**PHASE:5**

**PUBLIC TRANSPORTATION EFFICIENCY ANALYSIS**

**TEAM MEMBERS:**

NAME: Murali V

REG NO:721221104040

NAME: Ragul M

REG NO:721221104044

NAME: Aswin S

REG NO:721221104009

NAME: Arun S

REG NO:72122110400

NAME: Abishek S

REG NO:721221104004

**Table of Contents:**

* Abstract
* Module Outline

# Innovation

* Machine Learning Algorithms
* Data Collection
* Data Preprocessing and Cleaning
* IBM Cognos Visualizations
* Data Analysis
* Conclusion

**ABSTRACT:**

For modern urban living, efficient and accessible public transportation systems are essential. This project undertakes a thorough "Public Transportation Analysis" to examine the existing health of public transit networks, determine how they affect communities, and suggest data driven improvements. This analysis aims to improve the efficiency and sustainability of public transportation, ultimately seeking to create more connected and liveable communities. It does this through data collection, route optimization, and community participation.

**MODULE OUTLINE:**

1. **Data Collection and Preprocessing:** Data Sources: Collect data from various sources such as GPS trackers on vehicles, ticketing systems, weather APIs, and traffic data. Data Cleaning: Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies. Data Integration: Combine data from multiple sources to create a unified dataset for analysis.
2. **Exploratory Data Analysis (EDA):** Visualize and explore the data to gain insights into passenger behaviour, traffic patterns, and system performance. Identify key trends and anomalies in the data.
3. **Demand Forecasting:** Develop models to forecast passenger demand for different routes and times. Utilize time series analysis and machine learning algorithms for accurate predictions.
4. **Route Optimization:** Optimize bus/train routes based on historical data and real-time traffic information. Minimize travel time, reduce congestion, and improve efficiency.
5. **Real-time Passenger Information:** Create a system to provide real-time updates to passengers, including estimated arrival times and service disruptions. Enhance the passenger experience and increase overall satisfaction.
6. **Maintenance Scheduling:** Implement predictive maintenance models to schedule maintenance activities for vehicles and infrastructure. Reduce downtime and ensure the safety and reliability of public transportation systems.
7. **Performance Monitoring:** Develop dashboards and KPIs to monitor the performance of public transportation systems. Continuously assess and improve system efficiency and reliability.
8. **Sustainability Analysis:** Analyse the environmental impact of public transportation systems. Explore ways to reduce emissions and promote sustainability.
9. **Stakeholder Engagement:** Collaborate with transportation authorities, city planners, and other stakeholders to implement data-driven solutions. Ensure that the project aligns with the goals and needs of the community.
10. **Evaluation and Feedback Loop:** Continuously evaluate the effectiveness of implemented solutions. Collect feedback from passengers and stakeholders for further improvements.

# INNOVATION:

* **Real-time Data Integration:**

Implement a system that collects real-time data from buses, trains, and other public

transportation vehicles. Analyze this data to optimize routes, predict delays, and improve overall efficiency.

# Smart Ticketing Solutions:

Develop a smart ticketing system that uses contactless cards or mobile apps for payments.

This can reduce congestion at ticket counters and make the payment process more convenient for passengers.

# User Experience Analysis:

Conduct surveys and usability studies to gauge the experience of public transportation users. Use this feedback to make improvements in areas like accessibility, comfort, and safety.

# Route Optimization:

Use machine learning algorithms to optimize bus and train routes based on historical data. This can help reduce travel times and energy consumption.

# Accessibility for Persons with Disabilities:

Ensure that public transportation is fully accessible to people with disabilities. Analyze current accessibility and propose improvements.

# Safety and Security Enhancements:

Analyze security measures in place and suggest innovations to enhance passenger safety. This could involve surveillance technologies, emergency response systems, or incident prediction algorithms.

# MACHINE LEARNING ALGORITHMS:

Here are machine learning methods for predicting service disruptions and analyzing passenger sentiment in public transport:

## Predicting Service Disruptions:

**ARIMA**: A tool for forecasting transport delays.

**Isolation Forest and One-Class SVM**: Detect unusual events in transport data.

**Recurrent Neural Networks (RNNs):** Good for predicting delays**.**

## Analyzing Passenger Sentiment:

**VADER and Text Classification**: Understand if passenger feedback is positive, negative, or neutral.

**LDA:** Find common topics in feedback.

**Word Embeddings**: Helps to understand the context of words.

**Deep Learning (CNNs and RNNs):** Understand sentiment in text.

**BERT**: A powerful tool for understanding language.

**Random Forest and Gradient Boosting:** Useful for classifying text.

**DATA COLLECTION:**

* Collect transportation data from the provided source. Ensure that the data includes information about routes, schedules, delays, ridership, and any other relevant metrics.

**DatasetLink:**[**https://www.kaggle.com/datasets/rednivrug/unisys?select=20140711.CSV**](https://www.kaggle.com/datasets/rednivrug/unisys?select=20140711.CSV)

**DATA PREPROCESSING AND CLEANING:**

* Clean the collected data to ensure its quality and accuracy.

**#importing data set**

import pandas as pd

import numpy as np

import sklearn

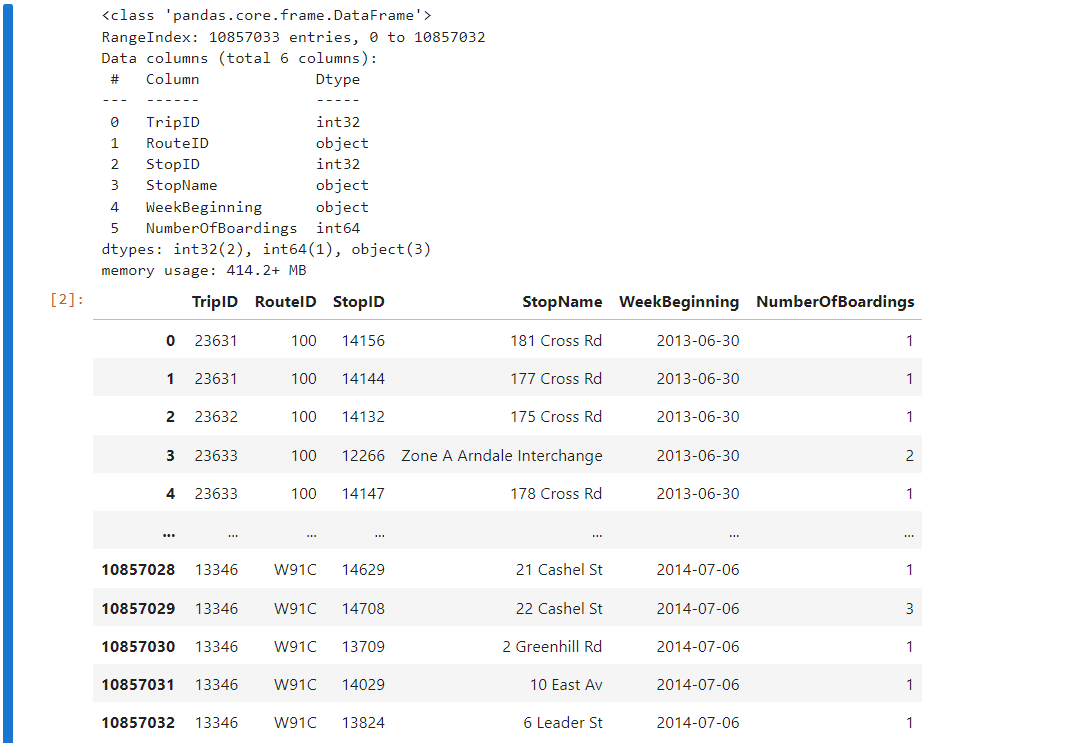
from sklearn.preprocessing import StandardScaler

data = pd.read\_csv("public transport.CSV",dtype={'TripID': int, 'RouteID': str, 'StopID': int, 'WeekBeginning': str})

data.info()

data

**output:**



**#cleansing the data set**

data['StopName'].fillna('Unknown', inplace=True)

data.drop\_duplicates(subset=['TripID', 'StopID', 'WeekBeginning'], keep='first', inplace=True)

data['TripID'] = data['TripID'].astype(int)

data['RouteID'] = data['RouteID'].astype(str)

data['StopID'] = data['StopID'].astype(int)

data['WeekBeginning'] = pd.to\_datetime(data['WeekBeginning'])

#stored the cleaned data into another file

data.to\_csv('cleanddataset.csv', index=False)

**Exploratory Data Analysis (EDA):**

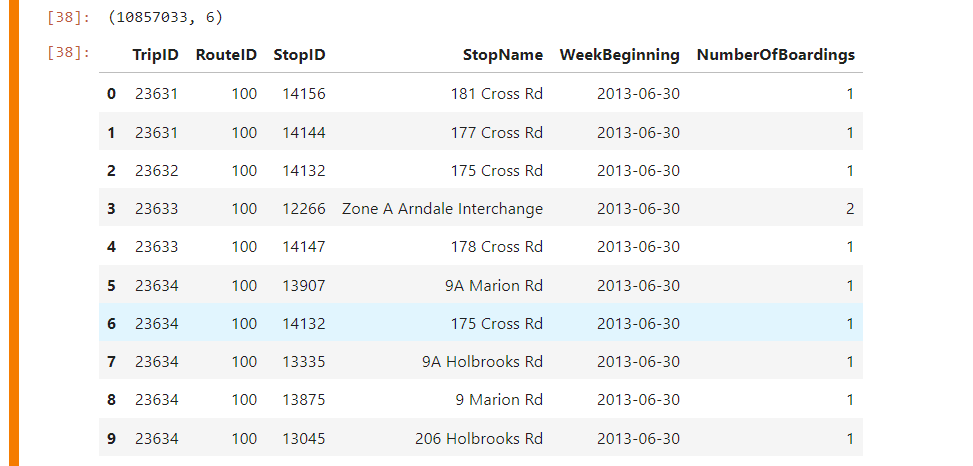
data = pd.read\_csv('cleanddataset.csv')

data.shape

data.head(10)

data.sample(5)

