Flight Delay and Performance Analytics System

DWV Final Report

Malak Sayed Ahmed, Faris Alaaeldin

ESLSCA 2025

## 1. Introduction

This project aims to explore and visualize patterns in airline flight performance data using a structured data analysis pipeline. The primary focus is on arrival delays, cancellations, and operational performance across U.S. carriers, routes, and dates.

## 2. Dataset Overview

* **Source**: A sample of 10,000 U.S. domestic flight records
* **File**: flights\_sample\_10k.csv
* **Records**: 10,000 rows
* **Attributes**: 32 columns
* **Key Columns**:
  + FL\_DATE, AIRLINE, ORIGIN, DEST, DEP\_DELAY, ARR\_DELAY, CANCELLED, CANCELLATION\_CODE, DIVERTED, DISTANCE

## 3. Data Preparation

**Tools Used:**

* **Python (pandas, matplotlib, seaborn)**
* **SQLite (via sqlite3)**
* **Tableau Public**

**Cleaning Steps:**

* Converted FL\_DATE to datetime
* Dropped rows with null critical fields (DEP\_DELAY, ARR\_DELAY)
* Filled missing CANCELLATION\_CODE with "None"
* Created DAY\_OF\_WEEK using Python's dt.day\_name()
* Added airline names using a lookup table in SQLite
* Exported cleaned data to flights\_for\_tableau.csv for visualization

## 4. Data Transformation

* Categorized delays into bins using ARR\_DELAY
* Extracted new columns:
  + DAY\_OF\_WEEK (from FL\_DATE)
  + DELAYED flag (1 if ARR\_DELAY > 15, else 0)
* Normalized delay metrics for ML
* Imported data into SQLite for efficient SQL querying

## 5. Analysis Process

**SQL Queries:**

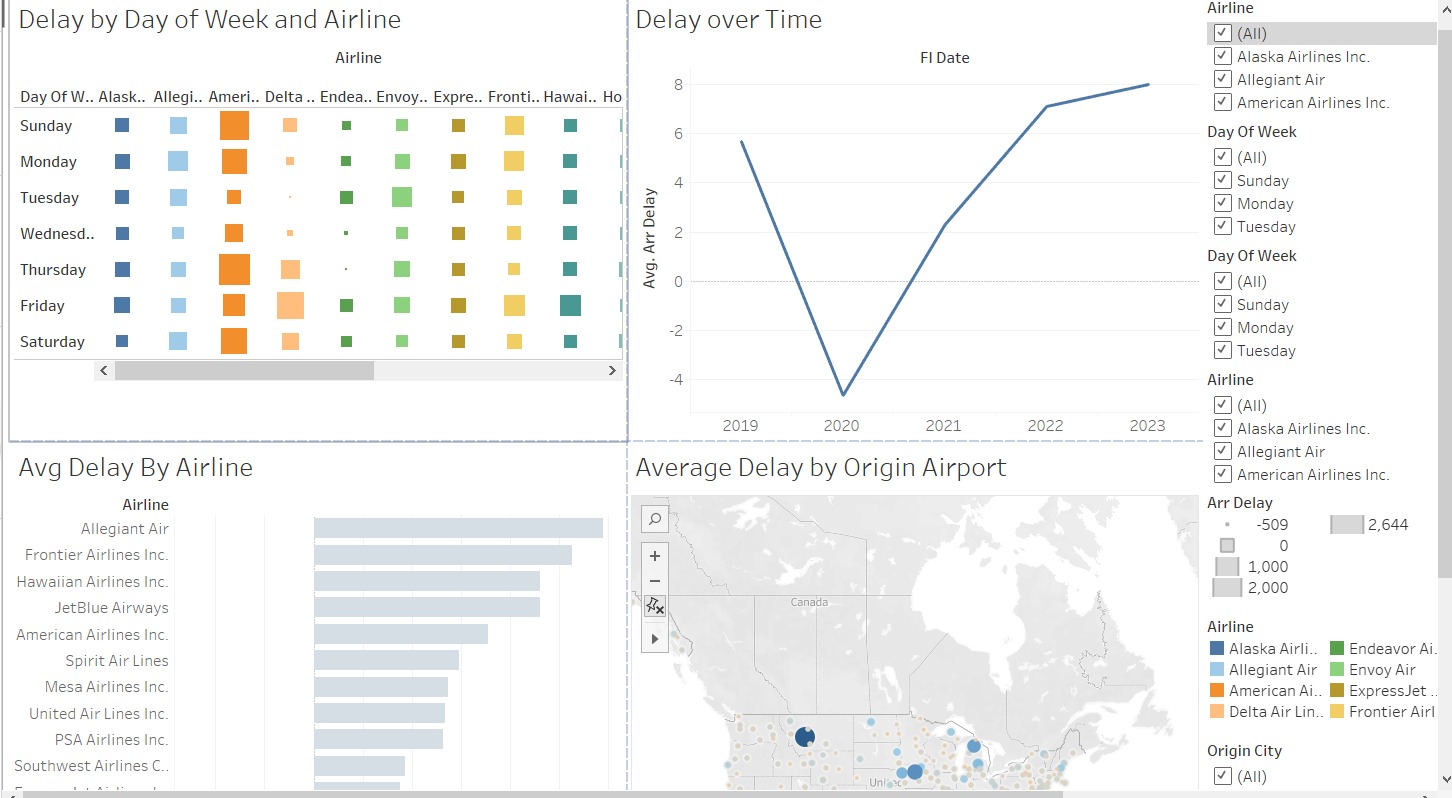
* Average delay per airline
* Top 10 most delayed routes
* Cancellations by reason
* Delay patterns by weekday
* Diverted vs. non-diverted flights

