# PHASE 3: DEVELOPMENT

# TOPIC: PUBLIC TRANSPORT EFFIENCY ANALYSIS

## INTRODUCTION:

At this pivotal stage, we delve into the heart of our endeavor by meticulously loading and preprocessing the dataset, laying the foundation for sophisticated analyses and meaningful insights. This phase serves as a crucial bridge, connecting theoretical foundations with practical implementations, propelling us closer to the realization of our project's objectives..

## DATA PREPROCESSING :

Data preprocessing in public transportation analysis is a crucial step that involves cleaning, transforming, and organizing raw transportation data to make it suitable for analysis. Public transportation systems generate vast amounts of data from various sources, such as ticketing systems, GPS trackers, sensors, and schedules. Preprocessing this data is necessary to extract meaningful insights, improve data quality, and ensure that it's ready for analytical and modeling tasks. Here are the key aspects of data preprocessing in public transportation analysis:

1. Data Collection:

- Data collection involves gathering information from various sources, such as fare collection systems, vehicle sensors, passenger counts, and scheduling systems. This raw data can be in different formats and structures.

2. Data Cleaning:

- Data cleaning is the process of identifying and correcting errors, inconsistencies, and missing values in the dataset. This can include dealing with duplicated records, removing outliers, and addressing data entry errors.

3. Data Integration:

- Public transportation data often comes from different sources and in various formats. Data integration involves merging, aligning, and transforming data so that it can be analyzed as a cohesive dataset.

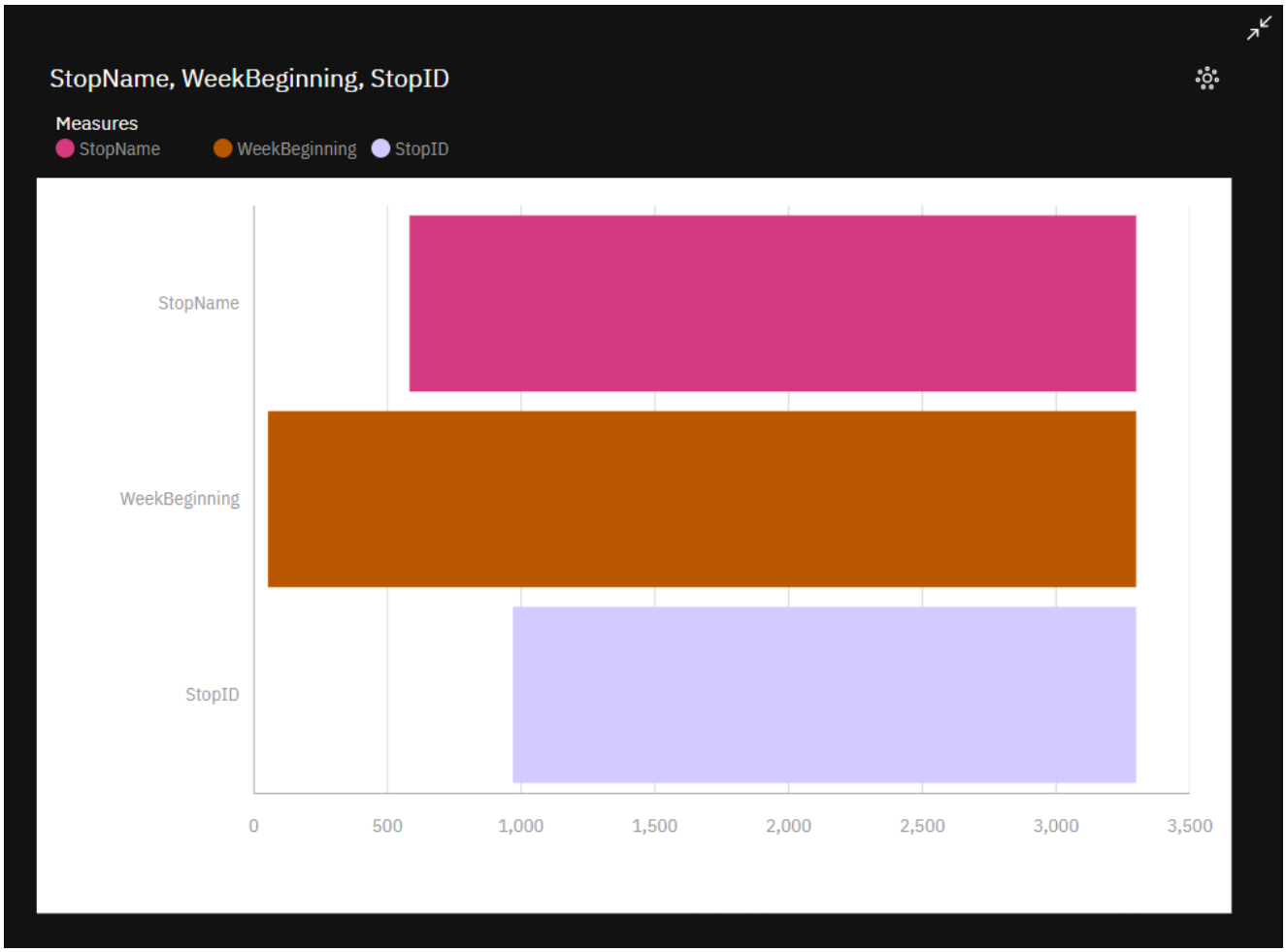
4. Data Transformation:

