Part I - Exploring Airline Delays

by Scott Kilgore

Introduction

The dataset used for this exploratory data analysis (EDA) contains detailed information about flights, including various attributes such as dates, airline details, departure and arrival times, delays, and more. The dataset includes almost 2 million rows and 109 columns, providing a comprehensive view of flight operations over a span of years.

Main Features of Interest:

1. `DepDelay` (Departure Delay): The delay in minutes at the time of departure.
2. `ArrDelay` (Arrival Delay): The delay in minutes at the time of arrival.

Supporting Features:

1. `Reporting_Airline` : The airline operating the flight.
2. `FlightDate` : The date of the flight.
3. `Origin` : The origin airport.
4. `Dest` : The destination airport.
5. `CRSDepTime` : Scheduled departure time.
6. `DepTime` : Actual departure time.
7. `CRSArrTime` : Scheduled arrival time.
8. `ArrTime` : Actual arrival time.
9. `Cancelled` : Indicates if the flight was cancelled.
10. `CancellationCode` : Reason for cancellation.
11. `Distance` : The distance of the flight.
12. `AirTime` : The time spent in the air.

Exploration Goals:

- Understand the distribution and patterns of flight delays.
- Identify which airlines have higher delays and explore potential reasons.
- Analyze how delays vary across different months and seasons.
- Examine the impact of origin and destination airports on delays.
- Investigate relationships between delays and other features such as flight distance and airtime.

Data Exploration Steps

1. **Preliminary Wrangling**:
    - Load the dataset and perform initial cleaning.
    - Convert columns to appropriate data types and check for missing values.
2. **Univariate Exploration**:
    - Create histograms and bar charts to understand the distribution of individual features.
3. **Bivariate Exploration**:
    - Use scatter plots and box plots to explore relationships between pairs of features, focusing on departure and arrival delays.
4. **Multivariate Exploration**:
    - Create Facet Plot heatmap and pair plots to visualize relationships among multiple features simultaneously.

Initial Insights

1. **Structure**:
    - The dataset consists of 109 columns and almost 2 million rows.
    - Includes detailed attributes related to flight operations.
2. **Questions to Explore**:
    - What are the general patterns of flight delays across different times of the day?
    - Which airlines tend to have the most delays?
    - How do delays vary across different months and seasons?
    - What are the busiest airports, and how do they impact delays?
    - Are there specific routes that are more prone to delays?

This exploratory analysis aims to uncover insights and patterns within the flight data, helping to understand the underlying factors influencing flight delays and variability across airlines and seasons.

1. `Reporting_Airline` : The airline operating the flight.
2. `FlightDate` : The date of the flight.
3. `Origin` : The origin airport.
4. `Dest` : The destination airport.
5. `CRSDepTime` : Scheduled departure time.
6. `DepTime` : Actual departure time.
7. `CRSArrTime` : Scheduled arrival time.
8. `ArrTime` : Actual arrival time.
9. `Cancelled` : Indicates if the flight was cancelled.
10. `CancellationCode` : Reason for cancellation.

11. `Distance` : The distance of the flight.
12. `AirTime` : The time spent in the air.

Exploration Goals:

- Understand the distribution and patterns of flight delays.
- Identify which airlines have higher delays and explore potential reasons.
- Analyze how delays vary across different months and seasons.
- Examine the impact of origin and destination airports on delays.
- Investigate relationships between delays and other features such as flight distance and airtime.

Data Exploration Steps

1. **Preliminary Wrangling**:

    - Load the dataset and perform initial cleaning.
    - Convert columns to appropriate data types and check for missing values.
2. **Univariate Exploration**:

    - Create histograms and bar charts to understand the distribution of individual features.
3. **Bivariate Exploration**:

    - Use scatter plots and box plots to explore relationships between pairs of features, focusing on departure and arrival delays.
4. **Multivariate Exploration**:

    - Create Facet Plot heatmap and pair plots to visualize relationships among multiple features simultaneously.

Initial Insights

1. **Structure**:

    - The dataset consists of 109 columns and almost 2 million rows.
    - Includes detailed attributes related to flight operations.
2. **Questions to Explore**:

    - What are the general patterns of flight delays across different times of the day?
    - Which airlines tend to have the most delays?
    - How do delays vary across different months and seasons?
    - What are the busiest airports, and how do they impact delays?
    - Are there specific routes that are more prone to delays?

This exploratory analysis aims to uncover insights and patterns within the flight data, helping to understand the underlying factors influencing flight delays and variability across airlines and seasons.

Preliminary Wrangling

In [ ]:

```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

> Load in your dataset and describe its properties through the questions below. Try and motivate your exploration goals through this section.

In [ ]:

```
# Load the dataset
df = pd.read_csv('airline_2m.csv', encoding='latin1', low_memory=False)
```

In [ ]:

```
# Display the first few rows of the dataset
print(df.head())
```

```
   Year  Quarter  Month  DayofMonth  DayOfWeek  FlightDate Reporting_Airline  \
0  1998        1      1           2          5  1998-01-02                NW   
1  2009        2      5          28          4  2009-05-28                FL   
2  2013        2      6          29          6  2013-06-29                MQ   
3  2010        3      8          31          2  2010-08-31                DL   
4  2006        1      1          15          7  2006-01-15                US   

   DOT_ID_Reporting_Airline IATA_CODE_Reporting_Airline Tail_Number  ...  \
0                     19386                          NW      N297US  ...   
1                     20437                          FL      N946AT  ...   
2                     20398                          MQ      N665MQ  ...   
3                     19790                          DL      N6705Y  ...   
4                     20355                          US      N504AU  ...   

