AeroBI Explorers

MIS 587 – Group 4

Team Members

Bharambe Neha

Giradkar Jay

Joshee Minita

Kshirsagar Tejaswini

Singh Kanika

Eller

Table of Contents

1. Overview of Client and Problem Statement

Background

The airline industry plays a critical role in global transportation, facilitating millions of journeys every day. However, flight delays and cancellations are common occurrences that can cause operational inefficiencies and inconvenience to passengers, disrupt travel plans, and incur significant costs for both airlines and travelers alike. Understanding the causes and patterns of flight delays and cancellations is crucial for airlines to improve operational efficiency, enhance customer satisfaction, and minimize financial losses.

Business Problem

Despite advancements in technology and operational procedures, flight delays and cancellations continue to plague the airline industry. Airlines face the challenge of accurately predicting and managing these disruptions to minimize their impact on passengers and maintain a competitive edge in the market. The business problem revolves around the need to develop effective strategies and solutions to reduce the frequency and severity of flight delays and cancellations, ultimately improving the overall travel experience for passengers while optimizing operational performance and cost-effectiveness for airlines.

Potential project objectives include:

1. Analyzing historical flight data to identify patterns and trends related to delays and cancellations.
2. Analyzing the correlation between various factors and occurrences of delay and cancellations.
3. Identifying sentiment in customer reviews to uncover key insights and improve customer satisfaction.

By addressing these objectives, the project aims to empower airlines with actionable insights and solutions to minimize the occurrence and impact of flight delays and cancellations, ultimately enhancing the overall efficiency and reliability of air travel operations.

2. Dataset Description

Dataset

Flight Delay & Cancellation:

- Flight delays and cancellations for January 2019 – August 2023 across the US.
- Dataset – 3M records, 30+ columns, picked from Kaggle.

Airline Reviews:z

- Airline experiences through Reviews
- Dataset– 23k+ records, 20 columns, picked from Kaggle.

Airports:

- Airport information across the US
- Dataset – 380 records, scraped through the web to gather information about all the airports.

Data Dictionary

ATTRIBUTE	DATA TYPE	DESCRIPTION
Actual Arrival	int	Time when the flight arrived
Actual Departure	int	Time when the flight departed
Actual Elapsed Time	int	Elapsed Time of Flight, in Minutes
Airline Name	varchar(100)	Name of the Airline
Airline SKey	int	Surrogate key for Airline dimension
Airport Code	varchar(50)	Code which is assigned to the Airport
Airport Name	varchar(50)	Name of the airport
Airport Skey	int	Surrogate key for Airport dimension
Arrival Delay	int	Difference between the actual and schedule arrival time
Cabin Staff Service	int	Parameter for the review
Calendar Month Name	nvarchar(20)	
Calendar Year	int	
Cancellation Code	varchar(50)	Code associated with the cancellation
Cancellation Reason	nvarchar(100)	Description of the reason for the cancellation
Cancellation Skey	int	Surrogate key for Cancellation dimension
Cancelled	int	If the flight was cancelled, then 1 else 0.
City	varchar(50)	Name of the city where airport is located
Date	date	Date
Date Key	int	Surrogate key for Date dimension
Day_of_Week	nvarchar(20)	
Delay_due_carrier	int	Carrier Delay, in Minutes
Delay_due_Late_Aircraft	int	Late Aircraft Delay, in Minutes
Delay_due_NAS	int	National Aviation System Delay, in Minutes
Delay_due_Security	int	Security Delay, in Minutes
Delay_due_weather	int	Weather delay, in minutes

Departure Delay	int	Difference between the actual and schedule departure time
Destination City	nvarchar(255)	City of destination of flight journey
Diverted	int	If the flight was diverted, then 1 else 0.
DOT Code	nvarchar(255)	Code associated with the tires of flight
Flight Number	nvarchar(255)	Unique number assigned to the flight
Food Beverages	int	Parameter for the review
Ground Service	int	Parameter for the review
IATA code	varchar(50)	Code assigned to the Airline
IATA Code	nvarchar(255)	Code assigned to the Airline
Inflight Entertainment	int	Parameter for the review
Origin	varchar(50)	Departure Airport code
Origin City	nvarchar(255)	City of origin of flight journey
Overall rating	int	Rating provided by the customer
Review Date	date	Date when the review was published
Review Description	ntext	Description of the review
Review ID	int	ID associated with the review
Review Skey	int	Surrogate key for Review dimension
Review Title	nvarchar(255)	Title for the review
Scheduled Arrival	int	Time when the flight was schedule for arrival
Scheduled Departure	int	Time when the flight was schedule for departure
Scheduled Elapsed Time	int	Scheduled Elapsed Time of Flight, in Minutes
Seat Comfort	int	Parameter for the review
Seat Type	nvarchar(255)	States the type of seat such as economy, business, etc.
State	varchar(50)	Name of the state where airport is located
Taxi In	int	Taxi In Time, in Minutes
Taxi Out	int	Taxi Out Time, in Minutes
Type of Traveler	nvarchar(255)	
ValidFrom	date	Date when the name of airport became effective
ValidTo	date	Date till the name of airport was effective
Value for money	int	Parameter for the review
Weekend	int	
Wheels Off	int	Wheels Off Time (local time: hhmm)
Wheels On	int	Wheels On Time (local time: hhmm)
Wifi Connectivity	int	Parameter for the review

