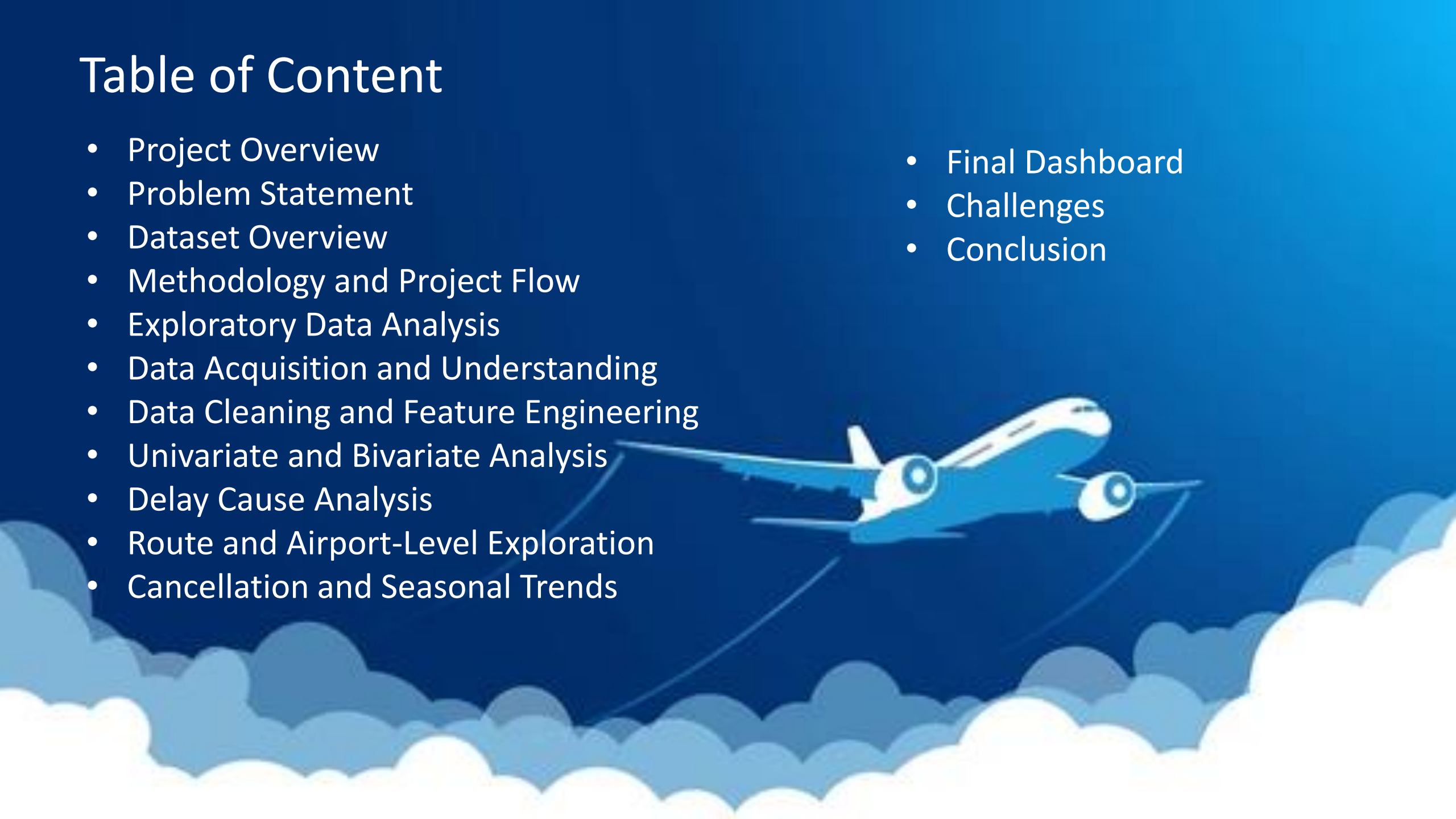
AIRFLY INSIGHTS

Data Visualization and Analysis of Airline and Airport Operations

Created By: Unnati Saxena

Table of Content



Project Overview

Airfly Insights explores large-scale U.S. airline data to understand delays, cancellations, and route performance. It uses data analytics and visual dashboards to help identify operational issues and improve flight efficiency.

Problem Statement

The objective of this project is to analyze large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons using data visualization techniques. The goal is to help understand airline and airport-level performance and contribute to actionable insights using visual analysis.

Dataset Overview

- Rows: 484,549
- Columns: 36 (flight times, delays, airlines, airports, status)
- Size: Large multi-year U.S. flight dataset
- Strengths: Rich features, detailed delay causes, high volume for reliable analysis.

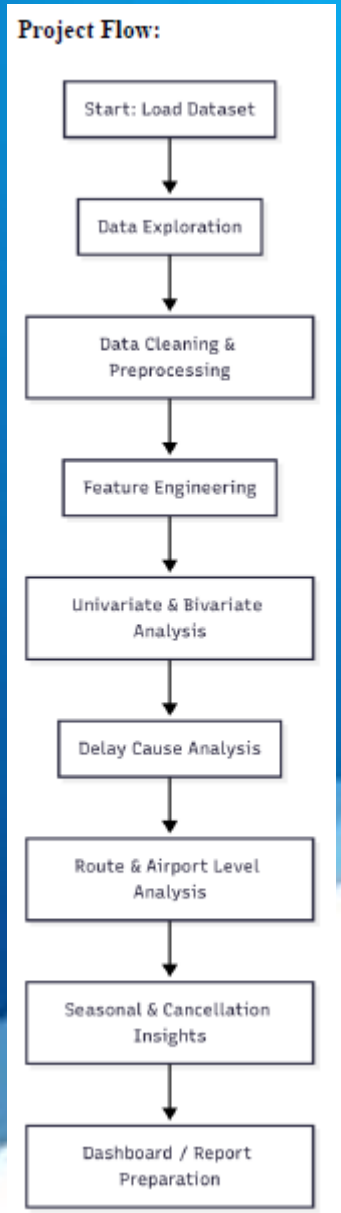
Reason of Choosing this Dataset

- The dataset is large and real-world, providing meaningful opportunities for practical data analysis.
- It contains diverse features such as delays, routes, airports, cancellations, and timings, allowing multi-dimensional insights.
- The dataset allows studying seasonal patterns, route performance, and airline efficiency, making it suitable for a complete data visualization project.

