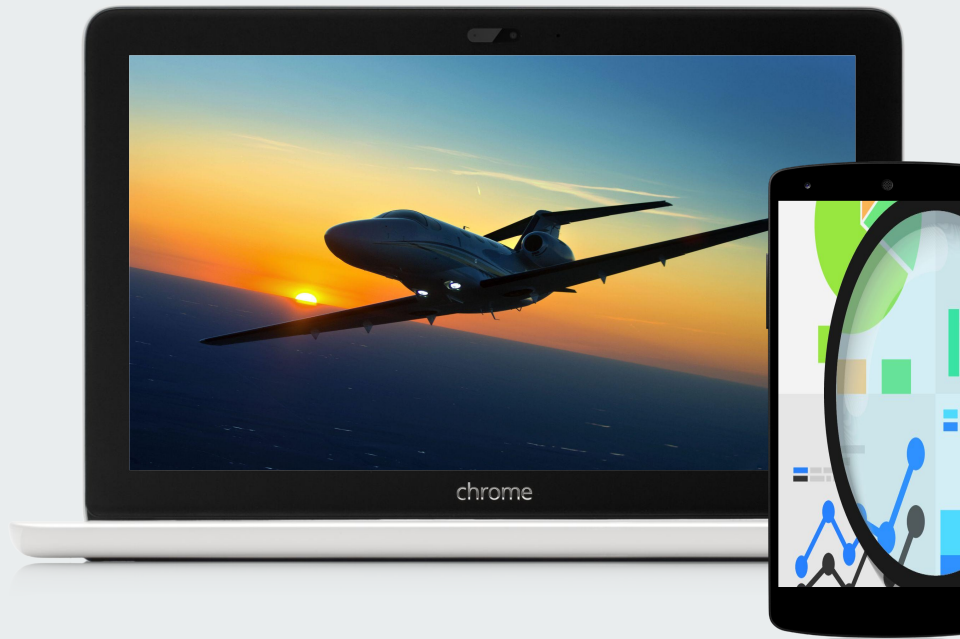




Flight Data Analysis

Data Epic



Outline

Introduction-----

Data Dictionary-----

Data Cleaning Steps-----

Features Created-----

Data Aggregation-----

Exploratory Analysis-----

Data Visualization and Insight Generated--



INTRODUCTION

The **U.S. Department of Transportation's (DOT)** Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in **DOT's monthly Air Travel Consumer Report** and in this dataset of 2015 flight delays and cancellations.

While performing this Analysis, we aim to:

Clean the Data: Ensuring accuracy and reliability in our analysis.

Conduct Exploratory Analysis: Gaining insights into patterns, trends, and anomalies.

Perform Feature Engineering: Creating new features to enhance the depth of our analysis.

Aggregate Data: Summarizing metrics by relevant categories, such as airlines.

Visualize Data: Translating raw data into meaningful insights through visual representations.

Travelers and the **aviation industry** will ultimately profit from this thorough approach, which will also reveal the precise details of the 2015 flight delays and cancellations and open the door for informed **decision-making**.

Data Dictionary: Flight.csv

S/N	Column Name	Description
1	DAY_OF_WEEK	The numerical representation (1 for Monday, 2 for Tuesday, etc.)of the day of the week when the flight occurred.
3	FLIGHT_NUMBER	A unique identifier assigned to a specific flight.
4	TAIL_NUMBER	The aircraft tail number, a unique alphanumeric code assigned to each aircraft.
5	ORIGIN_AIRPORT	The name of the airport from which the flight departed.
6	DESTINATION_AIRPORT	The name of the airport to which the flight arrived.
7	SCHEDULED_DEPARTURE	The scheduled departure time of the flight.
8	DEPARTURE_TIME	The actual departure time of the flight.

Data Dictionary: Flight.csv

S/N	Column Name	Description
9	DEPARTURE_DELAY	The difference in minutes between the actual and scheduled departure times. A positive value indicates a delay.
10	WHEELS_OFF	The time when the aircraft's wheels left the ground during takeoff.
11	SCHEDULED_TIME	The scheduled duration of the flight.
12	TAXI_OUT	The Taxi-out time is defined as the time spent by a flight between its actual off-block time (AOBT) and actual take-off time (ATOT)
13	ELAPSED_TIME	The actual elapsed time of the flight, from wheels-off to wheels-on.
14	AIR_TIME	The time the aircraft is in the air, excluding taxi time.
15	DISTANCE	The distance covered by the flight, often measured in miles or kilometers.
16	WHEELS_ON	he time when the aircraft's wheels touched the ground during landing.

