Public Transportation Analysis

2023 Naan Mudhalvan - IBM Data Analytics with Cognos
Group 1 - Project 8
College : NM001 - College of Engineering Guindy
Proj_200340_Team_2

Members: Abinithi R, Abirami S V, Adithya R U, Akshaya G R, Sai Rishi A N
Faculty Mentor : Dr. G Geetha

PHASE 3

DEVELOPMENT PART 1

PROBLEM DEFINITION :

Analyse public transportation data to assess **service efficiency**, **on time performance**, and
passenger feedback.
Provide insights that **support transportation improvement initiatives** and enhance the overall public transportation experience.

ANALYSIS STEPS

DATA PREPROCESSING

❖Cleaning and Preprocessing the Dataset:

Handling Missing Values

Missing data can significantly affect the performance of machine learning models. There are several methods to handle missing values, including:

- ***Removing Rows:*** Rows with missing values can be removed, but this might result in losing valuable data.
- ***Filling with Mean/Median/Mode*:** Filling missing values with the mean (average), median (middle value), or mode (most frequent value) of the respective column.
- ***Advanced Imputation Techniques*:** Using advanced techniques such as K-nearest neighbors imputation or regression imputation to predict missing values based on other features.

IMPORTING NECESSARY LIBRARIES

In [13]:

```
import pandas as pd
import numpy as np
```

LOADING DATASET

In [3]:

```
data = pd.read_csv("C:\\Users\\AbiramiSV\\Downloads\\Dataset\\PublicTransportDataset.CSV",
```

DISPLAYING FIRST 20 ROWS

In [4]:

```
data.head(20)
```

Out[4]:

	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
5	23634	100	13907	9A Marion Rd	2013-06-30 00:00:00	1
6	23634	100	14132	175 Cross Rd	2013-06-30 00:00:00	1
7	23634	100	13335	9A Holbrooks Rd	2013-06-30 00:00:00	1
8	23634	100	13875	9 Marion Rd	2013-06-30 00:00:00	1
9	23634	100	13045	206 Holbrooks Rd	2013-06-30 00:00:00	1
10	23635	100	13335	9A Holbrooks Rd	2013-06-30 00:00:00	1
11	23635	100	13383	8A Marion Rd	2013-06-30 00:00:00	1
12	23635	100	13586	8D Marion Rd	2013-06-30 00:00:00	2
13	23635	100	12726	23 Findon Rd	2013-06-30 00:00:00	1
14	23635	100	13813	8K Marion Rd	2013-06-30 00:00:00	1
15	23635	100	14062	20 Cross Rd	2013-06-30 00:00:00	1
16	23636	100	12780	22A Crittenden Rd	2013-06-30 00:00:00	1
17	23636	100	13383	8A Marion Rd	2013-06-30 00:00:00	1
18	23636	100	14154	180 Cross Rd	2013-06-30 00:00:00	2
19	23636	100	13524	8C Marion Rd	2013-06-30 00:00:00	3

DROPPING RECORDS HAVING DUPLICATE VALUES

In [5]:

```
data.drop_duplicates(inplace=True)
```

FILLING MISSING VALUES WITH MEAN

In [6]:

```
data.fillna(data.mean(), inplace=True)
```

PRINTING FIRST FEW ROWS

In [7]:

```
print(data.head())
```

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

   NumberOfBoardings
0                  1
1                  1
2                  1
3                  2
4                  1
```

GENERATING DESCRIPTIVE STATISTICS OF DATASET

In [8]:

```
print(data.describe())
```

```
             TripID        StopID  NumberOfBoardings
count  1.085723e+07  1.085723e+07       1.085723e+07
mean   2.952100e+04  1.366132e+04       4.743737e+00
std    1.960938e+04  1.971760e+03       9.382286e+00
min    7.900000e+01  1.000100e+04       1.000000e+00
25%    1.191700e+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
```

GENERATING CONCISE SUMMARY OF DATASET

In [9]:

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10857234 entries, 0 to 10857233
Data columns (total 6 columns):
 #   Column             Dtype 
---  ------             ----- 
 0   TripID             int64 
 1   RouteID            object
 2   StopID             int64 
 3   StopName           object
 4   WeekBeginning      object
 5   NumberOfBoardings  int64 
dtypes: int64(3), object(3)
memory usage: 579.8+ MB
None
```

