# VISUALIZING FLIGHT TRAFFIC TRENDS: ANALYZING THE BUSIEST AIRLINE ROUTES FROM 2015 TO 2019

## ABSTRACT:

The aviation sector plays a vital role in bringing people, products, and financial systems across the globe. Understanding the fundamental patterns in air traffic is important to enhance its operational efficiency, ensure customer satisfaction, and address logistical challenges, given the large number of flights each year. Analyzing airline data through visualization helps us discover patterns, evaluate performance metrics, and detect areas to improve, ultimately helping in wise decision-making in this industry.

The airline sector produces a lot of data, including flight timetables and disruptions, as well as passenger volume and route efficiency. This extensive data source offers a chance to utilize data visualization methods to discover more practical insights. Users can readily grasp important information and base their decisions on data analytics by processing unclean, noisy data into easy-to-understand and engaging visual displays. The reason for examining airline data for this study highlights its importance in worldwide transportation and the potential for data analytics to ignite innovation. Visualization is not just for simplifying large datasets, but also for enhancing communication, which makes it easier to analyze airline trends and aviation traffic dynamics.

Furthermore, the reason for choosing this dataset from 2015 to 2019 is the growing importance of past patterns in air transportation. This analysis lays some groundwork to predict possible obstacles and enhance passenger satisfaction. In addition, analyzing data such as flight frequencies and time days can assist in finding out the most heavily traveled routes, operational challenges, and seasonal trends.

## DATASET:

### Introduction:

For this assignment, we took the Airline Delay Analysis dataset from Kaggle repositories: https://www.kaggle.com/datasets/sherrytp/airline-delay-analysis. "Why this dataset?". We chose this dataset because it provides a comprehensive analysis of airline delays, and detailed information about flights, routes, delays, and other key metrics across multiple years. By using this dataset, we aim to compare the airline performance over different years, address the delay trends, and analyze the average arrival delay for most used routes in the world.

Our dataset covers two important factors of big data: **Volume** and **Variety**. It contains **9 GB** of data, distributed across **12 CSV files**, representing airline performance data from **2009 to 2020**. Each file consists of more than **5.8 million rows** and **25-30 columns**, providing granular insights into various flight attributes. The dataset is a massive collection of 69 million records after combining all files that represent flight information for past years. There are categorical data like airline names, routes, and origins, as well as numerical data such as delay in minutes, distance, and flight count.

Diverse attributes like flight delays, air traffic, cancellations, some diversion, and operation data are present. The dataset spans 12 years, including time series analysis of trends like the impact of weather, daily changes, or any global incidents. This dataset provides a good foundation for performing a wide range of analytics tasks, referring from historical performance to predictive modeling. While addressing the challenges we faced regarding storage, processing, and analyzing.

### Data exploration, processing, and cleaning:

To prepare the dataset for creating a graph, there are several steps such as data exploration, data processing, and data cleaning are performed to ensure the accuracy and relevance of the visualization. The dataset covers multiple years and consists of over 5.8 million rows per file. To create a visualization, we selected data for the top 5 most busy routes in the world and then compared flight arrivals, and delays, between 2015 and 2019.

## Data Cleaning and Processing:

Due to the large size of the dataset, it was difficult to open on any local computer and caused performance issues when trying to load on Google Colab. Initially, we downloaded the zip file from the Kaggle repository and then we decided to upload the file on Google Drive. This solution helped to resolve storage issues and provided a reliable way to access and process the data efficiently. We used Python as a programming language, then imported the different libraries such as Pandas, NumPy, matplotlib, and seaborn for data cleaning and processing. After uploading the data, we mounted the drive with colab using the mount function to access the data from the required files.

The dataset includes multiple unnecessary columns which are removed to streamline the analysis. Additionally, handling NaN or Null values posed challenges for predictions. To address this issue, we replaced the Null values with "0" to ensure the data maintained a usable structure and provided some context for missing entries. Data was also aggregated to summarize the total number of flights per route per year.

