Final Internship Report

Project Title: **AirFly Insights — Data Visualization and Analysis of Airline Operations**

Organization: **National Institute of Technology , Agartala**

Intern: **Adrika Saha**

➢ Project Overview

The *AirFly Insights* project aimed to analyze large-scale airline flight data to uncover operational trends, delay patterns, and cancellation causes through advanced data visualization techniques. The goal was to generate actionable insights to enhance airline operations and decision-making. This project aims to build a **clean**, **structured dataset** to support downstream analysis and modeling of **flight delays**, **cancellations**, and **route performance.**

The analysis covered flight schedules, delays, cancellations, and route-level performance, ultimately culminating in an interactive dashboard/report summarizing findings.

➢ Objectives

- Understand and preprocess aviation datasets for analysis.
- Explore trends in flight delays, cancellations, and operations.
- Visualize performance metrics using diverse visualization tools.
- Deliver meaningful insights to support data-driven decisions.
- Compile all findings into a cohesive visual report and presentation.

➢ Dataset Description

- **Source:** Kaggle Airlines Flights Dataset
- **Records:** Over 60 million flight records
- **Features:** Date, Airline, Flight Number, Origin, Destination, Departure/Arrival Time, Delay Types, Cancellation Codes, and Route Information.

- **Tools and Technologies:**

Programming Language	Python
Data Handling	Pandas, NumPy
Visualization	Matplotlib, Seaborn, Plotly, Folium
Dashboarding	Power BI
Documentation & Reporting	Jupyter Notebook/Vs Code, GitHub, PDF

- **Week-wise Progress Report**

WEEK 1: Project Initialization and Dataset Setup

1. Defining Objectives, KPIs, and Workflow

- The initial phase focused on establishing the scope and direction of the project, centered on airline delay and cancellation analysis.
- A detailed workflow was structured to guide the implementation process:
  Data Ingestion → Preprocessing → Feature Engineering → Exploratory Data Analysis (EDA) → Modeling.
- Key Performance Indicators (KPIs) were identified to measure airline efficiency, including metrics such as on-time performance, cancellation frequency, and delay duration.
- This step provided a clear analytical roadmap to ensure consistent progress throughout the project.

2. Data Loading and Integration

- The raw dataset was imported from the specified workspace path: /Volumes/airfly_workspace/default/airfly_insights/airfly_raw_data.csv using pandas.read_csv().
- The dataset consisted of:
  - Rows: 484,552
  - Columns: 30

- The data contained flight-level records, capturing various parameters such as flight number, origin and destination airports, departure and arrival times, delay status, and cancellation flags.

3. Dataset Exploration and Quality Assessment

A preliminary data audit was conducted to understand the structure and quality of the dataset.
Checks performed included:

- Schema and Data Types:
  Identified a mix of int, float, and object types. Time-related fields were stored as integers (HHMM format), while date fields appeared as strings.
- File Size: Approximately 90 MB.
- Missing Values: Detected in columns such as *ArrTime*, *DepTime*, *Org_Airport*, *Dest_Airport*, and *Cancelled*.
- Duplicate Records: Noted some repeated entries across *FlightNum*, *TailNum*, and *Date* combinations.

Findings: These inconsistencies were documented for resolution during the data preprocessing phase.

4. Sampling and Memory Optimization

To enhance computational efficiency and reduce processing time:

- A random sample of 5,000 rows was extracted for rapid testing and inspection.
- Data type optimization was applied by downcasting numeric columns (e.g., from int64 to int32) and converting categorical variables to category type.
- Time conversion operations (e.g., HHMM to datetime) were streamlined to execute once during preprocessing, minimizing redundant parsing overhead.
- These optimizations collectively improved performance, enabling faster iterations during exploratory and modeling stages.

