6610 Group Work

From: Helena Mabey, Isuri Rajapaksa, Lauren Thelen
SUBJECT: Statistics for Business Analysis / United Airline Delays
DATE: 02/08/2025

This analysis examines the impact of weather conditions on United Airlines flight delays between Denver (DIA) and Chicago (ORD) in 2000. The intent is to determine whether weather factors (e.g., snow, precipitation, wind speed) influence increased delays. Additionally, it reviews seasonal trends in delays in conjunction with potential weather-related factors.

Executive Summary

Major Findings:

Weather factors like humidity and snow were linked to longer delays, and rain also had an effect. Delays were longest in July and August averaging around 44 minutes. Additionally, it was found that the presence of snow had a greater impact on delays than the amount of snowfall. Based on the given case data, there could be other factors like air traffic, maintenance, operations, and holiday congestion contributing a role in flight delays. This could be validated as only approximately 21% of delays were affected by weather factors.

Analytical overview

The initial data contained values that were outside of the scope of this analysis, including destination location outside of Chicago and airlines other than United Airlines. This data was removed prior to conducting our analysis. It was also found the data had null values that would impact the analysis so these were replaced with zero values as to not impact the results. Lastly, prior to beginning the analysis, a final data cleaning was completed to remove any other inconsistencies found within the data.

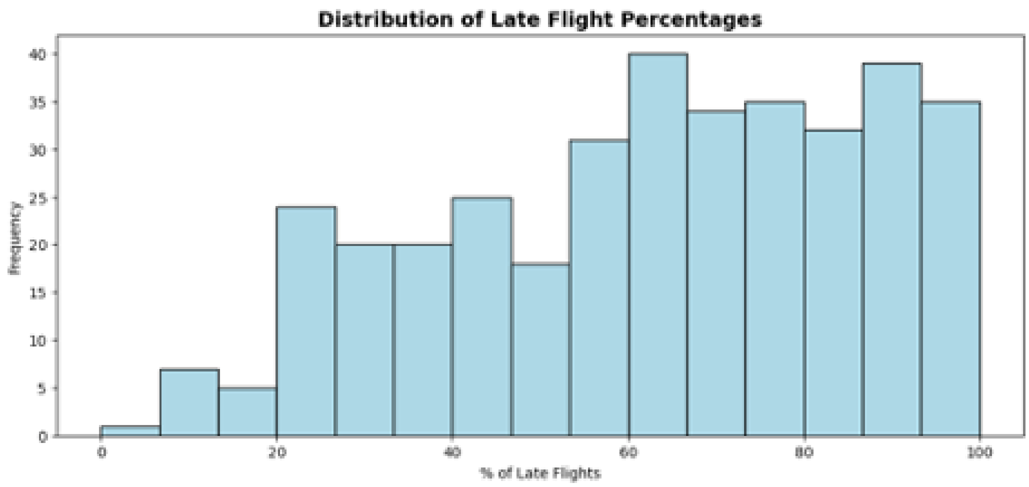
After the initial data cleaning, weather factors were reviewed to determine those with the largest impact. Once those were found, deeper analysis was completed to determine trends and correlation. This information led to the determinations of the impact of weather factors on flight delays.

Appendix

Youtube Link: **https://youtu.be/J7RR36krQ-4**

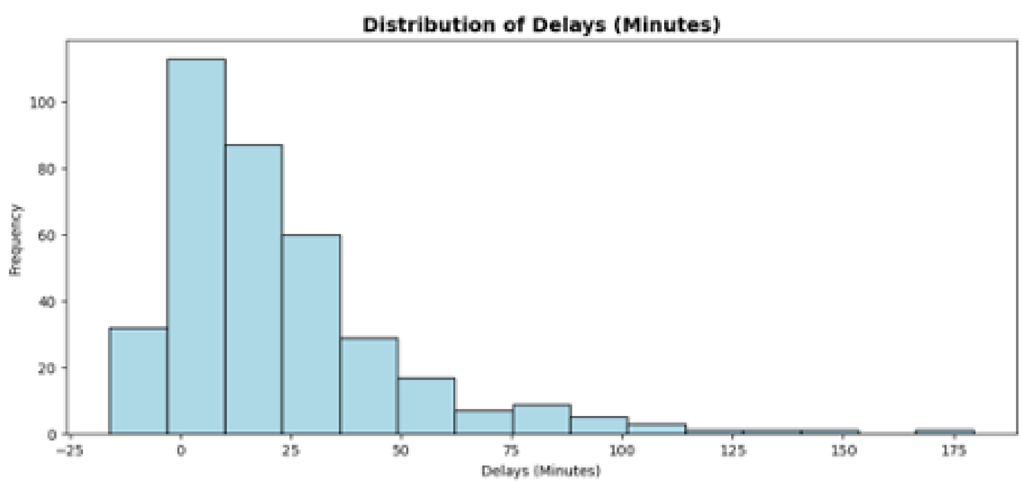
Appendix A: OUTPUT

Figure 1: Distribution of Late Flight Percentages



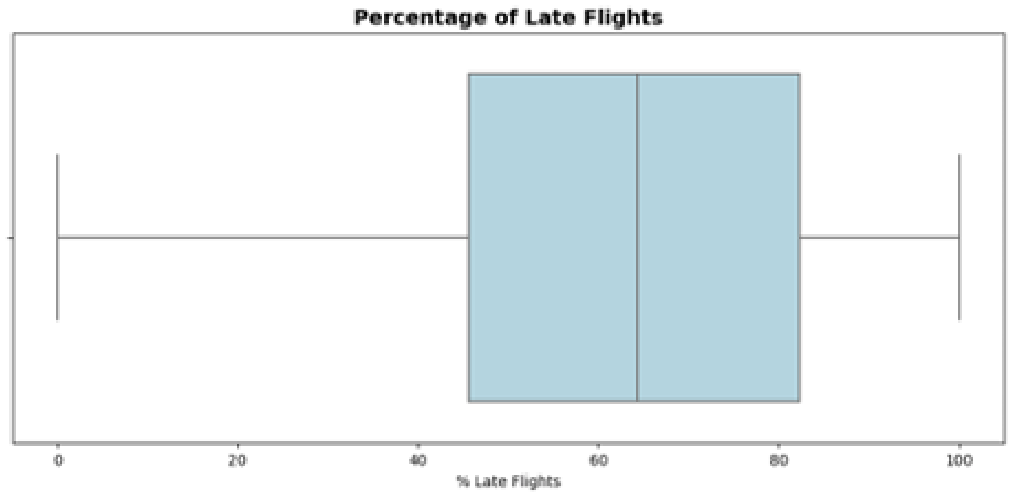
Analyzing the frequency and the percentage of late flights.

Figure 2: Distribution of Delays (Minutes)



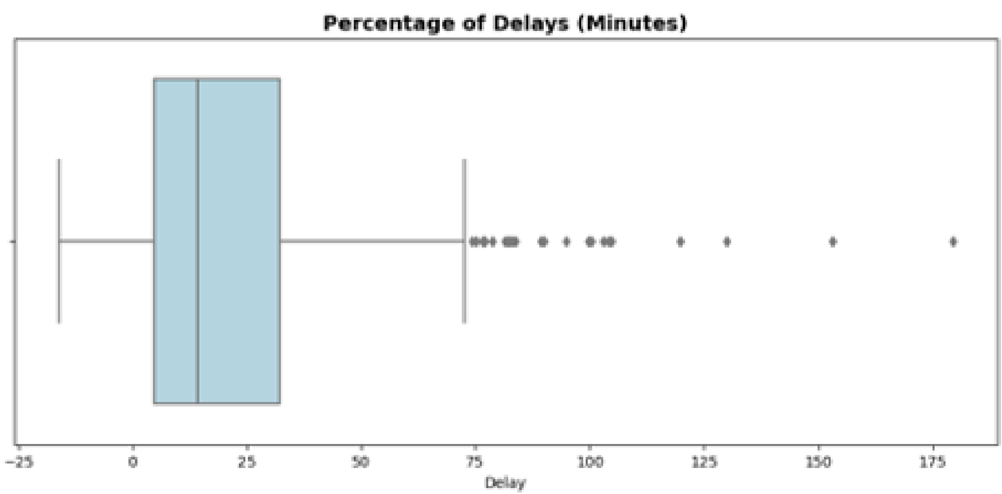
Analyzing the frequency compared to the length of delays in minutes.

