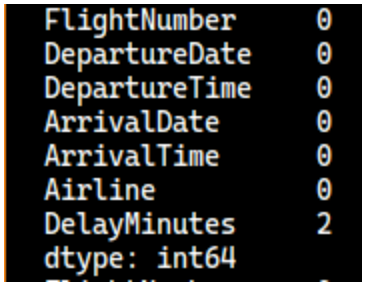
AnalysisReport-2347107

Aryan Majhi 2347107

AKASA - Task 1 (Python + SQL)

Summary of data cleaning and normalization steps.

- There were 2 missing values in the column DelayMinutes which was filled with mean value of that column



- The datatypes of columns with dates and times were changed accordingly with the help of pd.to_datetime(), datetime.strptime() and datetime.strftime() functions.

```
# Converting to Date
df['DepartureDate']=pd.to_datetime(df['DepartureDate'],format="%m/%d/%Y")
df['ArrivalDate']=pd.to_datetime(df['ArrivalDate'],format="%m/%d/%Y")
# print(">\n",df.dtypes)
# print(df.info())


# Converting to Time
df['DepartureTime'] = df['DepartureTime'].apply(convert_24hr)
df['ArrivalTime'] = df['ArrivalTime'].apply(convert_24hr)
```

- There were time errors which means that the arrival time of the flights were after the flights' departed which is physically impossible. Hence, the columns were swapped for those which had time errors and only one row was correct in the whole

dataset. There were conditions where we checked, that if the dates are same and the arrival time is greater than departure time then swap the values.

```
for index,row in df.iterrows():
    if row['DepartureDate']==row['ArrivalDate']: # If the depa
        if row['ArrivalTime'] > row['DepartureTime']: # If the a
            # Swap the departure and arrival time for that index
            temp=row['DepartureTime']
            df.at[index, 'DepartureTime'] = row['ArrivalTime']
            df.at[index, 'ArrivalTime'] = temp
    elif row['DepartureDate'] < row['ArrivalDate']: # If the d
        # Swap the departure and arrival date for that index
        temp=row['ArrivalDate']
        df.at[index, 'ArrivalDate'] = row['DepartureDate']
        df.at[index, 'DepartureDate'] = temp
```

- The time format used is 24-hour format, and the date is in YYYY-MM-DD format.

	FlightNumber	DepartureDate	DepartureTime	ArrivalDate	ArrivalTime	Airline	DelayMinutes
0	AA1234	2023-09-01	08:30	2023-09-01	10:45	American Airlines	15.0
1	DL5678	2023-09-01	13:15	2023-09-01	15:30	Delta	5.0
2	UA9101	2023-09-01	17:00	2023-09-01	19:15	United Airlines	25.0
3	AA1234	2023-09-01	08:30	2023-09-01	22:45	American Airlines	30.0
4	DL5678	2023-09-02	14:00	2023-09-02	16:10	Delta	25.0
5	UA9101	2023-09-02	17:00	2023-09-02	19:15	United Airlines	20.0
6	AA1234	2023-09-02	20:30	2023-09-03	10:45	American Airlines	60.0
7	DL5678	2023-09-03	13:00	2023-09-03	15:30	Delta	10.0
8	UA9101	2023-09-03	15:00	2023-09-03	17:20	United Airlines	25.0
9	AA1234	2023-09-03	08:30	2023-09-03	10:00	American Airlines	15.0
10	DL5678	2023-09-04	12:30	2023-09-04	14:40	Delta	25.0
11	UA9101	2023-09-04	19:00	2023-09-04	21:15	United Airlines	45.0

