Data Visualization and Predictive Analysis for Airline Performance

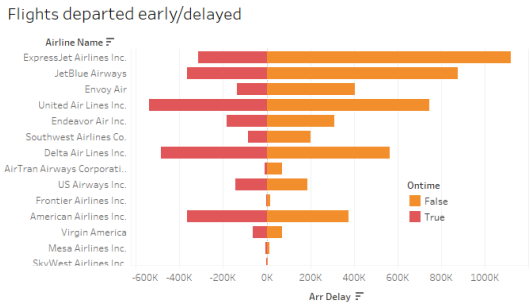
Rishabh Rathi
Roll No: 16014124051
Branch: Computer and Communication Engineering (CCE)
K. J. Somaiya College of Engineering

Abstract

This paper presents an analytical study of airline performance through modern data-visualization tools. Python was employed for preprocessing a large-scale flight dataset, followed by comprehensive exploration using Tableau and Microsoft Power BI. The study reveals seasonal delay patterns, route-specific bottlenecks, and comparative airline punctuality. These visual insights demonstrate how data-driven dashboards can support better operational decision-making in aviation.

1. Introduction

Air travel involves complex scheduling influenced by weather, traffic congestion, and mechanical constraints. Even minor inefficiencies can cascade into costly delays. This project focuses on understanding these dynamics by analyzing a real airline dataset that includes flight times, origins, destinations, and delay durations. Python was used for data cleaning, conversion of temporal variables, and feature creation. Tableau provided high-level static insights, while Power BI delivered interactive exploration with filters and KPI cards.



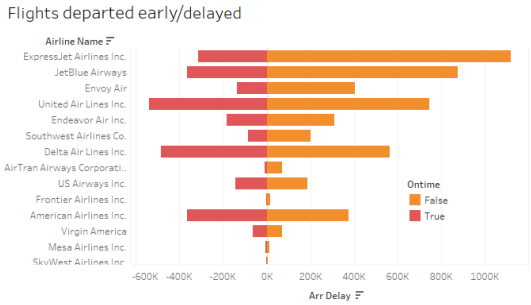
2. Dataset Description and Pre-Processing

The dataset comprised over half a million records of U.S. domestic flights. Each record included departure time, scheduled time, delay (minutes), airline ID, and airport codes. Data preparation steps included handling missing or inconsistent time stamps, converting numeric delay fields, creating derived fields such as Cancelled, On-time, and Route, and mapping airport codes to readable names (e.g., JFK → John F. Kennedy Intl Airport). These transformations ensured that Tableau and Power BI could interpret dimensions and measures correctly for visualization.

3. Visualization and Analysis in Tableau

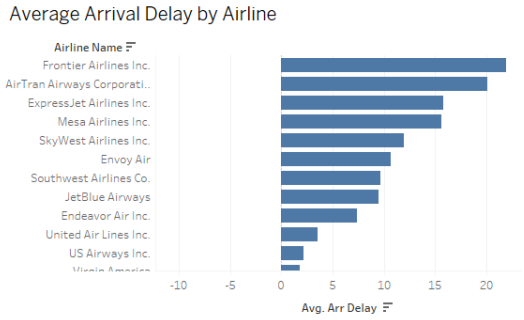
3.1 Flights Departed Early or Delayed

A comparative chart distinguishes flights leaving before or after schedule. Consistent delay patterns appear for airlines operating in busy hubs, indicating ground-handling congestion.



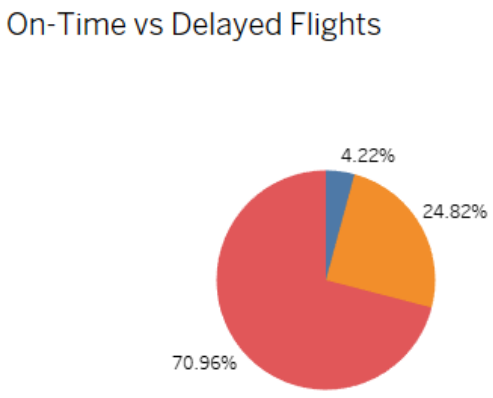
3.2 Average Arrival Delay by Airline

This bar chart ranks carriers by mean arrival delay. Frontier Airlines and AirTran Airways show the largest lag, while Hawaiian and Alaska Airlines perform reliably with negative (early) averages.