**output:**



data.shape

**output:**

****

data.columns

**output:**



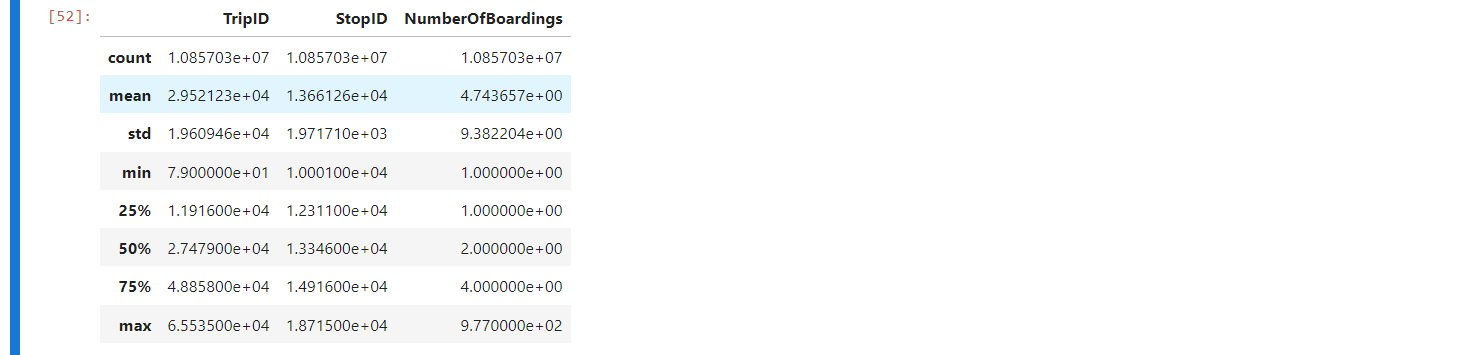
pd.isnull(data).sum()

**output:**



data.describe()

**output:**



data.nunique()

**output:**



**Visualization and Analysis:**

##can assign the each chart to one axes at a time

fig,axrr=plt.subplots(2,2,figsize=(15,15))

ax=axrr[0][0]

ax.set\_title("No of Boardings")

data['NumberOfBoardings'].value\_counts().sort\_index().head(20).plot.bar(ax=axrr[0][0])

ax=axrr[0][1]

ax.set\_title("WeekBeginning")

data['WeekBeginning'].value\_counts().plot.area(ax=axrr[0][1])

ax=axrr[1][0]

ax.set\_title("most Busiest Route")

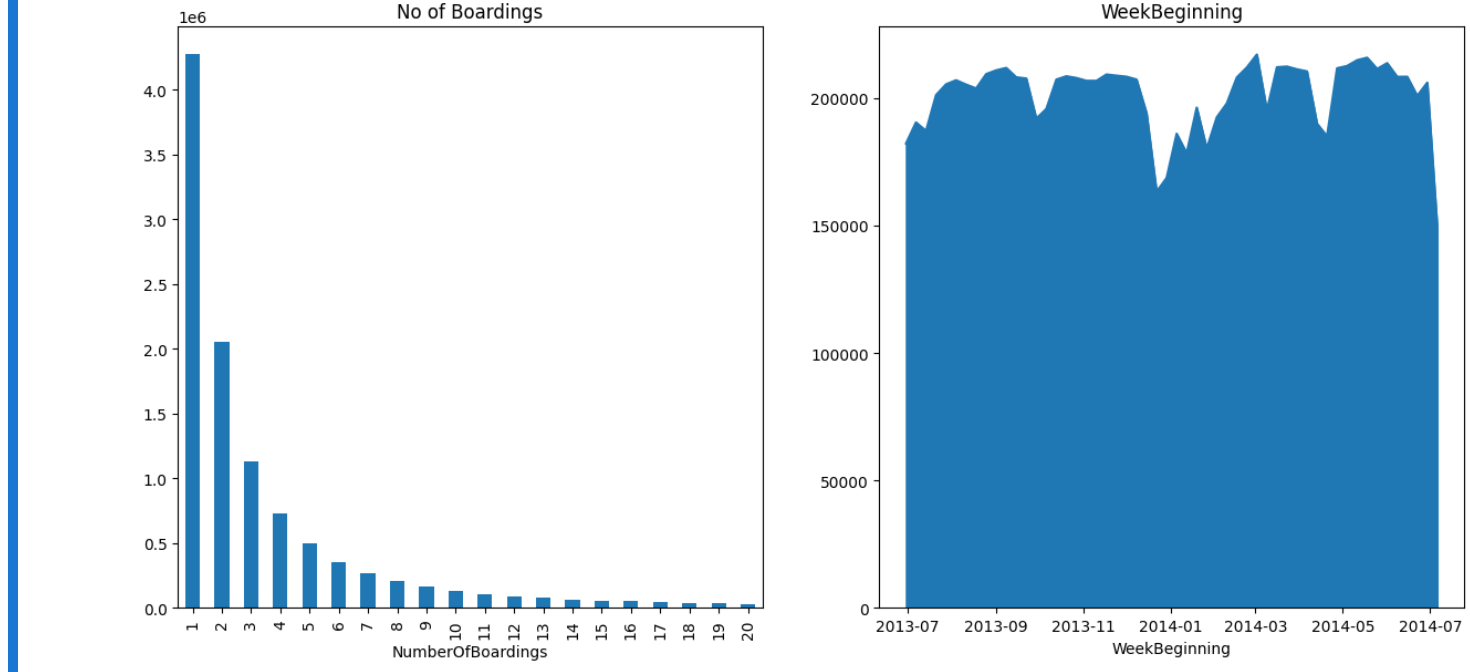
data['RouteID'].value\_counts().head(10).plot.bar(ax=axrr[1][0])

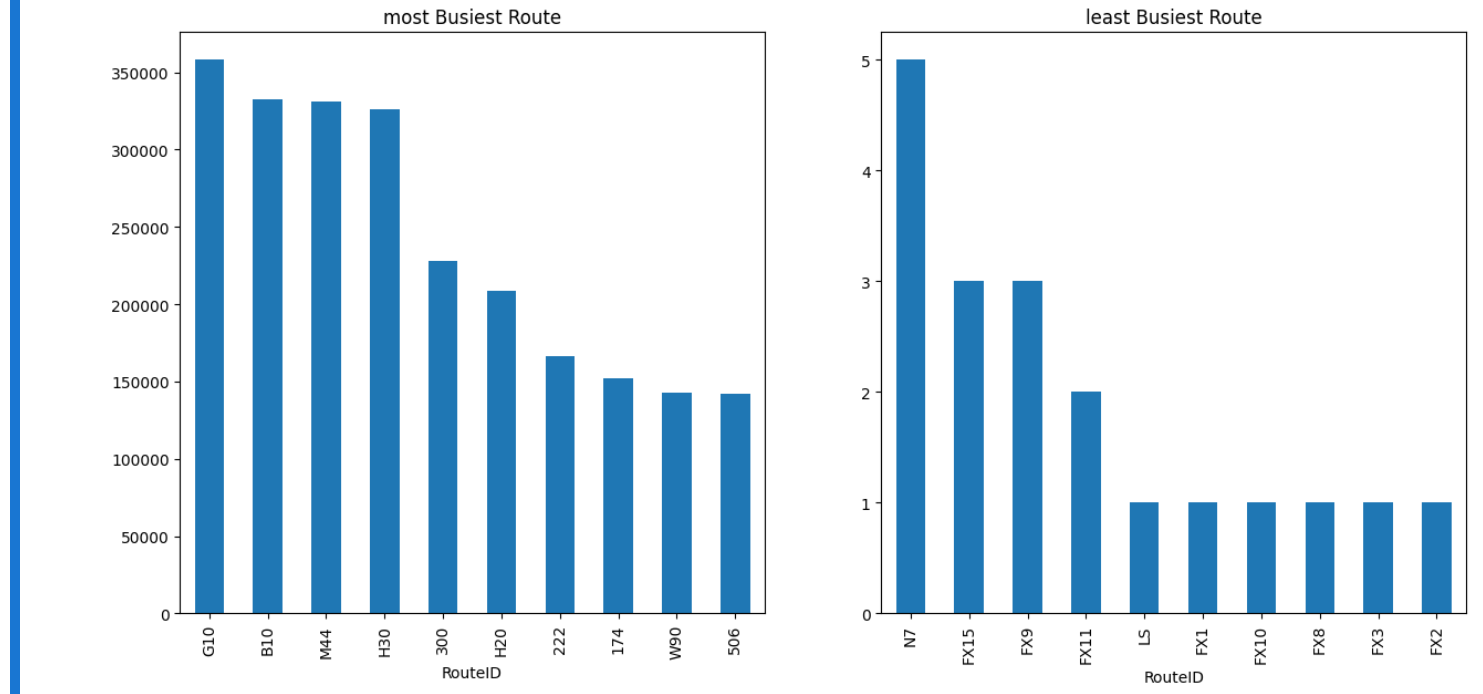
ax=axrr[1][1]

ax.set\_title("least Busiest Route")

data['RouteID'].value\_counts().tail(10).plot.bar(ax=axrr[1][1])

