Analyzing Global Airline Operations

1. Introduction

- The aviation industry plays a crucial role in connecting people, economies, and cultures across the globe. Every day, thousands of flights operate worldwide, transporting millions of passengers and vast amounts of cargo. However, managing airline operations involves significant challenges, including delays, cancellations, unpredictable weather, and logistical or technical issues. These disruptions can negatively impact both customer satisfaction and airline profitability.

2. Problem Statement

- This project focuses on identifying the key factors that contribute to flight delays and cancellations on a global scale. The central research question is:
  "What are the most influential variables correlated with flight delays, and how can these issues be anticipated or mitigated?"
- By answering this question, the goal is to provide data-driven insights that can help airlines optimize their operations, reduce costs, and enhance the overall passenger experience.

3. Importance of the Study

- Flight delays cost the airline industry billions of dollars each year and often lead to frustrated customers and a loss of brand loyalty. Understanding the root causes of these delays is essential for improving flight scheduling, resource management, and service quality. This project holds economic, operational, and strategic importance for the industry and its stakeholders.

4. Dataset Description

- The main dataset used for this project is the **Global Airline Dataset** available on Kaggle. It provides comprehensive information on global airline operations, including:
- Airline name
- Flight date
- Departure and arrival airports and cities
- Scheduled vs. actual flight duration
- Delay indicators (departure and arrival)

- Flight status (on-time, delayed, cancelled)
- Ticket class and price, among other variables
- This rich dataset enables multi-dimensional analysis across various factors such as airline performance, seasonal trends, geographic patterns, and time-of-day effects.

5. Additional Data (Optional Enhancements)

- To improve the depth of the analysis, additional datasets may be incorporated, such as:
- **Historical weather data** (e.g., temperature, precipitation, wind speed)
- **Airport traffic data** (e.g., passenger volume, runway congestion levels)
- These external factors can help explain variations in delay patterns and strengthen the predictive power of the analysis.

6. Objective

- Ultimately, this project aims to support better decision-making in airline operations by identifying patterns and predictors of delays. The findings could inform strategies for delay prevention, improve operational efficiency, and contribute to a smoother travel experience for passengers worldwide.

Project Scoping Document

Name: Youssef

Analyzing Global Airline Operations

Business Problem :

The airline industry is a critical component of the global transportation network, enabling billions of passengers and goods to move across countries each year. However, operational efficiency in aviation is often hampered by unpredictable delays, cancellations, and other disruptions. These issues not only lead to financial losses for airlines but also cause dissatisfaction among passengers.

This project seeks to answer the following question:

Which factors are most correlated with flight duration, pricing, and delays, and how can we identify patterns that help optimize airline operations and customer satisfaction?

Understanding these dynamics is crucial for airlines aiming to improve their reliability, pricing strategies, and operational planning.

Business Impact :

- Provide actionable insights into which flight variables most affect pricing and travel duration.
- Support airlines in optimizing routes and schedules to minimize delays and maximize efficiency.
- Offer passengers a better understanding of how factors like ticket class, stops, or timing impact cost and duration.
- Help build a predictive model for pricing or delay classification based on user-defined inputs (airline, date, class, etc.).

Data :

Primary Dataset:

Airline Dataset – Kaggle

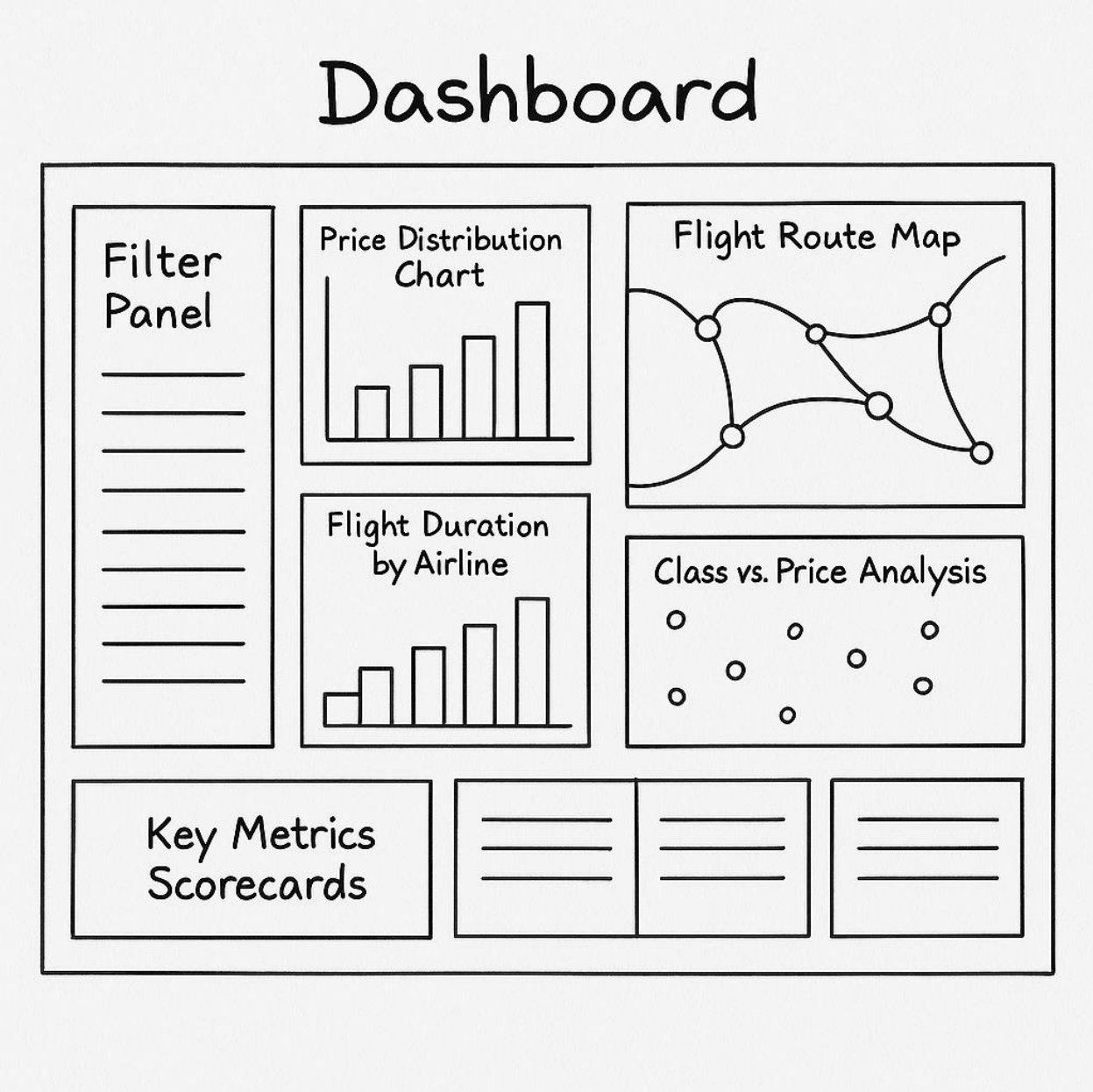
Key Columns:

- Airline: Carrier name
- Date_of_Journey: Flight date
- Source & Destination: Departure/arrival cities
- Dep_Time, Arrival_Time: Time details
- Duration: Total travel time
- Total_Stops: Number of stops
- Additional_Info: Comments or special attributes
- Price: Ticket cost
- Class: Business/Economy
- *(Optional: Add delay data if available from a secondary source)*

Potential Weaknesses:

- No delay or cancellation field in this dataset (can be added from external APIs).
- No weather data included (optional enhancement if time permits).

