GOVERNMENTCOLLEGEOFENGINEERINGERODE



B.EElectronicsandCommunicationEngineering

Public TransportationAnalysis

|  |  |
| --- | --- |
| **Nameofthe Students:** | **UniversityRegister no:** |
| **TeamLeader:**  Gohul P | 731121106014 |
| **TeamMembers:** |  |
| Monish C M | 731121106033 |
| DhiyaRohith S | 731121106009 |
| Luthdran C S | 731121106029 |

Underthementor of

**Dr.M.Poongothai**

**Department of Information Technology (IT) DepartmentofElectronicsandCommunicationEngineering**

GovernmentCollegeofEngineering

Erode,PO,nearVasaviCollege,TamilNadu-638316, Affiliated to Anna University, Chennai.

# ANALYSE OBJECTIVE:

* This analysis seeks to address the complex and pressing issues that plague public transportation networks, which impact the quality of service, ridership numbers, environmental sustainability, and the economic viability of these systems.
* This analysishave explored how people are travelling from different stops in Adelaide Metropolitan area and the rate at which passengers on each bus route are increasing. Finally created a predictive model to find the load of passengers on public Bus transport system in future.
* To visualize on-time performance, passenger feedback, and service efficiency metrics.
* The primary objectives of data preprocessing are to prepare raw data for analysis or machine learning tasks, ensuring that the data is in a suitable format for these tasks.

# LOADINGTHE DATASET:

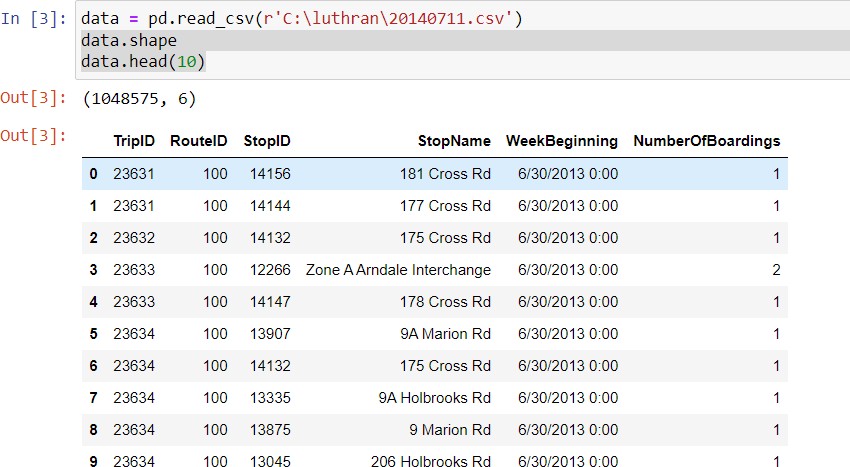
* Loading a dataset is a fundamental step in data analytics, where you import your data into the chosen data analysis tool or software for further exploration and analysis. The specific steps for loading a dataset can vary depending on the software or programming language you are using. Here's a general overview of how to load a dataset:
* **ChooseYourDataAnalysisTool**



* **PrepareYourDataFile**
* **ImportYourData**
* a.**Python(usingPandas)**
* b.**R**
* Loadingandpreprocessingdatasetsarecritical stepsinthedataanalyticsprocess,and they carry significant importance for several reasons:
* **DataQuality Assurance**



* **DataUnderstanding**
* **DataConsistency**
* **MissingDataHandling**
* **FeatureEngineering**



# DATAPREPROCESSING:

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

Data Preprocessing can be done in four different ways. Data cleaning/cleaning, data integration, data transformation, and data reduction are the four categories.

# Data Cleaning

Data cleaning is a crucial step in the data analysis process, as it involves identifying and rectifying errors, inconsistencies, and inaccuracies in the dataset. Cleaning data is essential for accurate and reliable analysis. Here are some common data cleaning methods and techniques:

-Handlingmissingdata

-HandlingDuplicate sets.

