Analyzing Airline Flight Performance: Cancellations & Delays

Investigation Overview

This investigation aims to explore various aspects of airline flight data to uncover insights into flight performance, delays, and cancellations. The analysis will be guided by the following research questions:

1. What percentage of flights are canceled or diverted out of the total number of flights?
2. Is there a relationship between the number of cancellations and the time of year (quarters) or days of the week?
3. How do the causes of cancellations vary by quarter and day of the week?
4. What is the distribution pattern of flight delays?

Dataset Overview

The airline dataset provides a detailed record of flight information, compiled from multiple flights across different years. It includes a wide range of features related to various aspects of flight operations:

1. General Flight Information: This includes temporal details such as Year, Quarter, Month, Day of the Month, and Day of the Week.
2. Origin and Destination Information : Origin, OriginCityName, OriginState, OriginStateFips, OriginStateName, OriginWac: Various details about the origin airport, including codes, city name, state name, and geographic information.
3. Dest, DestCityName, DestState, DestStateFips, DestStateName, DestWac: Details about the destination airport, including codes, city name, state name, and geographic information.
4. Departure Information: CRSDepTime, DepTime, DepDelay, DepDelayMinutes, DepDel15, DepartureDelayGroups: Scheduled and actual departure times, along with various delay metrics, providing insights into how on-time or delayed departures were.
5. Arrival Information: CRSArrTime, ArrTime, ArrDelay, ArrDelayMinutes, ArrDel15, ArrivalDelayGroups: Scheduled and actual arrival times, along with various delay metrics, providing insights into arrival performance.
6. Cancelled, CancellationCode, Diverted: Indicators of whether a flight was canceled or diverted, along with codes explaining the reason for cancellation.
7. Flights, Distance, DistanceGroup: Metrics related to the flight count and the distance covered, with DistanceGroup likely categorizing flights into distance ranges.
8. Delay Breakdown: CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay: These columns break down the reasons for delays into categories like carrier issues, weather, air traffic control (NAS), security, and delays due to late aircraft.

```
# 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 sb
import os

%matplotlib inline

# load in the dataset into a pandas dataframe
print(os.getcwd())
os.chdir('C:/Users/User/Udacity/data_analysis/finalProject')
df = pd.read_csv('./airline_2m/airline_2m.csv',encoding='ISO-8859-1')
```

```
C:\Users\User

C:\Users\User\AppData\Local\Temp\ipykernel_3268\1997674437.py:4:
DtypeWarning: Columns (69,76,77,84) have mixed types. Specify dtype
option on import or set low_memory=False.
  df = pd.read_csv('./airline_2m/airline_2m.csv',encoding='ISO-8859-
1')
```

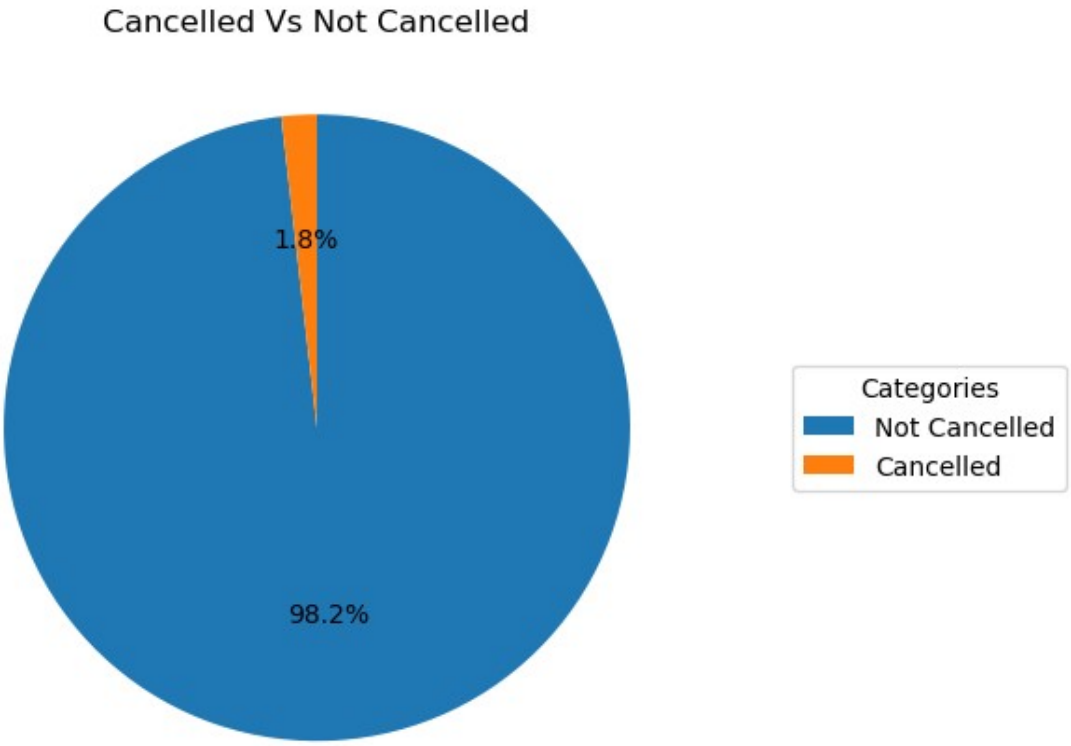
```
# convert cut, color, and clarity into ordered categorical types
ordinal_var_dict = {'cut': ['Fair','Good','Very
Good','Premium','Ideal'],
                    'color': ['J', 'I', 'H', 'G', 'F', 'E', 'D'],
                    'clarity': ['I1', 'SI2', 'SI1', 'VS2', 'VS1',
'VVS2', 'VVS1', 'IF']}

for var in ordinal_var_dict:
    ordered_var = pd.api.types.CategoricalDtype(ordered = True,
                                                categories =
ordinal_var_dict[var])
    diamonds[var] = diamonds[var].astype(ordered_var)

# data wrangling, removing diamonds with inconsistent or missing data.
incorrect_depth = (np.abs(2 * diamonds['z'] / (diamonds['x'] +
diamonds['y']) - diamonds['depth']/100) > 0.1)
no_size_info = ((diamonds['x'] == 0) & (diamonds['y'] == 0))
diamonds = diamonds.loc[-incorrect_depth & -no_size_info,:]
```