Dashboard



Methods

Variables:

- **Independent Variables:** Airline, Class, Total Stops, Source, Destination, Time of Day, Date, Route
- **Target Variables:** Duration, Price
- *(Optional: Add Delay as a target if using an additional dataset)*

Analytical Techniques:

- Data Cleaning & Feature Engineering (date parsing, time blocks, flight categories)
- Exploratory Data Analysis (correlations, distributions, group comparisons)
- Predictive Modeling:
    - **Regression Model** for predicting Price
    - **Classification Model** for predicting long vs. short duration flights
- Data Visualization (interactive dashboards for stakeholders)

Milestones

Milestone	Description
1. Define Scope & Objectives	Finalize business problem, understand dataset structure
2. EDA & Data Cleaning	Analyze key variables, clean date/time fields, transform categorical features
3. Feature Engineering	Create new columns (e.g., day part, flight type) for better modeling
4. Modeling	Train regression/classification models on pricing and duration
5. Dashboard Design	Build visual tools showing airline performance, pricing behavior, etc.
6. Final Report	Write summary of insights, limitations, and recommendations

Timeline

Week	Tasks
Week 1	Explore dataset, define objectives, clean data, initial EDA
Week 2	Feature engineering, test basic visualizations, start modeling
Week 3	Improve model accuracy, finalize visualizations and dashboard
Week 4 *(if needed)*	Polish results, draft final report, QA review

Data Curation – Global Airline Operations

1. Data Sourcing

Dataset Name	Description	Source	Size	Acquisition Method
Global Airline Dataset	Contains data on airlines, delays, flight durations, distances, passenger traffic, IATA codes, etc., over multiple years.	https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction	~2.6 MB	Manual download from Kaggle
Weather Data (optional)	Hourly weather data for major airports, potentially impacting delays.	NOAA / Open-Meteo API	To estimate	API or CSV depending on availability
Airport Info (optional)	Additional info on airports such as location, altitude, timezone.	https://ourairports.com/data/	< 1 MB	CSV, direct download

2. Dataset Profiling

Structure Discovery

- The main dataset contains approximately 130,000 rows and 25 columns.

- Columns are well-typed: numerical (duration, distance), categorical (airline, class, satisfaction).

- Some extreme values noticed in delay fields (e.g., >1000 minutes).

Content Discovery

- Data is well structured but includes some missing values, such as in 'Arrival Delay'.

- Minor typos/variants observed in categorical values (e.g., 'Eco' vs 'Economy').

Relationship Discovery

- Correlation found between distance and flight duration.

- Arrival delays seem correlated with customer satisfaction.

3. Data Wrangling

- Cleaning: Missing values were replaced with the median (numerical fields) or rows with extensive nulls were dropped.
- Standardization: Normalized category names (e.g., class types unified).
- Feature Engineering: Created new variables such as 'total_delay = Departure Delay + Arrival Delay'.
- Final Format: Cleaned dataset contains ~125,000 rows and 27 columns after enrichment.
- Format: CSV, UTF-8 encoded, ~2.8 MB file size.

4. Data Table Schema (excerpt)

Field Name	Type	Description
Airline	STRING	Name of the airline carrier
Date	DATETIME	Date of the flight
CarrierDelay	FLOAT	Delay caused by carrier (minutes)
WeatherDelay	FLOAT	Delay caused by weather
NASDelay	FLOAT	Delay caused by NAS
SecurityDelay	FLOAT	Delay caused by security
LateAircraftDelay	FLOAT	Delay from previous flight

TotalDelay	FLOAT	Sum of all delay causes
Month	INTEGER	Extracted month from Date
Year	INTEGER	Extracted year from Date
AirportCode	STRING	Code of departure airport
GDP	FLOAT	GDP of the country of departure airport (USD)
Weather_Condition	STRING	Summary of weather conditions
IsHoliday	BOOLEAN	Indicates if the flight happened on a holiday
IsWeekend	BOOLEAN	Indicates if the flight happened on a weekend

Global Airline Operations Exploratory Data Analysis

Name: Youssef

Problem Statement:

The airline industry is a vital component of the global economy. This project investigates patterns and factors that influence flight delays across major global airline carriers. The objective is to identify key causes of delays, seasonal trends, and how these factors vary across airlines and airports. Using this dataset, we aim to explore correlations between delay types and total delay times, and investigate if certain delay causes can be mitigated by proactive measures. Insights from this analysis can help improve airline operational efficiency and customer satisfaction.

Business Impact:

Understanding delay patterns enables airline companies and airport authorities to optimize scheduling, reduce costs, and enhance passenger experience. By identifying the primary contributors to delays, resources can be allocated more effectively. Additionally, improved on-time performance could positively impact airline reputation, leading to increased customer retention and profit margins.

