Internship Report

Data Visualization Internship

Submitted by

Abhinand P. K

Internship at

Infosys Springboard

Date: November 14, 2025

Contents

1. Introduction

This report presents the work carried out during my internship at Infosys Springboard in the domain of Data Visualization. The project, titled AirFly Insights, focuses on analyzing large-scale airline flight data to understand operational performance across airlines and airports. The primary objective is to explore delay patterns, flight cancellations, and route-level trends using effective data preprocessing and visualization techniques.

Through tools such as Python and Power BI, the project aims to extract meaningful insights that highlight the key factors influencing airline operations. These insights can support data-driven decision-making, improve understanding of flight behavior, and contribute to optimizing overall airline performance.

2. 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.

3. Software Requirements / Tools and Technologies

This project makes use of a range of data processing, visualization, and documentation tools to analyze large-scale airline flight data efficiently. The choice of tools ensures scalability, clarity of insights, and ease of presenting results.

Tools and Technologies

- **Data Handling: pandas, numpy**
  These Python libraries were used for data loading, preprocessing, cleaning, and manipulation. `pandas` provides powerful data structures like DataFrames that simplify handling large CSV files, while `numpy` offers optimized numerical operations essential for efficient computations on large datasets.

- **Visualization: matplotlib, seaborn, plotly**
  These libraries were used to generate meaningful graphs and visual summaries of flight delays, cancellations, and route patterns. `matplotlib` enables fundamental plotting, `seaborn` enhances statistical visualization with cleaner aesthetics, and `plotly` provides interactive charts that allow better exploration of trends and patterns.

- **Optional Dashboard: Power BI**
  Power BI was used optionally to create interactive dashboards that combine multiple visualizations into a unified reporting interface. This helps stakeholders explore data dynamically and obtain real-time insights from the processed dataset.

- **Documentation: Jupyter Notebook, PDF, GitHub**
  `Jupyter Notebook` was used for step-by-step analysis combining code, visual outputs, and explanations. Final reports were compiled into PDF format for submission. GitHub served as a version-control and repository platform to maintain the project code, visualizations, and documentation systematically.

4. Milestone 1: Data Foundation and Cleaning

Milestone 1 focused on establishing a strong foundation for the project by setting up the dataset, understanding its structure, and preparing it for advanced analysis. This phase ensured that the data was clean, well-structured, and optimized for efficient processing in later stages of the project. The tasks completed in this milestone are outlined below.

Week 1: Project Initialization and Dataset Setup

During the first week, the primary objective was to become familiar with the dataset and define the workflow for the entire project.

- **Define goals, KPIs, and workflow:** Key project goals and performance indicators were identified, including delay pattern detection, cancellation analysis, and route-based performance metrics. A clear workflow was established covering data acquisition, preprocessing, analysis, visualization, and reporting.
- **Load CSVs using pandas:** The airline dataset was imported using the `pandas` library. Due to its large size, efficient data-loading strategies were applied to ensure smooth processing.
- **Explore schema, types, size, and nulls:** A detailed exploratory check of the dataset structure was performed to understand column types, data distribution, dimensionality, and the extent of missing values. This step helped identify key preprocessing requirements.
- **Perform sampling and memory optimizations:** Since the dataset contained millions of records, sampling techniques and memory-efficient data types were used to reduce computational load. Unnecessary columns were examined, and datatype optimization was carried out to improve performance.

Week 2: Preprocessing and Feature Engineering

The second week focused on cleaning the dataset and creating additional features needed for deeper analysis.

- **Handle nulls in delay and cancellation columns:** Missing values in key delay-related attributes were treated using either imputation methods or removal based on their relevance, ensuring the reliability of analysis.
- **Create derived features: Month, Day of Week, Hour, Route:** New attributes were engineered from existing datetime and flight information. These derived features helped identify time-based trends and route-specific behaviors.
- **Format datetime columns:** All timestamp fields were converted into proper datetime formats to enable accurate extraction of date components and smooth temporal analysis.
- **Save preprocessed data for fast reuse:** After cleaning and feature engineering, the processed dataset was saved to avoid repeating heavy preprocessing steps. This improved efficiency during visualization and analysis.

Milestone 1 provided a clean, structured, and optimized dataset, forming the basis for meaningful visual exploration in subsequent milestones.