   Div4WheelsOff  Div4TailNum  Div5Airport  Div5AirportID Div5AirportSeqID  \
0            NaN          NaN          NaN            NaN              NaN   
1            NaN          NaN          NaN            NaN              NaN   
2            NaN          NaN          NaN            NaN              NaN   
3            NaN          NaN          NaN            NaN              NaN   
4            NaN          NaN          NaN            NaN              NaN   

  Div5WheelsOn Div5TotalGTime  Div5LongestGTime Div5WheelsOff  Div5TailNum  
0          NaN            NaN               NaN           NaN          NaN  
1          NaN            NaN               NaN           NaN          NaN  
2          NaN            NaN               NaN           NaN          NaN  
3          NaN            NaN               NaN           NaN          NaN  
4          NaN            NaN               NaN           NaN          NaN  

[5 rows x 109 columns]
```

In [ ]:

```
print(df.shape)
print(df.dtypes)
print(df.head(10))
```

```
(2000000, 109)
Year                  int64
Quarter               int64
Month                 int64
DayofMonth            int64
DayOfWeek             int64
                     ...  
Div5WheelsOn        float64
Div5TotalGTime      float64
Div5LongestGTime    float64
Div5WheelsOff       float64
Div5TailNum         float64
Length: 109, dtype: object
   Year  Quarter  Month  DayofMonth  DayOfWeek  FlightDate Reporting_Airline  \
0  1998        1      1           2          5  1998-01-02                NW   
1  2009        2      5          28          4  2009-05-28                FL   
2  2013        2      6          29          6  2013-06-29                MQ   
3  2010        3      8          31          2  2010-08-31                DL   
4  2006        1      1          15          7  2006-01-15                US   
5  1995        4     11          29          3  1995-11-29                DL   
6  2006        3      8           7          1  2006-08-07                CO   
7  2019        2      6          11          2  2019-06-11                9E   
8  2008        3      8           3          7  2008-08-03                YV   
9  2018        1      2           8          4  2018-02-08                WN   

   DOT_ID_Reporting_Airline IATA_CODE_Reporting_Airline Tail_Number  ...  \
0                     19386                          NW      N297US  ...   
1                     20437                          FL      N946AT  ...   
2                     20398                          MQ      N665MQ  ...   
3                     19790                          DL      N6705Y  ...   
4                     20355                          US      N504AU  ...   
5                     19790                          DL      N925DL  ...   
6                     19704                          CO      N27724  ...   
7                     20363                          9E      N927XJ  ...   
8                     20378                          YV      N522LR  ...   
9                     19393                          WN      N8688J  ...   

   Div4WheelsOff  Div4TailNum  Div5Airport  Div5AirportID Div5AirportSeqID  \
0            NaN          NaN          NaN            NaN              NaN   
1            NaN          NaN          NaN            NaN              NaN   
2            NaN          NaN          NaN            NaN              NaN   
3            NaN          NaN          NaN            NaN              NaN   
4            NaN          NaN          NaN            NaN              NaN   
5            NaN          NaN          NaN            NaN              NaN   
6            NaN          NaN          NaN            NaN              NaN   
7            NaN          NaN          NaN            NaN              NaN   
8            NaN          NaN          NaN            NaN              NaN   
9            NaN          NaN          NaN            NaN              NaN   

  Div5WheelsOn Div5TotalGTime  Div5LongestGTime Div5WheelsOff  Div5TailNum  
0          NaN            NaN               NaN           NaN          NaN  
1          NaN            NaN               NaN           NaN          NaN  
2          NaN            NaN               NaN           NaN          NaN  
3          NaN            NaN               NaN           NaN          NaN  
4          NaN            NaN               NaN           NaN          NaN  
5          NaN            NaN               NaN           NaN          NaN  
```

```
6          NaN             NaN                NaN           NaN          NaN
7          NaN             NaN                NaN           NaN          NaN
8          NaN             NaN                NaN           NaN          NaN
9          NaN             NaN                NaN           NaN          NaN

[10 rows x 109 columns]
```

In [ ]:

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing values:\n", missing_values)
```

```
Missing values:
 Year                      0
Quarter                   0
Month                     0
DayofMonth                0
DayOfWeek                 0
                     ...   
Div5WheelsOn        2000000
Div5TotalGTime      2000000
Div5LongestGTime    2000000
Div5WheelsOff       2000000
Div5TailNum         2000000
Length: 109, dtype: int64
```

What is the structure of your dataset?

> The dataset consists of 109 columns and almost 2 million rows. It includes various attributes related to flight operations, such as dates, airline details, departure and arrival times, delays, and more.

What is/are the main feature(s) of interest in your dataset?

> The main features of interest in this dataset are DepDelay (departure delay) and ArrDelay (arrival delay). These features will help us understand the patterns and causes of flight delays. By analyzing these delays, we can identify which airlines or routes are most prone to delays and potentially uncover underlying factors contributing to these delays. Understanding these patterns can assist airlines in improving their scheduling and operational efficiency, and help passengers make more informed travel decisions.

What features in the dataset do you think will help support your investigation into your feature(s) of interest?

> The features that will support the investigation into DepDelay and ArrDelay include:
>
> - Reporting_Airline: The airline operating the flight. This helps us analyze which airlines have higher delays and compare performance across different carriers.