Figure 3: Percentage of Late Flights



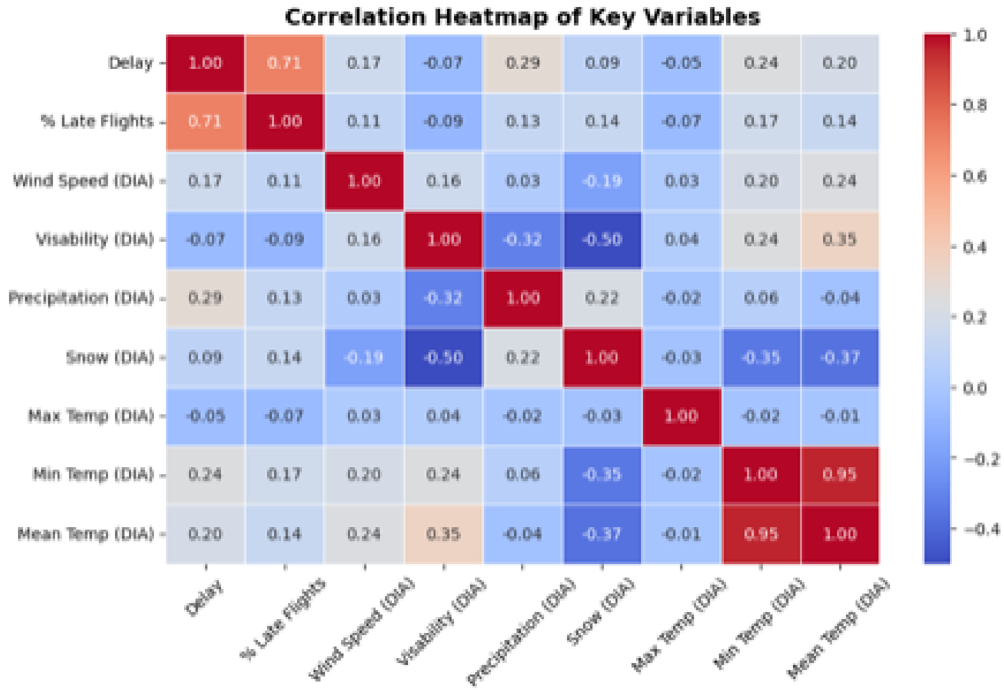
Analyzing what percentage of flights were late.

Figure 4: Percentage of Delays (Minutes)



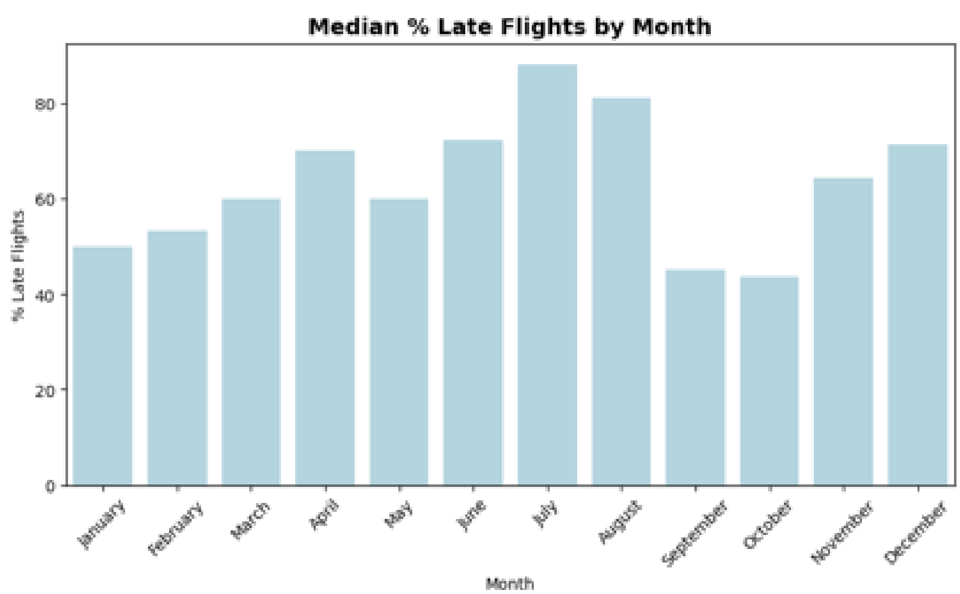
Analyzing the percentage of delayed flights in minutes.

Figure 5: Correlation Heat Map of Key Variables



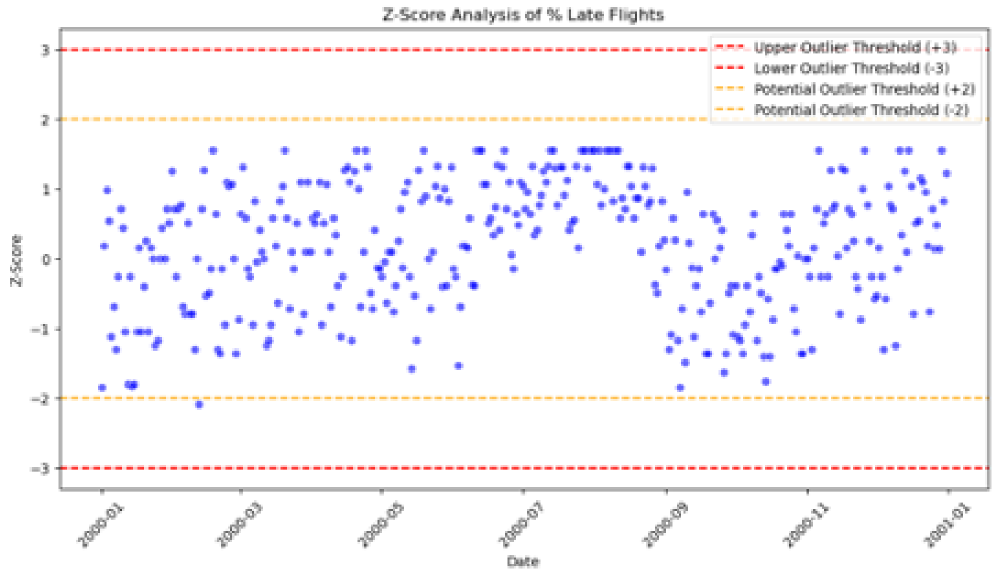
Understanding the correlation between flight delays, percentage of late flights, wind speed, visibility, precipitation, snow, maximum temperature, minimum temperature, and mean temperature at DIA.

Figure 6: Median Percentage of Late Flights by Month



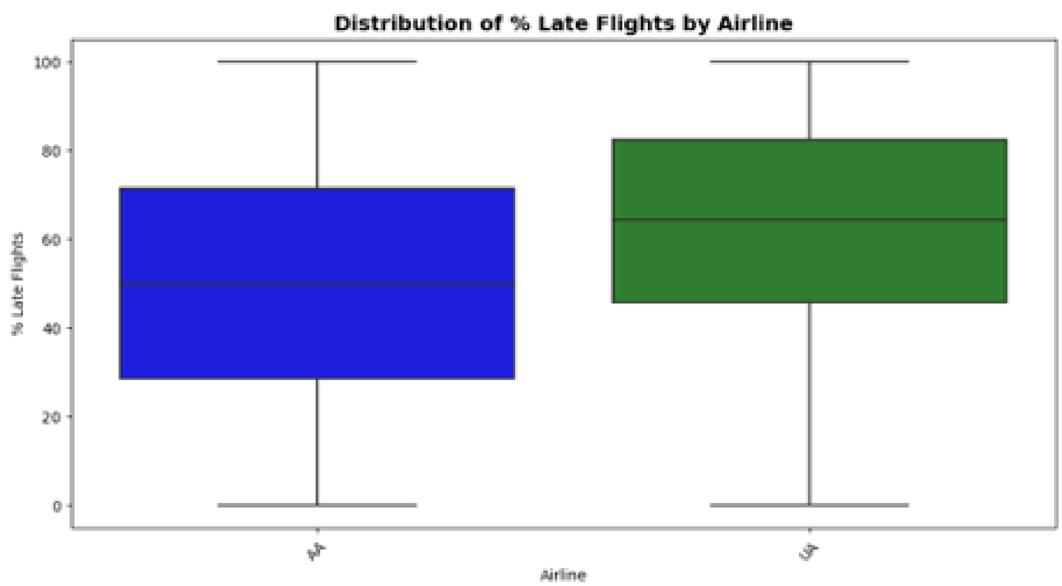
Analyzing the percentage of late flights and what month they occurred in.

Figure 7: Z-Score Analysis of the Percentage of Late Flights

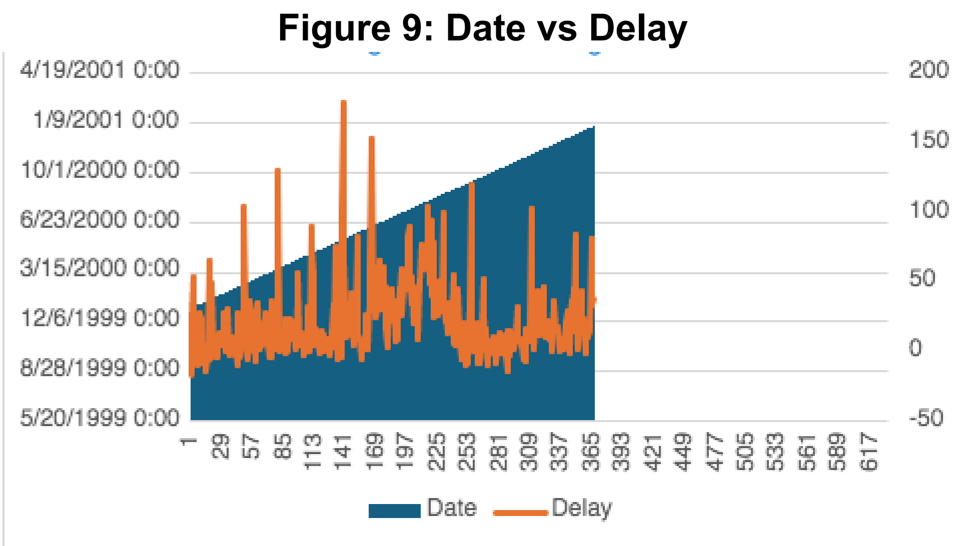


Comparing the Z-score with the date of the late flight occurrences.

