AnalysisReport-2347107

Aryan Majhi 2347107

AKASA - Task 1 (Python + SQL)

Github repo : https://github.com/aryanmajhi75/AKASA

Native System Configuration

- AMD Ryzen 5 7000 series
- 16 GB RAM
- Ubuntu 24.04 LTS

Dependencies Used

- Python libraries used
    - mysql-connector
    - pandas
    - datetime
    - matplotlib.pyplot
    - numpy
    - seaborn
    - scripy.stats
- Docker image for mysql 8.1
- Python Interpreter version 3.12.3

Setup Instructions

> Note: the instructions are based on my own system i.e., Ubuntu Linux 24.04 LTS, if you are using any other system then follow the instructions for that OS.

- Install docker using cli, for ubuntu it works like this:

```
sudo apt-get update
sudo apt-get install ./docker-desktop-<arch>.deb
```

For installation guide for other OS, follow the documentation :
https://docs.docker.com/engine/install/

- Pull mysql 8.1 image to docker

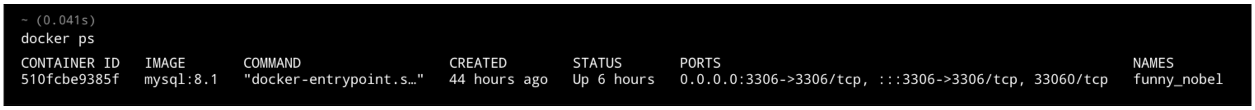
```
docker pull mysql:8.1
```

- Run docker image mysql 8.1 with credentials and port number

```
docker volume create akasa

docker volume inspect akasa

docker run -d \
-e MYSQL_ROOT_PASSWORD=1234 \
-e MYSQL_PASSWORD=1234 \
-e MYSQL_USER=aryan \
-e MYSQL_DATABASE=aviation_data \
-v mysql_volume:/var/lib/mysql \
-p 3306:3306 \
mysql:8.1
```

- Find the name of the image (last column), for my case it is called **funny_nobel**

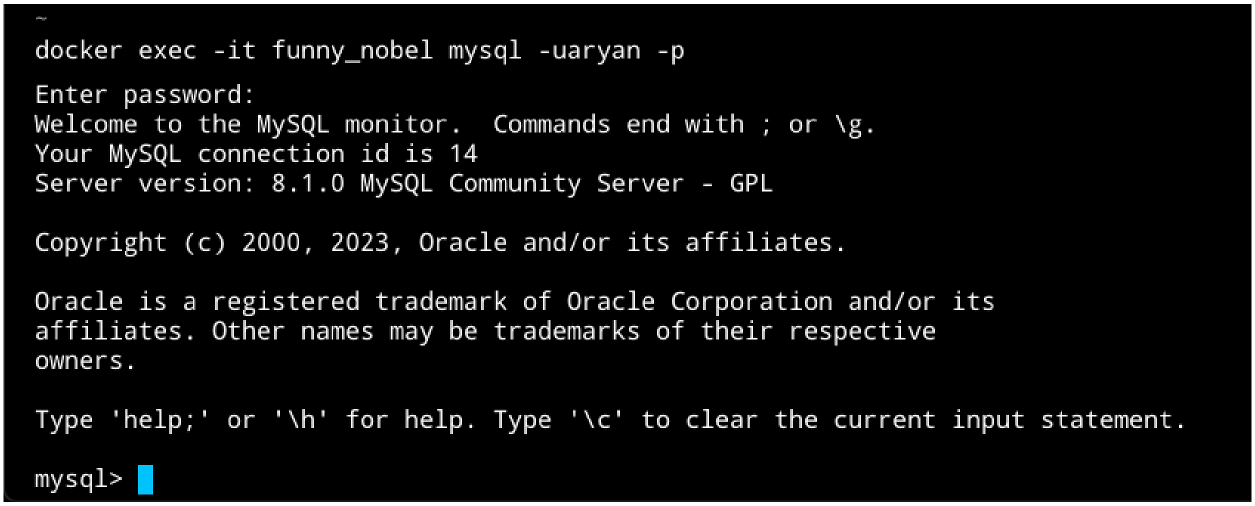
```
docker ps
```



```
~ (0.041s)
docker ps
CONTAINER ID   IMAGE       COMMAND                  CREATED        STATUS        PORTS                                                   NAMES
510fcbe9385f   mysql:8.1   "docker-entrypoint.s…"   44 hours ago   Up 6 hours    0.0.0.0:3306->3306/tcp, :::3306->3306/tcp, 33060/tcp   funny_nobel
```

- Execute the docker image with the name of the image

```
docker exec -it funny_nobel mysql -uaryan -p
```



```
~
docker exec -it funny_nobel mysql -uaryan -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 8.1.0 MySQL Community Server - GPL

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

- If the docker image was stopped, we need to start the image again for mysql to run.

