Interview Task – Data Engineering & Analytics

Dhanwin KB - 2347221 - Christ (Deemed to be University)

Analysis Report

Overview: This project analyzes flight data to derive insights about delays, airline performance, and scheduling patterns. It includes data cleaning, normalization, analysis, and visualization components.

Requirements:
Python 3.7+
pandas
sqlite3
seaborn
matplotlib

Install the required packages:
pip install pandas sqlite3 seaborn matplotlib

Instructions :
1.Place your flight data CSV file in the project directory and name it flights.csv.
2.Run the main script: python main.py
3.Follow the prompts to handle duplicate and inconsistent entries.
4.The script will generate:
  *A cleaned and normalized CSV file: transformed_dataset.csv
  *An SQLite database: flights.db
  *Several visualizations of the data
  *A summary of insights in the console output

File Descriptions:
main.py: The main script that orchestrates the data processing and analysis pipeline.
flights.csv: The input data file.
transformed_dataset.csv: The cleaned and normalized output data file.
flights.db: SQLite database containing the processed data.

Functions:
load_and_preprocess_data(): Loads and initially processes the data.
handle_missing_values(): Deals with NaN values in the dataset.
remove_duplicates(): Identifies and removes duplicate entries.
handle_inconsistent_entries(): Identifies and handles inconsistent time entries.
plot_delay_histogram(data) : Generates Histogram for Delay Distribution
plot_delay_by_airline(): Visualizes delays by airline over time.
calculate_average_delay_per_airline(): Computes average delays for each airline.
analyze_delay_by_departure_time(): Analyzes the relationship between departure time and delays.
plot_delay_distribution_by_airline(): Visualizes the distribution of delays by airline.
save_to_sqlite(): Saves the processed data to an SQLite database.

A sample dataset was provided to clean, normalize, and analyze for deriving relevant insights. The dataset included various flight details such as flight numbers, departure dates and times, arrival dates and times, airlines, and delays.

The dataset had the following issues to be handled:

a. **Inconsistent date/time formats** for Departure/Arrival dates (MM/DD/YYYY) and times (HH AM/PM).
b. **Missing values** in the DelayMinutes column.
c. **Duplicate flight entries** for some flights.
d. **Inconsistent time entries**, with ArrivalTime later than DepartureTime on the same day in some cases.

How it was handled :

For handling **inconsistent date/time formats,** Data was read into a pandas dataframe and stored a copy of the dataframe as backup. The to_datetime() function was used to convert dates from MM/DD/YYYY. By default, to_datetime() automatically converts dates into the YYYY-MM-DD format. Departure and Arrival Times were also converted using to_datetime(), and then formatted into a 24-hour format using the dt.strftime('%H:%M') function.

For handling **missing values**, NaN values in the DelayMinutes column were filled using the median of the respective Airline group. The median was chosen as a measure because it is less affected by outliers compared to the mean, providing a more accurate central tendency for flight delays. Unlike the mode, which represents the most frequent value, the median gives a better representation of the typical delay duration, especially in cases where delay distributions are skewed or have extreme values.

For handling **duplicate flight entries**, The user is shown the list of duplicate entries and prompted to review and keep/remove the entries. Duplicate entries were identified based on the combination of the FlightNumber, DepartureDate, and DepartureTime columns.

Duplicate Entries:

```
Duplicates found in the following entries:
  FlightNumber DepartureDate DepartureTime ArrivalDate ArrivalTime           Airline  DelayMinutes FlightDuration  FlightDuration (Minutes)
0       AA1234    2023-09-01        08:30  2023-09-01       10:45  American Airlines         15.0         02:15                     135.0
3       AA1234    2023-09-01        08:30  2023-09-01       22:45  American Airlines         30.0         14:15                     855.0
```

Unique Entries :

```
6       AA1234    2023-09-02        20:30  2023-09-03       10:45  American Airlines         60.0         14:15                     855.0
9       AA1234    2023-09-03        08:30  2023-09-03       10:00  American Airlines         15.0         01:30                      90.0
```