Methodology and Project Flow

The project follows a structured workflow starting with **data loading and initial exploration** to understand the dataset's structure, types, and missing values. Next, **data cleaning and preprocessing** are performed using pandas, including handling nulls, removing duplicates, converting date fields, and standardizing numeric columns. New features such as Month, DayNumber, Hour, Route, and OnTime are created for deeper analysis.

After preparation, **exploratory data analysis** is conducted using visualizations like bar charts, heatmaps, box plots, and line graphs to identify trends in delays, routes, airports, and seasonal patterns. Finally, all insights are compiled into **interactive Power BI dashboards** and summarized in the project's final documentation and presentation.

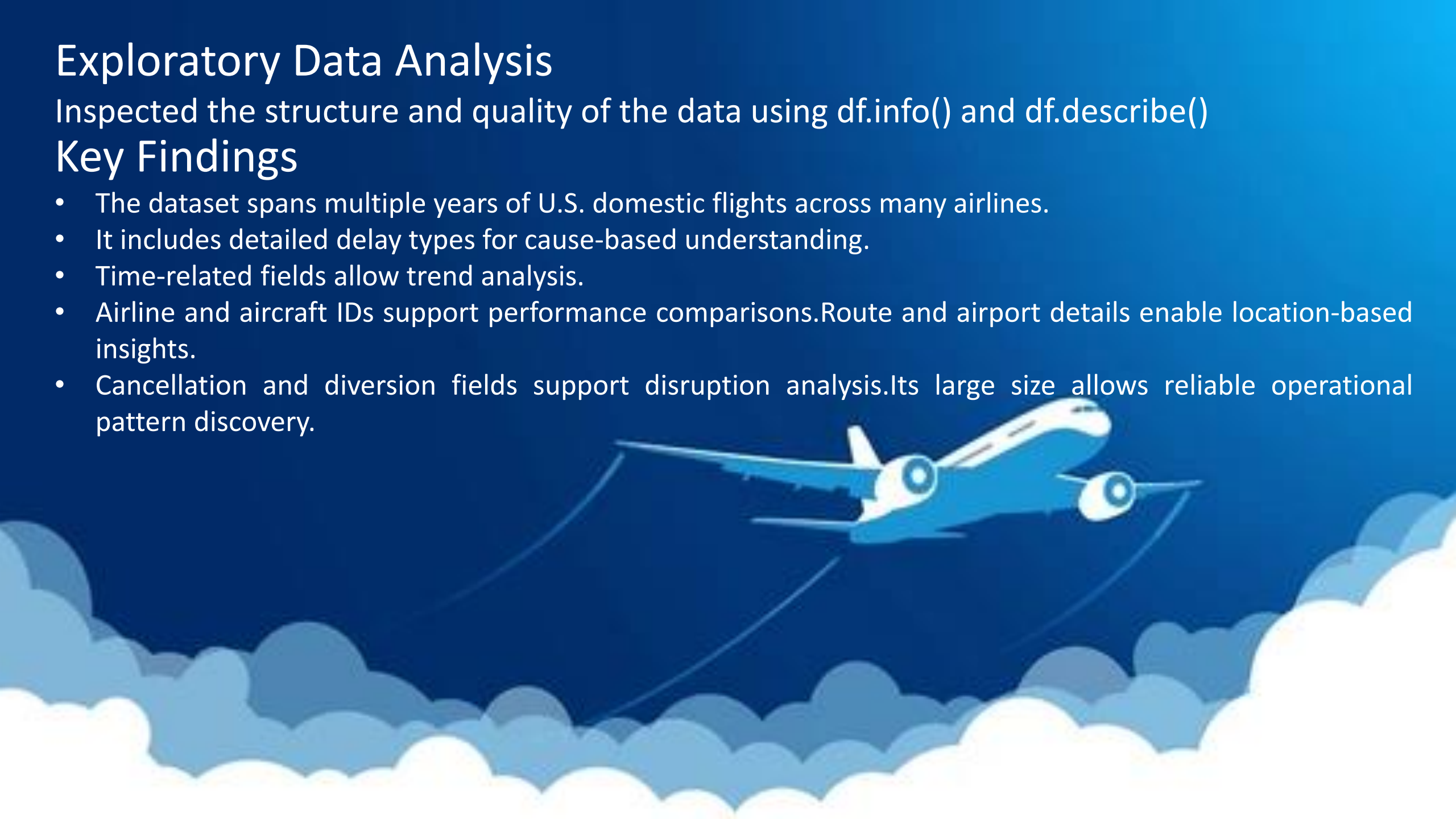


Exploratory Data Analysis

Inspected the structure and quality of the data using df.info() and df.describe()

Key Findings

- The dataset spans multiple years of U.S. domestic flights across many airlines.
- It includes detailed delay types for cause-based understanding.
- Time-related fields allow trend analysis.
- Airline and aircraft IDs support performance comparisons.Route and airport details enable location-based insights.
- Cancellation and diversion fields support disruption analysis.Its large size allows reliable operational pattern discovery.



Data Acquisition and Understanding

Dataset Source (Kaggle)

