AirFly Insights – Data Visualization and Analysis of Airline Operations

Internship Report

Submitted by: *Yuvadharanie M*
For: *Infosys Springboard 6.0 Internship*

Introduction

The aviation industry is one of the most dynamic and data-driven sectors in the world. Airlines operate millions of flights annually, and each flight generates a large amount of data — including departure and arrival times, delays, cancellations, weather impacts, and operational performance. Analyzing this data effectively can help airlines improve efficiency, customer experience, and scheduling accuracy.

The **AirFly Insights** internship project was designed to analyze large-scale flight data and uncover meaningful patterns through **data visualization and analytical storytelling**. This seven-week internship provided hands-on exposure to data cleaning, feature engineering, exploratory data analysis (EDA), and dashboard creation.

All analyses were performed using **Google Colab** with **Python** as the primary language, along with data visualization libraries such as **Matplotlib**, **Seaborn**, **Plotly**, and **Folium**. A final interactive **Power BI dashboard** was also created to summarize key findings.

The internship helped me develop both **technical expertise** and **analytical reasoning**, strengthening my ability to translate raw data into actionable insights.

Executive Summary

The AirFly Insights project was carried out in a structured weekly manner, starting from **data understanding** and progressing through **cleaning, transformation, visualization**, and finally **dashboard development**.

The dataset used was the **Kaggle Airlines Flight Dataset**, containing over **480,000 flight records**. It included information such as flight dates, airline codes, origin and destination airports, delay durations, and cancellation reasons.

Each week's tasks built upon the previous one:

- **Weeks 1–2:** Data preparation, cleaning, and feature engineering.

- **Weeks 3–4:** Exploratory and delay cause analysis.
- **Week 5:** Airport and route-level study.
- **Week 6:** Seasonal and cancellation behavior.
- **Week 7:** Compilation, dashboard design, and reporting.

By the end of the project, I was able to create a **complete data analysis workflow**, build **meaningful visualizations**, and design a **Power BI dashboard** summarizing airline operational performance.

Week 1 – Data Acquisition and Understanding

The internship began with setting up the development environment in **Google Colab** and understanding the dataset's structure. The **Kaggle Airlines dataset**, containing 484,000+ records, was imported using pandas. I explored its shape, columns, and data types to get an overview of available information.

The dataset contained columns like:

- **FlightDate**
- **Airline and Flight Number**
- **Origin and Destination Airports**
- **Departure/Arrival Delays**
- **Cancellation Codes**
- **Delay Causes (Carrier, Weather, NAS, Security)**

I checked for **missing values, data inconsistencies**, and **duplicates** to assess data quality. I also analyzed the memory usage of each column and applied optimization techniques such as changing data types (e.g., int64 → int32) to make computations more efficient.

Key Tasks

- Imported dataset from Kaggle into Google Colab.
- Explored data using df.info(), df.describe(), and df.isnull().sum().
- Identified null values and duplicates for cleaning.
- Performed basic profiling for data understanding.

This week laid the foundation for all upcoming analysis, ensuring a clear understanding of the raw data and its limitations.

Week 2 – Data Cleaning and Feature Engineering

In Week 2, the focus shifted to preparing the dataset for meaningful analysis. Data cleaning was performed to handle missing values and inconsistent formats. Null values in delay columns were replaced with zeros, and missing categorical entries were filled with appropriate placeholders.

After cleaning, I implemented **feature engineering** — the process of creating new variables that enhance analytical capability. Key transformations included:

- **Creating 'Route'** as a combination of Origin–Destination.
- **Extracting Month and Day of Week** from the flight date to analyze time-based trends.
- **Converting time fields** (stored in HHMM format) to Python datetime objects for more precise calculations.
- **Adding an 'Hour' column** to study time-of-day patterns in delays or cancellations.

The cleaned and enriched dataset was exported as a new CSV file for future visualization.

Key Tasks

- Cleaned missing and inconsistent entries.
- Performed data type conversions for accuracy.
- Engineered new analytical features.
- Validated transformed data and stored cleaned dataset.

This week made the data **structured, consistent, and ready for analysis**, setting the stage for deeper insights in upcoming phases.

Week 3 – Exploratory Data Visualization

Week 3 introduced **data visualization** — the most crucial part of understanding and communicating insights. Using **matplotlib** and **seaborn**, I performed both univariate and bivariate analyses.

Univariate analysis focused on single-column exploration:

- The **most frequent airlines** operating flights.
- The **busiest months** for flight operations.
- The **distribution of flights** across routes.

Bivariate analysis compared relationships between two columns, such as:

- Delay durations by airline.
- Flight frequency versus month.
- Correlation between route and average delay.