For the sample dataset, upon comparing the unique entries, it was observed that the AA1234 flight departing at 20:30 arrived approximately 14 hours later, while the one departing at 08:30 arrived about 1.5 hours later.

```
Duplicates found in the following entries:
  FlightNumber DepartureDate DepartureTime ArrivalDate ArrivalTime           Airline  DelayMinutes FlightDuration  FlightDuration (Minutes)
0       AA1234    2023-09-01        08:30  2023-09-01       10:45  American Airlines         15.0         02:15                     135.0
3       AA1234    2023-09-01        08:30  2023-09-01       22:45  American Airlines         30.0         14:15                     855.0
```

In contrast, the duplicate entry with a departure time of 08:30 arrived in about 2.15 hours, which is close to the 1.30 hours we noted earlier for that same departure time. However, the second entry took approximately 14.15 hours during the same period, indicating a potential anomaly. It's possible that the departure time was recorded incorrectly for the second duplicate entry; if it had been 20:30, like the earlier unique entry, it would have aligned with the expected pattern and taken about 14:15 hours, as observed previously. So **for analysis, the second duplicate entry was dropped when prompted.**

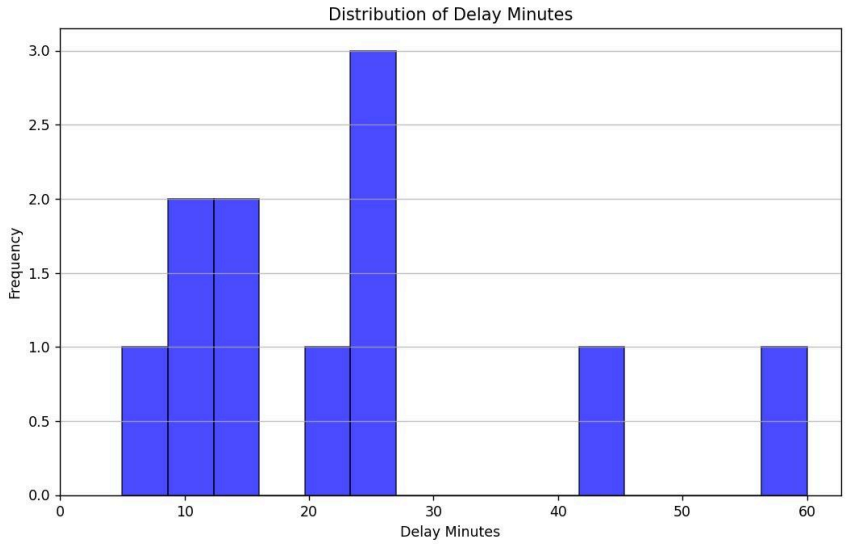
For handling **inconsistent time entries**, The dataset was scanned for entries where the ArrivalTime was recorded as being earlier than the DepartureTime on the same day, as well as for instances where the FlightDuration exceeded a reasonable threshold (e.g., 480 minutes). The entries were compiled and displayed for user review. The user was prompted to decide whether to remove these inconsistent entries. If the user chose to proceed, they could specify which entries they wished to keep.

Created a new column for **FlightDuration** by calculating the difference between **DepartureTime** and **ArrivalTime** on the same day. Converted time objects into datetime objects which combine both the date and the time, allowing us to perform subtraction. Subtraction between datetime objects is a built-in feature of Python's datetime module.