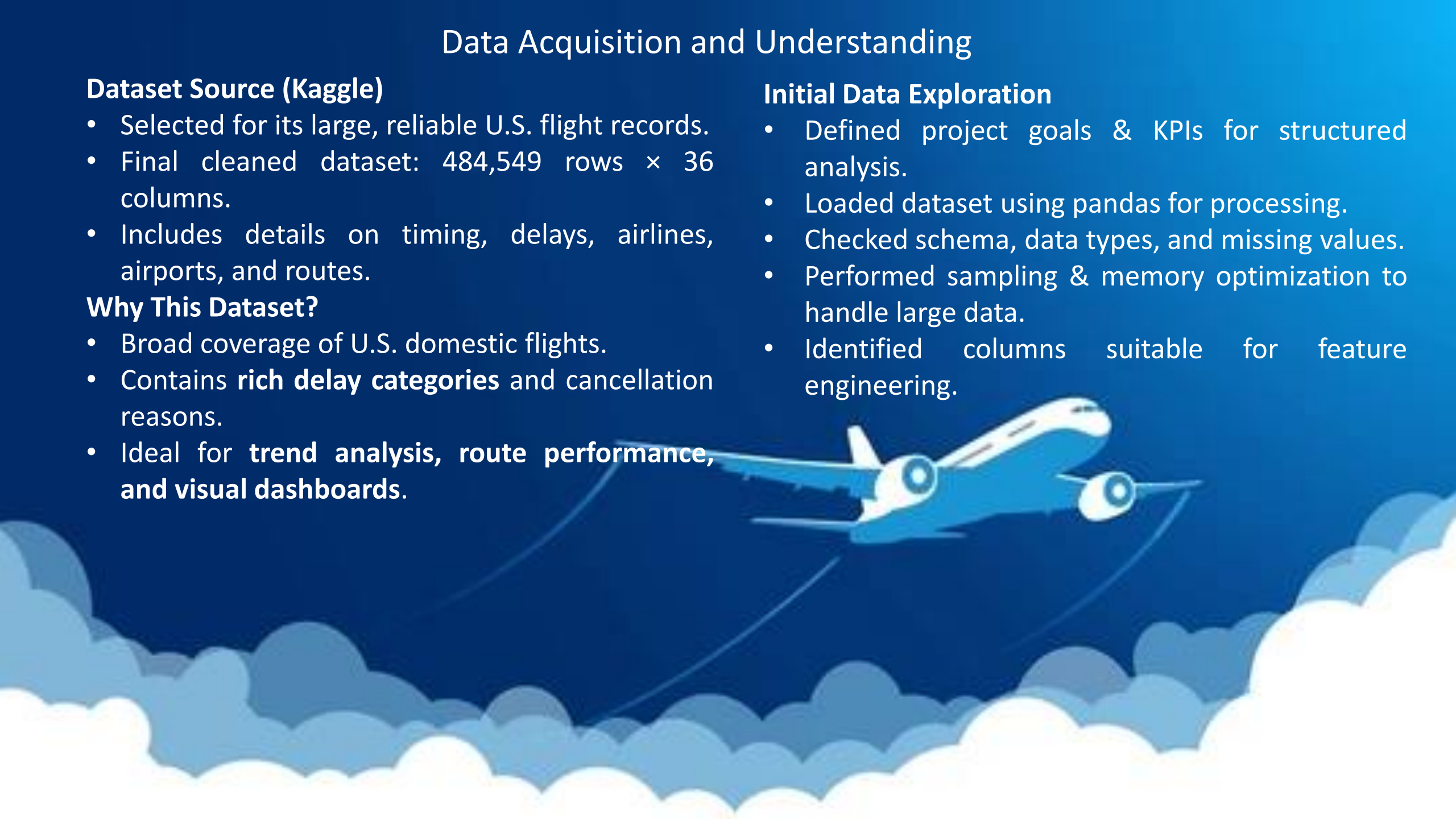
- Selected for its large, reliable U.S. flight records.
- Final cleaned dataset: 484,549 rows × 36 columns.
- Includes details on timing, delays, airlines, airports, and routes.

Why This Dataset?

- Broad coverage of U.S. domestic flights.
- Contains **rich delay categories** and cancellation reasons.
- Ideal for **trend analysis, route performance, and visual dashboards**.

Initial Data Exploration

- Defined project goals & KPIs for structured analysis.
- Loaded dataset using pandas for processing.
- Checked schema, data types, and missing values.
- Performed sampling & memory optimization to handle large data.
- Identified columns suitable for feature engineering.



Data Cleaning and Feature Engineering

- **Handled missing data** – Fixed null values in airport, delay, and cancellation fields.
- **Engineered new features** – Added Month, DayNumber, DepHour, Route, and OnTime for deeper insights.
- **Standardized time fields** – Converted date and time columns into consistent datetime formats.
- **Enhanced dataset usability** – Saved the transformed dataset for faster future analysis.

Oct 29, 2025 (<1s) 4: Fill the Nulls in Cancellation Python

```
#Fill Empty cells of Cancelled column with 0 or 1 according to the CancellationCode if N then 0 otherwise 1
print(df["Cancelled"].isnull().sum())
df['Cancelled'] = df['CancellationCode'].fillna('N').apply(lambda x: 0 if x == 'N' else 1)
```

13

Oct 29, 2025 (7s) 8: Format datetime columns Python

```
df['Date'] = pd.to_datetime(df['Date'],dayfirst=True)
display(df)
```

Oct 29, 2025 (1s) 9: Create derived features: Month, Day of Week, Ho... Python

```
df['Month'] = df['Date'].dt.month
df['DayOfWeek'] = df['Date'].dt.dayofweek + 1  # 1=Monday, 7=Sunday
df['Hour'] = (df['DepTime'] // 100).astype(int)
df['Route'] = df['Origin'] + '-' + df['Dest']
df['Duration'] = df['AirTime'].apply(lambda x: f"{x//60}h {x%60}m" if pd.notna(x) else "0h 0m")
```

Oct 29, 2025 (<1s) 12: Handling Nulls in Delays Python

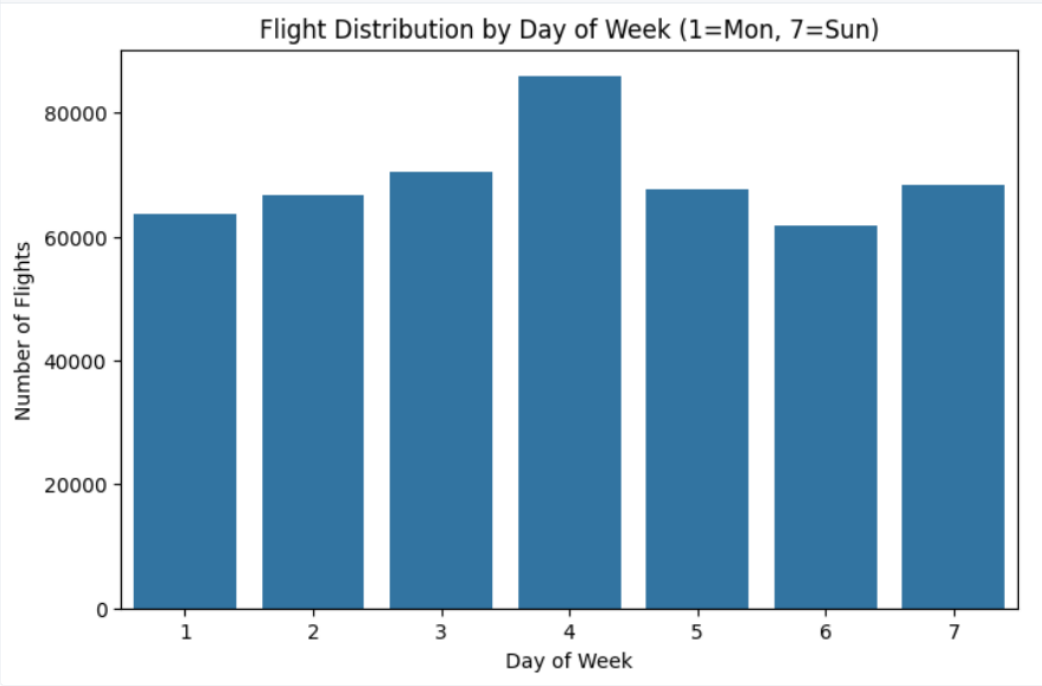
```
delay_cols = ['ArrDelay', 'DepDelay', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDelay',
'LateAircraftDelay']
df[delay_cols] = df[delay_cols].fillna(0)
```