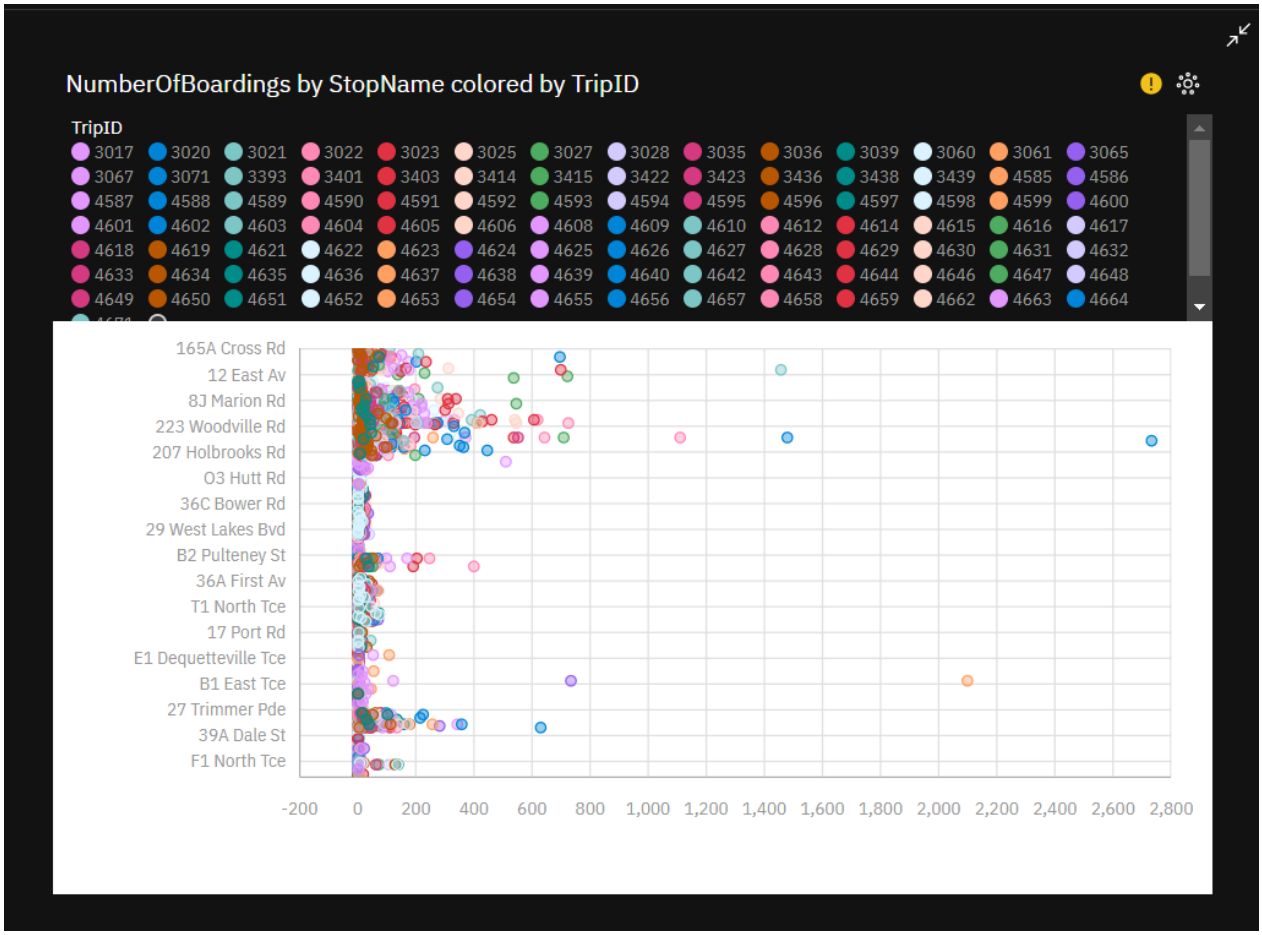
- Transformation tasks may include converting data into a standardized format, resampling temporal data, and aggregating data to different time intervals (e.g., hourly or daily) to align with analysis requirements.