We created a new column named “Year” and changed the date format from ‘YYYY-MM-DD’ to ‘YYYY’ which was useful to visualize the data year-wise. Dropping some columns as it is was necessary to shorten the process. After filtering unwanted columns and cleaning the data properly, all the files were combined into one data frame. To determine the busiest routes, we aggregated the flight data by grouping the dataset in the Route column which was a combination of ORIGIN and DEST. e.g. “data_combined['ORIGIN'] + ' → ' + data_combined['DEST']”.

After calculation, the data was sorted into descending order. This helps in identifying the routes with the highest number of flights. After sorting we filtered the top 5 busiest routes on a high flight count basis. This also gave some insights into the effect of the pandemic on the airline industry. This was necessary to narrow down the focus to the most used routes that gave meaningful insights. Once the routes were selected, the data was filtered to include only the flight count for the selected routes.

To analyze airline performance, two main key metrics are calculated to provide meaningful insights into flight trends and delays. The “avg_arr_delay” represents the average arrival for flights for each route year combination. Then “flight_count” captures the total number of flights for each route and year combination which is determined by counting occurrences of "FL_DATE" within each group. This highlights the volume of flights operating on specific routes over time. After calculating these aggregate metrics the data was adjusted into a structural format using the ‘.reset_index’ function. The attributes used for visualization are “Route”, “Year”, and “Flight Count” as they directly contribute to understanding the distribution and trends of flight activity across different routes over time. It can be easily argued that the selected subset of data and attributes was enough as it can roughly demonstrate the volume of flights across different routes.

Finally, the processed data, containing only top routes, was saved into a CSV file and was further visualized effectively in Tableau.

## VISUALIZATION:

Visualization is a crucial part of comprehending and analyzing extensive datasets by converting the missing and noisy ones into meaningful and structured information. Analyzing flight trends is important for the aviation sector to optimize resources, plan operations, and enhance passenger experiences.

Using a stacked car chart showcases important patterns, annual changes, and the impact of different routes on total traffic. This is also an ideal representation of displaying time-based information, providing a straightforward way to compare patterns across various aspects (such as routes and years). The final bar chart presents a year-by-year comparison of flight counts for each of the top 5 routes, clearly illustrating trends such as increases or plateaus in flight activity. By focusing on these key attributes and subsets, the graph provides actionable insights into the busiest flight routes and their traffic trends over a defined period.

The data includes three major attributes, namely the routes, years, and metrics of the flight count.

**Years:** Range of years (2015-2019).
**Routes:** The top five routes over the five years.
**Metrics:** Flight counts per route per year.

## Structure of the visualisation:

The X-Axis lists the 5 busiest airline Routes as follows:

1. OGG - HNL (Kahului to Honolulu)
2. LAX - LAS (Los Angeles to Las Vegas)
3. LGA - ORD (LaGuardia to Chicago O'Hare)
4. LAX - SFO (Los Angeles to San Francisco)
5. SFO - LAX (San Francisco to Los Angeles).

OGG – HNL: This route represents the two major cities in Hawaii.

LAX – LAS: This route reflects both leisure and business travel.

LGA – ORD: This is one major significant hub-to-hub route.

LAX – SFO and SFO – LAX: These routes also reflect both leisure and business travel, showing bidirectional traffic.

The Y-Axis shows the Flight Counts on each route, which is measured in thousands. This attribute from the dataset makes it easy to scale the differences in flight volume across each year and route.

## Color coding:

The stacked bar charts are distinguished by the year attribute with the lightest blue shade for 2015 and the darkest shade for 2019, thereby allowing users to analyze both flight count for a route and the yearly distribution.

## KEY INSIGHTS:

Several key insights include the top routes, variations through the years, and the growth and decline in flight counts.

