

# AirFly Insights: Flight Delay Analysis

## Documentation

### AirFly Insights: Data Visualization and Analysis of Airline Operations

#### Project Statement

The objective of this project was to analyze a large-scale U.S. domestic airline flight dataset to uncover operational trends, delay patterns, and cancellation reasons using data visualization and predictive modeling techniques. The ultimate goal was to provide **actionable insights** for stakeholders regarding airline and airport-level performance, route optimization, and operational efficiency.

**Dataset:** flights\_sample\_100k.csv (100,000 rows, 32 columns)

#### Key Performance Indicators (KPIs)

KPI	Description
<b>Average Arrival Delay</b>	Mean of ArrDelay by carrier and route.
<b>Cancellation Rate</b>	Proportion of flights canceled over total.
<b>On-Time Performance</b>	Percentage of flights with ArrDelay $\leq 0$ .
<b>Route Popularity</b>	Number of flights per Origin-Destination pair.
<b>Peak Departure Hour</b>	Hour of day with the highest departures.

## Week 1: Project Initialization & Data Foundation

**Goal:** Define the project scope, establish KPIs, load the raw dataset, and conduct initial data quality and exploratory checks.

### Tasks Performed:

1. **Project Scoping:** Defined the flow into phases (Cleaning, Univariate, Bivariate, Seasonal, Modeling).
2. **Data Loading:** Loaded flights\_sample\_100k.csv.
3. **Initial Quality Assessment:** Performed a missing value report to understand data completeness.

### Key Findings (Initial Data Quality):

The dataset, while large, showed predictable missingness concentrated in operational outcome columns, which must be addressed before analysis.

Column	Missing Count (out of 100,000)	Missing Percentage	Note
CANCELLATION_CODE	97,373	97.37%	Expected, as a code only exists for cancelled flights.
Delay Reasons (5 columns)	82,008	82.01%	Missing for flights that were on-time or early.

Column	Missing Count (out of 100,000)	Missing Percentage	Note
ARR_DELAY, ELAPSED_TIME, AIR_TIME	2,852	2.85%	Primarily missing for flights that were cancelled or diverted.

## Week 2: Data Cleaning & Feature Engineering

**Goal:** Establish a robust data cleaning pipeline, handle missing values systematically, engineer necessary temporal and operational features, and save the cleaned data for efficiency.

### Tasks Performed:

#### 1. Missing Value Imputation:

- DEP\_DELAY, ARR\_DELAY: Imputed missing values with **0** for all non-cancelled flights, as a missing value here likely means the actual time was not logged due to cancellation/diversion, or the delay was zero.
- Delay Reason Columns (DELAY\_DUE\_...): Imputed with **0** minutes, converting them into quantitative features.
- CANCELLATION\_CODE: Replaced NaNs with "**None**".

#### 2. Feature Engineering:

- **Route Feature:** Created ROUTE by concatenating ORIGIN and DEST (e.g., 'ATL-LAX').
  - **Temporal Features:** Extracted Month, DayOfWeek, and DepHour from date/time fields.
3. **Data Storage:** The cleaned dataset was saved as a **Parquet file** (cleaned\_flights.parquet) to ensure faster read/write speeds and reduced file size for all subsequent analysis steps.

## Week 3: Univariate and Bivariate Analysis

**Goal:** Analyze the distribution of single variables (Univariate) and the relationship between pairs of variables (Bivariate) to uncover initial performance insights.

### Univariate Insights:

- **Overall Performance:** The **Average Arrival Delay** across the entire sample was **4.35 minutes**, while the overall **Cancellation Rate** was **2.63%** (2,627 flights).
- **Traffic Volume:** **Tuesday** was identified as the day with the highest flight traffic volume (16,616 flights).
- **Busiest Airports:** **ATL, DFW, and ORD** were confirmed as the top three busiest airports in the sample for both origin and destination.

### Bivariate Insights (Delay Drivers):

1. **Delay Cause Severity:** When a delay occurs and a reason is recorded, the largest contributor to the **average delay duration** is **Late Aircraft (32.89 minutes)**, highlighting the cascading effect of delays through the network.
2. **Carrier Performance Ranking:**

- **Lowest Performance:** Allegiant Air (G4) had the highest average arrival delay (**14.47 min**) and departure delay (**15.61 min**), followed closely by JetBlue Airways.
- **Highest Performance:** Endeavor Air Inc. showed the best punctuality, with a negative average arrival delay (**-2.13 min**), indicating flights generally arrive early.

## Week 4: Temporal and Route-Specific Analysis

**Goal:** Investigate how time (season, hour of day) and route characteristics influence operational outcomes, focusing on seasonal congestion and risk areas.

### Temporal Analysis:

- **Intra-day Trends:** Median delay analysis showed that flights are most punctual in the early morning (median delay of **-8.0 minutes** from 5-11 AM), but delays **compound throughout the day**, peaking in the evening.
- **Seasonal Extremes (95th Percentile Delay):** The 95th percentile of arrival delay (representing severe delays) was highest in **July and December**. This suggests that while mid-winter months (Jan/Feb) have higher cancellation risks, peak travel/holiday months (Jul/Dec) experience the most extreme delays when a severe delay does occur.

### Route and Cancellation Analysis:

- **Top Risk Route:** The route **LGA → CLT** (LaGuardia to Charlotte) was identified as having the highest raw **count** of cancellations in the dataset.
- **Cancellation Cause Breakdown:** Cancellations are primarily driven by **Carrier-related issues (47.9%)** and **NAS congestion (34.1%)**, with Weather being the third leading cause.