```
docker start <image-name>
docker exec -it funny_nobel mysql -uaryan -p
```

- After mysql is setup, we need to make a table FlightSchedule and then fill the table.

```
// Create the table FlightSchedule
CREATE TABLE FlightSchedule (
    FlightNumber VARCHAR(10),
    DepartureDate VARCHAR(10),
    DepartureTime VARCHAR(8),
    ArrivalDate VARCHAR(10),
    ArrivalTime VARCHAR(8),
    Airline VARCHAR(50),
    DelayMinutes INT
```

```
);

// Insert data into the table
INSERT INTO FlightSchedule (FlightNumber, DepartureDate, Departu
VALUES
('AA1234', '09/01/2023', '08:30 AM', '09/01/2023', '10:45 AM',
('DL5678', '09/01/2023', '01:15 PM', '09/01/2023', '03:30 PM',
('UA9101', '09/01/2023', '05:00 PM', '09/01/2023', '07:15 PM',
('AA1234', '09/01/2023', '08:30 AM', '09/01/2023', '10:45 PM',
('DL5678', '09/02/2023', '02:00 PM', '09/02/2023', '04:10 PM',
('UA9101', '09/02/2023', '05:00 PM', '09/02/2023', '07:15 PM',
('AA1234', '09/02/2023', '08:30 PM', '09/03/2023', '10:45 AM',
('DL5678', '09/03/2023', '01:00 PM', '09/03/2023', '03:30 PM',
('UA9101', '09/03/2023', '03:00 PM', '09/03/2023', '05:20 PM',
('AA1234', '09/03/2023', '08:30 AM', '09/03/2023', '10:00 AM',
('DL5678', '09/04/2023', '12:30 PM', '09/04/2023', '02:40 PM',
('UA9101', '09/04/2023', '07:00 PM', '09/04/2023', '09:15 PM',
```

Since I’m using Ubuntu, the OS doesn’t recommend to install python packages to root, so I have an environment where the packages are saved

```
mkdir -p ~/.venvs
python3 -m venv ~/.venvs/<some-name>
~/.venvs/<some-name>/bin/python -m pip install <package-name>
```

The whole python script is running in a venv called **newenv**

```
python3 -m venv newenv
source newenv/bin/activate
```

To install any packages, we need to install it in the **newenv and in the venvs**

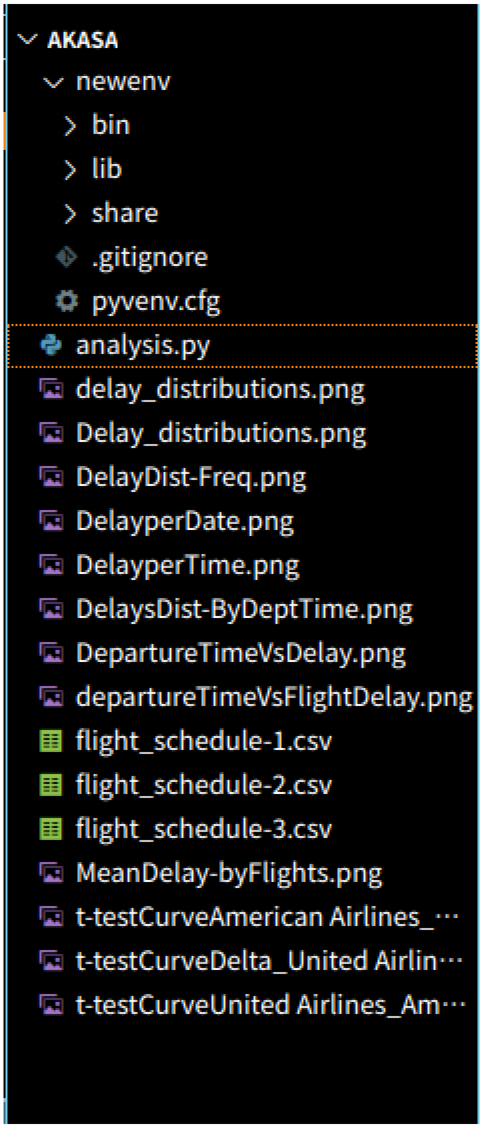
```
~/.venvs/mysql/bin/python -m pip install <package-name>
pip install <package-name>
```

After the installations are done, we need to run the python script named **analytics.py**