Cancellation Trends

1.8% of total trips were canceled. Analyzing cancellations by day of the week reveals that Fridays have fewer cancellations compared to other days, with the highest number of cancellations occurring on Tuesdays. When examining cancellations by quarter, it is evident that the number of cancellations is significantly higher in Q1 compared to other quarters

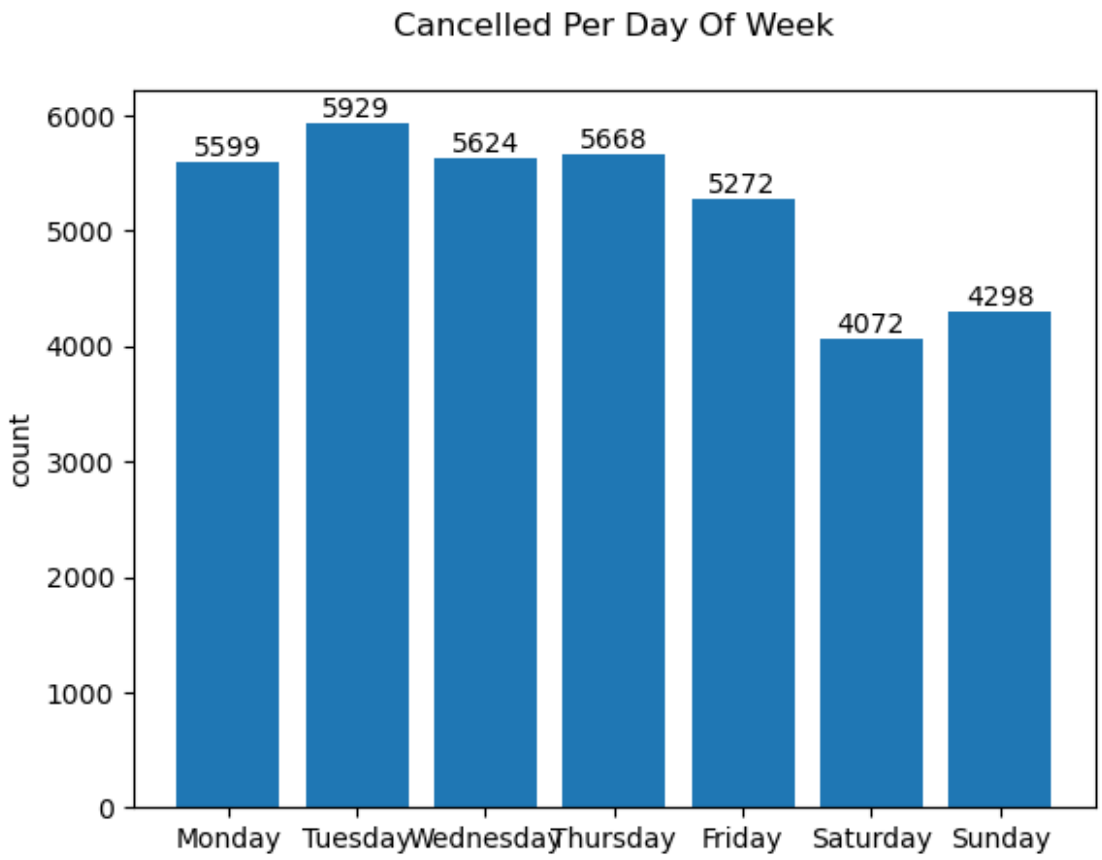
Cancelled vs Not Cancelled

```
pie_chart(cleaned_airline_df, 'Cancelled','Cancelled Vs Not
Cancelled', ['Not Cancelled', 'Cancelled'])
```



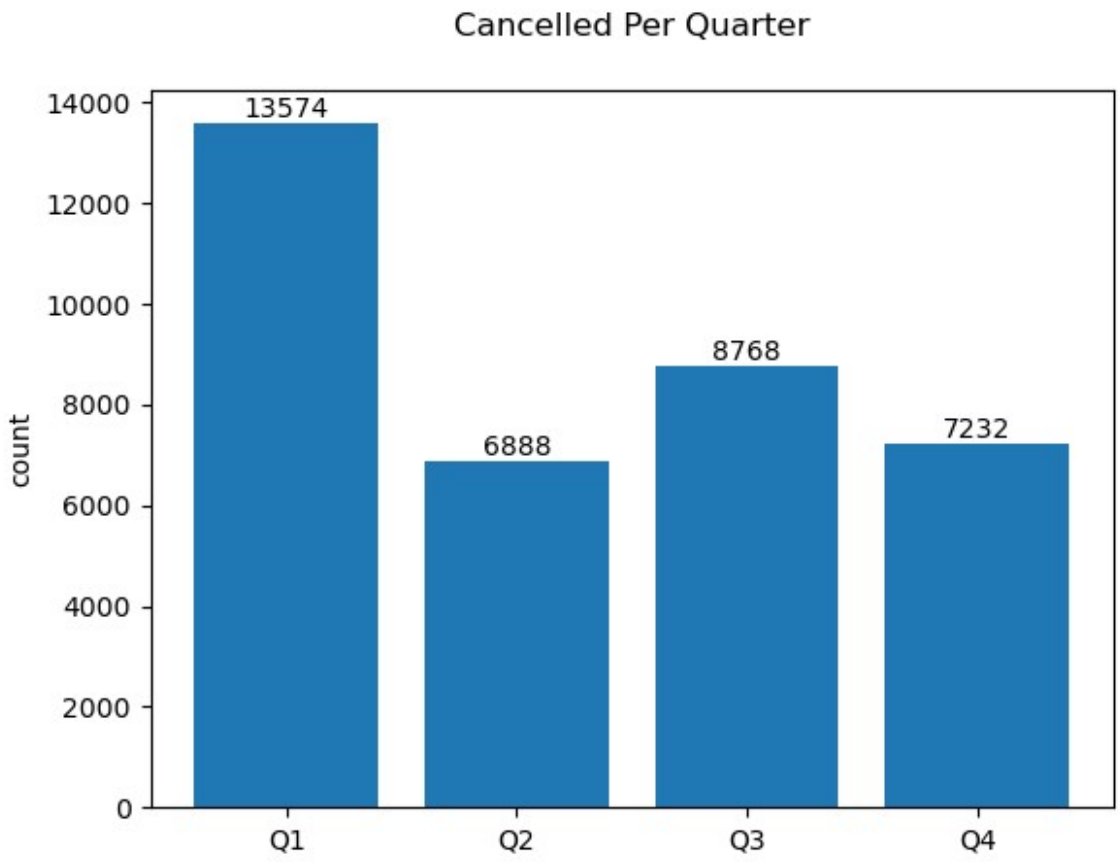
Cancellation Per week Day

```
bar_chart(canceled_airline_df.sort_values(['DayOfWeek']),
        'DayOfWeek_Desc',
        'Cancelled Per Day Of Week',)
```