Key Visuals

- Bar charts for airline-wise flight volume.
- Line plots for monthly trends.
- Box plots for delay comparison across airlines.

Observations

- Certain airlines maintained more consistent timings, showing higher reliability.
- Some routes showed recurring delay patterns, possibly due to traffic congestion or weather conditions.

By the end of the week, I had a strong understanding of the dataset's structure and variability, which guided later stages of the project.

Week 4 – Delay Cause Analysis and Time-Based Trends

This week's focus was on **understanding the causes of delays** and their **time-based patterns**. The dataset had four main delay categories:

1. **Carrier Delay** (airline-related)
2. **Weather Delay**
3. **NAS Delay** (air traffic control)
4. **Security Delay**

I grouped the data by airline and delay type to identify which carriers or regions were most affected by certain causes.

Time-based visualizations such as **line charts** and **heatmaps** were created to identify:

- Peak hours of delay occurrences.
- Days of the week with maximum delays.
- Average delay duration patterns.

Key Insights

- Carrier and weather delays contributed most to total delay time.
- Early morning flights had fewer delays compared to evening ones.
- Delays were more frequent during weekends and festive travel periods.

Tools Used

- pandas.groupby() for aggregations.
- seaborn.barplot() and sns.heatmap() for visual patterns.

This analysis revealed **operational inefficiencies** and helped connect delay trends with real-world causes like weather conditions and traffic congestion.

Week 5 – Route and Airport-Level Analysis

In Week 5, the analysis expanded to a **spatial and route-based level**. Each airport's average delay was calculated, and top routes were ranked by flight frequency and delay severity.

Using **Plotly** and **Folium**, I created **interactive heatmaps** and **geographical route maps** to visualize:

- The **top 10 busiest routes**.
- Airports with the **highest average delays**.
- Regional trends in flight punctuality.

Findings

- Some metropolitan airports showed longer delays due to heavy air traffic.
- Specific routes connecting high-traffic cities had the most congestion.
- Regional patterns helped link geographical areas to delay behaviors.

These interactive maps added a **visual storytelling layer**, helping interpret data geographically rather than numerically.

Week 6 – Seasonal and Cancellation Analysis

This week focused on analyzing **seasonal flight behavior** and **cancellation patterns**. Monthly and seasonal trends revealed how operational performance changed throughout the year.

I also explored **cancellation reasons** (Carrier, Weather, NAS, and Security). Using **bar charts and pie charts**, I identified:

- Winter months had increased weather-related cancellations.
- Carrier-related cancellations dominated during peak operational months.
- Security cancellations were minimal but consistent.

Analytical Steps

- Grouped data by month and cancellation reason.
- Calculated cancellation percentages per category.
- Visualized results for comparative analysis.

These insights illustrated how **seasonal factors influence flight reliability**, helping airlines plan resource allocation accordingly.

Week 7 – Dashboard Development and Final Compilation

The final week was dedicated to organizing all insights and presenting them in a **professional visual dashboard** using **Power BI**.

I refined every plot with appropriate titles, color schemes, legends, and labels to maintain visual consistency. The Power BI dashboard combined all major insights — including delay patterns, route congestion, seasonal cancellations, and airline performance metrics.

Dashboard Highlights

- Monthly delay trends (line chart).
- Cancellation breakdown by cause (pie chart).
- Top routes by average delay (bar chart).
- Flight volume comparison by airline (column chart).

The final deliverable was a **well-structured analytical report** that summarized seven weeks of exploration and visualization.

Tools and Technologies Used

Category	**Tools / Technologies**
Development Platform	Google Colab, Data Bricks
Programming Language	Python
Libraries Used	pandas, numpy, matplotlib, seaborn, plotly, folium
Dashboard Tool	Power BI
Dataset	Kaggle Airlines Flight Data (~484,000 records)
Documentation	Jupyter Notebook, Markdown, Final Report

Conclusion and Learnings

The AirFly Insights internship was a **complete data analytics journey**, starting from raw data handling to professional visualization. It provided a real-world understanding of how airline data can reveal operational trends and inefficiencies.

Through this internship, I learned to:

- Handle and clean large, complex datasets efficiently.
- Apply **feature engineering** for enhanced analytical value.
- Visualize insights using different chart types effectively.
- Use **Power BI** for creating interactive dashboards.
- Communicate technical findings through clear visual storytelling.

This experience not only improved my technical proficiency in Python but also enhanced my **analytical thinking, data interpretation, and presentation skills**. I now feel more confident in managing end-to-end data projects — from data preparation to actionable insights — a skill set essential for careers in **data analytics, business intelligence, and visualization**.