Public Transportation Analysis

Phase-3

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.

The code used :

#!/usr/bin/env python

# coding: utf-8

# In[2]:

#Importing necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# In[3]:

#Loading the dataset

data = pd.read\_csv("C:\\Users\\Maha\\Downloads\\Dataset\\PublicTransportDataset.CSV", low\_memory=False)

# In[4]:

#Displaying the first 20 rows

data.head(20)

# In[5]:

# Dropping records which have duplicate values

data.drop\_duplicates(inplace=True)

# In[6]:

# Filling missing values with mean

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

# In[7]:

# Printing the first few rows

print(data.head())

# In[8]:

# Generating descriptive statistics of the dataset

print(data.describe())

# In[9]:

# Generating concise summary of the dataset

print(data.info())

# In[11]:

# Shape of the dataset

print(data.shape)

# In[12]:

# Displaying first few rows after preprocessing

data.head()