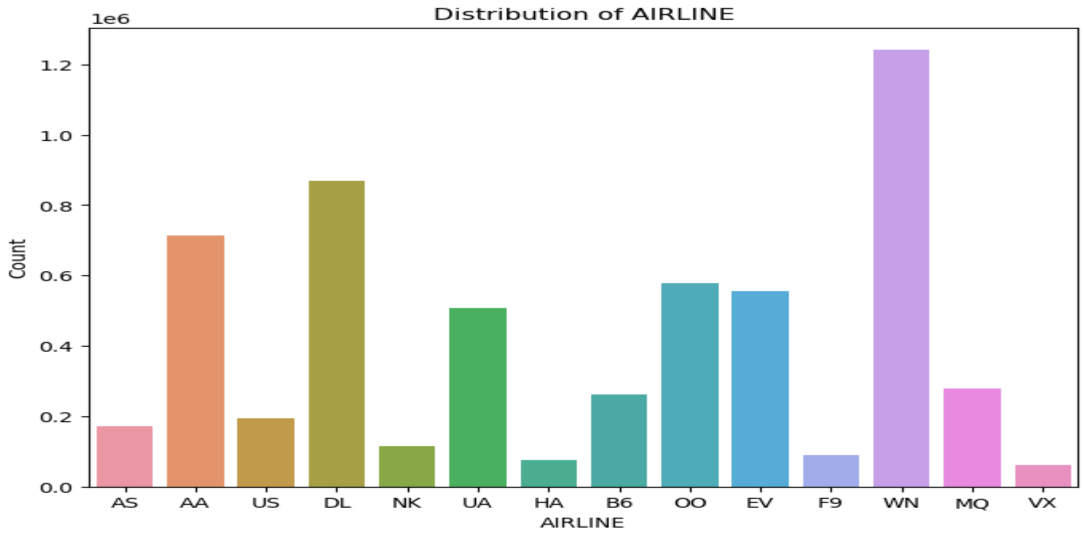
ASSIGNMENT 3

Flight dataset Analysis:

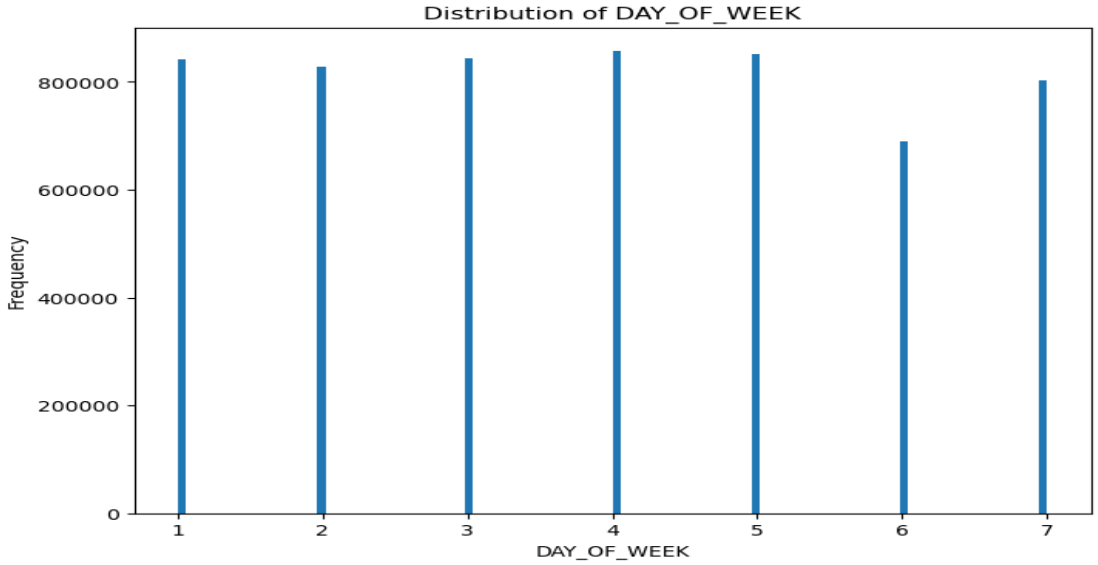
1.Describe any patterns or anomalies you observe from the univariate analysis

From Univariate analysis the distribution of airlines is as below and has less data related to some of the airlines as shown in below bar graph