Several automated data cleaning tools and libraries, such as OpenRefine,Trifactaand Python libraries like pandas, offer functions and methods for data cleaning.

In this project Python library named Pandas is used to clean data .The Python Pandas library is a powerful data manipulation and analysis library that is widely used for data cleaning and preparation. It offers functions and methods for tasks like handling missing values, data transformation, and filtering.

**Datatransformation:**

* Thisinvolvesconvertingthedataintoaformatthatissuitableforthe analysis task.
* Forexample,thismayinvolveconvertingcategoricaldatatonumerical data, or scaling the data to a suitable range.

# PROGRAM:

importpandasaspd

## #LoadthedatasetintoapandasDataFrame

df=pd.read\_csv('/content/Dataset.csv')

## #Removerowswithanymissingvaluesinanycolumn

df\_cleaned=df.dropna()

## #Or,fillmissingvaluesinspecificcolumns(ifapplicable)

**#df\_cleaned['NumberOfBoardings'].fillna(0,inplace=True) # Remove duplicate rows based on all columns**

df\_cleaned=df\_cleaned.drop\_duplicates()

## #Convertspecificcolumnstoappropriatedatatypes

df\_cleaned['TripID'] = df\_cleaned['TripID'].astype(int) df\_cleaned['NumberOfBoardings'] =

df\_cleaned['NumberOfBoardings'].astype(int)

## #Convertothercolumnsasneeded

**# Extract day, month, year, etc. from 'WeekBeginning' column** df\_cleaned['WeekBeginning']=pd.to\_datetime(df\_cleaned['WeekBeginning']) df\_cleaned['Year'] = df\_cleaned['WeekBeginning'].dt.yeardf\_cleaned['Month'] = df\_cleaned['WeekBeginning'].dt.monthdf\_cleaned['Day'] = df\_cleaned['WeekBeginning'].dt.day