Data Dictionary: Flight.csv

S/N	Column Name	Description
17	TAXI_IN	The time spent taxiing from the runway to the arrival gate after landing.
18	SCHEDULED_ARRIVAL	The scheduled arrival time of the flight.
19	ARRIVAL_TIME	The actual arrival time of the flight.
20	ARRIVAL_DELAY	The difference in minutes between the actual and scheduled arrival times. A positive value indicates a delay.
21	DIVERTED	A binary indicator (0 or 1) specifying whether the flight was diverted (1) or not (0).
22	CANCELLED	A binary indicator (0 or 1) specifying whether the flight was canceled (1) or not (0).
23	YEAR	The year in which the flight occurred.
24	MONTH	The month in which the flight occurred.

Data Dictionary: Flight.csv

S/N	Column Name	Description
25	DAY	The day of the month on which the flight occurred.
26	CANCELLATION_REASON	The reason for flight cancellation, if applicable.
27	AIR_SYSTEM_DELAY	The time delay caused by the air traffic control system.
28	SECURITY_DELAY	The time delay caused by security-related issues.
29	AIRLINE_DELAY	The time delay attributed to issues related to the airline.
30	LATE_AIRCRAFT_DELAY	The time delay caused by the aircraft arriving late from a previous flight.
31	WEATHER_DELAY	he time delay attributed to adverse weather conditions.

Data Cleaning

Data cleaning, also known as data cleansing, is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database.

For the flight dataset we will be using the following cleaning processes

Cleaning Process

Step 1: Check for duplicates and drop

Step 2: Outlier detection

Step 3: Data Transformations

Step 4: Rule based cleaning



Features Created

Feature engineering is the process of using domain knowledge to extract features from raw data¹. It is the process of transforming raw data into features that are suitable for machine learning models

For the flight data set, we will create a bunch of features to help in generating insights and to better train models for the flight dataset.

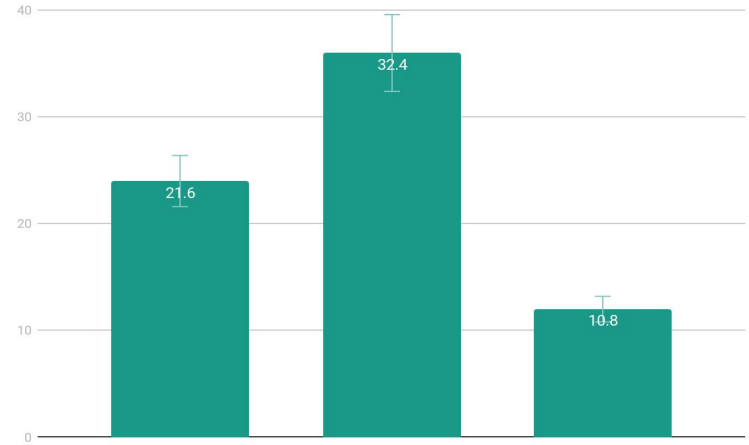
Some of the features created include:

- **DEPARTURE INDEX:** A ratio of departed time to scheduled departure
- **ARRIVAL INDEX:** A ratio of arrival time to schedule arrival
- **TOTAL DELAY:** A sum of the arrival delay and the departure delay
- **ROUTE:** A combination of the origin and destination airport
- **TIME OF DAY:** tells whether it is morning, afternoon or evening
- **SEASON:** This tells the season depending on the month
- **SPEED OF AIRPLANE:** A ratio of the distance travel to the elapsed time (in km/hr)