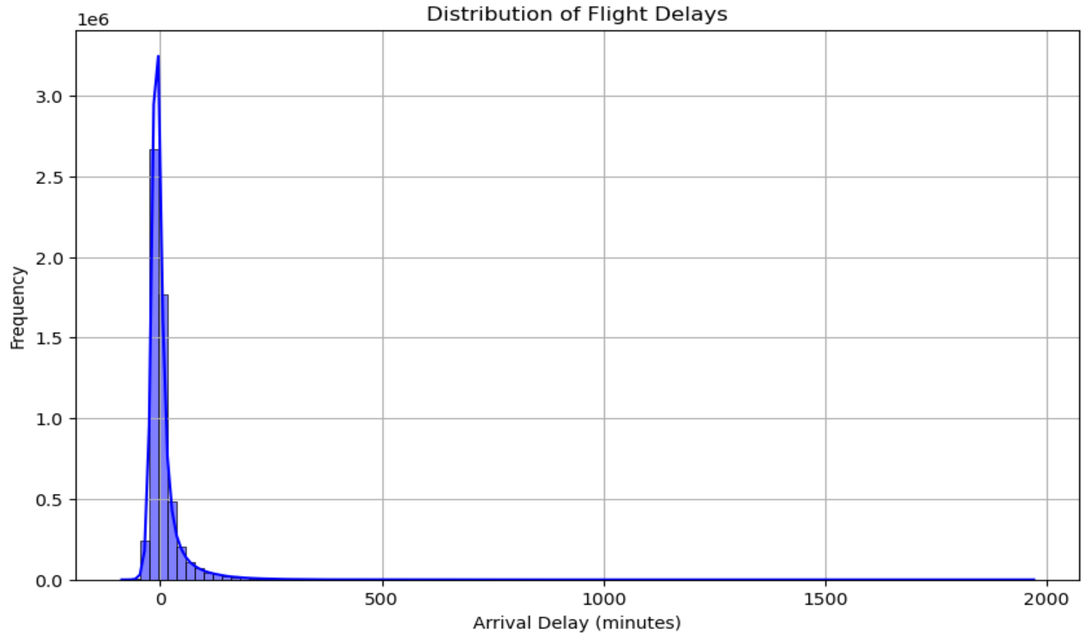


The above bar graph explains the distribution of airlines and it's frequency which is how often the same airline is used from the given flight data set.We can observe that the 'WN' airline has highest frequency which means most number of flights are operated by that airline and the next most operated airline is 'DL' , 'AA', 'OO' ,'EV'. Most of the airlines are operated in United states and the less frequent airlines are 'HA' followed by 'F9' ,'NK etc.

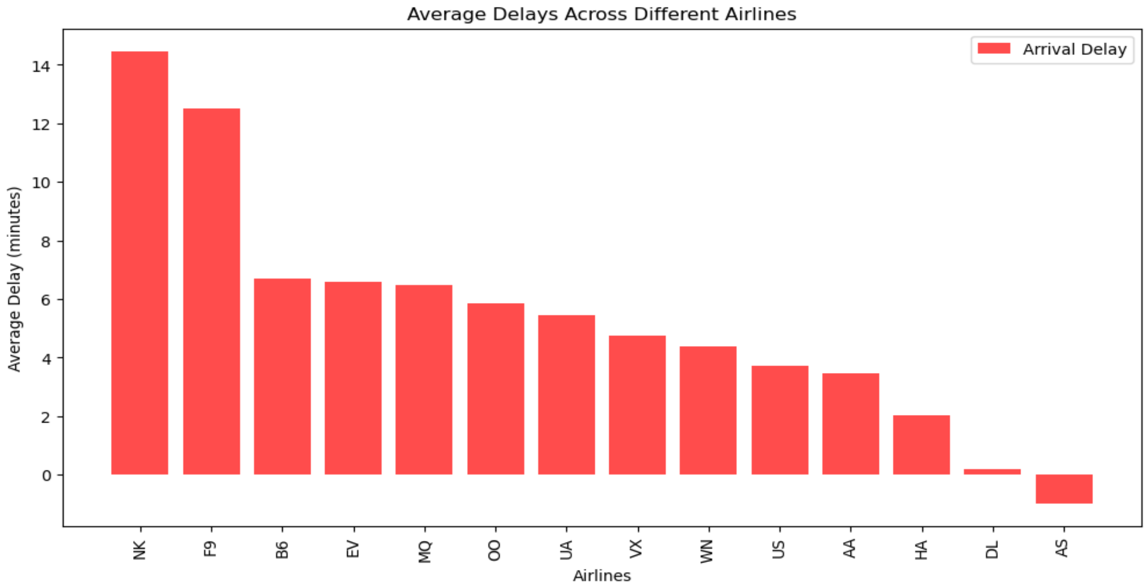
From this we can conclude that travellers prefer 'WN' , 'DL' airlines more compared to the other airlines.



