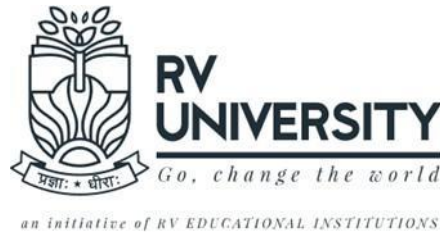


# RV UNIVERSITY, BENGALURU

## SCHOOL OF COMPUTER SCIENCE AND ENGINEERING



A Project Report On

### **Flight Data Analysis**

BSc (Honors)

In

School of Computer Science and Engineering

Submitted By

Team Member 01: 1RVU22BSC069 \_ PRABHAS BHAT

Team Member 02: 1RVU22BSC044 \_ KISHOR DESAI

**Data Visualization &  
Gen BI**

**Under the Guidance of**

**Vinod Kumar Raju**

Assistant Professor

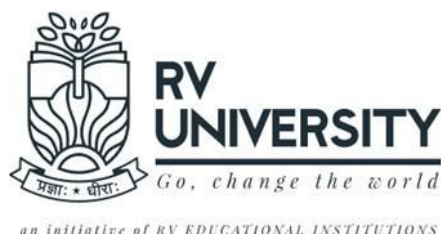
School of CSE

RV University, Bengaluru-

560059 2024-2025

# RV UNIVERSITY, BENGALURU-59

## SCHOOL OF COMPUTER SCIENCE AND ENGINEERING



### CERTIFICATE

Certified that the project work titled **Flight Data Analysis** is carried out by **Prabhas Bhat** (1RVU22BSC069), **Kishor Desai** (1RVU22BSC044), RV University, Bengaluru, **BSc (Hons) in the School of Computer Science and Engineering** of the RV University, Bengaluru during the year 2025-2026. It is certified that all corrections/suggestions indicated for the Internal Assessment have been incorporated in the project report. The Project report has been approved as it satisfies the academic requirements in respect of project work prescribed by the institution.

**Signature of Guide**

**External Viva:**

**Name of Examiners**

**Signature with Date**

1

2

**RV UNIVERSITY, BENGALURU-59**  
**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**DECLARATION**

**We**, Prabhas Bhat and Kishor Desai students of seventh semester BSc(Hons), SoCSE, RV University, Bengaluru, hereby declare that the project titled '**Flight Data Analysis**' has been carried out by us and submitted in partial fulfillment of **Bachelor of Science (Hons) in School of Computer Science and Engineering** during the year 2025-26.

Further, we declare that the content of the report has not been submitted previously by anybody or to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RV University will be the property of RV University, Bengaluru, and we will be one of the authors of the same.

Place: Bengaluru

Date:

**Name**

**Signature**

1. Prabhas Bhat (1RVU22BSC069)

2. Kishor Desai (1RVU22BSC044)

## ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project.

First, we take this opportunity to express our sincere gratitude to the School of Computer Science and Engineering, RV University, for providing us with a great opportunity to pursue our bachelor's degree in this institution.

A special thanks to our Program Director, **Dr. Lokanayaki K**, and Dean - **Dr.Shobha G**, for their continuous support and providing the necessary facilities with guidance to carry out mini project work.

We would like to thank our guide, Prof, **Vinod Kumar Raju, Assistant Professor**, School of Computer Science and Engineering, RV University, for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

*Signature of Student*

USN:

Name:

## Abstract

This project presents a data visualization–driven analysis of the **2015 U.S. Flights Delay Dataset** published by the U.S. Department of Transportation. The dataset consists of a large-scale fact table with **5.8 million flight records** and over **50 attributes** when integrated with airline and airport dimension tables, amounting to approximately **1.3 GB of data**. The study aims to analyze flight delays, cancellations, and traffic patterns across airlines, airports, time periods, and geographic regions in the United States using descriptive analytics and interactive visualizations.

The analysis addresses airline-specific, airport-specific, and temporal business questions, including identifying frequently delayed airlines, busiest and most delay-prone airports, seasonal delay trends, cancellation behavior, and aircraft-level delay contributions. Geographic insights are derived by examining flight operations across U.S. states, while airport-level analysis evaluates peak traffic periods and delay variations by month and airline. By converting large volumes of aviation data into actionable insights, the project demonstrates how data visualization and analytics can support informed decision-making to improve operational efficiency and passenger experience.

## Table of Contents:

	Page No
Abstract	v
List of Tables	vi
List of Figures	vii

Page Title	Page No
<b>Chapter 1: Introduction</b>	1
1.1. General Introduction	1
1.2. Literature Survey	2
1.3 Problem Statement / Objectives	2
<b>Chapter 2: System Design</b>	3
2.1 ER Schema Diagram	3
<b>Chapter 3: Software Requirements</b>	4
3.1. Functional Requirements	4
3.2. Non-Functional Requirements	4
3.3. Hardware Requirements	4
3.4. Software Requirements	4
<b>Chapter 4: Implementation and Testing</b>	5
4.1. Data Cleaning	5
4.2 Calculated Measures	7
4.3 Business Questions	8
<b>Chapter 5: Results and Discussion</b>	9
5.1. Results	9

<b>Chapter 6: Conclusion and Future Work</b>	11
<b>References [ APA format ]</b>	-
<b>Appendices</b>	12
<b>Appendix 1: Screenshots</b>	12
<b>Appendix 2: Source code</b>	-
<b>Appendix 3: Plagiarism details</b>	-

<b>List of Figures</b>		
<b>Figure No</b>	<b>Figure Name</b>	<b>Page No</b>
1.	ER Diagram	3
2.		