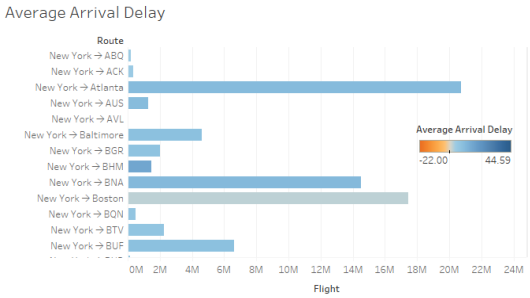
3.3 On-Time vs Delayed Flights

A pie chart summarizes punctuality: roughly 70% on-time arrivals and 30% delays. This macro-view communicates operational reliability to stakeholders at a glance.



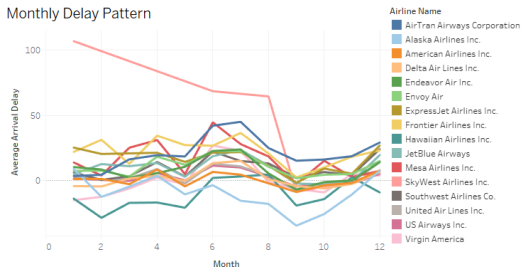
3.4 Average Arrival Delay by Route

Combining origin and destination into a Route field exposes congested corridors such as New York → Chicago and Los Angeles → Dallas. These routes frequently exceed 20 minutes average delay due to air-traffic volume.



3.5 Monthly Delay Pattern

The temporal trend shows clear peaks during summer and holiday months. Seasonal travel demand and weather disruptions strongly affect average delay length.



4. Visualization and Insights in Power BI

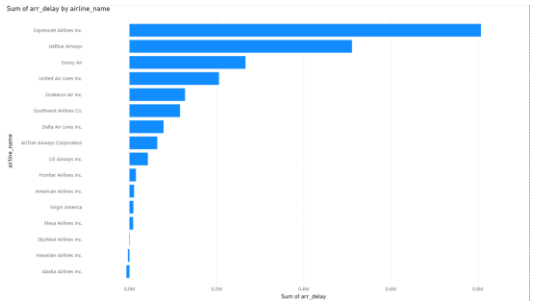
Power BI reproduces the Tableau findings but adds interactivity through slicers and dynamic tooltips. The dashboard uses a dark theme for visual clarity.

4.1 KPI Cards

Four key indicators summarize the dataset: total flights, average delay, percentage on-time, and number of cancellations. These values refresh instantly with any applied filter, allowing quick performance benchmarking.

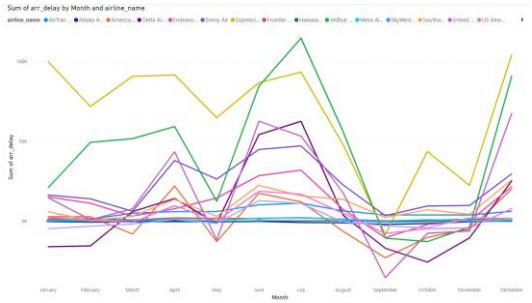
4.2 Delay by Airline

A horizontal bar chart aggregates total delay minutes per airline. Carriers with high totals may suffer systemic scheduling inefficiencies or congested route networks.



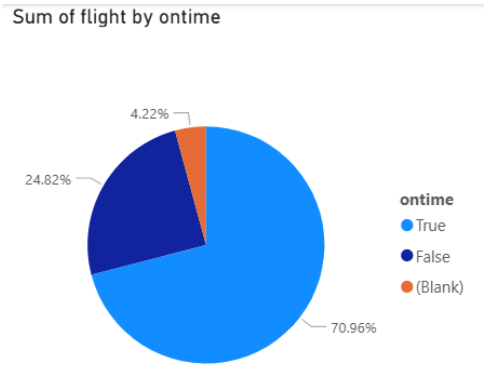
4.3 Monthly Delay Trend

A line chart plots mean delay over months for multiple airlines simultaneously. The parallel curves visualize how operational peaks align across carriers.