2015 US Flight Wrapped

Statistical summary and insight generated
of the flight data



Most Used Airline in 2015

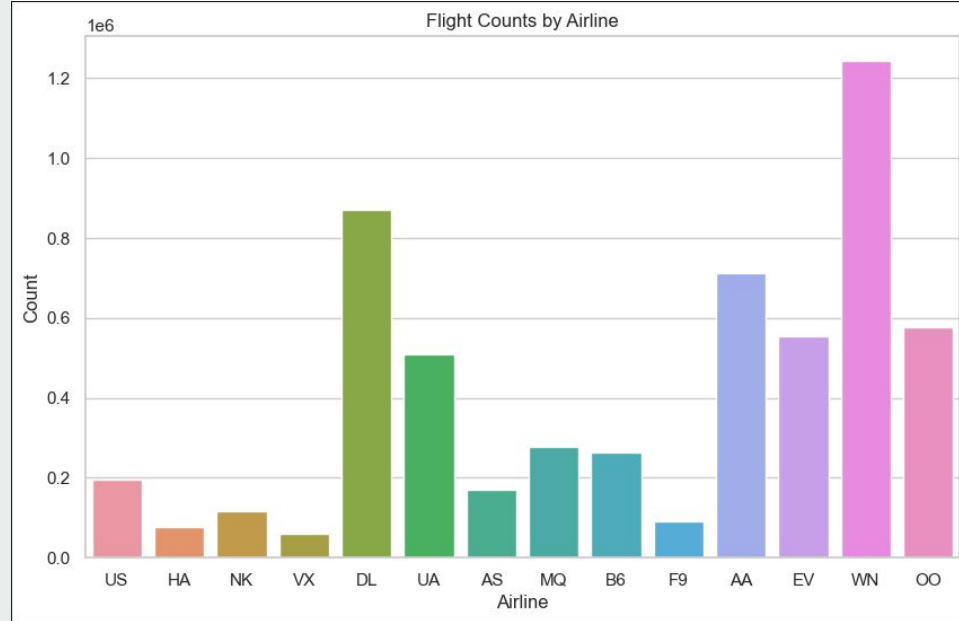
➤ Southwest Airlines Co.(WN)

With over 1,242,403 no of flights

RUNNER-UP

➤ Delta Air Lines Inc.

With over 870,275 no of flights



Day of the Week with most Flight



Thursday

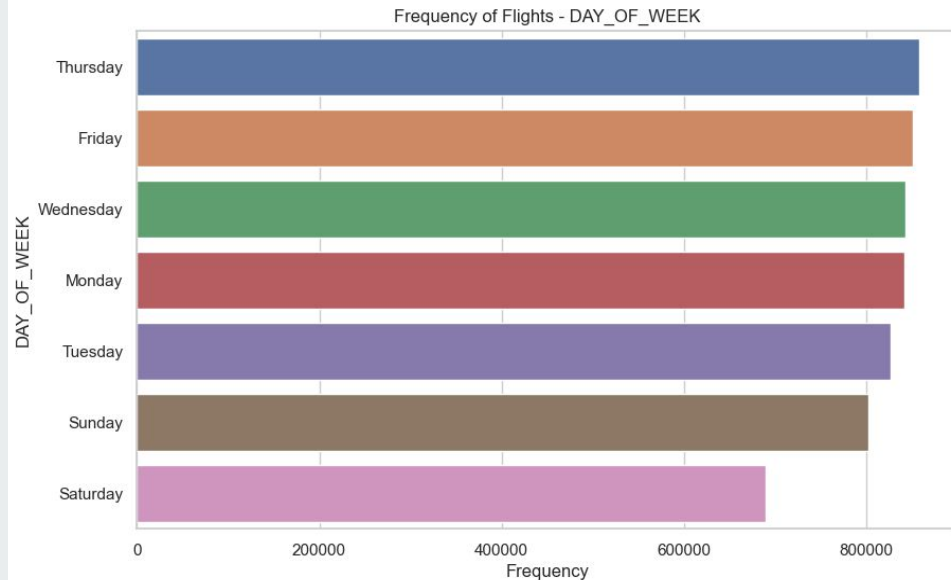
With over 857,886 no of flights on this day


RUNNER-UP



Friday

With over 851,387 no of flights on this day.





