Task 6.1

Balanced flight cancellation and delay - 2019-2023

Summary of Data Source:

The dataset is sourced from the U.S. Department of Transportation's Bureau of Transportation Statistics, covering Airline Flight Delay and Cancellation Data from August 2019 to August 2023. The dataset is accessible on Kaggle and comprises two CSV files: 'airport_location.csv' and 'preprocessed.csv'.

Key Variables:

The dataset encompasses crucial variables, including flight routes (origin, destination), time ranges for events (minutes, local time), and limited information on delay and cancellation reasons/attributions. There are 35 columns in total, containing details such as airport code, latitude, longitude, airline information, flight numbers, departure and arrival times, delays, cancellations, and more.

Collection Methodology:

The data collection method is detailed as "COLLECTION METHODOLOGY," and the true source is provided as https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ. This source is the official platform of the U.S. Department of Transportation, ensuring the reliability and authenticity of the data.

Reasons for Choosing the Dataset:

The primary goal of selecting this dataset is to conduct a comprehensive analysis of Flight Delay and Cancellation Metrics. Specific objectives include:

1. **Departure Delays, Arrival Delays, and Cancellation Rates:**

   Analyze patterns and trends in departure and arrival delays.

   Investigate the frequency and reasons behind flight cancellations.

2. **Airline Performance:**

   Evaluate the on-time performance of different airlines.

   Examine average delay times and cancellation rates for each airline.

3. **Distance and Route Analysis:**

   Explore the relationship between the distance of flights and delays.

   Analyze delays on specific routes or between particular cities.

4. **Time Analysis:**

   Study the impact of departure and arrival times on delays.

   Identify peak periods of delay occurrences.

5. **Airport Analysis:**

   Utilize the 'airport_location.csv' file to analyze delays and cancellations at various airports.

6. **Reasons for Delay:**

   Investigate the causes of delays, including aircraft issues, weather conditions, and other contributing factors.

Understanding Data:

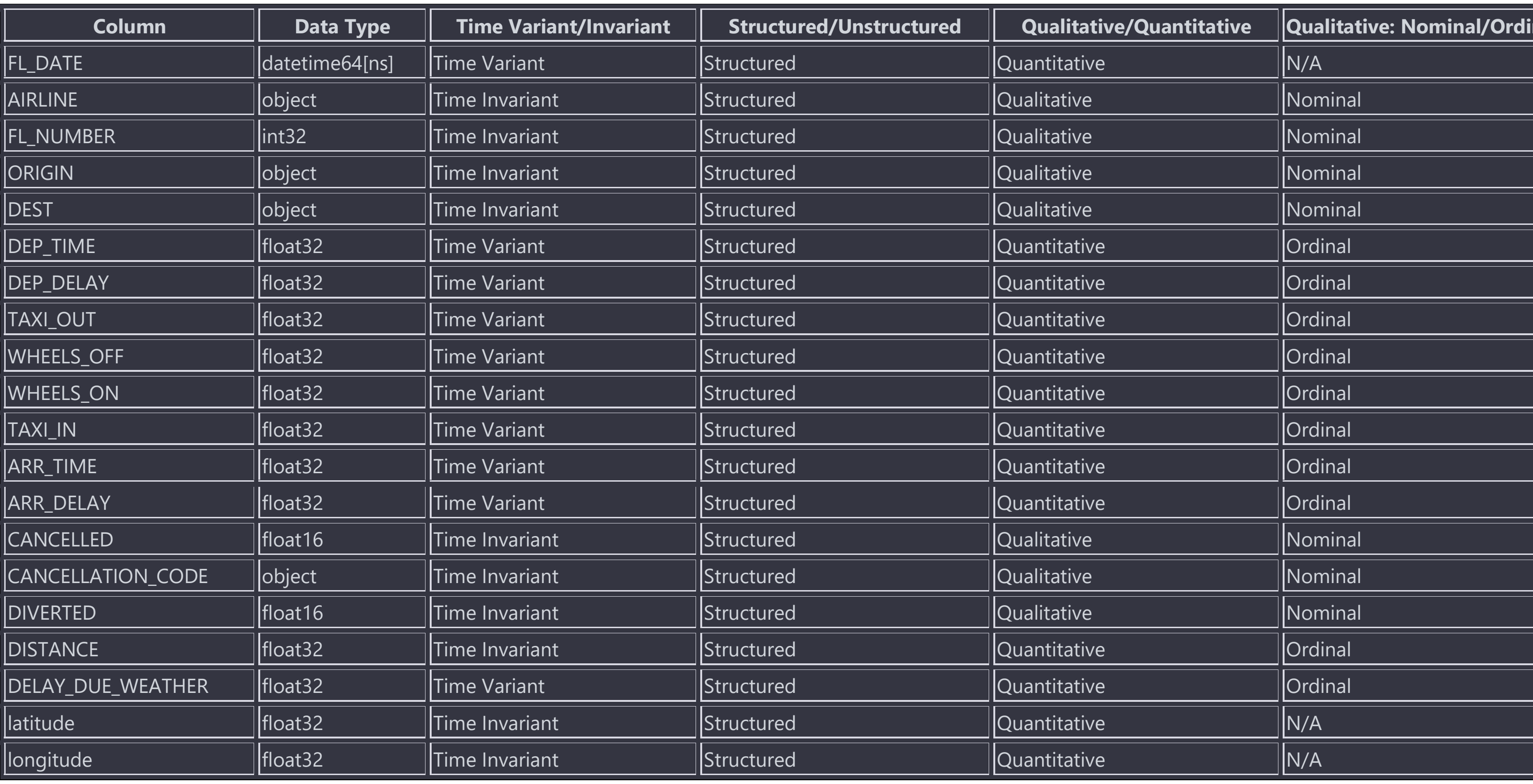
The dataset underwent a series of preprocessing steps to optimize it for analysis. Initially, irrelevant columns such as 'AIRLINE_CODE', 'AIRLINE_DOT', 'DOT_CODE', 'CRS_ELAPSED_TIME', 'ELAPSED_TIME', 'DELAY_DUE_CARRIER', 'ORIGIN_CITY', 'DEST_CITY', 'CRS_DEP_TIME', 'CRS_ARR_TIME', 'AIR_TIME', 'DELAY_DUE_NAS', 'DELAY_DUE_SECURITY', and 'DELAY_DUE_LATE_AIRCRAFT' were identified and subsequently removed.