SHAPE OF DATASET

In [11]:

```
print(data.shape)
```

```
(10857234, 6)
```

DISPLAYING FIRST FEW ROWS AFTER PREPROCESSING

In [12]:

```
data.head()
```

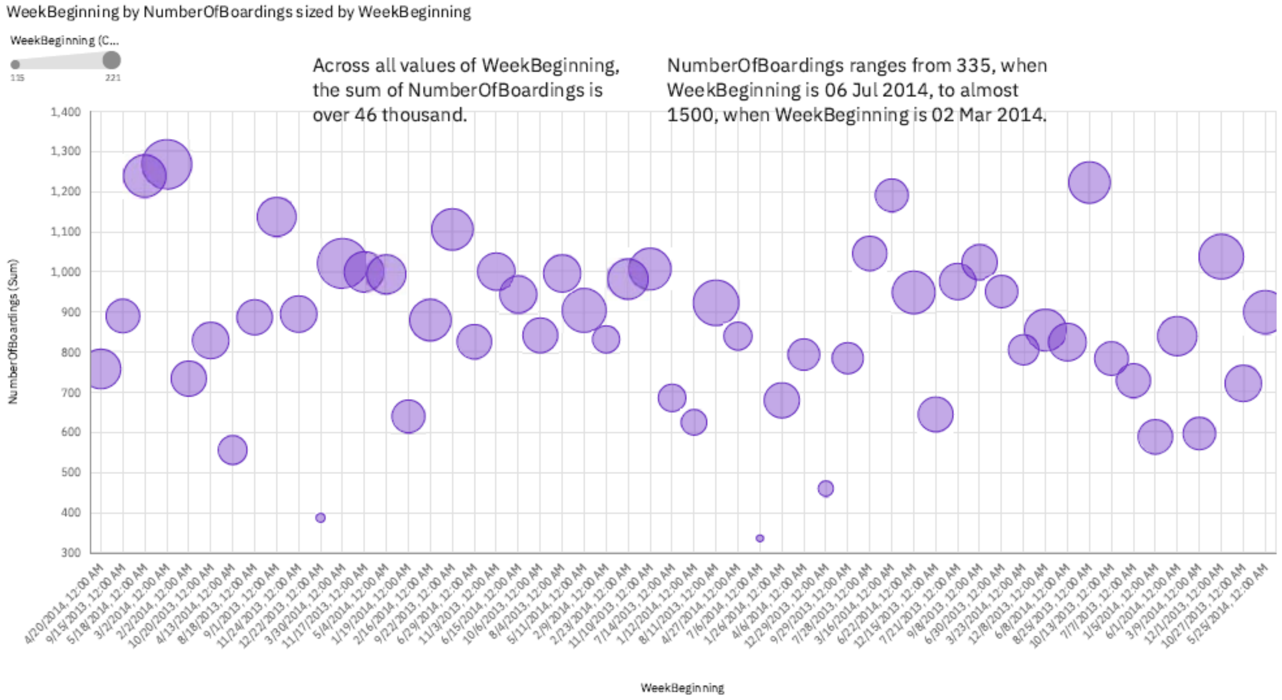
Out[12]:

	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 [ ]:

