PUBLIC TRANSPORTATION EFFICIENCY ANALYSIS

PHASE II

7145\_UNITEDINSTITUTEOFTECHNOLOGY\_Proj\_212990\_Team\_1

Team Mentor: [kavitha@uit.ac.in](mailto:kavitha@uit.ac.in)

Member Name: Mithun V

rajakuttimithun07@gmail.com

**Aim:**

* To analyse the public transportation efficiency,

**Steps or algorithm for analysis**

1. **Data Collection and Cleaning**:

- Collect the dataset containing the relevant fields (tripID, RouteID, StopID, Stopname, Week beginning , Number of boardings).

- Cleaning the dataset by removing duplicates, handling missing data, and correcting any obvious errors.

- The dataset is got from the given kaggle link: https://www.kaggle.com/datasets/rednivrug/unisys?select=20140711.CSV

2. **Data Exploration:**

- Exploring the dataset to gain a better understanding of its characteristics.

- Calculating summary statistics for the Number of boardings to get an initial sense of the data distribution.

3. **Data Preprocessing:**

- Formatting and standardizing the data.

- Converting data types to appropriate formats.

- Grouping the data by relevant attributes (e.g., RouteID, Weekbeginning) for aggregation.

4. **Aggregation and Summarization:**

- Aggregating data to calculate relevant metrics, such as average daily boardings per route or stop.

- Summarizing the data to identify trends or patterns in public transport usage over time.

5. **Visualization:**

- Creating visualizations (e.g., line charts, bar charts, heatmaps) to present the data effectively.

- Visualizing trends in public transport usage over time, identify peak days, or visualizing RouteID-specific performance.

6. **Performance Metrics:**

- Defining performance metrics to measure public transport efficiency. This will include:

- Average boardings per trip or route.

- Boardings per stop.

- Comparing peak vs. off-peak usage.

- Trends in weekly or monthly boardings.

- Performance metrics are essential for evaluating and assessing the efficiency of a public transportation system. With a dataset containing tripID, RouteID, StopID, Stopname, Week beginning, and Number of boardings, we can calculate several key performance metrics to measure the effectiveness and quality of the service.

7. **Hypothesis Testing:**

- Specific hypotheses about factors affecting public transport efficiency to perform statistical tests to validate or reject these hypotheses.

- Hypothesis testing is done to compare different groups or conditions and draw conclusions about the system's performance

- A hypothesis test using StopID as the basis to compare the average number of boardings at different stops is done. Below is sourcecode of a two-sample t-test to determine if there is a significant difference in the average boardings between two different stops (StopID A and StopID B)

**Source code:**

import pandas as pd

from scipy import stats

df = pd.read\_csv("20140711.CSV ")

stop\_A\_data = df[df['StopID'] == 'StopID A']['Number of boardings']

stop\_B\_data = df[df['StopID'] == 'StopID B']['Number of boardings']

t\_stat, p\_value = stats.ttest\_ind(stop\_A\_data, stop\_B\_data, equal\_var=False)

alpha = 0.05

if p\_value < alpha:

print("There is a significant difference in the average boardings between StopID A and StopID B.")

else:

print("There is no significant difference in the average boardings between StopID A and StopID B.")

print("T-statistic:", t\_stat)

print("P-value:", p\_value)

8. **Machine Learning Models:**

- To predict future public transport usage or identify factors influencing it, Predictive models using machine learning techniques like random forest and natural language processing will be performed.

- **Service disruptions**

For predicting service disruptions we are using classification algorithms like **Random Forest** Below is the source code for predicting service disruption using Random Forest,

**Source code:**

import pandas as pd

from sklearn.preprocessing import LabelEncoder

data = pd.read\_csv("20140711.CSV ")

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

label\_encoders = {}

categorical\_features = ["TripID", "RouteID", "StopID", "Stopname"]

for feature in categorical\_features:

label\_encoders[feature] = LabelEncoder()

data[feature] = label\_encoders[feature].fit\_transform(data[feature])

X = data[["TripID", "RouteID", "StopID", "Stopname", "Week beginning", "Number of boardings"]]

y = data["ServiceDisruptionLabel"]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

clf = RandomForestClassifier(n\_estimators=100, random\_state=42)

clf.fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

confusion = confusion\_matrix(y\_test, y\_pred)

print("Accuracy:", accuracy)

print("Classification Report:\n", report)

print("Confusion Matrix:\n", confusion)

new\_data = pd.DataFrame({"TripID": [1], "RouteID": [2], "StopID": [3], "Stopname": [4], "Week beginning": ["2023-01-01"], "Number of boardings": [100]})

new\_data["TripID"] = label\_encoders["TripID"].transform(new\_data["TripID"])

prediction = clf.predict(new\_data)