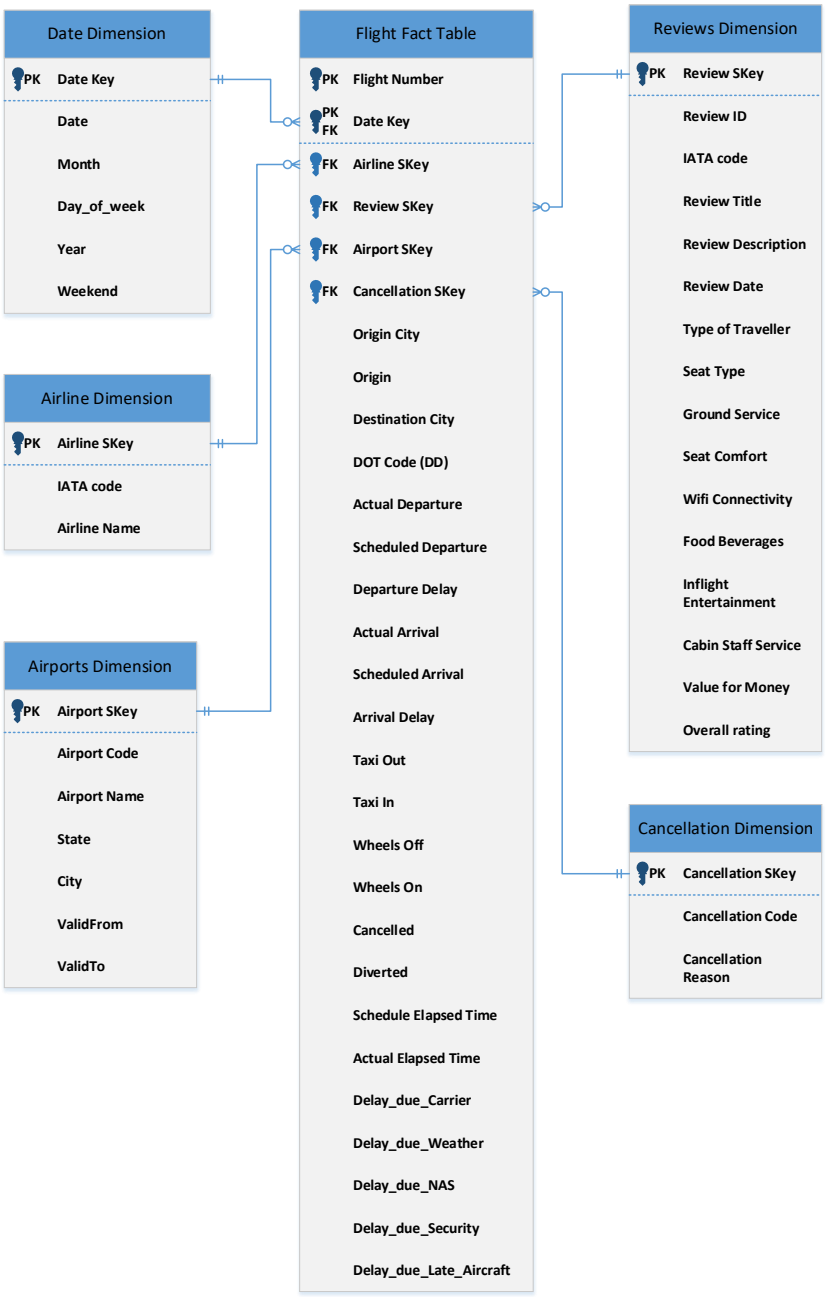
The business process it represents is flight operations and management. This involves scheduling flights, managing cancellations and diversions, analyzing delays, and providing services to travelers.

3. Data Warehouse Design & Implementation

Dimension Modeling Process

1. **Identifying the business process:**
   Flight operations and management encompass a multifaceted business process crucial for the smooth functioning of airlines. At its core, it involves meticulously scheduling flights to optimize resources and meet demand while considering factors like aircraft availability, crew schedules, and airport slots. It extends to managing unforeseen events such as cancellations and diversions due to weather, technical issues, or airspace restrictions. Moreover, analyzing delays is imperative for identifying patterns and implementing strategies to minimize disruptions in the future. Additionally, flight operations entail providing impeccable services to travelers, ensuring their safety, comfort, and satisfaction throughout their journey.
2. **Declaring the grain:**
   The granularity of the data captured in the fact table revolves around individual flights, encompassing detailed information such as flight number, origin and destination city, scheduled departure and arrival times, actual departure and arrival times, Schedule and Actual Elapsed Time, and any relevant operational metrics.
3. **Identifying dimensions:**
   The dimensions for the business process are Airline, Airports, Date, Cancellation, and review dimensions. The Airline dimension holds information of IATA codes and airline name. The Airports dimension provides insights into the various airports involved in the process, including their name, state, and city. Date dimension facilitates the analysis of temporal trends, seasonality, and day-of-week effects on flight operations. The Cancellation dimension offers insights into the reasons behind flight cancellations, such as weather events, technical issues, or operational constraints. Lastly, the Review dimension encompasses airline, customer feedback, and satisfaction ratings, enabling airlines to gauge customer sentiment.
4. **Identifying facts:**
   In the fact table for flight operations and management, key facts encompass various elements crucial for understanding and analyzing flight performance. These include flight travel time metrics such as scheduled departure time and actual departure time. Additionally, elapsed time, representing the duration of the flight from takeoff to landing, serves as a fundamental indicator of flight efficiency and operational performance. By capturing these essential facts in the fact table, airlines can assess flight timeliness, identify trends in departure and arrival times, and optimize scheduling and resource allocation to enhance overall operational efficiency and customer satisfaction.