**output:**





**My Source code:**[**https://drive.google.com/file/d/1UoqWbuUc8eYu3XAmw-lo8MQuFM013HKZ/view?usp=drive\_link**](https://drive.google.com/file/d/1UoqWbuUc8eYu3XAmw-lo8MQuFM013HKZ/view?usp=drive_link)

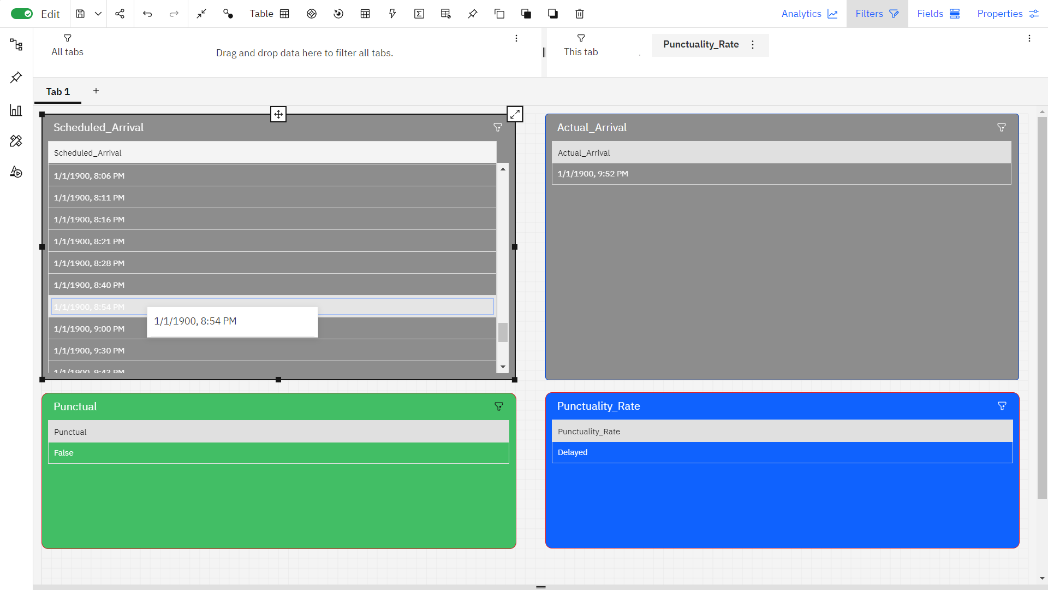
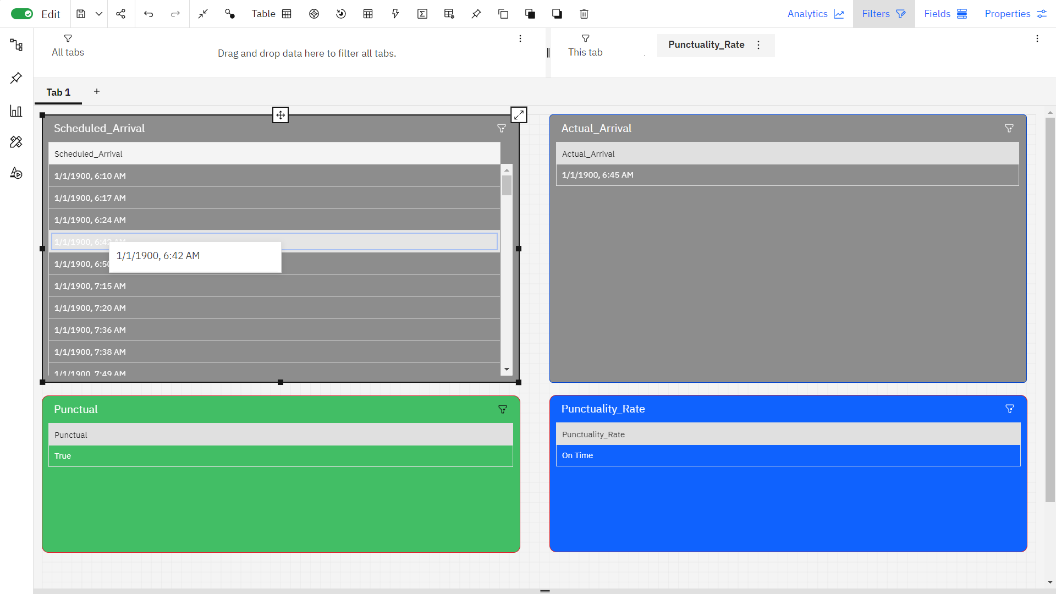
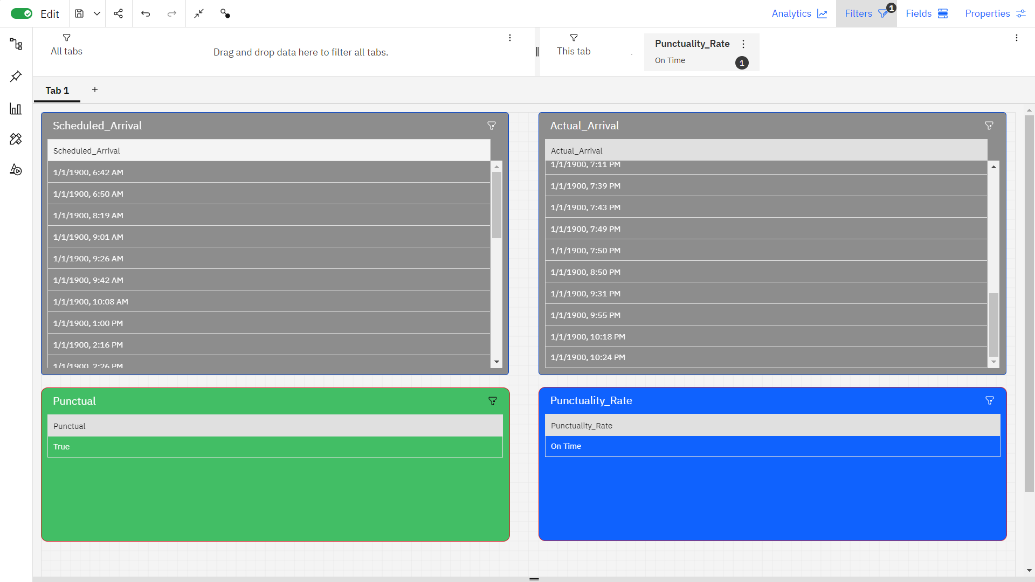
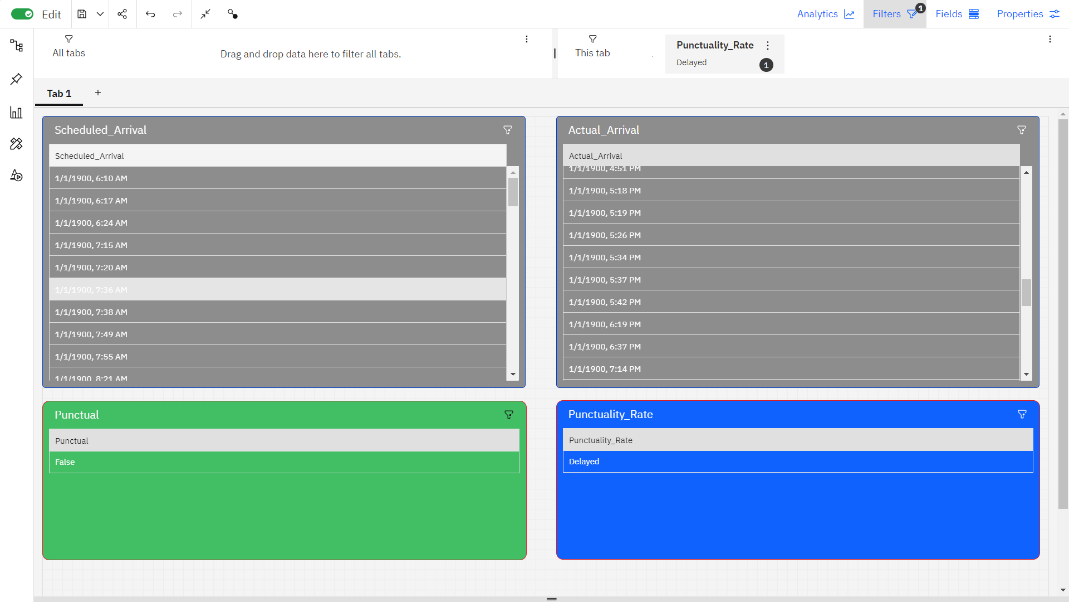
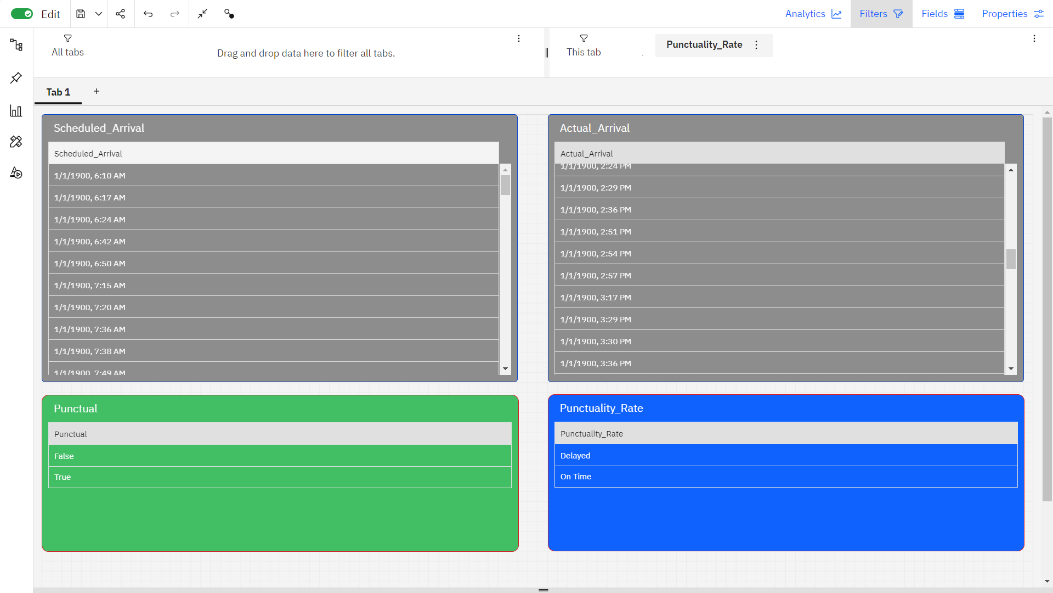
**My data sets:** [**https://drive.google.com/drive/folders/1aM2-ROjLiMaz6aXkKMvARcEmaeV6fYow?usp=drive\_link**](https://drive.google.com/drive/folders/1aM2-ROjLiMaz6aXkKMvARcEmaeV6fYow?usp=drive_link)

