Problem Set 4

Motivation behind the topic and data

In this study I chose to work on flight delay data which in my opinion seemed more fruitful for this project. Flight delays can lead to significant financial losses and impact the broader economy. Understanding the causes and patterns of these delays is crucial for improving the efficiency of the aircraft system. This allows for better scheduling, reduced airport congestion, and enhanced overall operational effectiveness. Delays are also a major source of frustration for passengers. By anticipating and managing delays, airlines can enhance customer service and satisfaction. A deeper understanding of the factors contributing to delays can help aircraft make strategic decisions to curb them and provide passengers with information to avoid potential problems.

Introduction

There are several primary reasons that may lead a flight to delay: Carrier delay, weather delay, previous flight delay, departure delay. For carrier delay, I want to first display the average delay caused by each carrier. Then I want to see the relationship between carrier delay and the departure hour (for example: are flights that depart in the evening more subject to carrier delay than those departs in the morning? This can measure the working effectiveness of the carrier employees during different hours). Moreover, I want to explore the relationship between the month and weather (In which season of the year are flights most subject to weather delay? Weather is more dependent on season rather than the time within a day). Also, I want to discuss during which hours are planes most subject to previous flight delay and departure delay (since these factors are more related to busy hours rather than the season or the year).

The dataset contains plains that are scheduled to travel from LAX to JFK, the route of which goes across the whole America Land. This is representative of other flights in the United States. Also, this enables us to ignore the factor of flight duration since the flight duration is not likely to vary a lot. The missing values in the data will be treated as “zeros” for purposes of maintaining our sample. The negative values in the delays in minutes will be treated as “zero” too since a flight that left earlier or arrived earlier has actually been delayed for 0 minutes. Since the dataset has been already downloaded, the following code snippet was used to read into the working directory.

```
# import data into the environment and Set missing values to 0
flight_data = pd.read_csv('lax_to_jfk.csv', skiprows = 0).fillna(0)
```

Results and findings

Table 1 gives the descriptive statistics of delay time in minutes. Notably, the arrival delay and departure delay have negative minimum values, which is an indication that some flights left their origin earlier than scheduled and some arrived to their destination earlier than scheduled. The average carrier delay is around (M =2.33 minutes, SD = 21.57). The standard deviation is significantly large which implies the observations are deviating so much from the mean. The average arrival delay for all the flights is approximately, M = 12.82 minutes, SD = 36.40. The average departure delay is around, M = 9 minutes, SD = 35.63. The mean delay due to security reasons was 0.09 minutes, SD = 3.41 and the average delay due to weather was 0.13 minutes, SD = 2.71. Notably, weather and security did not subject the planes to too much delays.

Figure 1 shows the annual distribution of average arrival delays for all flights. There are two notable peaks in mean arrival delays. The first one occurs around 1990, with delays over 600 minutes, and the second one around 2016, with approximately 700 minutes of delay. Between these peaks, the delays fluctuate but are significantly lower than the peak years. The results suggest that there might be specific events or systemic issues in those peak years that caused such high delays. It could be beneficial to investigate what happened in the aviation industry during those times to understand the reasons behind these spikes. This kind of data can be very useful for identifying patterns and potential areas for improvement in flight scheduling and management. These delays witnessed over the years can be attributed to various factors including security, weather, aircraft capacity, technical issues, procedures, and maybe also propagated delays. Comprehending these factors can help in identifying trends and potential improvements in airline and airport operations to minimize delays. For instance, analyzing the spike in delays around 1990 and 2016 could reveal specific systemic issues or external events that led to increased delays during those years.

Table 1: Descriptive statistics

	ArrDelay	ArrDelayMinutes	CarrierDelay	DepDelay	DepDelayMinutes	SecurityDelay	WeatherDelay
count	2855.00	2855.00	2855.00	2855.00	2855.00	2855.00	2855.00
mean	3.97	12.82	2.33	9.00	10.84	0.09	0.13
std	40.99	36.40	21.57	35.63	34.97	3.41	2.71
min	-73.00	0.00	0.00	-19.00	0.00	0.00	0.00
25%	-16.00	0.00	0.00	-3.00	0.00	0.00	0.00
50%	-3.00	0.00	0.00	0.00	0.00	0.00	0.00
75%	12.00	12.00	0.00	6.00	6.00	0.00	0.00
max	682.00	682.00	680.00	728.00	728.00	168.00	109.00