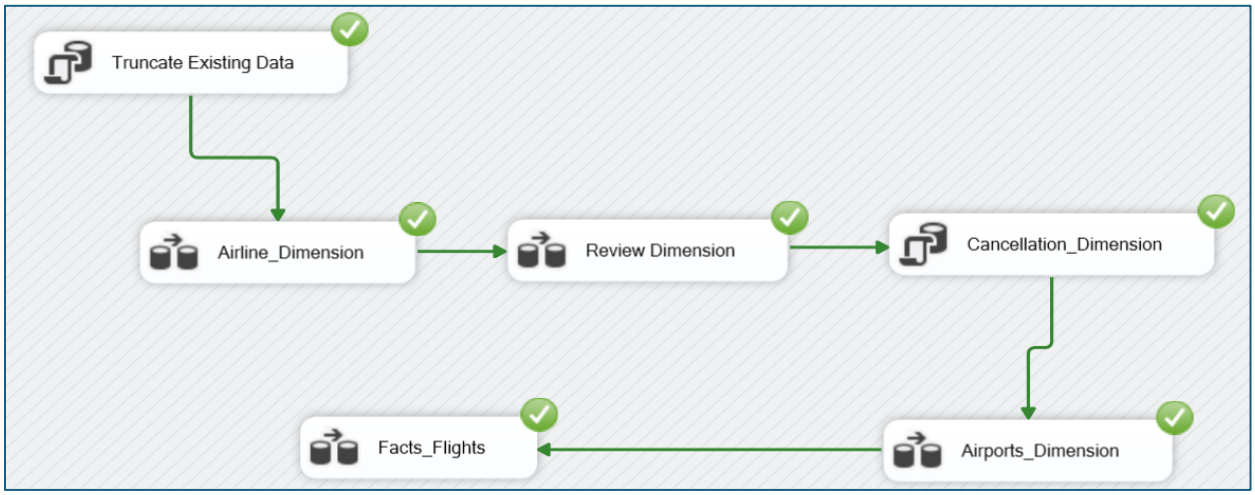
Star Schema

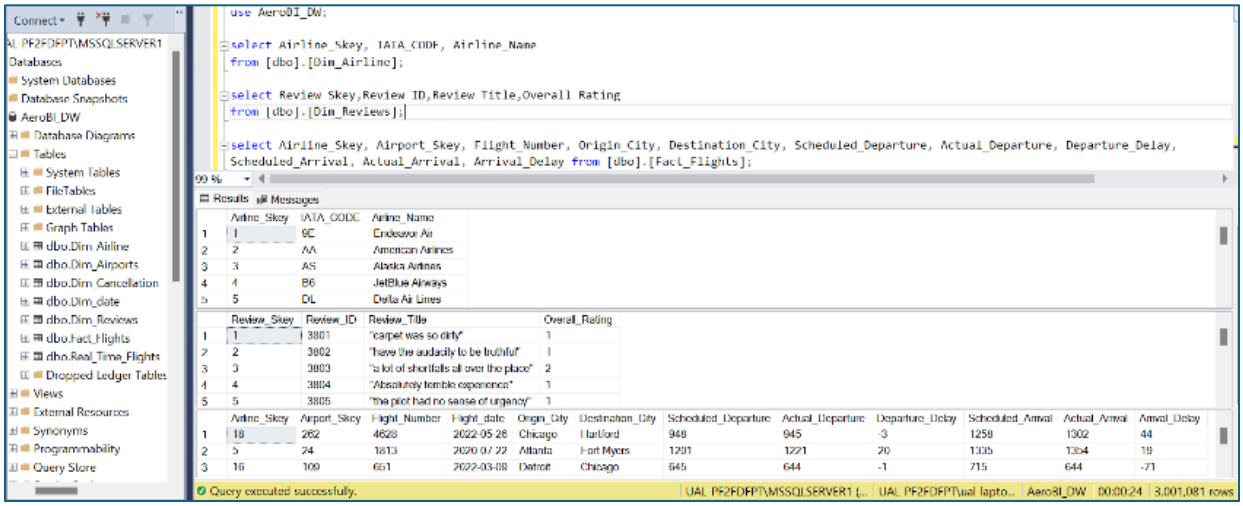


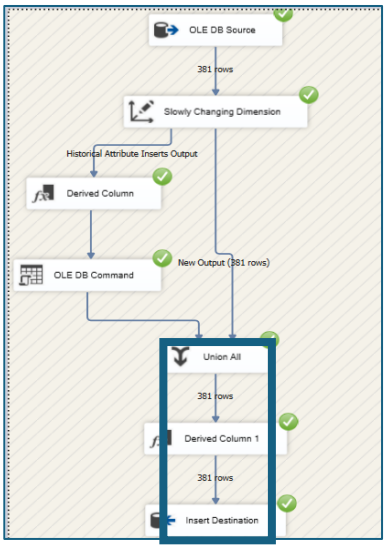
The Star Schema has 5 dimensions with 1 fact table. The fact table has measures like Actual and Scheduled Arrival/Departure, Cancelled etc. Surrogate keys are used as primary key in the dimension tables and foreign in the fact table. The primary (candidate) key for the fact table was the combination of Flight Number and Flight Date. This helps the data warehouse to be independent of operational changes in the source database. It also has a degenerate dimension – DOT_Code which contains the serial number of the tyre in an aircraft. There are 2 role-playing dimensions – City (as Origin and Destination) and Date (as Review and Flight date). Reviews are mapped as a 1-M relationship to the fact table.

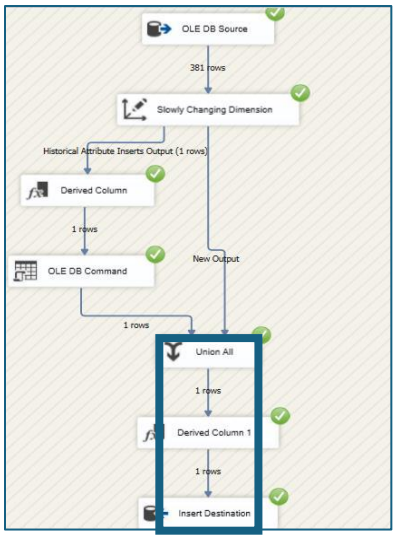
ETL Implementation

The data warehouse was implemented in SQL server via ETL using SSIS. The sources used were in the format excel and csv files which were used as flat file sources in the ETL. The ETL used 1 SQL task which truncated all destination tables. Data flows were used for each dimension table. The fact table was populated using all dimension tables where components like LookUp and Derived columns were used. One of the key components of our data warehouse was implementing Slowly Changing Dimensions (Type 2) for Airports dimension. This was done using the SCD component in SSIS. Airport Name was chosen as Historical attribute and the business key was Airport Code. Effective date management helped to identify the active record. As you can see below, a new Skey is generated for maintaining the change showing that it's independent of the changes in the source.










```
use AeroBI_DW;

SELECT Airport_Code, Airport_Name, City, State
