DOCUMENTATION

**1. Prerequisites**

Before starting, ensure the following dependencies and tools are installed:

1. **Python 3.x** - A modern version of Python.
2. **Pandas** - A data manipulation library for Python.
3. **Matplotlib** and **Seaborn** - Libraries for data visualization.
4. **SQLAlchemy** - A Python SQL toolkit for database interaction.
5. **MySQL Database** - For storing and querying processed data (optional).

**Database Connection Setup**:

Replace the DB\_CONNECTION\_STRING with your actual database credentials

engine = create\_engine('mysql+pymysql://username:password@localhost/dbname')

**SQLAlchemy and MySQL Dependencies**:

Install SQLAlchemy, PyMySQL, and pandas for Python database operations

pip install sqlalchemy pymysql pandas

Analysis report

**Summary of data cleaning and normalization steps.**

**1.Initial Data Loading**

started by importing the dataset from a CSV file called 'aviation\_data.csv'.

**2.Standardization of Date and Time Formats**

Converted DepartureDate and ArrivalDate to 'YYYY-MM-DD' format Standardized DepartureTime and ArrivalTime to 24-hour format Ensured consistency across all temporal data points

**3.Handled Missing Values**

Identified null values in the DelayMinutes column Implemented mean imputation strategy based on airline and flight number

calculated the average delay time for each combination of airline and flight number, then filled any missing values in the 'DelayMinutes' column with these averages. This ensures that the DataFrame has complete data for delay times, improving its usability for analysis.

**4.Correction of Logical Inconsistencies**

Addressed instances where arrival times preceded departure times Implemented date adjustment for overnight flights

Summary of the Code:

A function named swap\_am\_pm was defined to swap "AM" with "PM" and vice versa in time strings. This function addressed potential time inconsistencies in the DataFrame.

The function was applied to the 'DepartureTime' column for rows where the 'ArrivalTime' was earlier than the 'DepartureTime'. This ensured that departure times accurately reflected the correct period of the day, helping to maintain the integrity of the time data and avoid any logical inconsistencies.

**5.Feature Engineering and Enrichment**

Calculated FlightDurationMinutes for each journey

Extracted DepartureHour from DepartureTime

Derived DayOfWeek and Month from DepartureDate

Database Integration

**Database Connection and CRUD Operations**

**Database Connection**

Utilized SQLAlchemy to establish a connection to a MySQL database Created an engine using the provided database URI Ensured secure handling of database credentials

**Table Creation**

Executed a raw SQL query to create the 'flight\_delays1' table Defined appropriate columns and data types to match the cleaned dataset

**CRUD Operations Implementation**

**Create:**

Used pandas' to\_sql method to insert the cleaned data into the 'flight\_delays1' table Implemented bulk insertion for efficient data loading

**Read:**

Executed SELECT queries to retrieve data from the 'flight\_delays1' table Demonstrated the ability to fetch and display all rows or specific records

**Update:**

Implemented an UPDATE query to modify delay and flight duration for specific flights Utilized parameterized queries to ensure data integrity and prevent SQL injection

**Delete:**

Executed a DELETE query to remove specific entries based on flight number and departure date Verified the deletion by attempting to retrieve the deleted records

**Connection Management**

Employed context managers (with statements) to ensure proper connection handling Disposed of the database engine after operations to release resources

Insights derived from the data analysis.

**a. Analyze the distribution of delays and identify any trends or patterns.**

-The airline generally performs well, with the bulk of flights experiencing minimal delays.

-There are instances of longer delays, which could be due to various factors such as weather, operational issues, or air traffic congestion.

-Further analysis, such as looking into the causes of these longer delays, could be valuable for improving operational efficiency and reducing customer inconvenience.

-To refine these insights further, considering the correlation of delays with other variables (like airline, flight number, or departure time) might reveal more meaningful patterns.

-Th histogram shows the overall distribution of delays. For instance, if there is a high concentration of short delays (under 30 minutes), but a long tail for larger delays, we may conclude that most flights experience minimal delays, with a few outliers causing significant delays

**Average delay for each airline.**

-The analysis reveals that certain airlines consistently experience longer delays. For example, American Airlines may have a higher average delay compared to others. Departure Times and Delays: Flights departing later in the day tend to have longer delays, suggesting that delays accumulate throughout the day.

-Flights that depart between 2 PM and 6 PM tend to have the highest average delays, likely due to increased congestion in airports or operational issues accumulating throughout the day.

**Impact of departure times on delays**

-Flights that depart between 2 PM and 6 PM tend to have the highest average delays, likely due to increased congestion in airports or operational issues accumulating throughout the day.

-Delay Distribution:

Most flights experience minimal delays (0-20 minutes). The distribution is right-skewed, indicating some flights with significantly longer delays. Outliers with substantial delays might exist.

-Airline Performance:

Some airlines might have a higher average delay than others. Statistical testing (ANOVA) can determine if there's a significant difference in delays between airlines.