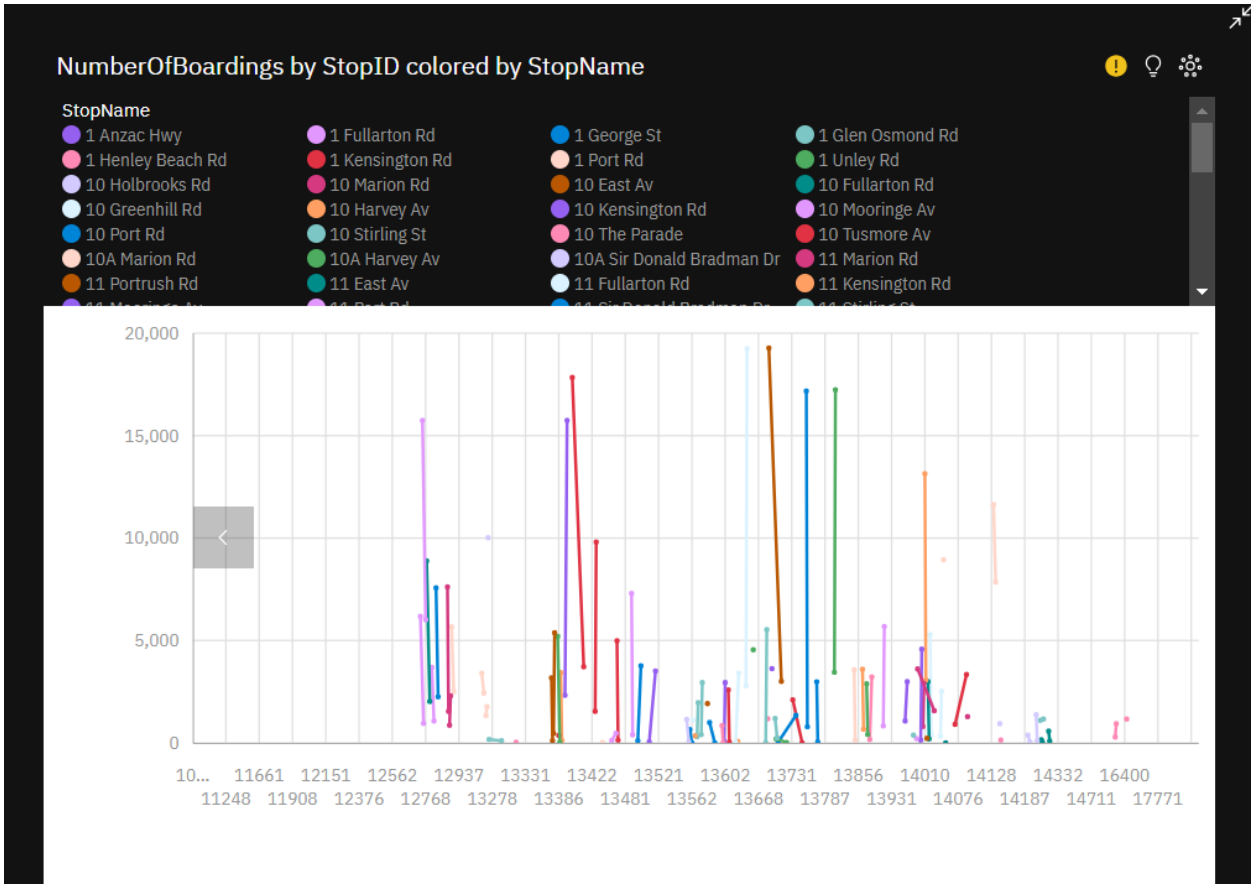
5. Geospatial Data Processing:

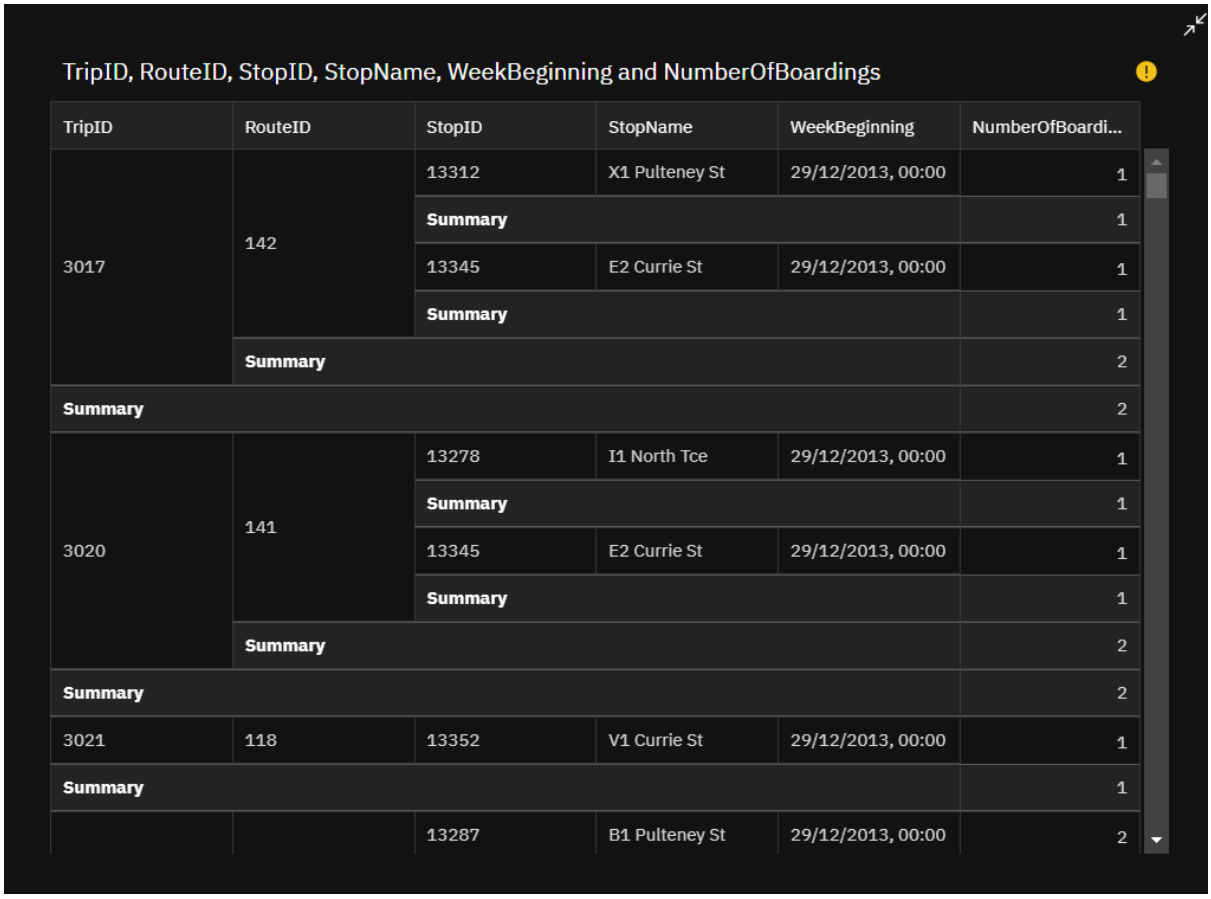
- Public transportation analysis often involves geospatial data, including GPS coordinates, routes, and geographical boundaries. Preprocessing may involve geocoding, spatial indexing, and the calculation of distances or travel times between locations.