- A new column was formed to store the flight duration of each flight which had two different conditions - Same day and Different Days. For that the devised methods are :
    - For the same date, duration of the flight is 24 hours - (departure time of the flight - arrival time of the flight), since departure time is greater than arrival time. Since, departure time of the flight - arrival time of the flight is the halting time i.e., the time spent in the airport.
    - For the flights in different days, 48 hours - (duration of the flight is (24 hours - arrival time) + the departure time of the next day). Since, duration of the flight is (24 hours - arrival time) + the departure time of the next day is the halting time i.e., the time spent in the airport.

```
for index,row in df.iterrows():
  # For the same date, duration of the flight is (24 hours - (departure time of the flight - arrival time)) of
  if row['DepartureDate'] == row['ArrivalDate']:
    departure_seconds = pd.to_timedelta(row['DepartureTime']+":00").total_seconds()
    arrival_seconds = pd.to_timedelta(row['ArrivalTime']+":00").total_seconds()

    # Storing the data in a new column
    df.at[index, 'FlightDuration(in Hrs)'] = round(((86400-(departure_seconds - arrival_seconds)) / 3600),2)
  else:
    # For the flights not in the same day, duration of the flight is (48 hours - (24 hours - arrival time) + t
    arrival_seconds= (pd.Timedelta('1 day')-pd.to_timedelta(row['ArrivalTime']+":00")).total_seconds()
    departure_seconds = pd.to_timedelta(row['DepartureTime']+":00").total_seconds()

    # Storing the data in a new column
    df.at[index, 'FlightDuration(in Hrs)'] = round(((172800-(departure_seconds + arrival_seconds)) / 3600),2)
```

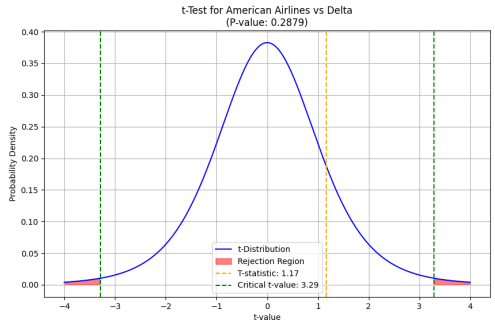
- The flight duration is stored as hours so 2.25 hours would mean 2 hours and 15 minutes.

	FlightNumber	DepartureDate	DepartureTime	ArrivalDate	ArrivalTime	Airline	DelayMinutes	FlightDuration(in Hrs)
9	AA1234	2023-09-03	10:00	2023-09-03	08:30	American Airlines	15.0	22.50
0	AA1234	2023-09-01	10:45	2023-09-01	08:30	American Airlines	15.0	21.75
10	DL5678	2023-09-04	14:40	2023-09-04	12:30	Delta	25.0	21.83
1	DL5678	2023-09-01	15:30	2023-09-01	13:15	Delta	5.0	21.75
7	DL5678	2023-09-03	15:30	2023-09-03	13:00	Delta	10.0	21.50
4	DL5678	2023-09-02	16:10	2023-09-02	14:00	Delta	25.0	21.83
8	UA9101	2023-09-03	17:20	2023-09-03	15:00	United Airlines	25.0	21.67
2	UA9101	2023-09-01	19:15	2023-09-01	17:00	United Airlines	25.0	21.75
5	UA9101	2023-09-02	19:15	2023-09-02	17:00	United Airlines	20.0	21.75
6	AA1234	2023-09-03	20:30	2023-09-02	10:45	American Airlines	60.0	14.25
11	UA9101	2023-09-04	21:15	2023-09-04	19:00	United Airlines	45.0	21.75
3	AA1234	2023-09-01	22:45	2023-09-01	08:30	American Airlines	30.0	9.75

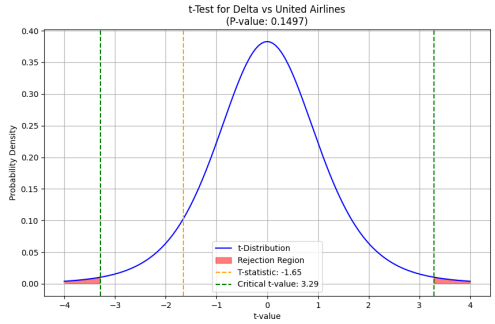
Insights derived from the data analysis.

- The airline delta has the least delays suggesting that the airline is efficient.
- The airline American Airlines has the highest delay which suggests that they are not at all efficient.
- The flights with departure time after 19:15 are prone to long delays anywhere between 30-60 mins.
- The flights in the morning are less prone to delays and even if there is a delay it is significantly lesser than evening flights. The delays can range between somewhere between 0-20 mins.

