Airline Delay Prediction

Table of contents:

Abstract:

This project aims to develop a predictive model for airline delays using historical flight data. The dataset includes flight information such as departure and arrival times, delay durations, carrier details, and weather conditions. By cleaning, transforming, and analyzing the data, the goal is to predict delays based on patterns in the data, providing insights into factors contributing to flight delays. This project will also involve building and evaluating machine learning models to forecast delays, which could assist airlines in improving operations and minimizing delay impacts on passengers.

Introduction:

Flight delays are a persistent issue in the aviation industry, causing inconvenience for passengers and significant financial losses for airlines. Predicting airline delays in advance can help both passengers and airlines make informed decisions to mitigate the impact of these delays. This project aims to build a predictive model using historical flight data to forecast delays and understand the factors contributing to them.

By analyzing flight schedules, carrier information, and external factors such as weather conditions, the project seeks to uncover patterns that affect flight punctuality. Key factors such as departure and arrival times, flight routes, and seasonal trends are evaluated to develop a robust model capable of predicting delays before they occur.

The outcome of this project could provide airlines with insights into improving operations and reducing delays, benefiting passengers by offering more accurate flight information. Additionally, this predictive model can be extended to assist airport management in allocating resources more efficiently and preparing for potential disruptions.

Dataset Overview:

The dataset used in this project consists of historical flight records from various U.S. airlines, sourced from the Bureau of Transportation Statistics. It includes a range of attributes relevant to predicting flight delays, such as:

Day of Week and Month: Temporal variables indicating the day of the week and the month of the flight.
Flight Number: A unique identifier for each flight.
Airline Carrier: The airline operating the flight (e.g., Delta, American Airlines).
Origin and Destination Airports: The airports from which the flight departs and arrives.
Scheduled Departure and Arrival Times: Planned departure and arrival times for each flight.
Actual Departure and Arrival Times: The recorded departure and arrival times.
Delay Duration: The delay time in minutes, which is the primary target variable for prediction.
Delay : It indicates whether the flight has a delay of more than 15 minutes or not.
Cancellation: It indicates whether the flight is canceled or not.
Diverted: It indicates whether the flight is diverted or not.

This dataset spans several years and covers a large number of domestic flights across various airports, providing a broad base for analysis and predictive modeling. It includes information on both delayed and on-time flights, making it suitable for classification and regression tasks.

Data Collection

The dataset was obtained from the U.S. Department of Transportation's Bureau of Transportation Statistics, which provides comprehensive flight data. Additionally, we integrated weather data from a publicly available weather API to capture environmental factors that could influence flight delays. Data was collected for several years.

Data Integration

The dataset comprises 33495946 data of monthly flight records from 2022, 2023, and 2024, initially stored in separate CSV files. To facilitate a comprehensive analysis, we integrated these datasets into a single, unified DataFrame. This integration involved two key steps:

1. Loading Multiple CSV Files: We loaded the monthly flight data into DataFrames for each year. Loading data in smaller chunks (by month) allows for more flexibility and improves performance when handling large datasets.

2. Concatenating DataFrames: After loading the monthly data for each year, we used the concat() function from pandas to combine them into yearly DataFrames. Subsequently, we merged the yearly DataFrames (2022, 2023, and 2024) into one unified DataFrame. This process enabled us to perform data operations efficiently across multiple years, simplifying filtering, grouping, and analysis without the need to reference multiple smaller DataFrames.

By integrating the monthly and yearly data, we created a cohesive dataset that allows for more comprehensive analysis of flight operations and delays over time, enabling us to identify trends and patterns across different time periods in Colorado.

Data Exploration

To ensure the dataset was accurate, consistent, and ready for predictive modeling, we undertook several preparation and cleaning steps. These steps involved handling missing data, integrating external datasets, transforming features, and performing data quality checks to enhance the dataset's utility.

Data Cleaning

Data cleaning is a critical process to ensure that the dataset is accurate, reliable, and ready for analysis. It involves identifying and resolving errors, inconsistencies, and missing values. Below are the key steps taken during the cleaning process:

- **Removing Duplicates:** Duplicate records can distort analysis by introducing redundant data. To ensure each flight record was unique, we removed duplicate rows from the dataset.
- **Handling Missing Data:** Missing data can negatively impact the quality and consistency of the dataset. We identified missing values in key columns such as DEP_TIME, DEP_DELAY_NEW, ARR_TIME, ARR_DELAY_NEW, and CANCELLATION_CODE. We used appropriate techniques, such as imputing missing values where relevant or excluding records where data was not available and filled the missing values with median value.

Data Reduction

After the initial data cleaning, we applied data reduction techniques to streamline the dataset by removing unnecessary information while preserving essential insights. This made the dataset more manageable for analysis. The key steps involved in data reduction were:

1. Dropping Columns with Excessive Missing Data: We removed the CANCELLATION_CODE column, which contained a large number of missing values and was irrelevant to our analysis of overall flight performance. This helped reduce the dataset size and eliminate unnecessary information.

2. Filling Missing Numeric Values: Instead of removing rows with missing data, we filled missing values in numeric columns using the median, which is a robust measure of central tendency. This approach ensured that no data was lost, and the filled values accurately represented the central trends of each column.

3. Checking Data Quality: After handling missing values, we verified that the dataset was complete by using the isnull().sum() function, confirming that no missing values remained in any column.

4. Renaming Columns: To improve readability and consistency, we renamed several columns (e.g., OP_UNIQUE_CARRIER to CARRIER, ORIGIN_CITY_NAME to ORIGIN_CITY) using the rename() function, enhancing the clarity of the dataset.

By applying these techniques, we reduced the dataset size from 33,495,946 entries to 31,973,403 entries, making it more efficient to process without losing valuable insights.

Data reduction was essential for improving the overall quality and efficiency of our analysis by focusing on key aspects of the data and ensuring it was both manageable and informative.

Data Type Identification

Identifying the types of data in the dataset is crucial for applying the appropriate analysis techniques. Categorical data is used for grouping and filtering, while numerical data is typically summarized and used in statistical models. This distinction ensures that our analysis is meaningful and relevant.

- **Categorical Columns**

Categorical data consists of discrete values, often representing labels or categories. In our dataset, the categorical columns include flight dates, carriers, and locations, which are essential for grouping and filtering during analysis.

- **Numerical Columns**

Numerical data consists of continuous or discrete numerical values. The numerical columns in our dataset include flight numbers, scheduled and actual times, delays, and indicators for cancellations and diversions.

By identifying and categorizing data types, we were able to apply the correct analysis methods, ensuring accurate insights. Categorical data was used for grouping and filtering, while numerical data was summarized and used for statistical modeling, leading to more informed and precise analysis.

Statistical Summary

We generated a statistical summary of the numerical columns using the describe() function, providing key statistics such as count, mean, standard deviation, and percentiles:

- The mean departure delay (DEP_DELAY_NEW) is approximately 16.64 minutes, but most flights have delays below 13 minutes, as shown by the 75th percentile.
- Most flights are neither canceled nor diverted, with values for these columns predominantly being 0.
- The distribution of flight times aligns with typical flight schedules, with the median scheduled departure and arrival times falling within expected ranges.

- **Measures of Central Tendency**

  To understand the central tendency of the normalized delay columns, we calculated:

  Departure Delay:

  Mean: 0.235

  Median: 0.0 (over half of the flights had no delay)

  Mode: 0.0 (on-time departure is the most common)

  Arrival Delay:

  Mean: 0.227

  Median: 0.0 (most flights arrived on time)

  Mode: 0.0 (on-time arrival is the most frequent outcome)

- **Measures of Dispersion**

  We also calculated the variance, standard deviation, and range to understand the spread of the normalized delays:

  Departure Delay:

  Variance: 0.18

  Standard Deviation: 0.424

  Range: 1.0 (data scaled between 0 and 1)

Arrival Delay:

Variance: 0.176

Standard Deviation: 0.419

Range: 1.0

These measures indicate that the variation in delays is relatively small after normalization, with both departure and arrival delays showing a standard deviation of less than 0.5.