# 1. Introduction

## 1.1 General Introduction

The aviation industry plays a vital role in global connectivity, economic growth, and mobility. With millions of flights operating every year, even minor inefficiencies can lead to significant delays, cancellations, and passenger inconvenience. Flight delays not only impact customer satisfaction but also result in increased operational costs for airlines and congestion at airports. Hence, analyzing historical flight data is essential to understand delay patterns, identify operational bottlenecks, and support data-driven decision-making.

This project focuses on the **2015 U.S. Flights Delay Dataset** published by the U.S. Department of Transportation. The dataset contains detailed information on flight operations, including schedules, actual departure and arrival times, delays, cancellations, and their causes. By combining a large fact table with airline and airport dimension tables, the project enables comprehensive analysis across airlines, airports, time periods, and geographic regions. Data visualization techniques are used to transform complex, large-scale data into meaningful insights that are easy to interpret.

## 1.2 Literature survey

Previous studies in aviation analytics emphasize the importance of structured flight data for understanding operational performance. Most publicly available airline datasets follow a **star schema** structure, consisting of a central fact table (flights) linked to multiple dimension tables (airlines and airports). This design enables efficient querying and aggregation for large datasets.

These are some key details of the dataset and its important terminologies:

- Key columns in such datasets include **IATA codes**, which are standardized identifiers assigned by the International Air Transport Association (IATA).
- Airline IATA codes are two-letter codes (e.g., AA for American Airlines), while airport IATA codes are three-letter codes (e.g., JFK for John F. Kennedy International Airport). These codes ensure consistency and interoperability across aviation systems.
- Temporal columns such as YEAR, MONTH, DAY, and DAY\_OF\_WEEK are commonly used in time-series analysis to study seasonal trends and daily patterns.
- Delay-related columns like DEPARTURE\_DELAY, ARRIVAL\_DELAY, and categorized



delay causes (weather, airline, security, air system) are frequently analyzed in prior research to identify the root causes of inefficiencies.

- Geographic attributes such as STATE, LATITUDE, and LONGITUDE support spatial analysis and mapping of flight operations.

### **1.3 Problem Statement & Objectives**

#### **Problem Statement:**

Despite the availability of large volumes of flight data, extracting meaningful insights remains challenging due to data size, complexity, and dimensionality. Airlines and airports require clear, actionable metrics to understand where delays occur most frequently, which entities contribute to them, and how these patterns change over time and location.

The primary objectives of this project are:

- To analyze and visualize flight delay and cancellation patterns across airlines and airports.
- To identify the most frequently delayed airlines and the busiest and most delay-prone airports.
- To study temporal trends such as monthly and daily variations in delays and cancellations.
- To evaluate airline-specific performance in terms of flight volume, cancellation rates, and delay causes.
- To derive geographic insights by analyzing flight operations across U.S. states.

