Sukanya

Capstone Project - 3

NJ Transit Delays-An Analysis

Train Delays

- Ensuring smooth operation of transport services is necessary for providing efficient reliable passenger transport service

- For transportation authorities like NJ Transit, understanding the underlying patterns and causes of train delays is crucial for optimizing service reliability, minimizing disruptions, and enhancing overall passenger satisfaction.

Sukanya

Dataset Source

url:
https://www.kaggle.com/datasets/pranavbadami/nj-transit-amtrak-nec-performance/data

Why Customer Retention

"NJ Transit delays are not just an inconvenience; they represent a systemic failure in our public transportation system that needs urgent attention."

- Transit Activist, Michael Brown

Sukanya

Exploration & Analysis

Sukanya

Data Exploration

- The dataset has 27 months of data.
- The data under analysis has 98698 records and 13 columns.

```
1   data.shape
```

```
(98698, 13)
```

```
1   data.columns
```

```
Index(['date', 'train_id', 'stop_sequence', 'from', 'from_id', 'to', 'to_id',
       'scheduled_time', 'actual_time', 'delay_minutes', 'status', 'line',
       'type'],
      dtype='object')
```

Sukanya

Data Preprocessing

```
1   data.dtypes
```

```
date                object
train_id            object
stop_sequence      float64
from                object
from_id              int64
to                  object
to_id                int64
scheduled_time      object
actual_time         object
delay_minutes      float64
status              object
line                object
type                object
dtype: object
```

- The date is object type in the dataset
- Changed the object type of date to date format
- Changed the actual_time and schedule_time columns to date time instead of objects
- Nulls are dropped.

```
1   data.dropna(inplace=True)
2   data.shape
```

```
(87172, 15)
```

Sukanya

Data Wrangling-Data Reduction & Type Conversion

```
#Changing date columnns to dates

data['scheduled_time'] = pd.to_datetime(data['scheduled_time'])
data['actual_time'] = pd.to_datetime(data['actual_time'])

# Extract day of the week from 'actual_time'
data['day_of_week'] = data['actual_time'].dt.dayofweek

# Extract hour of the day from 'actual_time'
data['hour_of_day'] = data['actual_time'].dt.hour
```

Sukanya

One Hot Encoding

```
# Define preprocessing steps for numerical and categorical features

numeric_features = ['delay_minutes']
categorical_features = ['day_of_week', 'hour_of_day', 'from', 'to', 'line']
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
```

Assigning numerical values to categorical values

Sukanya

Unsupervised Machine Learning

Sukanya

Model Training

Clustering models used:

- K-Means
- PCA (Principal Component Analysis)
- T-SNE(t-distributed Stochastic Neighbor Embedding)
- GMM (Gaussian Mixture Model)

Sukanya

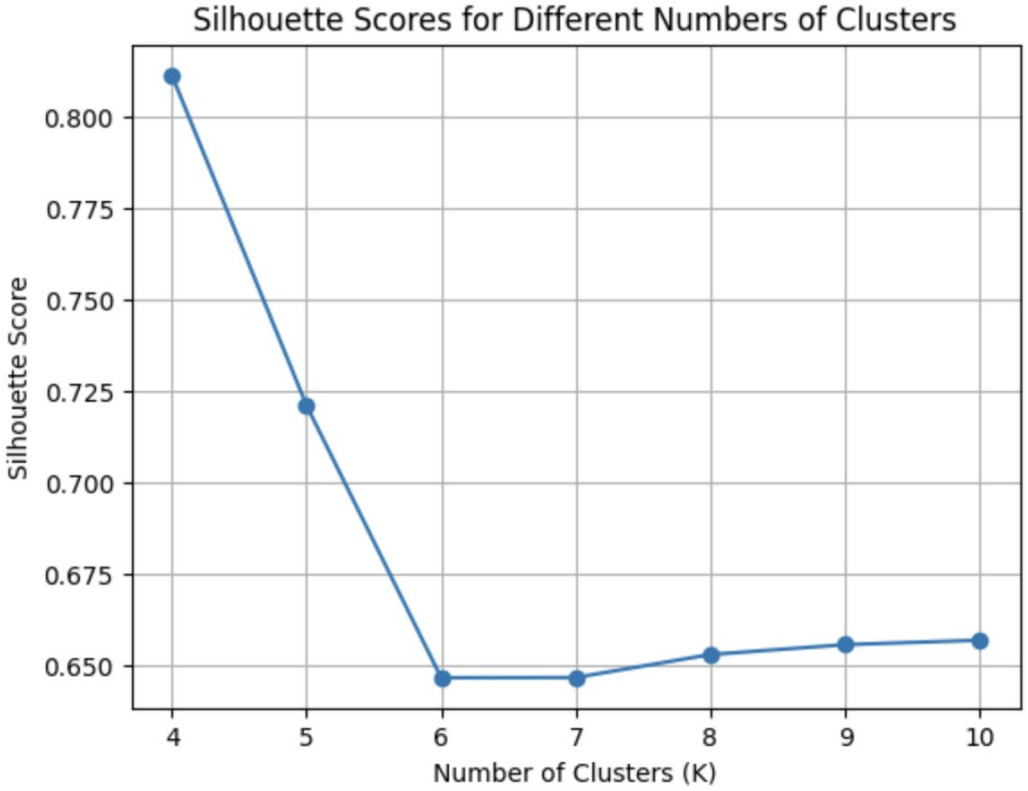
Features Selection

```
selected_features = ['day_of_week', 'hour_of_day', 'from', 'to', 'line', 'delay_minutes']
selected_data = data[selected_features]
selected_data.columns

selected_data['day_of_week']=selected_data['day_of_week'].astype(int)

selected_data['hour_of_day']=selected_data['hour_of_day'].astype(int)
```