Data Normalization

In this step, we applied data normalization to ensure that the values of numerical columns like DEP_DELAY (departure delay) and ARR_DELAY (arrival delay) were scaled to a common range. This is critical for machine learning algorithms that rely on distance metrics or feature comparisons. Normalization ensures that all values, typically between 0 and 1, are on a similar scale, improving comparability and algorithm performance.

- **Applying Min-Max Normalization**

  We used the MinMaxScaler from the sklearn.preprocessing module to scale the values of the delay columns. By applying Min-Max normalization, we brought the values of DEP_DELAY and ARR_DELAY to a uniform range, typically between 0 and 1.

Data normalization ensures that the values are uniformly scaled, which is essential for improving machine learning model performance. By normalizing the DEP_DELAY and ARR_DELAY columns, we achieved a consistent scale across features, allowing for more accurate comparisons and better results in predictive modeling.

Data Transformation

To prepare the dataset for machine learning models, we applied data transformation techniques to standardize numerical features and encode categorical data. This process ensures that all features are in a suitable format for model training and analysis.

1. Standardizing Numerical Features

Numerical features in the dataset, such as flight numbers (FL_NUM) and delays (DEP_DELAY, ARR_DELAY), vary in scale. To handle this, we applied standardization using the StandardScaler from sklearn.preprocessing. Standardization ensures that each numerical feature has a mean of 0 and a standard deviation of 1, allowing all features to contribute equally during analysis and model training.

2. One-Hot Encoding Categorical Features

Categorical data, such as CARRIER and ORIGIN_CITY, cannot be directly used by machine learning algorithms. To convert these into a numerical format, we applied One-Hot Encoding, which transforms each category into a binary column. For example, each unique airline carrier is represented as a separate binary column, allowing categorical features to be included in the analysis.

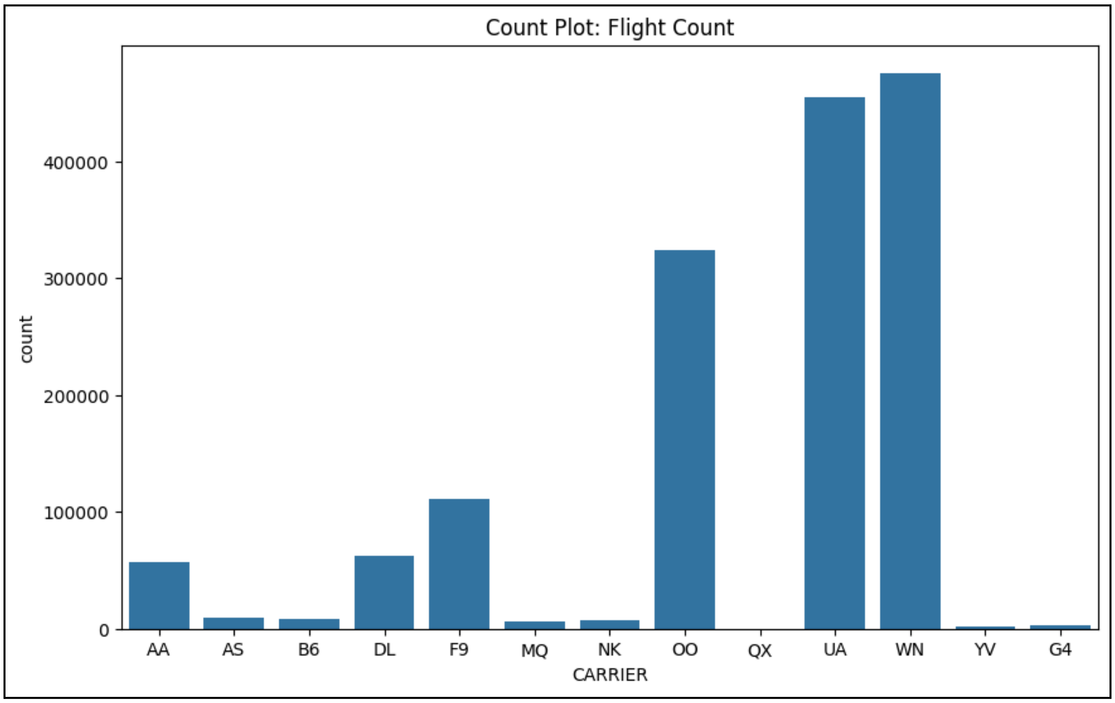
Data transformation is essential for preparing the dataset for machine learning. By standardizing numerical columns and encoding categorical columns, we ensured that our data is in a consistent and usable format, improving model performance and enabling easier comparisons across features.

Data Visualization:

Data visualization plays a crucial role in understanding patterns, trends, and insights within the dataset. Through different types of visualizations, we can explore the relationships between various features, identify distributions of delays, and analyze the performance of flights across time periods. Below are some of the key visualizations that were created to help interpret the data effectively.

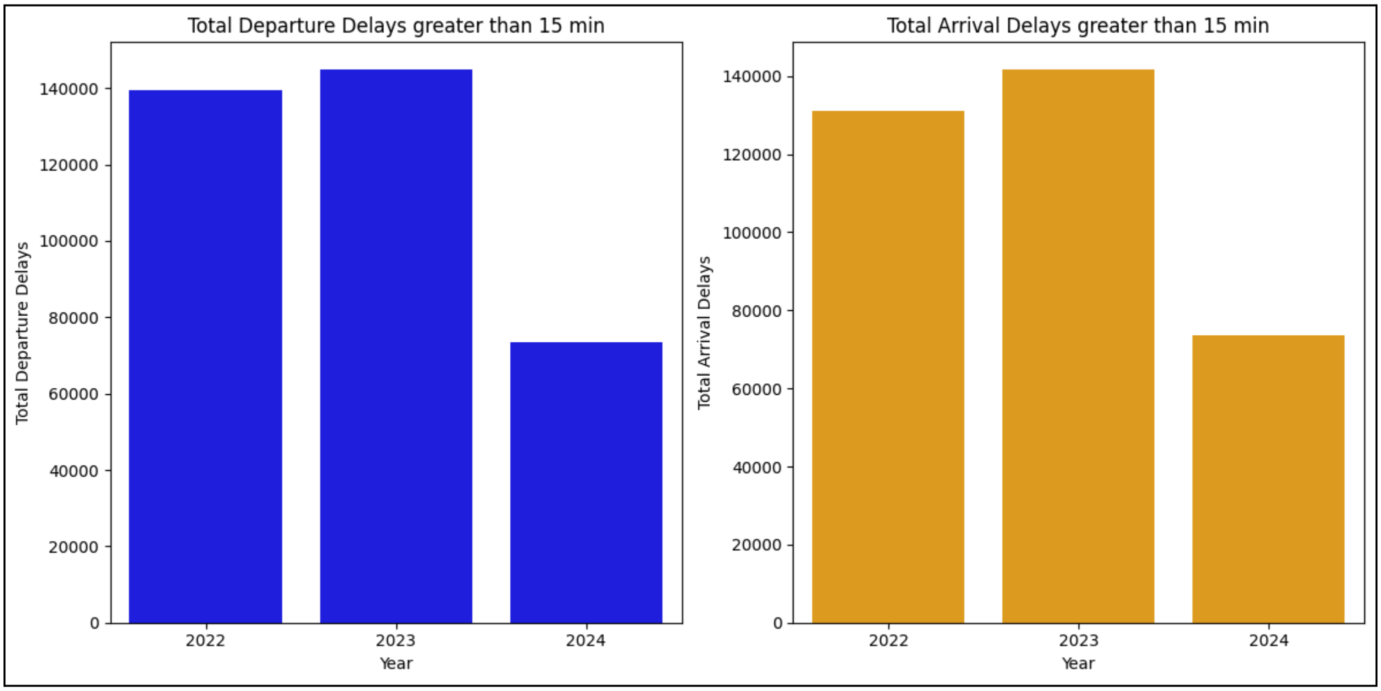
1. Count Plot: Flight Count by Carrier

A count plot shows the number of flights operated by each carrier. The plot helps compare the frequency of flights across different airlines, offering insights into which carriers have the highest or lowest flight volume. From the chart, it's clear that certain airlines, such as UA and WN, dominate in terms of flight count, while others have significantly fewer flights.