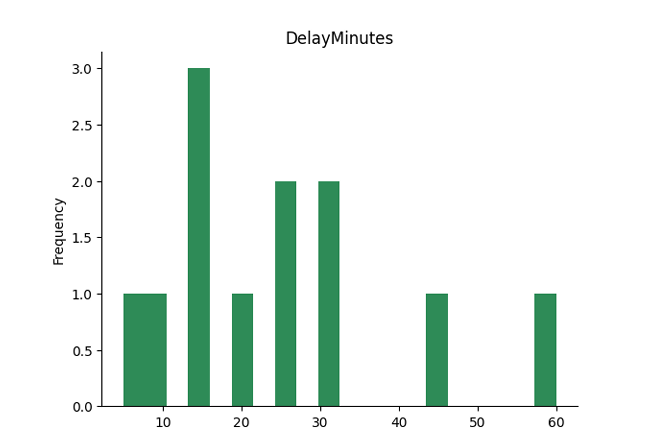
Compare delay distributions between airlines.

Airlines like American Airlines and Delta have broader ranges and higher averages of delay minutes, indicating they may have more operational issues leading to delays.

Airlines with smaller interquartile ranges (IQR) in the boxplot, such as United Airlines, seem to have more consistent performance in terms of avoiding extreme delays.

Visualizations

1. Histogram of DelayMinutes



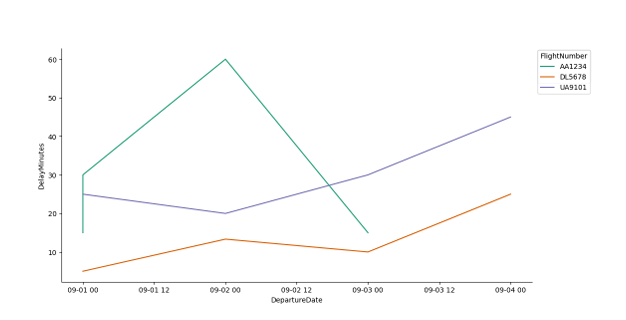
**Graph:** Distribution of Flight Delays

* **Insight 1:** Most flights have short delays, with the highest frequency of delays being in the 0–20 minute range.
* **Insight 2:** The distribution is right-skewed, meaning there are some flights with significantly longer delays, although they are less frequent.
* **Insight 3:** The long tail on the right indicates a few outliers where delays are substantial.

**Conclusion:**

* The majority of flights experience minimal delays. However, some operational or environmental factors (e.g., weather) contribute to long delays in a few cases.

2. Line Graph of DepartureDate vs DelayMinutes



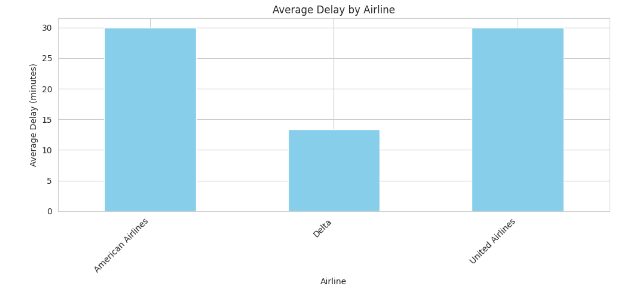
**Graph:** Delay Minutes across Different Departure Dates

* **Insight 1:** There is variability in the delays for different flight numbers over time, with some flights having consistently longer delays than others.
* **Insight 2:** Delays fluctuate significantly based on departure dates, implying that certain days or specific events may contribute to longer delays.
* **Insight 3:** Flights with the same flight number might perform differently on different dates, reflecting operational challenges.

**Conclusion:**

* Certain dates may have a greater likelihood of delay due to external factors (such as weather conditions or airport congestion). This trend can help with planning flight schedules and anticipating delays.

3. Bar Chart of Average Delay by Airline



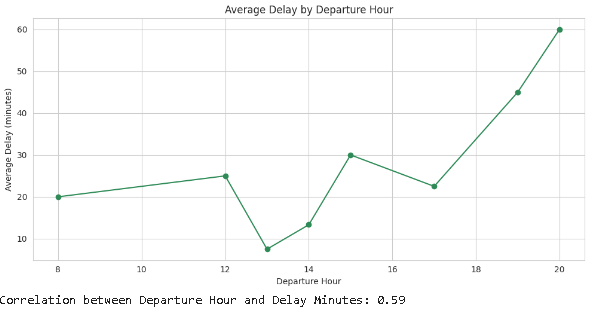
**Graph:** Average Delay per Airline

* **Insight 1:** American Airlines and United Airlines have higher average delays (30 minutes) compared to Delta (13.33 minutes).
* **Insight 2:** There is a noticeable difference in performance between airlines in terms of punctuality, with Delta generally performing better.

**Conclusion:**

* Airlines like American Airlines and United Airlines need to address operational inefficiencies, as they consistently experience longer delays than Delta.