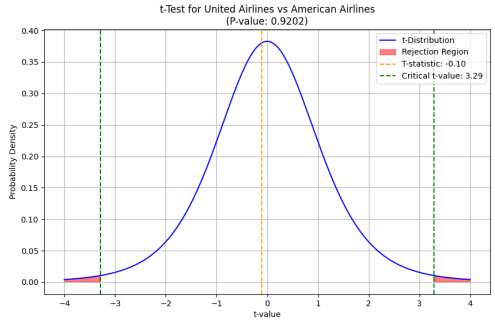
- Comparing American Airlines and Delta:
  T-statistic: 1.165997668006996, P-value: 0.2878631950549536
  Conclusion : Fail to reject the null hypothesis.



- Comparing Delta and United Airlines:
  T-statistic: -1.651445647689541, P-value: 0.14973733521583907
  Conclusion : Fail to reject the null hypothesis.



- Comparing United Airlines and American Airlines:
  T-statistic: -0.1044465935734187, P-value: 0.9202189075644182
  Conclusion : Fail to reject the null hypothesis.



- There is no significant difference in the mean delay times between the airlines, as found from the t-test for the airlines.
- Correlation between Departure Time (in minutes) and Delay Minutes: 0.6248542726807503. This means that the correlation is strongly positive and correlation values are between -1 and 1, hence 0.62 is pretty good correlation.
- The flights departing later time of the day has higher delays is suggested by the correlation and is evident by the graph as well.

Visualizations (e.g., bar charts, histograms) illustrating the key findings.

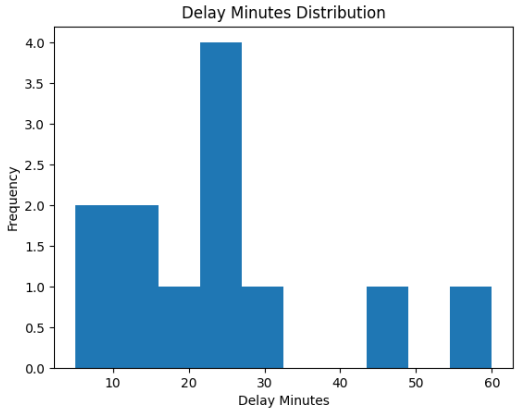
This plot shows the delay minutes distribution.

The bar plot shows how many times the delay of a specific time has occurred.

It is observed that the highest frequency of delay in minutes are somewhere between 20 - 30 mins.

The second highest is the range of 0 - 15 mins

and the lowest frequencies of delays are other than these ones : 15-20 mins, 45-50 mins, and 55-60 mins.

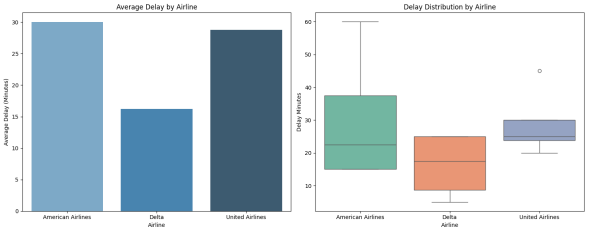


The left plot is for average delay of flights according to the flight number

American Airlines has the highest average delay at around 30 mins, followed by United Airlines with average delay at around 28 mins.

The right plot shows the delay distribution of the airlines in a box-plot to understand the distribution of the data points.

The median of American Airlines is somewhere in the early 20s, Delta has median below 20 and United Airlines have the median slightly above 20.

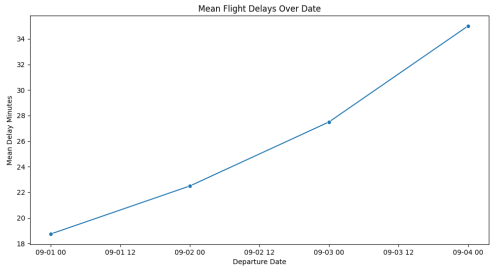


There is a single outlier in the United Airlines which is at around early 40s.

This graph depicts the mean flight delay for the departure dates.

We can see that the delay increases for the days coming.

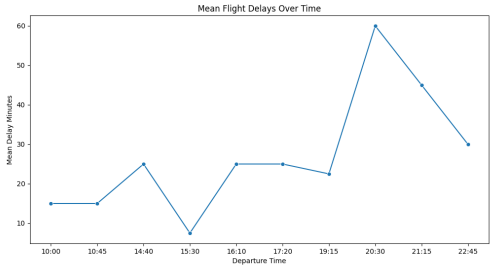
The 3 days are divided into 2 halves so that we can see if the delay changes in the afternoon. But the mean delay keeps increasing.