```
python3 analytics.py
```

After running all the graphs are stored in the same folder as png files and csv files for checking the changes in the dataset after every section - 1,2,3

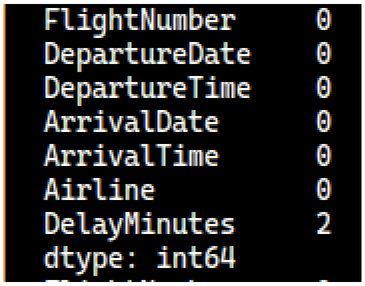
The folder structure will be as follows:



```
AKASA
  newenv
    bin
    lib
    share
    .gitignore
    pyvenv.cfg
  analysis.py
  delay_distributions.png
  Delay_distributions.png
  DelayDist-Freq.png
  DelayperDate.png
  DelayperTime.png
  DelaysDist-ByDeptTime.png
  DepartureTimeVsDelay.png
  departureTimeVsFlightDelay.png
  flight_schedule-1.csv
  flight_schedule-2.csv
  flight_schedule-3.csv
  MeanDelay-byFlights.png
  t-testCurveAmerican Airlines_…
  t-testCurveDelta_United Airlin…
  t-testCurveUnited Airlines_Am…
```

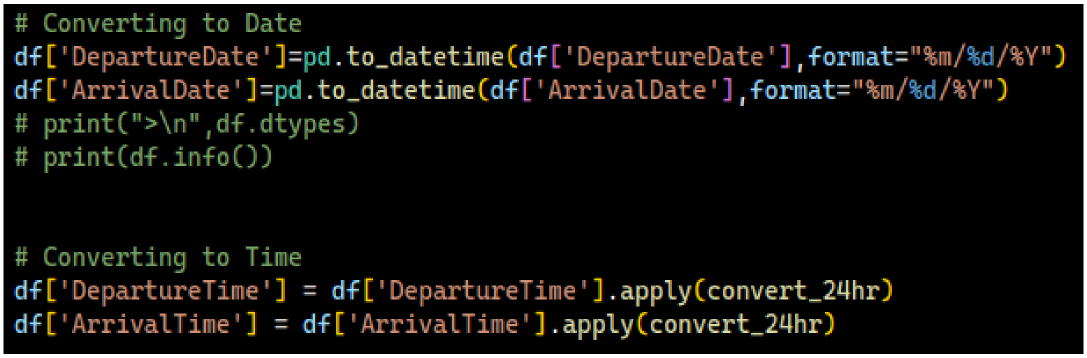
Summary of data cleaning and normalization steps.

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



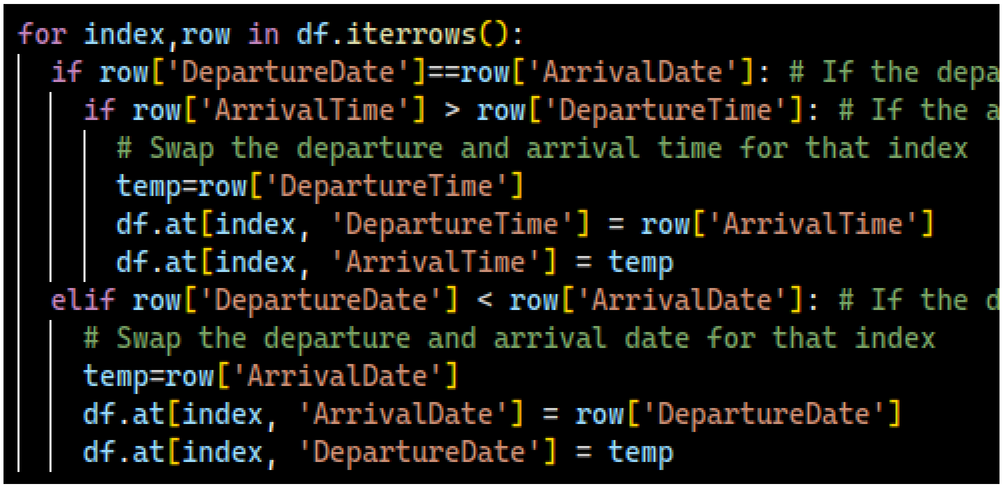
```
FlightNumber     0
DepartureDate    0
DepartureTime    0
ArrivalDate      0
ArrivalTime      0
Airline          0
DelayMinutes     2
dtype: int64
```

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



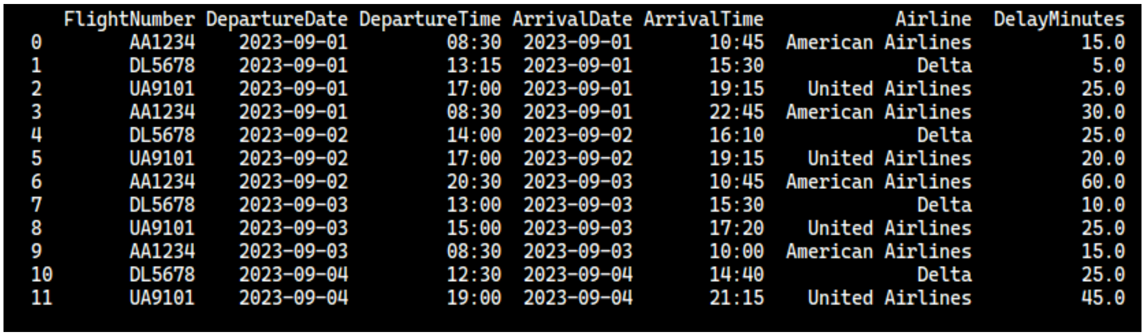
```
# Converting to Date
df['DepartureDate']=pd.to_datetime(df['DepartureDate'],format="%m/%d/%Y")
df['ArrivalDate']=pd.to_datetime(df['ArrivalDate'],format="%m/%d/%Y")
# print(">\n",df.dtypes)
# print(df.info())


