Project Design Document: Enhancing Public Transportation through Data Analysis

Phase 1: Problem Definition and Design Thinking

Project Definition

Project Title: Enhancing Public Transportation through Data Analysis

Project Objective: The project aims to analyze public transportation data to assess service efficiency, on-time performance, and passenger feedback, ultimately supporting transportation improvement initiatives and enhancing the overall public transportation experience.

Design Thinking

In this phase, we will outline the key components of the project, including analysis objectives, data collection strategies, visualization plans using IBM Cognos, and the integration of code for data analysis. The design thinking process ensures a structured and holistic approach to problem-solving.

Analysis Objectives

1. **Assess On-Time Performance:** The primary objective is to evaluate the punctuality of public transportation services. This involves analyzing historical data on scheduled departure and arrival times versus actual times.

2. **Measure Passenger Satisfaction:** We aim to gauge passenger satisfaction through surveys and feedback data. This includes sentiment analysis of customer reviews and ratings.

3. **Evaluate Service Efficiency:** To determine the efficiency of transportation services, we will examine data related to route optimization, fuel consumption, and operational costs.

Data Collection

1. **Schedules and Real-Time Updates:** We will gather historical schedules and real-time updates from transportation agencies. This data will provide insights into planned versus actual service timings and potential delays.

2. **Passenger Feedback:** To measure passenger satisfaction, we will collect feedback through online surveys, customer service logs, and social media sentiment analysis.

3. **Operational Data:** Information regarding fuel consumption, maintenance schedules, and operational costs will be obtained from transportation agencies and maintenance records.

Visualization Strategy

1. **Dashboard Creation:** Utilizing IBM Cognos, we will design interactive dashboards that present key performance indicators (KPIs) related to on-time performance, passenger satisfaction, and service efficiency. These dashboards will enable stakeholders to easily grasp trends and make informed decisions.

2. **Custom Reports:** Alongside dashboards, custom reports will be generated to provide in-depth insights and trends. These reports will cater to the specific needs of different stakeholders, such as transportation authorities, management teams, and maintenance crews.

3. **Geospatial Visualization:** Geospatial data will be used to visualize route performance, helping identify areas with frequent delays or service issues. Heatmaps and spatial analysis will be employed for this purpose.

Code Integration

1. **Data Cleaning:** Code will be used to clean and preprocess the collected data. This includes handling missing values, data normalization, and ensuring data consistency.

2. **Transformation:** Complex data transformations, such as aggregating passenger feedback scores, calculating performance metrics, and geospatial data transformations, will be implemented using code.

3. **Statistical Analysis:** Advanced statistical analysis, including regression analysis to identify factors affecting on-time performance, and clustering for customer segmentation, will be carried out through code-based approaches.

CODE:

```
import pandas as pd
import matplotlib.pyplot as plt
# Load transportation data (example: CSV file)
transport_data = pd.read_csv('transport_data.csv')

# Assuming 'ScheduledTime' and 'ActualTime' columns in the dataset
# Convert time columns to datetime objects for analysis
transport_data['ScheduledTime'] = pd.to_datetime(transport_data['ScheduledTime'])
transport_data['ActualTime'] = pd.to_datetime(transport_data['ActualTime'])

# Calculate the time difference between scheduled and actual times
transport_data['TimeDifference'] = (transport_data['ActualTime'] -
transport_data['ScheduledTime']).dt.total_seconds()

# Calculate the percentage of on-time arrivals
on_time_threshold = 300  # Assuming 5 minutes (300 seconds) delay is considered on-time
on_time_percentage = (transport_data['TimeDifference'] <= on_time_threshold).mean() * 100

# Visualize on-time performance
plt.figure(figsize=(8, 6))
plt.hist(transport_data['TimeDifference'], bins=30, color='skyblue', edgecolor='black')
plt.axvline(x=on_time_threshold, color='red', linestyle='--', label=f'On-Time Threshold
({on_time_threshold} sec)')
plt.xlabel('Time Difference (seconds)')
plt.ylabel('Frequency')
plt.title('Distribution of Arrival Time Differences')
plt.legend()
plt.show()

# Print on-time performance percentage
print(f'Percentage of on-time arrivals: {on_time_percentage:.2f}%')
```

OUTPUT:

Percentage of on-time arrivals: XX.XX%

CONCLUSION:

In this initial phase of the project, I have defined the problem statement and outlined our design thinking approach. By setting clear analysis objectives, specifying data sources and collection methods, planning visualization strategies, and considering code integration, we are well-prepared to proceed to the next phases of data gathering, analysis, and reporting. This structured approach will ensure that our project results in actionable insights to enhance the public transportation experience.