Cancellation Per Quarter

```
bar_chart(canceled_airline_df.sort_values(['Quarter']),
        'Quarter_Desc',
        'Cancelled Per Quarter',)
```

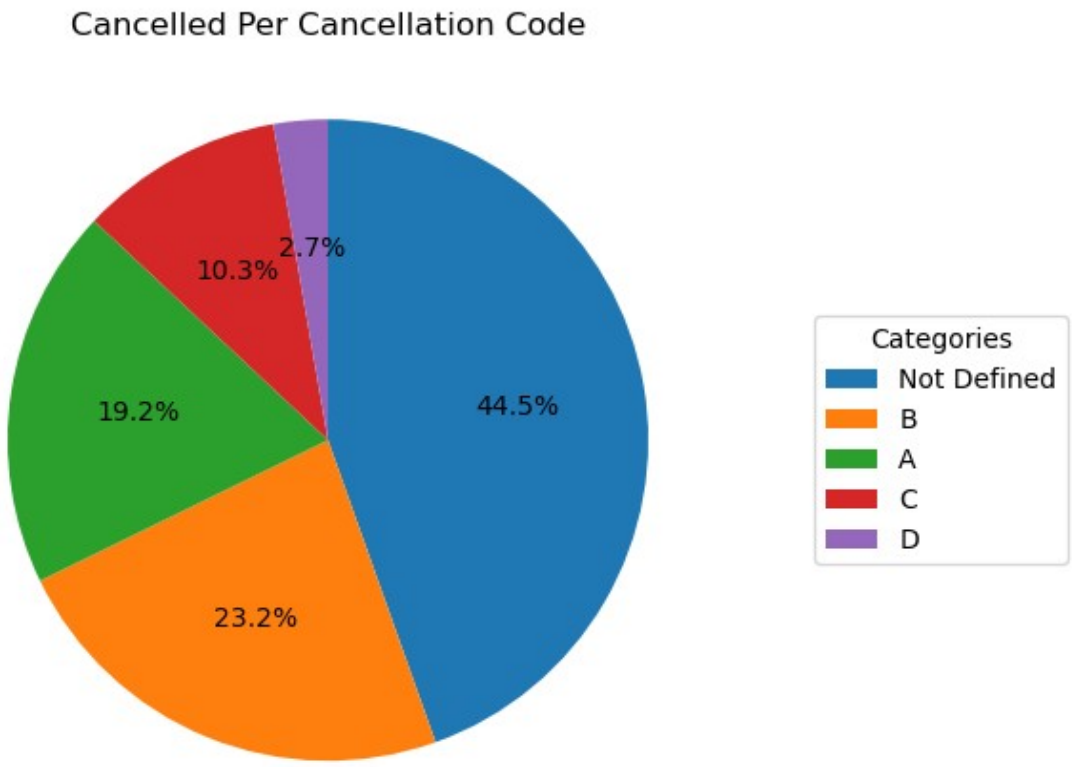


Cancellation based On Cancellation Code

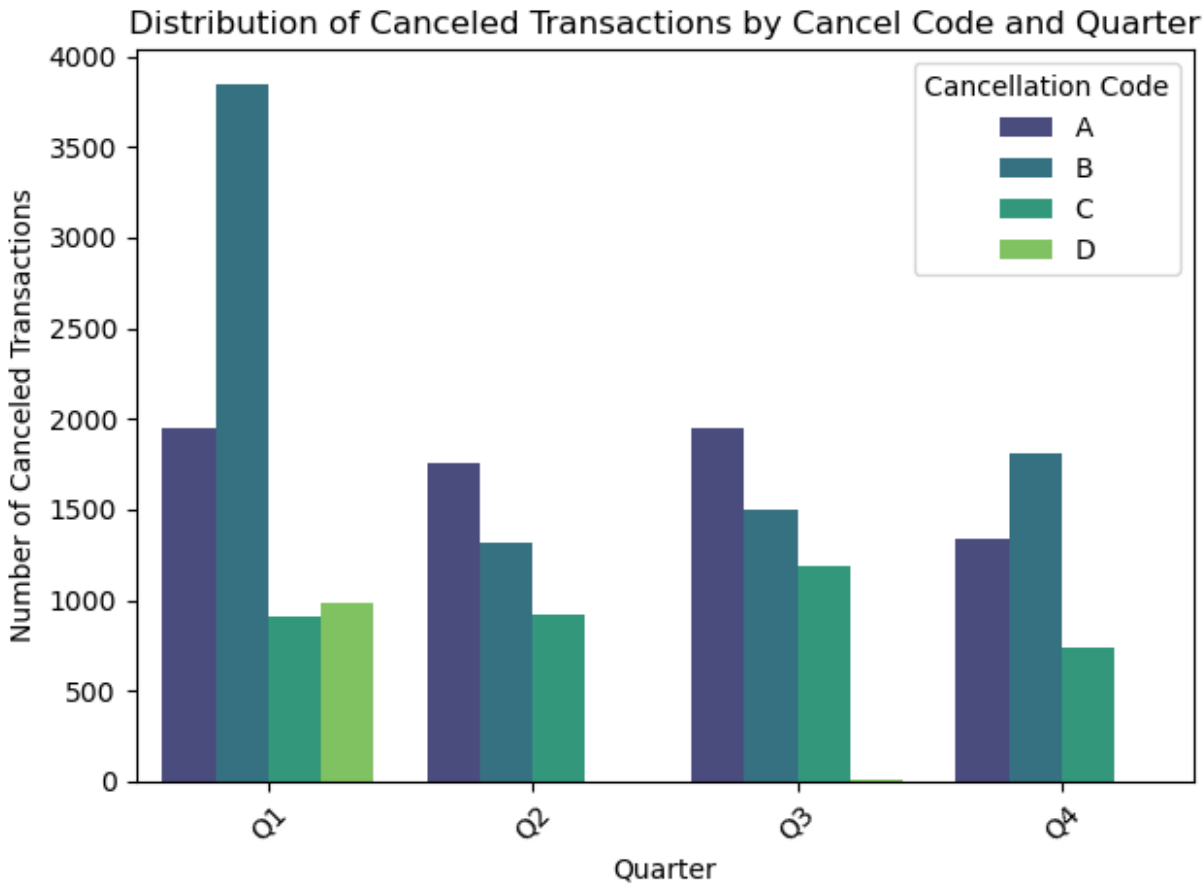
Examining the causes of cancellations reveals that 44% of the cancellations are attributed to unspecified reasons. Causes A and B account for approximately 40% of the cancellations.