# 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.



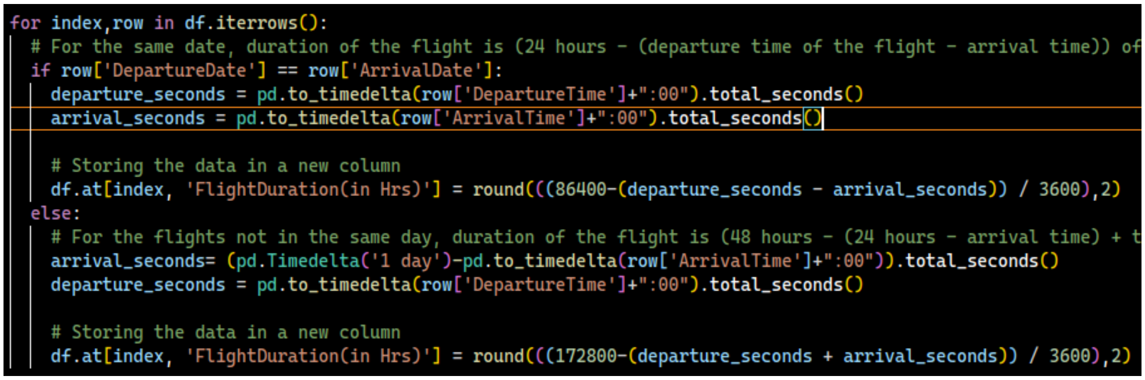
```
for index,row in df.iterrows():
    if row['DepartureDate']==row['ArrivalDate']: # If the depa
        if row['ArrivalTime'] > row['DepartureTime']: # If the a
            # Swap the departure and arrival time for that index
            temp=row['DepartureTime']
            df.at[index, 'DepartureTime']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



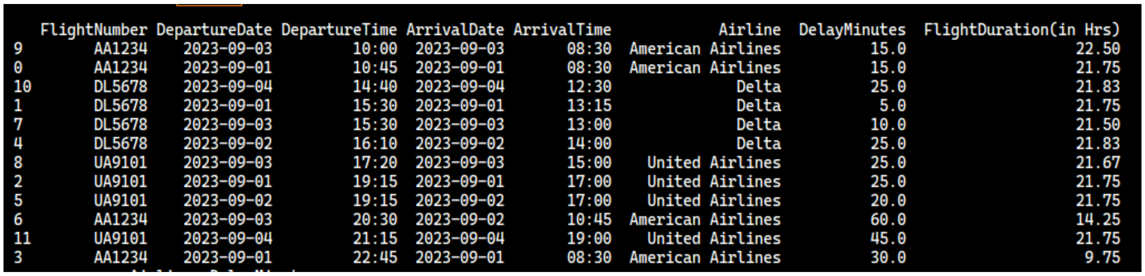
```
   FlightNumber DepartureDate DepartureTime ArrivalDate ArrivalTime            Airline  DelayMinutes
0        AA1234    2023-09-01         08:30  2023-09-01       10:45  American Airlines          15.0
1        DL5678    2023-09-01         13:15  2023-09-01       15:30              Delta           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.