5. Milestone 2: Visual Exploration and Delay Trends

Milestone 2 focused on visually exploring the dataset to uncover early operational patterns and delay behaviors across airlines, airports, and time periods. This stage involved generating multiple univariate and bivariate visualizations and interpreting the resulting insights. The tasks were distributed across two weeks: general visual exploration followed by an in-depth delay analysis.

Week 3: Univariate and Bivariate Visual Analysis

Week 3 involved developing a wide range of visualizations to understand the distribution and relationships within the airline dataset.

- **Top airlines, routes, and busiest months:** Visualizations were created to identify the most active airlines, frequently used routes, and months with the highest volume of flights. These observations helped establish an initial understanding of operational intensity.
- **Flight distribution by day, time, and airport:** The dataset was analyzed to observe how flights are distributed across different days of the week, hours of the day, and various airports. This provided insights into peak operational periods and airport traffic patterns.
- **Use of bar charts, histograms, boxplots, and line plots:** A variety of plotting techniques were applied to study flight counts, delay distributions, and time-based trends.
  - Bar charts to compare activity across airlines, airports, and months
  - Histograms to understand the frequency distribution of delays
  - Boxplots to detect variability and outliers in delay values
  - Line plots to capture temporal variations in flight operations

These visual tools provided an essential understanding of both individual attributes and relationships between different flight factors.

Week 4: Delay Analysis — Airline and Weather

Week 4 focused specifically on understanding delay behaviors by examining different delay types and their distribution across airlines, airports, and time periods.

- **Delay causes across airlines:** Comparative analysis revealed how different airline carriers performed with respect to delays. This helped identify carriers experiencing frequent operational or scheduling challenges.
- **Carrier, weather, and NAS delays:** The dataset includes multiple delay categories, such as carrier delays, weather delays, and National Airspace System (NAS) delays. Visual exploration of these categories provided clarity on which delay types most frequently contributed to total delay time.
- **Delays by time of day and airport:** Visualizations were generated to show how delay intensity changes throughout the day and varies across airports. This helped in identifying delay-prone time periods and airports that frequently encounter operational disruptions.

Overall, Milestone 2 provided a strong visual understanding of core operational patterns and delay behaviors. The insights gained here laid the groundwork for more advanced route-level, seasonal, and cancellation analysis in subsequent milestones.

6. Milestone 3: Route, Cancellation, and Seasonal Insights

Milestone 3 focused on investigating route-level performance, airport congestion patterns, and seasonal variations that influence flight operations. This phase also explored cancellation trends and the underlying factors contributing to operational disruptions. The tasks were divided into two key components: route and airport-level analysis, followed by seasonal and cancellation-focused exploration.

Week 5: Route and Airport-Level Analysis

During Week 5, a detailed analysis of flight routes and airport activity was carried out to understand operational behavior across the airline network.

- **Top 10 origin–destination pairs:** The dataset was examined to identify the top ten most frequently traveled origin–destination routes. This helped highlight high-traffic corridors and provided a foundation for evaluating route-specific delays and traffic intensity.
- **Delay heatmaps by airport and route:** Heatmaps were generated to visually represent delay levels across airports and routes. These visualizations clearly showed which airports experience the most congestion and which routes consistently encounter higher delays.
- **Maps showing busiest airports and average delays:** Geospatial maps were created to visualize major airports along with their corresponding average delay times. This provided a geographic understanding of flight density patterns and delay hotspots across the network.

Week 6: Seasonal and Cancellation Analysis

Week 6 involved exploring how cancellations and delays vary across seasons, as well as identifying the major categories contributing to flight cancellations.

- **Monthly cancellation trends:** Monthly analysis revealed noticeable fluctuations in cancellation patterns. Certain months showed higher cancellation frequencies, reflecting seasonal operational challenges.
- **Cancellation types: carrier, weather, security, NAS:** The cancellation data was broken down into categories such as carrier issues, weather disturbances, security events, and NAS (National Airspace System) constraints. Comparing these categories helped identify which factors were the most dominant contributors to flight cancellations.
- **Impact of holidays or winter months:** An examination of holiday seasons and winter months showed their influence on operational reliability. Weather-related disruptions and increased travel demand during holidays contributed to spikes in delays and cancellations.

Overall, Milestone 3 provided a comprehensive understanding of how route characteristics, airport activity, and seasonal factors affect flight delays and cancellations. This milestone highlighted recurring patterns and problem areas, offering deeper insight into the operational dynamics of airline networks.