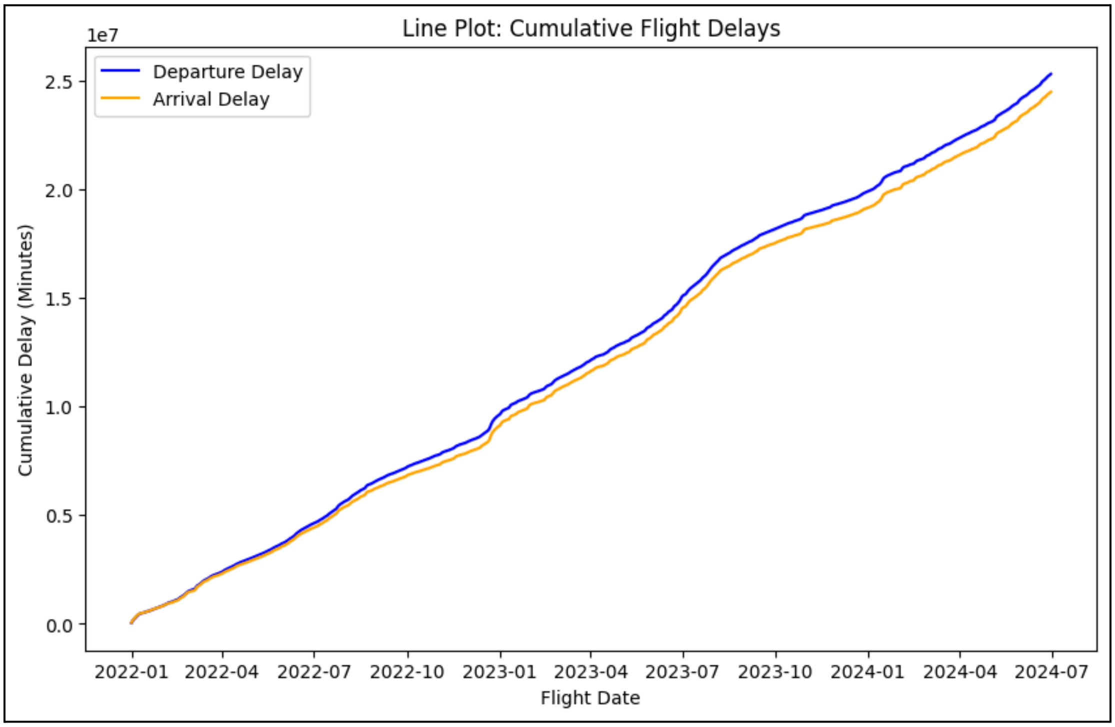
2. Bar Plot: Total Departure and Arrival Delays by Year

A bar plot was used to display the total departure and arrival delays across different years. This visualization helps analyze how total delays have evolved over the years. By displaying the delays side-by-side, we can quickly see trends and identify the years with the most severe delays. This plot shows the total departure delays over 15 minutes for each year. In 2023, there were the highest total departure delays, slightly higher than in 2022. The year 2024 shows a significant drop in departure delays compared to the previous two years.



3. Line Plot: Cumulative Flight Delays Over Time

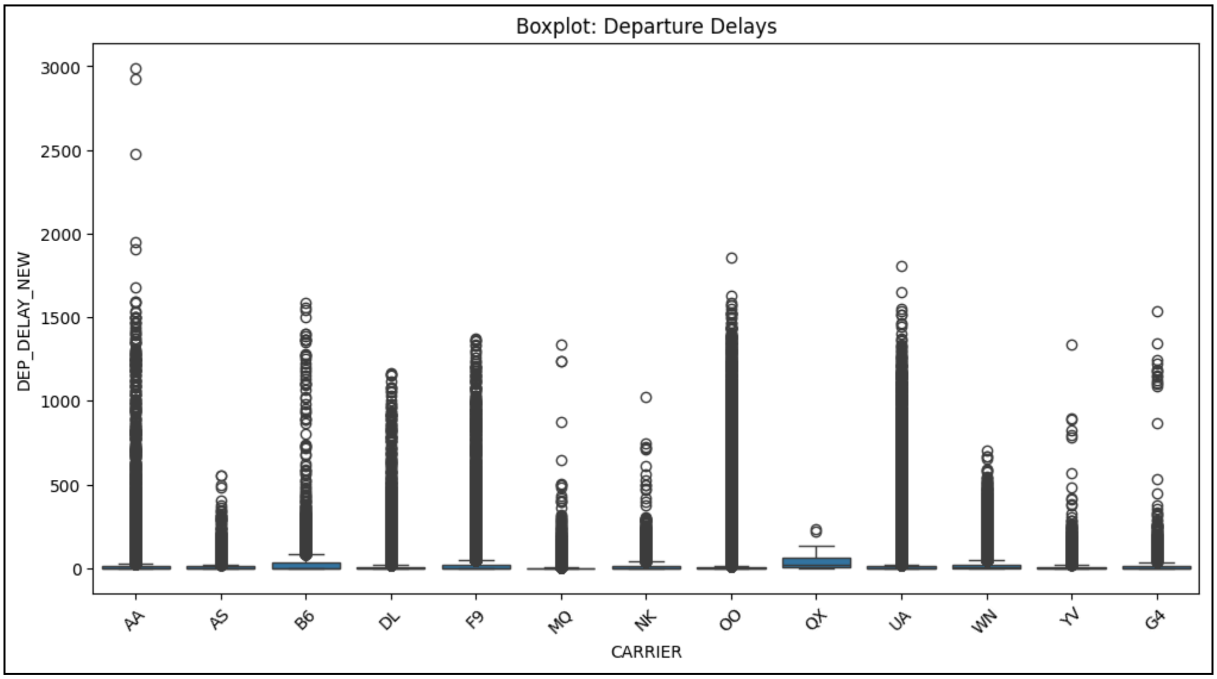
A line plot was created to show the cumulative departure and arrival delays over time. This line plot helps us visualize how delays have accumulated over the years, giving insight into long-term trends in flight delays. The blue line represents cumulative departure delays, while the orange line represents cumulative arrival delays. We can see that delays have steadily grown over time, with no significant periods where delays decreased or leveled off. By comparing cumulative departure and arrival delays, we can infer that departure delays slightly outpace arrival delays over time.



4. Boxplot: Distribution of Departure Delays by Carrier

A boxplot visualizes the distribution of departure delays for each carrier, showing the spread and potential outliers in the delay data. By highlighting the outliers, it shows which airlines occasionally experience significant operational issues leading to major delays.

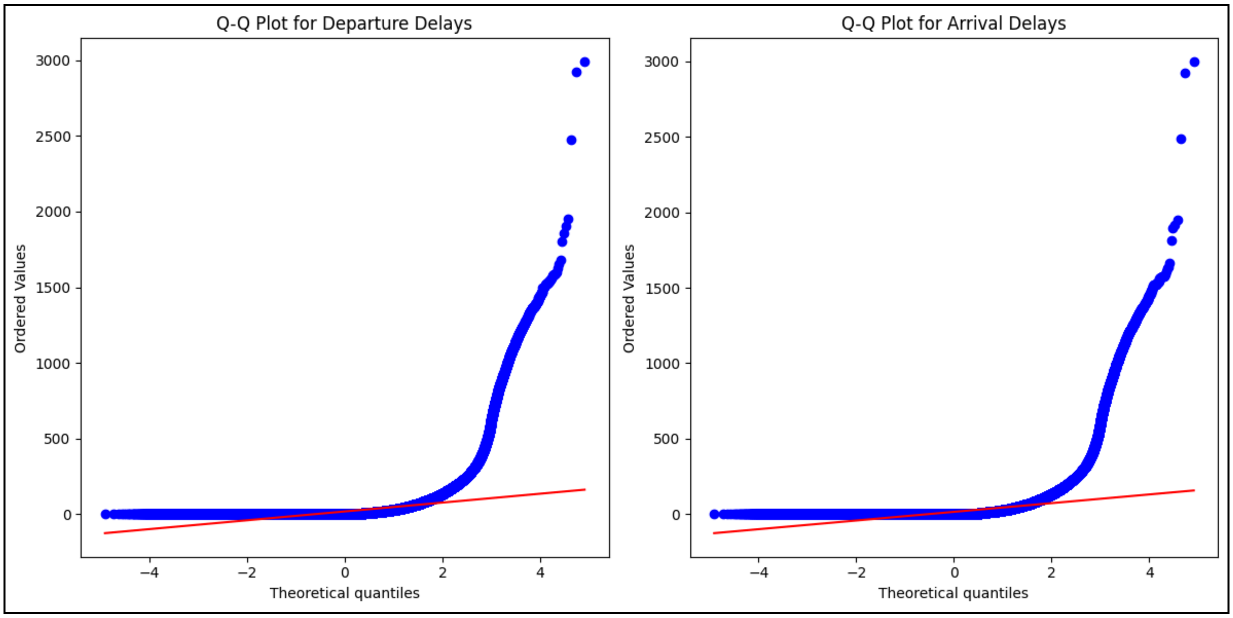
The x-axis represents the different carriers, while the y-axis shows the departure delays in minutes. Most carriers have small median delays, meaning that a significant portion of their flights are either on time or experience minor delays. Certain carriers, such as AA, F9, NK, and OO, have numerous outliers, indicating that while most flights experience short delays, some flights face significant delays of over 1,000 minutes. QX appears to have relatively fewer outliers and more consistent delays, with a much smaller range compared to other carriers.