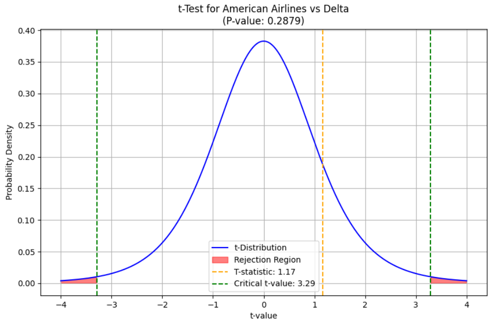


```
   FlightNumber DepartureDate DepartureTime ArrivalDate ArrivalTime           Airline  DelayMinutes  FlightDuration(in Hrs)
9        AA1234    2023-09-03         10:00  2023-09-03       08:30  American Airlines          15.0                   22.50
0        AA1234    2023-09-01         10:45  2023-09-01       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
```

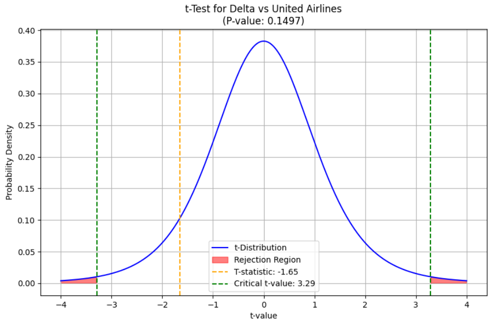
Insights derived from the data analysis.

- 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.

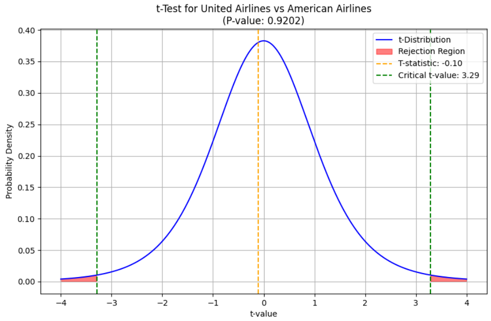
- Comparing American Airlines and Delta:
  T-statistic: 1.165997668006996, P-value: 0.2878631950549536
  Conclusion : Fail to reject the null hypothesis.



- 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.



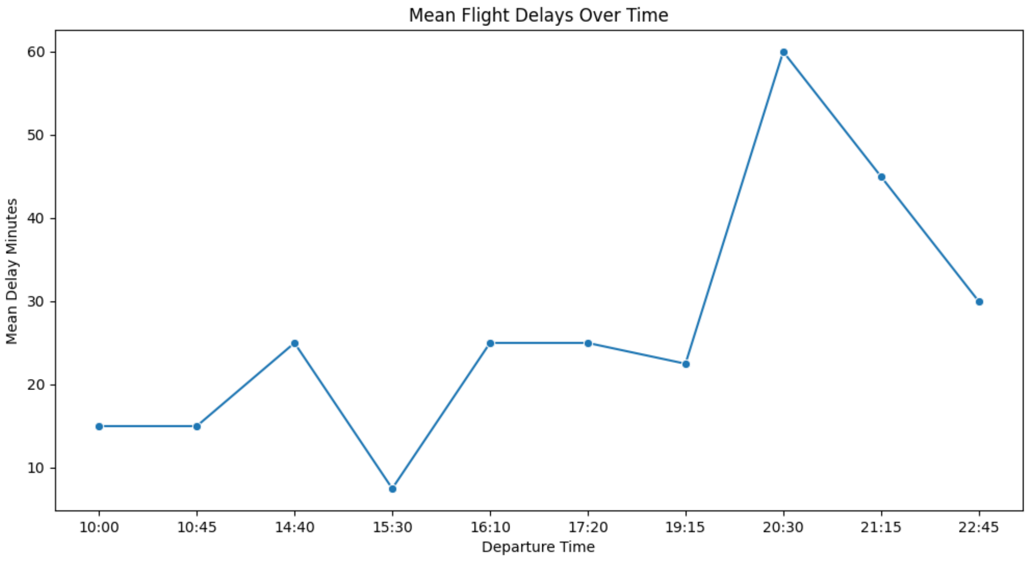
- There is no significant difference in the mean delay times between the airlines, as found from the t-test for the airlines.
- Correlation between Departure Time (in minutes) and Delay Minutes: 0.6248542726807503. This means that the correlation is strongly positive and correlation values are between -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.