Queries were documented in analysis\_queries.sql.

**Python EDA:**

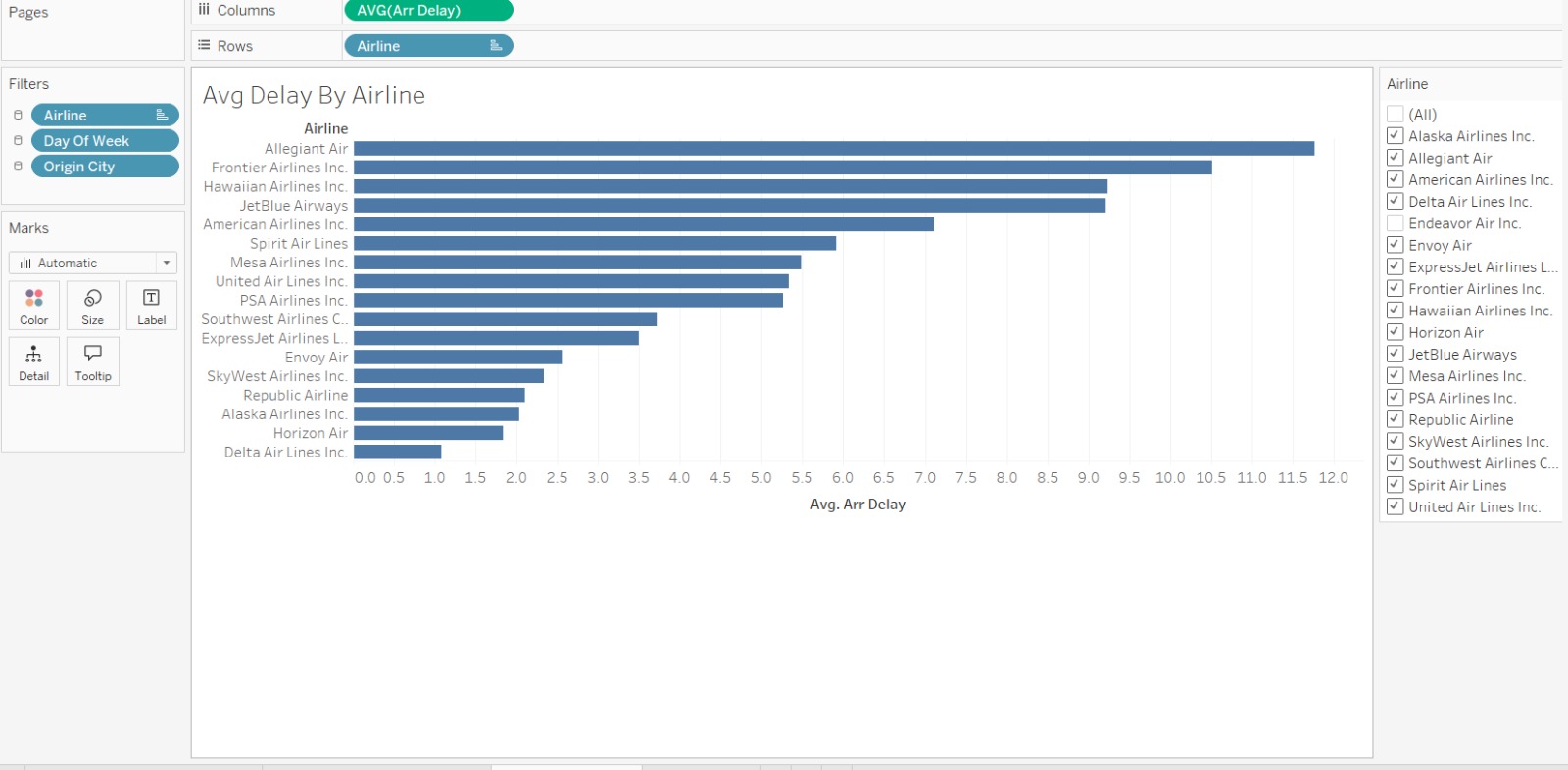
* Histograms of delay distribution
* Boxplots of delay by airline
* Bar charts of delay by day
* Random forest classifier for predicting delays