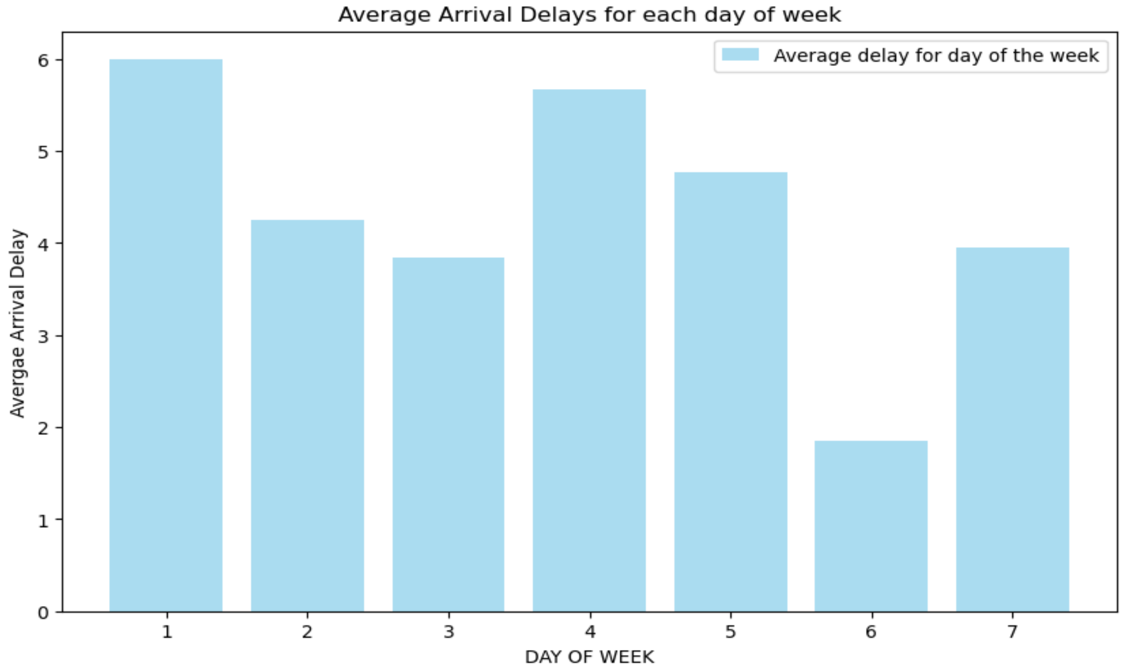
The above bar graph shows the frequency of distribution of days in the weeks.Here we can see that Day 6 i.e Saturday is least this indirectly says that more people may not often travel on Saturdays.