The plot depicts the mean delay over the departure time.

We can find a pattern here that suggests that flights in the evening will have significantly more delays than in the mornings.

Whereas the flights in the morning have very less delay.

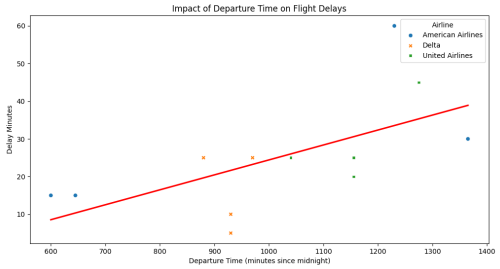


This plot is based on the correlation of the departure time and delay minutes.

The red line shows the trend also known as the trend line

It is a positively correlated graph.

Delta have departure time in the afternoon and hence the delay is very less. Suggesting that time to be the optimal for traveling for passengers.



The paired t-test for United and American Airlines

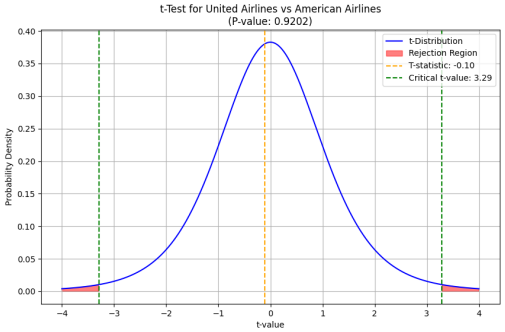
The curve is used for determining if the null hypothesis should be rejected or accepted.

The green lines are the critical values beyond which we reject the null hypothesis.

The yellow line is for showing the t-statistic value calculated for the paired t-test.

The red area is the rejection region.

Since the t-statistic value (-0.10) is in the acceptance region, we accepted the null hypothesis for United and American Airlines



The paired t-test for United and American Airlines

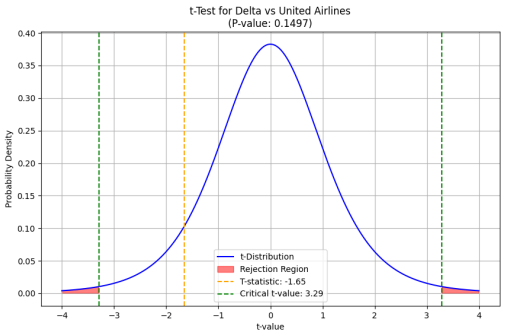
The curve is used for determining if the null hypothesis should be rejected or accepted.

The green lines are the critical values beyond which we reject the null hypothesis.

The yellow line is for showing the t-statistic value calculated for the paired t-test.

The red area is the rejection region.

Since the t-statistic value (-1.65) is in the acceptance region, we accepted the null hypothesis for Delta and United Airlines



The paired t-test for United and American Airlines

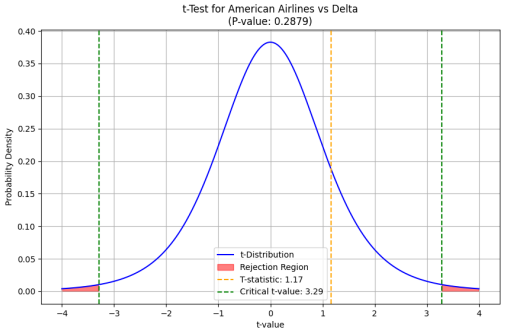
The curve is used for determining if the null hypothesis should be rejected or accepted.

The green lines are the critical values beyond which we reject the null hypothesis.

The yellow line is for showing the t-statistic value calculated for the paired t-test.

The red area is the rejection region.

Since the t-statistic value (1.17) is in the acceptance region, we accepted the null hypothesis for American Airlines and Delta.



Recommendations based on the analysis.

- The delays on the evening should be minimized by American Airlines and United Airlines.
- Implementation of predictive model should be used to predict delays and counter plan for those delays.
- Airlines with high delays should review their schedule and make appropriate changes.
- There should be higher number of flights assigned in the morning to avoid delays such that it doesn’t bottleneck the operations. The numbers of flights should be changed in such a manner that the flights in the morning operates with minimal delays and in that way the number of flights delayed will decrease.