Technology Stack

- Language: Python
- Libraries: Pandas, NumPy
- Environment: Jupyter Notebook / VS Code
- Storage: Local workspace directory

WEEK 2: Data Preprocessing and Feature Engineering

1. Handling Missing and Null Values

To ensure data completeness and maintain consistency across delay and cancellation-related fields, missing values were systematically treated as follows:

- Delay Columns (ArrDelay, DepDelay, CarrierDelay, WeatherDelay, etc.):
  Replaced all null entries with 0, indicating no delay recorded.
- Cancelled:
  Missing values were filled with 0, assuming the flight was not canceled.
- CancellationCode:
  Filled with 'None' for all missing cases to represent flights without a cancellation reason.

This approach standardized the dataset, making it suitable for statistical analysis and machine learning without introducing data bias.

---

2. Creation of Derived Features

New engineered features were introduced to enable temporal, route-based, and operational analysis. These derived attributes enriched the dataset and provided deeper analytical insights.

Feature	Description
Month	Extracted from the Date column to analyze monthly flight trends.
DayOfWeekNum	Encoded as numeric values (0–6) representing Monday–Sunday.
DepHour	Derived from converted DepTime to capture hourly flight patterns.
Route	Formed by concatenating Origin and Dest airport codes (e.g., *IND–BWI*) for route-level performance analysis.

These engineered variables played a key role in visualizations and performance trend identification during exploratory analysis.

---

3. Datetime Formatting and Standardization

Accurate handling of date and time fields was crucial for delay calculations and trend analysis. The following transformations were applied:

- Date Parsing: Converted to datetime objects using dayfirst=True to correctly interpret *DD–MM–YYYY* format.
- Time Fields (DepTime, ArrTime, CRSArrTime): Transformed from HHMM integers into datetime.time objects to facilitate duration computations and temporal grouping. This ensured temporal consistency across all records for subsequent modeling tasks.

---

4. Saving the Preprocessed Dataset

After completing data cleaning and feature engineering, the refined dataset was stored in multiple formats for flexibility and performance optimization:

- CSV Format: /Volumes/airfly_workspace/default/airfly_insights/flights_cleaned.csv
- Parquet Format (Optional): Used for faster read/write operations in downstream processing and EDA stages.

This structured output served as the master dataset for further analysis.

---

5. Feature Dictionary

A comprehensive feature dictionary was prepared to describe each variable's meaning and purpose for clarity and documentation.

Column	Description
DayOfWeek	Day of the week (1 = Monday, 7 = Sunday)
Date	Flight date
DepTime	Actual departure time (HH:MM)
ArrTime	Actual arrival time (HH:MM)
CRSArrTime	Scheduled arrival time
UniqueCarrier	Airline code
Airline	Airline name
FlightNum	Flight number
TailNum	Aircraft tail number

Column	Description
ActualElapsedTime	Actual flight duration (minutes)
CRSElapsedTime	Scheduled flight duration (minutes)
AirTime	Airborne time (minutes)
ArrDelay	Arrival delay (minutes)
DepDelay	Departure delay (minutes)
Origin	Origin airport code
Org_Airport	Origin airport name
Dest	Destination airport code
Dest_Airport	Destination airport name
Distance	Distance between airports (miles)
TaxiIn	Taxi-in duration (minutes)
TaxiOut	Taxi-out duration (minutes)
Cancelled	1 if flight canceled, else 0
CancellationCode	Reason for cancellation
Diverted	1 if flight diverted
CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay	Specific delay causes
Month, DayOfWeekNum, DepHour, Route	Derived features for time- and route-based analysis

---

Technology Stack

- Programming Language: Python
- Libraries Used: Pandas, NumPy, Datetime
- Data Storage Formats: CSV, Parquet
- Development Environment: Jupyter Notebook / VS Code

WEEK 3: Univariate and Bivariate Visual Analysis

The third week was dedicated to performing visual exploratory data analysis (EDA) on the cleaned and feature-enriched dataset.
The goal was to identify significant trends, patterns, and dependencies related to flight volume, delays, cancellations, and route behaviors.