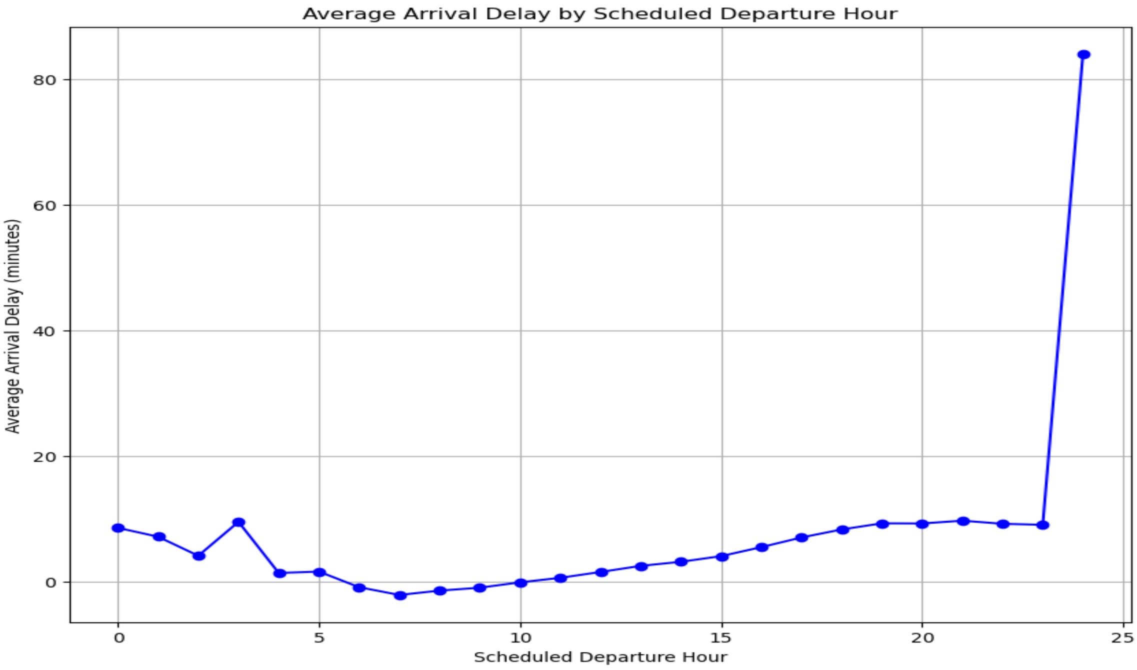
The above bar plot shows Distribution of flight delays in minutes and how frequently the delays are occurring. The delays which are negative i.e means early flights are more frequent as shown in plot and the



The above bar plot shows the average delays across different airlines. From the above graph we can observe that the "NK" airlines has the highest average delay and "F9" airlines is in second place with highest average delay whereas "B6","EV","MQ" airlines has almost same average delay. "DL" airlines is with least average delay and it seems mostly on time with minute delay and ‘AS’ airlines flights average delay is negative i.e it has no delay and the flights are arrived a head of the scheduled arrival