**IBM COGNOS VISUALIZATIONS**

**On-Time Performance Dashboard:**

The On-Time Performance Dashboard is designed to provide a visual representation of the punctuality and reliability of a service.

**Screenshots:**

****

In the On-Time Performance Analysis Dashboard, the following key information is presented:

**1. Actual Arrival Time:** This is the time at which a service, such as a bus, train, or flight, actually arrives at its destination. It is recorded in real-time and provides an accurate indication of when the service reaches its final stop.

**2. Scheduled Arrival Time:** The scheduled arrival time represents the expected or planned arrival time for the service. It's the time passengers were informed to expect the service to arrive.

**3. Punctuality Rate:** The punctuality rate is a percentage that indicates how often the service arrives on time, i.e., within the acceptable time window defined for punctuality. It's calculated by comparing the actual arrival time to the scheduled arrival time.

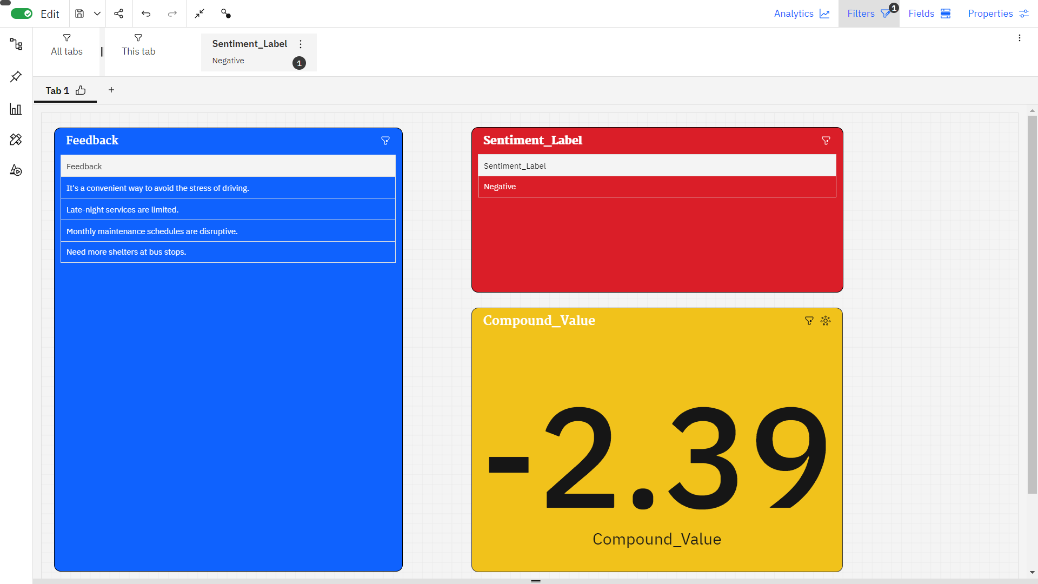
**4. Punctual or Delayed:** This is a categorical indicator that classifies whether the service's arrival was punctual or delayed. If the actual arrival time falls within the defined punctuality time window, it's labeled as "Punctual." If it's outside that time window, it's labeled as "Delayed."

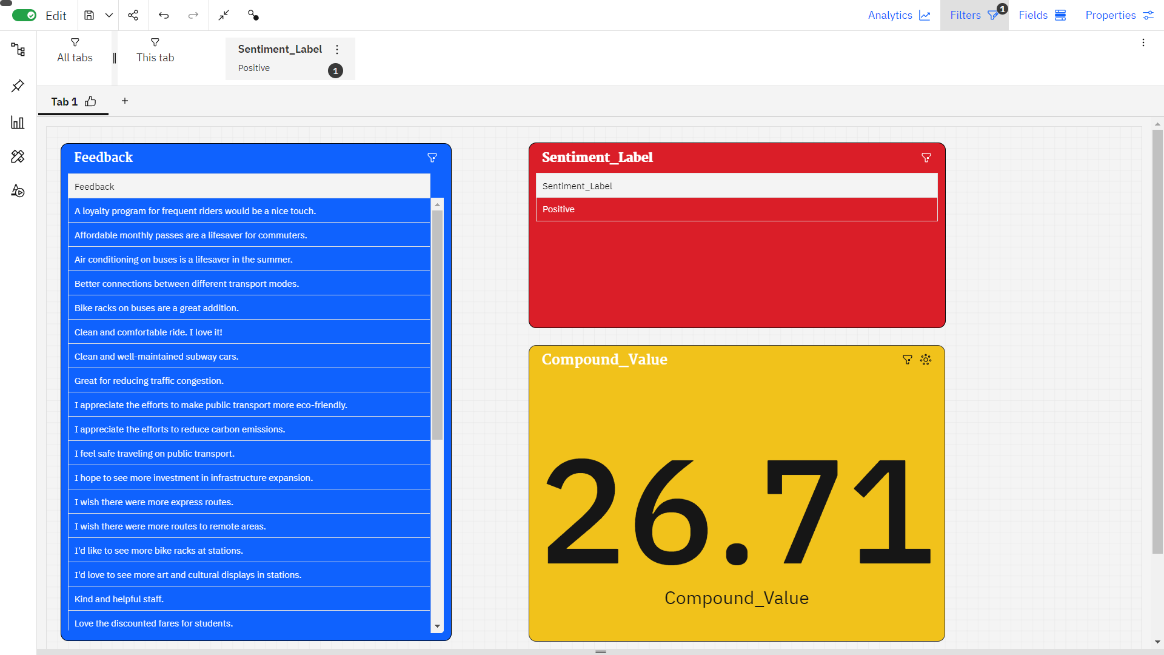
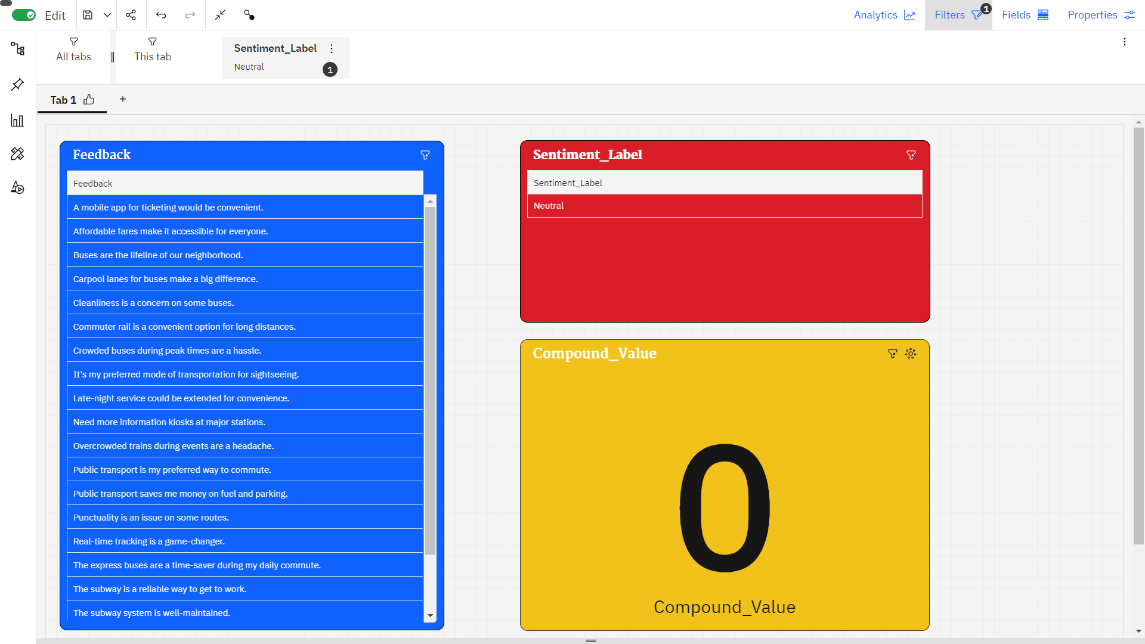
**Dashboard Link:** [**https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my\_folders%2Fdashboard%2Fpuntuality&action=view&mode=dashboard**](https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2Fdashboard%2Fpuntuality&action=view&mode=dashboard)