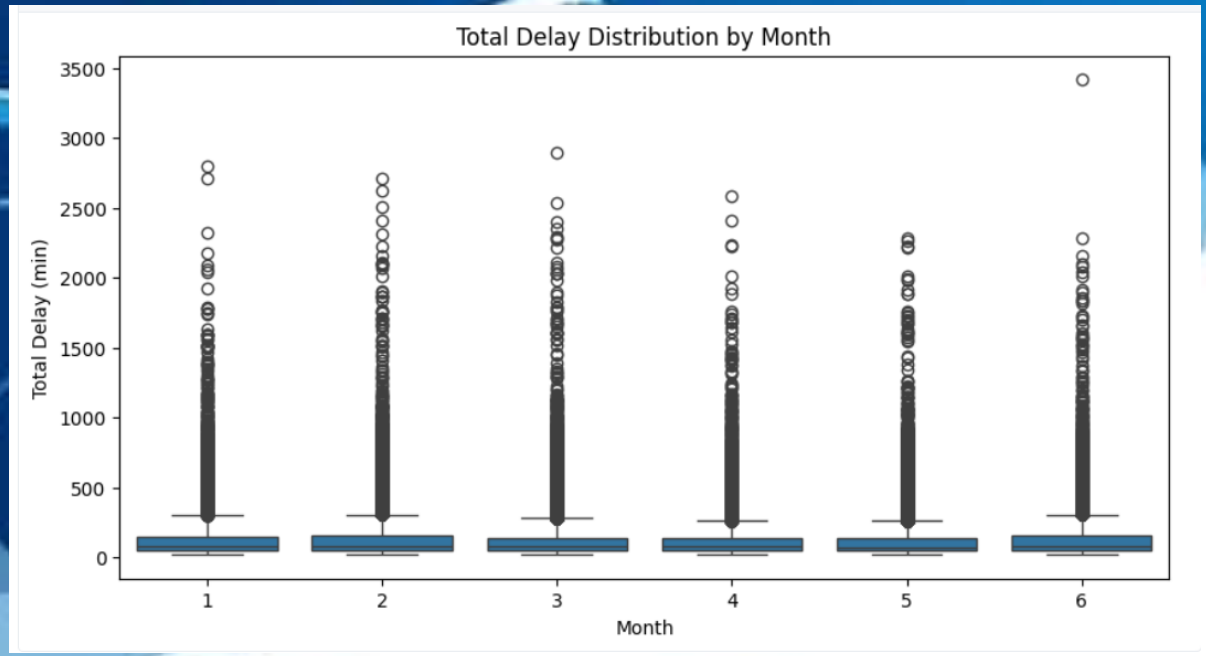
Oct 29, 2025 (7s) 14: Save preprocessed data for fast reuse

```
df.to_csv("/Volumes/workspace/default/airlines/Flight_delay_cleaned.csv")
```

Univariate and Bivariate Analysis

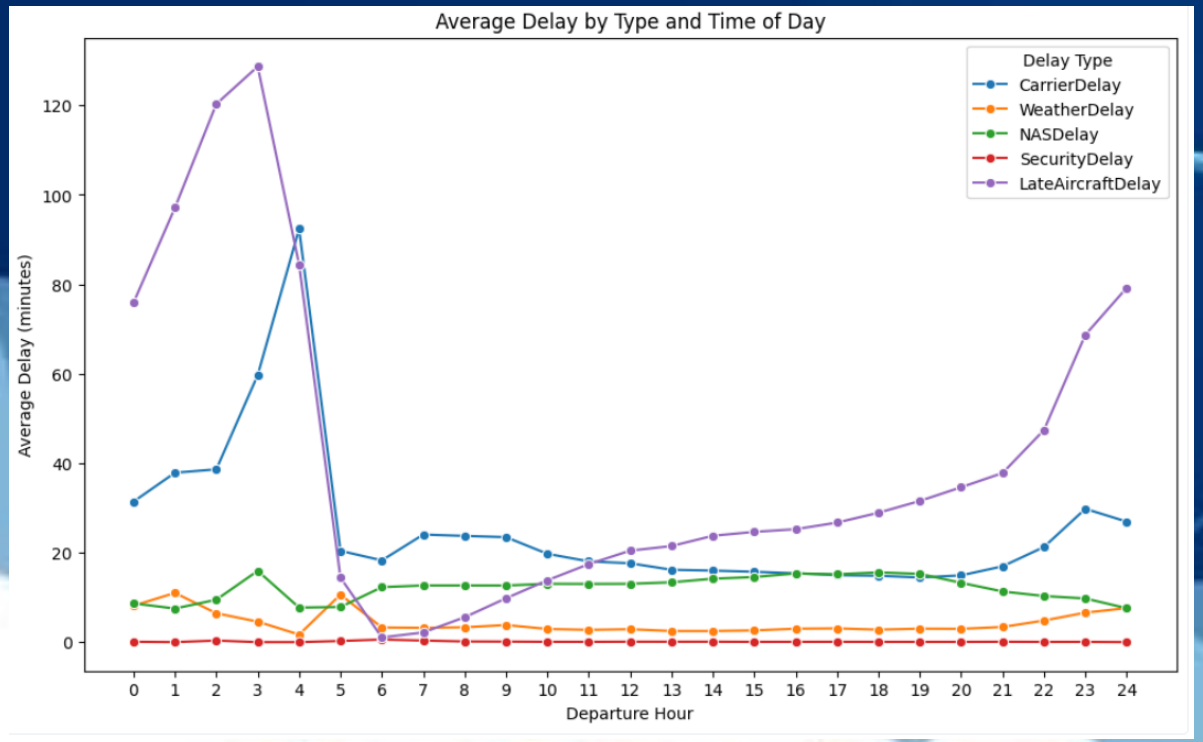
Univariate Analysis – Explored single columns such as airlines, months, delay types, and departure hours to understand distributions.
Bivariate Analysis – Compared pairs of variables (Airline vs Delay, Airport vs Cancellations, Time vs Delay) to detect patterns and relationships.
Visuals Used – Bar charts, line plots, histograms, scatter plots, and boxplots.
Outcome – Found peak delay hours, busiest airlines/routes, and early signs of seasonal patterns.