Notably, in Quarter 1, cause B is the predominant reason for cancellations, with a markedly higher count than in other quarters. Cause D is exclusively observed in Quarter 1. In contrast, cause A is the primary reason for cancellations in both Quarter 2 and Quarter 3.

```
pie_chart(canceled_airline_df.sort_values(['CancellationCode']),
        'CancellationCode',
        'Cancelled Per Cancellation Code')
```




```
# Plotting
sns.barplot(data=canceled_airline_df_agg, x='Quarter_Desc', y='Count',
hue='CancellationCode', palette='viridis')

# Title and labels
plt.title('Distribution of Canceled Transactions by Cancel Code and
Quarter')
plt.xlabel('Quarter')
plt.ylabel('Number of Canceled Transactions')


plt.legend(title='Cancellation Code')
plt.xticks(rotation=45)
plt.tight_layout()
```



Finding:

1. 1.8% of all flights were canceled.
2. Cancellations were most frequent on Tuesdays and least common on Fridays.
3. The first quarter of the year saw the highest number of cancellations, significantly more than other quarters.
4. 44% of cancellations were due to unspecified reasons.
5. Causes A and B together accounted for approximately 40% of the total cancellations.
6. In Quarter 1, cause B was the leading reason for cancellations, with a significantly higher count than in other quarters.
7. Cause A was the primary reason for cancellations in Quarters 2 and 3.

Distribution of Delay

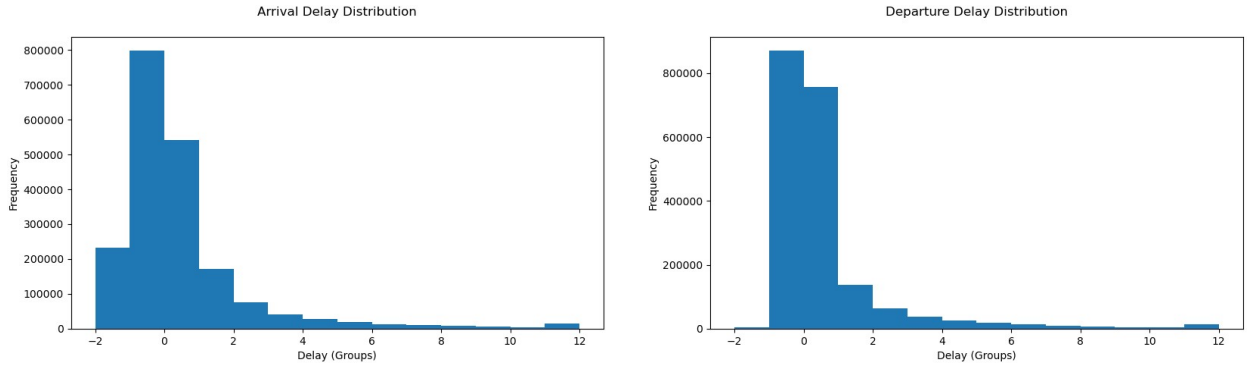
Diamond prices in the dataset take on a very large range of values, from about $300 at the lowest, to about $19,000 at the highest. Plotted on a logarithmic scale, the distribution of diamond prices takes on a multimodal shape.

Distribution Of Arrival Delay Groups and DepartureDelayGroups

The histogram below illustrates the distribution of arrival and departure delays. It shows that most data points are concentrated at the lower end of the range, with only a small number