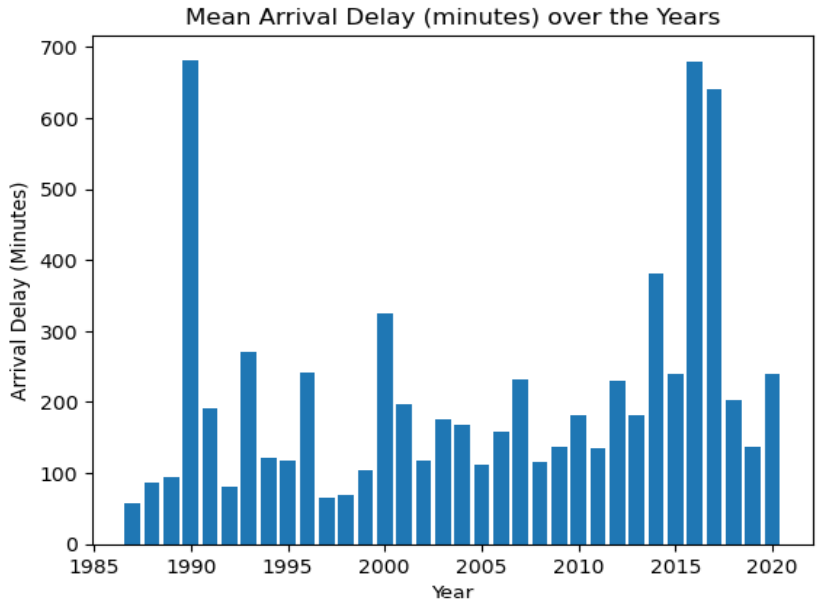

Figure 1: Mean arrival time (minutes) by year

Figure 2, shows the distribution of carrier delays by the departure hour of the aircraft. Most flights have shorter average delays, as indicated by the cluster of data points near the bottom of the chart. There are significant outliers with higher delays. One peak occurs around midnight with a delay close to 700 minutes, and another in the evening at 19:12 with a delay around 600 minutes (figure 2). Most of the carrier delays were witnessed between 04:48 in the morning to the evening at around 19:12. The peak around midnight could be due to fewer staff, security checks, weather delays, or operational limitations during late hours. The evening peak might be related to increased air traffic during that time, which can cause longer wait times for takeoff clearance. These insights suggest that departure time is a crucial factor in predicting potential delays. Airlines and airports might use this data to optimize schedules and reduce delay times. For instance, they could increase staffing during peak delay hours or adjust flight schedules to avoid congested departure times.