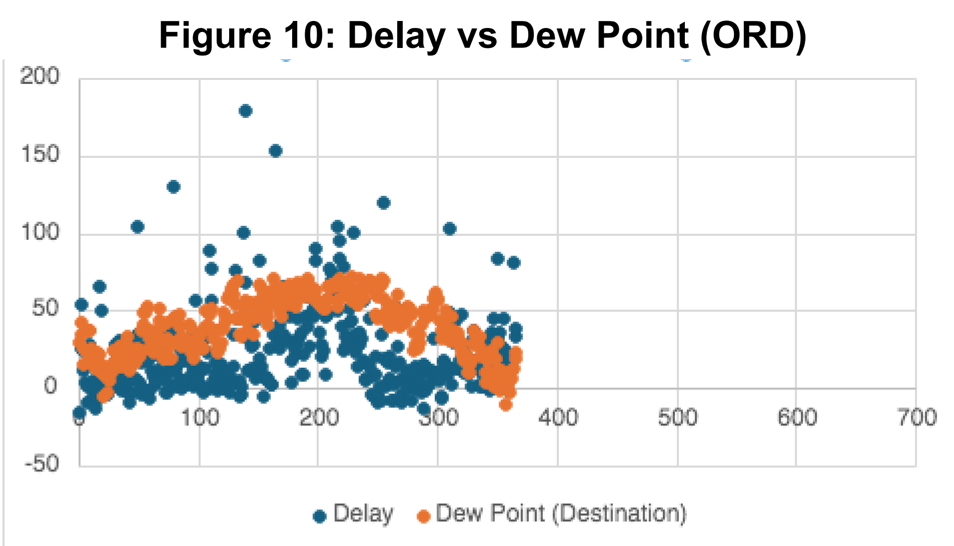
Figure 8: Distribution of the Percentage of Late Flights by Airline



Comparing the percentage of late flights between United Airlines and American Airlines.

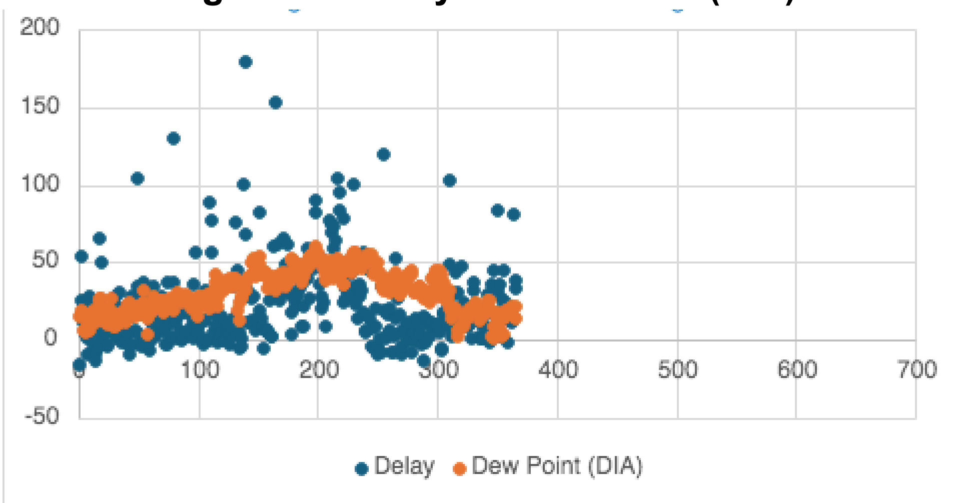


Comparing the date of the flights and the length of the delay.



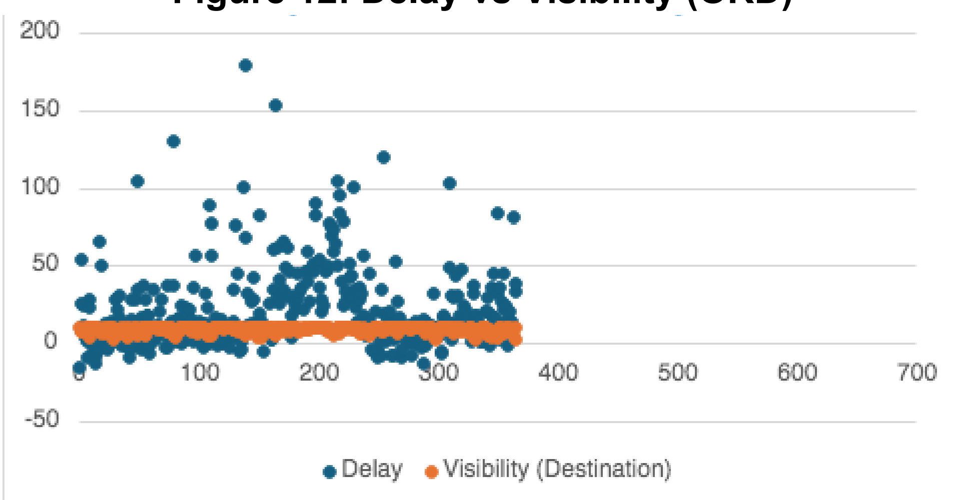
Comparing the length of the delay with the dew point in ORD, the arrival location.

Figure 11: Delay vs Dew Point (DIA)

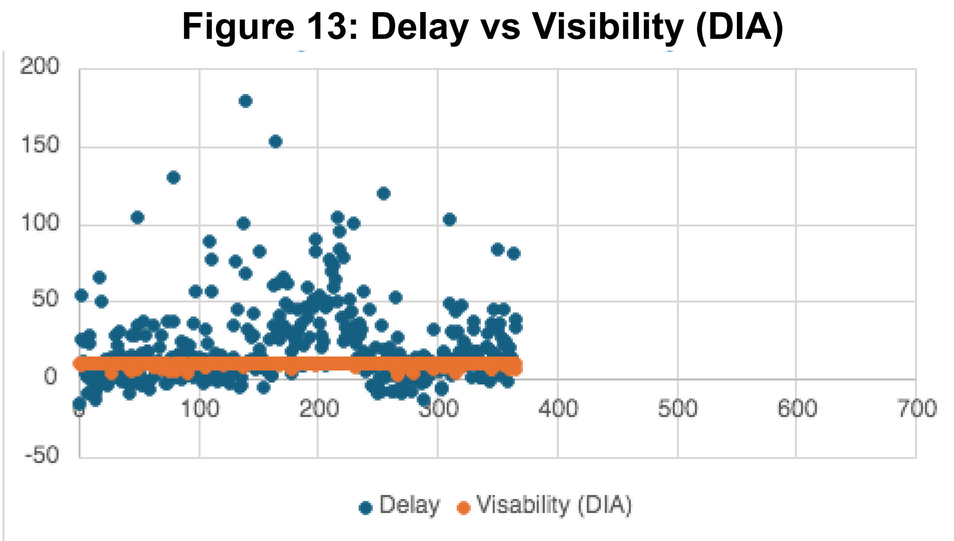


Comparing the length of delay with the dew point at DIA, the departure location.

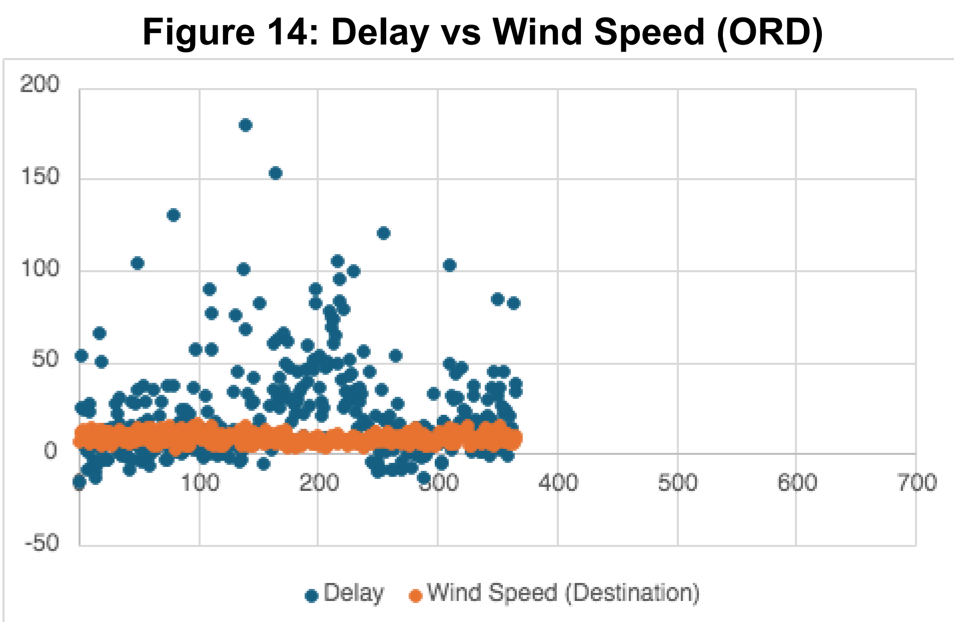
Figure 12: Delay vs Visibility (ORD)



Comparing length of delay to the visibility level at ORD.