## 2.System Design

### 2.1 ER Schema Diagram

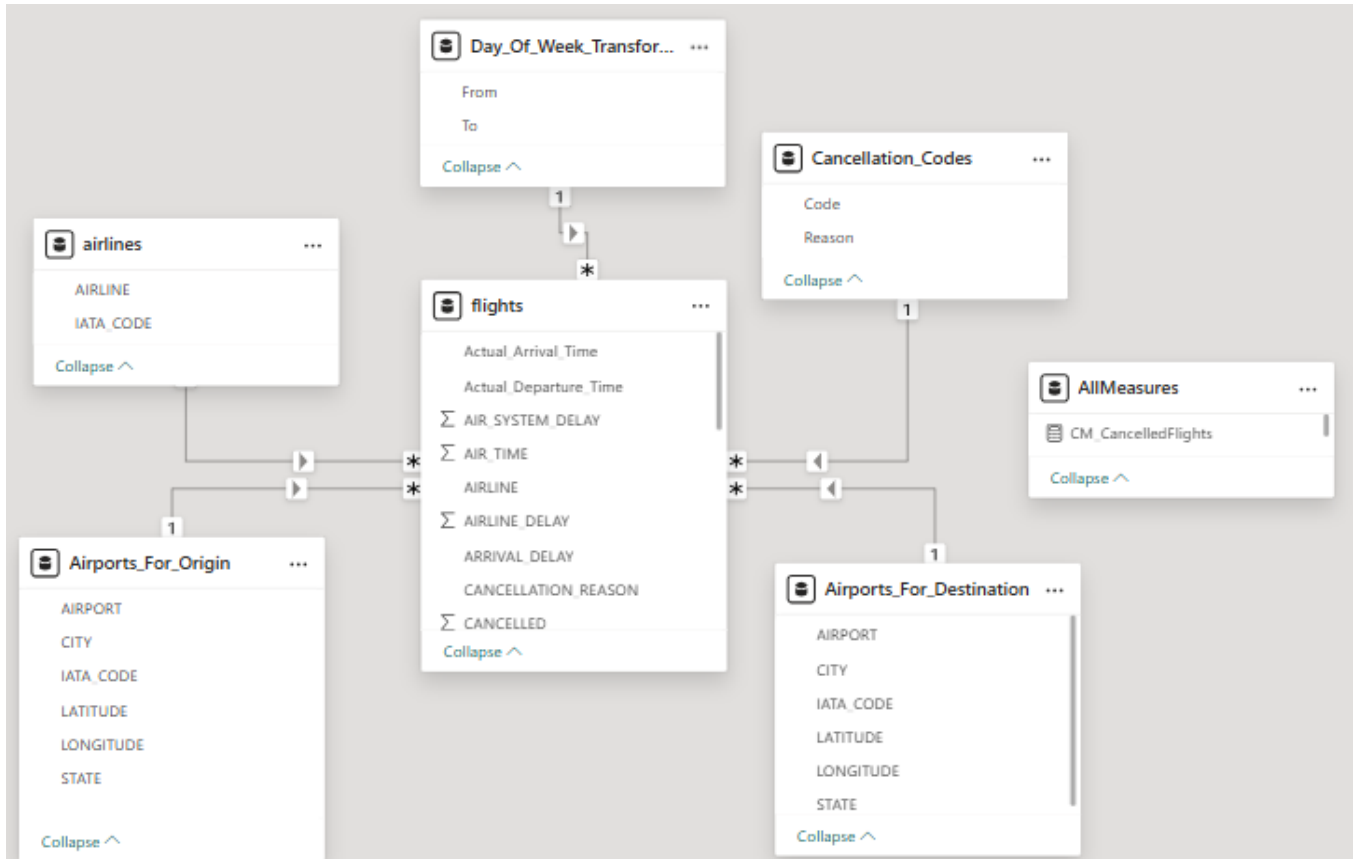


Figure-1: ER diagram

The ER diagram represents the data model for analyzing the 2015 Flights Delay Dataset. It consists of 1 fact table and 5 dimension tables.

#### Fact Table: Flights

- Central table that stores detailed flight-level data.
- Contains actual departure and arrival times, delay durations, cancellation status, and delay causes.
- Acts as the main source for all calculations and analysis.
- Connected to all dimension tables using keys such as airline code, airport codes, day of week, and cancellation reason.

#### Dimension Table-1: Airlines

- Stores airline-related information.
- Uses **IATA airline code** (two-letter code) to uniquely identify each airline.
- Enables analysis of delays, cancellations, and flight volume by airline.

#### Dimension Table-2: Airports\_For\_Origin

- Contains details of departure (origin) airports.

- Includes airport name, city, state, latitude, and longitude.
- Used to analyze departure delays, traffic volume, and geographic patterns.

**Dimension Table-3: Airports\_For\_Destination**

- Contains details of arrival (destination) airports.
- Includes airport name, city, state, latitude, and longitude.
- Used to analyze arrival delays and destination-based performance.

**Dimension Table-4: Day\_Of\_Week\_Transformation**

- Converts numeric day values into meaningful day names.
- Helps in analyzing flight trends across weekdays and weekends.

**Dimension Table-5: Cancellation\_Codes**

- Maps cancellation codes (A, B, C, D) to readable cancellation reasons.
- Helps understand why flights were cancelled (weather, airline, security, etc.).

**Measures Table: AllMeasures**

- Stores calculated metrics such as total cancelled flights.
- Keeps business logic separate from raw data for cleaner analysis.

**Relationships**

- All dimension tables have a **one-to-many** relationship with the Flights fact table.
- This structure follows a **star schema**, optimized for reporting and data visualization.

## 3. Software Requirements

### 3.1 Functional Requirements

The system must be capable of importing and processing large-scale flight data efficiently. It should support data cleaning, transformation, and integration of multiple tables such as flights, airlines, and airports. The system must allow users to perform analytical queries to identify delays, cancellations, flight volumes, and performance metrics across airlines, airports, time periods, and geographic regions. Interactive data visualizations such as charts, maps, and dashboards should be generated to answer predefined business questions. Additionally, the system should support filtering and slicing of data using parameters like airline, airport, month, and day of the week.

### 3.2 Non- Functional Requirements

Despite the availability of large volumes of flight data, extracting meaningful insights remains challenging due to data size, complexity, and dimensionality. The system should be scalable to handle large datasets exceeding millions of records without significant performance degradation. It must provide acceptable query response times for aggregation and visualization tasks. The system should be reliable, ensuring accurate calculations and consistent results across repeated analyses. Usability is important, and visualizations should be intuitive and easy to interpret for non-technical users. The system should also be maintainable, allowing future enhancements such as adding new datasets or metrics with minimal changes.

### **3.3 Hardware Requirements**

The analysis requires a system capable of handling large datasets and performing complex aggregations. A minimum of 8 GB RAM is recommended, with 16 GB RAM preferred for smoother performance. A multi-core processor (Intel i5 / AMD Ryzen 5 or higher) is suggested to support data processing and visualization tasks. Adequate storage (at least 20 GB free disk space) is required to store raw datasets, processed files, and project artifacts. A stable internet connection is necessary for dataset download and cloud-based tools, if used.

### **3.4 Software Requirements**

The project requires a modern operating system such as Windows, Linux, or macOS. Data analysis and visualization tools like Power BI / Tableau are used for dashboard creation and interactive reporting. Python (with libraries such as Pandas and NumPy) may be used for data preprocessing and transformation. A database or data model environment is required to establish relationships between fact and dimension tables. Optional tools include cloud platforms or big data frameworks if scalability testing is performed.

## **4.Implementation and Testing**

### **4.1 Data Cleaning**

To ensure data accuracy, consistency, and suitability for analysis, multiple data cleaning and transformation steps were performed across all tables in the dataset using Power Query. These steps helped standardize formats, handle missing or inconsistent values, and prepare the data for modeling and visualization.

#### **Airlines Table**

- Promoted the first row to column headers to correctly define airline attributes.
- Ensured consistency of airline IATA codes for establishing relationships with the Flights fact table.