FROM [AeroBI_DW].[dbo].[Airports]
WHERE Airport_Code = 'SDF';

SELECT [Airport_Code], [Airport_Name], [City], [State], [Airport_Skey], [ValidFrom], [ValidTo]
FROM [AeroBI_DW].[dbo].[Dim_Airports]
WHERE Airport_Code = 'SDF';
```

08 %

Results | Messages

	Airport_Code	Airport_Name	City	State
1	SDF	Standiford	Louisville	KY

	Airport_Code	Airport_Name	City	State	Airport_Skey	ValidFrom	ValidTo
1	SDF	Standiford	Louisville	KY	323	2024-04-23 19:50:13.000	9999-12-31 00:00:00.000

```
SELECT Airport_Code,Airport_Name,City,State
FROM [AeroBI_DW].[dbo].[Airports]
WHERE Airport_Code = 'SDF';

SELECT [Airport_Code], [Airport_Name], [City], [State], [Airport_Skey], [ValidFrom], [ValidTo]
FROM [AeroBI_DW].[dbo].[Dim_Airports]
WHERE Airport_Code = 'SDF';
```

108 %

Results | Messages

	Airport_Code	Airport_Name	City	State
1	SDF	Louisville Muhammad Ali International Airport	Louisville	KY

	Airport_Code	Airport_Name	City	State	Airport_Skey	ValidFrom	ValidTo
1	SDF	Standiford	Louisville	KY	323	2024-04-23 19:50:13.000	2024-04-23 19:57:28.000
2	SDF	Louisville Muhammad Ali International Airport	Louisville	KY	382	2024-04-23 19:57:28.000	9999-12-31 00:00:00.000

4. Data Preparation

Splitting City and State

In this section, the str.split() method is utilized to split the 'ORIGIN_CITY' column into separate 'ORIGIN_CITY' and 'ORIGIN_STATE' columns

```
data[['ORIGIN_CITY', 'ORIGIN_STATE']] = data['ORIGIN_CITY'].str.split(',', expand = True)
data.head(3)
```

Dropping Duplicates

This also covers the application of the **drop_duplicates()** method to eliminate duplicate rows based on specific columns, namely 'IATA_CODE' and 'AIRLINE'. Removing duplicates ensures data consistency and prevents skewness in subsequent analyses.

```
df_dim_airlines_derived = df_fact_flights[['IATA_CODE', 'AIRLINE']].drop_duplicates()
```