All analyses utilized the processed dataset that included derived features such as Month, DayOfWeekNum, DepHour, and Route.

---

1. Univariate Analysis

Objective

Univariate analysis was carried out to study the distribution and key characteristics of individual variables within the dataset. This helped in understanding central tendencies, dispersion, and frequency patterns of critical flight attributes.

Tasks Performed

1. Top Airlines:
   Counted the total number of flights operated by each airline to determine the busiest carriers.
2. Top Routes:
   Analyzed the frequency of flights between each origin–destination pair to identify high-traffic routes.
3. Busiest Months:
   Calculated monthly flight counts to detect seasonal peaks in air traffic.
4. Flights by Day of Week:
   Evaluated weekly flight trends to assess operational consistency and weekly travel demand.
5. Departure Hour Distribution:
   Examined the number of flights across different departure hours to determine peak flight times.
6. Flights by Origin Airport:
   Counted departures from each airport to identify the busiest origin airports.
7. Arrival Delay Distribution:
   Visualized the distribution of arrival delays to analyze punctuality performance and detect long-tail delays.

Visualization Techniques

- Bar Charts: Used for categorical comparisons such as Airline, Route, Month, and Origin.
- Histograms: Applied to visualize continuous distributions like Departure Hour and Arrival Delay.
- Boxplots: Utilized for identifying the spread, variability, and outliers in numeric features (e.g., Arrival Delay).

Insights

The univariate study revealed variations in flight frequencies across airlines and routes, distinct travel peaks in certain months, and visible outlier patterns in delay distributions that indicated irregular punctuality during specific periods.

---

2. Bivariate Analysis

Objective

Bivariate analysis was conducted to examine relationships between pairs of variables, uncovering patterns, dependencies, and correlations relevant to flight operations and delays.

Tasks Performed

1. Arrival Delay by Airline:
   Compared delay distributions among airlines using boxplots to assess performance disparities.
2. Average Arrival Delay by Month:
   Computed mean arrival delay for each month and visualized seasonal variations through line plots.
3. Optional Analyses:
   - Delay vs Departure Hour: To determine if delays were influenced by time of day.
   - Delay vs Distance: To explore whether longer flights exhibited higher delay tendencies.

Visualization Techniques

- Boxplots: To compare numeric variables against categorical ones (e.g., ArrDelay vs Airline).
- Line Plots: To visualize temporal delay trends (e.g., ArrDelay vs Month).

- Scatter Plots: To study numeric correlations, such as between ArrDelay and Distance.

Insights

The bivariate analysis helped uncover:

- Airlines with consistently higher average delays.
- Seasonal fluctuations in delay patterns.
- Time-of-day and route-related dependencies affecting punctuality.

These insights provided a strong foundation for further multivariate analysis and modeling in subsequent stages.

Technology Stack

- Language: Python
- Libraries: Pandas, Matplotlib, Seaborn
- Visualization Environment: Jupyter Notebook / VS Code
- Chart Types: Bar, Histogram, Boxplot, Line, and Scatter

WEEK 4: Delay Cause Analysis and Time-Based Trends

1. Objective

The fourth week was dedicated to a detailed investigation of flight delay causes and their temporal behavior. The primary goal was to identify which operational or environmental factors most influenced delays and how these patterns varied over different time periods such as hours of the day and days of the week.

---

2. Delay Categories Analyzed

The dataset included multiple delay-type columns, each representing a specific source of delay. The analysis primarily focused on the following four key categories:

1. Carrier Delay: Delays caused by the airline itself (e.g., maintenance issues, crew unavailability).
2. Weather Delay: Delays due to adverse meteorological conditions.
3. NAS Delay: Delays arising from the National Airspace System or air traffic control restrictions.

4. Security Delay: Delays caused by security checks or threats impacting flight schedules.

Each of these delay causes was examined individually and in aggregate to assess their contribution to total delay time.