The dataset was then analyzed using visualizations :

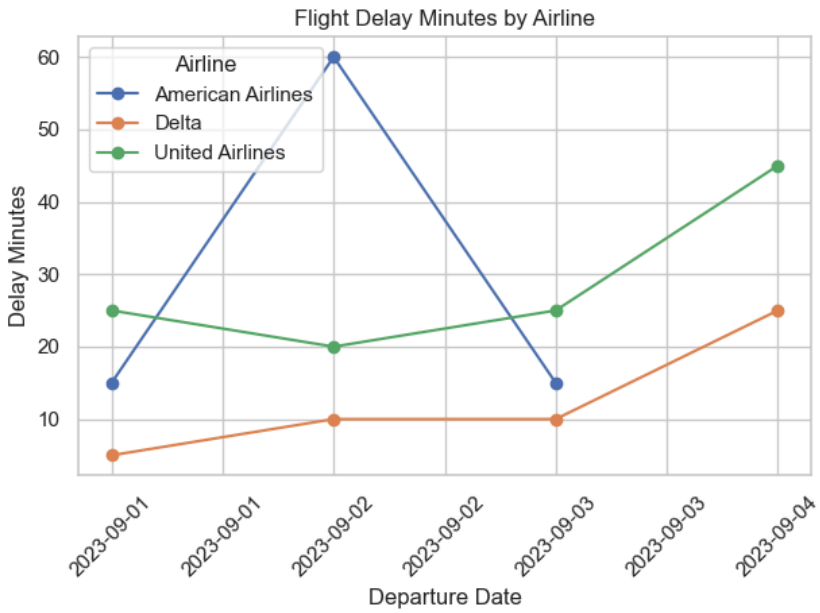
1. **To analyze the distribution of delays and to identify any trends or patterns :**

A **Histogram of DelayMinutes** was plotted to display the frequency distribution of delay minutes.



The x-axis represents delay duration in minutes, while the y-axis shows how often each delay length occurs.Our **most common delays fall in the 20-30 minute range**, occurring 3 times in our sample. Longer delays are less frequent but still present: we see 1 occurrence each in the 40-50 and 50-60 minute ranges.

Then a **Line Chart** was plotted with the delay times for multiple airlines on the same chart for delay comparison.



It was observed that **Delta Airlines recorded the shortest delays** among the three airlines, followed by United Airlines and then American Airlines.However, **delays across all airlines show an upward trend toward the end**, indicating a lack of improvement in delay management for these carriers. The limited data indicates **a significant spike in delays for American Airlines on September 2, 2023**, suggesting a particularly problematic day for the airline. Meanwhile, **September 3, 2023, was challenging for both Delta and United Airlines, as both experienced delays**, contributing to a broader issue for these carriers.

2. **Calculated the average delay for each airline :**

```
Average Delay per Airline:
             Airline  AverageDelay (in Minutes)
0  American Airlines                      30.00
1              Delta                      12.50
2    United Airlines                      28.75
```

Excluding the significant delay of 60 minutes recorded by American Airlines on September 2, 2023, **United Airlines had the highest average delay at 28.75 minutes. Delta Airlines had the lowest average delay at 12.50 minutes.**