- **Passenger Sentiment Analysis:**

Passenger sentiment analysis using **Natural Language Processing** (NLP) can help us understand the sentiments and opinions of passengers based on their feedback. However, for sentiment analysis, we would typically need a dataset that contains the text feedback provided by passengers, alongside other relevant information like TripID, RouteID, StopID, Stopname, Week beginning, and Number of boardings. Below the steps to perform passenger sentiment analysis using Python and common NLP libraries are stated.

**Source code:**

import pandas as pd

import spacy

from nltk.sentiment.vader import SentimentIntensityAnalyzer

import matplotlib.pyplot as plt

df = pd.read\_csv("20140711.CSV ")

df['Feedback'] = df['Feedback'].str.lower().str.replace('[^a-z\s]', '', regex=True)

nlp = spacy.load("en\_core\_web\_sm")

df['Tokenized\_Feedback'] = df['Feedback'].apply(lambda x: [token.text for token in nlp(x)])

analyzer = SentimentIntensityAnalyzer()

def analyze\_sentiment(text):

sentiment = analyzer.polarity\_scores(text)

if sentiment['compound'] >= 0.05:

return 'positive'

elif sentiment['compound'] <= -0.05:

return 'negative'

else:

return 'neutral'

df['Sentiment'] = df['Feedback'].apply(analyze\_sentiment)

sentiment\_distribution = df['Sentiment'].value\_counts()

trip\_sentiments = df.groupby('TripID')['Sentiment'].value\_counts()

plt.bar(sentiment\_distribution.index, sentiment\_distribution.values)

plt.xlabel('Sentiment')

plt.ylabel('Count')

plt.show()

**VADER (Valence Aware** analysis tool that is specifically attuned to sentiments expressed in social media. **VADER** uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. **VADER** not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.**Dictionary and sEntiment Reasoner)** is a lexicon and rule-based sentiment. Hence the VADER analysis tool is used to analyse the passenger sentiment.

9. **Recommendations and Insights:**

- Based on our analysis, insights and recommendations to improve public transport efficiency will be provided.

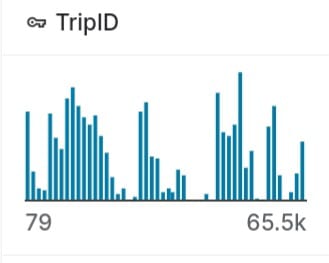
- Suggesting route adjustments, schedule changes, or infrastructure improvements as needed.

10. **Iterate and Refine:**

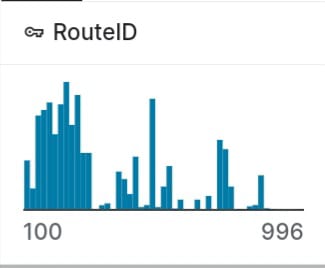
- Continuously monitoring public transport data and iterating on your analysis to adapt to changing conditions and evaluate the impact of implemented recommendations.

Remember that the specific analysis you perform will depend on the goals and questions you have regarding public transport efficiency. Adjust the steps accordingly to meet your objectives and the complexity of your dataset.

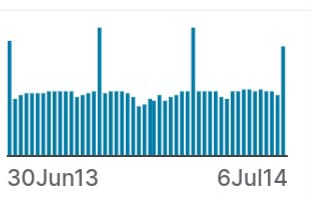
11. **visualisation of dataset:**



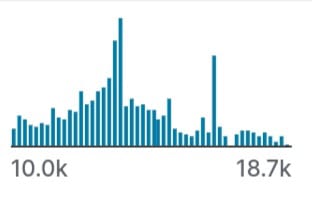
ROUTEID VISUALISATION



WEEK BEGINNING VISUALISATION



STOPID VISUALISATION



NUMBER OF BOARDINGS VISUALISATION



Conclution:

In conclusion, public transport efficiency analysis is a multifaceted process that involves data preprocessing, the calculation of various performance metrics, geospatial analysis, and hypothesis testing. By applying these techniques to a dataset containing trip and passenger information, transportation authorities and analysts can gain valuable insights into system performance, identify areas for improvement, and make data-driven decisions to enhance public transportation services. The specific analysis methods and metrics chosen will depend on the objectives and goals of the analysis, as well as the unique characteristics of the public transport system under evaluation