#### **Airports (Origin and Destination) Tables**

- Promoted the first row as column headers.
- Appended a custom query to add an additional row representing non-IATA airports, assigned with a placeholder code 'ZZZ', to handle flights with missing or invalid airport

codes.

- Removed the *Country* column, as all airport records in the dataset belong to the United States.
- Maintained geographic attributes such as city, state, latitude, and longitude for spatial analysis.

### **Flights Table**

- Promoted the first row as column headers.
- Created a custom Date column by combining *Year*, *Month*, and *Day* columns.
- Removed the *Year* column since all records belong to the year 2015.
- Removed *Month* and *Day* columns as their information was already captured in the Date column.
- Merged the Flights table with the Day of Week Transformer table to convert numeric day values into meaningful textual representations.
- Removed the original numeric *Day of Week* column after transformation.
- Created a custom column to replace numeric values in the Origin Airport field with the placeholder code 'ZZZ' for non-IATA entries, and removed the original column.
- Added a new custom column for Destination Airport using the same logic and removed the old column containing numeric values.
- Converted scheduled and actual time columns from numeric HHMM format into proper time data type using custom columns.
- Removed the original numeric time columns after conversion.
- Reordered columns for better readability and logical grouping.

### **Day of Week Transformer Table**

- Created a custom transformation table to map numeric day values to textual day names.

### **New Row Airports Query**

- A separate custom query was created to assign the placeholder code 'ZZZ' to non-IATA airports.
- This ensured referential integrity between the Flights table and Airport dimension tables without data loss.

## 4.2 Calculated Measures

To support analytical insights and answer the defined business questions, several calculated measures were created using DAX. These measures enable aggregation, comparison, and percentage-based analysis of flight operations, delays, and cancellations.

- Total Flights (CM\_TotalFlights)

This measure calculates the total number of flights in the dataset and serves as the base metric for most percentage calculations.

Formula: COUNTROWS(flights)

- Cancelled Flights (CM\_CancelledFlights)

This measure computes the total number of flights that were cancelled.

Formula: COUNTROWS(FILTER(flights, flights[CANCELLED] = 1))

- Cancelled Flights Percentage (CM\_CancelledFlights%)

This measure represents the proportion of cancelled flights relative to the total number of flights.

Formula: DIVIDE([CM\_CancelledFlights], [CM\_TotalFlights], 0)

- Delayed Arrival Flights (CM\_DelayedArrivalFlights)

Calculates the number of flights that experienced a positive arrival delay.

Formula: COUNTROWS(FILTER(flights, flights[ARRIVAL\_DELAY] > 0))

- Delayed Departure Flights (CM\_DelayedDepartureFlights)

Calculates the number of flights that experienced a positive departure delay.

Formula: COUNTROWS(FILTER(flights, flights[DEPARTURE\_DELAY] > 0))

- Total Delayed Flights (CM\_TotalDelayedFlights)

Counts flights that were delayed either at departure or arrival.

Formula: COUNTROWS(FILTER(flights, flights[ARRIVAL\_DELAY] > 0 || flights[DEPARTURE\_DELAY] > 0))

- Delayed Flights Percentage (CM\_DelayedFlights%)

Indicates the proportion of delayed flights compared to the total flights operated.

Formula: DIVIDE([CM\_TotalDelayedFlights], [CM\_TotalFlights], 0)

- Total Flight Delay (CM\_TotalFlightDelay)

Computes the cumulative delay time by summing arrival and departure delays across all flights.

Formula: SUMX(flights, flights[ARRIVAL\_DELAY]) + SUMX(flights, flights[DEPARTURE\_DELAY])

### 4.3 Business Questions

The business problems identified for this project were addressed using a structured set of dashboards, each designed to answer a specific category of questions. In total, five dashboards were created: a General Overview dashboard, an Airline Overview dashboard, an Airline Detailed dashboard, a Departure Airport dashboard, and an Arrival Airport dashboard. Together, these dashboards provide airline-level, airport-level, temporal, and operational insights.

1. The **General Overview dashboard** addresses high-level questions related to overall delay patterns in the aviation system. It identifies airlines with frequent delays, highlights the busiest airports based on take-offs and landings, determines airports with the highest number of delayed flights, and analyzes monthly trends to identify periods with the least and most delays.
2. The **Airline Overview and Airline Detailed dashboards** focus on airline-specific performance. These dashboards answer questions related to the total number of flights operated by an airline, overall delay and cancellation rates, breakdown of cancellation reasons, monthly delay behavior, and identification of operational inefficiencies linked to specific airlines or aircraft. Filters allow comparison across different time periods for deeper analysis.
3. The **Departure Airport dashboard** is designed to analyze departure-side operations at airports. It answers questions related to peak traffic times during the day, airlines contributing most to departure delays, month-wise variation in departure delays, and destinations associated with higher departure delays.
4. The **Arrival Airport dashboard** complements the departure analysis by focusing on arrival-side operations. It answers questions regarding landing volumes across time, airlines causing higher arrival delays, variation of arrival delays across months, and identification of airports with higher average arrival delays.

We made 5 separate dashboards to answer these business questions in particular:

1. The **first business question (BQ-1)** focuses on identifying **which airlines experience delays most frequently**. A dedicated dashboard was created to compare delay performance across airlines using aggregated delay metrics. This dashboard enables quick identification of airlines with consistently higher delay rates, supporting comparative performance evaluation at an industry level.
2. The **second business question (BQ-2)** addresses the identification of the **Top 5 airports by number of take-offs and landings**. A separate dashboard was designed to highlight airport traffic volume from both departure and arrival perspectives. This analysis helps determine the busiest airports and provides insight into traffic concentration within the aviation

network.