Provide a summary of the key findings from the data

- There are total of 12 observations (rows) and 8 variables (columns).
- There are no duplicate entries in the dataset.
- There were missing values which were handled.
- The statistical summary of Delay Minutes **(after preprocessing)** :
    - Number of non null values are 12
    - The average delays in minutes is 25
    - The amount of dispersion(spread) of data in the dataset is 15.07
    - The minimum delay is 5 minutes
    - The maximum delay is 60 minutes (1 hour)
    - First Quartile(25%) is 15.0, means that 25% of the data points are less than 15.0.
    - Second Quartile(50%) is 25.0, means that 50% of the data points are less than or equal to 25.
    - Third Quartile(75%) is 26.25, means that 75% of the data points are less than or equal to 26.25
- The missing values are handled by using the mean method, because if the columns of NaN would have been replaced with 0 it would be unrealistic as in real-world scenario there are always delays.On an average the delay is 25 mins.
- For determining the significant difference between times between the airlines, I have used the approach of t-tests. T-tests are ideal while dealing with groups and since there are three different airlines, hence 3 different groups.
    - **Null hypothesis (H0)** is there is no significant difference in the mean delay times between the airlines.
    - **Alternative Hypothesis (H1)** is there is a significant difference in the mean delay times between the airlines.
    - The alpha value is taken as standard 0.05
    - We are performing paired t-test as we can find the difference between delays for any two airlines. The groups are:

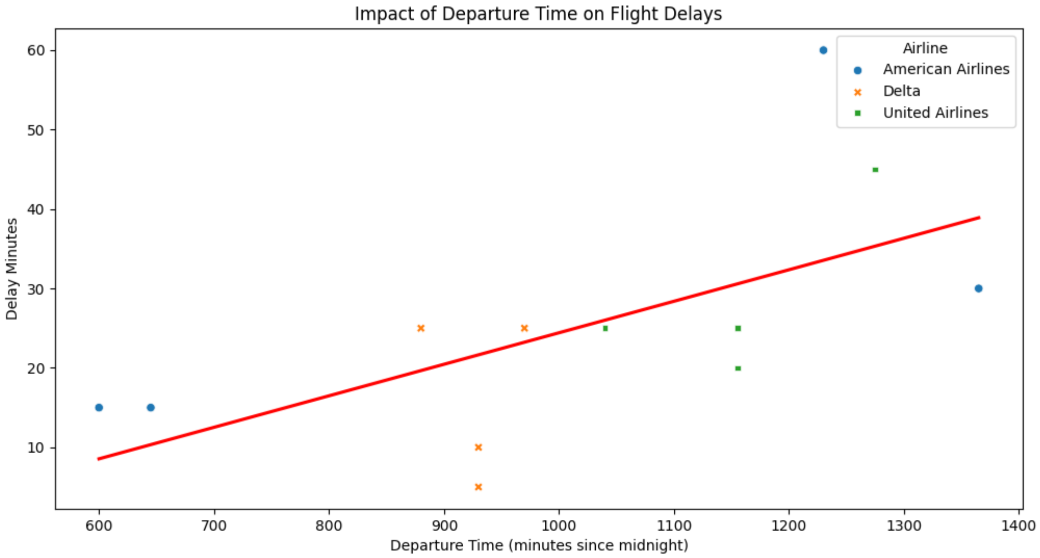
        - American Airlines, Delta
        - Delta, United Airlines
        - United Airlines, American Airlines
    - After **comparing American Airlines and Delta**:
        - T-statistic: 1.165997668006996, P-value: 0.2878631950549536
        - Conclusion is **fail to reject the null hypothesis.**
    - After **comparing Delta and United Airlines**:
        - T-statistic: -1.651445647689541, P-value: 0.14973733521583907
        - Conclusion is **fail to reject the null hypothesis.**
    - After **comparing United Airlines and American Airlines**:
        - T-statistic: -0.1044465935734187, P-value: 0.9202189075644182
        - Conclusion is **fail to reject the null hypothesis.**
    - We accept the Null Hypothesis for all the airlines. So, there is no significant difference in delays between the airlines.

Analyze the impact of departure times on delays.