Time of the Year with the most trip

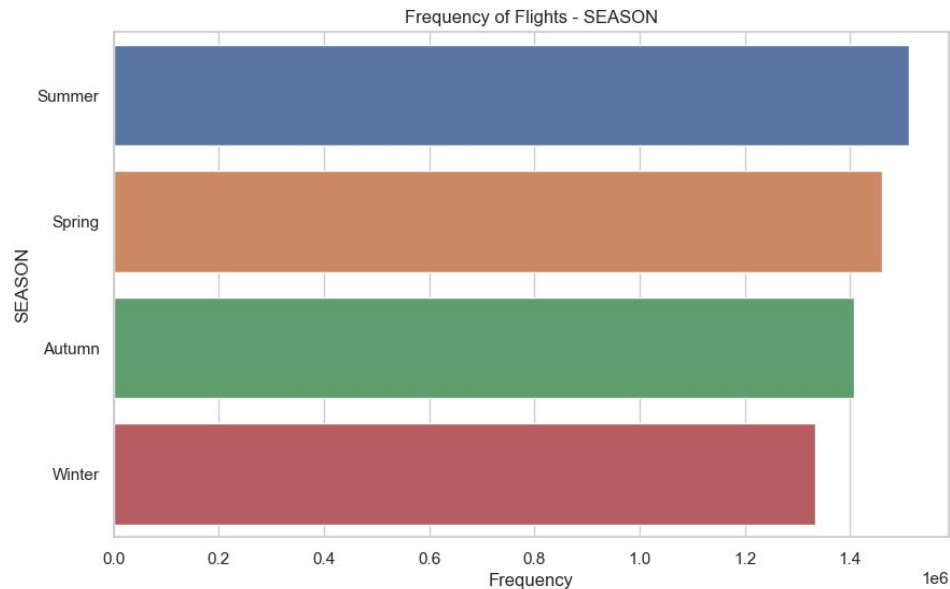
➤ Summer

With over 1,511,187 no of flights in this period

RUNNER-UP

➤ Spring

With over 1,461,030 no of flights in this period.

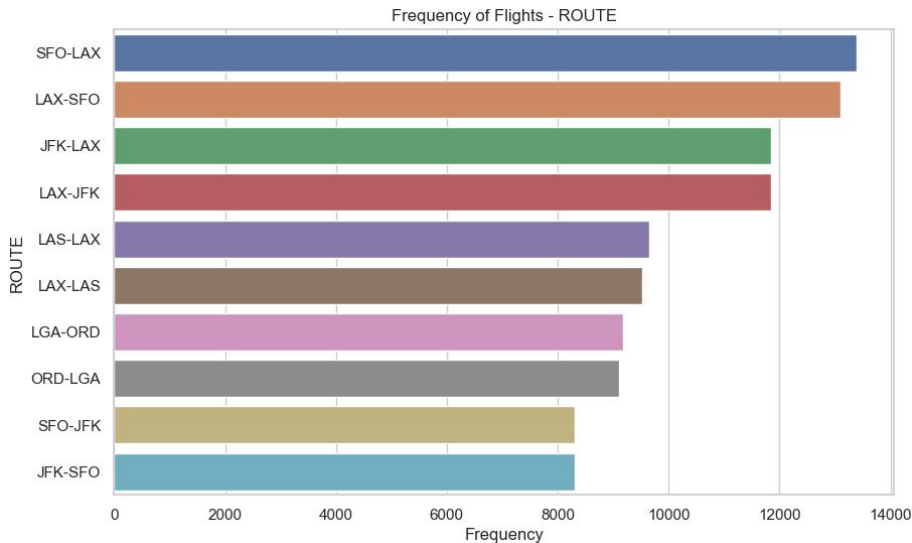


Most Used Route

- **San Francisco International Airport to Los Angeles International Airport (SFO-LAX)**
With over 13,400 made through the route

RUNNER-UP

- **Los Angeles International Airport to San Francisco International Airport (LAX-SFO)**
With over 13,109 made through the route.