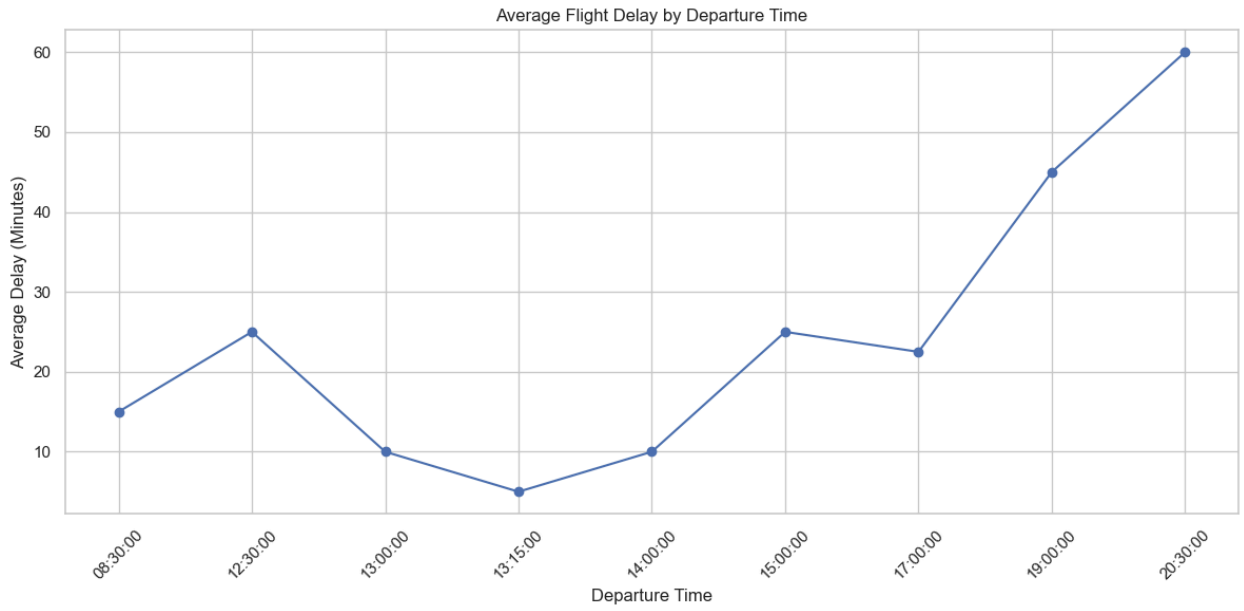
3. **To Identify relationships between flight delays and departure times (e.g., are flights departing later in the day more likely to be delayed?)**

The average delay for respective departure time was calculated.

	Departure Time	Average Delay (Minutes)
0	08:30:00	15.0
1	12:30:00	25.0
2	13:00:00	10.0
3	13:15:00	5.0
4	14:00:00	10.0
5	15:00:00	25.0
6	17:00:00	22.5
7	19:00:00	45.0
8	20:30:00	60.0

From the calculations, it is possible that flight delays increase as departure times get later in the day.

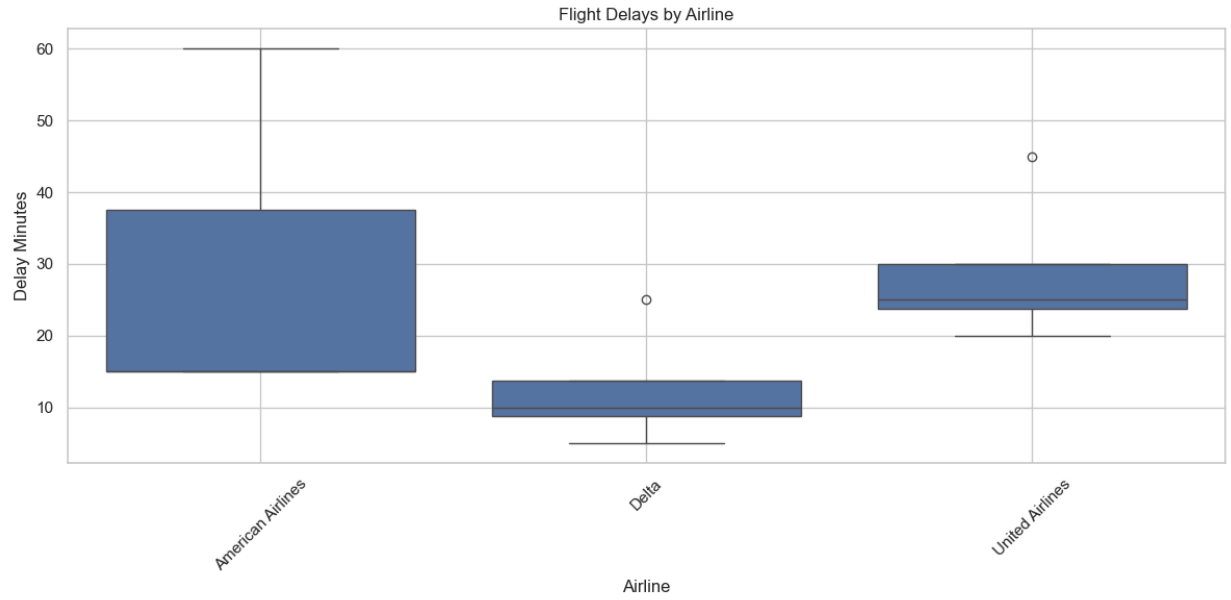
A **line chart** between Average Flight Delay by Departure Time was plotted.