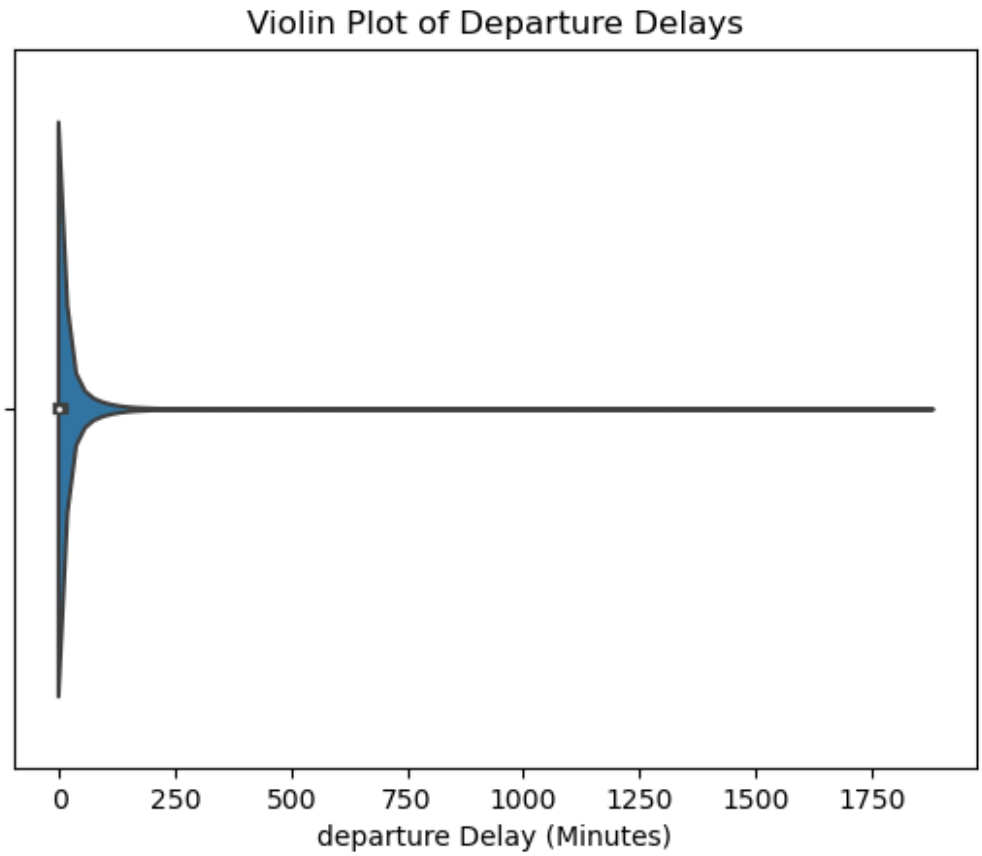
extending into the higher end. This distribution is highly skewed to the right, indicating that while most flights experience minimal or no delays, there are a few significant delays that create a long tail on the right side of the chart

```
two_hist_chart(cleaned_airline_df,
               ['ArrivalDelayGroups', 'DepartureDelayGroups'],
               ['Arrival Delay Distribution', 'Departure Delay
Distribution'],
                ['Delay (Groups)', 'Delay (Groups)'])
```

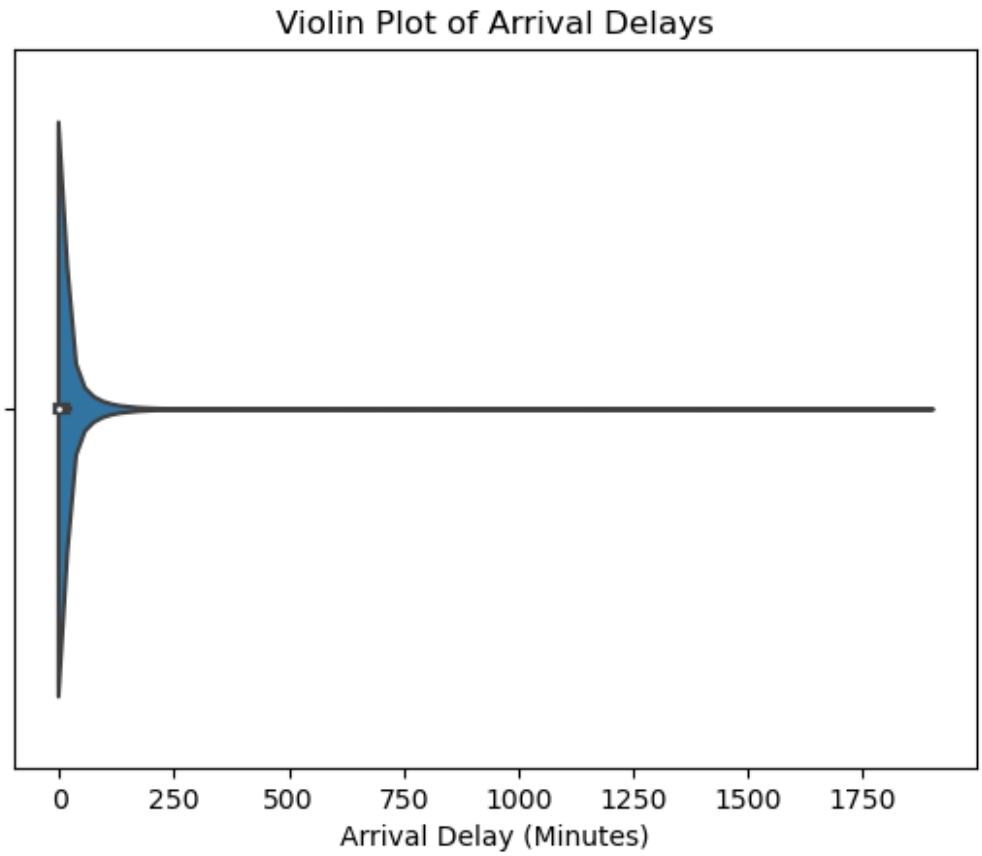


The violin plots for arrival and departure delays clearly demonstrate a highly skewed distribution, with most data concentrated around lower values. Delays up to 30 minutes fall predominantly within the first two delay groups, with a few extreme outliers. The overall delay data is categorized into 12 groups, with approximately 70% of the occurrences concentrated in the first three groups, as depicted in the pie chart below