Following the column removal, data types were adjusted to optimize memory usage. Specifically, the 'FL_DATE' column was converted to the datetime data type, and other fields originally stored as 64-bit were transformed to 32-bit where applicable. This two-step process of deleting unnecessary columns prior to adjusting data types contributes to a streamlined and more memory-efficient dataset. There were several missing values identified for the 'DELAY_DUE_WEATHER' column, but they were not imputed. There were no duplicates discovered. Finally, the two CSV files were merged, culminating in a dataset that is both refined and structured for more effective analysis.

Data Limitations:

The dataset analysis is constrained by potential missing or inaccurate data, limited temporal scope, and variable coverage. Ethical considerations involve privacy, bias mitigation, security, transparency, stakeholder impact, data ownership, and legal compliance, requiring meticulous handling to ensure responsible and unbiased insights from the flight delay and cancellation data.

DATA PROFILE:



Column	Data Type	Time Variant/Invariant	Structured/Unstructured	Qualitative/Quantitative	Qualitative: Nominal/Ordi
FL_DATE	datetime64[ns]	Time Variant	Structured	Quantitative	N/A
AIRLINE	object	Time Invariant	Structured	Qualitative	Nominal
FL_NUMBER	int32	Time Invariant	Structured	Qualitative	Nominal
ORIGIN	object	Time Invariant	Structured	Qualitative	Nominal
DEST	object	Time Invariant	Structured	Qualitative	Nominal
DEP_TIME	float32	Time Variant	Structured	Quantitative	Ordinal
DEP_DELAY	float32	Time Variant	Structured	Quantitative	Ordinal
TAXI_OUT	float32	Time Variant	Structured	Quantitative	Ordinal
WHEELS_OFF	float32	Time Variant	Structured	Quantitative	Ordinal
WHEELS_ON	float32	Time Variant	Structured	Quantitative	Ordinal
TAXI_IN	float32	Time Variant	Structured	Quantitative	Ordinal
ARR_TIME	float32	Time Variant	Structured	Quantitative	Ordinal
ARR_DELAY	float32	Time Variant	Structured	Quantitative	Ordinal
CANCELLED	float16	Time Invariant	Structured	Qualitative	Nominal
CANCELLATION_CODE	object	Time Invariant	Structured	Qualitative	Nominal
DIVERTED	float16	Time Invariant	Structured	Qualitative	Nominal
DISTANCE	float32	Time Invariant	Structured	Quantitative	Ordinal
DELAY_DUE_WEATHER	float32	Time Variant	Structured	Quantitative	Ordinal
latitude	float32	Time Invariant	Structured	Quantitative	N/A
longitude	float32	Time Invariant	Structured	Quantitative	N/A