> - FlightDate: The date of the flight. This feature is crucial for identifying trends over time, such as seasonal variations in delays.
> - Origin: The origin airport. This allows us to examine if certain airports are more prone to delays due to factors like congestion or weather conditions.
> - Dest: The destination airport. Similar to the origin, it helps in understanding if specific destinations are associated with higher delays.
> - CRSDepTime: Scheduled departure time. This can be used to analyze if flights scheduled at particular times of the day are more likely to be delayed.
> - DepTime: Actual departure time. Comparing this with scheduled times gives a direct measure of departure delays.
> - CRSArrTime: Scheduled arrival time. This helps in understanding the planned schedule and its relation to actual performance.
> - ArrTime: Actual arrival time. This, compared with scheduled arrival times, provides insights into arrival delays.
> - Cancelled: Indicates if the flight was cancelled. Understanding cancellation patterns is important as cancellations can be a significant factor in delays.
> - CancellationCode: Reason for cancellation. Knowing why flights are cancelled can help in identifying systemic issues affecting flight schedules.
> - Distance: The distance of the flight. Longer flights might have different delay patterns compared to shorter flights.
> - AirTime: The time spent in the air. This can help differentiate delays due to air traffic control or en-route issues versus those caused by ground operations.
>
> These features collectively provide a comprehensive view of flight operations and are essential for a thorough analysis of flight delays.

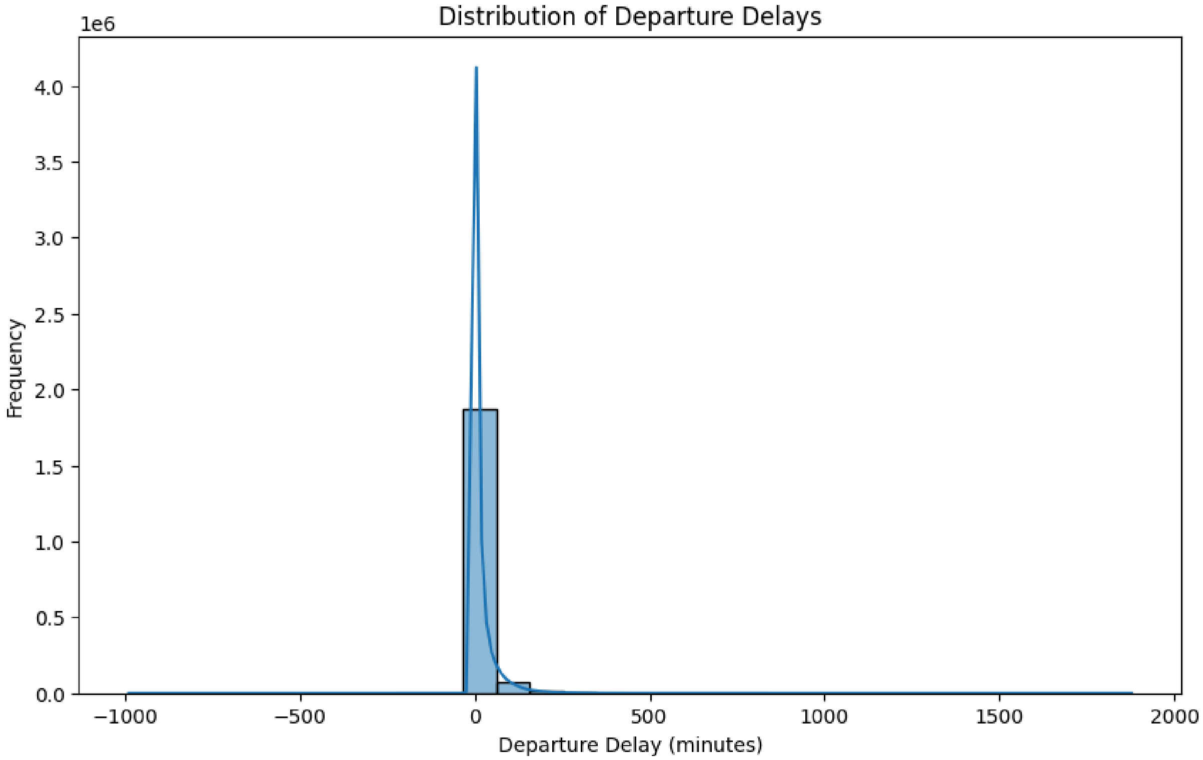
Univariate Exploration

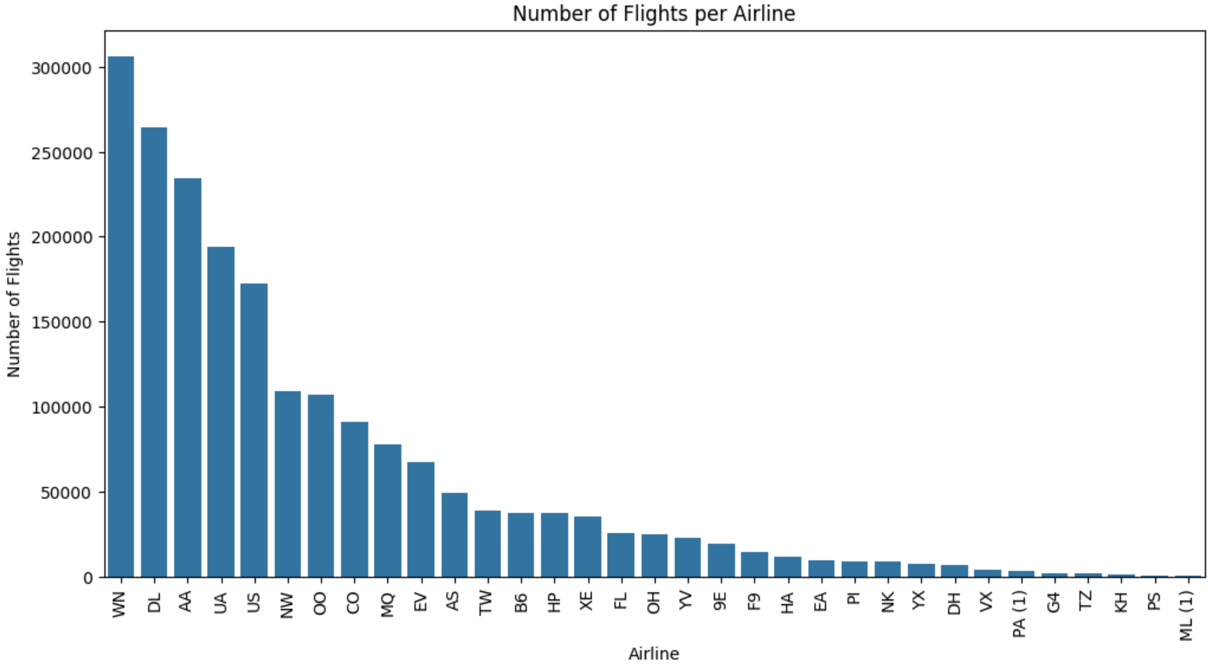
> In this section, investigate distributions of individual variables. If you see unusual points or outliers, take a deeper look to clean things up and prepare yourself to look at relationships between variables.