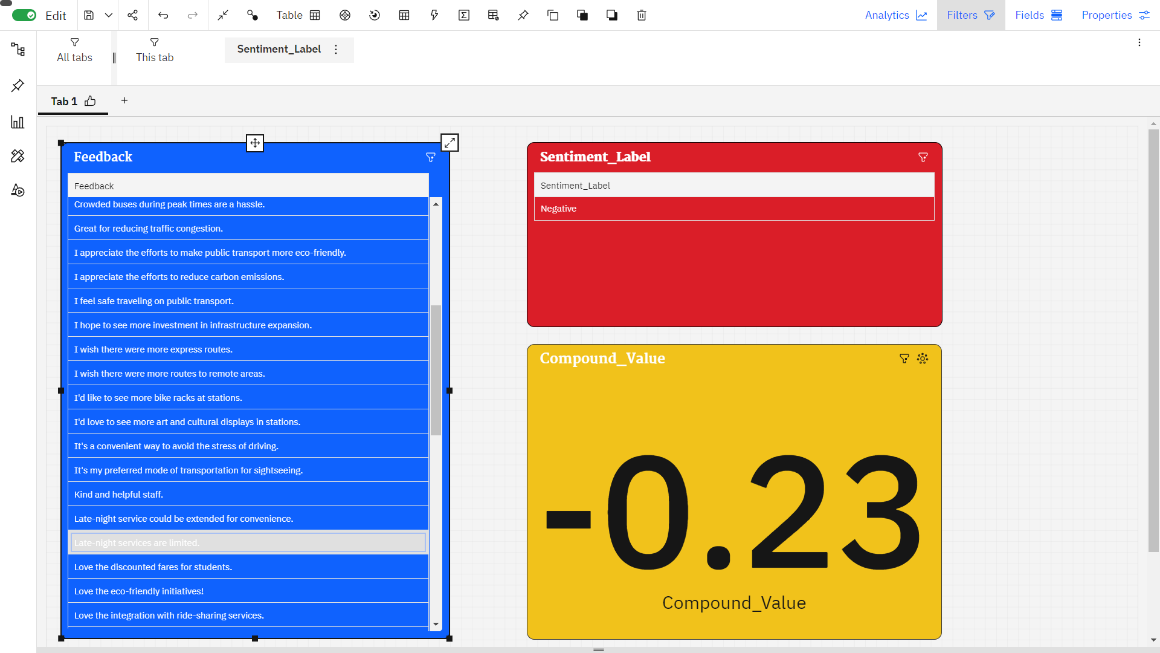
**Passenger Feedback Dashboard:**

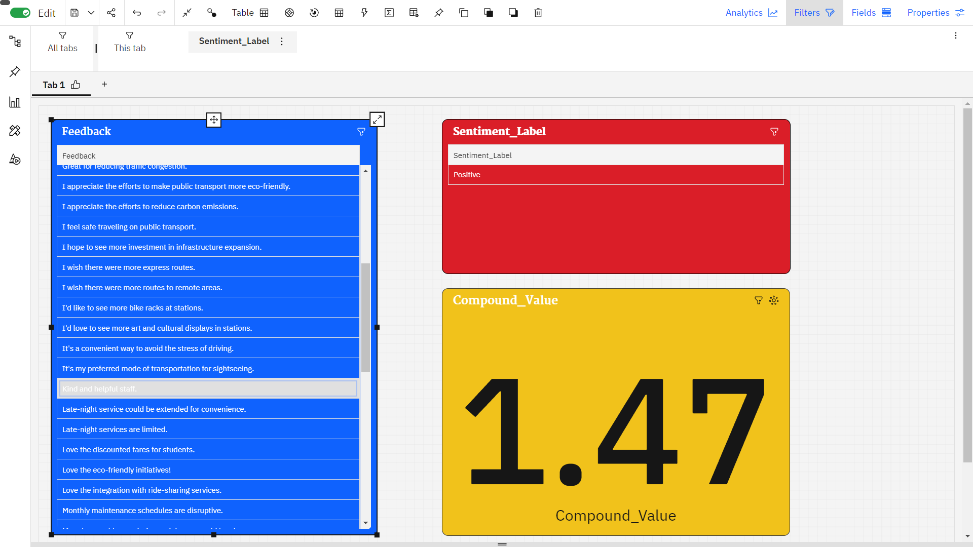
The Passenger Feedback Dashboard is designed to provide a comprehensive overview of customer sentiment and feedback for a service.

**Screenshots:**









In the Passenger Feedback Dashboard, feedback data collected from various sources is processed and analysed to provide valuable insights into passenger sentiment and satisfaction. Here's a brief explanation of key components:

**1. Feedback Data:** This is the raw data collected from passengers through surveys, social media, customer support interactions, app/website feedback, comment cards, online forums, and other channels. It includes comments, ratings, and other textual or numerical feedback provided by passengers.

**2. Sentiment Analysis**: Sentiment analysis is a natural language processing (NLP) technique used to classify the sentiment or emotional tone of the feedback data. It categorizes the feedback into three main sentiment categories:

**Positive:** Feedback with a positive tone, indicating satisfaction or happiness with the service.

**Negative:** Feedback with a negative tone, highlighting dissatisfaction, concerns, or issues.

**Neutral:** Feedback that doesn't express strong positive or negative sentiment, often considered as factual or objective statements.

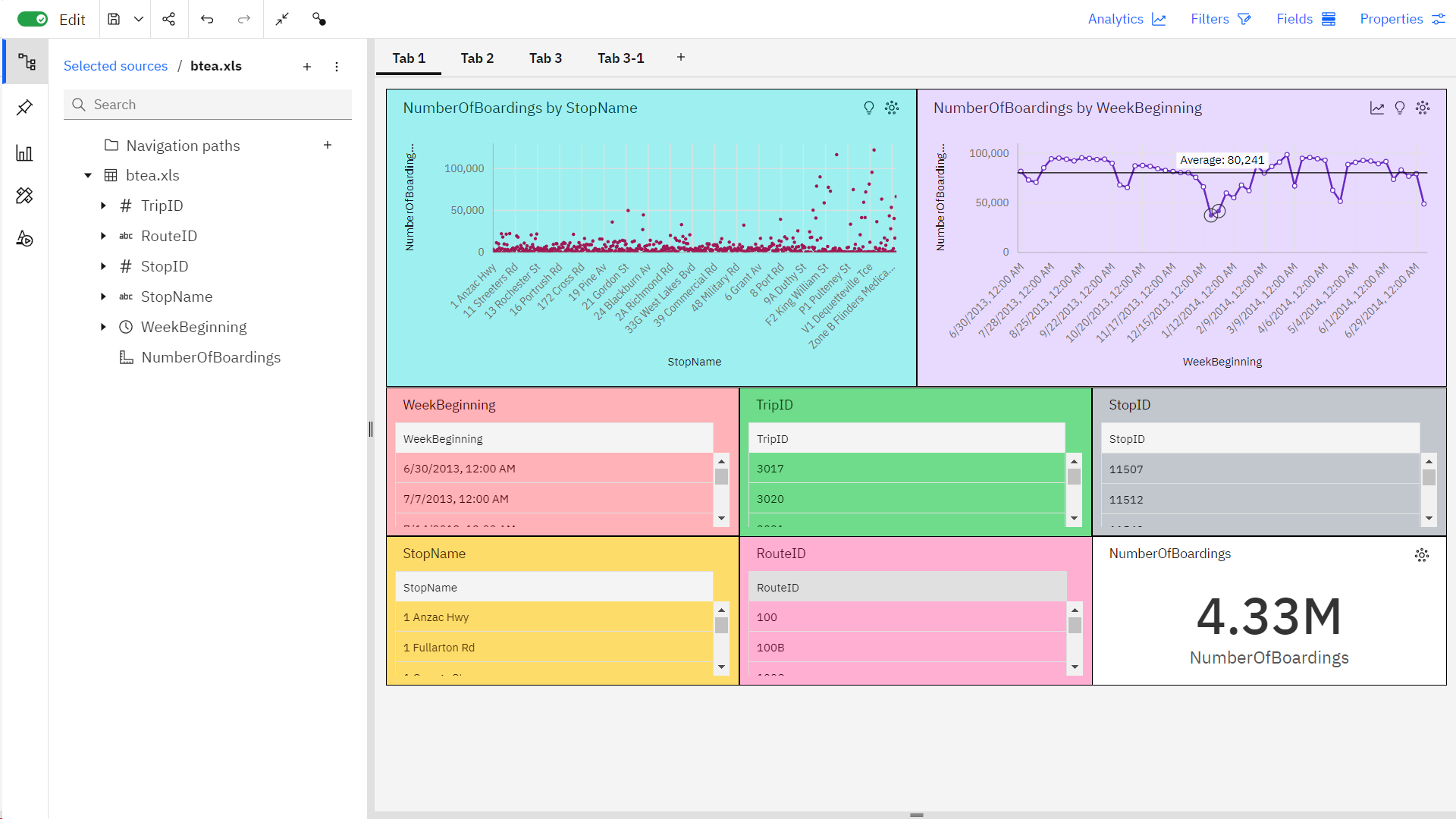
**3. Compound Value** :The compound value is a single numerical score that represents the overall sentiment or polarity of a piece of feedback. It's often calculated using sentiment analysis libraries and algorithms. The compound value takes into account the entire context of the text and assigns a score that ranges from -1 (most negative) to 1 (most positive). A score around 0 suggests a neutral sentiment.