K-Means with PCA Output



It is observed that for k=4 has the highest silhouette scores

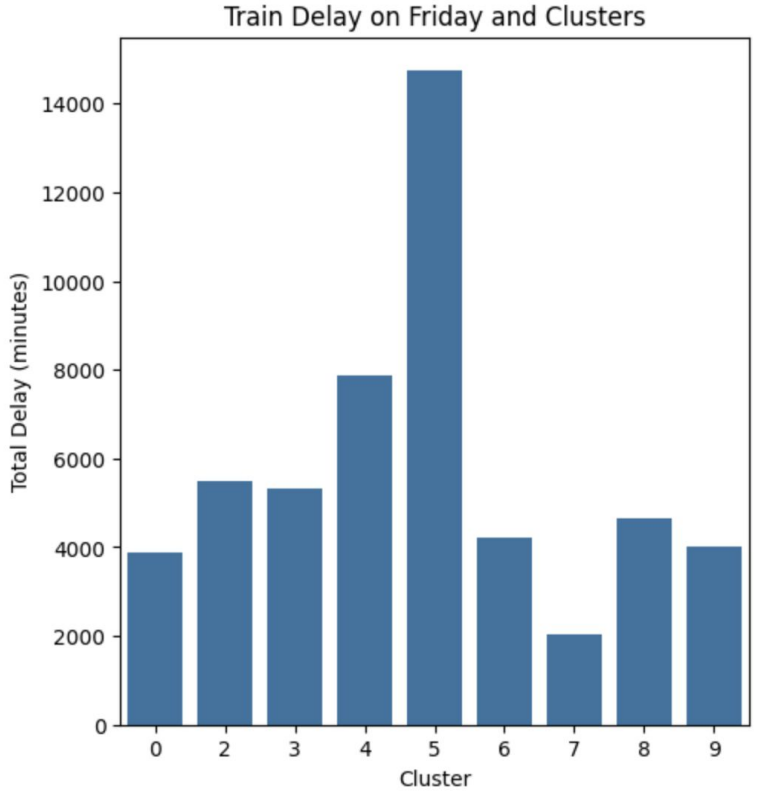
Sukanya

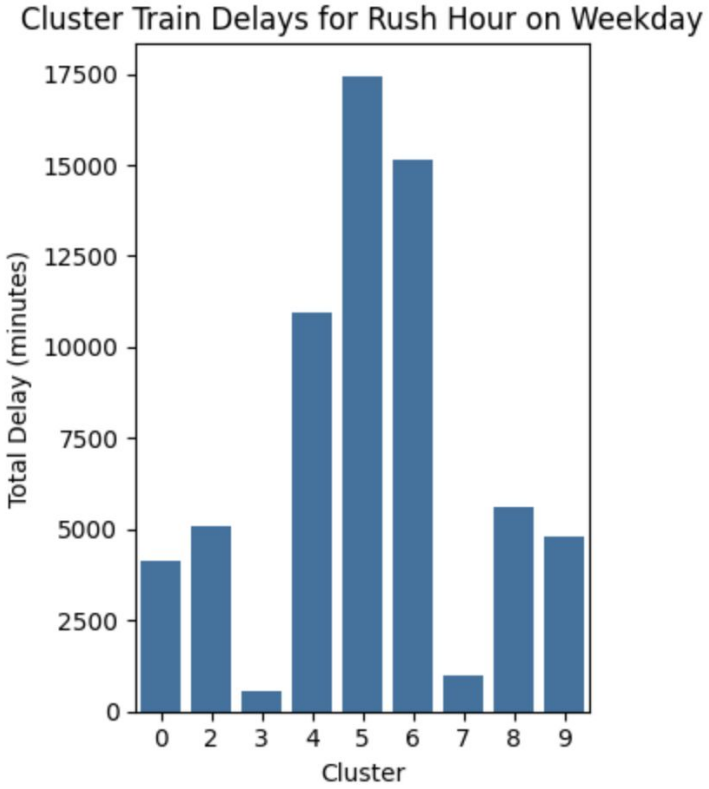
T-SNE with GMM

It is observed that for k=10 has the highest silhouette scores

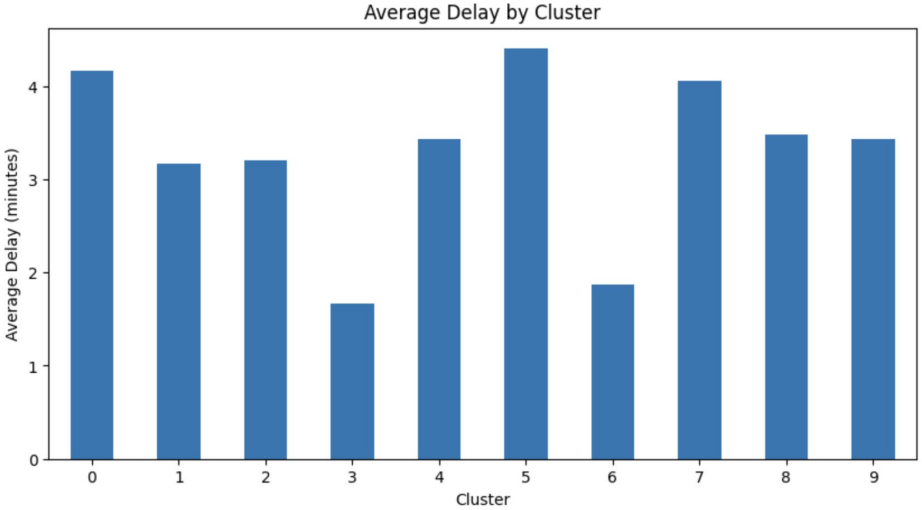
Sukanya

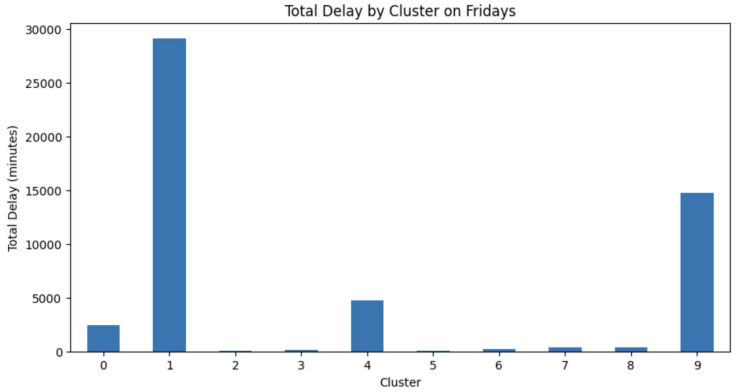
Output Visualizations for K Means & PCA

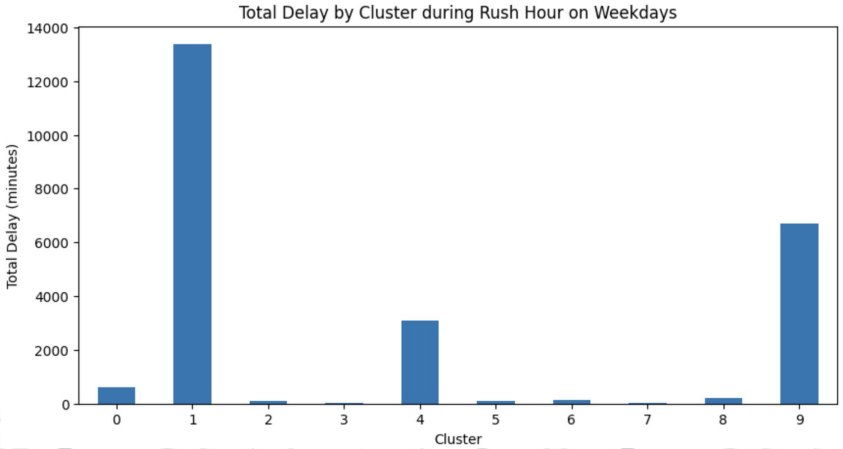




Output Visualizations for T-SNE & GMM







Sukanya

Challenges & Recommendations

- The dataset was huge and was constantly facing issues with timeout errors in google collab. Due to this, was able to check only for specific month file for limited k values ranging from 4 to 10

- Utilizing predictive outcomes derived from one month's data can serve as a foundational strategy for extrapolating insights to optimize operations across broader timeframes. By leveraging these predictions, particularly during peak rush hours on weekdays, transportation authorities can proactively introduce additional trains with increased frequencies. This proactive approach aims to alleviate congestion, enhance reliability, and improve the transportation experience for passengers.

Sukanya

Thank You