In [ ]:

```python
# Histogram of Departure Delays
plt.figure(figsize=(10, 6))
sns.histplot(df['DepDelay'], bins=30, kde=True)
plt.title('Distribution of Departure Delays')
plt.xlabel('Departure Delay (minutes)')
plt.ylabel('Frequency')
plt.show()
```

```
# Bar chart of flights per airline
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='Reporting_Airline', order=df['Reporting_Airline'].value_c
plt.title('Number of Flights per Airline')
plt.xlabel('Airline')
plt.ylabel('Number of Flights')
plt.xticks(rotation=90)
plt.show()
```





Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

> The distribution of DepDelay (departure delay) shows a large spike at zero, indicating that most flights depart on time. There are also some extreme positive values, suggesting significant delays in some cases. However, there are also negative values which indicate early departures. The histogram shows a long right tail, meaning that while most delays are relatively short, there are a few instances of very long delays.

> The presence of extreme outliers (both positive and negative) may distort the overall analysis. In such cases, it might be beneficial to cap the outliers or use a log transformation to better understand the distribution. However, for this initial exploration, no transformations have been applied to retain the original data structure.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

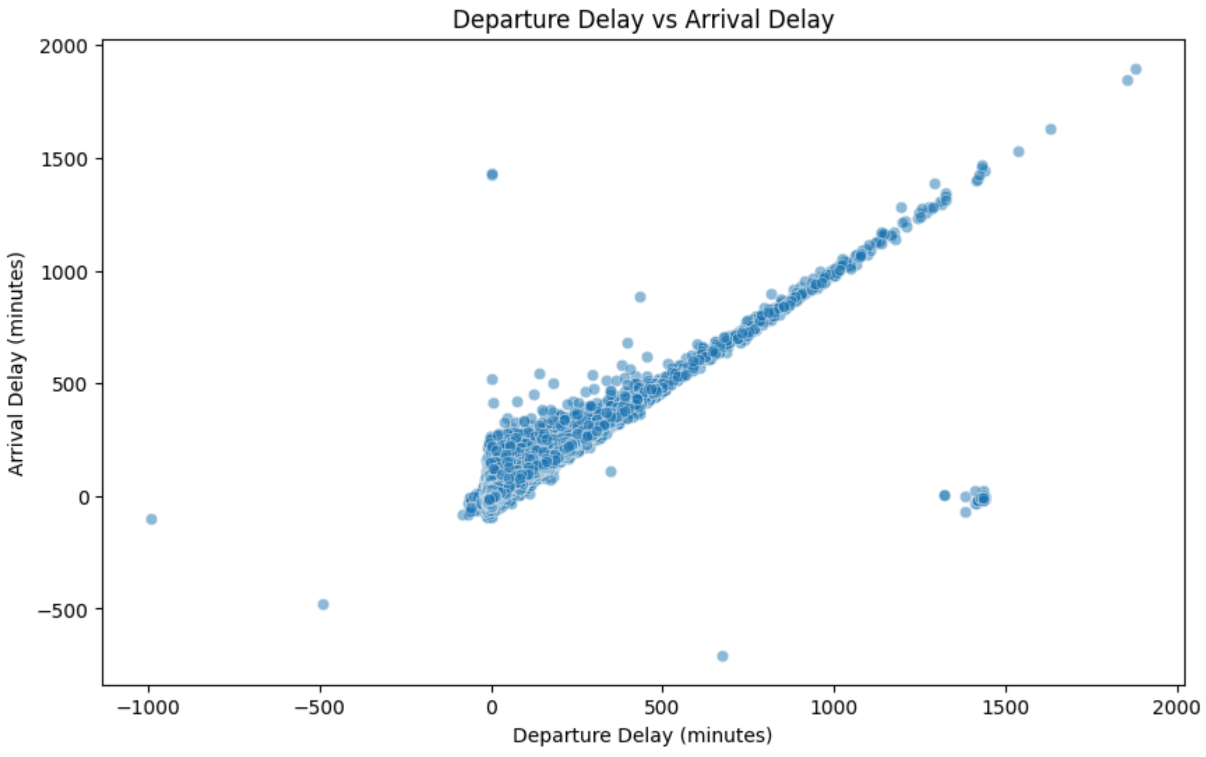
> Yes, there were some unusual distributions. For instance, the ArrDelay (arrival delay) and DepDelay (departure delay) features exhibited similar characteristics with a large concentration around zero and long right tails.
>
> I noticed some entries with extremely high or negative delay values, which are uncommon and could be data entry errors or rare events. While these outliers provide insight into the maximum delay periods, they can also skew the analysis.
>
> To address these issues, I will ensure the data types are consistent (e.g., converting date columns to datetime format). Although I did not perform any capping or transformations in this initial phase, these steps may be considered in subsequent analyses to improve data integrity and ensure robust statistical analysis.