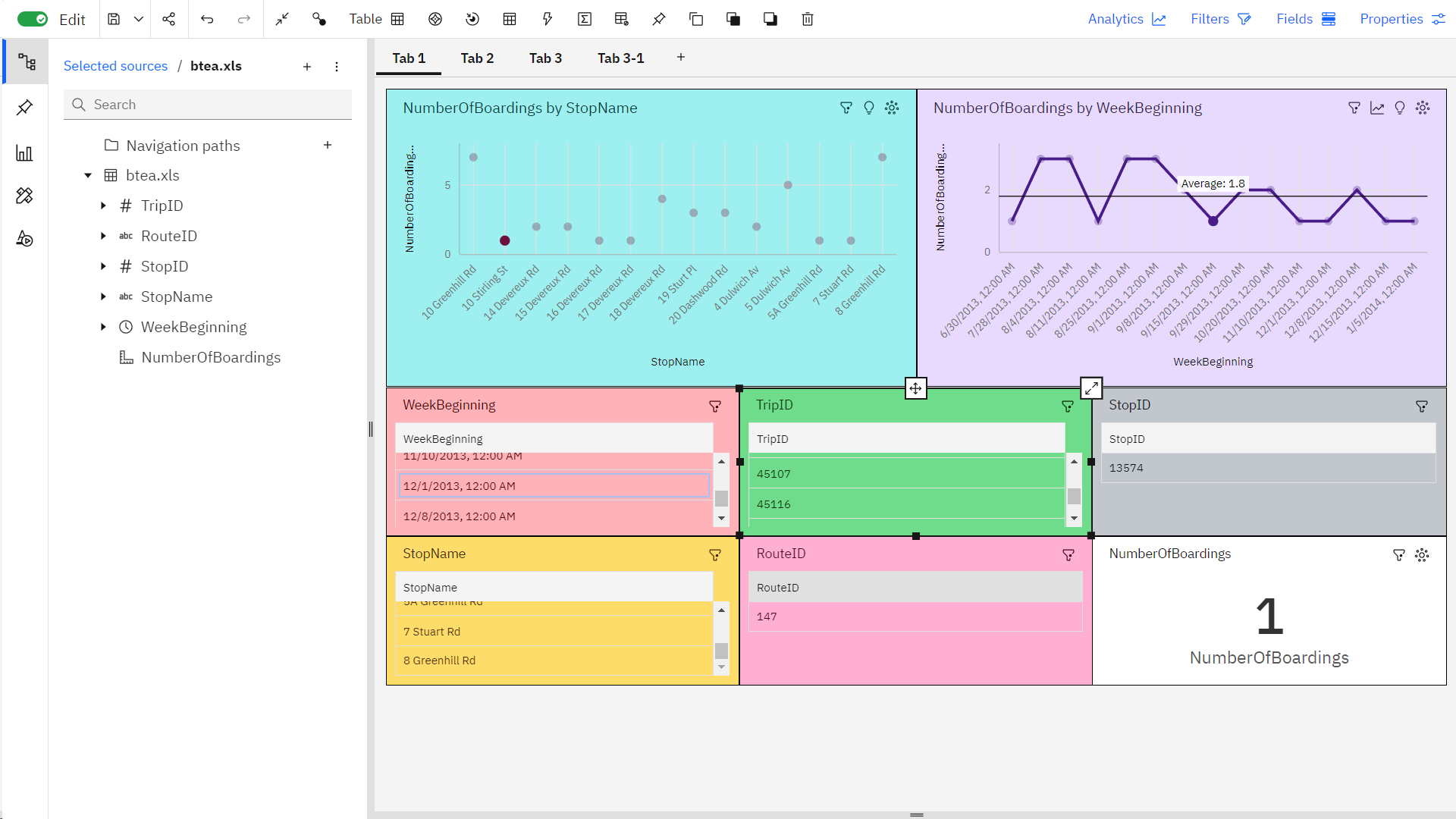
For instance, if a passenger provides feedback like, "The service was excellent, but the wait times were frustrating," sentiment analysis would classify "excellent" as positive, "frustrating" as negative, and the overall compound value would reflect the mixed sentiment of the feedback.

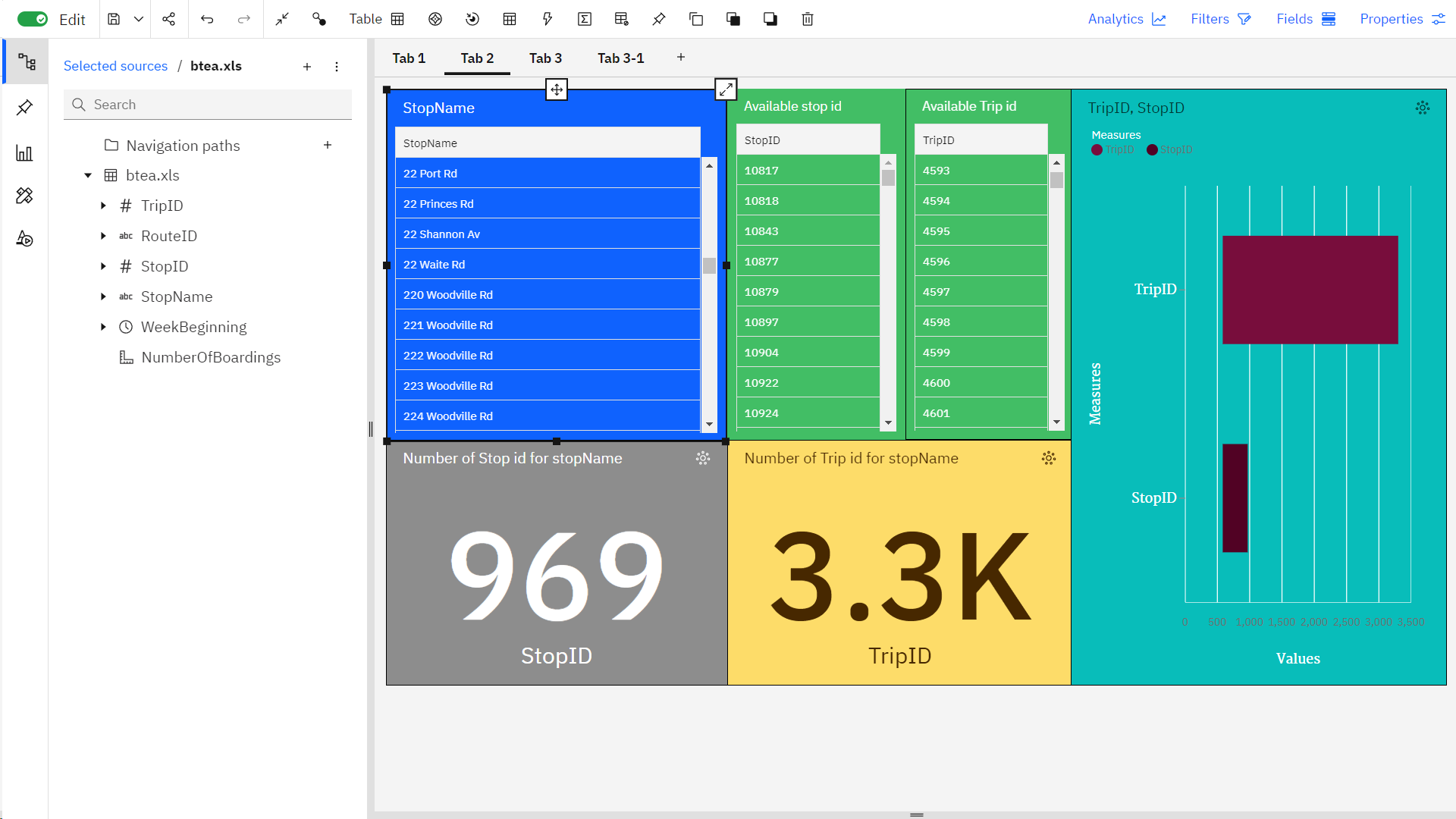
**Dashboard Link:** [**https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my\_folders%2Fdashboard%2Ffeedback&action=view&mode=dashboard**](https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2Fdashboard%2Ffeedback&action=view&mode=dashboard)