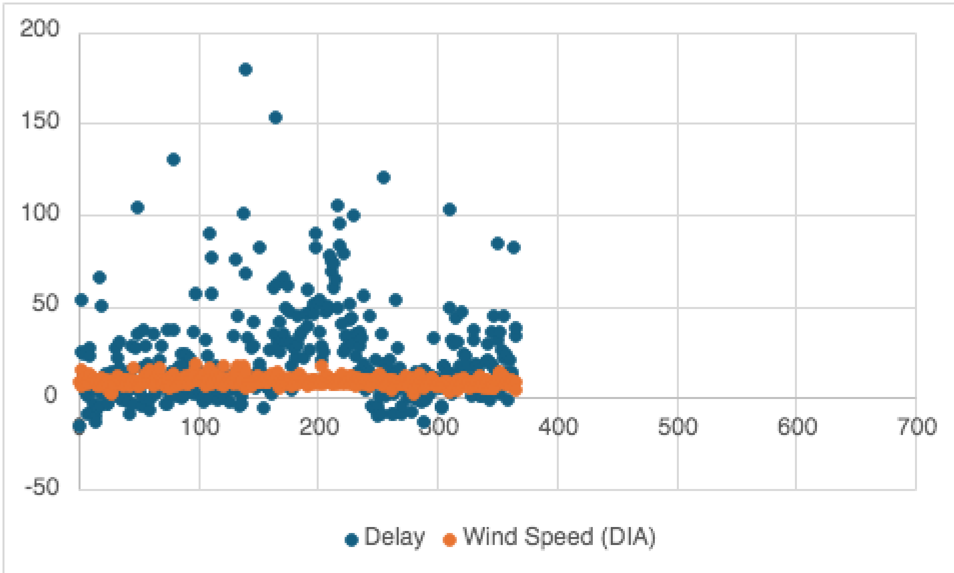


Comparing the length of delay with the visibility level at DIA.



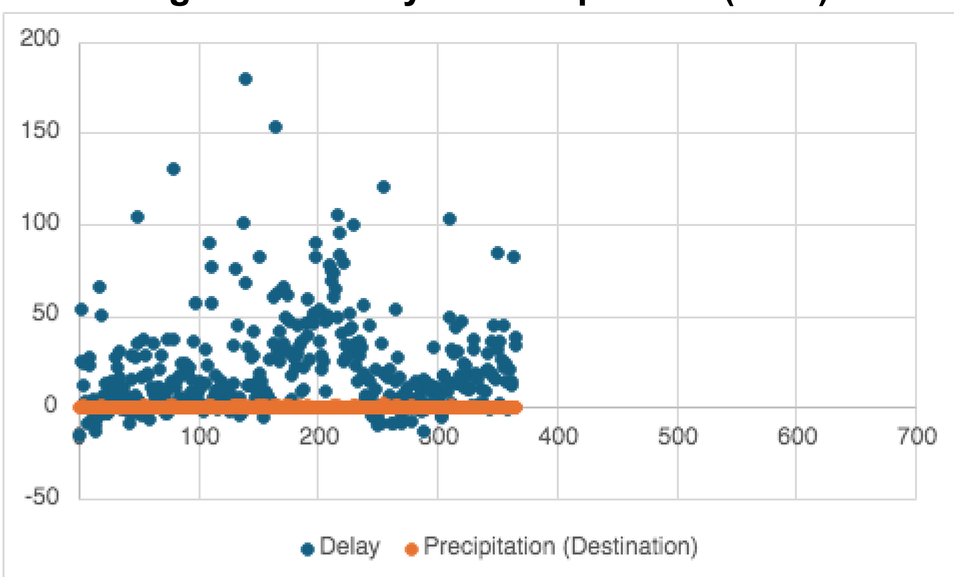
Comparing the length of delay with the wind speed at ORD.

Figure 15: Delay vs Wind Speed (DIA)

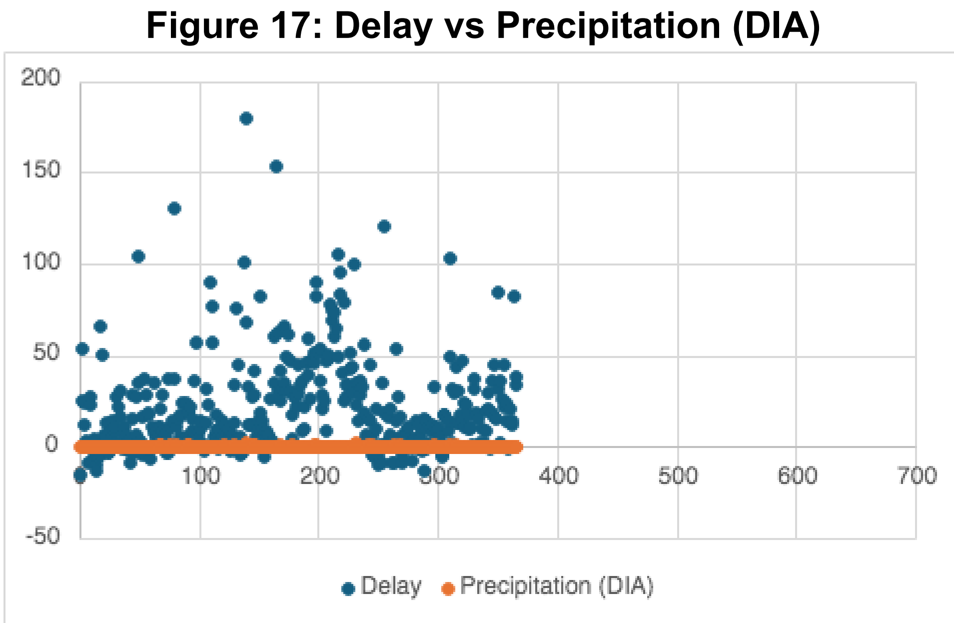


Comparing the length of delay with the wind speed at DIA.

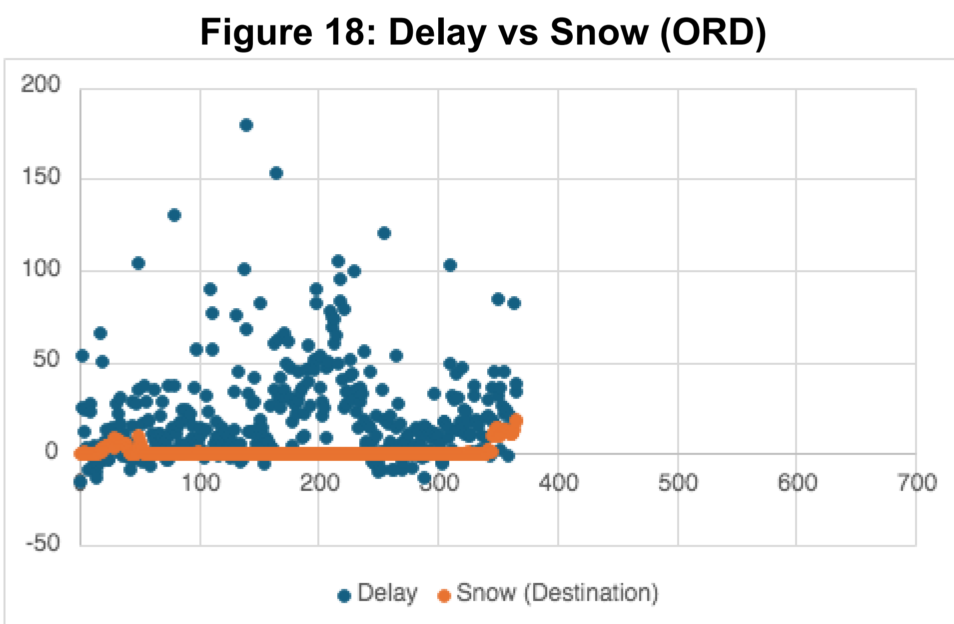
Figure 16: Delay vs Precipitation (ORD)



Comparing the length of delay with the level of precipitation at ORD.