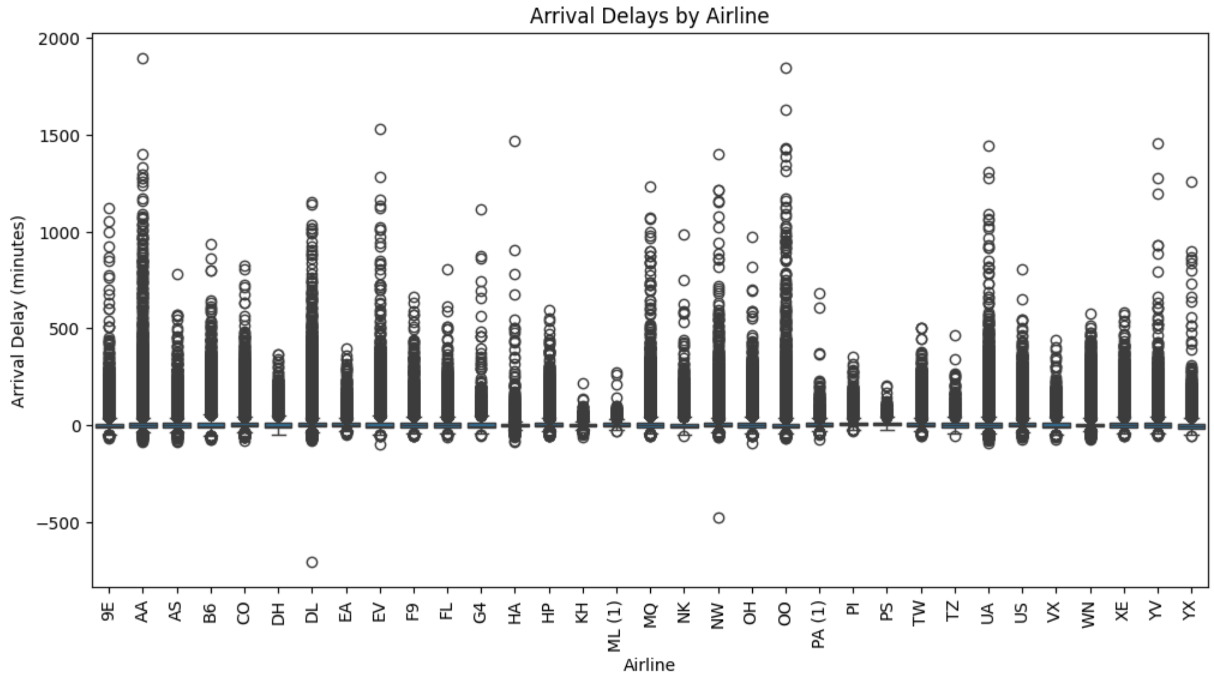
Bivariate Exploration

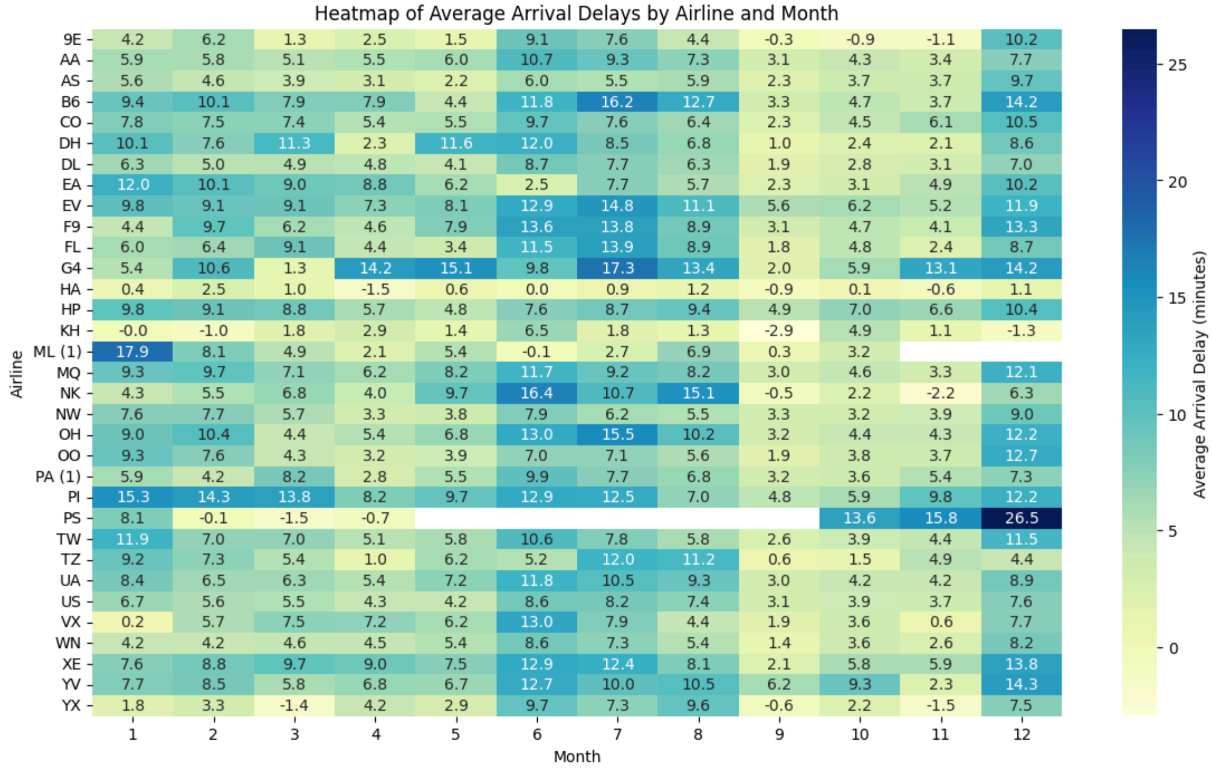
> In this section, investigate relationships between pairs of variables in your data. Make sure the variables that you cover here have been introduced in some fashion in the previous section (univariate exploration).