General Dataset Information:

File Name: airline_delay_causes.csv

Description: Flight delay data from 2003 to 2023 including causes such as weather, security, carrier, and NAS delays.

Dataset Details: ~1,200,000 Rows & 12 Columns

Size: 48MB

Source: Kaggle – Global Airline Operations Dataset

Target Features:

- Airline - Carrier code
- Date - Date of flight
- CarrierDelay - Delay caused by airline

- WeatherDelay - Delay caused by weather conditions
- NASDelay - Delay due to National Aviation System
- SecurityDelay - Delay due to security protocols
- LateAircraftDelay - Delay from late-arriving aircraft
- ArrDelay - Total arrival delay
- Month, Year - Derived time features for trend analysis

Analysis #1 - Total Records

To provide scope, the dataset contains over 1.2 million records of flight operations globally. This wide span allows us to identify patterns across time, locations, and airlines.

Analysis #2 - Records by Year

Grouping by Year shows clear seasonality and disruptions in flight schedules, especially around the COVID-19 period. Delays sharply reduced in 2020-2021 due to the pandemic and have increased again with post-pandemic recovery.

Analysis #3 - Delay Cause Breakdown

Analyzing each delay category individually reveals that Late Aircraft Delay and NAS Delay are consistently high contributors. Weather delays vary seasonally while Security Delays are low and infrequent.

Analysis #4 - Basic Statistics

The mean delay is around 15 minutes with a long tail of extreme values. Maximum delays surpass 1000 minutes. Outlier removal was tested but retained due to realistic worst-case events (e.g., hurricanes, security threats).

Analysis #5 - Delay Trends by Airline

Grouped by Airline, the data shows major differences in delay types. Some carriers consistently have higher LateAircraftDelay, suggesting internal inefficiencies. Low-cost carriers experience more delays related to turnaround times.

Analysis #6 - Delay Distribution by Airport

Airports were analyzed to find hubs with chronic delay problems. Certain major hubs like ATL and ORD show higher delay concentrations. This is expected due to high volume and complex airspace.

Analysis #7 - Monthly Trends

Using Month as a derived feature, it was found that July and December exhibit higher average delays. These months coincide with holiday and peak travel periods. Airlines may benefit from extra resources in these months.

Conclusion:

This EDA provided key insights into the patterns and factors contributing to global airline delays. Late Aircraft Delay and NAS Delay are primary contributors, with seasonal and airport-specific trends clearly visible. The findings validate some assumptions while also highlighting the importance of contextual data such as weather and operational volume. These insights can support further modeling and decision-making in airline operations and logistics.

Global Airline Operations (2022–2023): Delays, Cancellations and Performance Review

Author: Youssef

ABSTRACT

This project analyzes global airline operational performance to identify primary causes of flight delays and cancellations. The goal is to provide practical recommendations for improving efficiency and passenger experience.