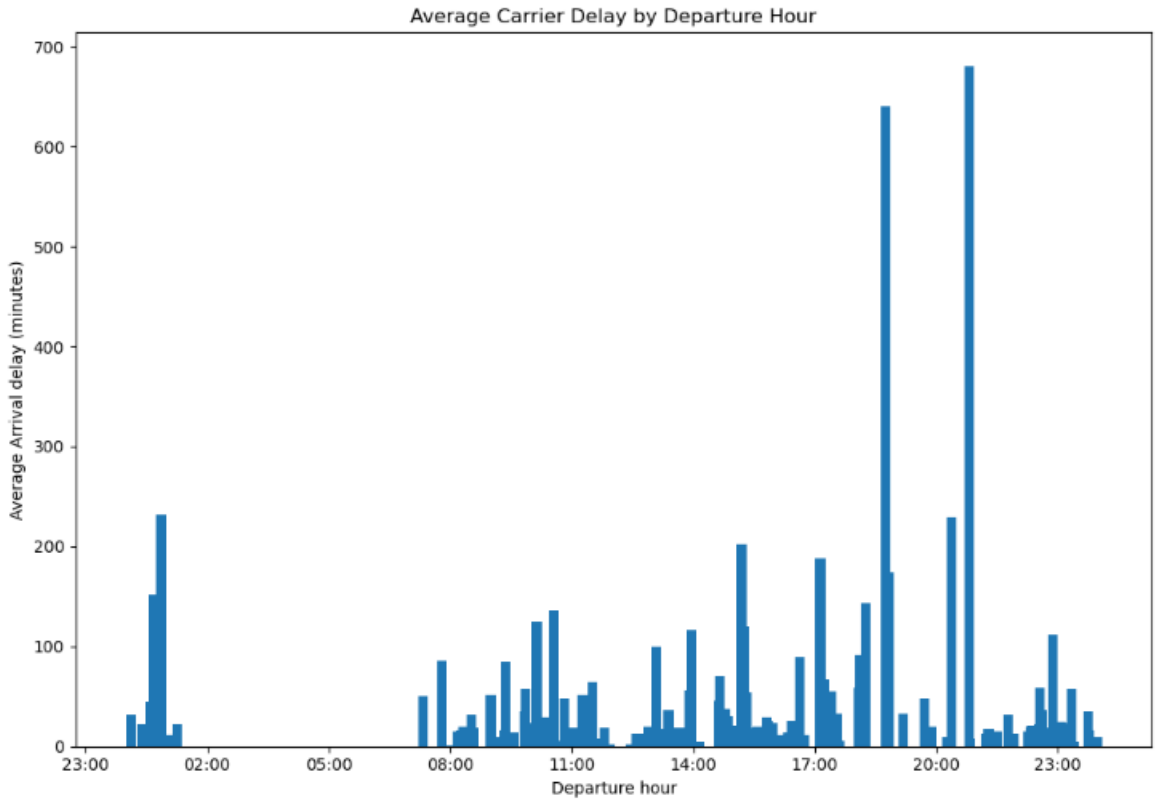

Figure 2: Average carrier delay by departure hour

Figure 3 shows the average delays related to weather for each month of the related years of data collection. The graph shows that December has the highest average weather-related delays, reaching more than 100 minutes (see figure 3). Since December is in the winter, it suggests that winter is the season most subject to flight delays. From January through November, flight delays were relatively low, with no other month approaching the magnitude of December's delays. Some months like March, June, August, and October did not witness flight delays. This could be attributed to stable weather conditions during the period, improved operational efficiency, lower traffic volume which reduces congestion, historical findings, and adjusted seasonal strategies based on seasonal demand. This data implies that winter weather conditions, such as snow and ice, can significantly disrupt flight schedules, leading to longer delays. Airlines may need to

adjust their schedules or have contingency plans ready for the winter months, especially December, to minimize the impact on passengers.