7. Milestone 4: Report and Presentation

Milestone 4 focused on consolidating the visual analysis into a coherent report and presentation. This stage involved combining visualizations, preparing dashboards and slides, and documenting insights for final delivery. The tasks were carried out over two weeks: visual report/dashboard preparation followed by final documentation and presentation.

Week 7: Visual Report or Dashboard Preparation

During Week 7, the visualizations generated in previous milestones were organized into a coherent and interpretable format.

- **Combined plots into a coherent storyline:** Visuals from earlier analysis were sequenced logically to convey key operational patterns, delay trends, and insights effectively.
- **Prepared reports and Power BI dashboards:** Static documents (Markdown and slides) as well as interactive dashboards in Power BI were created to present the findings.
- **Ensured clarity in plots:** All visualizations were enhanced with appropriate labels, titles, legends, and axis descriptions to maximize interpretability.

Week 8: Documentation and Final Presentation

During Week 8, the final deliverables were completed and all insights were documented professionally.

- **Final report completed:** A comprehensive report (PDF) documenting all analysis, visualizations, and key insights was prepared.
- **Slide deck created:** A concise slide deck was built to communicate the analysis workflow, findings, and conclusions effectively.

- **Visual walkthrough recorded:** Key insights were summarized, and a guided walkthrough of visualizations was prepared to enhance understanding during presentations.

8. Design

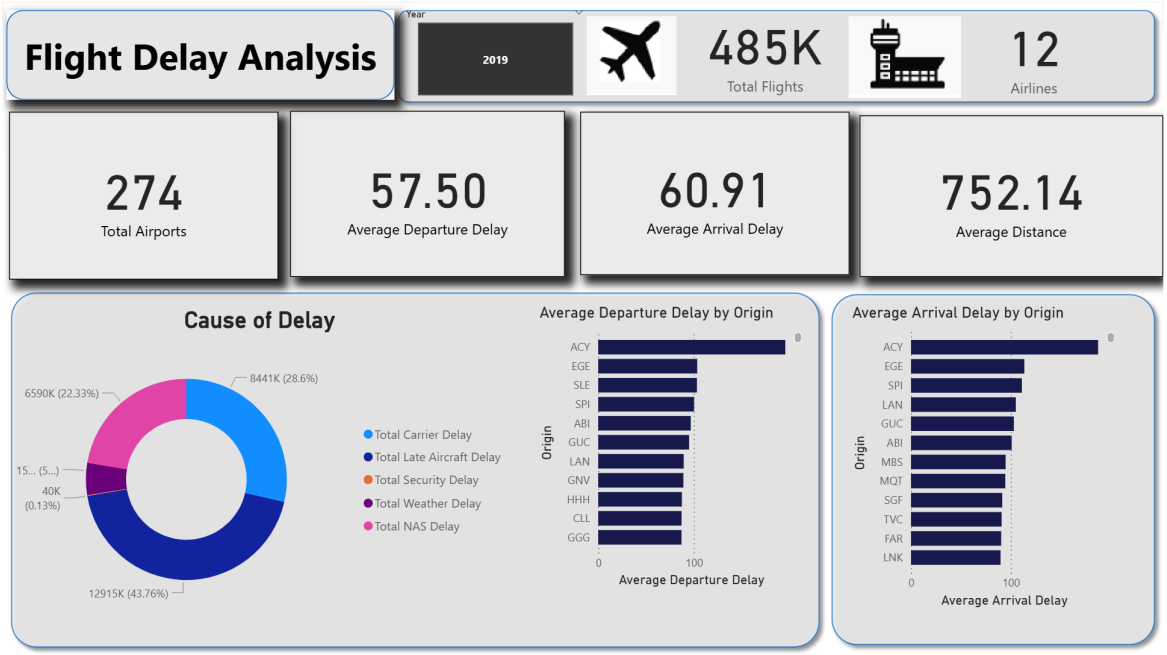

Figure 8.1: Final dashboard of flight analytics

Figure 8.2: Final dashboard of flight analytics

9. Conclusion

This internship provided a comprehensive exposure to data analytics and visualization, allowing me to apply theoretical knowledge to real-world datasets. Through the creation of dashboards and detailed visual analyses, I developed a deeper understanding of operational patterns, delay trends, and data-driven decision-making. The experience strengthened my technical skills in data manipulation, visualization, and interpretation, while also enhancing my problem-solving, analytical thinking, and professional reporting abilities. Overall, the internship has significantly contributed to my preparedness for a career in data analytics and business intelligence.