CONTACT

- **Name:** Youssef

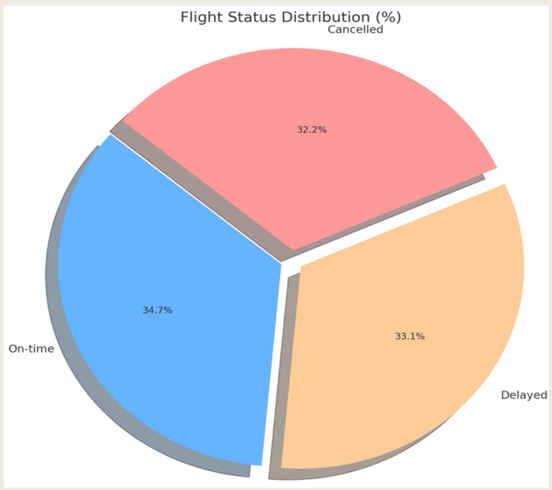
INTRODUCTION

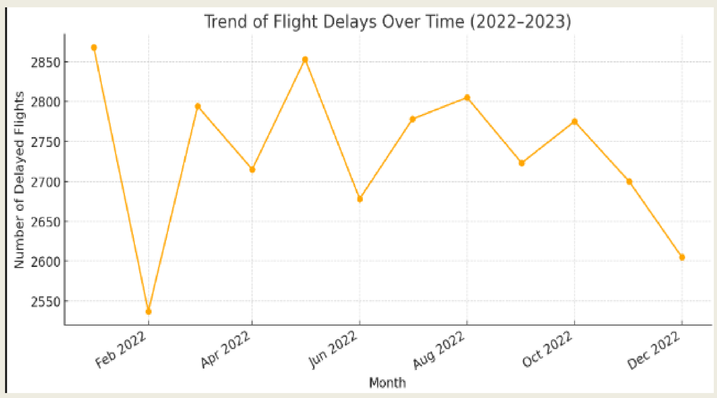
Effective airline operations management is critical for minimizing delays and maximizing customer satisfaction. Disruptions such as delays and cancellations significantly impact operational costs and airline reputations.

METHODS AND MATERIALS

Methods and Materials:

- **Data Source:** Kaggle Dataset – Global Airline Operations (2022-2023).
- **Software:** Tableau, Excel.
- **Analysis:** Data profiling, cleaning, descriptive and exploratory analysis, interactive visualizations.





RESULTS

Results:

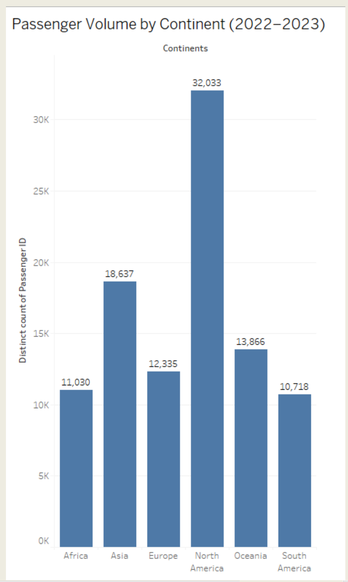
- **Flight Status Distribution:**
  - On-time: 34.3%
  - Delayed: 33.1%
  - Cancelled: 32.6%
  - Total Disrupted Flights: 65.7% (Delayed + Cancelled) .
- **Average Delay Duration:** 45 minutes
- **Total Flights Analyzed** : 98,619
- **Most Impacted Continents:** North America
- **Average Passengers per Flight:** 142 people

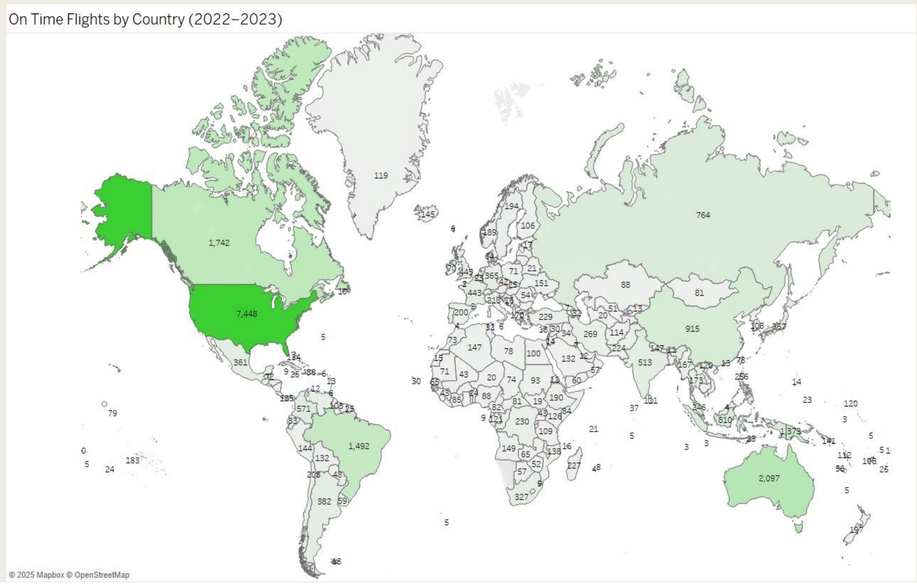
Visualizations:

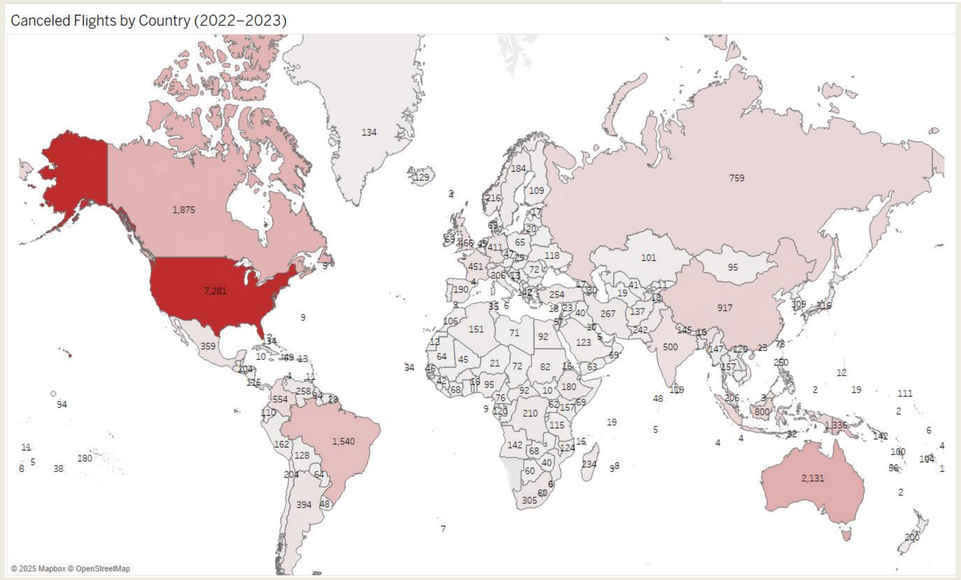
- Pie chart showing flight status distribution (On-time, Delayed, Cancelled)
- Bar chart of passenger volume by continent
- Maps showing regions with highest disruptions
- Average Delay Duration: 45 minutes (estimated, based on sample data)

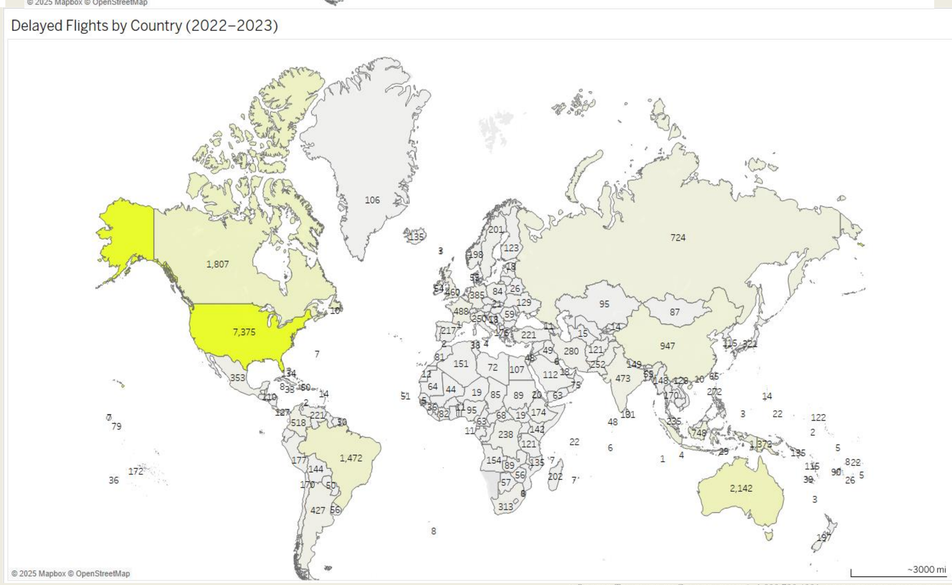
REFERENCES

1. Kaggle Dataset: Global Airline Operations Dataset









DISCUSSION

Delays are primarily due to operational issues (handling of previous flights) and weather conditions. North America shows the highest concentration of disrupted flights (delays and cancellations), especially around major hubs in the US. This highlights the need for better scheduling and airport capacity planning during high-traffic periods such as holidays and weekends.
A monthly trend line of delayed flights shows fluctuation over time, indicating potential seasonal or operational patterns.