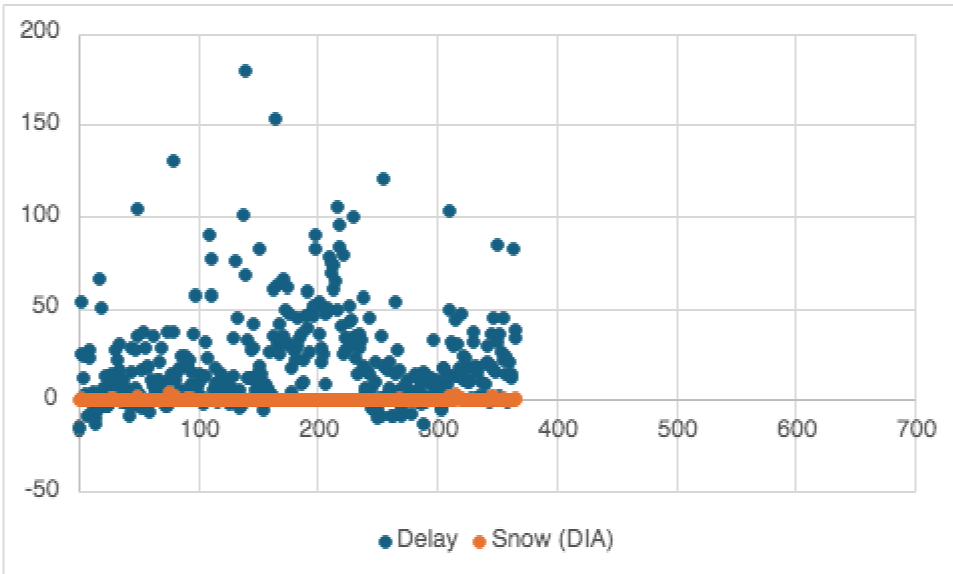


Comparing the length of delay with the level of precipitation at DIA.



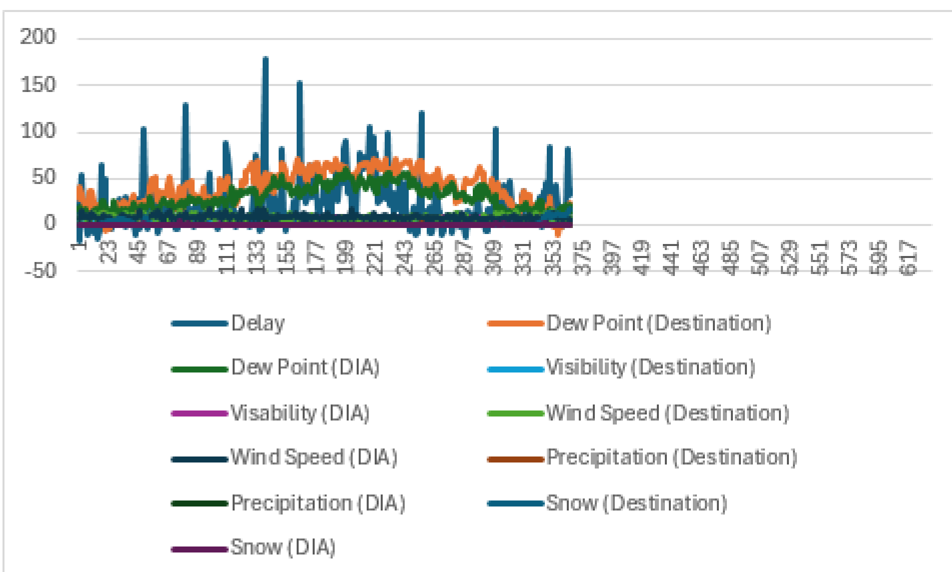
Comparing the length of the delay with the snow conditions at ORD.

Figure 19: Delay vs Snow (DIA)



Comparing the length of delay vs the snow conditions at DIA.

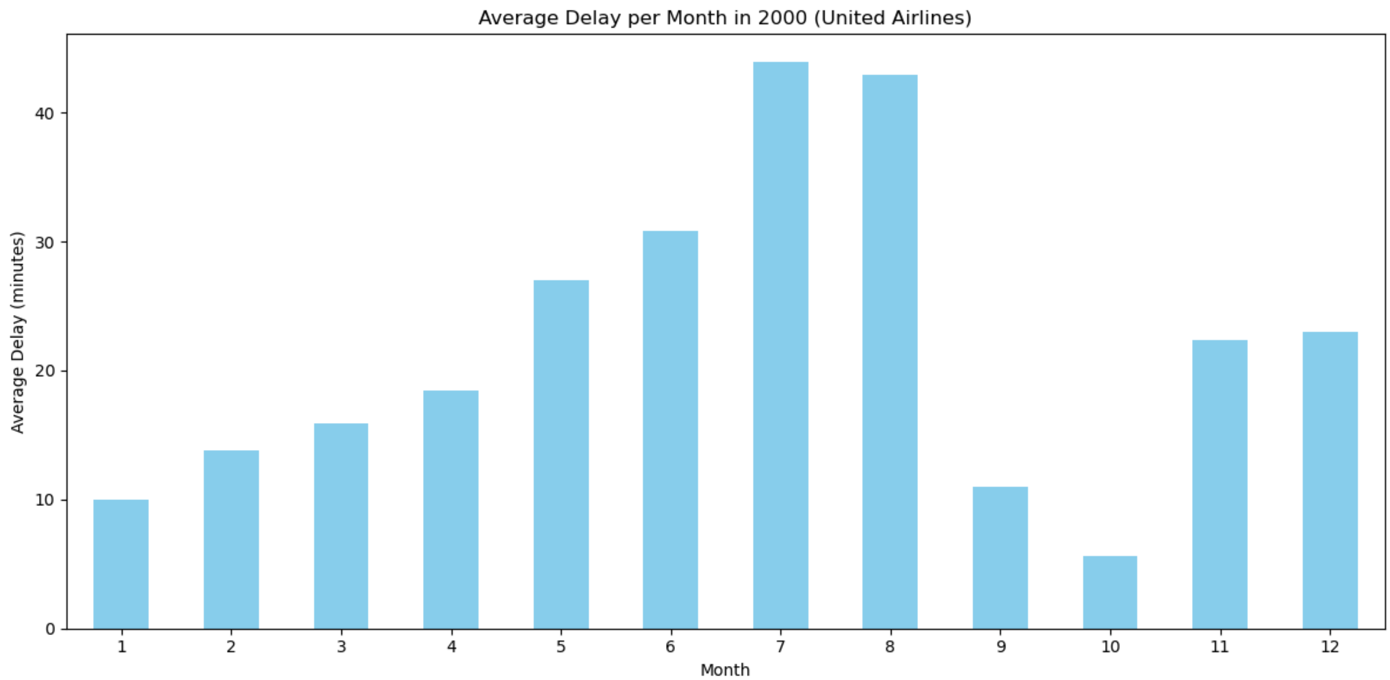
Figure 20: Delay vs Weather Conditions



Highlighting the length of delay with the present weather conditions.

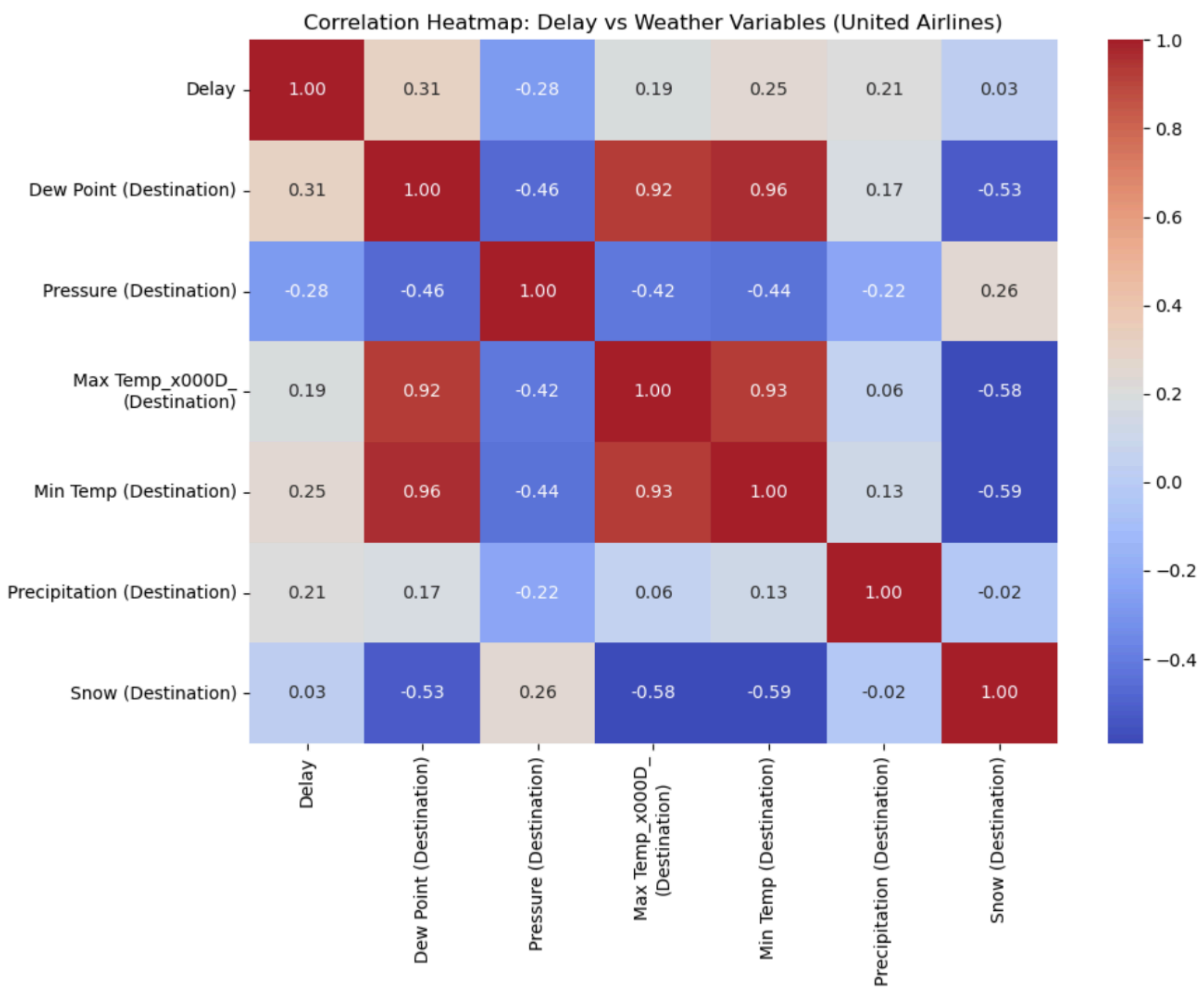
.