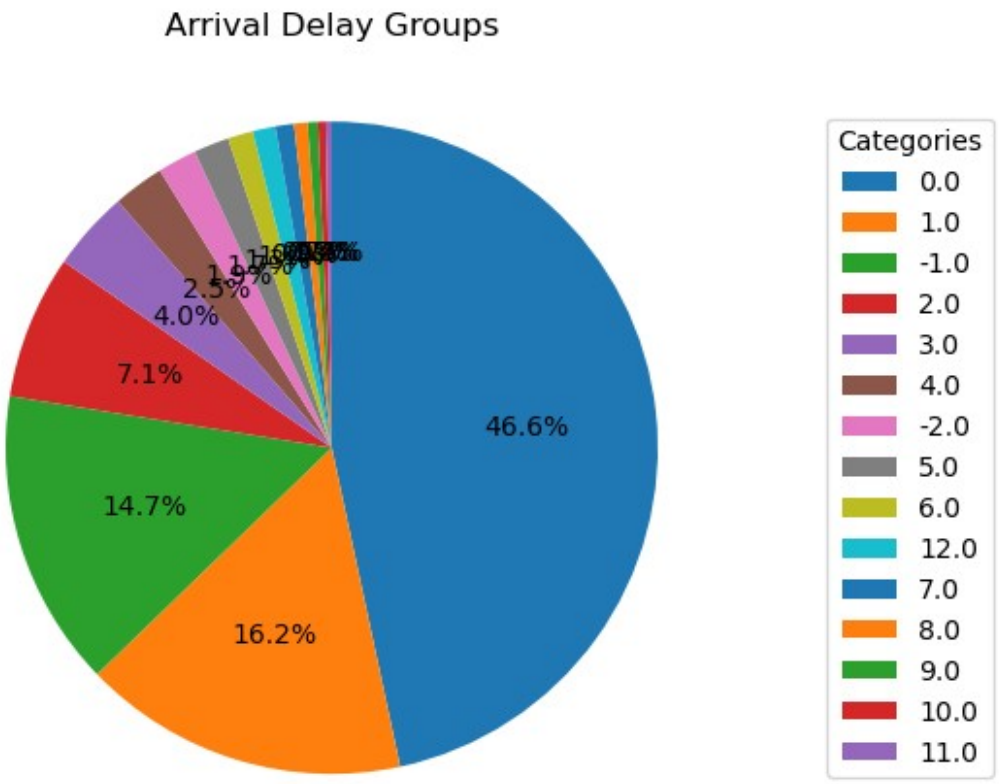
```
sns.violinplot(x=df['DepDelayMinutes'].dropna())
plt.xlabel('departure Delay (Minutes)')
plt.title('Violin Plot of Departure Delays')
plt.show()
```




```
sns.violinplot(x=df['ArrDelayMinutes'].dropna())
plt.xlabel('Arrival Delay (Minutes)')
plt.title('Violin Plot of Arrival Delays')
plt.show()
```

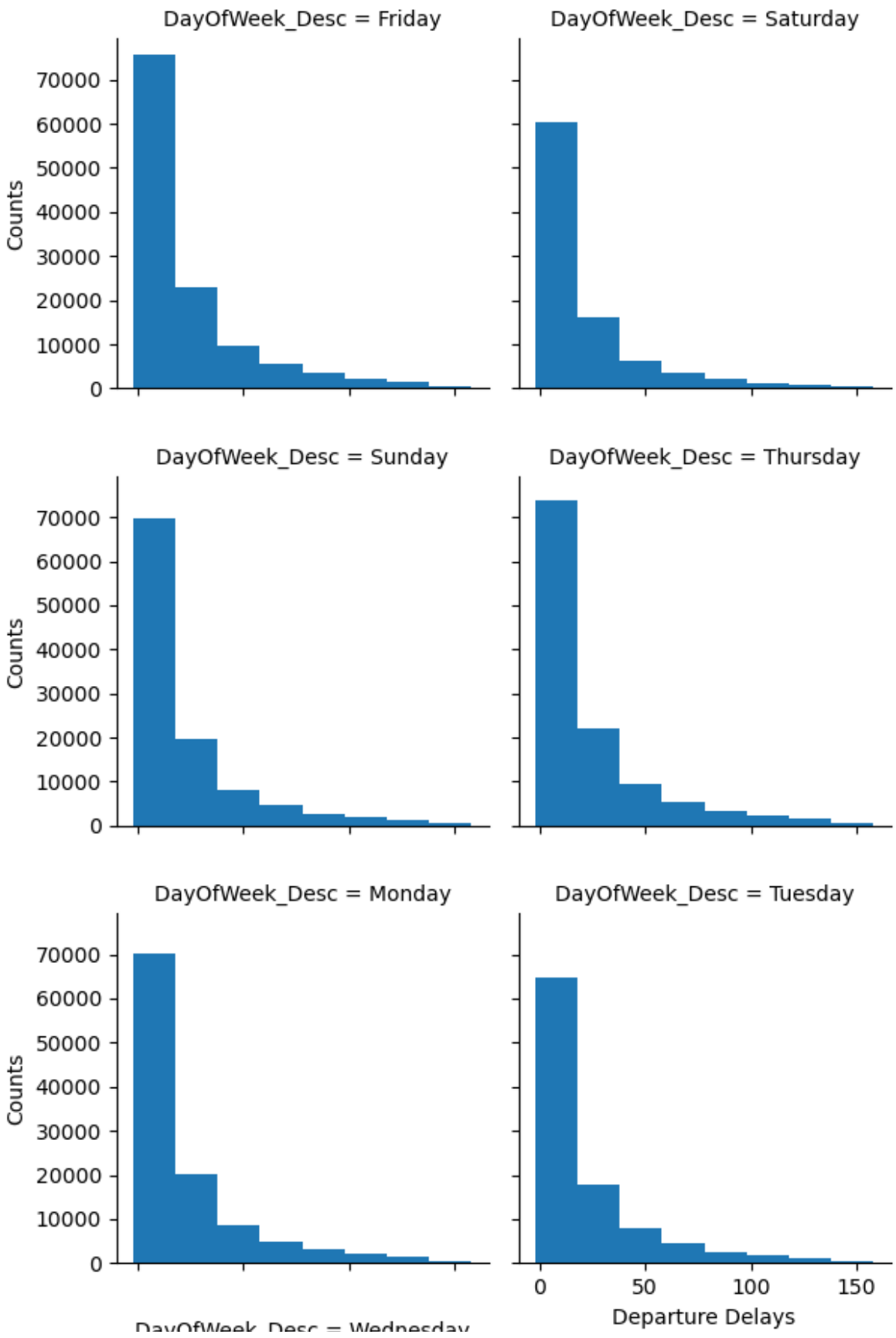



```
pie_chart(delay_airline_df, 'ArrivalDelayGroups','Arrival Delay
Groups')
```