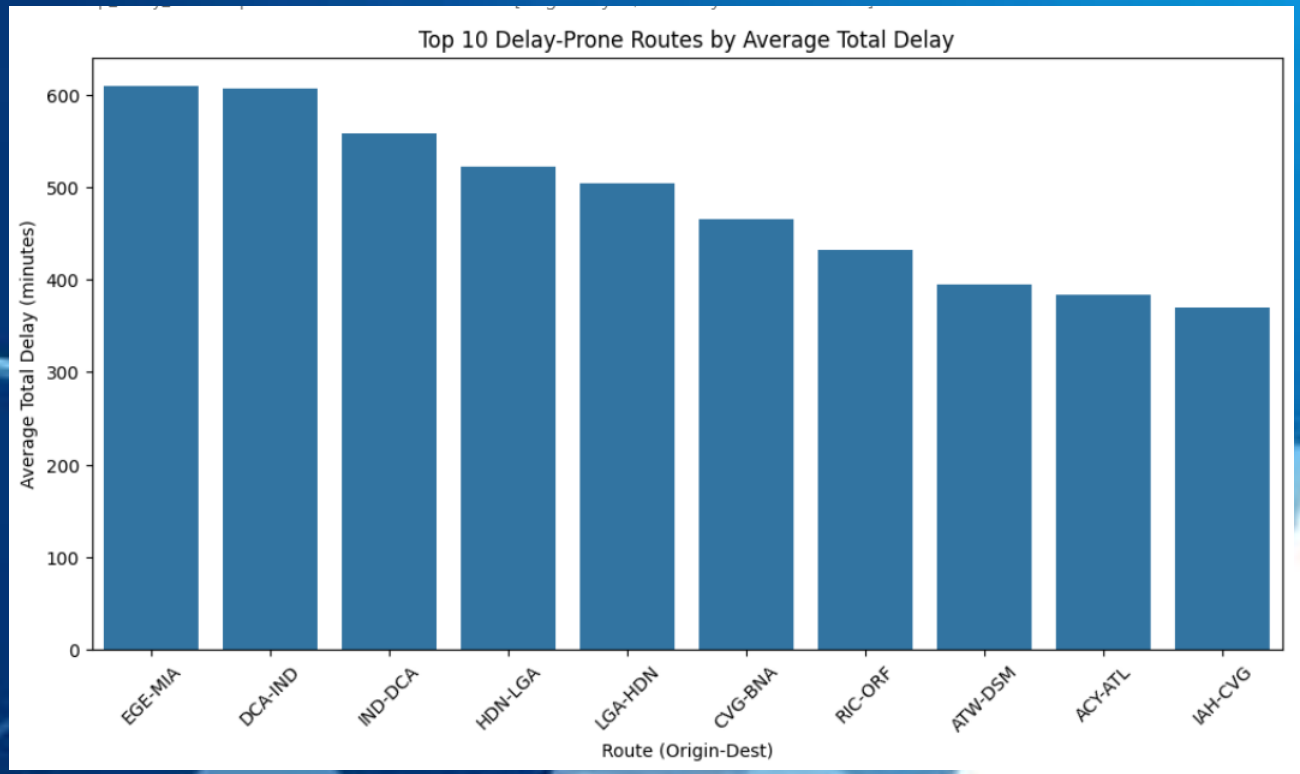




Delay Cause Analysis

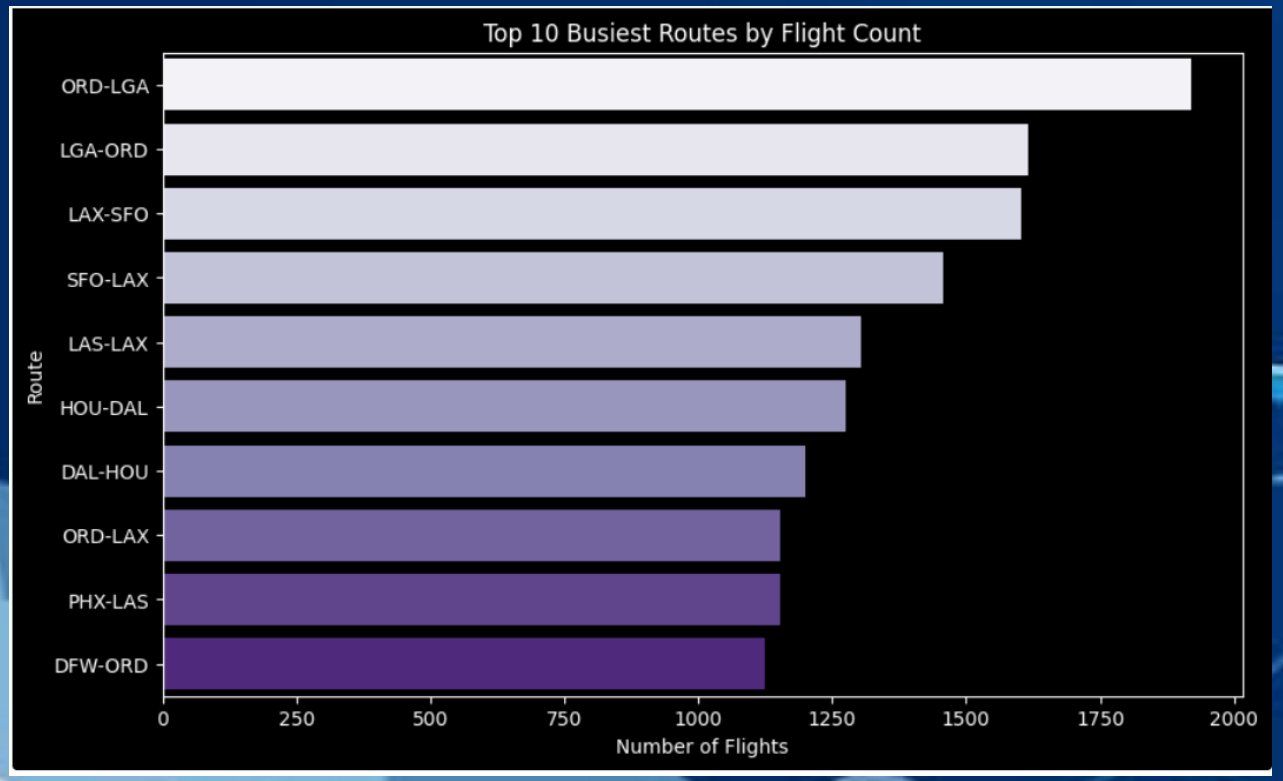
- **Analyzed delay categories** – Carrier, NAS, Weather, Security, and Late Aircraft.
- **Compared delays across** airlines, airports, and departure hours.
- **Identified top delay-prone routes and airlines** with consistently higher averages.
- **Used visuals** to highlight peak delay times and major contributing factors.