---

3. Analytical Approach

A systematic approach was employed to group and analyze delay data across various dimensions:

- Grouping by Airline and Delay Type:
  Using pandas.groupby(), delays were aggregated per airline and delay category to determine which carriers or regions experienced the highest frequency and duration of specific delay types.
- Time-Based Analysis:
  Delay patterns were further explored across time intervals to identify recurring trends and high-impact periods.
  The time dimensions considered were:
  - Hour of the Day – to identify peak hours of delay occurrences.
  - Day of the Week – to highlight operational days with maximum delay rates.
  - Month/Season – to observe longer-term temporal patterns and festive travel impacts.

---

4. Visualization and Interpretation

To effectively interpret delay behaviors, multiple visualization techniques were applied:

- Bar Charts (seaborn.barplot) – to compare the average delay duration across airlines and delay types.
- Heatmaps (sns.heatmap) – to visualize delay density across different hours of the day and days of the week, helping identify high-risk time windows.
- Line Charts – to reveal cyclical and seasonal delay trends over time.

These visualizations provided a clear understanding of how delays varied both by cause and time period.

---

5. Key Insights and Observations

The findings from this analysis highlighted several crucial operational trends:

- Carrier and Weather Delays accounted for the largest share of total delay minutes, indicating that both airline operations and weather disruptions significantly affected flight punctuality.
- Early Morning Flights generally experienced fewer delays, suggesting smoother air traffic and reduced congestion during those hours.
- Evening Flights showed higher delay frequencies, often due to cumulative schedule shifts and airspace congestion.
- Delays Peaked During Weekends and Festive Seasons, reflecting higher passenger volumes and increased flight density.

These insights not only explained operational inefficiencies but also helped relate delay patterns to real-world factors such as weather conditions, traffic congestion, and peak travel behavior.

---

6. Tools and Technologies Used

Category	Tools/Functions
Data Aggregation	pandas.groupby(), pandas.mean()
Visualization	seaborn.barplot(), sns.heatmap(), matplotlib.pyplot
Analysis Environment	Python (Jupyter Notebook / VS Code)
Libraries Used	Pandas, Seaborn, Matplotlib

---

7. Outcome

The week’s analysis successfully established a data-driven understanding of delay causation. It provided valuable operational insights that could support airline performance optimization, resource allocation, and schedule planning for improving punctuality and minimizing disruptions in future flight operations.

.

Week 5: Route and Airport-Level Analysis

Tasks Completed:

- Identified **Top 10 Origin-Destination pairs** with maximum traffic.
- Created **heatmaps** for delay intensity by airport and route.
- Used **Folium** maps to visualize the busiest airports and average delay times geographically.

Technologies Used: Seaborn.
Outcome: Discovered congestion-prone airports and routes; mapped regional delay patterns visually for better interpretation.

WEEK 6: Seasonal and Cancellation Analysis

1. Objective

The sixth week concentrated on examining seasonal flight behavior and cancellation patterns across the operational year.
The primary goal was to understand how external and internal factors — such as weather conditions, operational loads, and airspace management — influenced the frequency and causes of flight cancellations.

---

2. Focus Areas

Two major analytical components were addressed during this week:

1. Seasonal Trend Analysis:
   Studying how monthly and seasonal variations affected flight volumes, punctuality, and cancellations.
2. Cancellation Cause Analysis:
   Investigating the primary reasons behind flight cancellations and their proportional contribution throughout the year.

The delay and cancellation records were analyzed across four major categories available in the dataset:

- Carrier Cancellations – operational or technical issues within the airline.
- Weather Cancellations – adverse weather events such as fog, storms, or heavy rain.
- NAS Cancellations – air traffic control and airspace management restrictions.
- Security Cancellations – cancellations due to safety or security-related concerns.