Cleaning Airline Names

In this section, unnecessary suffixes such as 'Inc.' and 'Co.', along with extra white spaces are addressed using the str.replace() and str.strip() methods. Additionally, the value_counts() method is employed to inspect unique airline names in the dataset.

```
data1['AIRLINE'] = data1['AIRLINE'].str.replace('Inc.', '')
data1['AIRLINE'] = data1['AIRLINE'].str.replace('Co.', '')
data1['AIRLINE'] = data1['AIRLINE'].str.strip()
data1['AIRLINE'].value_counts()
```

Date Manipulation in Flight Data

The final technique covers the manipulation of date values within flight data. The 'FL_DATE' column, initially containing **string** dates, is **converted** to **datetime** format using **pd.to_datetime().** Furthermore, a new 'FL_YEAR' column is created by extracting the year component from the 'FL_DATE' column, enabling temporal analysis based on yearly trends.

```
# Convert 'FL_DATE' column to datetime format
df_fact_flights['FL_DATE'] = pd.to_datetime(df_fact_flights['FL_DATE'])

# Extract year from 'FL_DATE' column and add it as a new column 'FL_YEAR'
df_fact_flights['FL_YEAR'] = df_fact_flights['FL_DATE'].dt.year
```

5. Data Exploration

Handling Missing Data

Transitioning to the handling of missing data, this section focuses on exploratory data analysis of flight records. Addressing missing values is crucial for maintaining data integrity and ensuring the reliability of subsequent analyses.

Missing Values before Processing

It starts by showing missing value counts for 'ARR_DELAY' and 'DEP_DELAY' before processing, setting the stage for choosing suitable imputation or deletion methods.

```
print("Missing values before processing:")
print(df[['ARR_DELAY', 'DEP_DELAY']].isnull().sum())
```

Imputation and Deletion Strategies

Two primary strategies are outlined for handling missing values:

1. **Imputation**: Missing values in the 'ARR_DELAY' column are replaced with 0 using the fillna() method, preserving data continuity.

```
df['ARR_DELAY'].fillna(0, inplace=True)
```

2. **Deletion**: Rows with missing values in the 'DEP_DELAY' column are removed using the dropna() method, ensuring data quality by eliminating incomplete observations.

```
df.dropna(subset=['DEP_DELAY'], inplace=True)

print("Missing values after processing:")
print(df[['ARR_DELAY', 'DEP_DELAY']].isnull().sum())
```

Missing Values after Processing

Following the implementation of imputation and deletion strategies, the chapter revisits the counts of missing values for 'ARR_DELAY' and 'DEP_DELAY' columns. This comparison highlights the effectiveness of the chosen techniques in mitigating missing data issues.

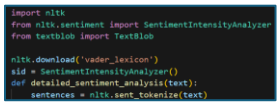
Impact and Significance

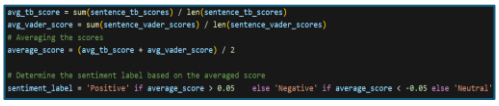
The significance of accurate missing data handling is highlighted, emphasizing its pivotal role in facilitating robust flight data analysis. Ensuring data integrity through meticulous preprocessing lays the foundation for meaningful insights and informed decision-making.

Sentiment Analysis of Customer Reviews

The primary goal of this project is to perform sentiment analysis on customer reviews to gauge sentiments such as satisfaction or disappointment expressed by customers. The project utilizes two Python libraries for text processing and sentiment analysis.

- **TextBlob**: Provides APIs for NLP tasks like POS tagging and sentiment analysis.
- **NLTK (Natural Language Toolkit):** Supports comprehensive NLP functionalities, including text preprocessing, tokenization, and sentiment analysis.





The VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon, integral to the NLTK library, is tailored for sentiment analysis of social media texts, offering a 'compound' score that indicates overall sentiment polarity.

Implementation Strategy

The project combines TextBlob and NLTK to process and analyze customer reviews deeply. Here's how the provided code snippet facilitates this:

1. **Function Definition and Tokenization**
   - A function, detailed_sentiment_analysis, processes text input.
   - Text is tokenized into sentences using nltk.sent_tokenize for detailed sentiment examination.

2. **Sentiment Scoring per Sentence**
   - Two lists, sentence_tb_scores and sentence_vader_scores, store scores from TextBlob and VADER.
   - TextBlob calculates sentiment polarity per sentence.
   - VADER provides a compound sentiment score for each sentence.

3. **Calculating Average Sentiment Scores**
   - Average scores from TextBlob and VADER are computed separately, reflecting linguistic and contextual sentiment insights.

4. **Combining and Finalizing Sentiment Score**
   - An overall average sentiment score is derived from both TextBlob and VADER scores.
   - Sentiments are labeled as 'Positive', 'Neutral', or 'Negative' based on specific thresholds, refining the sentiment analysis accuracy.