### Top routes:

The busiest routes are as follows:

LAX – SFO and SFO – LAX: These routes reflect high demand for travel between Southern and Northern California. The increase in high counts says that these routes are highly critical for both leisure and business travel.

OGG – HNL: Depicts the importance of inter-island travel between the two major cities in Hawaii.

### Year-wise variations:

The variations in routes through the years are as follows:

LAX – LAS: This route shows stable traffic across these five years which highlights the increasing demand.

LGA – ORD: This route shows mild variations and, therefore, shows possible impact from operational changes, seasonal demand, or external factors.

### Growth and decline:

OGG – HNL and SFO – LAX: These routes show a subtle increase with yearly variations.

## ADVANTAGES OF THE STACKED BAR CHART:

There are various advantages to using a stacked bar chart because they are

**Comparison across categories:** Easy for comparison of total values across various categories, in this case, the total flight counts across different routes throughout the years.

**Efficient use of space:** Stacking data within a single bar into a compact space makes it an efficient use of space.

**Visualizing temporal data:** We can analyze and visualize changes in proportions or totals in datasets dealing with time-series trends.

**Multidimensional analysis:** It allows multidimensional analysis of the data, by comparing the x-axis and examining yearly distributions within each bar.

## ROUTE SPECIFIC ANALYSIS:

OGG - HNL (Kahului to Honolulu):
The flight count attribute remained steady at around 10,000 in count every year from 2015 to 2019, with a slight increase observed in 2019.

LAX - LAS (Los Angeles to Las Vegas):
The flight count for this route has consistently drifted 12,000 annually, showing very less variations throughout 5 years, and can be concluded as a stable, high-demand connection.

LGA - ORD (LaGuardia to Chicago O'Hare):
This route exhibits slight variations in flight counts across the years, ranging between 10,000 and 11,000 flights annually. These fluctuations suggest periodic changes in the number of flights operated over time.

LAX - SFO (Los Angeles to San Francisco):
The flight count for this route ranges above 12,000 flights per year, maintaining a steady and prominent presence for five years.

SFO - LAX (San Francisco to Los Angeles):
This route is similar to LAX – SFO, with a similar flight count exceeding 12,000 flights per year.