VISUALIZATIONS IN COGNOS

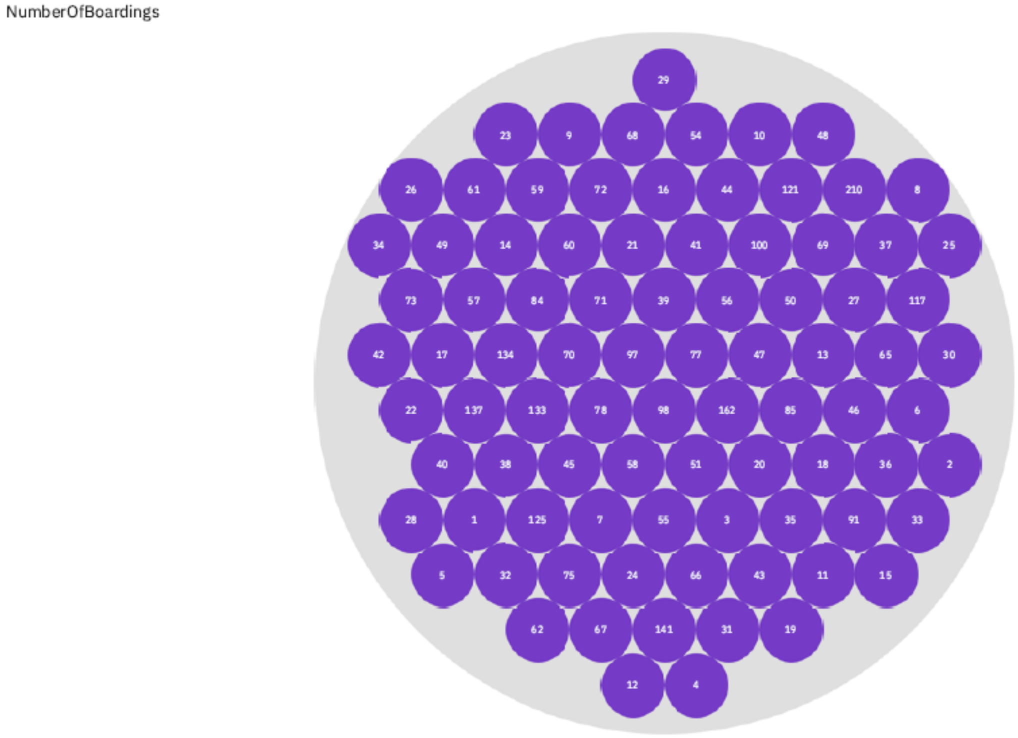
1. Bubble plot of WeekBeginning by NumberOfBoardings sized by WeekBeginning



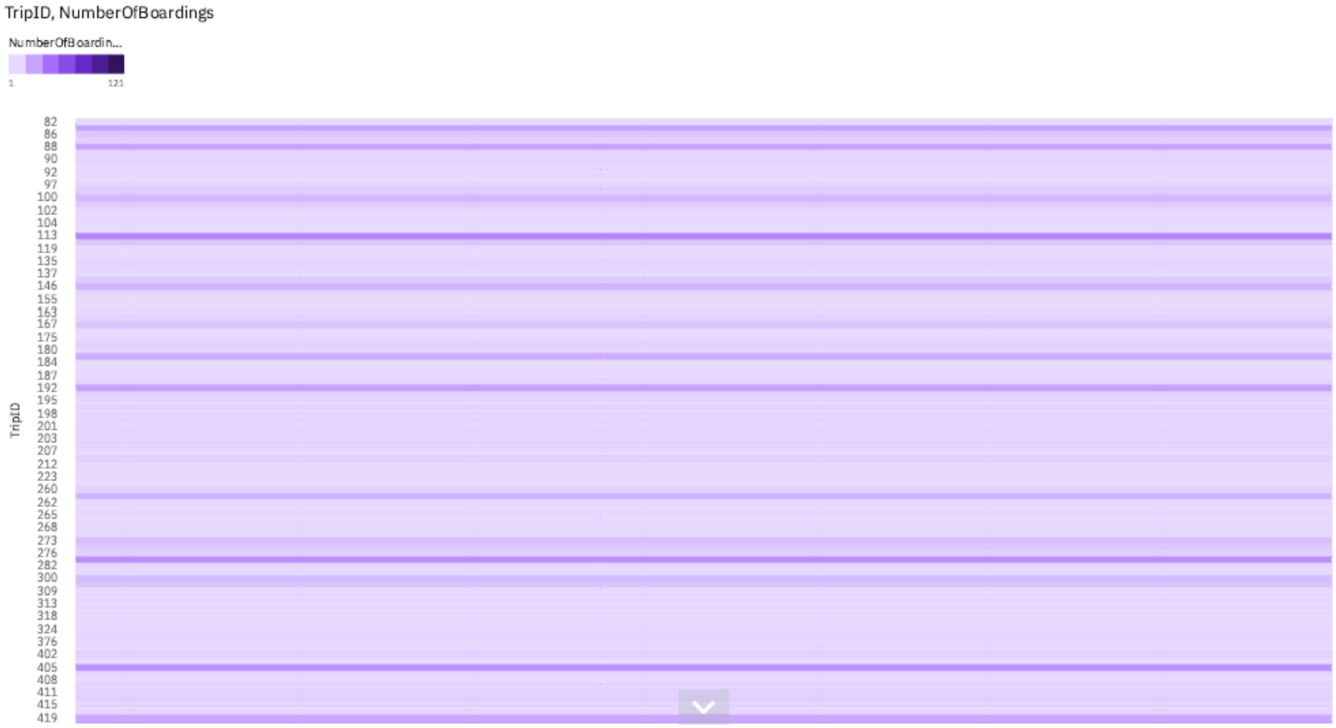
Across all values of WeekBeginning, the sum of NumberOfBoardings is over 46 thousand.