3. The **third business question (BQ-3)** focuses on determining the **Top 5 airports with the highest number of delayed flights**. A dedicated dashboard analyzes delay counts separately for arrivals and departures, allowing identification of airports that are most affected by operational inefficiencies or congestion-related delays.
4. The **fourth business question (BQ-4)** analyzes **monthly delay behavior** to identify periods with the **least and most number of delays**. A standalone dashboard was created to study temporal patterns across the calendar year, helping uncover seasonal trends and peak delay periods in flight operations.
5. The **fifth business question (BQ-5)** examines **aircraft-level delay contribution** for a selected airline. A dedicated dashboard allows users to filter by airline and analyze individual **aircraft tail numbers**. This dashboard highlights both the **frequency of delays** and the **total delay duration (in minutes)** caused by specific aircraft, enabling deeper operational insights and identification of high-impact assets.

## 5.Results and Discussion

### 5.1 Results

The results highlight airline performance, airport congestion, seasonal trends, and aircraft-level operational issues, providing a comprehensive understanding of delay behavior across the U.S. aviation system.

1. The overall analysis indicates that the dataset contains approximately **5.82 million flights**, of which **46.47% experienced delays**, while **1.54% were cancelled**. The cumulative delay time exceeds **79 million minutes**, emphasizing the scale of operational inefficiencies present in the system. Airline-wise analysis shows that high-volume carriers such as **Southwest Airlines, Delta Air Lines, and American Airlines** operate the largest number of flights, which naturally exposes them to higher absolute delay counts.
2. Airport-level results demonstrate that **Hartsfield–Jackson Atlanta International Airport, Chicago O’Hare International Airport, and Dallas/Fort Worth International Airport** rank among the busiest airports in terms of both take-offs and landings. These same airports also appear consistently in the list of airports with the highest number of delayed arrivals and



departures. This strong overlap suggests that **airport congestion is a major contributing factor to delays**, particularly at high-traffic hubs where runway and gate capacity are under constant pressure.

3. Temporal analysis reveals clear **seasonal delay patterns**. The number of delayed flights peaks during **summer months (June and July)** and remains high through **late winter and early spring**, while relatively fewer delays are observed during early autumn months. Cancellation analysis further shows that **weather-related issues account for more than half of all cancellations (54.35%)**, followed by airline-related causes (28.11%) and air traffic control constraints (17.52%). This highlights the dominant role of external environmental factors in flight disruptions, particularly during adverse seasonal conditions.
4. Aircraft-level analysis provides deeper operational insight by identifying specific **tail numbers** that contribute disproportionately to delays.
5. Further airport-specific analysis of arrivals and departures reveals that **departure delays are generally higher than arrival delays**, with average take-off delays exceeding average landing delays. Certain airports such as **Orlando International Airport, Chicago O'Hare, and Honolulu International Airport** exhibit higher average departure delays, while smaller or regional airports show higher arrival delays. Additionally, peak flight volumes are concentrated during **morning and evening hours**, which aligns with business travel.

Overall, the results demonstrate that flight delays are influenced by a combination of **airline operations, airport congestion, seasonal factors, and aircraft-level characteristics**.

## 6.Conclusion and Future work

This project successfully analyzed the **2015 U.S. Flights Delay Dataset** using data visualization and business intelligence techniques to uncover meaningful insights into airline performance, airport congestion, and temporal delay patterns. By integrating large-scale flight data with airline and airport dimensions, the study identified key contributors to delays and cancellations, including high-traffic airports, seasonal effects, and specific aircraft within airline fleets. The dashboards developed in this project transformed complex aviation data into intuitive insights, enabling effective comparison and performance evaluation across airlines and airports.