Month with the Most Trip



July

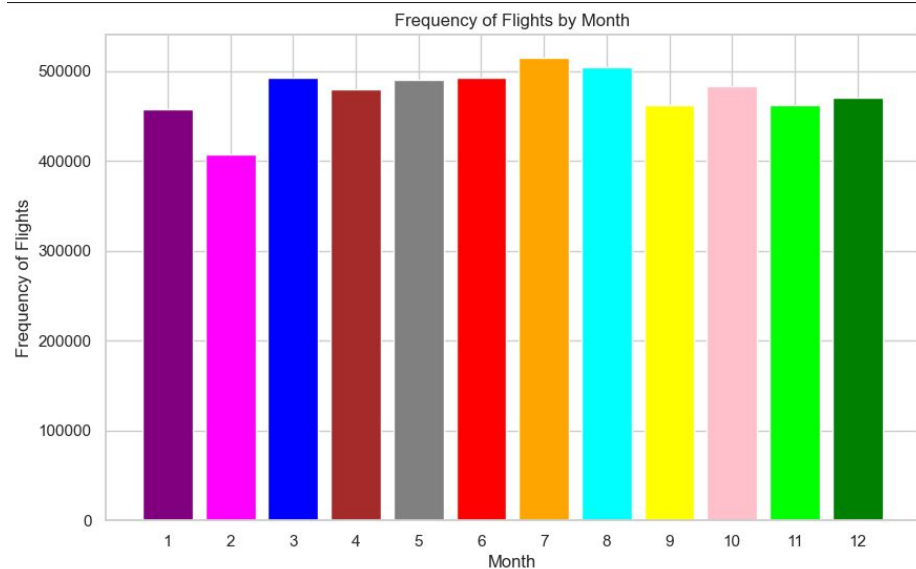
With about 514,384 flights

RUNNER-UP



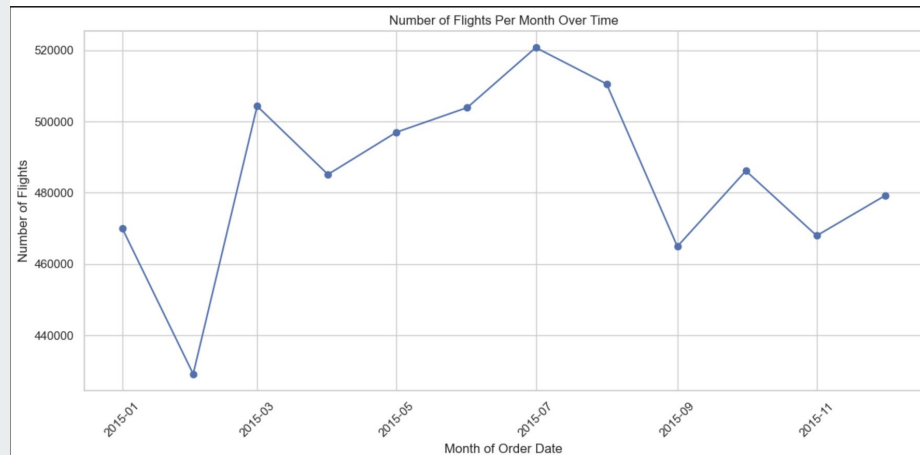
August

With about 503,956 flights



Monthly Flight Trend

A visual representation of the time series analysis of the trend over the year.



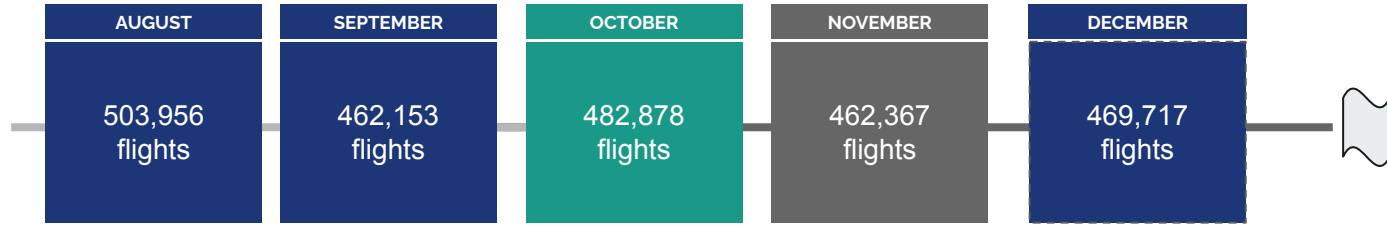


Number of Flight Completed for each Month





Number of Flight Completed for each Month



Other Statistical summaries

- Shortest Distance and Route: **31 km(WRG-PSG)**
- Longest Distance and Route: **4983 km(HNL-JFK)**
- Highest Departure Delay: **3959 minutes**
- Lowest Departure Delay: **-162 minutes**
- Highest Arrival Delay: **1971 minutes**
- Lowest Arrival Delay: **-87 minutes**
- Mean Total Delay: **18.11 minutes**
- Median Total Delay: **-11.0 minutes**
- Mode of Total Delay: **-20.0 minutes**
- Percentage of Cancelled Flights: **1.54%**
- Percentage of Diverted Flights: **0.26%**