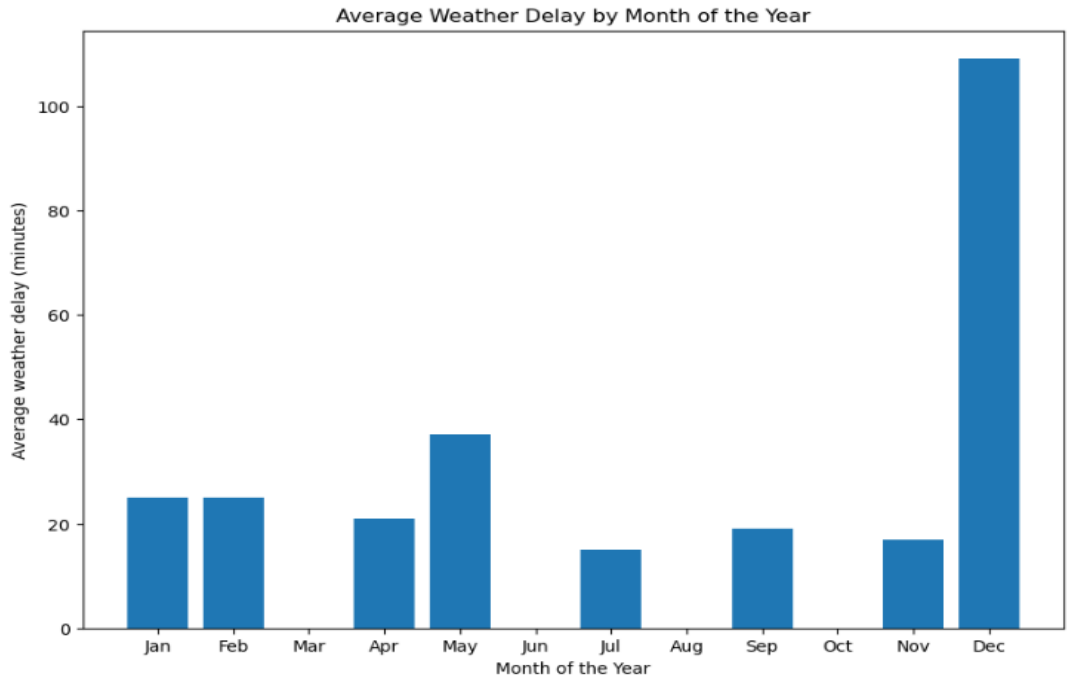

Figure 3: Monthly delays related to weather

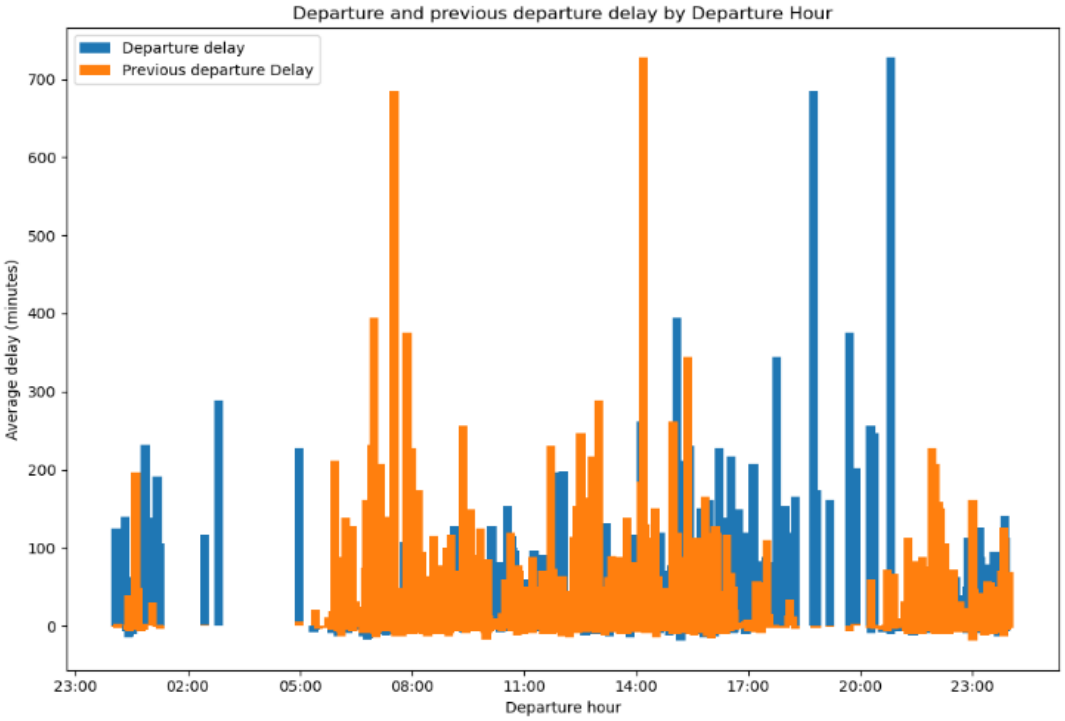

Figure 4: Departure and previous flight delays by departure hour

Figure 4 shows the distribution of departure and previous departure delays by departure hour. The chart shows a noticeable concentration of delays during the early hours of the morning, between 04:48 (early morning) and 19:12 (later in the evening). This suggests that planes are most subject to previous flight delay and departure delay during these hours. Departure delay - represented by blue dots, departure delays are spread across all departure hours but are particularly concentrated in the early morning hours. Previous departure delay - represented by orange dots, these delays are more scattered throughout the day but have a significant concentration in the early morning hours, with delays ranging from approximately 100 to over 500 minutes.