Summary Statistics:

```
+-------------------+-----------------+----------------+----------------+---------------+----------------+----------------+----------------+----------------+-
|              | FL_NUMBER     | DEP_TIME      | DEP_DELAY     | TAXI_OUT      | WHEELS_OFF    | WHEELS_ON     | TAXI_IN       | ARR_TIME      | ARR_DELAY     | CANCELLED     | DIVERTED      | DISTANCE           |
+-------------------+-----------------+----------------+----------------+---------------+----------------+---------------+----------------+----------------+----------------+---------------------+-------------------+-------------------+
| count         | 1.551842e+06   | 791105.000000  | 790821.000000  | 779254.000000  | 779254.000000  | 775721.000000  | 775721.000000  | 775722.000000  | 774141.000000  | 1551842.0      | 1.551842e+06
| mean          | 2.567300e+03   | 1331.823120    | 11.158143      | 16.662516      | 1351.786011    | 1461.620239    | 7.665515       | 1465.657471    | 4.280709       | NaN            | 1.147270e-03   | 7.802709e+02
| std           | 1.767008e+03   | 500.233185     | 52.397831      | 9.212326       | 501.073425     | 526.752625     | 6.253695       | 531.288391     | 51.487415      | 0.0            | 3.387451e-02   | 5.582019e+02       | 32.28
| min           | 1.000000e+00   | 1.000000       | -84.000000     | 1.000000       | 1.000000       | 1.000000       | 1.000000       | 1.000000       | -89.000000     | 0.0            | 0.000000e+00   | 2.100000e+01       | -2.71891
| 25%           | 1.081000e+03   | 918.000000     | -6.000000      | 11.000000      | 931.000000     | 1049.000000    | 4.000000       | 1053.000000    | -16.000000     | 0.0            | 0.000000e+00   | 3.690000e+02       | 0
| 50%           | 2.207000e+03   | 1325.000000    | -2.000000      | 14.000000      | 1335.000000    | 1500.000000    | 6.000000       | 1504.000000    | -7.000000      | 0.5            | 0.000000e+00   | 6.420000e+02       | 0
| 75%           | 3.913000e+03   | 1740.000000    | 7.000000       | 19.000000      | 1752.000000    | 1907.000000    | 9.000000       | 1912.000000    | 7.000000       | 1.0            | 0.000000e+00   | 1.014000e+03       | 0
| max           | 8.819000e+03   | 2400.000000    | 2368.000000    | 186.000000     | 2400.000000    | 2400.000000    | 214.000000     | 2400.000000    | 2221.000000    | 1
```

Key Questions:

- What is the overall frequency of flight delays and cancellations in the dataset?
- Are there specific time periods, airlines, or routes with higher rates of delays or cancellations?
- What are the primary reasons for flight delays and cancellations (e.g., weather, technical issues)?
- How does the time of day, airline, or distance impact delays?
- Which airlines have the best on-time performance?
- Are there specific airlines consistently facing higher delays or cancellations?
- Are there specific routes with a higher likelihood of delays?
- Is there a correlation between the distance of the flight and the likelihood of delays?
- Are there peak times during the day or specific days with higher delays?
- How do delays vary by seasons?
- Which airports experience the highest delays or cancellations?
- Are certain airports more prone to specific types of delays?
- Can we predict the likelihood of a flight being delayed based on historical data?
- Are there specific factors that are strong predictors of delays?
- How can airlines mitigate the impact of common causes of cancellations?
- Are there clusters of airports with similar delay patterns?
- How do delays propagate through airline networks?
- How do delays and cancellations affect passenger satisfaction?
- What are the financial implications for passengers due to delays?