Figure 21: Average Delay per Month Bar Chart



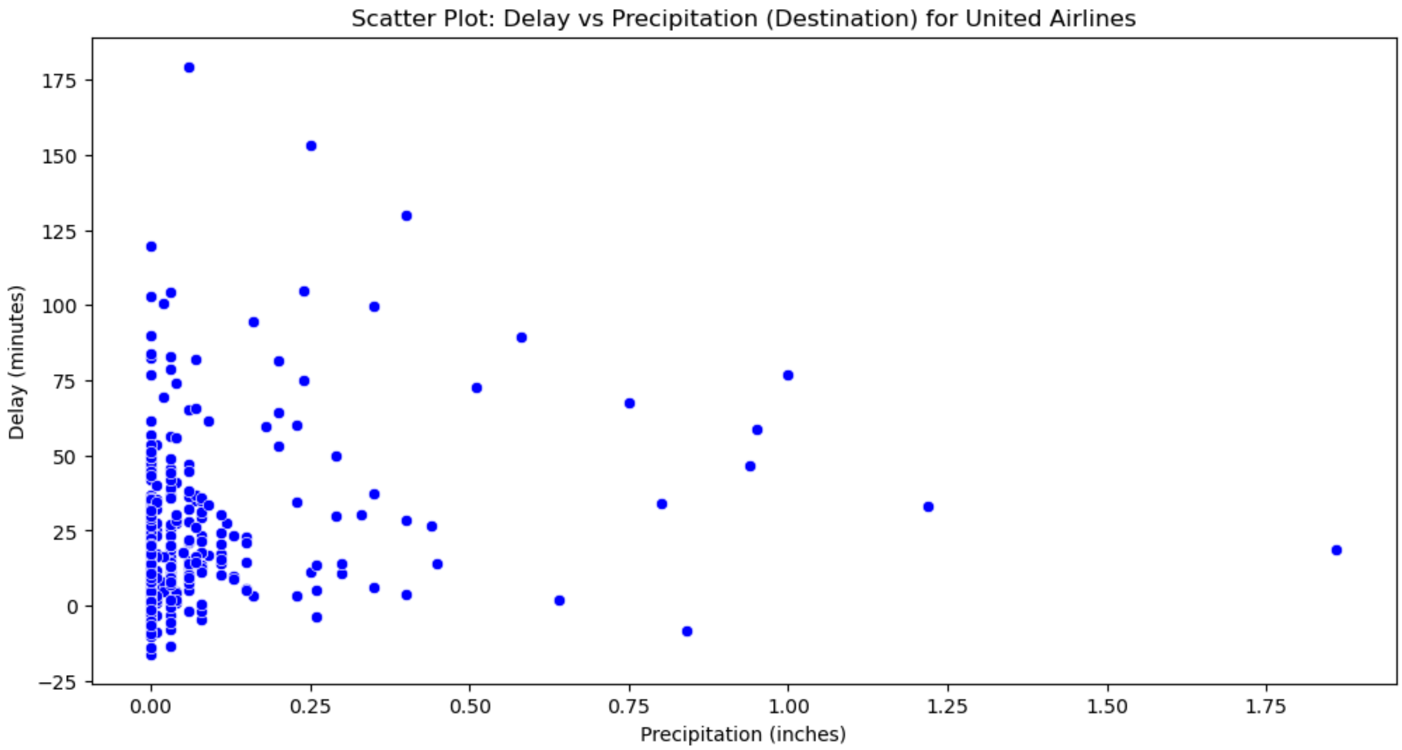
Analyzing the average length of flight delay and the month that they occurred.

Figure 22: Delay Vs Weather Variables Correlation Heat Map



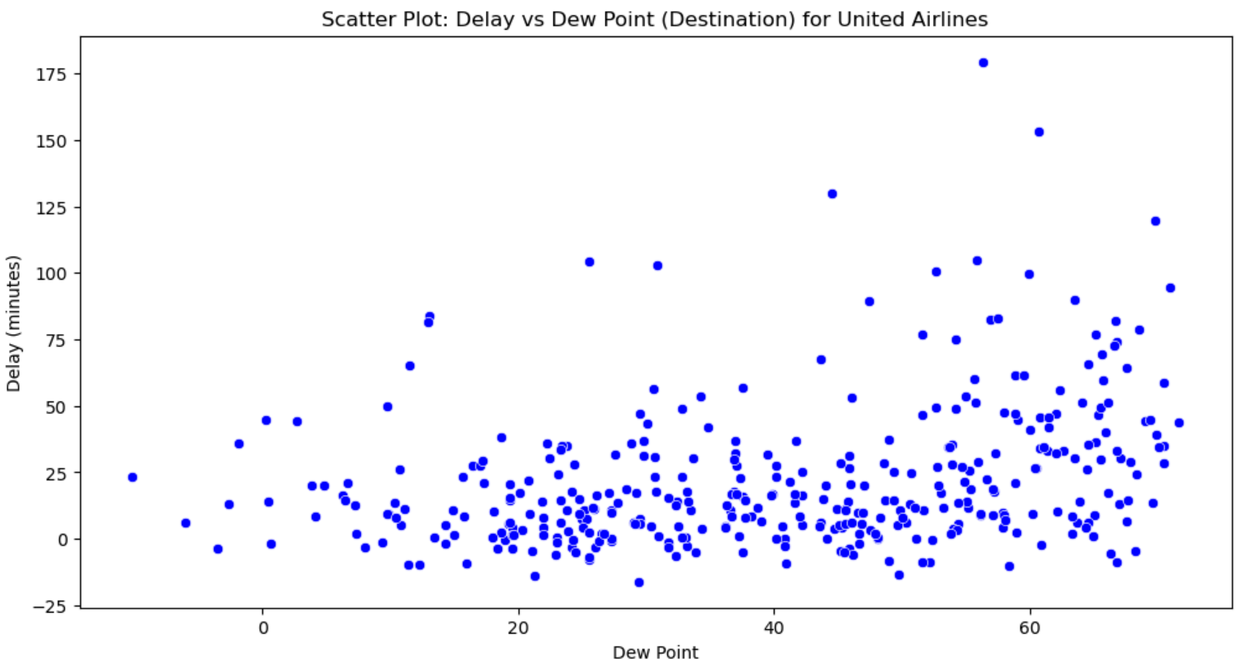
Understanding the correlation between flight delays, dew point, pressure, maximum temperature, minimum temperature, precipitation and snow levels.