This figure shows the delays for time of the day and we can deduce the following points from it:

- The delay pattern is non-linear
- Flights departing after 19:15 have high chances of delay of 25 min or more
- The morning flights have less delays
- The evening flights have more delays
- From this graph, the unusual amount of delays are:
  - At 15:30 with less than 10 mins delay
  - At 20:30 with almost an hour delay



We can infer the following from this graph:

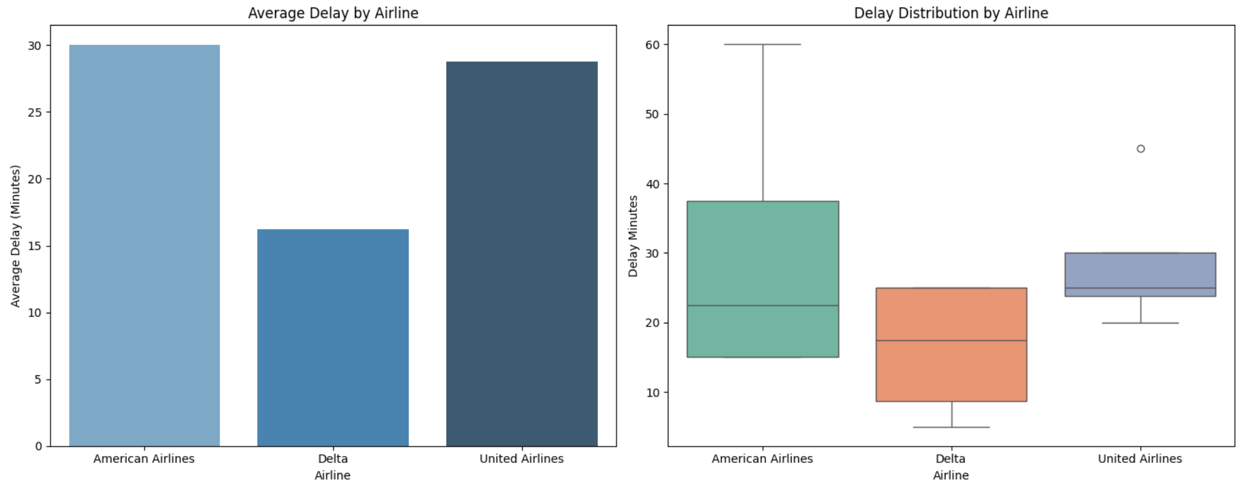
- There is a positive correlation i.e., for increase in x-axis, there is increase in y-axis.
- The delay increases with time i.e., the later in the day it is, the more is the chances of having long delays.
- For American airlines, there is an outlier with around 60 mins delay.
- For United airlines, the delay is within the range of 20-30 mins mostly.
- For Delta, the delay is very less for any given time within the range of 10-30 mins, the highest being below 30 mins

Compare delay distributions between airlines.

- Delay and average delay per airline are as follows:

Airlines	Delay (in mins)	Average (in mins)
American Airlines	120	30.0
Delta	65	16.25
United Airlines	115	28.75
Total	300	75

Visualize the average delay by airline and the delay distribution using appropriate charts.



- Average delay by Airline
  - **American Airlines** has the highest average delay, approximately **30 minutes**.
  - **United Airlines** has a slightly lower average delay, just under **30 minutes**.
  - **Delta** has the lowest average delay, roughly **15 minutes**.
- Delay Distribution by Airline
  - **American Airlines**:
    - The median delay is around **20 minutes**.

        - The **box** spans from about **15 to 40 minutes**, representing the **interquartile range** (IQR), meaning 50% of the delay times are within this range.
        - The **whiskers** extend up to **60 minutes**, indicating the general spread of delays, without any significant outliers.
    - **Delta**:
        - The **median delay** is around **15 minutes**.
        - The **interquartile range** IQR for Delta is much narrower, ranging from **5 to 25 minutes**, indicating that Delta's delay times are more consistent and less variable.
        - There are no major outliers in Delta's delay distribution.
    - **United Airlines**:
        - The **median delay** is about **25 minutes**.
        - The **interquartile range** IQR ranges from **20 to 30 minutes**, showing a somewhat consistent delay pattern, but the box is narrower than American Airlines.
        - There is one **outlier** beyond the whiskers, indicating that at least one flight had an unusually long delay.

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

This plot (Fig 1) 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

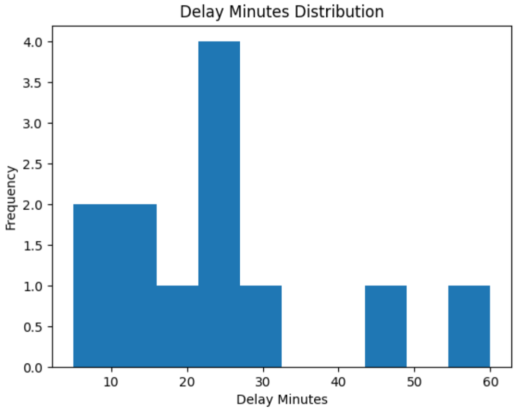


Fig 1: Delay Minutes Distribution

55-60 mins.