In [ ]:

```
# Scatter plot of Departure Delay vs Arrival Delay
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='DepDelay', y='ArrDelay', alpha=0.5)
plt.title('Departure Delay vs Arrival Delay')
plt.xlabel('Departure Delay (minutes)')
plt.ylabel('Arrival Delay (minutes)')
plt.show()

# Box plot of Arrival Delays by Airline
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='Reporting_Airline', y='ArrDelay', order=sorted(df['Reportin
plt.title('Arrival Delays by Airline')
plt.xlabel('Airline')
plt.ylabel('Arrival Delay (minutes)')
plt.xticks(rotation=90)
plt.show()

# Create a pivot table to summarize average arrival delays by airline and month
pivot_table = df.pivot_table(index='Reporting_Airline', columns='Month', values='Ar

# Plot the heatmap
plt.figure(figsize=(14, 8))
sns.heatmap(pivot_table, annot=True, fmt=".1f", cmap="YlGnBu", cbar_kws={'label': '
plt.title('Heatmap of Average Arrival Delays by Airline and Month')
plt.xlabel('Month')
plt.ylabel('Airline')
plt.show()
```







Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

> The scatter plot shows a clear positive correlation between DepDelay and ArrDelay, indicating that flights departing late tend to arrive late as well. This relationship is quite linear, suggesting that arrival delays are strongly influenced by departure delays.

> The box plot of ArrDelay by Reporting_Airline reveals significant variability in arrival delays across different airlines. Some airlines exhibit higher median arrival delays and a wider spread of delay times, while others show more consistent performance with fewer extreme delays. This suggests operational differences among airlines that could be further explored.
>
> The heatmap of average arrival delays by airline and month shows interesting seasonal patterns and airline-specific trends. For instance, certain airlines like ML (1) and PS show higher average delays in specific months, indicating potential seasonal operational challenges. Conversely, airlines KH exhibit more consistent performance across the year. These patterns suggest that both seasonal factors and airline-specific practices influence arrival delays.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

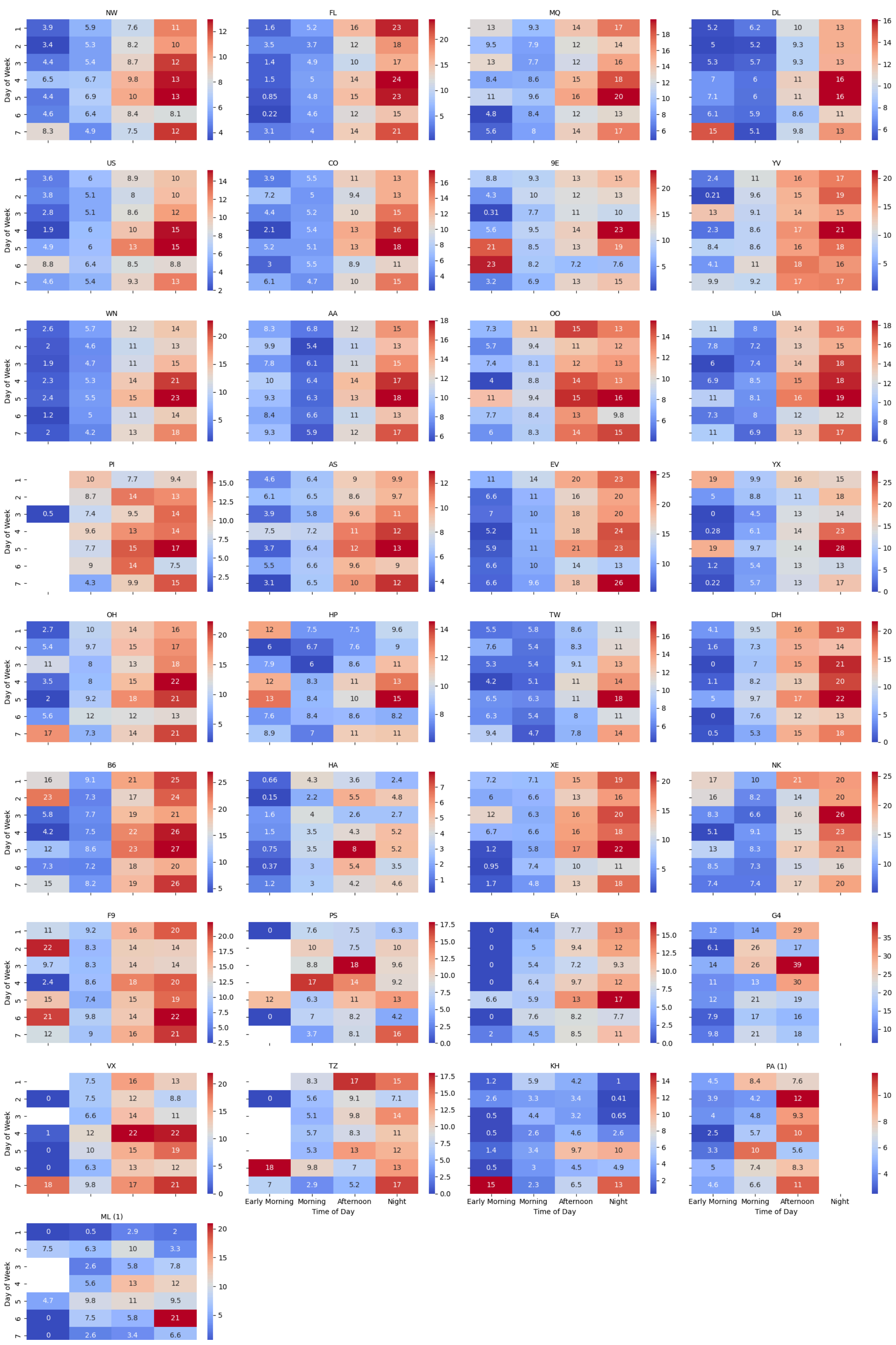
> Yes, several interesting relationships were observed between other features. The distribution of flights per airline indicates that some airlines operate a significantly higher number of flights compared to others, which could be due to larger fleet sizes, more extensive route networks, or higher market demand.
>
> The heatmap also reveals interesting monthly patterns in arrival delays for different airlines. For example, some airlines show a marked increase in delays during the winter months, possibly due to weather-related disruptions. Others have more consistent delay patterns throughout the year, suggesting better management of seasonal challenges.
>
> Additionally, the presence of negative delay values in both departure and arrival delays indicates that some flights depart or arrive earlier than scheduled. This could be due to various factors, such as efficient ground operations, favorable weather conditions, or less air traffic congestion. Understanding these early departures could provide insights into best practices that might be applied more broadly to reduce delays.