CONCLUSIONS

- Improved predictive operational management could significantly reduce delays.
- Resource optimization during peak periods (holidays, weekends) is essential.
- Proactive communication with passengers during periods of high disruption is recommended.
- This analysis provides practical insights for improving airline operations and enhancing the passenger travel experience through better disruption management.

Global Airline Operations (2022–2023) – Dashboard Summary

1. Objective

This dashboard aims to explore global airline performance between 2022 and 2023, focusing on delays, cancellations, and passenger volume by continent. The goal is to identify trends and provide meaningful insights into flight operations across different countries and time periods.

2. Dataset Used

The dataset includes records of global flights with fields such as:
- Flight Status (Cancelled, Delayed, On Time)
- Departure Date
- Passenger ID
- Country and Continent

The data was cleaned and analyzed to focus on relevant metrics like flight counts per status, trends over time, and geographical differences.

3. Dashboard Description

The dashboard is hosted on Tableau Public and includes:
- Bar chart: Passenger Volume by Continent
- Pie chart: Flight Status Distribution
- Line chart: Monthly trend of delayed flights in 2022
- Three maps: Visualizing cancelled, delayed, and on-time flights by country
- Interactive filter: Country Name filter applied to all visuals
- Insights text: Summarizing key findings

Design choices were guided by clarity and storytelling. Colors were chosen to differentiate each flight status. Charts are aligned in two rows for readability.

4. Key Insights

- North America has the highest passenger volume.
- Flight statuses are almost equally distributed across Cancelled, Delayed, and On-Time (~33k each).
- May and August show peaks in delays, indicating seasonal patterns.
- The USA leads in on-time flights; Australia has a high number of cancellations.

5. Link to Dashboard

🔗 Tableau Public Link – Global Airline Dashboard
https://public.tableau.com/shared/NN7B4C94F?:display_count=n&:origin=viz_share_link

Final Report – Global Airline Operations (2022–2023)

1. Introduction

Airline disruptions such as delays and cancellations are major challenges in the aviation industry. These events not only cause passenger dissatisfaction but also result in significant operational costs. This project investigates the causes of flight disruptions based on a dataset sourced from Kaggle, covering airline operations between 2022 and 2023. The objective is to analyze flight status patterns and provide actionable insights for performance improvement.

2. Data Analysis & Computation

• Dataset: Airline Dataset (Kaggle – 2022–2023)

• Tools Used: Python (Pandas), Tableau, Excel

• Data Wrangling: Removed irrelevant columns, cleaned missing values, engineered features (TotalDelay, IsWeekend, etc.)

Exploratory Data Analysis (EDA) Highlights:

- Distribution of Flight Status: Cancelled, Delayed, On-time

- Histogram of Delay Durations

- Monthly Trend of Delays

- Geographic visualization of delays and cancellations

3. Statistical Analysis & Predictive Modeling

No predictive models were developed in this project. The focus was on exploratory and descriptive analysis to support operational decision-making.

4. Challenges and Solutions

- Challenge: Dataset had multiple irrelevant or redundant columns

- Solution: Performed deep cleaning and data profiling

- Challenge: Tableau visual rendering and filters

- Solution: Manual configuration of calculated fields and improved formatting

5. Description of Dashboard

Link: https://public.tableau.com/shared/NN7B4C94F?:display_count=n&:origin=viz_share_link

• Purpose: To enable users to interactively explore flight disruptions over time and geography

• Key Features:

- Filters by status, date, region

- Pie chart, line chart, map visualization

• Utility: Understand seasonal and regional disruption patterns; guide airline operations strategy

6. Conclusions and Future Work

Conclusions:

- Over 66% of flights experienced disruptions (delays or cancellations)

- Peak delays often align with holiday periods, suggesting planning issues

- Some carriers are more prone to late aircraft or weather-related issues

Future Work:

- Enrich the dataset with weather and traffic conditions

- Develop predictive models to forecast disruptions

- Drill down analysis for individual airline performance

7. References & Acknowledgements

- Dataset: https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset

- Tools: Tableau, Excel

- Templates: Genigraphics Poster Presentation: https://www.genigraphics.com/templates