NumberOfBoardings ranges from 335, when WeekBeginning is 06 Jul 2014, to almost 1500, when WeekBeginning is 02 Mar 2014.

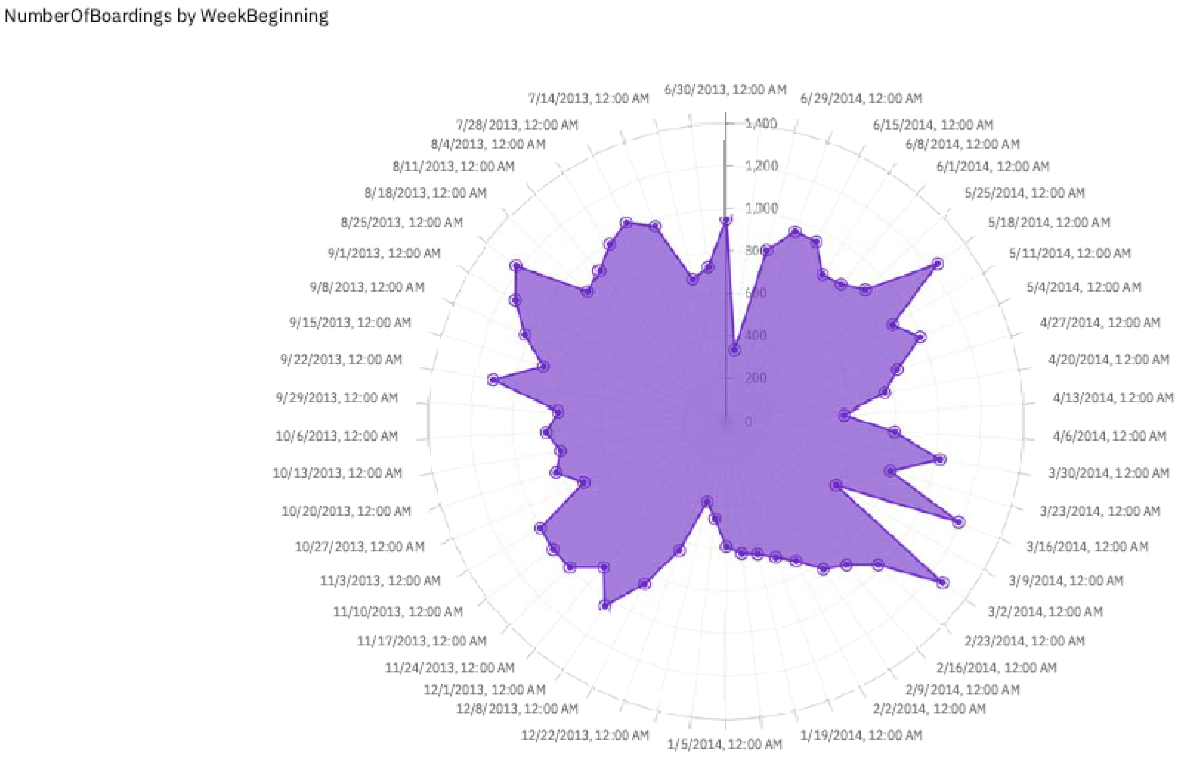
2. Hierarchy Bubble of NumberOfBoardings