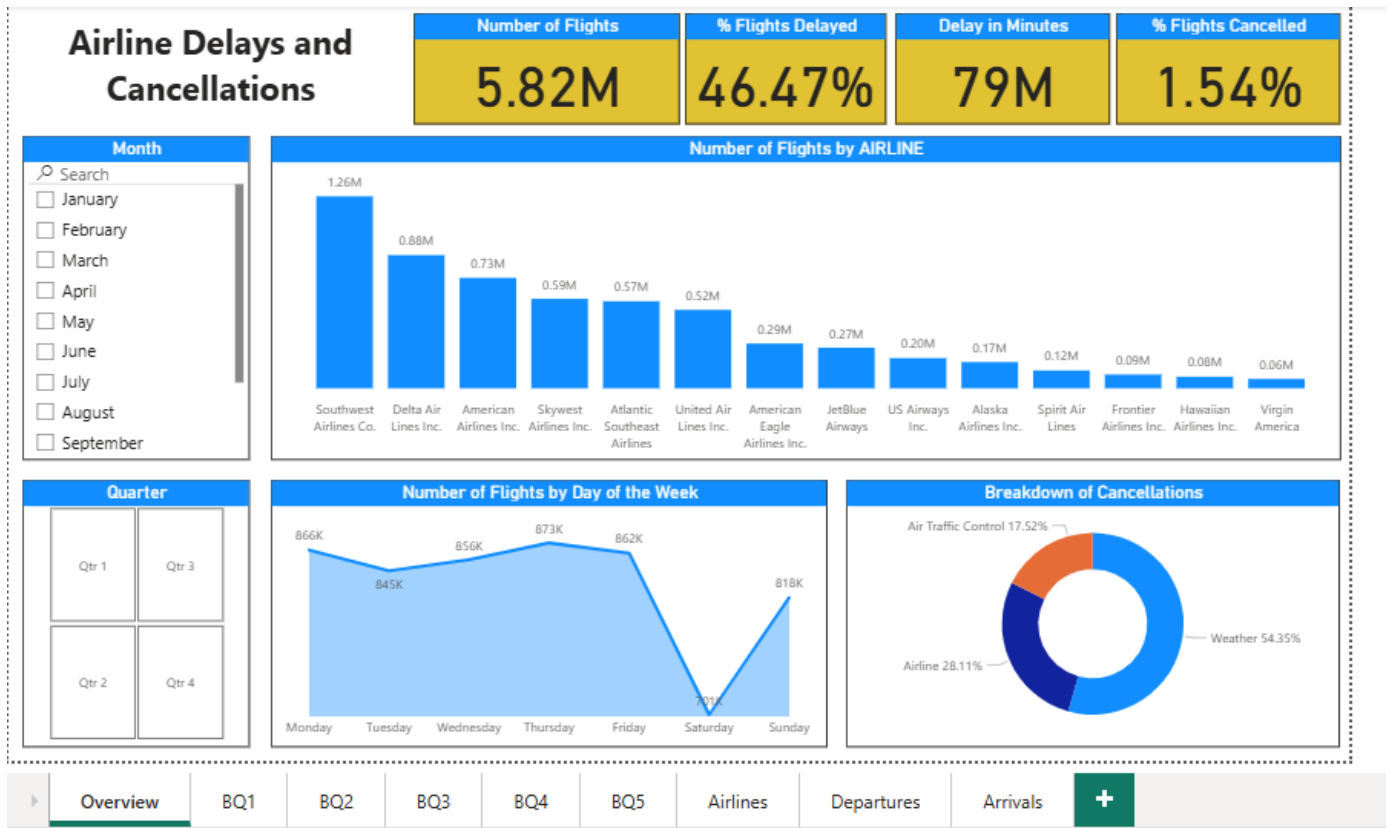
Future enhancements to this project could include :

- Incorporating **multi-year flight data** to study long-term trends and year-over-year performance changes.
- Predictive analytics and machine learning models could be developed to **forecast delays and cancellations** based on historical patterns, weather conditions, and traffic levels.
- Integration of real-time or near real-time data would further improve operational decision support.

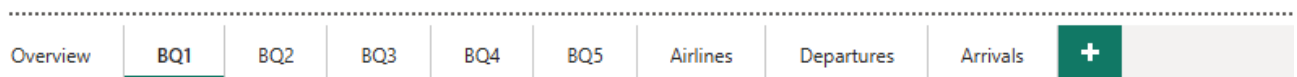
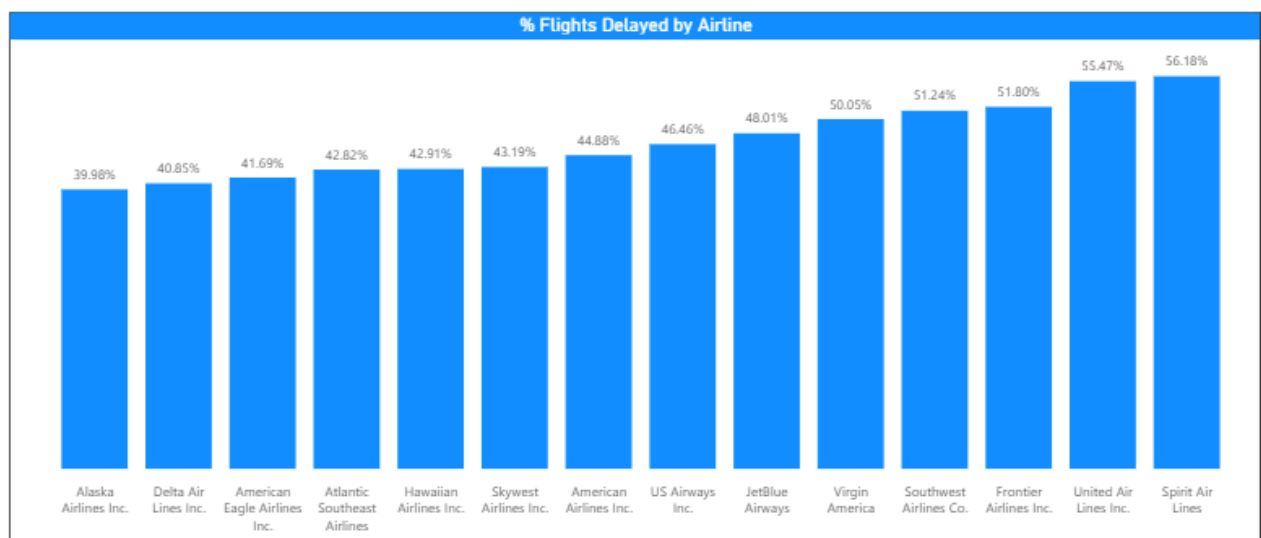
Additionally, expanding the geographic analysis to include **international flight data**, and incorporating external factors such as **weather forecasts, airport capacity metrics, and maintenance logs**, could provide deeper insights. Optimizing the dashboards for real-time monitoring and adding automated alerts for high-delay risk scenarios would further increase the practical value of the system.