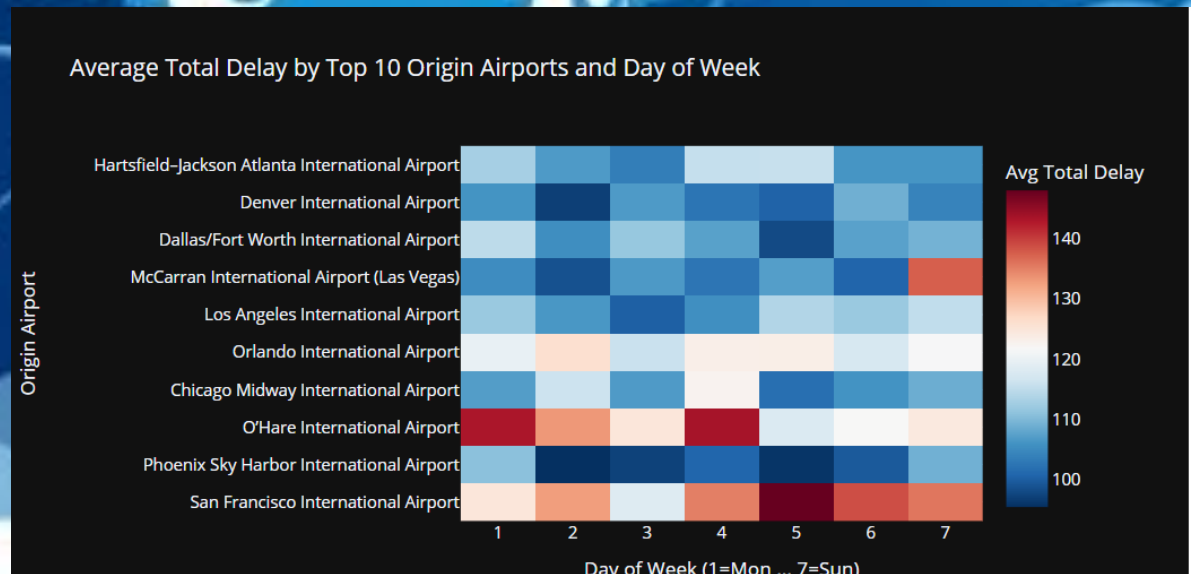




Route and Airport-Level Exploration

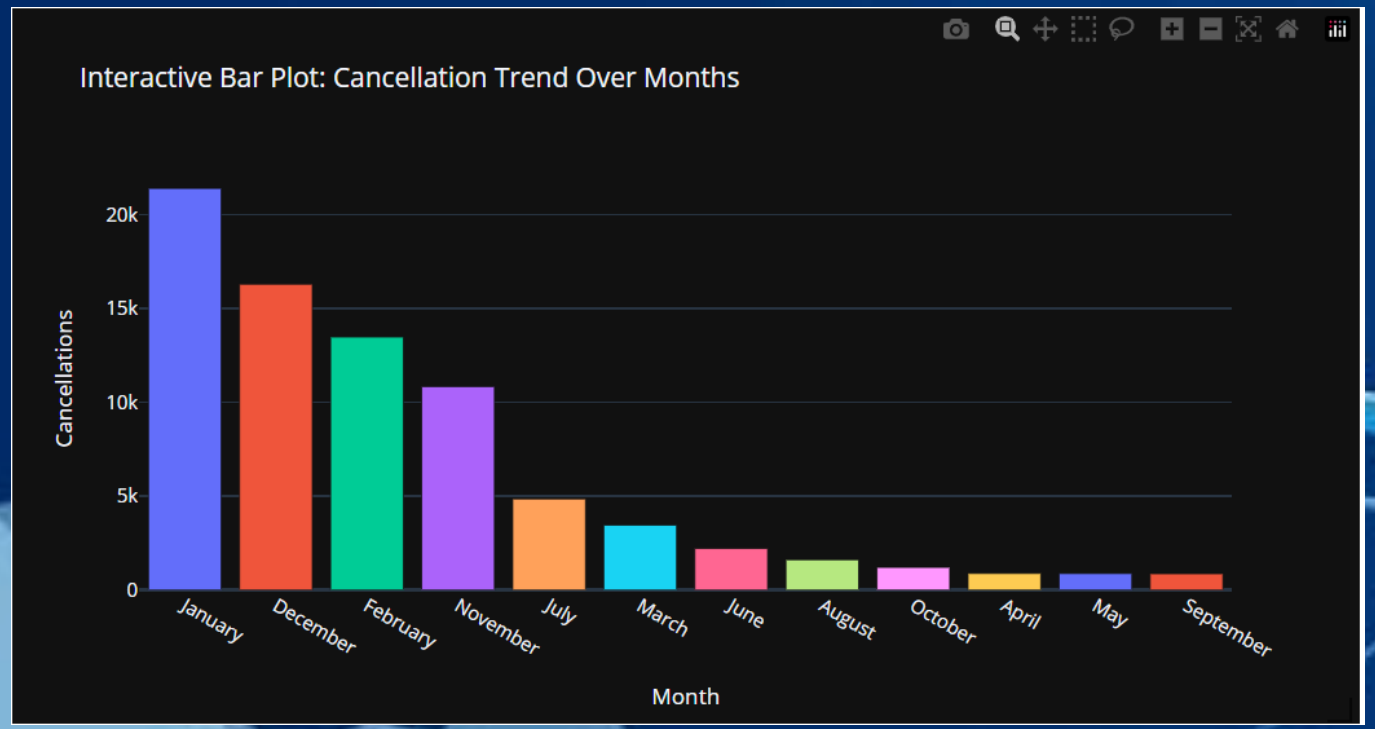
- **Identified busiest routes** using origin–destination flight counts.
- **Analyzed airport delays** to find the most congested or delay-heavy airports.
- **Compared route-level delays** to highlight efficient vs. problematic flight paths.
- **Checked how distance and congestion** influence overall flight performance.