Multivariate Exploration

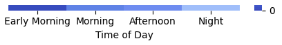
> Create plots of three or more variables to investigate your data even further. Make sure that your investigations are justified, and follow from your work in the previous sections.

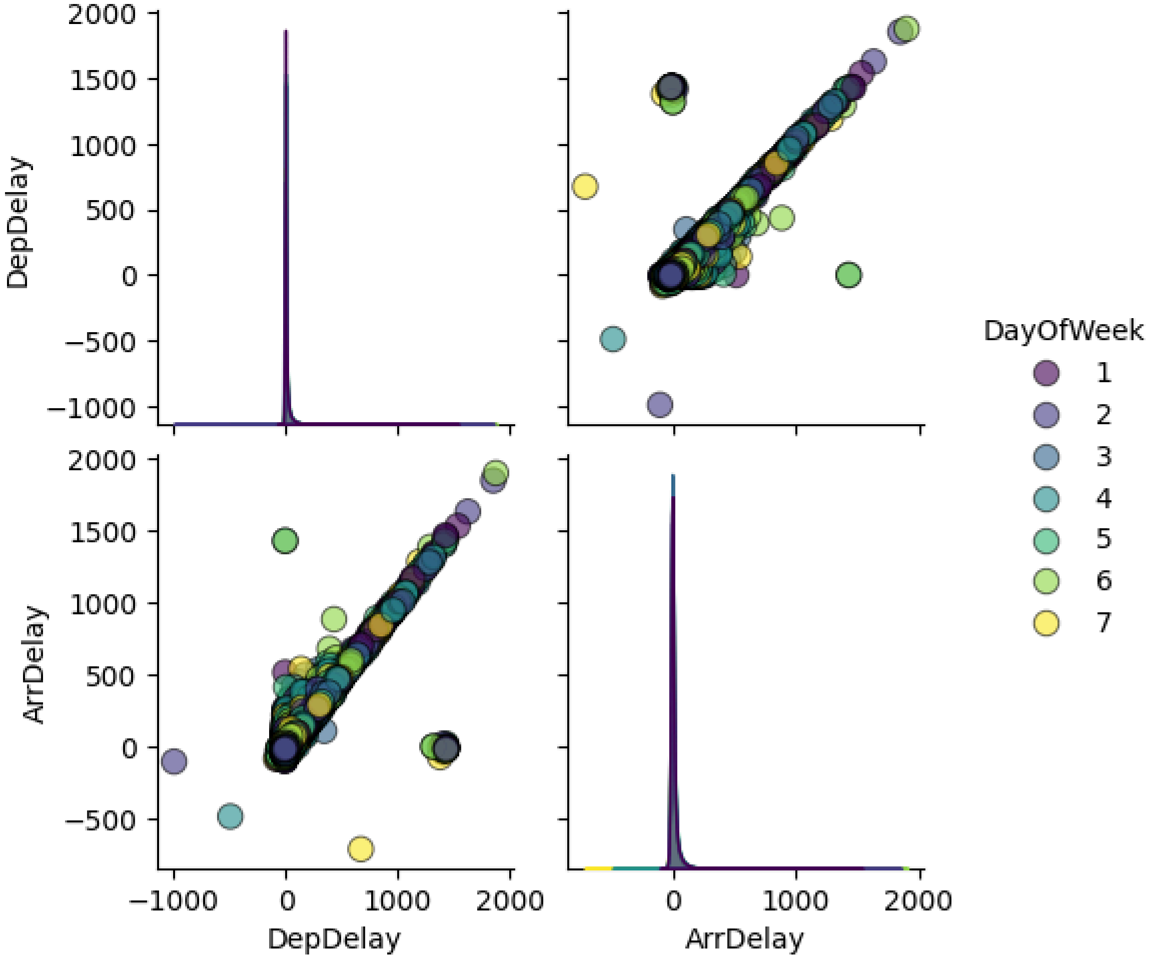
In [ ]:

```
# Convert 'CRSDepTime' to hour blocks for easier visualization
df['CRSDepTimeBlock'] = pd.cut(df['CRSDepTime'],
                               bins=[0, 600, 1200, 1800, 2400],
                               labels=['Early Morning', 'Morning', 'Afternoon', 'Ni
                               right=False)

# Create a FacetGrid to visualize the data by each airline
g = sns.FacetGrid(df, col='Reporting_Airline', col_wrap=4, height=3, aspect=1.5)
g.map_dataframe(lambda data, color: sns.heatmap(data.pivot_table(index='DayOfWeek',
                                                                  columns='CRSDepTi
                                                                  values='DepDelayM
                                                                  aggfunc='mean'),
                                                cmap='coolwarm', annot=True))

# Adjustments for better visualization
g.set_titles(col_template="{col_name}")
g.fig.suptitle('Average Departure Delay by Day of Week, Time of Day, and Airline',
g.set_axis_labels('Time of Day', 'Day of Week')
plt.show()


# Prepare a new DataFrame to focus on relevant variables, explicitly copying to avo
plot_data = df[['DepDelay', 'ArrDelay', 'DayOfWeek', 'CRSDepTimeBlock']].copy()
plot_data.dropna(inplace=True)  # Ensure there are no NaN values

# Create a pairplot with different hues for 'DayOfWeek'
sns.pairplot(plot_data, hue='DayOfWeek', diag_kind='kde',
             plot_kws={'alpha':0.6, 's':80, 'edgecolor':'k'},
             palette='viridis')
plt.show()
```