5. Q-Q Plot: Normality of Departure and Arrival Delays

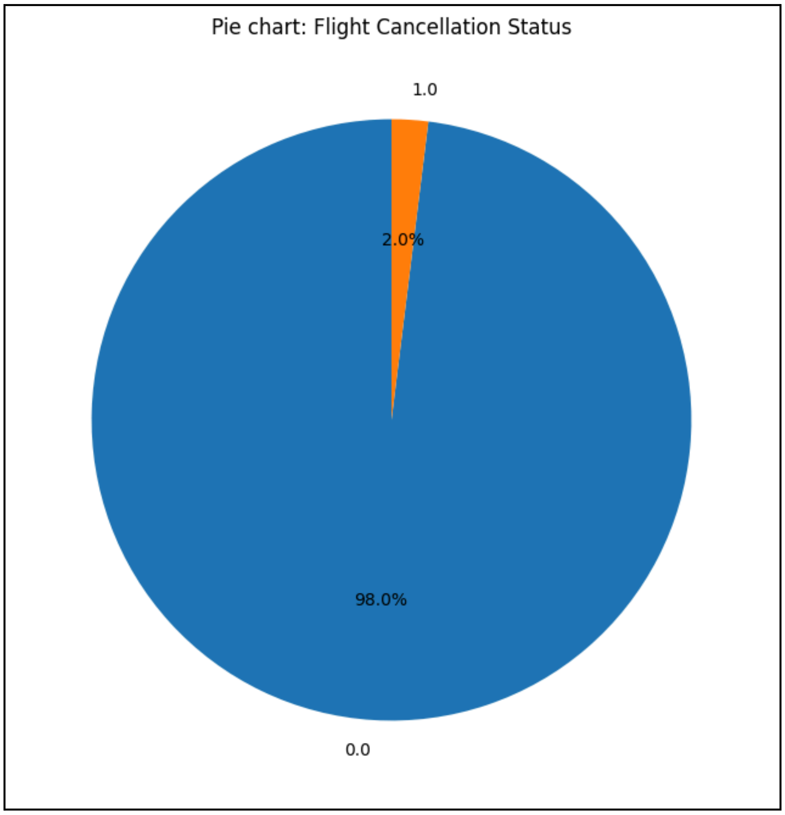
A Q-Q plot compares the distribution of departure and arrival delays to a normal distribution, allowing us to assess whether the delay data follows a normal distribution. It highlights the fact that both departure and arrival delays do not follow a normal distribution due to the presence of extreme outliers. Both departure and arrival delays exhibit heavy tails, with many delays far exceeding what would be expected under a normal distribution.These heavy tails are indicative of outliers flights that experienced extreme delays compared to the majority.

The data points cluster along the line for smaller values, meaning that most flights have relatively short delays, but the presence of extreme delays skews the overall distribution. This suggests that when analyzing flight delays, assuming a normal distribution may not be appropriate, and alternative statistical models that account for skewness or heavy tails should be considered.



6. Pie Chart: Flight Cancellation Status

A pie chart was created to show the proportion of canceled versus non-canceled flights. This pie chart quickly conveys the proportion of canceled versus non-canceled flights, allowing stakeholders to understand the reliability of flight operations. The blue portion of the pie chart represents flights that were not canceled (98% of the flights). The orange portion represents flights that were canceled (2% of the flights). The vast majority of flights (98%) in the dataset were not canceled, indicating that flight cancellations are relatively rare. Only 2% of the flights were canceled, which is a small fraction of the total flights in the dataset.



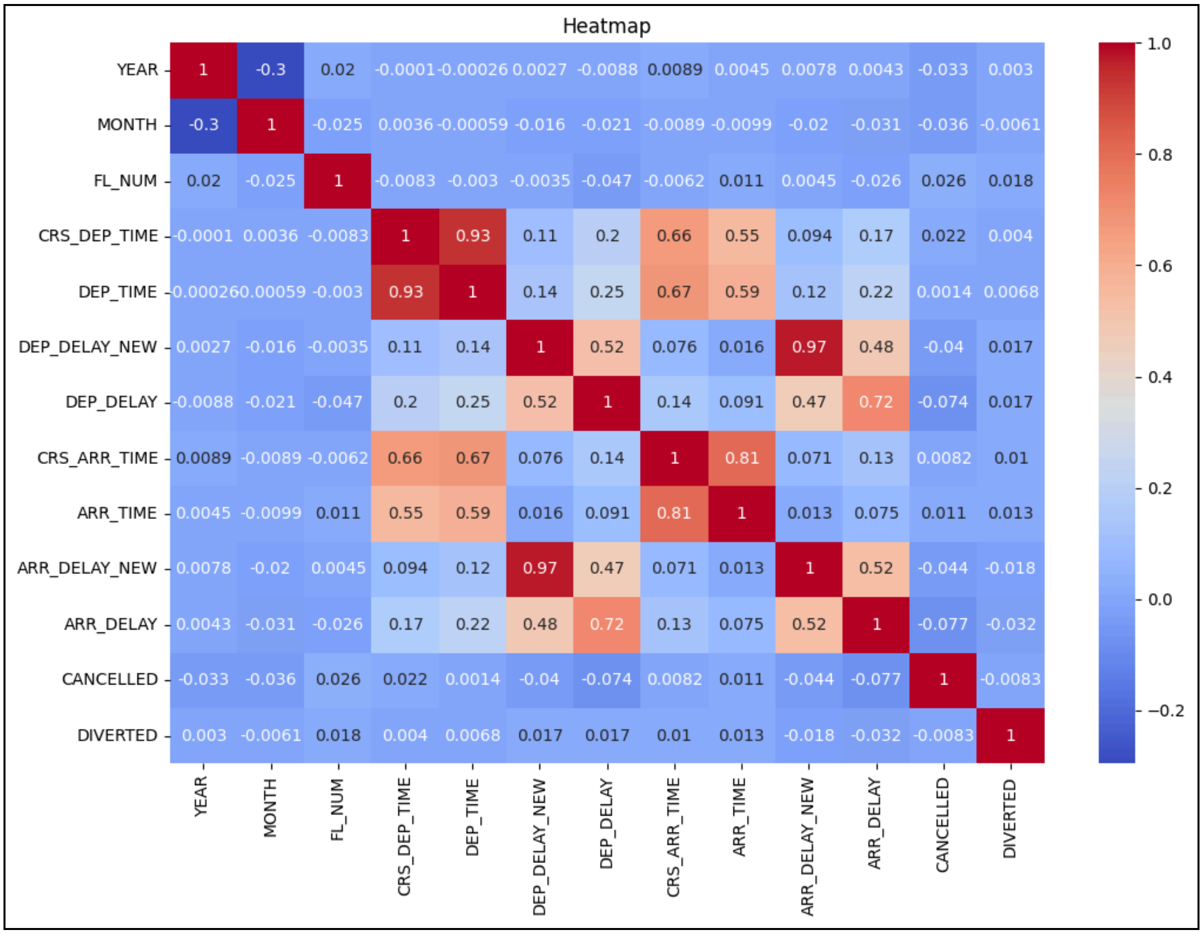
7. Heatmap: Correlation Between Numerical Features

A heatmap was generated to show the correlation between various numerical features in the dataset. The heatmap helps quickly identify relationships between variables, enabling us to focus on features that may be influencing each other. The color scale ranges from blue (negative correlation) to red (positive correlation).