Figure 23: Delay Vs Precipitation



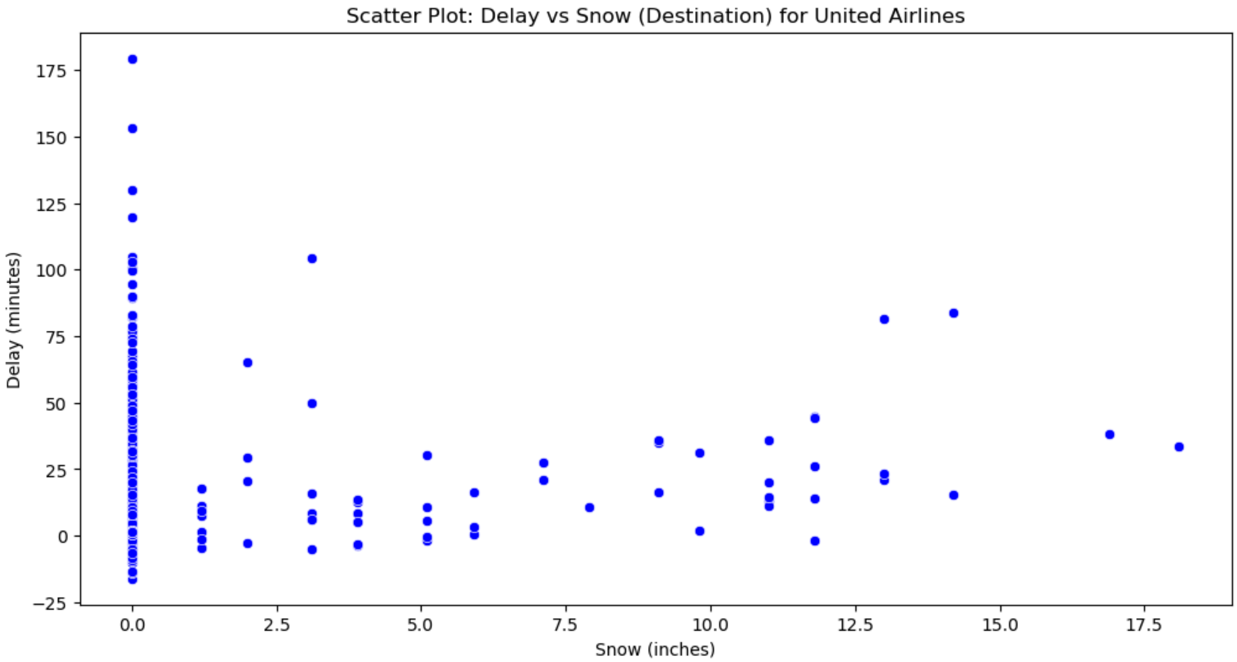
Comparing the length of delay to the precipitation level.

Figure 24: Delay Vs Dew Point



Comparing the flight delay length to the dew point at ORD.

Figure 25: Delay Vs Snow



Comparing the length of flight delay to the snow level at ORD.

Figure 26: Delay Vs Weather Variables

Analyzing the flight delay compared to the weather conditions.

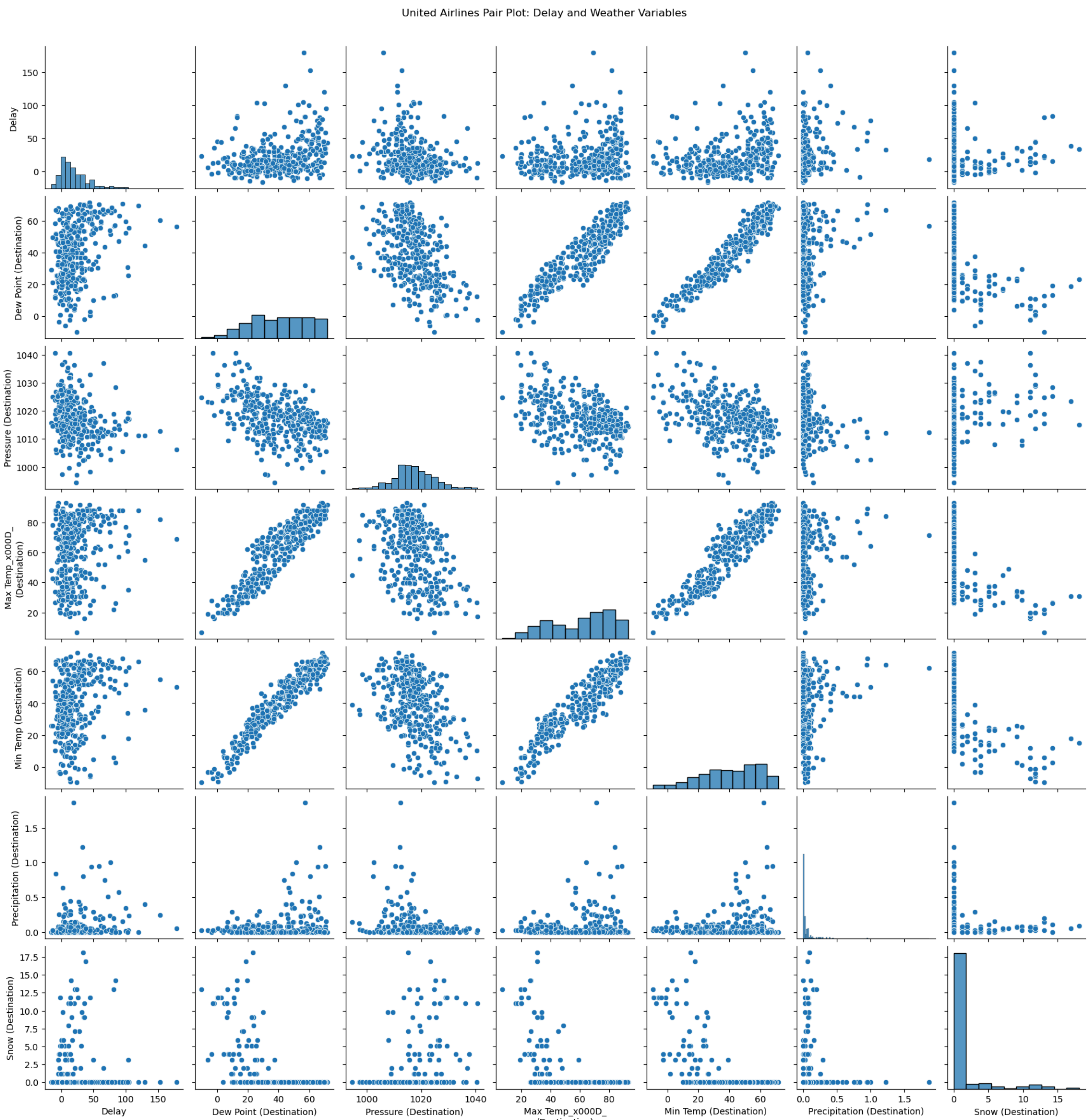
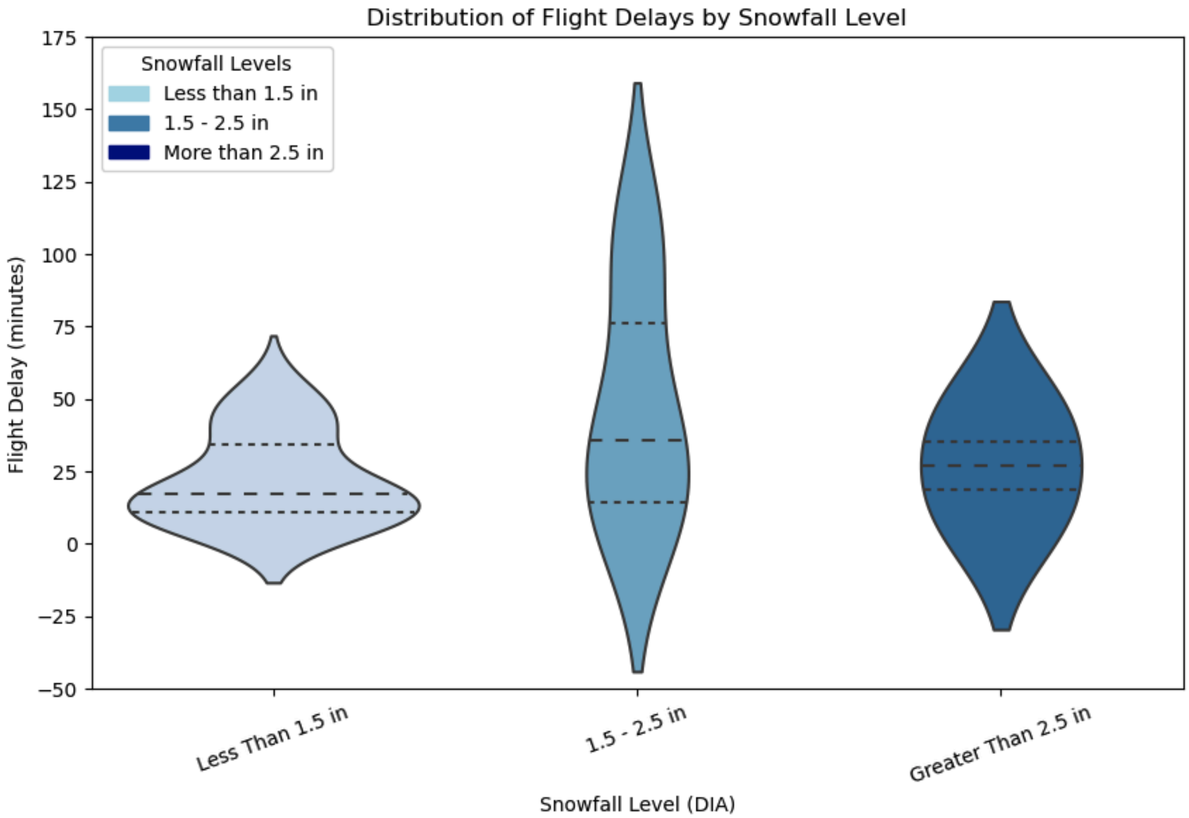


Figure 27: Flight Delay Vs Snowfall



This violin plot shows the distribution and variability of flight delays across different snowfall levels.