## Week 5 & 6: Predictive Modeling and Advanced Visualization

**Goal:** Finalize feature engineering, set up the data for a classification model, assess data quality for modeling (PSI), and create advanced comparative visualizations.

### Predictive Modeling Setup

- **Target:** A binary target variable, `delay_class` (1 if `ARR_DELAY` > 15 min, 0 otherwise), was chosen for classification modeling.
- **Data Preparation:** The dataset was split chronologically into **80% Training (80,000 rows)** and **20% Testing (20,000 rows)** to evaluate the model's ability to generalize to future flights.
- **Feature Set:** After one-hot encoding, the model utilized **751 features**, heavily weighted by the one-hot encoding of ORIGIN and DEST airports.

### Data Stability Assessment (PSI)

A crucial step for time-series splits is verifying that the distribution of key features remains stable between the training and testing periods.

Feature	Population Stability Index (PSI)	Interpretation
<code>DEP_DELAY</code>	0.177	<b>Small Change.</b> Since $0.1 < \text{PSI} < 0.2$ , the distribution of departure delays is considered reasonably stable, validating the robustness of the training data relative to the test data.

## **Advanced Visual Insights (7th\_Week\_Visuals\_AirFly\_Insights.pdf)**

1. **Airline & Month Cancellation Heatmap:** Visualized the **count** of cancellations by carrier and month.
  - **Insight:** While February and March are overall peaks, large carriers like **Southwest Airlines Co.** and **American Airlines Inc.** contribute the highest raw volume of cancellations due to their extensive network coverage.
2. **Airport Cancellation Rate Heatmap:** Visualized the **rate** of cancellations by origin airport per month.
  - **Insight:** This view showed that the highest cancellation *rates* are highly localized and not always found at the busiest hubs (ATL, DFW), indicating specific airport/route operational vulnerabilities beyond just traffic volume.
3. **Normalized Carrier KPIs (Radar Chart):** Plotted the top 4 busiest carriers against normalized metrics (Avg. Delays, Cancel Rate).
  - **Insight:** This comparative view confirmed that large carriers contribute the most to the overall magnitude of negative performance indicators, necessitating a focus on improving operational flow in the largest networks.

## **Tools, Technologies, and Libraries Used in the Project**

In this project it utilized a standard data science and visualization stack common for Python-based analytical projects, with specific tools chosen for data handling, cleaning, advanced visualization, and predictive modeling.

<b>Category</b>	<b>Tool / Library</b>	<b>Usage in Project</b>
<b>Programming Language</b>	<b>Python 3</b>	Core language used across all Jupyter Notebooks for data processing and analysis.
<b>Integrated Environment</b>	<b>Google Colab</b>	Primary environment for developing, documenting, and executing the weekly workflow (e.g., 1st_Week_AirFly_Insights.ipynb).
<b>Data Handling</b>	<b>Pandas</b>	Essential for data loading (.csv), cleaning, transformation, feature engineering, and high-performance operations (e.g., pd.read_csv, groupby(), missing value imputation).
	<b>NumPy</b>	Used for numerical operations, array manipulation, and mathematical functions, often integrated with Pandas.
<b>Data Storage</b>	<b>Parquet Format and CSV Format</b>	Used to save the cleaned data (cleaned_flights.parquet) for efficient, compressed, and faster read/write operations in subsequent weeks.
<b>Data Visualization</b>	<b>Matplotlib</b>	Base library for creating standard plots like bar charts,

<b>Category</b>	<b>Tool / Library</b>	<b>Usage in Project</b>
		line plots, and histograms (as seen in the notebooks).
	<b>Seaborn</b>	Used for statistical data visualization (e.g., potentially heatmaps, boxplots) to show relationships and distributions clearly.
	<b>Plotly (Express &amp; GO)</b>	Used for interactive and more advanced visualizations (as noted in 5th_Week_AirFly_Insights.ipynb), including route maps or advanced comparison charts.
	<b>Folium</b>	Specifically used for geographic data visualization (mapping) to analyze airport locations, routes, or average delays geographically.
<b>Machine Learning Modeling</b>	<b>Scikit-learn (sklearn)</b>	Core library for all modeling tasks, including:
	- train_test_split	Splitting data chronologically for modeling.
	- StandardScaler	Scaling features for distance-based models (if used).

<b>Category</b>	<b>Tool / Library</b>	<b>Usage in Project</b>
	- LogisticRegression	Likely used for the binary classification of delay_class.
	- RandomForestClassifier	Ensemble method used for robust classification of flight delays.
	-Metrics (accuracy_score, confusion_matrix, etc.)	Evaluating model performance.
<b>Advanced Visualization</b>	<b>NetworkX</b>	Could be used (as listed in 5th_Week_AirFly_Insights.ipynb imports) for visualizing the network structure of flights between airports.
<b>Reporting / Documentation</b>	<b>PDF &amp; Markdown</b>	Final delivery format for the project report and documentation (as requested and demonstrated by the example files).

## **Summary of Workflow per Week**

<b>Week</b>	<b>Focus</b>	<b>Key Libraries Used</b>
<b>Week 1-2</b>	Data Loading, Cleaning, Feature Engineering	Pandas, NumPy, Python's built-in datetime
<b>Week 3-4</b>	Univariate, Bivariate, and Seasonal Analysis	Pandas, Matplotlib, Seaborn
<b>Week 5-6</b>	Predictive Modeling and Advanced Visualization	Scikit-learn, Pandas, Plotly, Folium