**GRAPHS DEVELOPED USING IBM:**




NumberOfBoardings by StopName colored by TripID
TripID
3017 3020 3021 3022 3023 3025 3027 3028 3035 3036 3039 3060 3061 3065
3067 3071 3393 3401 3403 3414 3415 3422 3423 3436 3438 3439 4585 4586
4587 4588 4589 4590 4591 4592 4593 4594 4595 4596 4597 4598 4599 4600
4601 4602 4603 4604 4605 4606 4608 4609 4610 4612 4614 4615 4616 4617
4618 4619 4621 4622 4623 4624 4625 4626 4627 4628 4629 4630 4631 4632
4633 4634 4635 4636 4637 4638 4639 4640 4642 4643 4644 4646 4647 4648
4649 4650 4651 4652 4653 4654 4655 4656 4657 4658 4659 4662 4663 4664
165A Cross Rd
12 East Av
8J Marion Rd
223 Woodville Rd
207 Holbrooks Rd
O3 Hutt Rd
36C Bower Rd
29 West Lakes Bvd
B2 Pulteney St
36A First Av
T1 North Tce
17 Port Rd
E1 Dequetteville Tce
B1 East Tce
27 Trimmer Pde
39A Dale St
F1 North Tce
-200 0 200 400 600 800 1,000 1,200 1,400 1,600 1,800 2,000 2,200 2,400 2,600 2,800


NumberOfBoardings by StopID colored by StopName
StopName
1 Anzac Hwy
1 Fullarton Rd
1 George St
1 Glen Osmond Rd
1 Henley Beach Rd
1 Kensington Rd
1 Port Rd
1 Unley Rd
10 Holbrooks Rd
10 Marion Rd
10 East Av
10 Fullarton Rd
10 Greenhill Rd
10 Harvey Av
10 Kensington Rd
10 Mooringe Av
10 Port Rd
10 Stirling St
10 The Parade
10 Tusmore Av
10A Marion Rd
10A Harvey Av
10A Sir Donald Bradman Dr
11 Marion Rd
11 Portrush Rd
11 East Av
11 Fullarton Rd
11 Kensington Rd
20,000
15,000
10,000
5,000
0
10...
11661
12151
12562
12937
13331
13422
13521
13602
13731
13856
14010
14128
14332
16400
11248
11908
12376
12768
13278
13386
13481
13562
13668
13787
13931
14076
14187
14711
17771



TripID, RouteID, StopID, StopName, WeekBeginning and NumberOfBoardings

| TripID | RouteID | StopID | StopName | WeekBeginning | NumberOfBoardi... |
|---|---|---|---|---|---|
| 3017 | 142 | 13312 | X1 Pulteney St | 29/12/2013, 00:00 | 1 |
| | | **Summary** | | | 1 |
| | | 13345 | E2 Currie St | 29/12/2013, 00:00 | 1 |
| | | **Summary** | | | 1 |
| | **Summary** | | | | 2 |
| **Summary** | | | | | 2 |
| 3020 | 141 | 13278 | I1 North Tce | 29/12/2013, 00:00 | 1 |
| | | **Summary** | | | 1 |
| | | 13345 | E2 Currie St | 29/12/2013, 00:00 | 1 |
| | | **Summary** | | | 1 |
| | **Summary** | | | | 2 |
| **Summary** | | | | | 2 |
| 3021 | 118 | 13352 | V1 Currie St | 29/12/2013, 00:00 | 1 |
| **Summary** | | | | | 1 |
| | | 13287 | B1 Pulteney St | 29/12/2013, 00:00 | 2 |

## CODE USED:

```
import pandas as pd
# Load the dataset (replace 'nm_dataset.csv' with the actual path to your dataset)
data = pd.read_csv('nm_dataset.csv', low_memory=False, dtype={'route id': str})
# Check for and remove duplicate rows
data.drop_duplicates(inplace=True)
# Check for and remove rows with missing data
data.dropna(inplace=True)
# Reset the index after removing rows
data.reset_index(drop=True, inplace=True)
# Save the cleaned data to a new file (if needed)
data.to_csv('done_data.csv', index=False)
```

## CONCLUSION:

In the culmination of Phase3, we stand on the precipice of innovation, armed with a meticulously curated dataset and a wealth of preprocessing insights. This phase has been a testament to our team's dedication, expertise, and collaborative spirit. As we conclude this stage, we do so with a profound sense of accomplishment and a keen anticipation for the transformative journey that lies ahead.

## TEAM MEMBERS:


1. SRIVISHWA P -2021115109
2. SURYA NARAYANAAN K C -2021115113
3. SWASTHIKA LAKSHMI PRIYA R -2021115115
4. VISHNU PRIYA K -2021115121
5. SIVA L -2021115330