Departure Delay Distribution by Day Of Week

```
FacetGrid(delay_airline_df_copy[(delay_airline_df_copy['DepDelayMinute
s']>0) & (delay_airline_df_copy['DepDelayMinutes']<150)],
          value_column='DepDelayMinutes',
          class_column='DayOfWeek_Desc',
          bin_size=20,
          title='Flights Delays',
          xyLabels=['Departure Delays', 'Counts'])
```


DayOfWeek_Desc = Friday
DayOfWeek_Desc = Saturday
DayOfWeek_Desc = Sunday
DayOfWeek_Desc = Thursday
DayOfWeek_Desc = Monday
DayOfWeek_Desc = Tuesday
DayOfWeek_Desc = Wednesday
Counts
0
10000
20000
30000
40000
50000
60000
70000
0
50
100
150
Departure Delays

```
cleaned_airline_df['ArrDelayMinutes'].describe()
```

```
count    1.958922e+06
mean     1.179442e+01
std      3.197121e+01
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      1.000000e+01
max      1.898000e+03
Name: ArrDelayMinutes, dtype: float64
```

From the above:

1. The average arrival delay in minutes across all flights is ~ 11 min
2. The minimum delay in minutes is ~ 0 min
3. 25% of the flights had no arrival delay - 25th Percentile (25%)
4. 50% of the flights had no arrival delay - 50th Percentile (50%)
5. 75% of the flights had no arrival delay - 70th Percentile (70%)
6. The maximum delay in minutes is ~ 1898 minutes (about 31.6 hours)

```
cleaned_airline_df['DepDelayMinutes'].describe()
```

```
count    1.963932e+06
mean     1.049667e+01
std      3.196467e+01
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      7.000000e+00
max      1.878000e+03
Name: DepDelayMinutes, dtype: float64
```

From the above:

1. The average departure delay in minutes across all flights is ~ 10 min
2. The minimum delay in minutes is ~ 0 min
3. 25% of the flights had no departure delay - 25th Percentile (25%)
4. 50% of the flights had no departure delay - 50th Percentile (50%)
5. 75% of the flights had 7 minutes departure delay - 70th Percentile (70%)
6. The maximum delay in minutes is ~ 1878 minutes (about 31.3 hours)

Finding

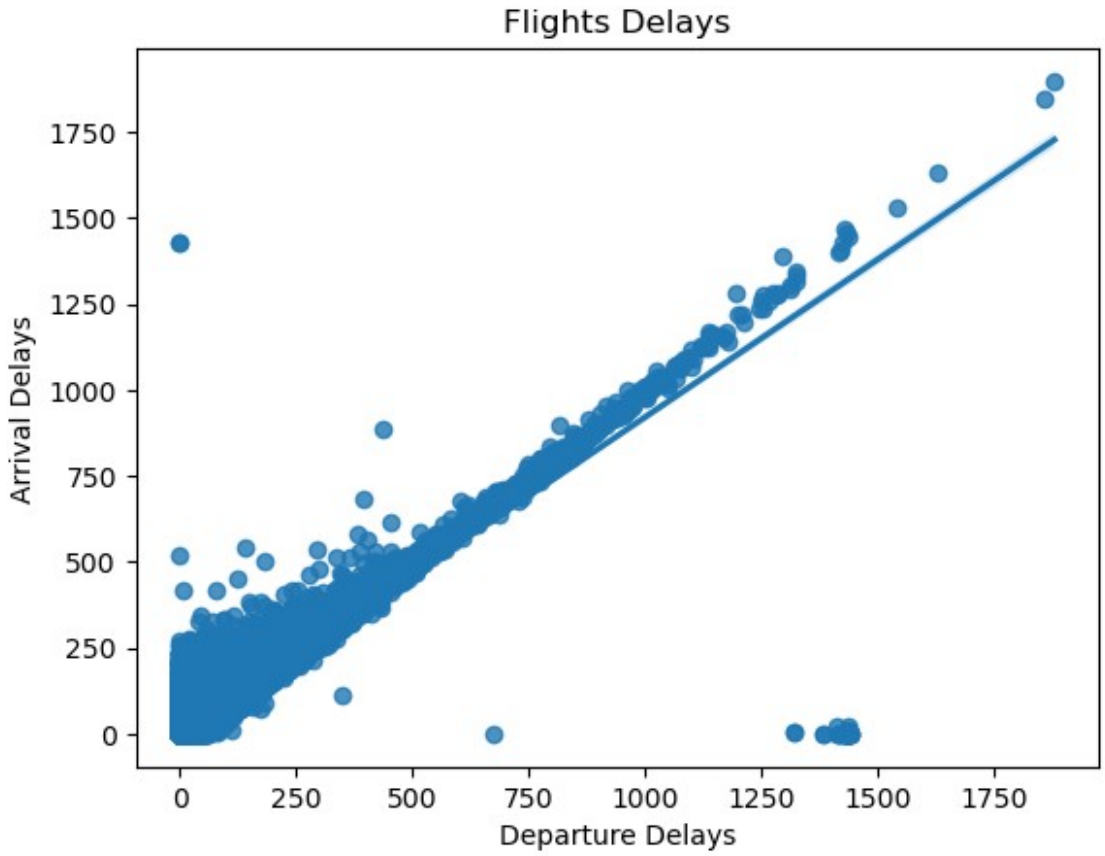
The analysis of flight delays reveals a highly skewed distribution, where the majority of flights experience little to no delay, reflected by the concentration of data at the lower end of the range. However, a small number of flights suffer from significant delays, which extend the distribution and create a long tail. This indicates that while most flights are punctual, attention should be given to the outliers, as they represent instances of considerable delay that could impact overall

operational efficiency and passenger satisfaction. The findings emphasize the need for targeted strategies to address these extreme delays to improve overall airline performance