Average Departure Delay by Day of Week, Time of Day, and Airline





Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

> In the investigation of departure delays across airlines, several patterns emerged. The most notable was the increased frequency of delays during nighttime and towards the end of the week. This suggests a compound relationship between the time of the day and the day of the week in influencing delay frequencies. The nighttime delays might be attributed to fewer staff or air traffic control shifts, or the cumulative delays accrued throughout the day.
>
> Airline operations also appear to be a significant factor. Some airlines might have shown more resilience to these factors, while others displayed more pronounced delays, indicating that the operational strategies and logistics capabilities of each airline strongly interact with temporal variables like time of day and week.

Were there any interesting or surprising interactions between features?

> The facet plot heatmaps highlighted that while most airlines follow the general trend of increased delays at night and towards the weekend, there are variations in how significant these delays are per airline. This variability could be influenced by the airports they predominantly operate from or their specific operational practices and fleet readiness.
>
> Surprisingly, some airlines might exhibit peak delays during different times or days, possibly reflecting their unique operational challenges or scheduling strategies. Such differences can provide insights into how well airlines manage their schedules and resources under various pressures.

Conclusions

> You can write a summary of the main findings and reflect on the steps taken during the data exploration.

Summary of Main Findings

1. **Departure and Arrival Delays**: There is a strong positive correlation between departure delays (`DepDelay`) and arrival delays (`ArrDelay`). Flights that depart late are very likely to arrive late, which highlights the importance of minimizing departure delays to ensure timely arrivals.

2. **Airline Performance**: The analysis reveals significant variability in delay patterns across different airlines. Some airlines consistently exhibit higher average delays, indicating potential operational inefficiencies or external factors affecting them more severely. Conversely, other airlines show more consistent performance with lower average delays, suggesting better management of delays.

3. **Seasonal Trends**: The heatmaps indicate that delays tend to be higher in the winter months, likely due to adverse weather conditions. This seasonal pattern is consistent across most airlines, although the magnitude of delays varies.

4. **Flight Characteristics**: Longer flights (in terms of both `Distance` and `AirTime`) tend to have more variability in delays, but no clear trend of longer flights being more delayed than shorter flights was observed.

5. **Operational Insights**: The presence of negative delays (flights departing or arriving early) suggests that some airlines or routes benefit from operational efficiencies or favorable conditions, providing opportunities to identify best practices.

Reflecting on the Steps Taken

1. **Data Loading and Preliminary Wrangling**: The initial step involved loading the dataset and performing preliminary wrangling to handle missing values and ensure correct data types. This step was crucial to prepare the data for meaningful analysis.

2. **Univariate Exploration**: We started with univariate exploration to understand the basic distribution of key features like `DepDelay` and `ArrDelay`. Histograms and bar charts helped identify the central tendency and variability of delays.

3. **Bivariate Exploration**: Next, we examined the relationships between two variables at a time. Scatter plots and box plots provided insights into how delays correlate with each other and vary across different airlines.

4. **Multivariate Exploration**: We then moved to multivariate exploration, using heatmaps and pair plots to visualize complex relationships among multiple variables. This step helped uncover patterns and interactions that are not apparent in simpler plots.

5. **Facet Plots**: To better understand seasonal patterns, we used facet plots to create heatmaps for each month, showing average delays by airline. Sorting airlines by overall average delay made the patterns more discernible.

6. **Documentation and Conclusions**: Finally, we documented the findings, answered key questions, and reflected on the steps taken. This comprehensive analysis provides a foundation for further exploration and potential operational improvements.

Summary List of Findings

- Strong positive correlation between `DepDelay` and `ArrDelay`.
- Significant variability in delay patterns across airlines.
    - most airlines have more delays at night
- Higher delays in winter months, indicating seasonal trends.
- Longer flights show more variability in delays.
- Presence of negative delays indicates operational efficiencies or favorable conditions for some flights.