## 7. Appendices

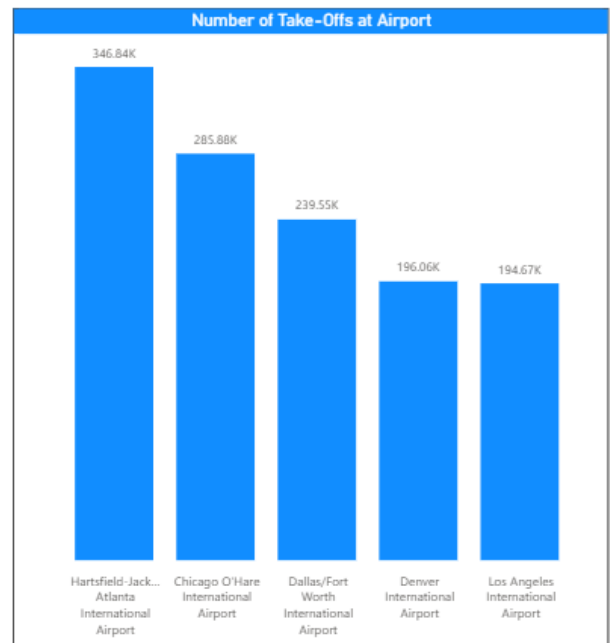
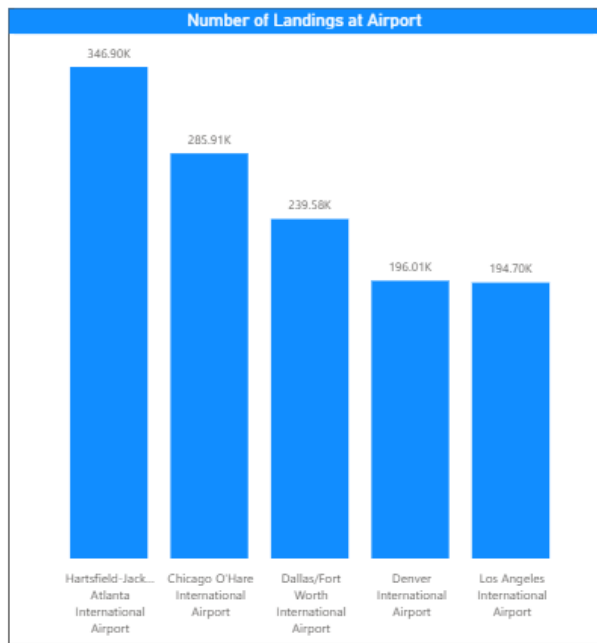
### Appendix 1: Screenshots



Flights of Which Airline are Delayed Frequently?

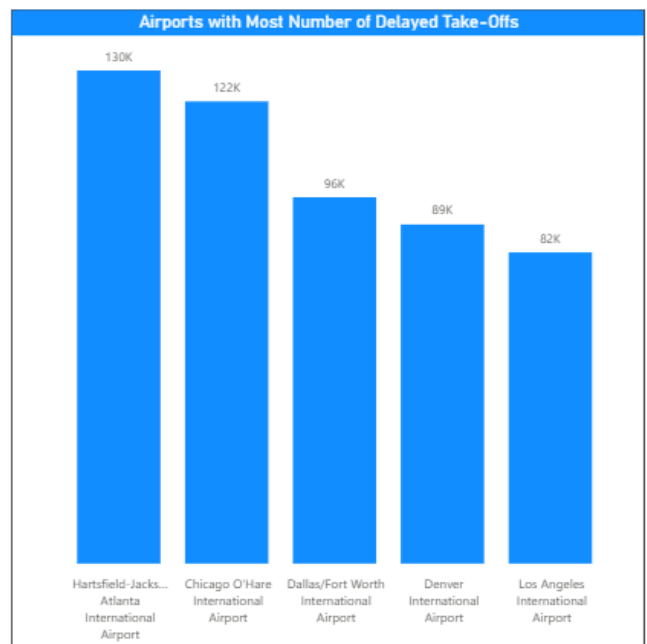
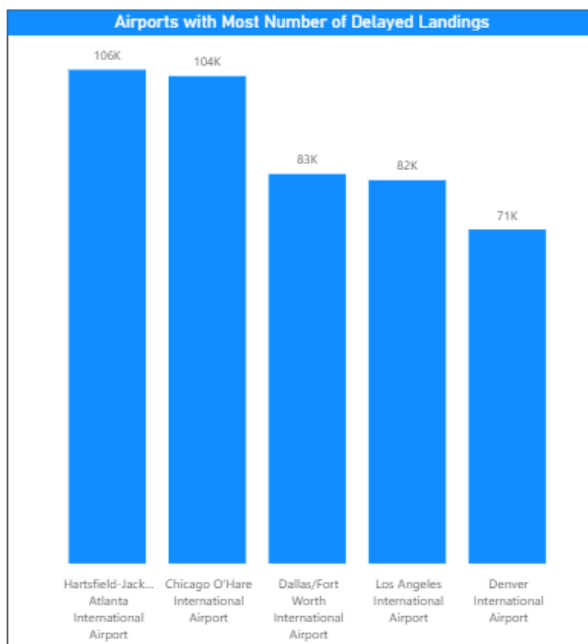