The values inside each cell represent the correlation coefficient between two variables, ranging from -1 to 1.
A value of 1 indicates a perfect positive correlation.
A value of -1 indicates a perfect negative correlation.
A value of 0 suggests no correlation.

Insights:

- Departure Time and CRS Departure Time (DEP_TIME and CRS_DEP_TIME) show a strong positive correlation (0.93). This suggests that the actual departure time is closely aligned with the scheduled departure time.
- Arrival Time and CRS Arrival Time (ARR_TIME and CRS_ARR_TIME) also show a strong positive correlation (0.81), indicating a similar trend for arrivals.
- Departure Delay New (DEP_DELAY_NEW) and Arrival Delay New (ARR_DELAY_NEW) are highly correlated (0.97), which means that flights experiencing departure delays are very likely to experience arrival delays as well.



8. Pairplot: Relationships Between Delays, Cancellations, and Diversions

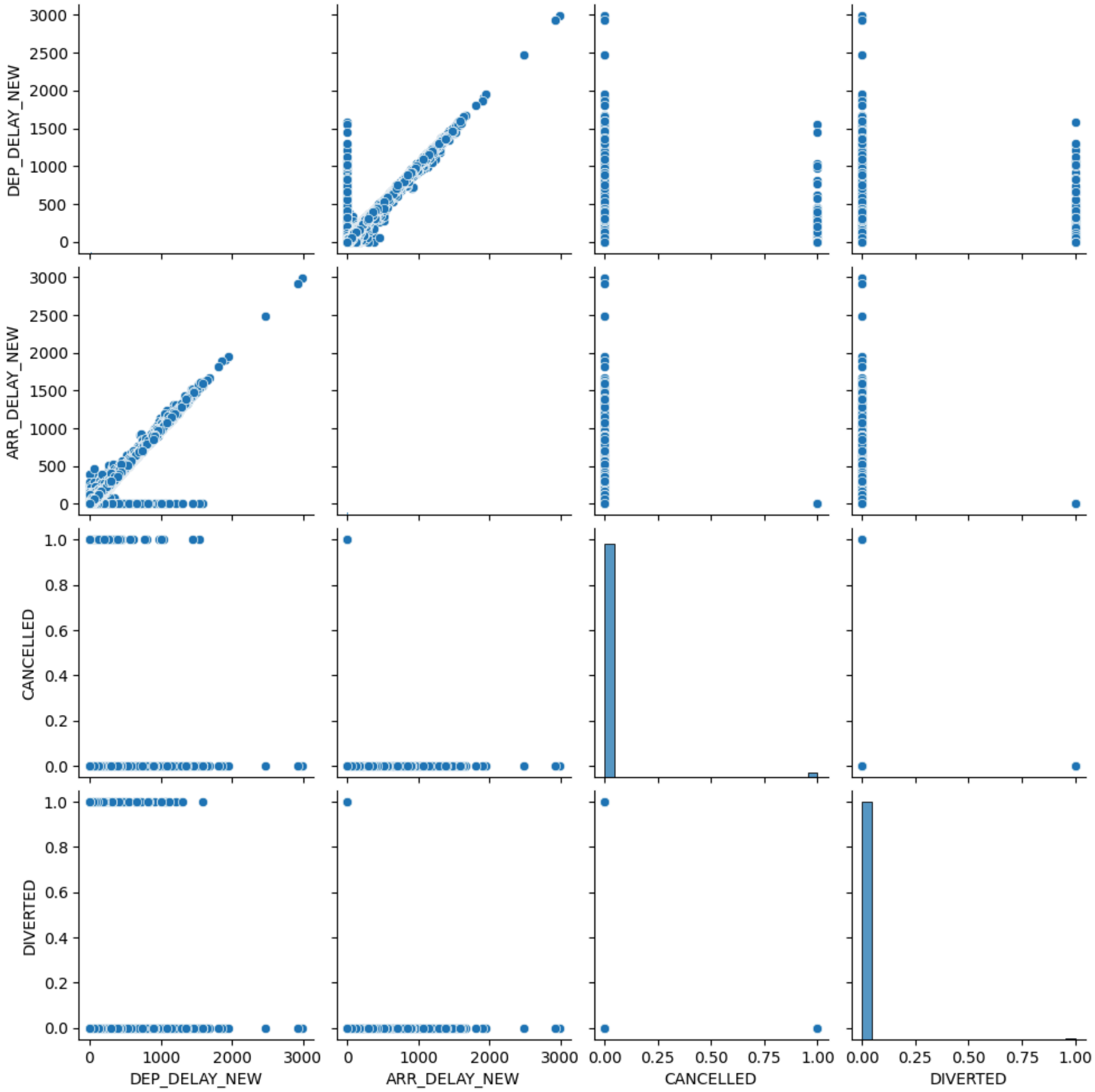
A pairplot allows us to visualize the pairwise relationships between multiple variables in the dataset. It creates a matrix of scatterplots for each combination of variables and shows the distribution of individual variables on the diagonal.

Insights:

- **Strong Correlation Between Delays:** Departure and arrival delays show a clear positive correlation—flights that depart late tend to arrive late.
- **Cancellations and Delays**: Most delayed flights are not canceled. Cancellations (CANCELED = 1) show no strong relationship with high delays.

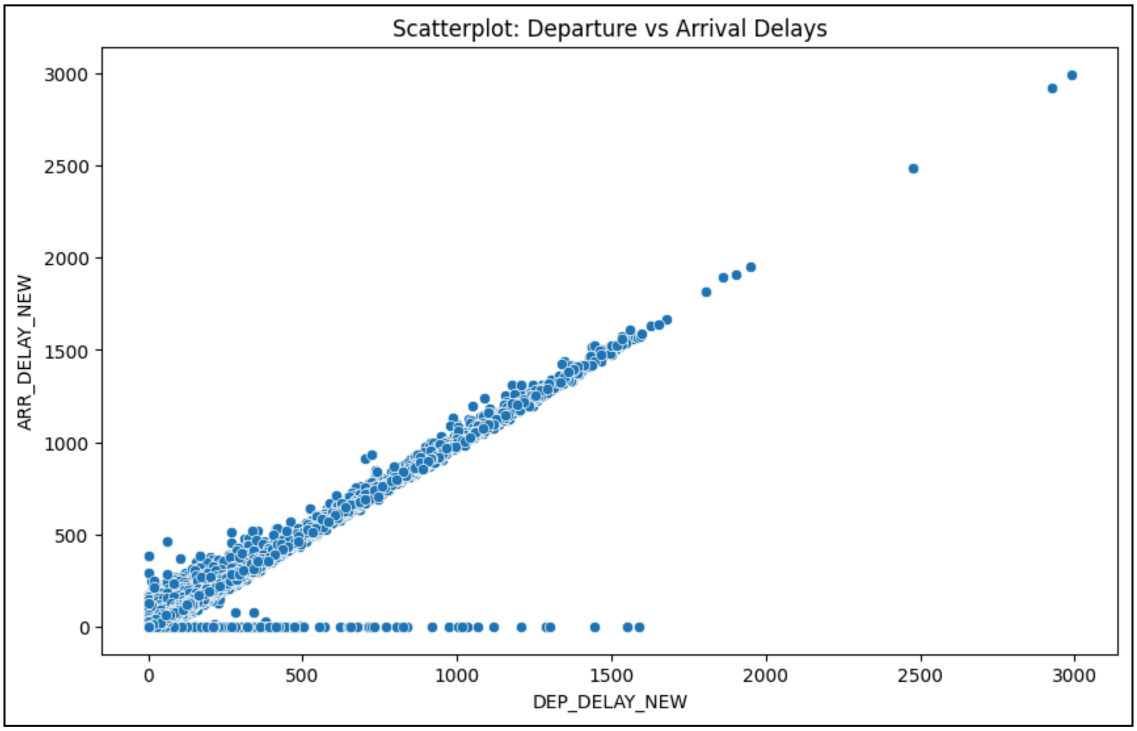
- **Diversions and Delays**: Most flights are not diverted (DIVERTED = 0). A few diverted flights have high delays, but diversions are rare overall.
- **Distribution**: Both departure and arrival delays are right-skewed, with most flights having minimal delays. Cancellations and diversions are infrequent, mostly skewed toward 0.