Data Aggregation



AVERAGE TAXI DURATION BY AIRLINE

AIRLINE	AVERAGE_TAXI_DURATION
str	f64
"AS"	21.497991
"US"	26.637014
"EV"	24.369651
"AA"	26.643592
"HA"	17.794501
"VX"	22.943029
"B6"	23.99386
"WN"	18.110806
"NK"	24.158753
"MQ"	25.628789
"DL"	24.817017
"F9"	24.820694
"OO"	25.075532
"UA"	25.919617

AVERAGE SPEED BASED ON AIRLINE

AIRLINE	AVERAGE_SPEED
str	f64
"HA"	243.792244
"MQ"	258.52292
"EV"	269.487384
"OO"	271.143919
"US"	329.428654
"DL"	335.149449
"WN"	346.40146
"AA"	350.654126
"B6"	351.931898
"NK"	364.06717
"F9"	367.108435
"VX"	373.845026
"AS"	374.173617
"UA"	377.010508

Data Aggregation

AVERAGE DISTANCE BY DAY OF THE WEEK

DAY_OF_WEEK	AVERAGE_DISTANCE
---	---
str	f64
Saturday	857.545201
Thursday	818.616981
Wednesday	811.784365
Monday	816.347052
Friday	817.928374
Tuesday	809.096095
Sunday	831.991691

PERCENTAGE OF FLIGHT DELAY FOR EACH AIRLINE

AIRLINE	PERCENTAGE_DELAYED
str	f64
"UA"	50.349928
"B6"	38.829644
"WN"	45.485419
"AS"	25.34233
"EV"	30.486063
"OO"	29.628069
"DL"	32.385972
"VX"	38.085852
"HA"	26.466454
"US"	32.113435
"MQ"	33.439893
"F9"	38.645476
"NK"	45.11667
"AA"	34.363428

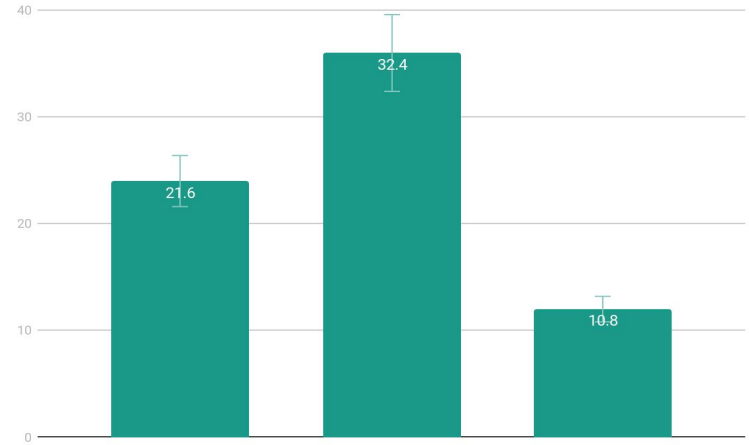
Data Aggregation

TOTAL FLIGHT BY DAY OF THE WEEK

DAY_OF_WEEK	TOTAL_FLIGHTS
---	---
str	u32
Thursday	872521
Monday	865543
Friday	862209
Wednesday	855897
Tuesday	844600
Sunday	817764
Saturday	700545