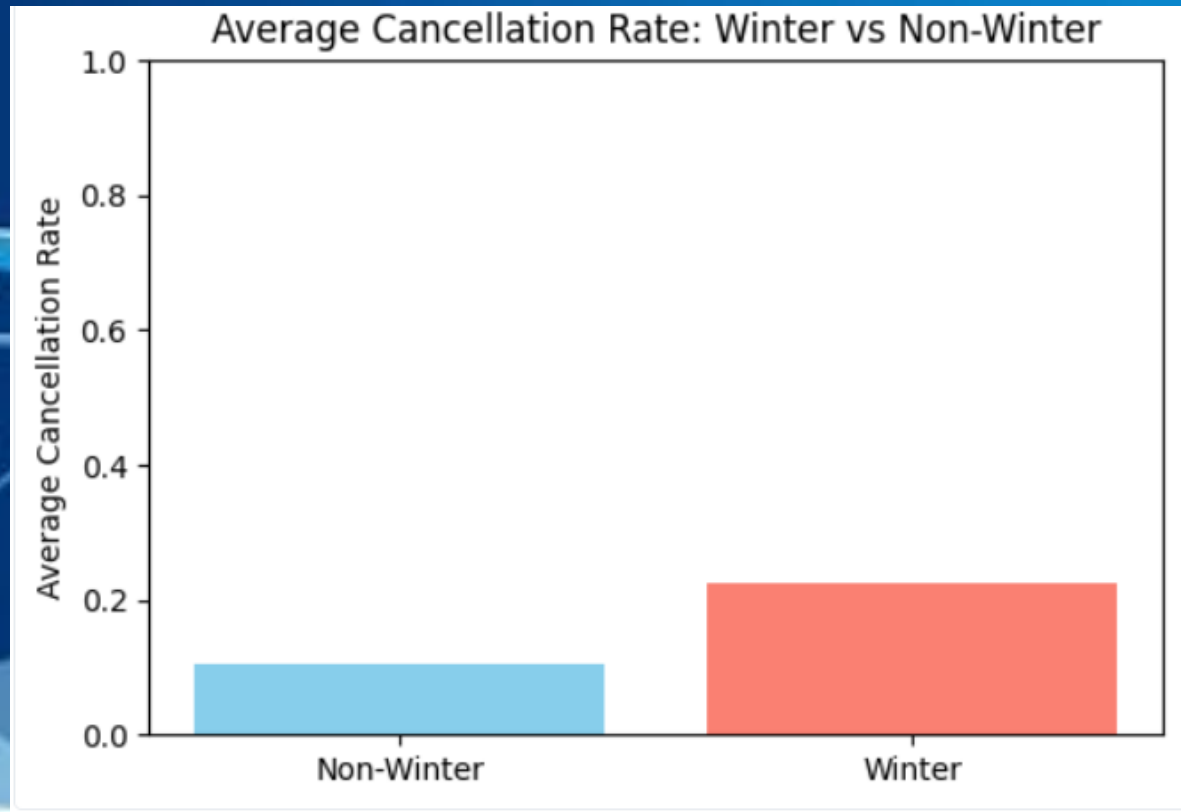




Cancellation and Seasonal Trends

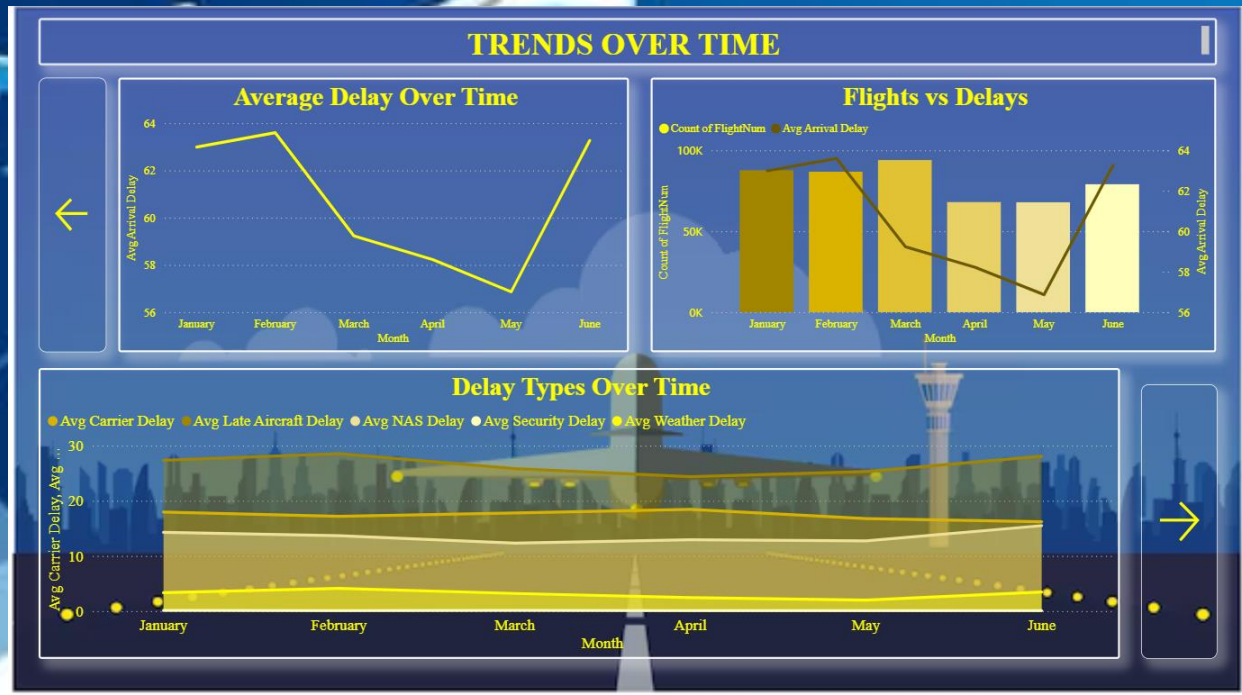
- **Studied monthly & seasonal cancellation patterns** to find peak disruption periods.
- **Analyzed cancellation reasons** (Carrier, Weather, NAS, Security).
- **Mapped cancellation codes** to correct causes for accurate classification.
- **Compared cancellations** across airlines and routes to detect highly affected segments.