The left plot (Fig2 left) 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 (Fig2 right) 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.

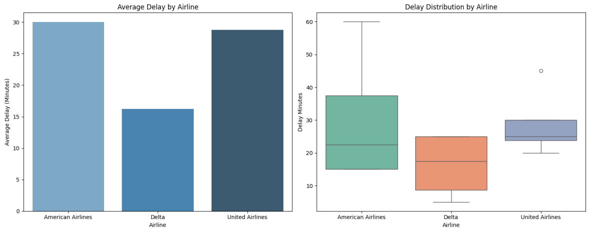


Fig 2 : (left) Average Delay by Airline, (right) : Delay Distribution by Airline

The plot (Fig 3) 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.

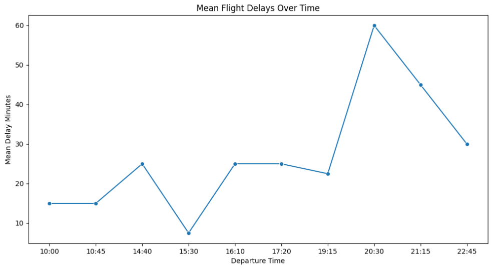


Fig 3: Mean Flights Over Time

This plot (Fig 4) 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.

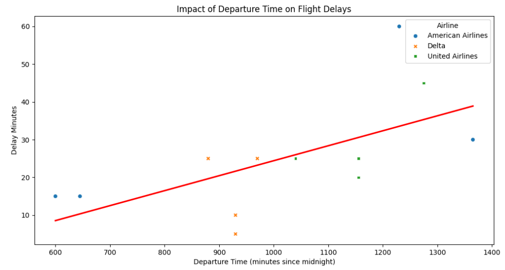


Fig 4: Impact of departure time on delays

For Fig 5,

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.

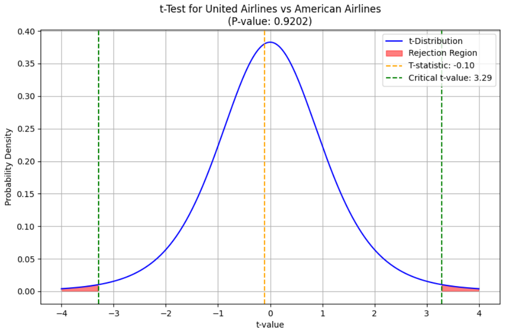


Fig 5: T-test for United Airlines and American Airlines

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

For Fig 6,

The paired t-test for Delta and United 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.

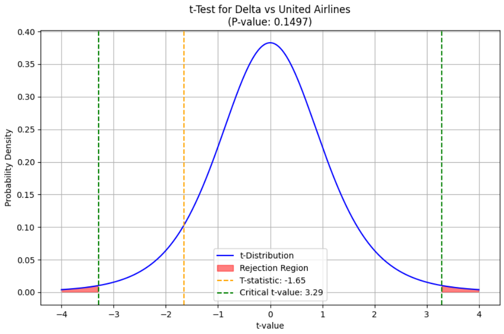


Fig 6: T-test for Delta and United Airlines

For Fig 7,

The paired t-test for Delta 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.

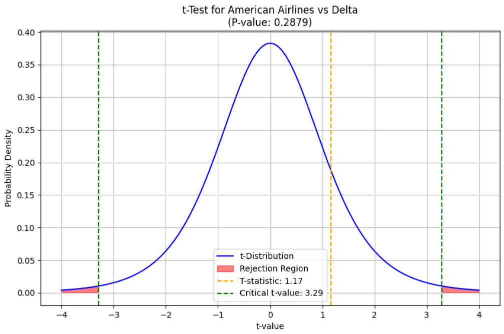


Fig 7: T-test for American Airlines and Delta

Recommendations based on the analysis.

- The American and United Airlines should increase staff for the later part of the day when the delay is more, since the delay for these airlines are highest.
- The airlines should review the time schedule of their flights and also change the timings so that delays can be minimized.
- The departure of one flight to another should have a significant time gap so that if any flight is late then the other flights after it are not affected.
- The airlines should first check their workforce and capacity and then schedule flights and bookings, as with lesser workforce it is difficult to manage and deliver on time.
- 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.
- 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.