## Airports with High Activity



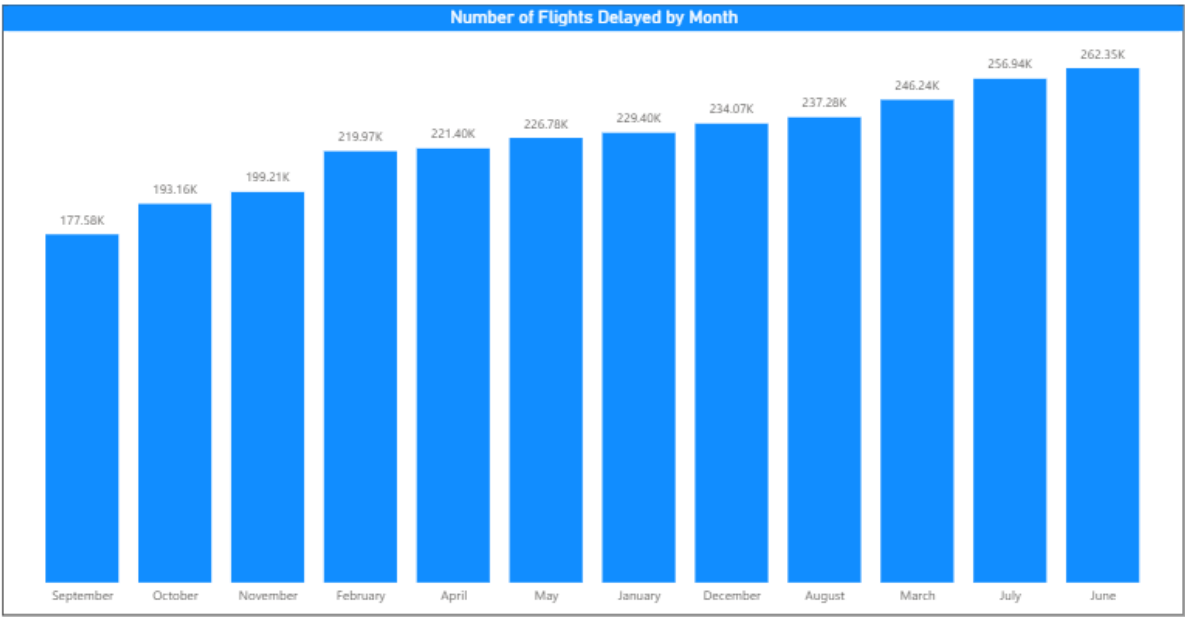
Overview	BQ1	<b>BQ2</b>	BQ3	BQ4	BQ5	Airlines	Departures	Arrivals	+	
----------	-----	------------	-----	-----	-----	----------	------------	----------	---	--

## Airports with Most Number of Delayed Flights



Overview	BQ1	BQ2	<b>BQ3</b>	BQ4	BQ5	Airlines	Departures	Arrivals	+	
----------	-----	-----	------------	-----	-----	----------	------------	----------	---	--

Months with Least and Most Number of Delays



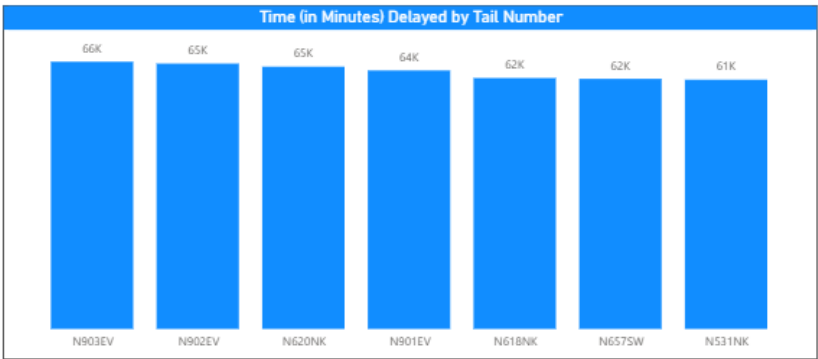
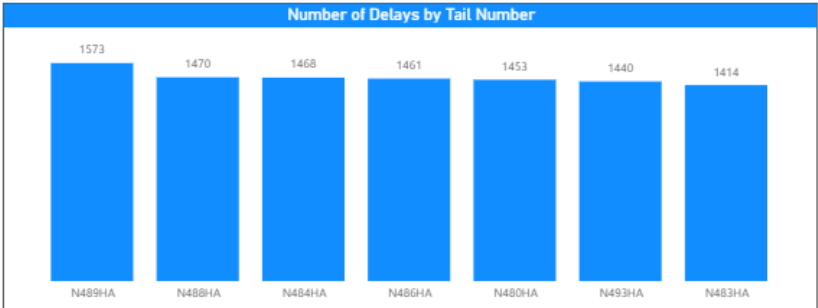
Overview	BQ1	BQ2	BQ3	BQ4	BQ5	Airlines	Departures	Arrivals	+
----------	-----	-----	-----	-----	-----	----------	------------	----------	---

For a Given Airline, which  
Aircrafts (Tail Numbers) cause  
Delays

Airline Operator

Search

- ☐ Alaska Airlines Inc.
- ☐ American Airlines Inc.
- ☐ American Eagle Airlines Inc.
- ☐ Atlantic Southeast Airlines
- ☐ Delta Air Lines Inc.
- ☐ Frontier Airlines Inc.
- ☐ Hawaiian Airlines Inc.
- ☐ JetBlue Airways
- ☐ Skywest Airlines Inc.
- ☐ Southwest Airlines Co.
- ☐ Spirit Air Lines
- ☐ United Air Lines Inc.
- ☐ US Airways Inc.
- ☐ Virgin America



Overview	BQ1	BQ2	BQ3	BQ4	BQ5	Airlines	Departures	Arrivals	+
----------	-----	-----	-----	-----	-----	----------	------------	----------	---