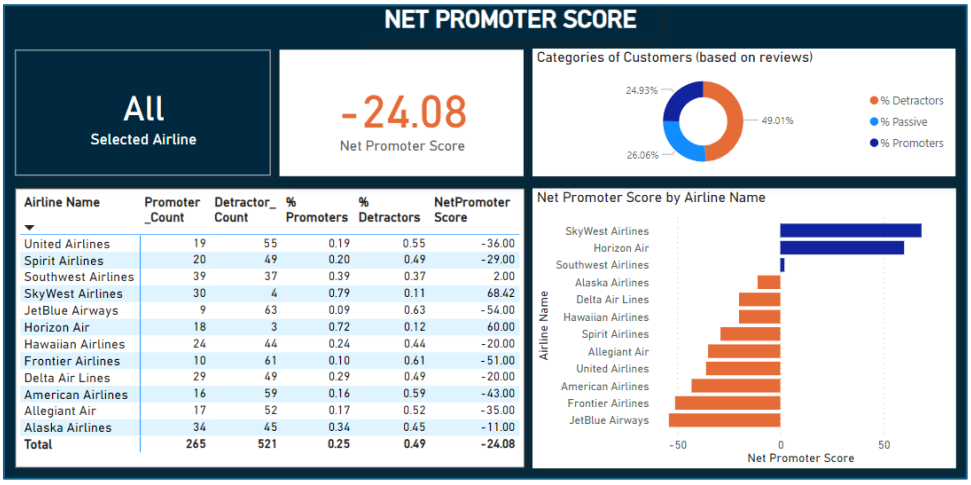
6. Data Analysis and Results

Data Visualization

Data visualization refers to the graphical representation of data and information. It involves creating visual elements such as charts, graphs, maps, and dashboards to communicate insights from datasets effectively. Data visualization plays a crucial role in data analysis and decision-making processes across various fields and industries, Overall, data visualization enhances understanding, enables better decision-making, and promotes data-driven insights across organizations.

Visualizations used in the project are as follows:

1. **Net Promoter Score (Overall)**



Airline Name	Promoter_Count	Detractor_Count	% Promoters	% Detractors	NetPromoter Score
United Airlines	19	55	0.19	0.55	-36.00
Spirit Airlines	20	49	0.20	0.49	-29.00
Southwest Airlines	39	37	0.39	0.37	2.00
SkyWest Airlines	30	4	0.79	0.11	68.42
JetBlue Airways	9	63	0.09	0.63	-54.00
Horizon Air	18	3	0.72	0.12	60.00
Hawaiian Airlines	24	44	0.24	0.44	-20.00
Frontier Airlines	10	61	0.10	0.61	-51.00
Delta Air Lines	29	49	0.29	0.49	-20.00
American Airlines	16	59	0.16	0.59	-43.00
Allegiant Air	17	52	0.17	0.52	-35.00
Alaska Airlines	34	45	0.34	0.45	-11.00
Total	**265**	**521**	**0.25**	**0.49**	**-24.08**

This dashboard consists of various visualizations related to understanding the Net Promoter Score for each airline. First, the card component on the top left displays the name of the selected airline for which the NPS is being calculated.
To calculate the NPS score, DAX queries are written to find the Promoters, Detractors and Passives count from the sentiment analysis label.

```
Promoter_Count =
CALCULATE(
    COUNTROWS('sentiment_analysis_Airline_review'),
    'sentiment_analysis_Airline_review'[sentiment_label] = "Positive"
)
```

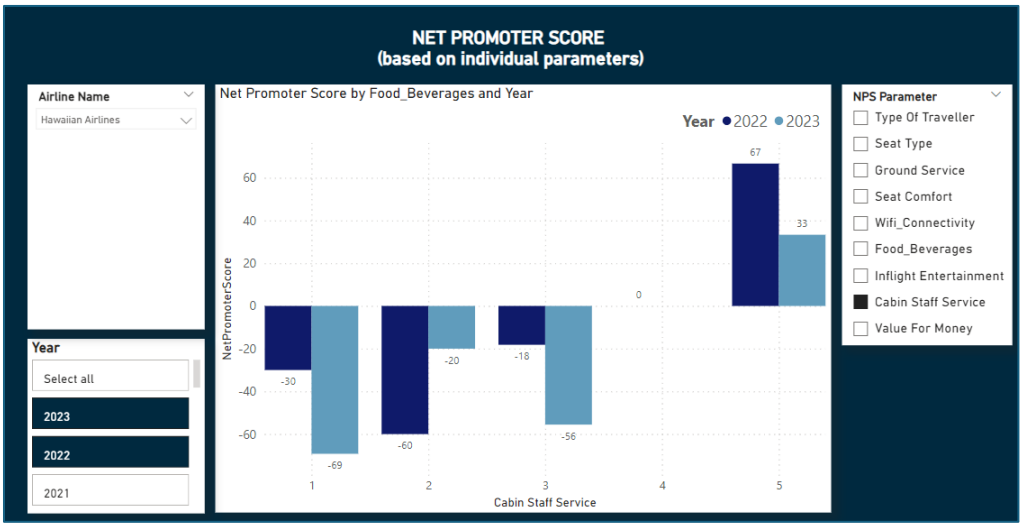
```
Passive_Count =
CALCULATE(
    COUNTROWS('sentiment_analysis_Airline_review'),
    'sentiment_analysis_Airline_review'[sentiment_label] = "Neutral"
)
```

```
Detractor_Count =
CALCULATE(
    COUNTROWS('sentiment_analysis_Airline_review'),
    'sentiment_analysis_Airline_review'[sentiment_label] = "Negative"
)
```

```
NetPromoterScore = ([% Promoters]-[% Detractors])*100
```