Regression Summary

United Airlines

```
Regression Summary for United Airlines:                          OLS Regression Results
==============================================================================
Dep. Variable:                  Delay   R-squared:                       0.212
Model:                            OLS   Adj. R-squared:                  0.199
Method:                 Least Squares   F-statistic:                     16.10
Date:                Sat, 08 Feb 2025   Prob (F-statistic):           2.01e-16
Time:                        14:47:32   Log-Likelihood:                -1678.4
No. Observations:                 366   AIC:                             3371.
Df Residuals:                     359   BIC:                             3398.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                        620.1143    203.985      3.040      0.003     218.960    1021.269
Dew Point (Destination)        1.1046      0.260      4.248      0.000       0.593       1.616
Pressure (Destination)        -0.5976      0.199     -3.003      0.003      -0.989      -0.206
Max Temp_x000D_
(Destination)    -0.5566      0.178     -3.131      0.002      -0.906      -0.207
Min Temp (Destination)        -0.0750      0.271     -0.277      0.782      -0.608       0.458
Precipitation (Destination)   11.6029      7.360      1.576      0.116      -2.872      26.078
Snow (Destination)             1.7637      0.522      3.376      0.001       0.736       2.791
==============================================================================
Omnibus:                      158.058   Durbin-Watson:                   1.446
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              740.971
Skew:                           1.828   Prob(JB):                    1.26e-161
Kurtosis:                       8.935   Cond. No.                     1.66e+05
==============================================================================
```

Summary Statistics
Delay and % Late Flights

```
count     366.000000
min       -16.066667
max       179.400000
mean       22.122451
median     14.248718
std        26.776391
var       716.975126
skew        1.970956
kurt        5.722543
Name: Delay, dtype: float64
```

```
count     366.000000
min         0.000000
max       100.000000
mean       62.731577
median     64.285714
std        24.690584
var       609.624949
skew       -0.348009
kurt       -0.807596
Name: % Late Flights, dtype: float64
```

Appendix B: CODE

Lauren -

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt #
data = {
"Day": list(range(1, 24)),
"Delay": [-16.07, 25.06, 53.47, 11.39, 23.53, 3.00, -9.29, 0.67, 27.38, 22.92, -8.00, -3.62, 4.29,
-9.71, -13.85, -5.93, 65.38, -4.64, 1.62, 49.93],
"Dew Point (Chicago)": [29.4, 42.2, 34.3, 25.9, 15.7, 23.9, 16, 23, 37.1, 37.4, 25.6, 19.5, 22,
12.3, 21.3, 22.9, 11.5, 21.1, 15, 9.8],
"Pressure (Chicago)": [1015.7, 1006.4, 1010.9, 1008.9, 1025.1, 1021.4, 1026.7, 1019.7, 1010,
994.6, 1005.1, 1020.2, 1023.7, 1040.5, 1024.9, 1025.5, 1037, 1021.3, 1017.4, 1015.7] }
df = pd.DataFrame(data)
# Set seaborn style
sns.set(style="whitegrid")#
Scatter Plot: Delay vs. Dew Point
plt.figure(figsize=(8, 5))
sns.scatterplot(x=df["Dew Point (Chicago)"], y=df["Delay"], color="blue", marker="o")
plt.title("Scatter Plot: Flight Delay vs. Dew Point (Chicago)")
plt.xlabel("Dew Point (Chicago)") plt.ylabel("Flight Delay (Minutes)")
plt.show()
Violin Chart: Delay Distribution by Dew Point
plt.figure(figsize=(8, 5))
sns.violinplot(x=df["Dew Point (Chicago)"], y=df["Delay"], inner="quartile", palette="coolwarm")
plt.title("Violin Chart: Flight Delay vs. Dew Point (Chicago)")
plt.xlabel("Dew Point (Chicago)") plt.ylabel("Flight Delay (Minutes)") plt.xticks(rotation=45)
plt.show()
Line Chart: Delay Over Time (Days)
plt.figure(figsize=(8, 5))
sns.lineplot(x=df["Day"], y=df["Delay"], marker="o", linestyle="-", color="red")
plt.title("Line Chart: Flight Delay Over Time")
plt.xlabel("Day")
plt.ylabel("Flight Delay (Minutes)")
plt.grid(True) plt.show()
```

Isuri -

```
import statsmodels.api as sm
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


# Load the data
file_path = '/Users/isurirajapaksa/Desktop/BANA 6610/Homework2/Airlines delay new1.xlsx'
df = pd.read_excel(file_path, sheet_name='Sheet1')

# Preprocessing- double check
df['Date'] = pd.to_datetime(df['Date'])
df['Month'] = df['Date'].dt.month

# All selected columns are numeric- double check
for col in weather_features + ['Delay']:
    ua_data[col] = pd.to_numeric(ua_data[col], errors='coerce')


# Filter data for United Airlines
ua_data = df[df['Airline'] == 'UA']

# Selecting relevant weather variables for regression
weather_features = ['Dew Point (Destination)', 'Pressure (Destination)', 'Max
Temp_x000D_\n(Destination)',
                'Min Temp (Destination)', 'Precipitation (Destination)', 'Snow (Destination)']

# Regression for United Airlines (with the support of CGPT, tested multiple times)
X_ua = sm.add_constant(ua_data[weather_features])
y_ua = ua_data['Delay']
model_ua = sm.OLS(y_ua, X_ua).fit()

print("\nUnited Airlines- Regression Summary:")
print(model_ua.summary())

# Bar Chart (Average delay per month for United Airlines)
avg_delay_per_month_ua = ua_data.groupby('Month')['Delay'].mean()
plt.figure(figsize=(12, 6))
avg_delay_per_month_ua.plot(kind='bar', color='skyblue')
plt.title('Average Delay per Month in 2000 (United Airlines)')
plt.xlabel('Month')
plt.ylabel('Average Delay (minutes)')
```

```
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()


# Heatmap (Correlation between delays and weather variables for United Airlines)

corr_matrix_ua = ua_data[['Delay'] + weather_features].corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix_ua, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap: Delay vs Weather Variables (United Airlines)')
plt.show()

#  Scatter Plot (Delay vs Precipitation for United Airlines)
plt.figure(figsize=(12, 6))
sns.scatterplot(data=ua_data, x='Precipitation (Destination)', y='Delay', color='blue')
plt.title('Scatter Plot: Delay vs Precipitation (Destination) for United Airlines')
plt.xlabel('Precipitation (inches)')
plt.ylabel('Delay (minutes)')
plt.show()

#  Scatter Plot (Delay vs Dew Point for United Airlines)
plt.figure(figsize=(12, 6))
sns.scatterplot(data=ua_data, x='Dew Point (Destination)', y='Delay', color='blue')
plt.title('Scatter Plot: Delay vs Dew Point (Destination) for United Airlines')
plt.xlabel('Dew Point')
plt.ylabel('Delay (minutes)')
plt.show()


#  Scatter Plot (Delay vs Snow for United Airlines)
plt.figure(figsize=(12, 6))
sns.scatterplot(data=ua_data, x='Snow (Destination)', y='Delay', color='blue')
plt.title('Scatter Plot: Delay vs Snow (Destination) for United Airlines')
plt.xlabel('Snow (inches)')
plt.ylabel('Delay (minutes)')
plt.show()


# Drop NaNs
ua_data_clean = ua_data.dropna(subset=weather_features + ['Delay'])

# Pair Plot for United Airlines
sns.pairplot(ua_data_clean, vars=['Delay'] + weather_features, height=2.5)
plt.suptitle('Pair Plot: Delay and Weather Variables (United Airlines)', y=1.02)
plt.show()
```

(Python code attached in submission)

- Python Notebook Code: Helena Mabey

  h_mabey_hw_2_part1.pdf
  Github Link:
  https://github.com/helenamabey/spring_stats_hw2_part1/tree/main

AI statement

As a team, we opted to use AI as a tool for our project. AI was useful during this project when it came to giving example codes, giving the steps to run certain codes and guiding us through creating certain graphs. We used ChatGPT to ask clarifying questions, explain codes and assist us throughout this case study. We are all relatively new to using Python, so using AI as a resource was very helpful in the completion of this assignment.

Responsibilities:

Person	**Responsibilities**
Helena	**Data Cleaning & Summary Stats** - Preprocess dataset - Compute basic stats ● Mean ● Median ● Mode ● Standard deviation - Analyze distribution of delays ● Histograms ● boxplots - Write findings for the executive summary - Create slides for this section - 3.5 min recording
Lauren	**Trends Over Time & Weather Correlation** - Identify seasonal trends (delays by month) - Perform correlation analysis (weather vs. delays) Graphs for all findings ● Scatterplots ● Violin chart ● line chart - Write findings for the executive summary - Create slides for this section 3.5 mins recording
Isuri	**Regression Analysis & Summary** - Run regression to quantify weather's impact ● Regression Summary Chart- ● Bar Chart ● Heat map ● Scatter plots ● Pair plot - Write the final section of the executive summary - Create slides for this section - 3.5 mins recording