## 6. Tableau Visualizations

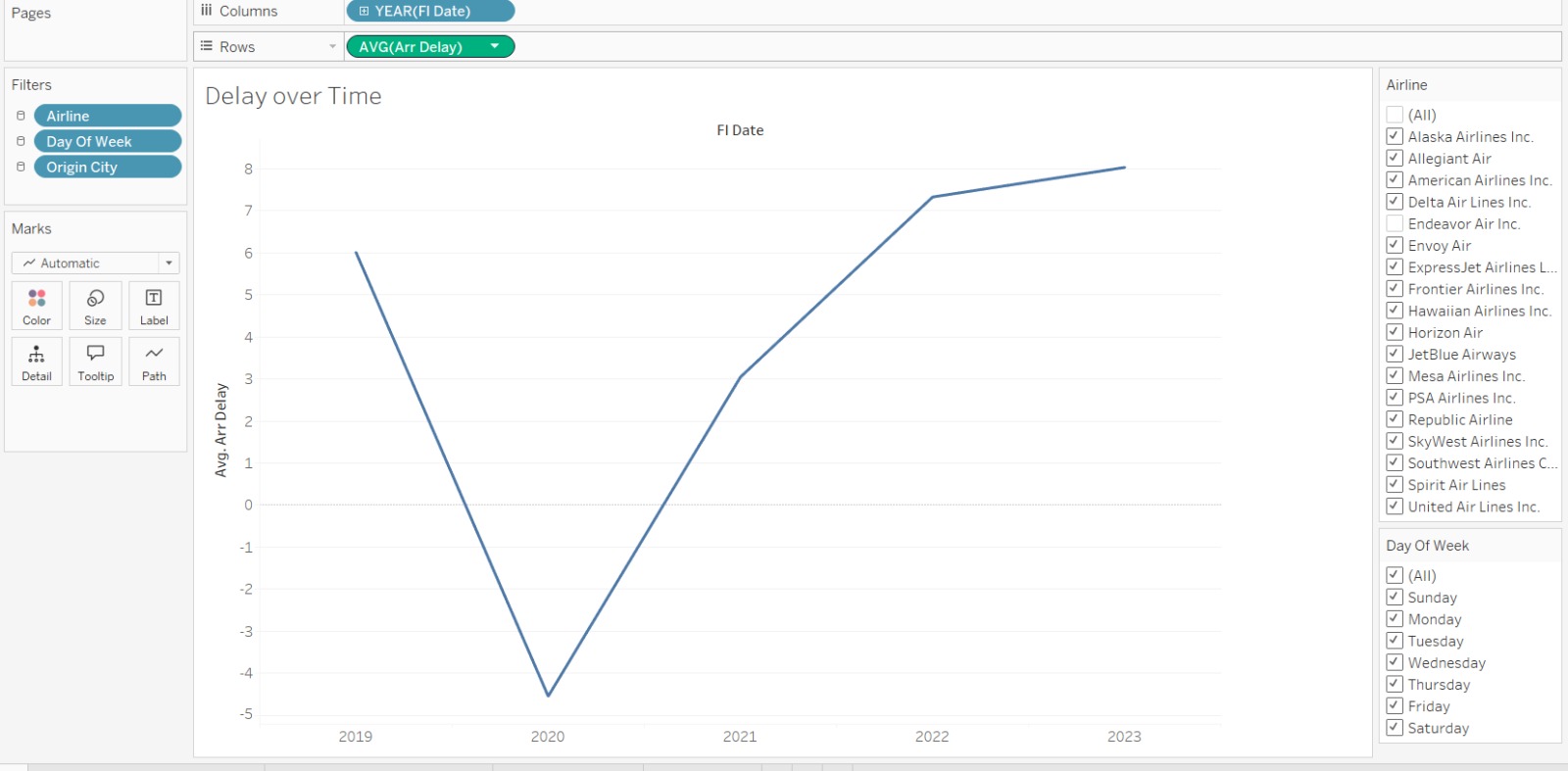
****

## Key Charts:

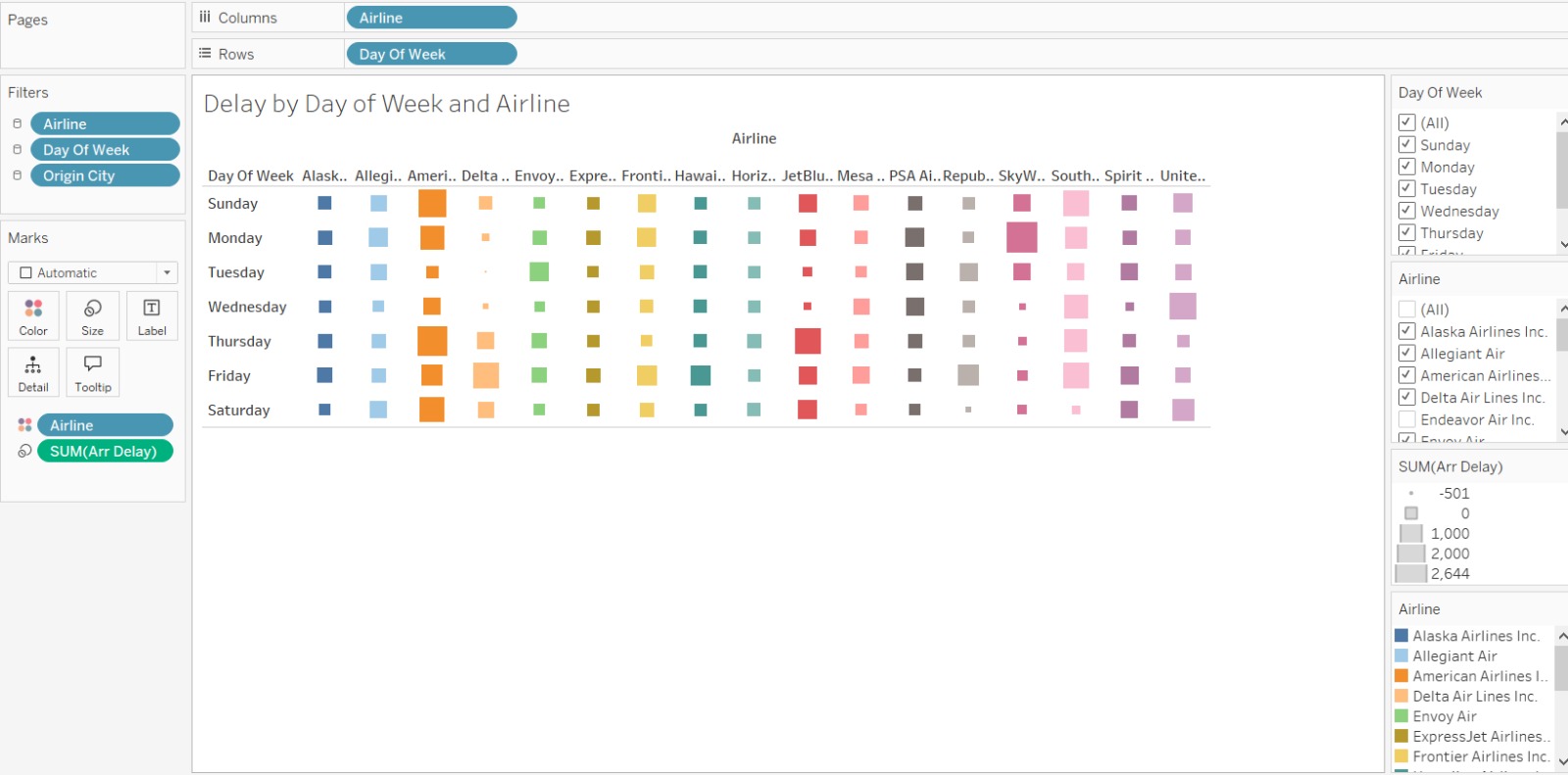
* **Bar Chart**: Average Delay by Airline