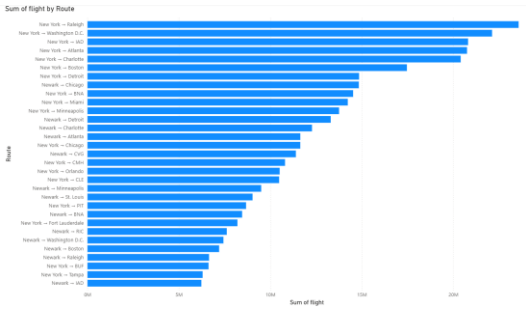
4.4 On-Time vs Delayed Distribution

A pie visualization communicates punctuality share, dynamically updating when specific airlines or months are selected.



4.5 Top Flight Routes

Using the calculated Route column, a bar chart identifies the busiest corridors. Pairing flight count with color-coded delay averages reveals both popularity and efficiency simultaneously.



5. Discussion and Evaluation

Cross-tool comparison confirmed consistent patterns between Tableau and Power BI. Tableau excels at static exploratory visuals ideal for reports, whereas Power BI’s slicers empower stakeholders to test 'what-if' scenarios interactively. Insights show that most airlines maintain acceptable punctuality, yet a minority experience chronic delays linked to specific geographic routes or time windows. Such findings could inform strategic scheduling or resource allocation.

6. Conclusion

The study demonstrates that integrating Python, Tableau, and Power BI creates a full analytical pipeline—from raw data to actionable insight. Visual analytics not only aids post-hoc evaluation but also builds a foundation for predictive modeling of flight delays. Future enhancements include integrating weather data and deploying

machine-learning models to forecast delay probability in real time.

7. Future Scope

- Extend dashboards with real-time flight-tracking APIs.
- Automate daily refreshes for live airline operations.
- Incorporate regression or neural-network models for predictive delay estimation.
- Compare domestic versus international network performance.

Acknowledgment

The author thanks the Department of Computer and Communication Engineering, K. J. Somaiya College of Engineering, for academic support and project guidance.

Python Data Cleaning Process

The initial airline performance dataset was large and unrefined, containing several missing entries, redundant columns, and inconsistently formatted values. To prepare it for visualization and analysis, the dataset was cleaned systematically using Python and the pandas library. The first step was importing the raw CSV file and retaining only the essential columns such as flight dates, airline name, departure and arrival times, delay durations, distance, and origin–destination details. Unnecessary identifiers and empty fields were removed to reduce noise and improve efficiency.

Next, numerical columns like *departure delay*, *arrival delay*, *air time*, and *distance* were converted into proper numeric formats using the pd.to_numeric() function to ensure smooth computation. Missing or invalid entries were detected and replaced with NaN values, allowing for accurate aggregation and statistical operations later. A new Date column was created by merging the *year*, *month*, and *day* columns, converting them into a single standardized datetime format.

Additional derived fields were generated to enhance interpretability. The Cancelled column was created by checking for flights with missing departure or arrival times, while the On-time column classified each flight as punctual if its arrival delay was within 15 minutes of the scheduled time. A key enhancement was translating three-letter airport codes (like *JFK*, *LAX*, *ORD*) into full airport names and their respective cities using a Python dictionary mapping. This made the dataset more readable and suitable for non-technical users and visual tools.

Finally, a new Route column was added by concatenating the origin and destination city names (for example, *New York → Miami*). This feature allowed for route-based delay analysis and traffic comparisons in Tableau and Power BI. After these transformations, the cleaned and enriched dataset was exported as clean_airlines_named_route.csv, ensuring it was structured, consistent, and ready for effective visualization and analysis.

Original Dataset Source:

https://www.kaggle.com/code/farzadnekouei/flight-data-eda-to-preprocessing/input