: (10857033, 6)
```

```
data.columns
```

output:

```
[50]: Index(['TripID', 'RouteID', 'StopID', 'StopName', 'WeekBeginning',  
          'NumberOfBoardings'],  
          dtype='object')
```

```
pd.isnull(data).sum()
```

output:

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

```
data.describe()
```

output:

```
[52]:
```

	TripID	StopID	NumberOfBoardings
count	1.085703e+07	1.085703e+07	1.085703e+07
mean	2.952123e+04	1.366126e+04	4.743657e+00
std	1.960946e+04	1.971710e+03	9.382204e+00
min	7.900000e+01	1.000100e+04	1.000000e+00
25%	1.191600e+04	1.231100e+04	1.000000e+00
50%	2.747900e+04	1.334600e+04	2.000000e+00
75%	4.885800e+04	1.491600e+04	4.000000e+00
max	6.553500e+04	1.871500e+04	9.770000e+02

```
data.nunique()
```

output:

```
[53]:
```

TripID	39282
RouteID	619
StopID	7397
StopName	4165
WeekBeginning	54
NumberOfBoardings	400
dtype:	int64

Visualization and Analysis:

```
##can assign the each chart to one axes at a time
```

```
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])
```

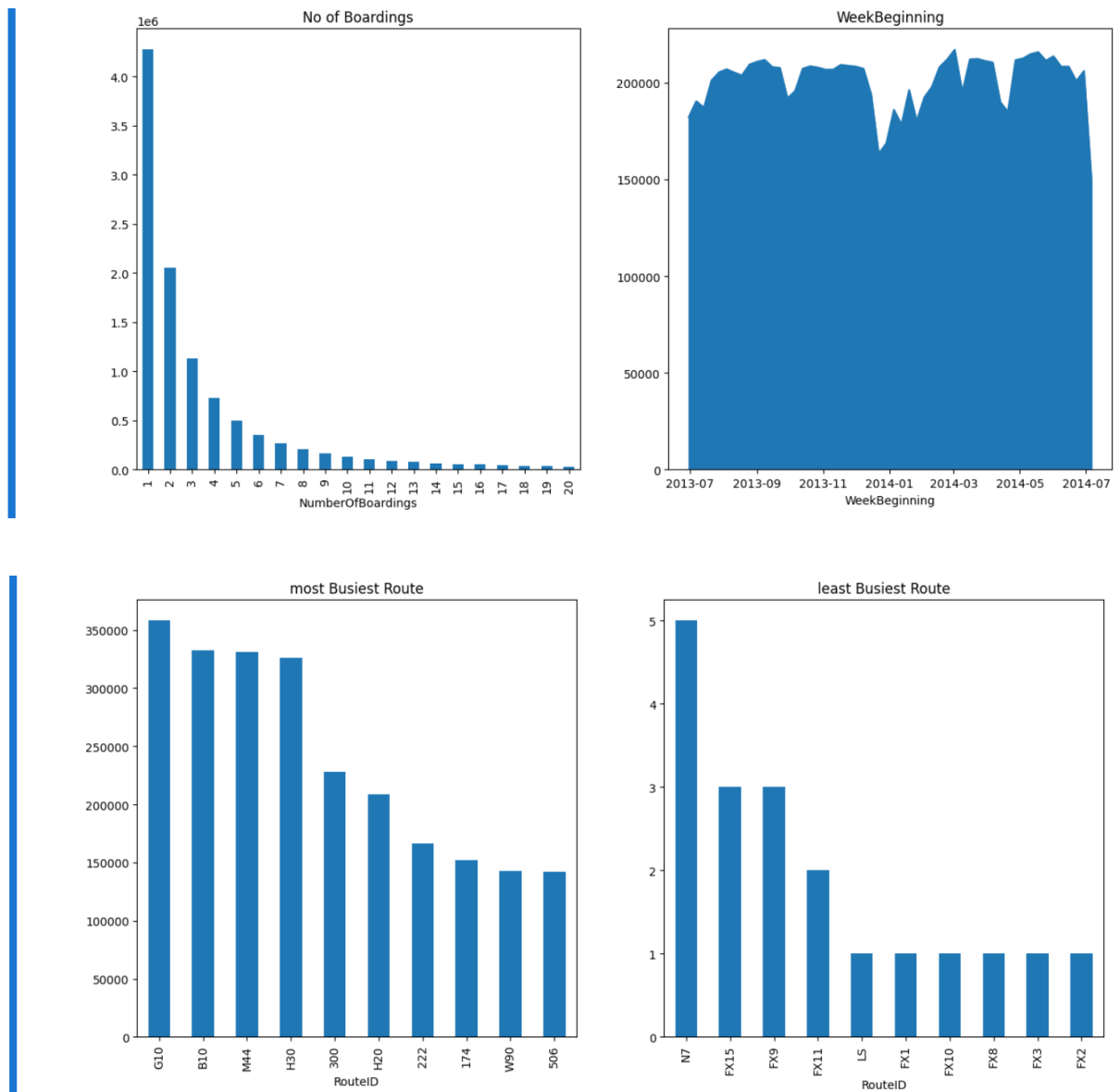
```
ax=axrr[1][0]
```

```
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])
```

output:



Project Conclusion:

The public transportation efficiency analysis project aims to provide valuable insights into the performance of public transportation systems. By using IBM Cognos for visualization, we can create interactive and informative dashboards that assist in making informed decisions to enhance public transportation efficiency, ultimately contributing to more sustainable and accessible urban environments.