The pairplot will show how these variables are related to one another, providing a quick way to spot trends and correlations between flight delays, cancellations, and diversions.



9. Scatterplot: Departure vs Arrival Delays

A scatterplot was used to compare departure and arrival delays. This scatterplot provides a visual representation of how delays at departure influence delays at arrival. The strong correlation shown here supports the observation that flights with significant departure delays are likely to also experience significant arrival delays. There are a few extreme outliers where both departure and arrival delays exceed 1,500 minutes, but these instances are rare.

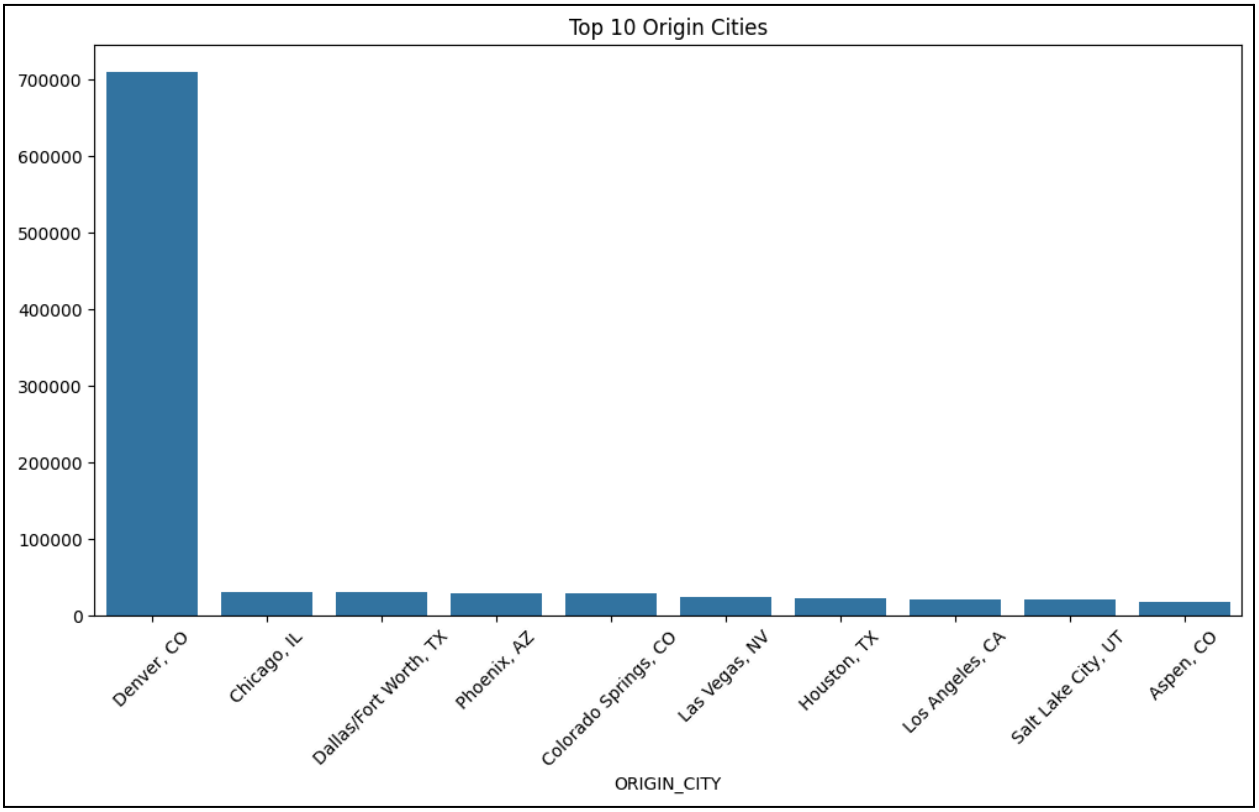


10. Bar Plot: Top 10 Origin Cities by Flight Count

A **bar plot** was used to display the top 10 origin cities based on the number of flights. This visualization helps to identify the cities that have the highest number of departing flights in the dataset.

Key Insights:

- Denver, CO is by far the busiest origin city in the dataset, with a flight count that dwarfs all other cities.
- The remaining top 9 cities have relatively similar flight counts, indicating that Denver's airport is a major hub in this dataset.
- The cities include major hubs across the U.S., such as Chicago, Dallas/Fort Worth, and Phoenix, but none come close to Denver's volume of flights.

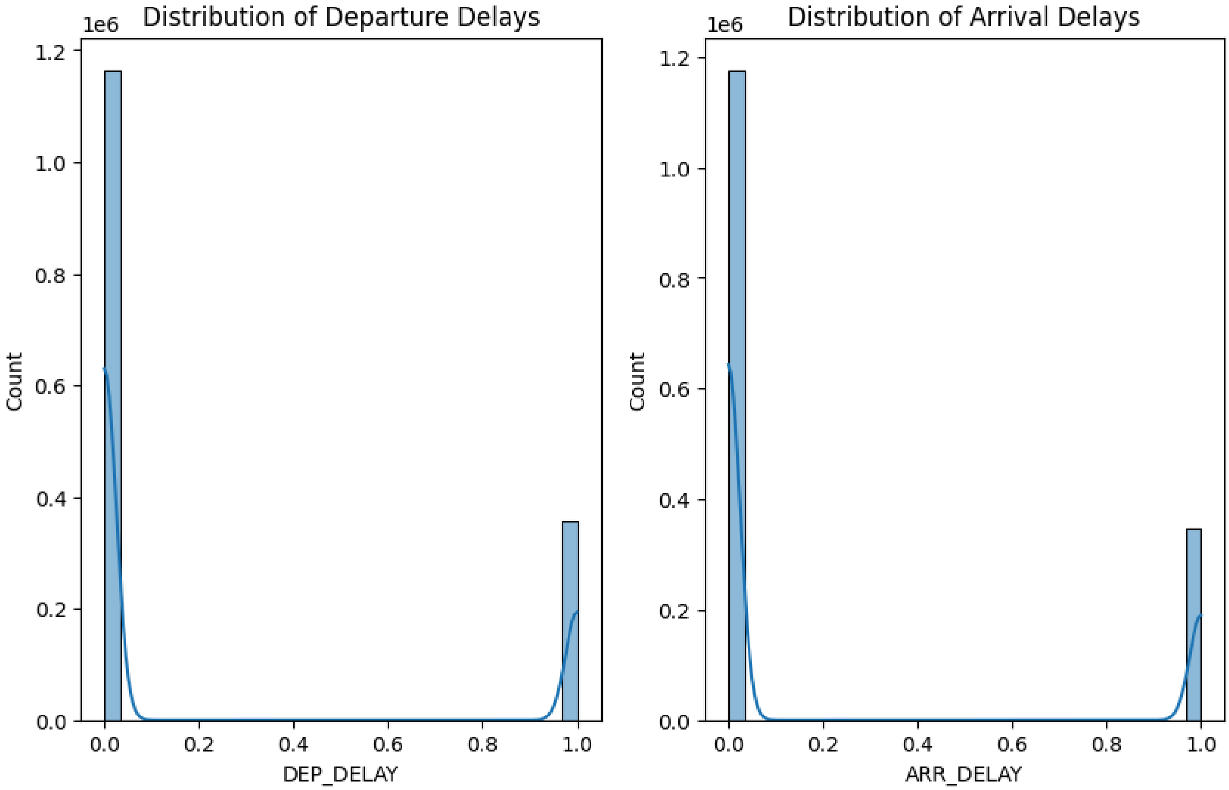


11. Histogram: Distribution of Departure and Arrival Delays

A histogram was used to visualize the distribution of departure and arrival delays. These histograms help visualize the frequency of delays and confirm that the dataset is dominated by flights with minimal delays.