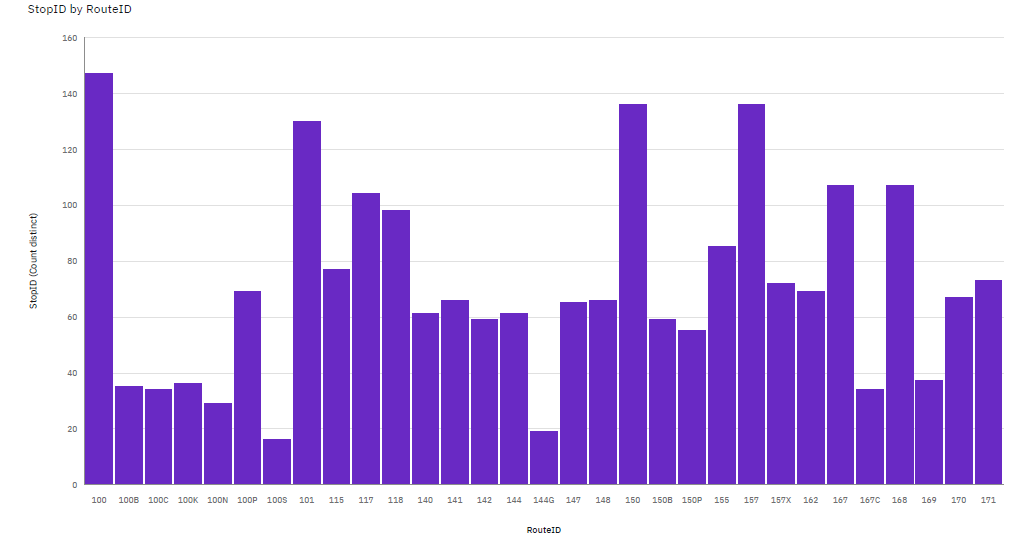
## #Extractotherfeaturesasneeded

**#SavethecleaneddatasettoanewCSVfile**

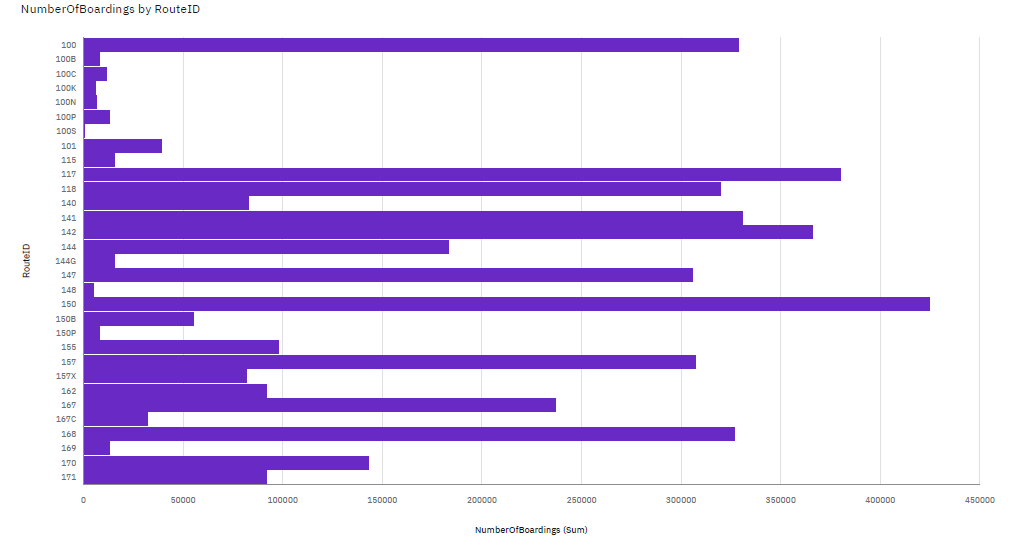
df\_cleaned.to\_csv('project\_dataset.csv',index=False)

# DATAVISUALIZATIONUSINGCOGNOSIBM:

* Data visualization is the graphical representation of data to help people understand the information contained in the data more easily. Visualizations make patterns, trends, and insights in the data more apparent. Here the analyzed data can be visualized by using cognosibm and matplot python library. The data can be visualized in the form of graphs,tables, charts,etc,
* IBM Cognos Analytics is a specific product within the IBM Cognos suite, focusing on analytics and business intelligence. Cognos Analytics is designed to help organizations transform their data into meaningful insights and actionable information.
* Cognos Analytics provides a wide range of data visualization options, including charts, graphs, and maps, to represent data in a visually appealing and informative way.