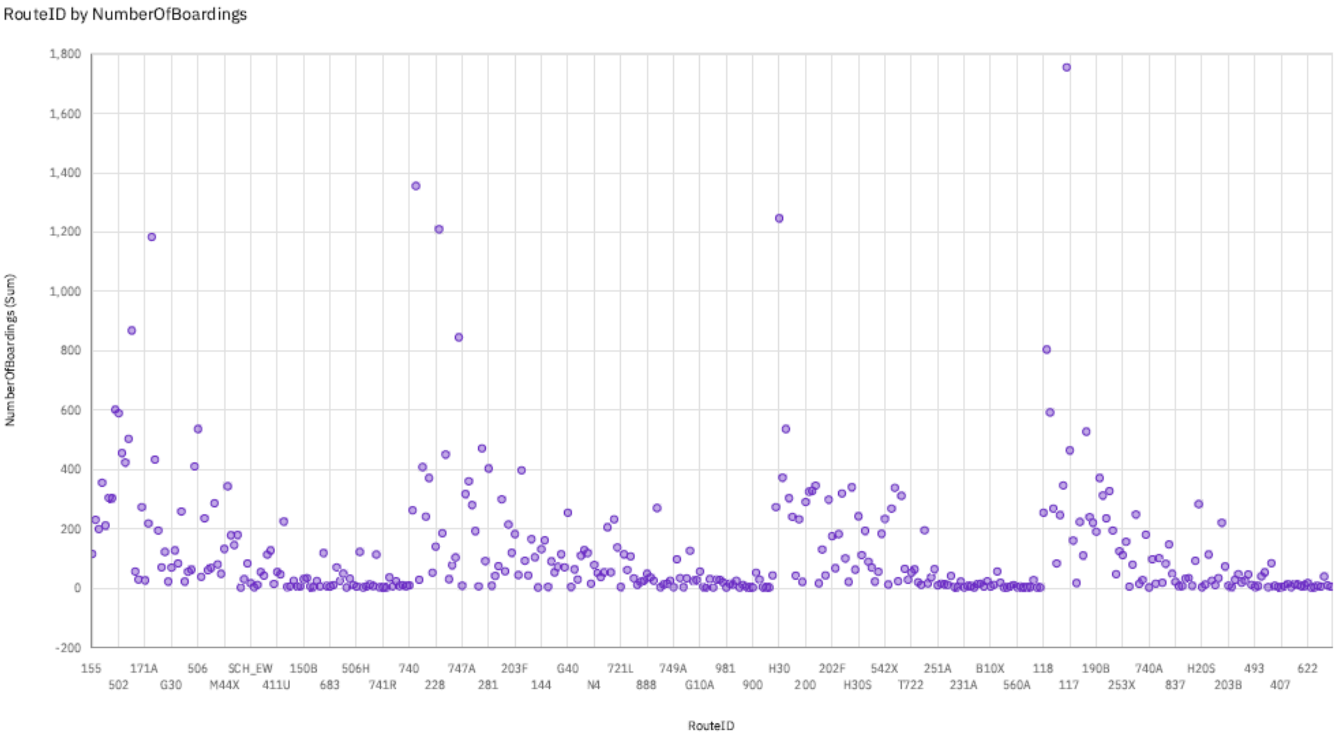
3. Heat Map of NumberOfBoardings by TripID



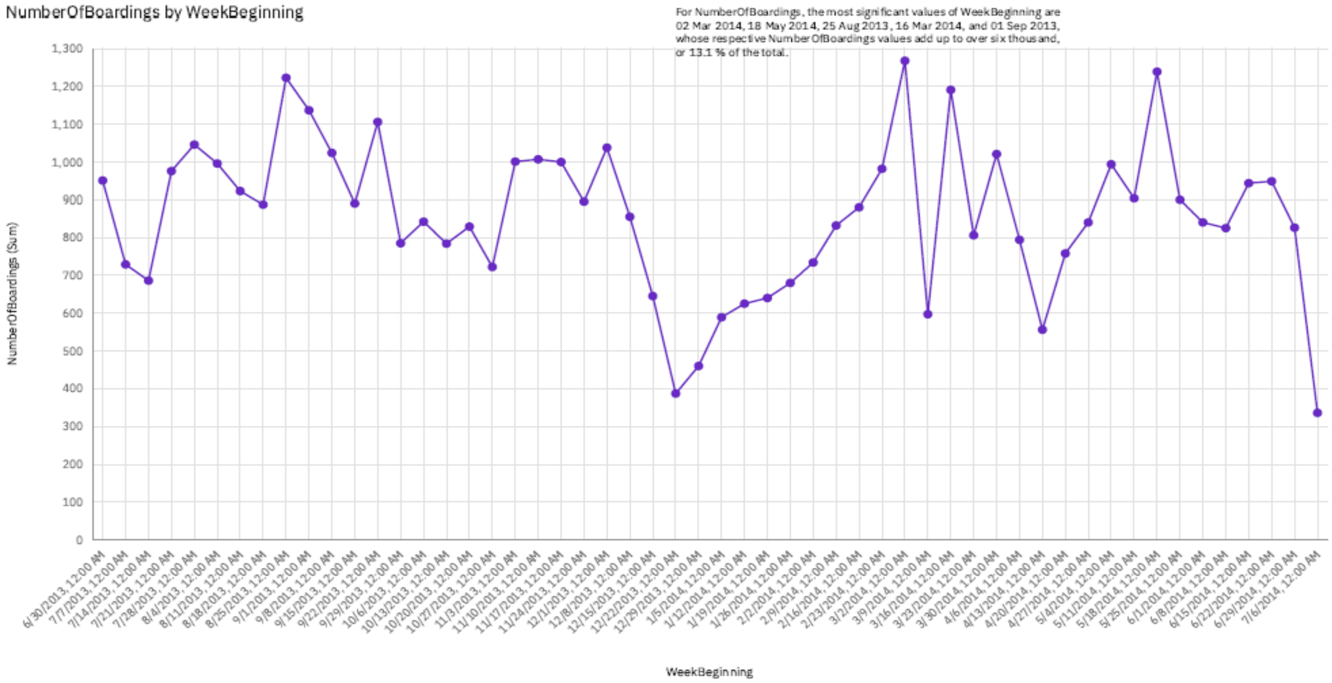
4. Radar of NumberOfBoardings by WeekBeginning