CANCELLATIONS, DELAY AND DIVERSIONS STUDY



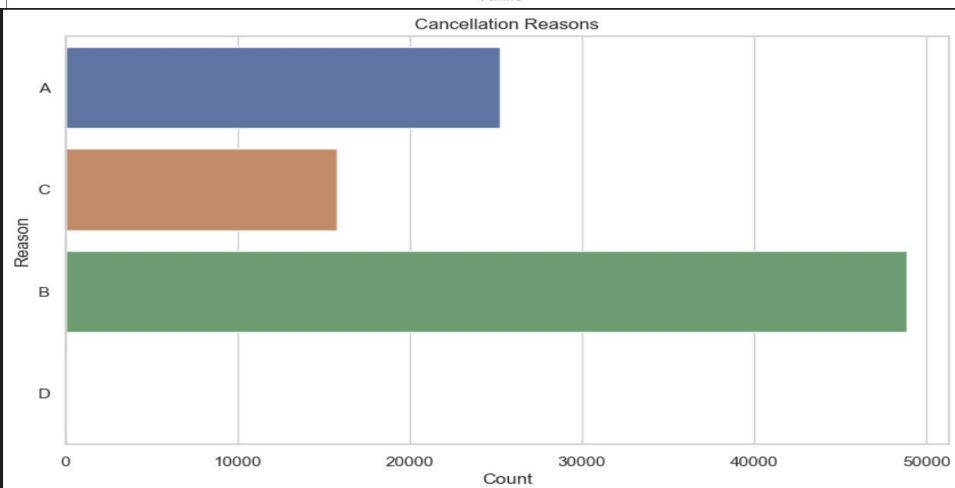
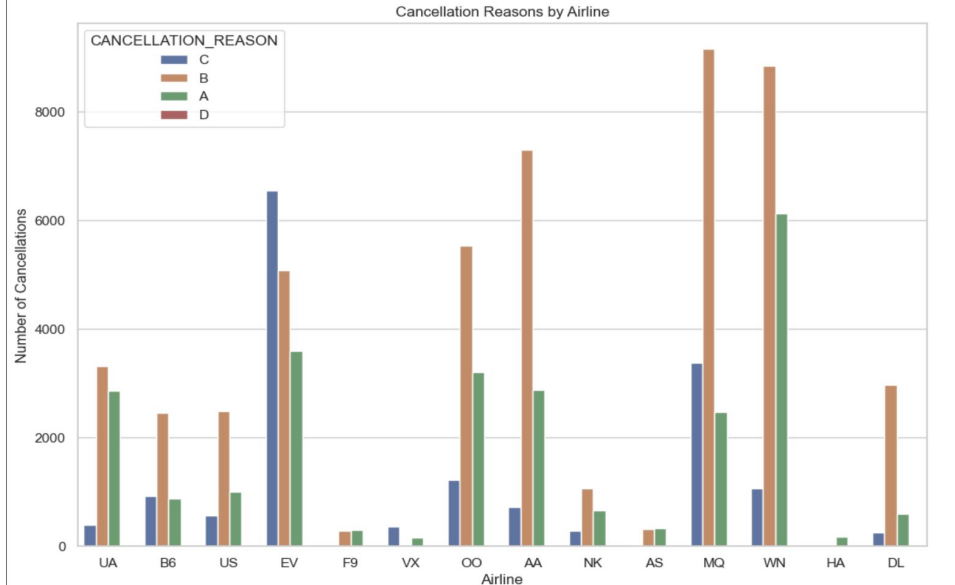
Intent

The study of flight cancellations, delays, and diversions is an essential component of airline and airport operational management. This analysis offers invaluable insights into the efficiency and reliability of air transportation services, which are critical for both service providers and consumers.

Cancellations

The flight data highlights the prevalence of different cancellation reason, such as A, B,C with B being the most common among airlines. Some airlines demonstrate notably higher cancellations due to specific reasons, suggesting areas for operational improvement and also area of low cancelation could be a hint for travelers on which airline to use.

These insights can guide travelers in choosing more reliable airlines(For example Hawaiian Airlines Inc., Frontier Airlines Inc.) and assist airlines in addressing the root causes of cancellations to bolster efficiency and customer satisfaction(For example, creating solution for Reasons B).