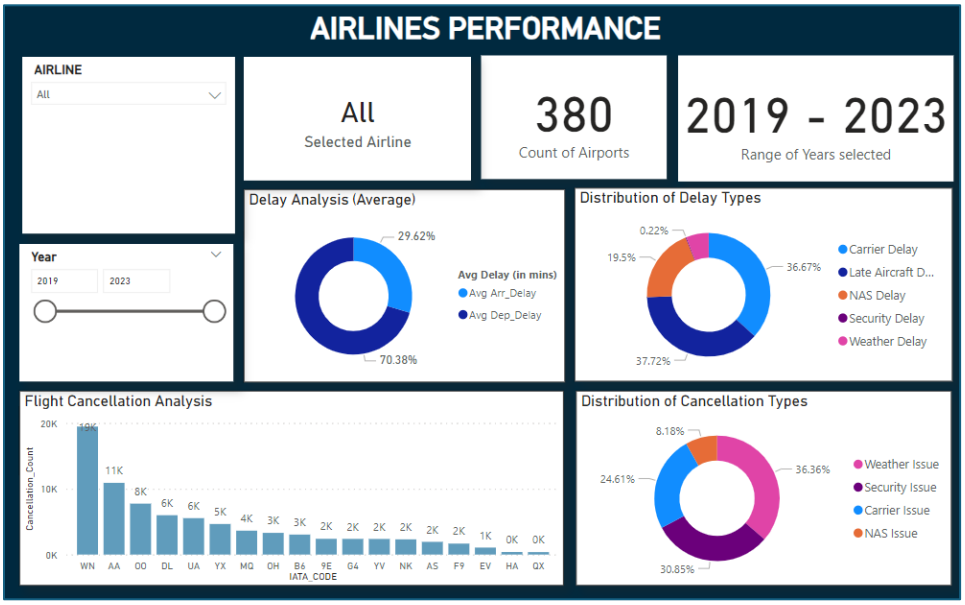
The above DAX queries allow us to evaluate the number of customers who can make (Promoters) or break (Detractors) brand image of the airlines.
This count of Promoters, Detractors and Passives is shown in a tabular format alongside the airline name. The table also acts as a slicer. The remaining visualizations depict information for the airline which is selected in the table. The dashboard also consists of a pie chart which shows the % Promoters, % Detractors and % Passives for individual airlines. The horizontal bar graph shows a comparative analysis of NPS for each airline.

2. **Net Promoter Score (by individual parameters)**



This dashboard represents NPS based on various parameters like seat type, ground service, wifi connectivity, food beverages, inflight entertainment, cabin staff service, value for money, etc. Left side has two slicers one for date dimension (year) and another for airlines. Selecting a particular airline and a year will give the NPS for that airline in the form of a bar chart in the middle of the dashboard. This will change based on which NPS parameter is selected from the list present on the right-hand side of the dashboard.
This dashboard can help the airlines to improve their performance in certain areas in order to provide better customer experience which will in turn help boost their business.