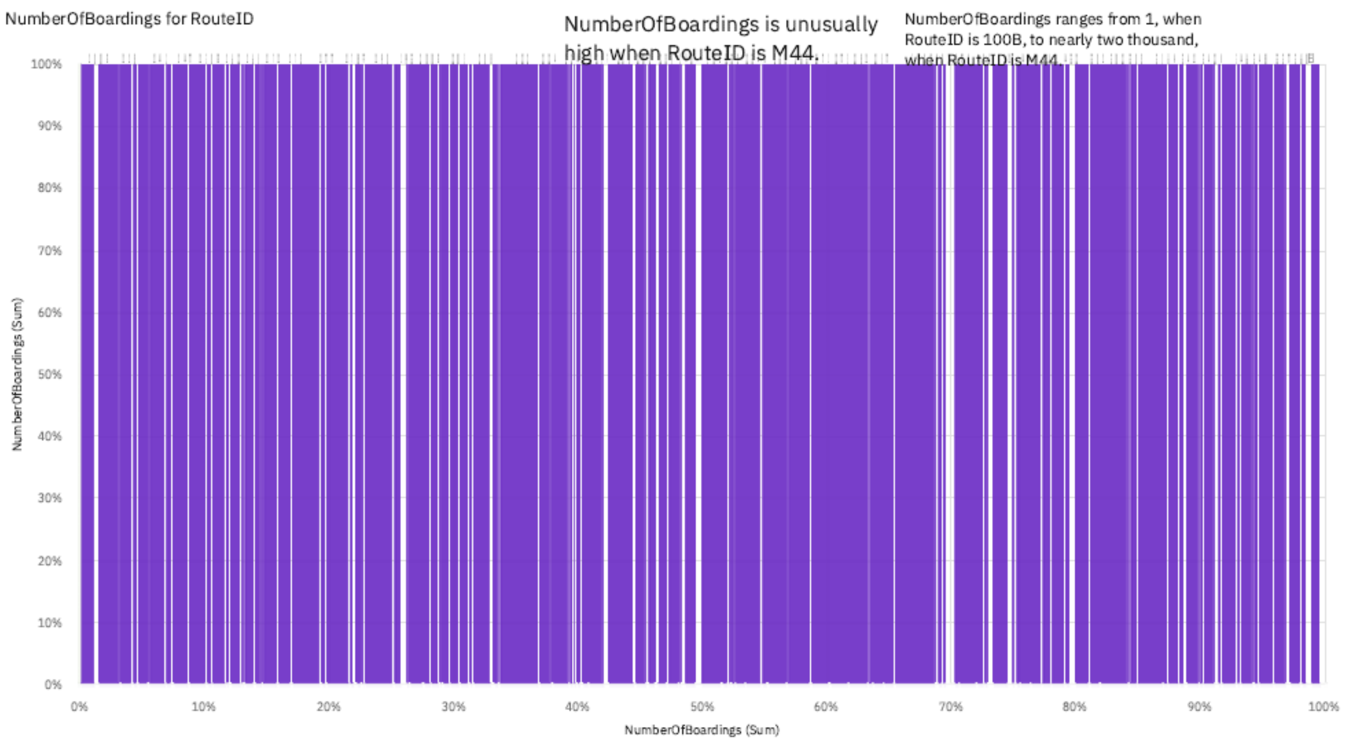
5. Scatter Plot of RouteID by NumberofBoardings