Key Insights:

- In both plots, the majority of flights experience little to no delay, as seen by the large spike around 0, meaning that most flights depart and arrive on time.
- There is a smaller spike around 1 for both departure and arrival delays, which likely represents flights with significant delays.
- The distribution is highly right-skewed, indicating that while most flights are on time, a smaller number of flights face substantial delays.
- The gap between 0 and 1 suggests that delays are either very small or very large, with few flights experiencing moderate delays.

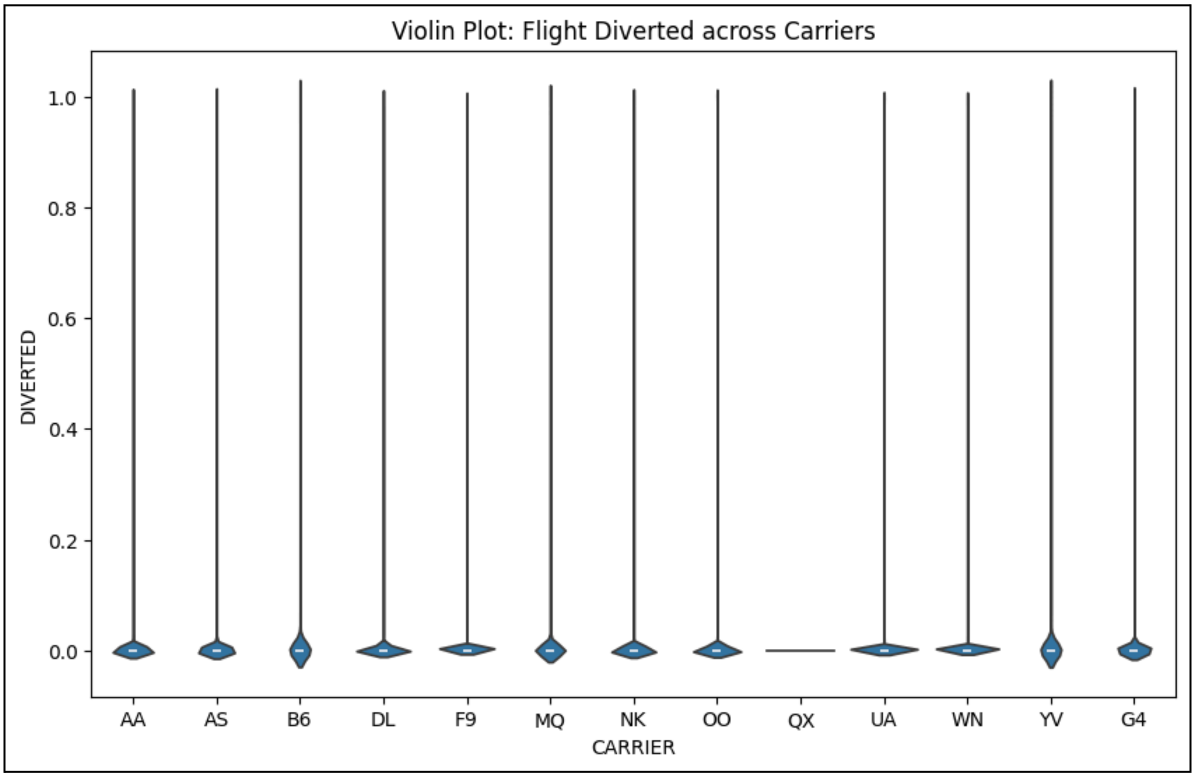


12. Violin Plot: Flight Diversions by Carrier

A violin plot was created to show the distribution of diverted flights across different carriers. This violin plot helps to visually assess the distribution of diversions across different airlines. It highlights that, although diversions do occur, they are generally infrequent for all carriers.

Key Insights:

- The majority of flights for all airlines have a value of 0 for DIVERTED, meaning that most flights are not diverted.
- There is a small amount of density near 1 (indicating diverted flights) for each airline, but diversions are relatively rare.
- The plot shows that diversions are fairly uniform across carriers, with no airline showing a significant deviation from others in terms of the likelihood of diversions.

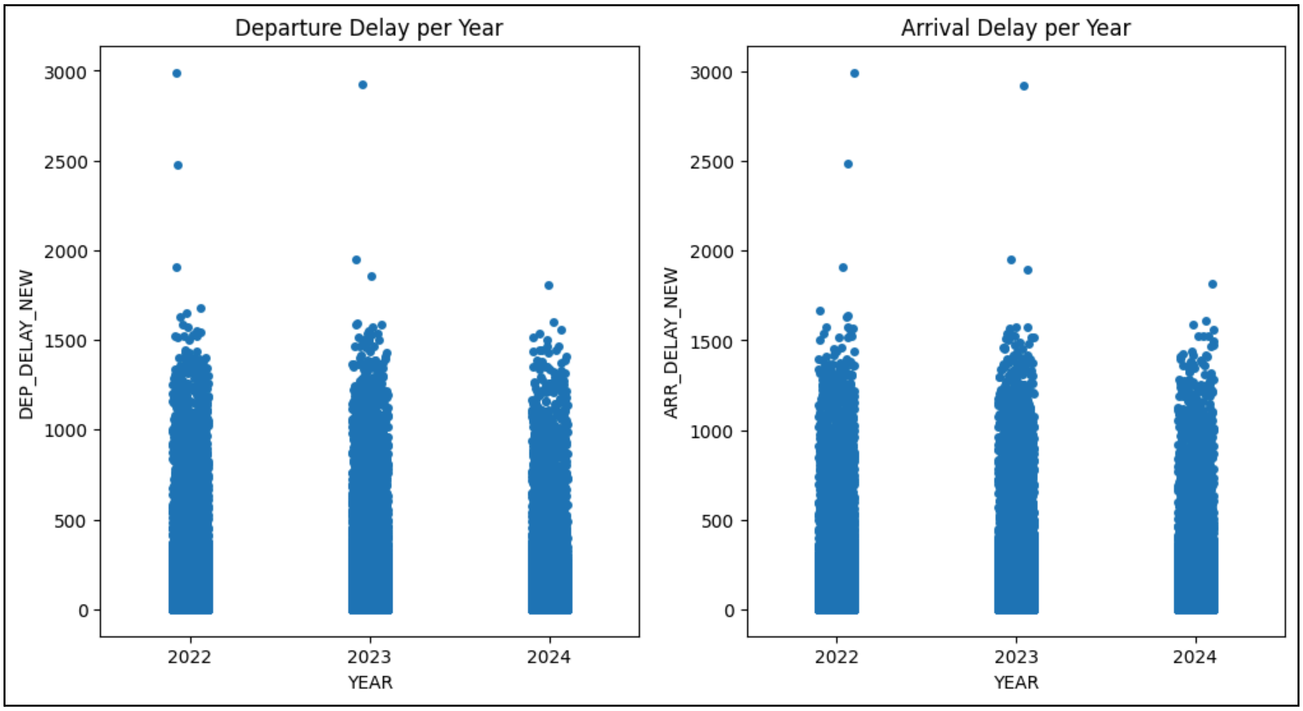


13. Strip Plot: Delays by Year

A strip plot was used to show the spread of departure and arrival delays by year. This type of plot is useful for visualizing the distribution of delays and any potential outliers across different years.

Key Insights:

- In both plots, the majority of flights have minimal delays, with most data points clustering at the lower end (closer to 0).
- There are several outliers each year, where flights experienced significant delays of over 1,500 to 3,000 minutes.
- The distribution of delays appears similar across all three years, with no drastic changes in the frequency or magnitude of delays from one year to the next.