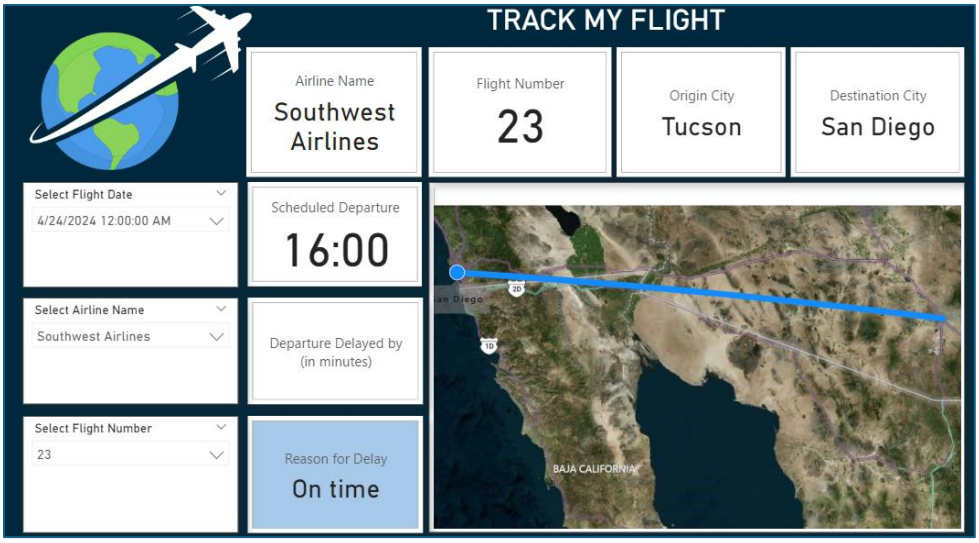
3. Airline Performance dashboard

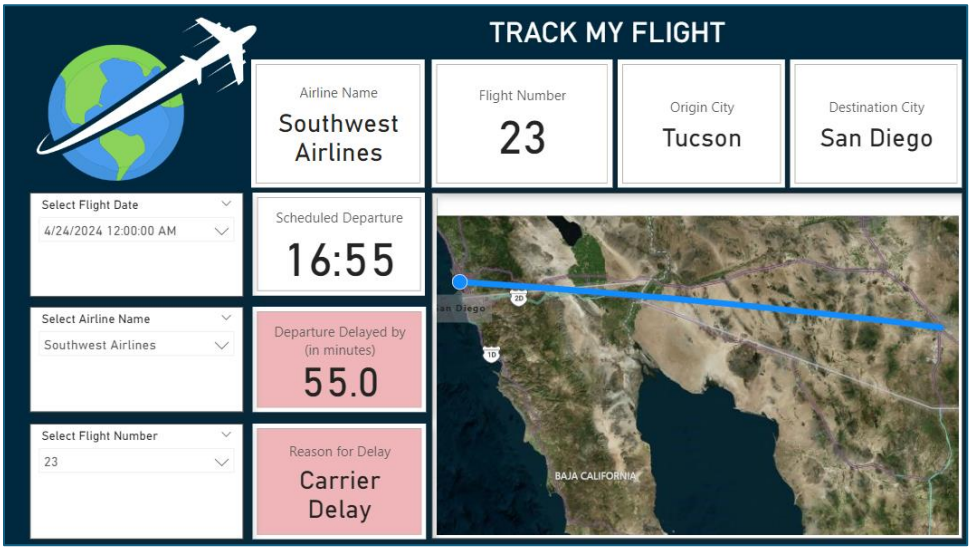


The above dashboard represents the airline performance based on delays and cancellations of the flights. The dashboard has two slicers namely airline and year. The three cards at the top displays the name of the selected airline, number of airports at which the airline is active, and the range of years selected. Based on the selections, the remaining visualizations change according to the data. The pie chart in the second row (middle) depicts the average departure delay and average arrival delay in minutes. The second donut chart in second row depicts the distribution of delay categories. There are majorly 5 delay reasons: Carrier delay, Late Aircraft delay, NAS delay, Security Delay, Weather Delay.
At the bottom, the bar chart represents the flight cancellation count per airline. Adjacent to that, the donut chart displays the division of cancellation categories. There are total of 4 cancellation reasons: Weather issue, Security issue, Carrier issue, NAS issue.

4. Real-time dashboard:

This is a real-time dashboard, which can help the customers to get real-time updates about the status of the flight. The dashboard contains the map which shows the route of the flight. Other cards include the scheduled departure, departure delay and reason for delay. The dashboard updates automatically when the changes are pushed by the ATC.





In all, looking at the visualizations, following is the analysis on the data,
For NPS (overall), SkyWest Airlines has the highest Net Promoter Score of 68.42 whereas JetBlue Airways has the least NPS of -54.00.

Similarly, the airline which has the maximum number of cancellations was from Southwest airlines. We got this insight from the airline performance dashboard. The major reason for the cancellation was Security issue.
Using this analysis, airlines can improve in the areas which are contributing to the cancellation and delay issues.

7. Business Implications

Net Promoter Score

NPS is a metric used in customer relationship management to gauge customer satisfaction and loyalty. A high NPS score can translate into increased customer loyalty and repeat business. Conversely, a low NPS score can indicate that customers are unhappy with the airline's service and are unlikely to recommend it to others. This can lead to an increase or decrease in revenue for the airline. This information can be used by:

- Airline Management: The visualization can help airlines identify areas where they need to improve their customer service such as on-time performance, baggage handling, or the customer experience at the airport.
- Marketing Teams: They can use it to target marketing campaigns to specific customer segments and tailor messaging accordingly.
- Investors: NPS can be a factor when evaluating an airline's financial status and prospects.
- Industry Analysts: They can use this data to compare airline performance and identify trends in the industry.
- Travelers: Travelers can use customers reviews and NPS score before booking a flight.

Airline Analysis

This dashboard provides analysis of cancellations and delays and highlights various factors that led to these cancellations or delays. This dashboard can be used by various personnel in the airline industry such as:

- Airline managers: This information can be used to identify areas for improvement and to make decisions about how to improve the airline's operations.
- Operations staff: They can use the dashboard to track the status of flights in real time. This can be used to monitor potential problems and to take steps to avoid them.
- Customer service staff: This information can be used to help them answer questions from passengers and to provide them with updates on the status of their flights.

Real Time Dashboard

Real time Analysis: Flight tracker provides real-time information about the status of flights, including departure and arrival times, and delays including their reason. This information can be helpful for:

- Passengers: Passengers can use flight trackers to track the status of their flights, to check in for flights, and to book flights.
- Airlines: Airlines can use flight trackers to track the location of their aircraft and crews, to monitor air traffic, and to communicate with passengers.
- Travel agents: Travel agents can use flight trackers to search for flights and to book flights for their clients.
- Freight forwarders: They can use flight trackers to track the status of their shipments.