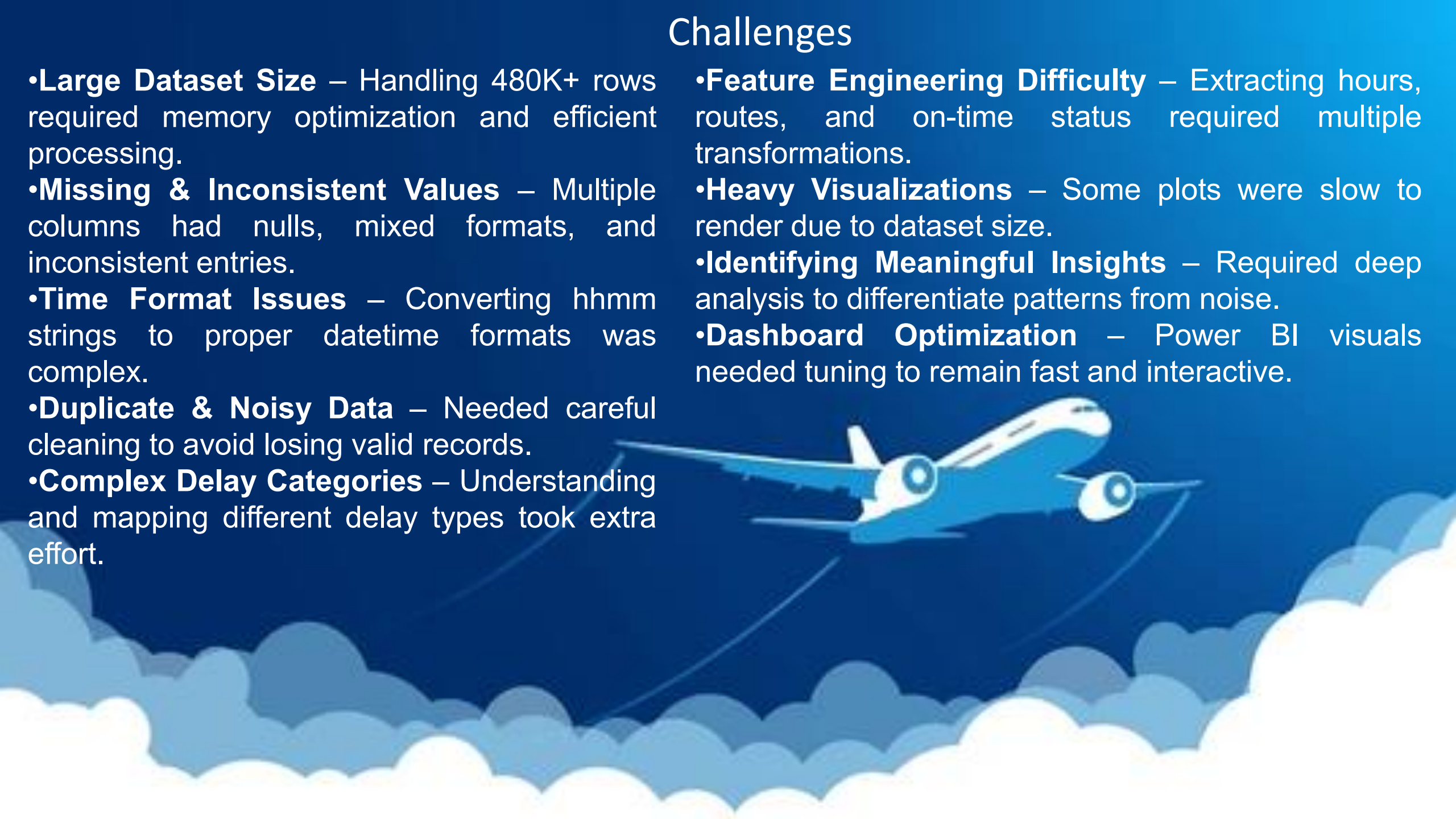
Final Dashboard

- **Imported cleaned data** into Power BI.
- **Transformed fields** (Month, Route, Delay Type) using Power Query.
- **Built model relationships** and added calculated columns.
- **Created visual dashboards** using bar, line, map, and pie charts.
- **Added KPIs** like total flights, delays, and cancellation rate.
- **Applied filters** for Airline, Airport, Month, and Delay Cause.
- **Published dashboard** for interactive exploration.



Challenges

- **Large Dataset Size** – Handling 480K+ rows required memory optimization and efficient processing.
- **Missing & Inconsistent Values** – Multiple columns had nulls, mixed formats, and inconsistent entries.
- **Time Format Issues** – Converting hhmm strings to proper datetime formats was complex.
- **Duplicate & Noisy Data** – Needed careful cleaning to avoid losing valid records.
- **Complex Delay Categories** – Understanding and mapping different delay types took extra effort.
- **Feature Engineering Difficulty** – Extracting hours, routes, and on-time status required multiple transformations.
- **Heavy Visualizations** – Some plots were slow to render due to dataset size.
- **Identifying Meaningful Insights** – Required deep analysis to differentiate patterns from noise.
- **Dashboard Optimization** – Power BI visuals needed tuning to remain fast and interactive.



Conclusion

This project turned raw flight data into clear insights on delays, cancellations, routes, and airline performance. Through cleaning, analysis, and dashboards, it revealed key patterns that can improve scheduling, reduce delays, and support better operational decisions.