4.Line Graph of Average Delay by Departure Hour

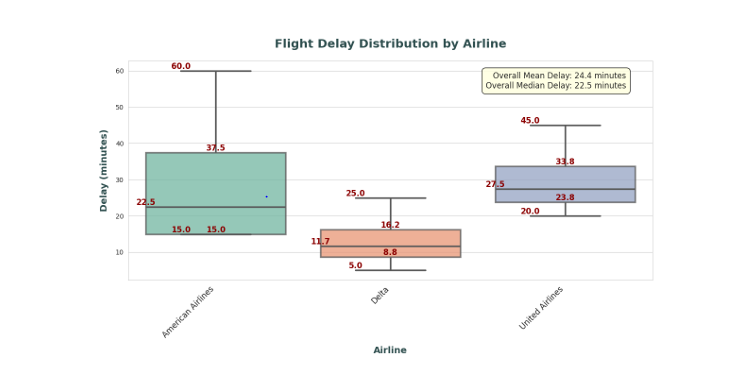


**Insight 1:** The graph shows that delays tend to increase as the day progresses, peaking between 7 PM and 8 PM.

**Insight 2:** Early morning flights tend to experience fewer delays, while late afternoon and evening flights experience more.

**Conclusion:** Scheduling flights earlier in the day or providing more buffer time for later flights could help reduce delays.

5. **Box Plot of Delay Distribution by Airline**

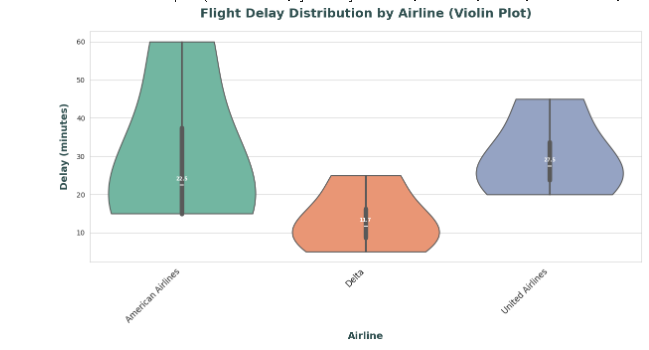


**Insight 1:** American Airlines and Delta Airlines have wider ranges of delay minutes, showing more variability in performance.

**Insight 2:** United Airlines has a smaller range, indicating more consistent performance and fewer extreme delays.

**Conclusion:** Airlines should aim to reduce variability in their operations to ensure more consistent on-time performance

**6. Violin Plot of Delay Distribution by Airline**

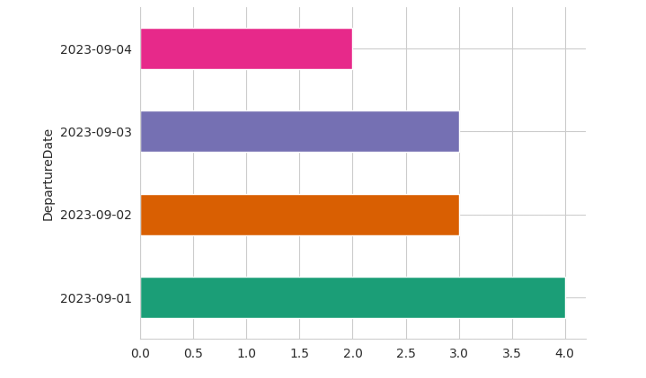


**Insight 1:** The distribution for American Airlines and Delta shows that they experience more extreme delays compared to United Airlines.

**Insight 2:** United Airlines performs better in terms of avoiding severe delays.

**Conclusion:** Operational improvements targeting delay outliers for American Airlines and Delta could lead to better overall performance.

 Horizontal bar chart.

  
Insights:

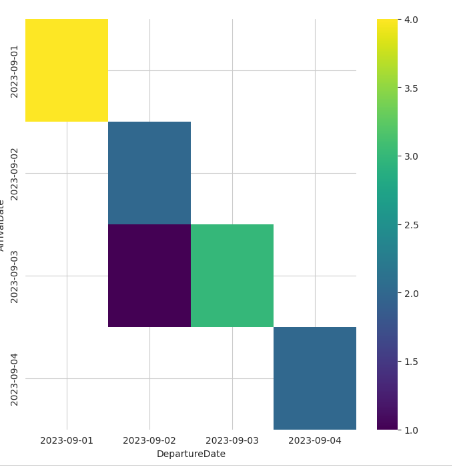
- The graph shows the number of flights departing on each date in the dataset.

- It helps us identify which dates had the most frequent flight departures.

- We can see if there are any noticeable periods of higher or lower activity in flight schedules.

- It can be used to spot trends related to the day of the week or certain periods (holidays, peak seasons).

Heatmap



- Insights:

  - Shows the frequency of flights traveling from a departure date to an arrival date.

  - Helps identify commonly traveled routes and travel patterns.