Relation Between Arrival Delay & Departure Delay

There is a clear linear relationship between arrival and departure delays, indicating that as departure delay increases, arrival delay also tends to rise proportionally. Consequently, the longer a flight is delayed at departure, the more likely it is to be delayed upon arrival. This direct correlation underscores the importance of minimizing departure delays, as they can have a cascading effect on arrival times, potentially disrupting schedules and affecting subsequent flights.

```
regression_scatter_plot(delay_airline_df,
             ['DepDelayMinutes', 'ArrDelayMinutes'],
             'Flights Delays',
             ['Departure Delays','Arrival Delays'])
```

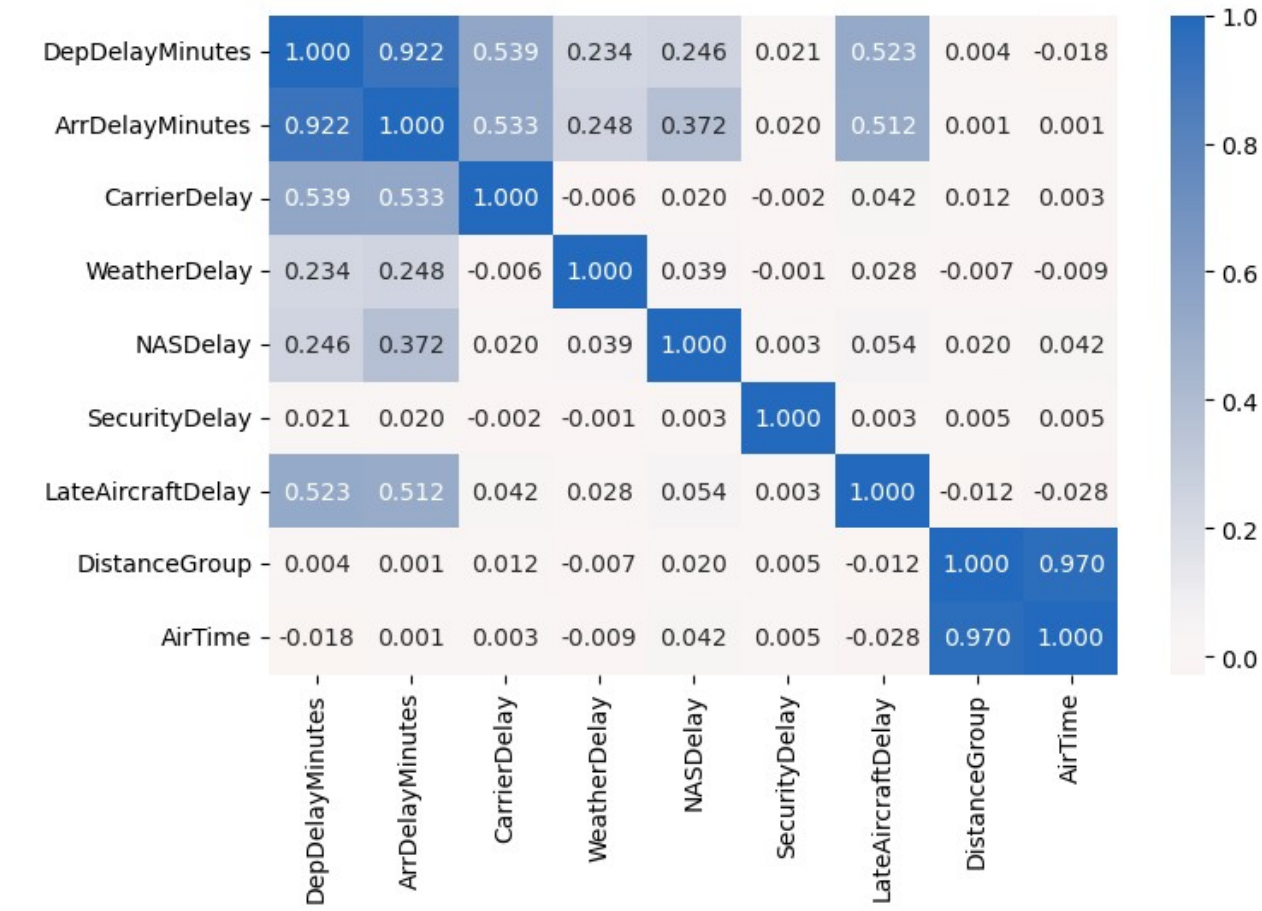


Correlation between delay Features, distance and flight time

The correlation matrix reveals that Arrival and Departure Delays exhibit the strongest correlation, indicating a close relationship between delays at departure and delays upon arrival. Late Aircraft Delay shows a moderate correlation with both arrival and departure delays, suggesting its significant but not predominant role in overall delays. Additionally, Distance

Group and AirTime have a very strong correlation, underscoring the expected link between flight distance and duration.

```
correlation_columns = [
'DepDelayMinutes',
'ArrDelayMinutes',
'CarrierDelay',
'WeatherDelay',
'NASDelay',
'SecurityDelay',
'LateAircraftDelay',
'DistanceGroup',
'AirTime',
]

# correlation plot

plt.figure(figsize = [8, 5])
sb.heatmap(delay_airline_df[correlation_columns].corr(), annot = True,
fmt = '.3f',
           cmap = 'vlag_r', center = 0)
plt.show()
```



Conclusion

Analyzing airline flight performance reveals key insights into both cancellations and delays. Cancellations, though relatively rare at 1.8% of total trips, show a higher frequency on Tuesdays and are notably more prevalent in the first quarter of the year. Causes for cancellations are often unspecified, with specific causes like A and B contributing significantly.

Delays present a more complex picture: arrival delays are strongly correlated with departure delays, emphasizing the cascading effect of late departures on arrival times. Late Aircraft Delays, while significant, are less impactful compared to the direct relationship between departure and arrival delays. Furthermore, Distance Group and AirTime exhibit a very strong correlation, reflecting the expected connection between flight distance and duration.

These findings highlight the importance of addressing departure delays to improve overall flight punctuality and managing cancellations effectively to minimize disruption. Understanding these patterns can aid in optimizing flight operations and enhancing passenger experience.