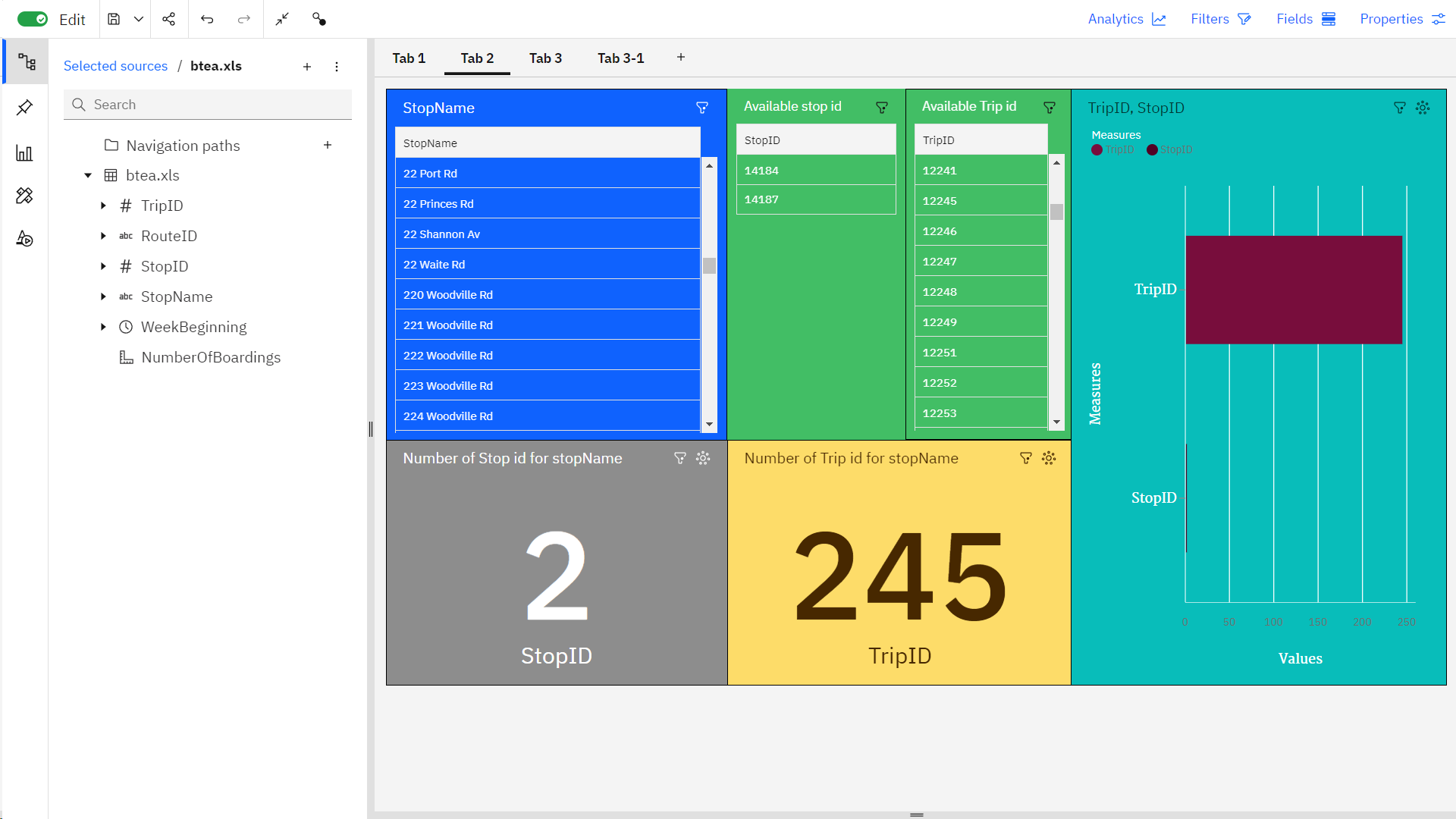
**Service Efficiency Metrics Dashboard**

The Service Efficiency Metrics Dashboard is designed to provide a comprehensive view of the efficiency and operational performance of a service, such as a transportation system.









**Dashboard Link:** [**https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my\_folders%2Fdashboard%2Fservice%2Befficiency&action=view&mode=dashboard**](https://us1.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2Fdashboard%2Fservice%2Befficiency&action=view&mode=dashboard)

**DATA ANALYSIS**

**3.1. Advanced Data Analysis with Python**

**# Importing data and preprocessing**

import pandas as pd

import numpy as np

import sklearn

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt # Add this import statement

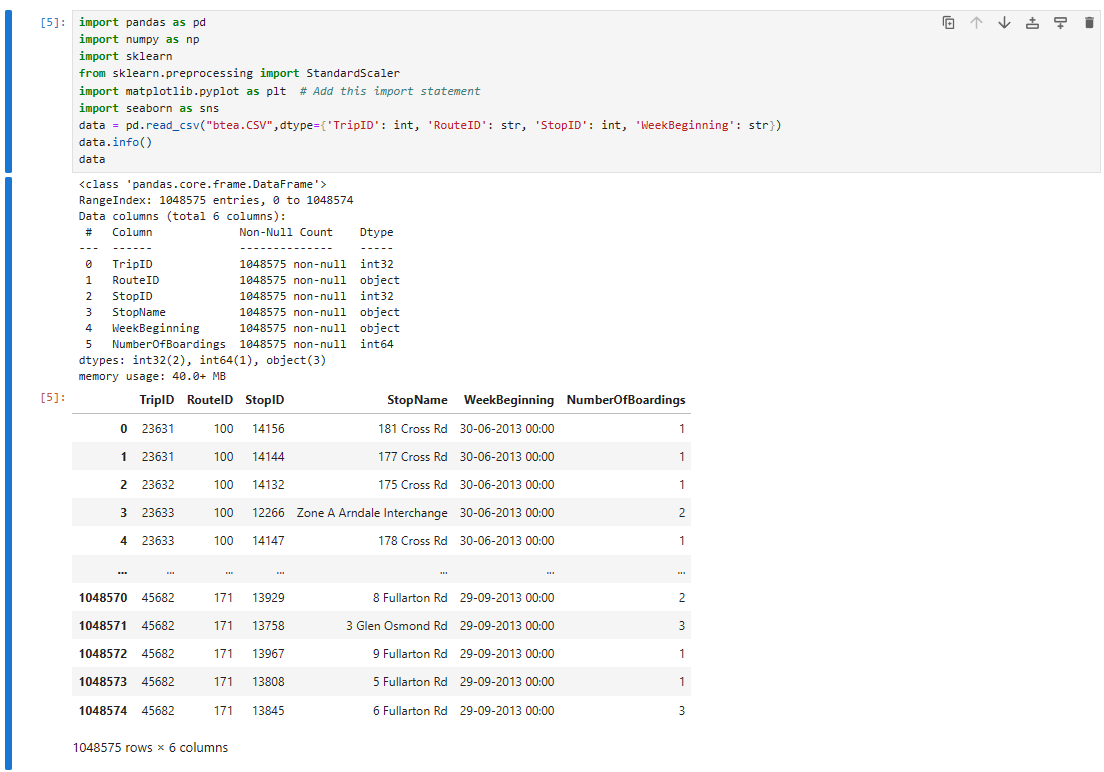
import seaborn as sns

data = pd.read\_csv("btea.CSV",dtype={'TripID': int, 'RouteID': str, 'StopID': int, 'WeekBeginning': str})

data.info()

data

**#output**



**#code**

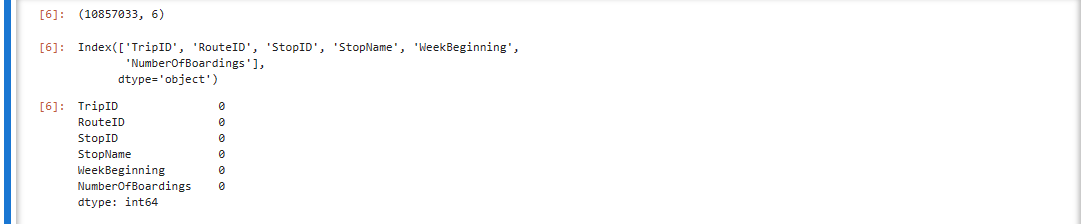
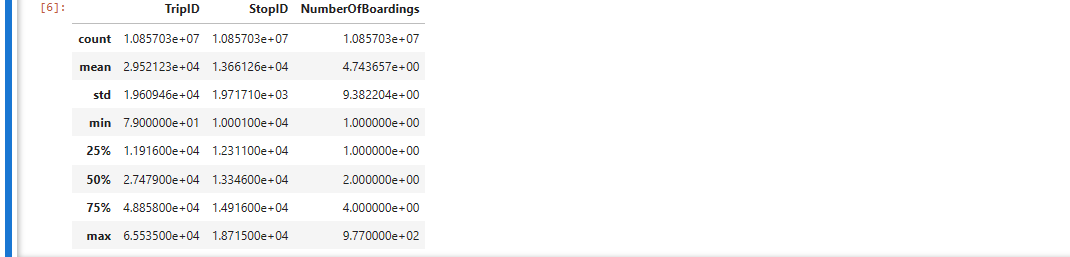
data.shape

data.columns

pd.isnull(data).sum()

data.describe()

**#output**

**#Data visualization**

plt.figure(figsize=(8, 6))

sns.histplot(data['TripID'], bins=20, kde=True, color='skyblue', label='TripID')

sns.histplot(data['StopID'], bins=20, kde=True, color='salmon', label='StopID')

plt.title('Distribution of TripID and StopID')

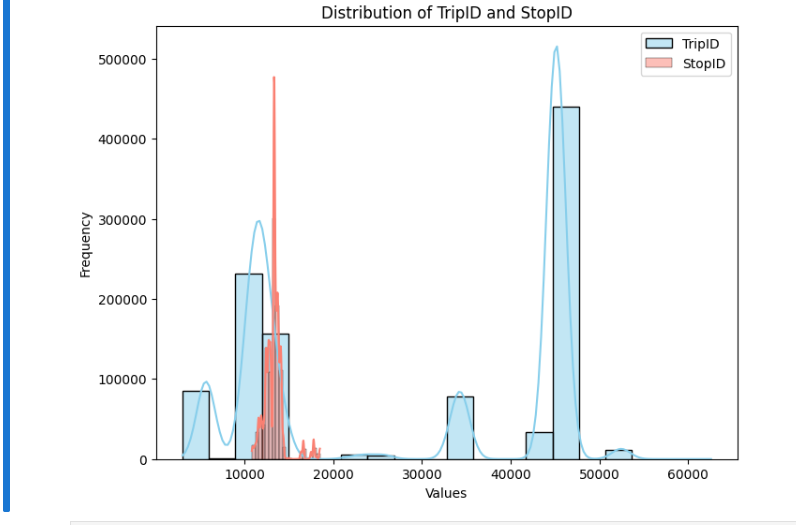
plt.xlabel('Values')

plt.ylabel('Frequency')

plt.legend()

plt.show()