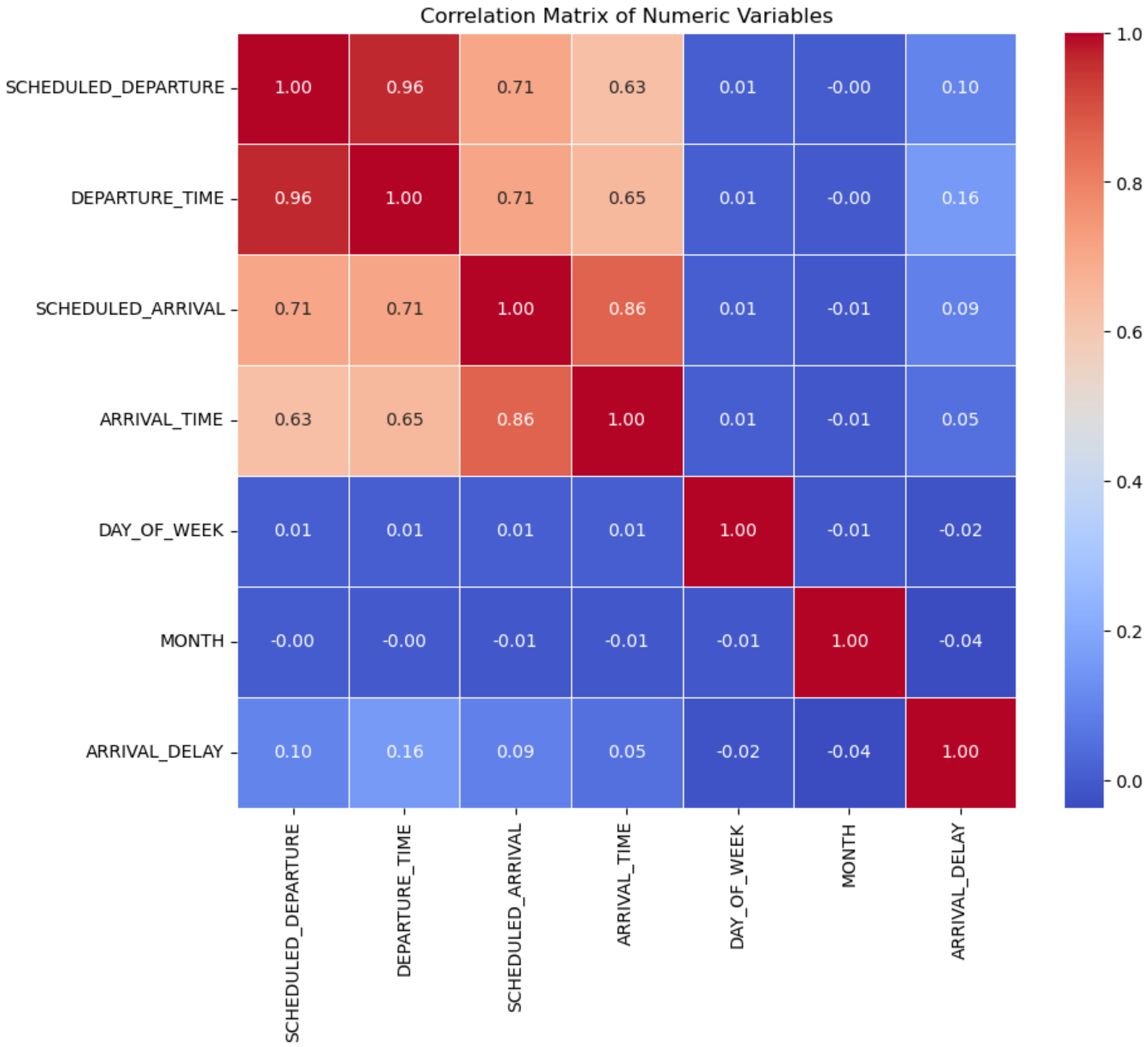


The above bar plot describes the Average Arrivals delays for each day in a week. From the above bar plot it is clear that the average delay of flight is high Monday which is the first day of the week and it is also high on Thursday and Friday of the week. The average arrival delay is less on the Saturdays and it would be the best day to travel to reach the destinations on time.



A line graph displaying the average arrival delay by scheduled departure hour . The average arrival delay rises with the scheduled departure hour, as the line graph illustrates.Accordingly, flights with later departure times have a higher chance of experiencing delays than those with earlier departure times. There are several reasons why this can be the case. One reason for delays could be an increase in flight traffic later in the day. There could also be an increase in weather delays later in the day. Furthermore, in an effort to prevent passenger bumping, airlines might be more inclined to postpone flights later in the day.

.