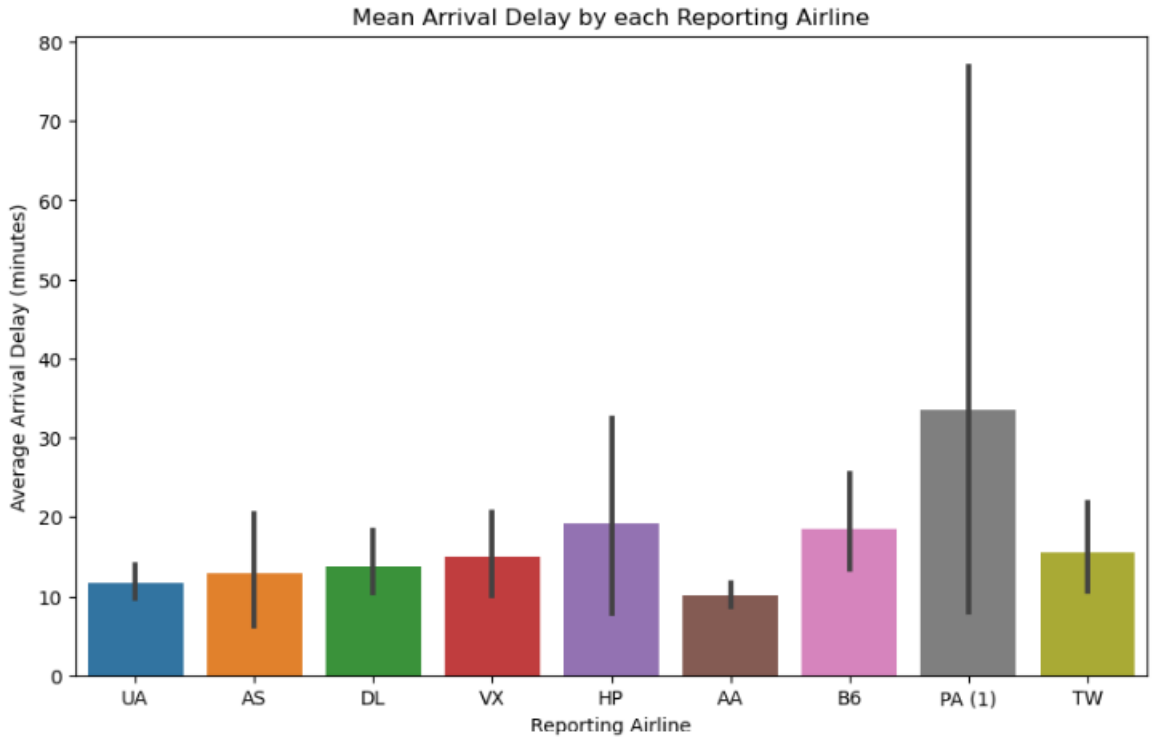

Figure 5: This shows the distribution of arrival delay in minutes by each airline that reported in the aircraft. As shown, PA (1) had the highest delays with HP and B6 slightly below it. AA had the lowest average delay time

Regression analysis

To achieve the purpose of fitting a regression model, and that month and day of the week cannot be treated as continuous variables, I decided to divide both month and day of the week into categories. The "month" column was divided into four quarter (Q1, Q2, Q3, and Q4). The arrival delay was predicted using the dummy variables. We under stand from few significant considerations, flights tend to be busiest on Fridays, Sundays, and Mondays. These days see higher demand due to business travel and leisure travelers starting or ending their trips. Conversely, Tuesdays and Wednesdays are generally less busy, as fewer people are flying mid-week, which can also make flights cheaper on these days. And we can consider the aircrafts to be moderately busy on Thursdays and Saturdays. Therefore, day of the week can be divided into busy, less busy and moderately busy. One of the categories will be a reference category to avoid multicollinearity in the model. The OLS results are as given in table 2.

Table 2: OLS estimation for predicting overall arrival delay time

OLS Regression Results

Dep. Variable:	ArrDelayMinutes	R-squared:	0.850
Model:	OLS	Adj. R-squared:	0.849
Method:	Least Squares	F-statistic:	1790.
Date:	Sat, 06 Apr 2024	Prob (F-statistic):	0.00
Time:	03:50:31	Log-Likelihood:	-11606.
No. Observations:	2855	AIC:	2.323e+04
Df Residuals:	2845	BIC:	2.329e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.0398	0.886	3.430	0.001	1.302	4.778
DepDelayMinutes	0.9632	0.009	112.686	0.000	0.946	0.980
DepTime	-0.0005	0.000	-1.120	0.263	-0.001	0.000
SecurityDelay	-0.1927	0.078	-2.460	0.014	-0.346	-0.039
LateAircraftDelay	-0.0173	0.022	-0.802	0.422	-0.060	0.025
less busy	-1.4480	0.636	-2.276	0.023	-2.696	-0.200
moderately busy	-0.8011	0.642	-1.247	0.213	-2.061	0.459
Q1	0.6416	0.752	0.854	0.393	-0.832	2.115
Q2	0.4615	0.755	0.611	0.541	-1.019	1.942
Q3	1.8346	0.741	2.476	0.013	0.382	3.288

Omnibus:	3098.945	Durbin-Watson:	1.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	762511.013
Skew:	4.972	Prob(JB):	0.00
Kurtosis:	82.442	Cond. No.	6.73e+03