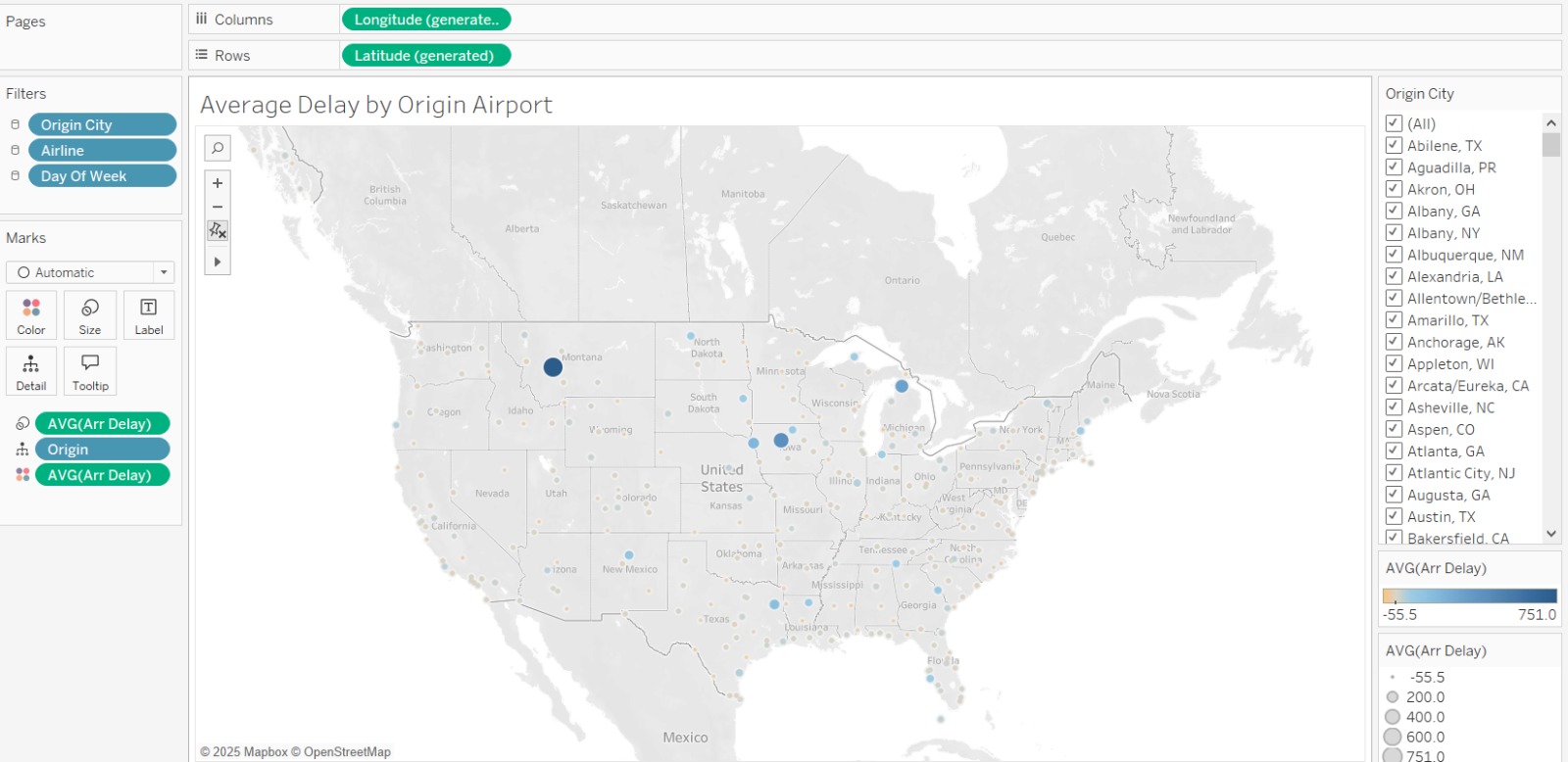
* **Line Chart**: Average Delay Over Time



* **Heatmap**: Delay by Airline and Day of Week



* **Map**: Average Delay by Origin Airports



## 7. Insights & Findings

* **Airline ATL** had the worst average delays, while **Airline Horizon Air** was most punctual
* **Thursday** and **Sunday** showed higher average delays than midweek
* Most cancellations were coded as "A" (Carrier-related)
* Some airports (like ORD and ATL) consistently had longer delays
* A basic classifier was able to predict delays using DEP\_DELAY and DISTANCE with 95% accuracy

## 8. Deliverables Summary

|  |  |
| --- | --- |
| File | Description |
| cleaned\_flights.csv | Cleaned dataset for analysis |
| flights.db | SQLite DB with both flights and airlines tables |
| analysis\_queries.sql | Documented SQL queries |
| flights\_for\_tableau.csv | Tableau-ready export with airline names |
| flights\_analysis.py | Python script for wrangling, EDA, and ML |
| dashboard.twbx | Tableau workbook file |
| *Screenshots* | For final presentation and report |

## 9. Conclusion

This project demonstrates a full data analysis lifecycle: collecting, cleaning, transforming, analyzing, and visualizing real-world airline data. Tableau provided a powerful platform to communicate the results in an interactive and visual way, enabling insights into operational reliability and areas for improvement in airline services.