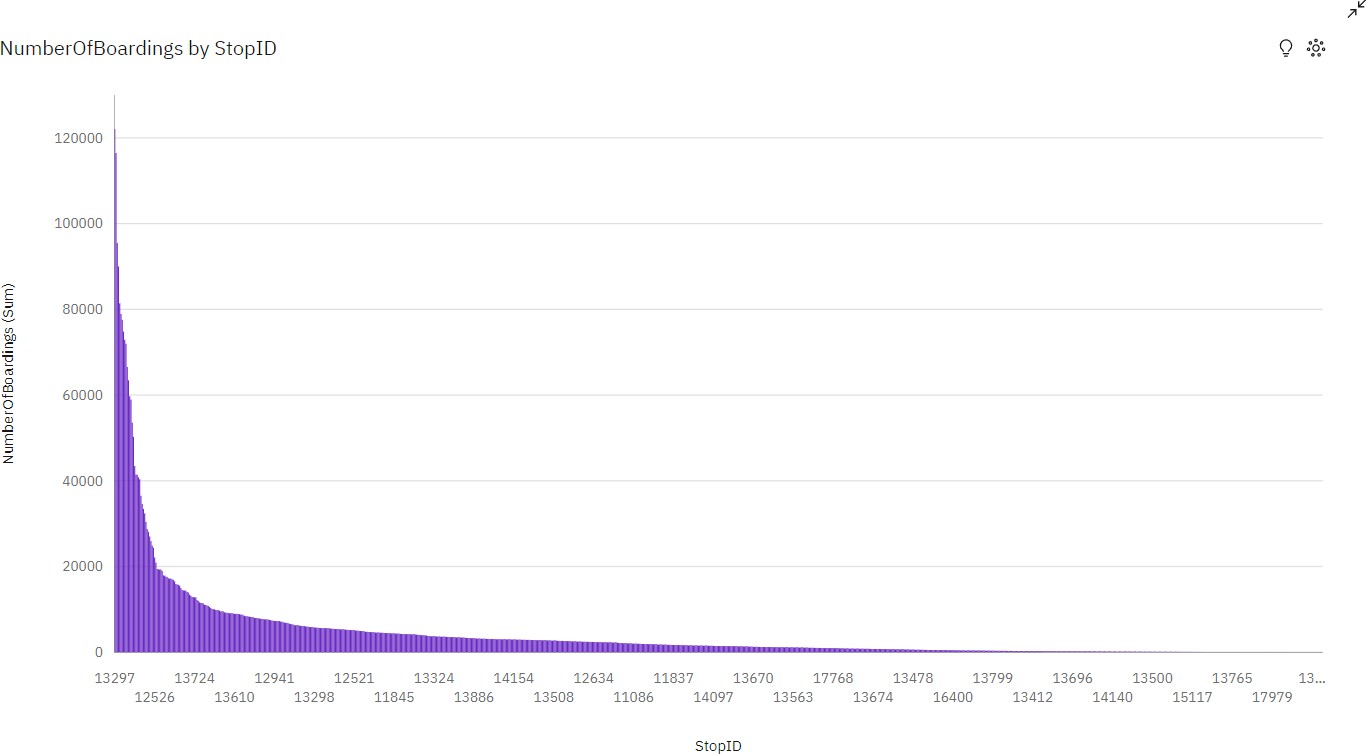


* RouteID 150 has the highest total NumberOfBoardings due to StopID 13297.
* StopID 13297 has the highest NumberOfBoardings at over 122 thousand, out of which RouteID 150 contributed the most at over 45 thousand.



* NumberOfBoardings is unusually high when RouteID is 150.
* It is projected that by 21 Sep 2014, 150 will exceed 118 in NumberOfBoardings by over two thousand.
* From 22 Sep 2013 to 29 Sep 2013, 171's NumberOfBoardings dropped by 90%.
* Across all values of RouteID, the sum of NumberOfBoardings is over 4.3 million.
* NumberOfBoardings ranges from 260, when RouteID is 100S, to nearly 425 thousand, when RouteID is 150.

Morethan1.2lakhsboardingsuseSTOPID13297



* NumberOfBoardings is unusually high when StopID is 13297 and 13278.
* Across all values of StopID, the sum of NumberOfBoardings is over 4.3 million.
* NumberOfBoardings ranges from 1, when StopID is 13277, to over 122 thousand, when StopID is 13297.
* NumberOfBoardings is unusually high when StopID is 13297 and 13278.
* Across all values of StopID, the sum of NumberOfBoardings is over 4.3 million.
* NumberOfBoardings ranges from 1, when StopID is 13277, to over 122 thousand, when StopID is 13297.
* For NumberOfBoardings, the most significant values of StopID are 13297 and 13278, whose respective NumberOfBoardings values add up to over 238 thousand, or 5.5 % of the total.

**CONCLUSION:**

In this part,the given data set is loaded. Then the loaded data is preprocessed which includes cleaning of data, removingduplicates, adding missing values.

Then the data is visualized by using cognos ibm analytics.

By using visualization dashboard ,it is able to identify the busiest route and least busiest route, how much passengers are boarding, on which trip I’d maximum no.of people’s travelled etc.

This visualization gave some variety of ideas about given data set.