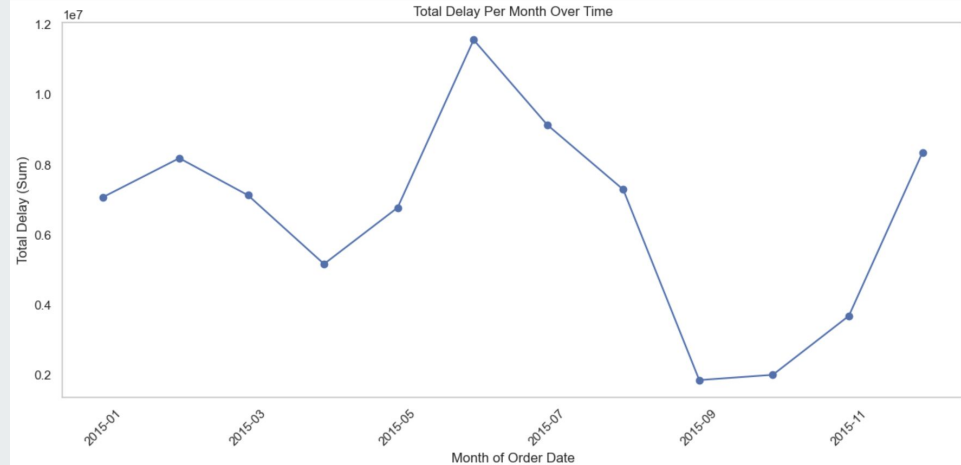
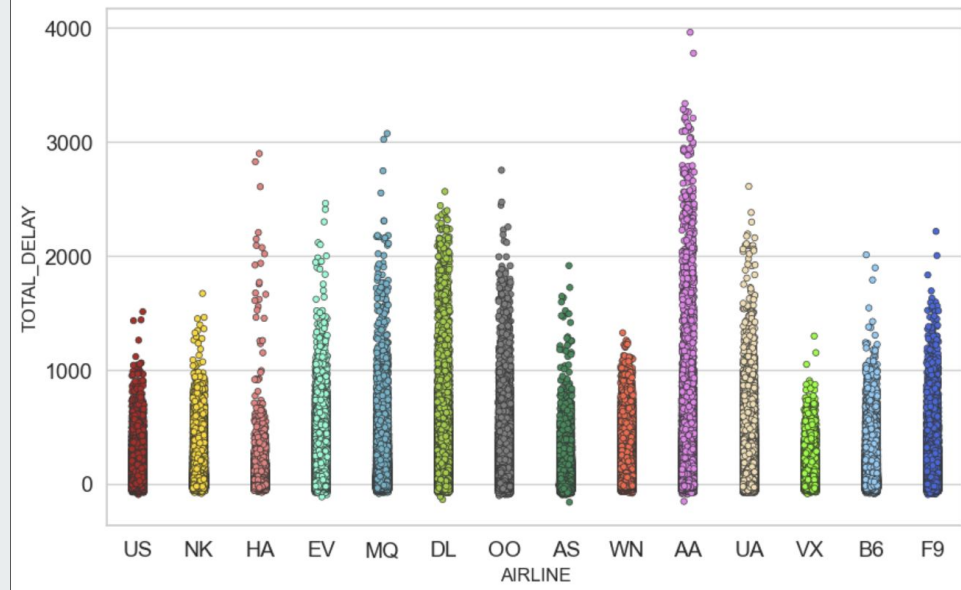


Delay

The analysis of flight data reveals that while delays are a common occurrence across all airlines, certain carriers experience them more frequently. Notably, American Airlines Inc. records the highest number of both departure and arrival delays, whereas Southwest Airlines Co tends to have the fewest.

Such visual data representations are instrumental in uncovering delay patterns, essential for airlines to enhance their operational strategies and punctuality, ultimately boosting passenger satisfaction. Additionally, they empower travelers to make informed choices about which airlines are more likely to maintain on-time schedules.

The Total Delay also neared the all time high towards december due to frequency of people traveling in that period for holidays.



OTHER INSIGHTS

- **Seasonal Variations:** Travelers and airlines could benefit from understanding how flight patterns vary with seasons and months. For instance, if certain months have higher flight volumes, it could indicate peak travel times, which might be due to holidays or seasonal tourism, it could also implicate harshness of weather in a particular season. Airlines could use this information to adjust capacity, while travelers might want to book in advance or expect higher prices during these times.

OTHER INSIGHTS

- **Cancellation Patterns:** High cancellation rates might reveal operational challenges or external factors affecting flights. For instance, a high rate of Reason B could suggest a need for improvement in that particular factor, prompting airlines to plan for potential disruptions and travelers to consider travel insurance or flexible booking options.

OTHER INSIGHTS

- **Delays and Airline Performance:** Comparing delay times across airlines could indicate efficiency and reliability. Airlines with lower average delays might be preferred by travelers who value punctuality. Airlines could use this insight to benchmark and drive operational improvements.

CONCLUSION

In conclusion, The flight data analysis of flight data offers valuable insights into airline operational performance, revealing variations in delay frequencies and durations, which is crucial for airlines to improve operations and customer service, and for passengers to make informed choices.