The line chart indicates that, **starting from 13:15:00, the average delay time shows a steep increase,** visually confirming a **positive relationship between later departure times and increased flight delays.**

4. **To determine if there is a significant difference in delays between different airlines.**

A box plot for **Flight Delays by Airline** was generated.



The box plot reveals a clear difference in delay times among the three airlines. **American Airlines has the highest median delay and the largest variability,** indicating a tendency toward **more severe delays.** In contrast, **Delta Airlines** shows the **shortest median delays and consistent performance. United Airlines occupies a middle ground in terms of both median delay and variability.**The noticeable differences in median values, interquartile ranges, and the presence of outliers indicate significant disparities in delay management across the airlines.

Delta Airlines:
Median Delay: 10 minutes
Variability: Small IQR indicates consistent performance, with only 1 outlier.

United Airlines:
Median Delay: 25 minutes
Variability: Larger IQR suggests greater variability in delays compared to Delta, also with 1 outlier.

American Airlines:
Median Delay: 37.5 minutes
Variability: Largest IQR indicates the highest variability in delays, but no outliers present.

The absence of lower (or upper whisker) indicates that the minimum or maximum is equal to the lower/ upper quartile.

Insights Summary :

- **Delta Airlines** recorded the shortest delays among the three airlines, followed by United Airlines and American Airlines. However, **there is an upward trend in delays across all airlines, indicating ongoing issues in delay management.**
- **American Airlines experienced a significant spike in delays** on September 2, 2023, particularly problematic for the airline. **On September 3, both Delta and United Airlines also faced delays, reflecting a broader challenge for these carriers.**
- Excluding the 60-minute delay for American Airlines on September 2, **United Airlines had the highest average delay at 28.75 minutes**, while American Airlines averaged 30.00 minutes overall. **Delta Airlines recorded the lowest average delay at 12.50 minutes.**
- The line chart shows that **from 13:15:00 onward, average delay times steeply increase, confirming a positive relationship between later departure times and increased delays.**
- Significant differences in delays between the airlines was found. **American Airlines had the highest median delay and the largest variability**, indicating **a tendency toward more severe delays**. **Delta Airlines showed the shortest median delays** and consistent performance.

Suggestions / Recommendations :

- **The peak delay times were identified in the evening hours (14:00 to 20:30)**, indicating high-traffic periods or conflicting scheduling. **Deploying additional resources or adjusting flight schedules during these times could help minimize delays and congestion**.
- **American Airlines** needs to **investigate the duplicate data entries that caused inconsistencies**, as well as **the significant delays experienced on September 2, 2023**, to implement targeted improvements. They should also **publicly address these delays and work on rebuilding customer trust, satisfaction, and reliability in their airline**. This could include **introducing loyalty programs and providing real-time updates on delays and estimated wait times.**
- Given that **United Airlines** had the highest average delay (excluding American Airlines' significant delay), **a detailed review of flight schedules and operations during peak hours** is essential. **Implementing pre-flight checks and quicker boarding processes could help reduce the chances of delays** as flights approach peak times.
- **Delta Airlines**' performance should be considered a benchmark for the other airlines. **Workshops to analyze Delta's operational strategies that contribute to shorter delays** should be organized, with the aim of **implementing similar practices across other carriers**.
- Additionally, airlines should **collaborate to assess whether their schedules are contributing to bottlenecks** that lead to runway congestion, prolonged security checks, and lengthy boarding processes.
- Moreover, **more comprehensive data is required, such as destination, distance, weather conditions, and specific causes of delays, to enhance the consistency and processing of the information.**