TITLE: PUBLIC TRANSPORTATION EFFICIENCY ANALYSIS USING COGNOS

PHASE 3:DEVELOPMENT PART 1

PROJECT DESCRIPTION

In this part,building the project by loading and preprocessing the dataset.Start building the public transportation efficiency analysis using IBM Cognos for visualization.

DEFINE ANALYSIS OBJECTIVES

1.Determine patterns of ridership, including peak hours, popular routes, and demographics of riders. This can help optimize service frequency and capacity.

2. Assess the punctuality of public transportation services. Define standards for on-time arrivals and departures, and analyze deviations from these standards.

3.Evaluate existing routes to identify underused or overcrowded segments. Optimize routes to reduce travel times and enhance overall system efficiency.

4. Examine service reliability by measuring factors such as the frequency of breakdowns, disruptions, and delays. This information can be used to improve maintenance and infrastructure.

5.Analyze the costs associated with operating public transportation services, including fuel, maintenance, labor, and infrastructure. Determine cost-effective measures to reduce expenses while maintaining or improving service quality.

6.Assess the environmental impact of public transportation, including emissions and fuel consumption. Identify opportunities for adopting more sustainable practices.

7.Measure customer satisfaction through surveys and feedback analysis. Understand the areas where passengers are most satisfied or dissatisfied to make targeted improvements.

8.Evaluate the accessibility of public transportation for all passengers, including those with disabilities. Identify areas where improvements are needed to make the system more inclusive.

9.Analyze safety-related incidents and accidents within the public transportation system. Develop strategies to minimize safety risks for passengers and employees.

10.Assess revenue collection methods and the efficiency of fare collection systems. Identify opportunities to reduce fare evasion and increase revenue.

DATA COLLECTION

Identify the sources of transportation data.This could include public transportation authorities government agenta or any other open portals.Here are some key sources of transportation data:

1.Cameras installed at key points on roadways to monitor traffic flow, congestion, and incidents.

2. Data from GPS-enabled devices and navigation apps, which track the location and movement of vehicles.

3.Mobile apps, such as Waze and Google Maps, collect data from users to provide real-time traffic information.

4.Public transportation authorities collect data on routes, schedules, ridership, and service disruptions.

5. Toll systems capture data on vehicles passing through toll booths, including entry and exit times.

6.In-road or above-road sensors that monitor traffic speed, volume, and congestion.

STEPS FOR DATA PREPROCESSING AND CLEANING

Data processing and cleaning in transportation data using Python and Pandas typically involves several steps:

1.Data Import: Import the transportation data into a Pandas DataFrame. You can read data from various formats like CSV, Excel, or databases.

```
import pandas as pd
df =pd.read_csv('transportation_data.csv')
```

2.Data Exploration: Examine the data to understand its structure, column names, and missing values.

```
df.head()  # View the first few rows
df.info ()  # Get data types and non-null counts
```

3.Handling Missing Values: Deal with missing data by either removing rows with missing values or imputing values.

df.dropna() # Remove rows with missing values
df.fillna(0) # Replace missing values with 0

4.Data Cleaning:Clean and standardize data by removing duplicates, correcting data types, and renaming columns.

df.drop_duplicates() # Remove duplicate rows
df['column_name'] = df['column_name'].astype(int) # Convert data types
df.rename(columns={'old_name': 'new_name'}, inplace=True) # Rename columns

5.Data Transformation: Apply transformations to columns, such as datetime parsing or creating new features.

df['date'] = pd.to_datetime(df['date_column']) # Convert to datetime
df['hour'] = df['timestamp_column'].dt.hour # Extract hour from timestamp

6. Data Filtering:Filter data based on specific criteria.

df_filtered = df[df['column_name'] > 100] # Filter rows where a column value is greater than 100

7.Data Aggregation: Aggregate data using groupby to gain insights.

mean_values = df.groupby('category')['value_column'].mean() # Calculate mean values by category

8.Data Export:Save the cleaned data to a new file or database.

df.to_csv('cleaned_transportation_data.csv', index=False) # Save to a new CSV file

PROGRAM

#loading and preprocessing of dataset

#import the required libraries

import pandas as pd

import matplotlib.pyplot as plt

Step 1: Load the dataset

data = pd.read_csv("transportdata.csv")

Step 2: Explore the data

Display the first few rows to get an overview

print(data.head())

```
# Step 3: Data Preprocessing
# Handle missing values (replace NaN with 0)
data = data.fillna(0)

# Data Cleaning (remove duplicates)
data = data.drop_duplicates()
# Step 4: Exploratory Data Analysis (EDA)
# Calculate statistics and generate plots to explore the data
# Example of how to calculate statistics (mean, median)
mean_NumberOfBoardings = data['NumberOfBoardings'].mean()
median_NumberOfBoardings = data['NumberOfBoardings'].median()
print("Mean Cases:", mean_NumberOfBoardings)
print("Median Cases:", median_NumberOfBoardings)
# Step 5: Data Visualization (optional)
column_name = 'NumberOfBoardings'
# Set the size of the plot
plt.figure(figsize=(2,10))
plt.hist(data[column_name],  color='blue',alpha=0.5)
plt.title('public transport analysis')
plt.xlabel(column_name)
plt.ylabel('TripID')
plt.xlim([2, 5])
plt.ylim([0,1000])
plt.show()
```

OUTPUT

```
   TripID RouteID  ...      WeekBeginning NumberOfBoardings
0   23631     100  ...  2013-06-30 00:00:00               1
1   23631     100  ...  2013-06-30 00:00:00               1
2   23632     100  ...  2013-06-30 00:00:00               1
3   23633     100  ...  2013-06-30 00:00:00               2
4   23633     100  ...  2013-06-30 00:00:00               1

[5 rows x 6 columns]
Mean Cases: 4.743736664421159
Median Cases: 2.0
```

PREPROCESSED DATA

It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data analysis task.

This may involve removing errors and inconsistencies,handling missing values,transforming the data into a consistent format and scaling the data to a suitable range.