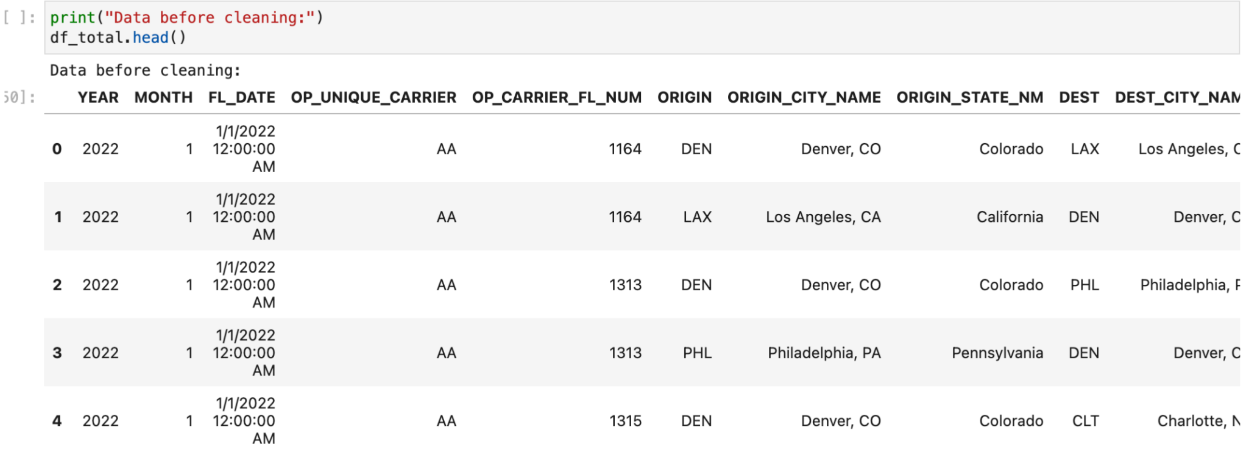
Dataset Comparison: Before and After Cleaning

This section highlights the differences between the dataset before and after the data cleaning and transformation process. It demonstrates how the data was cleaned, normalized, and prepared for analysis.

Before Cleaning

Prior to cleaning, the dataset exhibited several issues, such as missing values, unprocessed categorical features, and unscaled numerical data. Key observations include:

- The CANCELLATION_CODE column contained a significant number of missing values.
- Both numerical and categorical features were in their raw, unprocessed formats.
- Delay columns (DEP_DELAY_NEW, ARR_DELAY_NEW) had a wide range of values and were not normalized.
- Categorical features, such as OP_UNIQUE_CARRIER, were not encoded, and some columns could be renamed for clarity.



```
print("Data before cleaning:")
df_total.head()
```

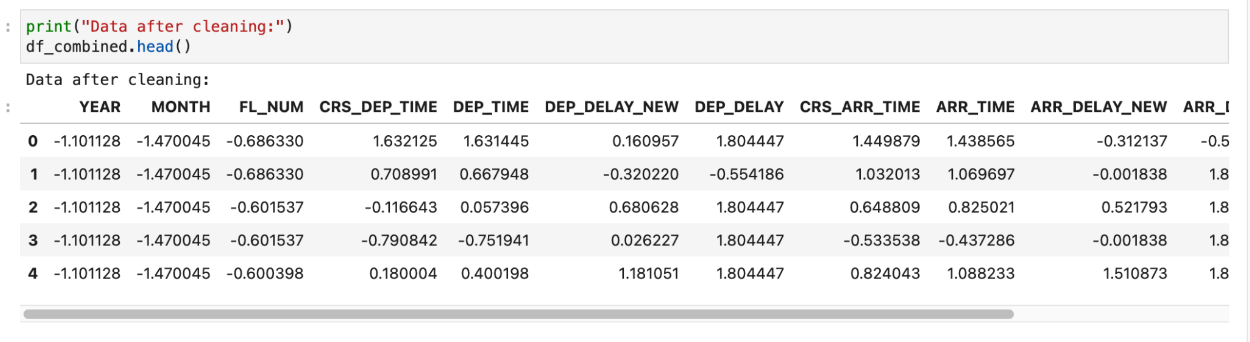
```
Data before cleaning:
```

	YEAR	MONTH	FL_DATE	OP_UNIQUE_CARRIER	OP_CARRIER_FL_NUM	ORIGIN	ORIGIN_CITY_NAME	ORIGIN_STATE_NM	DEST	DEST_CITY_NAM
0	2022	1	1/1/2022 12:00:00 AM	AA	1164	DEN	Denver, CO	Colorado	LAX	Los Angeles, C
1	2022	1	1/1/2022 12:00:00 AM	AA	1164	LAX	Los Angeles, CA	California	DEN	Denver, C
2	2022	1	1/1/2022 12:00:00 AM	AA	1313	DEN	Denver, CO	Colorado	PHL	Philadelphia, F
3	2022	1	1/1/2022 12:00:00 AM	AA	1313	PHL	Philadelphia, PA	Pennsylvania	DEN	Denver, C
4	2022	1	1/1/2022 12:00:00 AM	AA	1315	DEN	Denver, CO	Colorado	CLT	Charlotte, N

After Cleaning

Following the data cleaning and transformation process, the dataset was significantly improved and ready for analysis. Key changes include:

- The CANCELLATION_CODE column was removed due to excessive missing values.
- Numerical columns like DEP_DELAY and ARR_DELAY were standardized using the StandardScaler, bringing all values to a common scale.
- Categorical columns were one-hot encoded, converting them into a numerical format suitable for machine learning algorithms.
- Certain columns were renamed for clarity (e.g., OP_UNIQUE_CARRIER was renamed to CARRIER).



```
print("Data after cleaning:")
df_combined.head()
```

```
Data after cleaning:
```

	YEAR	MONTH	FL_NUM	CRS_DEP_TIME	DEP_TIME	DEP_DELAY_NEW	DEP_DELAY	CRS_ARR_TIME	ARR_TIME	ARR_DELAY_NEW	ARR_I
0	-1.101128	-1.470045	-0.686330	1.632125	1.631445	0.160957	1.804447	1.449879	1.438565	-0.312137	-0.5
1	-1.101128	-1.470045	-0.686330	0.708991	0.667948	-0.320220	-0.554186	1.032013	1.069697	-0.001838	1.8
2	-1.101128	-1.470045	-0.601537	-0.116643	0.057396	0.680628	1.804447	0.648809	0.825021	0.521793	1.8
3	-1.101128	-1.470045	-0.601537	-0.790842	-0.751941	0.026227	1.804447	-0.533538	-0.437286	-0.001838	1.8
4	-1.101128	-1.470045	-0.600398	0.180004	0.400198	1.181051	1.804447	0.824043	1.088233	1.510873	1.8

The cleaned dataset is now consistent, with standardized numerical values and properly encoded categorical features, making it ready for statistical analysis and machine learning tasks.

Importance of Data Cleaning

Cleaning and transforming the dataset is a critical step in preparing it for analysis. Raw data often contains missing values, inconsistencies, and mixed data types that can hinder the performance of machine learning models. By addressing these issues, we:

- Ensured consistency and completeness in the dataset.
- Made the data suitable for further analysis by standardizing numerical features and encoding categorical ones.
- Reduced complexity by eliminating unnecessary columns and addressing missing data.

These improvements enhance the dataset's quality and make it more reliable for generating accurate insights.

Project Status

At this stage of the project, significant progress has been made in cleaning and transforming the dataset. We have completed the following steps:

1. **Data Cleaning**: The dataset has been thoroughly cleaned by handling missing values, removing irrelevant columns such as CANCELLATION_CODE, and addressing inconsistencies in the data.

2. **Data Transformation**: We standardized the numerical features using the StandardScaler and applied one-hot encoding to categorical features, making the dataset suitable for machine learning models.
3. **Data Integration**: The monthly flight records for 2022, 2023, and 2024 have been combined into a unified dataset, allowing for comprehensive analysis across multiple years.
4. **Current Focus**: We are now preparing to perform exploratory data analysis (EDA) to uncover trends, correlations, and key insights from the cleaned and transformed dataset. This will be followed by model selection for predicting flight delays.

5. **Challenges**: So far, no major challenges have arisen during the data preparation process. All tasks have been completed according to schedule.