## TOOLS AND LIBRARIES USED:

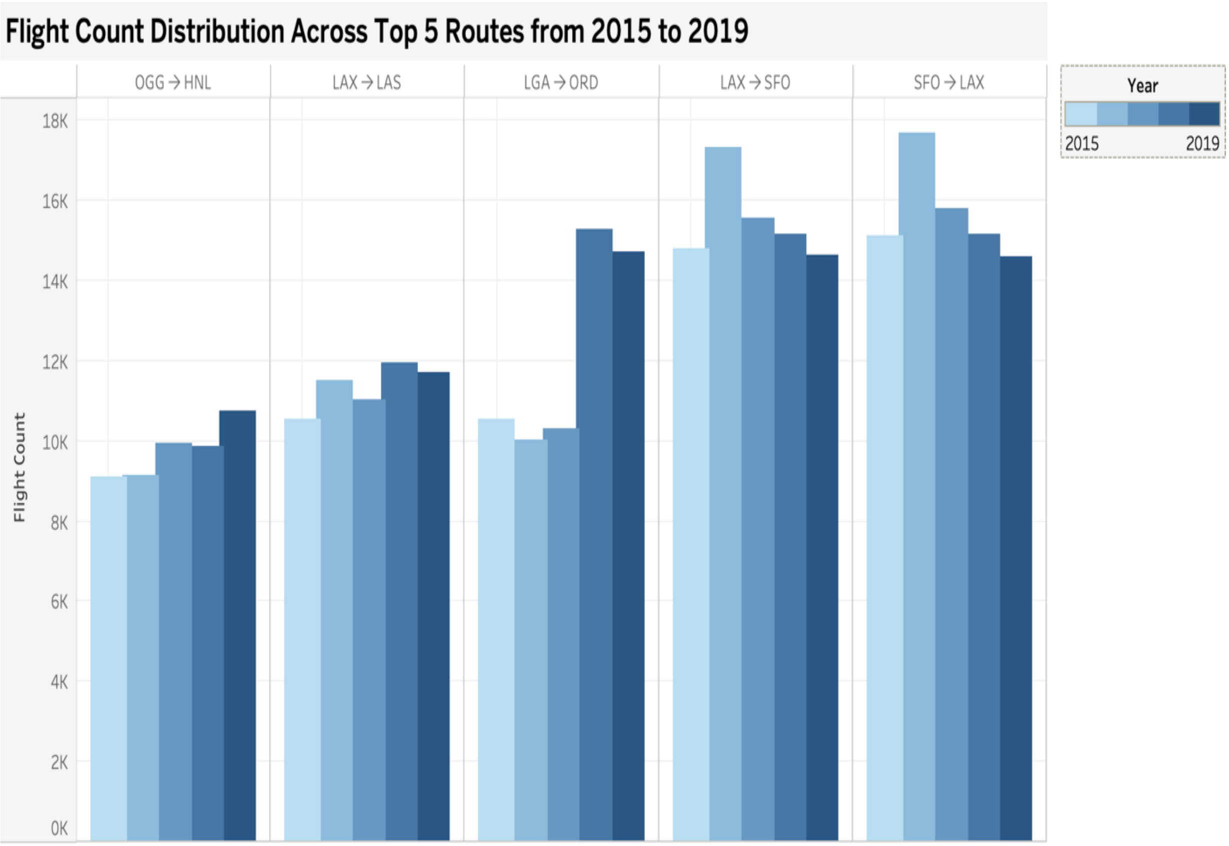
Google Colab and Tableau were primarily used to achieve this data visualization and exploration. Colab, with its cloud-based Python environment, was mainly used to load the data, clean and process it, and aggregate and calculate key attributes like total flight counts and average delays. The usage of Python libraries like numpy, pandas, and seaborn was efficient in handling large datasets.

On the other hand, Tableau was utilized to create a visualization of the data. By mapping the processed data into attributes like routes on the x-axis and flight count on the y-axis, Tableau enabled a detailed exploration of trends over five years from 2015 to 2019.

The combined usage of these tools facilitated both robust analysis and visual representation, providing insights into the flight counts of the top five routes in the United States from 2015 to 2019.

## OUR CHART:

Below is our chart that depicts the flight count distribution from 2015 to 2019 differentiated through shades of blue color highlighted where 2015 being the lightest shade and 2019 the darkest.



The plot of sum of Flight Count for Year broken down by Route. Color shows details about Year from 2015 to 2019. Details are shown for Route. The data is filtered on Flight Count. The Flight Count filter ranges from 9055 to 17683. The view is filtered on Route, which keeps LAX → LAS, LAX → SFO, LGA → ORD, OGG → HNL and SFO → LAX.

## CONCLUSION:

In conclusion, this report showcases flight data from 2015 to 2019 and discusses the patterns and changes in the most heavily used flight route with its corresponding flight count in the United States. This study also emphasizes the uniformity and fluctuations in air traffic patterns by merging data over five years and analyzing important statistics. From the visualization, it is evident that the routes OGG – HNL and LAX – LAS continue to have steady traffic.

Overall, from the visualization, the route SFO – LAX has the most flights out of the top five routes with an estimated number of flights for this route typically falling between 16,000 and 18,000 flights per year, peaking in the year 2016. The consistent number of flights on all routes underscores their importance in the US air travel system, with SFO – LAX being the most heavily traveled route, making it an important transportation link.

## REFERENCES:

1. Kaggle Dataset - https://www.kaggle.com/datasets/sherrytp/airline-delay-analysis

2. Autumn 2024 CSC1143 Data Management & Visualization (10661) - Prof. Suzanne Little https://loop.dcu.ie/course/view.php?id=66616