**#output**



**#Correlation matrix**

numerical\_data = data.select\_dtypes(include=[np.number])

corr\_matrix = numerical\_data.corr()

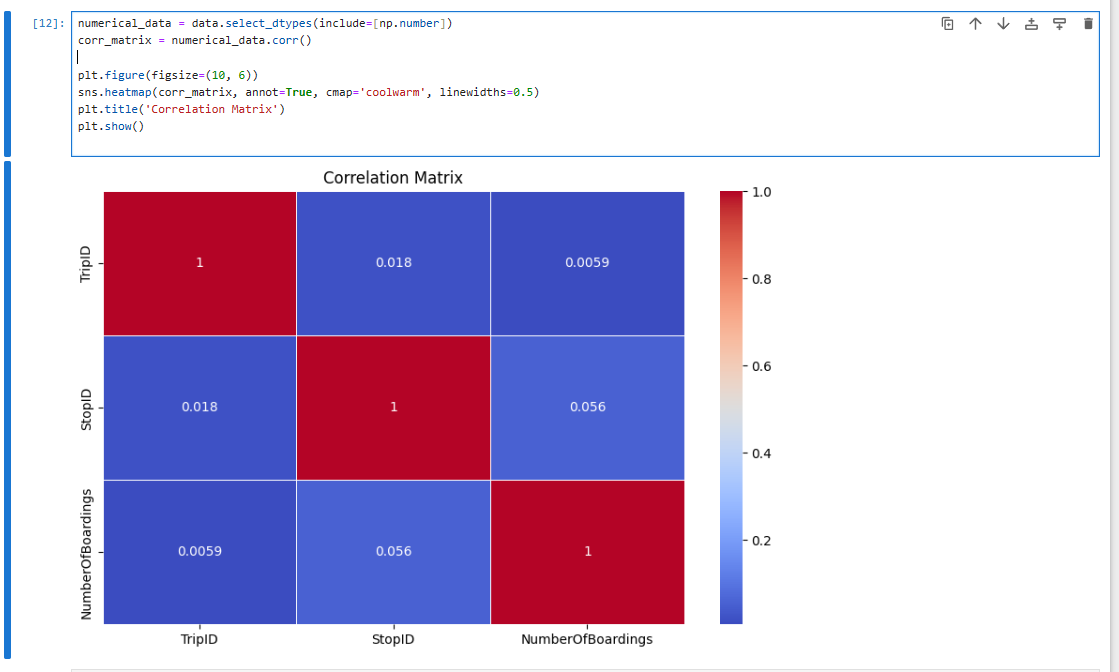
plt.figure(figsize=(10, 6))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Matrix')

plt.show()

**#output**



**3.2. Calculating Service Punctuality Rates**

**#Data processing and calculate punctuality Rate**

import pandas as pd

data = pd.read\_csv('arrival\_times.csv')

data['Scheduled\_Arrival'] = pd.to\_datetime(data['Scheduled\_Arrival'], format='%H:%M')

data['Actual\_Arrival'] = pd.to\_datetime(data['Actual\_Arrival'], format='%H:%M')

threshold = pd.Timedelta(minutes=20)

data['Punctual'] = (data['Actual\_Arrival'] - data['Scheduled\_Arrival']) <= threshold

data['Punctuality\_Rate'] = data['Punctual'].apply(lambda x: 'On Time' if x else 'Delayed')

data.to\_csv('punctuality.csv', index=False)

**#Calculating punctuality rate**

on\_time\_rate = (data['Punctual'].sum() / len(data)) \* 100

print(f'On-Time Performance Rate: {on\_time\_rate:.2f}%')

**#output**

****

**#Data Visualization**

import pandas as pd

import matplotlib.pyplot as plt

data = pd.read\_csv('punctuality.csv')

punctuality\_counts = data['Punctuality\_Rate'].value\_counts()

plt.figure(figsize=(2, 6))

plt.bar(punctuality\_counts.index, punctuality\_counts.values, color=['red', 'green'])

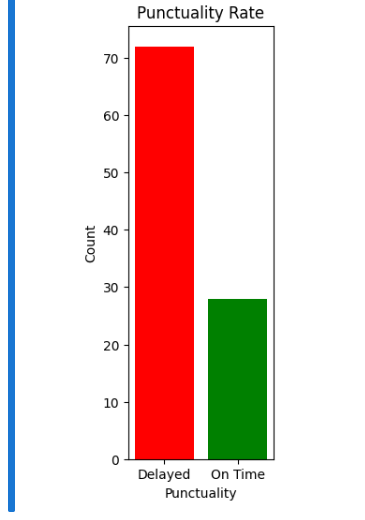
plt.title('Punctuality Rate')

plt.xlabel('Punctuality')

plt.ylabel('Count')

plt.show()

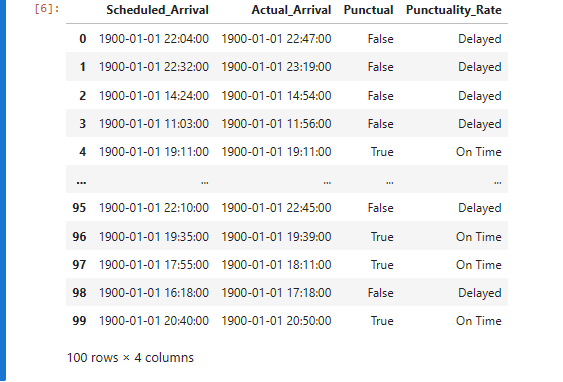
**#output**

****

data = pd.read\_csv('punctuality.csv')

data

**#output**

****

**Dataset:** [**Click here**](https://docs.google.com/spreadsheets/d/1G8g2e-KraZ7ZueVH4fqvpW8ceab617Nh/edit?usp=drive_link&ouid=109841781556534553902&rtpof=true&sd=true)

**Source code:** [**click here**](https://drive.google.com/file/d/1dRIDkI3_iMTeVWnotsMI1hOHm8NMZJmK/view?usp=drive_link)

**3.3. Sentiment Analysis on Passenger Feedback**

import pandas as pd

from nltk.sentiment.vader import SentimentIntensityAnalyzer

data = pd.read\_csv('feedback.csv')

sid = SentimentIntensityAnalyzer()

sentiments = []

for comment in data['Feedback']:

sentiment = sid.polarity\_scores(comment)

sentiments.append(sentiment)

sentiment\_labels = []

compound\_values = [] # To store the compound values

for sentiment in sentiments:

compound = sentiment['compound']

compound\_values.append(compound) # Append the compound value

if compound >= 0.05:

sentiment\_labels.append('Positive')

elif compound <= -0.05:

sentiment\_labels.append('Negative')

else:

sentiment\_labels.append('Neutral')

# Save the DataFrame to a new CSV file

data.to\_csv('feedback\_with\_sentiment.csv', index=False)

for i in range(len(data['Feedback'])):

print(f"Feedback: {data['Feedback'][i]}")

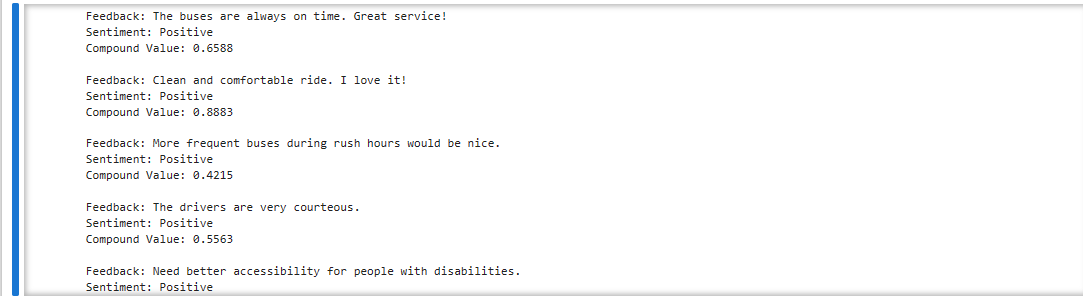
print(f"Sentiment: {sentiment\_labels[i]}")

print(f"Compound Value: {compound\_values[i]}\n")

data['Sentiment\_Label'] = sentiment\_labels

data['Compound\_Value'] = compound\_values

**#output**

****

**Dataset:** [**Click here**](https://docs.google.com/spreadsheets/d/1ndxgGwRdgksvIQuvvGUql96yaqnWxrY2/edit?usp=drive_link&ouid=109841781556534553902&rtpof=true&sd=true)

**Source code:** [**Click here**](https://drive.google.com/file/d/1k8denHJKKkGIv2nRj6M7bcq270SAcqSy/view?usp=drive_link)

**CONCLUSION:**

The public transportation efficiency analysis project aims to provide valuable insights into the performance of public transportation systems. By using IBM Cognos for visualization, we can create interactive and informative dashboards that assist in making informed decisions to enhance public transportation efficiency, ultimately contributing to more sustainable and accessible urban environments.