As shown the factors in the model explain 84.9% of the variation in arrival delays. The departure hour does not seem to significantly explain the variations in arrival delays ($t = -1.12$, $p = 0.263$). There is a positive relationship between departure delays and overall arrival delay times. The implication is that overall arrival delay increases with departure delay and falls with it. In the sense that, flights that did not delay their departure (i.e., departure delay ≤ 0) arrived earlier or on time. There is a negligible negative effect of departure time on arrival delay times. Analyzing the effect of day of the week on arrival delays, notably flights are more likely to arrive early when the aircrafts are less busy than when they are busier (Sunday, Monday and Fridays). This could be justified by the negative coefficients of regression for the dummies of day of the week. Analyzing the seasons of the year, we could identify a positive trend in the seasons. The coefficients of Qi are positive, which suggests that relative to the reference category (Q4 = October, November and December), more delays were witnessed in the first three quarters (January – September) than the last quarter of each year. These insights suggest that flights were most likely to arrive earlier in the last quarter of the year than in the first nine months of the year. They also suggest that during the aircrafts' busy days (Sundays, Mondays, and Fridays) the flights are expected delay more than other days of the week. Delays due to security checks have an impact on arrival times, a negative coefficient ($p = 0.032$) means arrival delay times reduce with increase in security delays, which is an unexpected result. We expect that any delay along the flight route will influence the arrival time except the planes left earlier than they were initially scheduled. Late aircraft delays are insignificant in explaining the variations in arrival times ($p = 0.422$), and thus, there is no impact of late aircraft delay on arrival delays.

Correlation analysis

	WeatherDelay	DepTime	CarrierDelay	SecurityDelay
WeatherDelay	1.000000	0.044504	0.050490	-0.002434
DepTime	0.044504	1.000000	0.061806	-0.002541
CarrierDelay	0.050490	0.061806	1.000000	0.028782
SecurityDelay	-0.002434	-0.002541	0.028782	1.000000

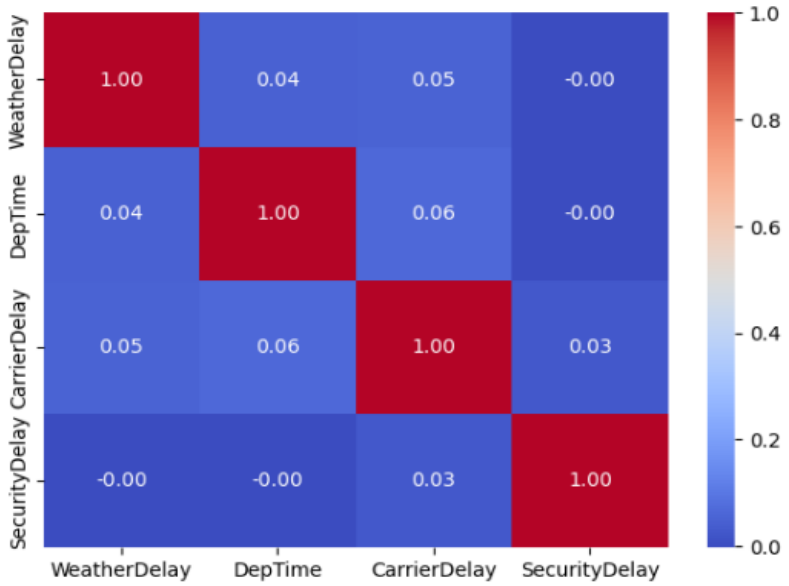

Figure 6: Correlation matrix

As shown in the correlation matrix and significantly heat map, there exists an overall weak relationship between the identified delay factors, with the strongest correlation recorded between departure time and the delays caused by carriers ($\rho = 0.0618$). All the correlations except between variables and themselves falls below 0.1 in absolute value. Therefore, there is no significant correlation between the flight delay factors.

Conclusion

Between 1990 and 2016, the airline sector saw two notable maxima in mean arrival delays, with delays exceeding 600 minutes and 700 minutes, respectively. Many causes, including weather, security, airplane capacity, technical problems, procedures, and propagated delays, were blamed for these delays. It is possible to find patterns and make changes to aircraft scheduling and management by analyzing these surges. With notable outliers experiencing longer delays, most planes have shorter average delays. Airlines and airports can use this information to optimize schedules and shorten delay times, as departure time is a critical component in predicting future delays. Winter appears to be the season most prone to aircraft delays, since December had the largest average weather-related delays, exceeding 100 minutes. There is a concentrated area of departure delays in the early morning hours with delays ranging from 100 minutes to 500 minutes. Based on departure hour, peak delays were witnessed at around midnight and another in the evening at 19:12. Most of the delays were witnessed between early morning (04:480 and later in evening (19:12). The monthly insights showed that the carriers witnessed the highest weather-related delays in December and the rest of the months had significantly lower delays. Also, the aircrafts were subject to previous flight delay and departure delay from 00:00 to 04:48. Flights are expected to delay their arrival more on Sundays, Mondays, and Fridays than other days.