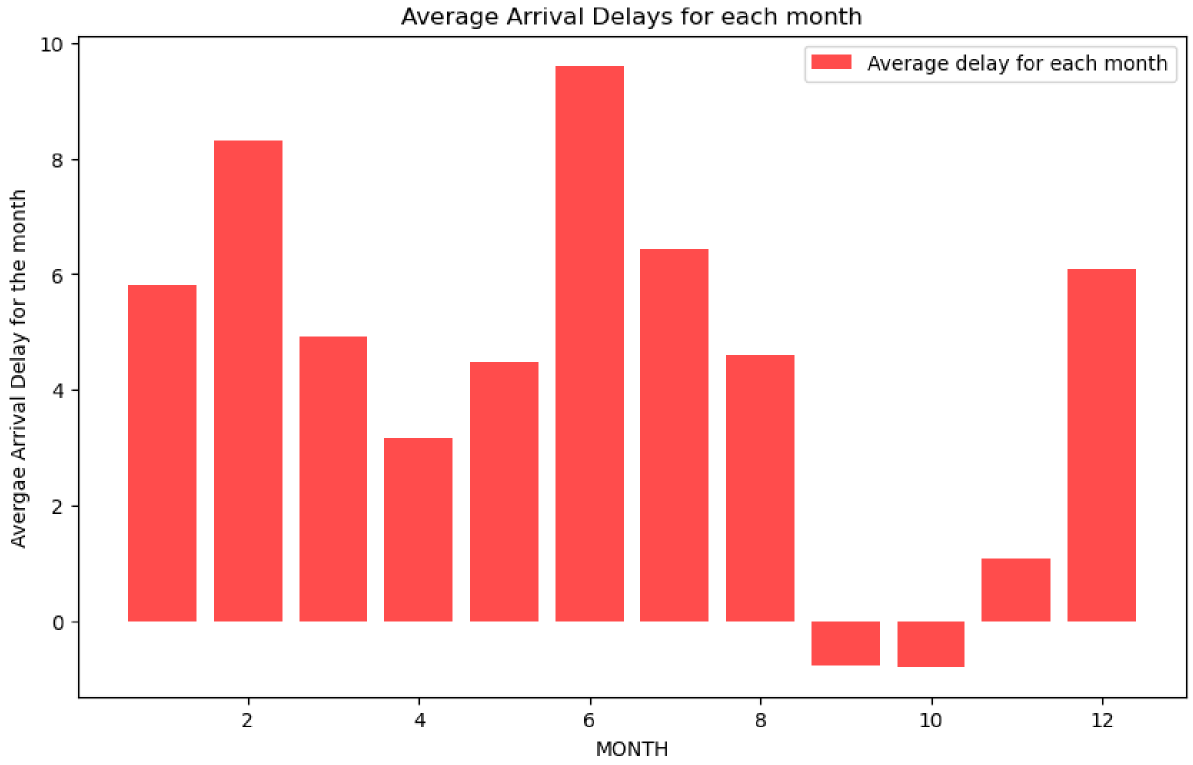


	SCHEDULED_DEPARTURE	DEPARTURE_TIME	SCHEDULED_ARRIVAL	ARRIVAL_TIME	DAY_OF_WEEK	MONTH	ARRIVAL_DELAY
SCHEDULED_DEPARTURE	1.00	0.96	0.71	0.63	0.01	-0.00	0.10
DEPARTURE_TIME	0.96	1.00	0.71	0.65	0.01	-0.00	0.16
SCHEDULED_ARRIVAL	0.71	0.71	1.00	0.86	0.01	-0.01	0.09
ARRIVAL_TIME	0.63	0.65	0.86	1.00	0.01	-0.01	0.05
DAY_OF_WEEK	0.01	0.01	0.01	0.01	1.00	-0.01	-0.02
MONTH	-0.00	-0.00	-0.01	-0.01	-0.01	1.00	-0.04
ARRIVAL_DELAY	0.10	0.16	0.09	0.05	-0.02	-0.04	1.00

The above images shows the correlation between various numeric values in the given dataset.The Arrival delay has not much correlation with the other numeric values in the given dataset and the arrival time and schedule time has more correlation.

The correlation matrix in the image shows that there are strong positive correlations between the following pairs of variables:
SCHEDULED_DEPARTURE and DEPARTURE_TIME
SCHEDULED_ARRIVAL and ARRIVAL_TIME
DEPARTURE_TIME and ARRIVAL_TIME

The correlation matrix also shows that there are weak negative correlations between the following pairs of variables:
DAY_OF_WEEK and ARRIVAL_DELAY
MONTH and ARRIVAL_DELAY



The above bar plot shows the Average arrivals delays for each month. From the above plot it is clear that in the month of June the average arrival delay is highest and the second highest is February month i.e more people are travelling in the Summer season which is in the month of June, July. February is the second highest and the people are travelling very less in the months of September and October as the average delay is very less in this months.

6.

Recommendations for airlines to reduce delays

Airlines may want to adjust their flight schedules in order to reduce the number of flights that are scheduled to depart during the peak delay times in the weak and also in the months. For example, airlines could schedule more flights to depart earlier in the day or later in the evening and also should schedule more flights in the months of February,June, July.

Airlines should also use the above data to estimate the delays ahead based on climatic changes and should do advanced weather forecasting and let the passengers now if there are any possibilities for delays so that the passengers can adjust their bookings.

Airlines should carefully manage overbooking to reduce the chances of involuntary denied boardings. Efficient passenger management can help prevent situations that lead to delays.

Airports may want to take steps to reduce congestion during the peak delay times.For example, airports could open more check-in counters and security lanes during this time period.

Advices to passengers:

Based on the above analysis passengers who are booking flights may want to consider the scheduled departure hour when choosing a flight. If possible, it is best to choose a flight that is scheduled to depart earlier in the day, as these flights are less likely to be delayed.

Passengers should also choose the best airlines which maintains the less average arrival delay times as seen in the above bar graphs.

In order to avoid the delays passengers should also choose the right time to travel i.e they should the Saturdays in the week to travel as the average delays are less in this days and also passengers should plan their travel on non busy months like September and October to avoid the delays.

Large, established airlines often have more resources and a more extensive network, which can provide more flexibility to manage delays and rebook passengers. So major airlines are recommended .

7.
Metrics for Linear Regression model:

```
Mean Absolute Error: 20.969851843347783
Mean Squared Error: 1510.6869777705226
Root Mean Squared Error: 38.86755687936306
R-squared: 0.01586705331668603
```

Name: Mannem Harsha Vardhan
Blazer id: hmannem
Course: EE590 Basics of Data Analytics and Machine Learning