---

3. Analytical Procedure

A step-by-step approach was followed to identify and visualize seasonal and categorical patterns:

1. Data Grouping:
   The dataset was grouped by both Month and CancellationCode using pandas.groupby() to compute the number of cancellations under each category per month.
2. Percentage Calculation:
   The proportion of cancellations for each reason was calculated relative to total monthly flights to enable comparative analysis across categories.
3. Visualization:
   Multiple visualization techniques were implemented to display results clearly:
   - Bar Charts: To compare cancellation counts and proportions across months and categories.
   - Pie Charts: To illustrate the percentage distribution of cancellation causes within each month or season.
   - Line Charts (Optional): Used to highlight changing cancellation trends over time.

---

4. Key Observations and Insights

The analysis revealed several important trends in airline operations and cancellation behaviors:

- Winter Season Impact:
  A noticeable increase in weather-related cancellations occurred during the winter months (particularly December to February). This aligned with seasonal conditions such as fog, low visibility, and storms affecting takeoffs and landings.
- Carrier-Related Dominance:
  Carrier-related cancellations were the most frequent overall, especially during peak operational months (summer and festive travel periods). This pattern indicated increased pressure on airline fleets and scheduling systems during high-demand periods.
- NAS (Air Traffic) Delays:
  NAS-related cancellations were moderate but showed slight increases during months with heavy traffic, suggesting congestion or regulatory airspace constraints.

- Security Cancellations:
  Although minimal, security-related cancellations remained consistent throughout the year, implying that they are rare but unavoidable due to mandatory safety measures.

These findings demonstrated how seasonal variability and external factors significantly influenced flight reliability.

---

5. Implications and Applications

The insights derived from this week's analysis have several operational implications:

- Airlines can use seasonal cancellation trends to plan proactive maintenance schedules and crew allocations ahead of high-risk periods.
- Enhanced weather forecasting integration could help reduce weather-related disruptions.
- Better resource and slot management during high-traffic months can mitigate carrier and NAS-related cancellations.
- Data-driven policy formulation can improve punctuality targets and passenger satisfaction by anticipating cancellation risks.

---

6. Tools and Technologies Used

Category	Tools / Methods
Data Aggregation & Calculation	pandas.groupby(), pandas.value_counts(), pandas.pivot_table()
Visualization	matplotlib.pyplot, seaborn.barplot(), plt.pie()
Environment	Python (Jupyter Notebook / VS Code)
Libraries Used	Pandas, Matplotlib, Seaborn

---

7. Outcome

This week's work provided a comprehensive understanding of cancellation dynamics and their seasonal dependencies.
The analysis emphasized the importance of strategic scheduling, weather preparedness, and operational resilience to improve airline performance and minimize disruptions during critical travel seasons.

Week 7: Visual Report and Dashboard Preparation

Tasks Completed:

- Combined all visualizations into a coherent narrative storyline.
- Designed an interactive **dashboard** using **Streamlit/Power BI** for intuitive exploration.
- Ensured clarity in titles, labels, legends, and annotations for readability.

Technologies Used: Power BI
Outcome: Developed a visually engaging dashboard ("AirFly Insights") presenting all analytical findings in a professional format.

Final Reflections and Key Takeaways

The **AirFly Insights internship** was an enriching end-to-end data analytics experience — from raw data preprocessing to meaningful visualization. It offered practical exposure to how real-world airline data can uncover operational patterns and performance inefficiencies.

Throughout this journey, I gained hands-on expertise in:

- Efficient data cleaning and preprocessing techniques.
- Applying feature engineering to enhance analytical depth.
- Creating insightful visualizations and interactive dashboards in **Power BI**.
- Translating technical analysis into clear, data-driven storytelling.

This internship not only strengthened my proficiency in **Python and data analytics tools** but also refined my **critical thinking, visualization, and communication skills**. It has prepared me to handle full-cycle data projects with confidence — transforming data into insights that drive smarter decisions.