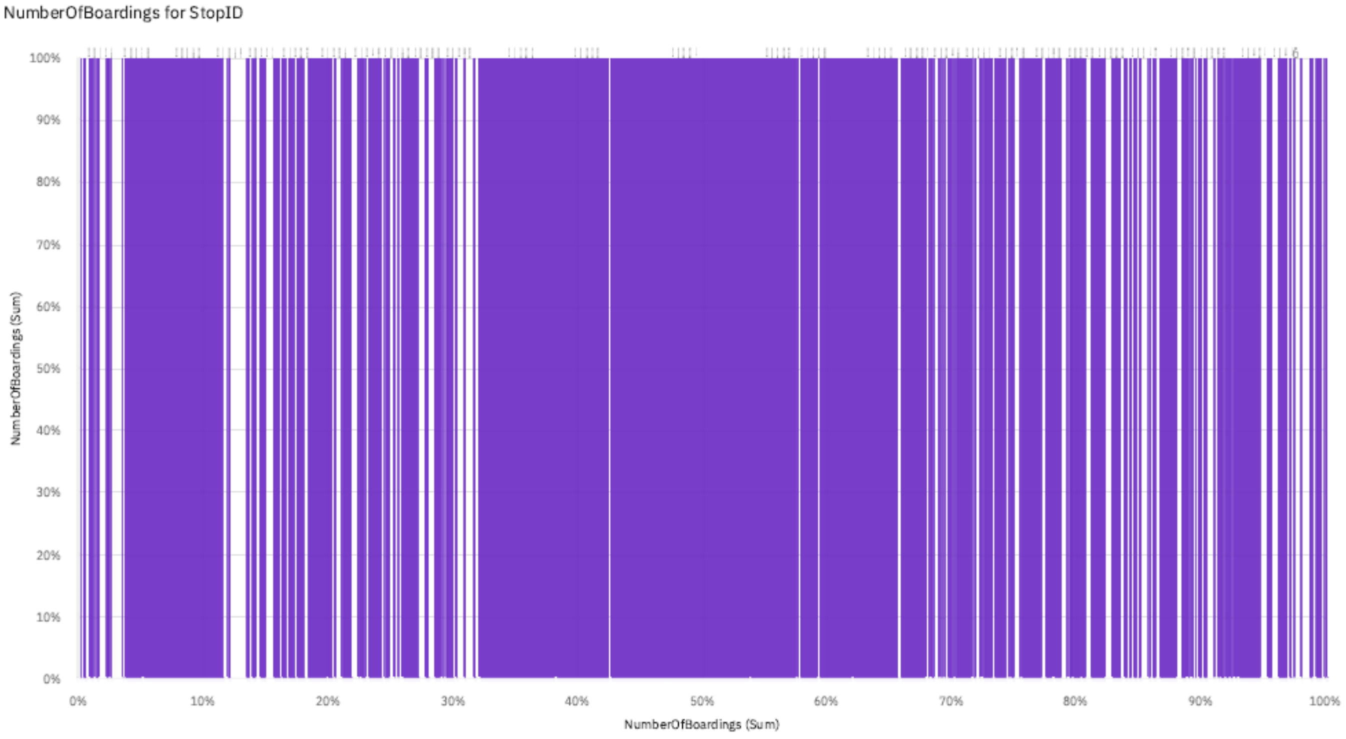
6. Line Graph of NumberOfBoardings by WeekBeginning


For NumberOfBoardings, the most significant values of WeekBeginning are 02 Mar 2014, 18 May 2014, 25 Aug 2013, 16 Mar 2014, and 01 Sep 2013, whose respective NumberOfBoardings values add up to over six thousand, or 13.1 % of the total.

7. Waterfall Plot for NumberofBoardings for RouteID



8. Waterfall Plot for NumberOfBoardings for StopID


NumberOfBoardings for StopID
NumberOfBoardings (Sum)
NumberOfBoardings (Sum)