AIRLINE DELAY AND CANCELLATION

Submitted by: Chitranshi Bhatnagar (22263775)

Niyati Datwani (22261488)

Group - DN

Dublin City University
Data Management and Visualization

Abstract

The aim of this project is to create a dashboard for delays and cancellation of flights. Demand for flights is increasing these days as people are traveling more, especially after the pandemic. Therefore, the aviation industry is experiencing shortages in staff to deploy when unforeseen events and weather requires them to switch up their schedules. So, the main point for doing this analysis is to help people in choosing an airline which is consistent in terms of service and punctuality. This analysis will help the customers make a well-informed decision while keeping their personal travel plans in mind.

Data Collection

We have collected the Airline delay and cancellation data of US aviation industry from Kaggle (https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018), which is a public dataset website. This data is provided by Bureau of Transportation Statistics which is storing real-time flight performances from 1987 and can be downloaded in CSV format. We are taking data from **2015 to 2018** to analyse flight delays and cancellations. The dataset contains details of Origin, Destination, Arrival, Delay and Cancelled timings, Weather details, Air time, Distance and other columns in an object and float data type. The dataset contains almost **24.3 million** records before cleaning and pre-processing. The size of original CSV file is **2.8 GB** and cleaned dataset was of **2.3 GB** which includes data of 4 years. **This addresses the volume aspect of Big Data.**

Data Exploration, pre-processing, Cleaning and Transformation

Data Exploration and pre-processing

Data exploration and pre-processing was performed in python using libraries (Pandas and Numpy) with the goal of separating years and months and removing all the null values. Firstly, we identify the data types of each column and dropped the rows which contained null values. Additionally, our dataset contains short form of airline names. So, we converted the short form into their full name.

Data Cleaning and Transformation

For Data Cleaning, we have used Python libraries (Pandas and Numpy). Firstly, we separated Flight date (FL_DATE) into month and year to analyse flight delays according to month. Then, we converted Arrival time (ARR_DELAY) and Departure time (DEP_DELAY) from minutes to hours for better representation of delays.

For transformation, we dropped some columns (CARRIER_DELAY, WEATHER_DELAY, Unnamed: 0, AIR_TIME, OP_CARRIER_FL_NUM) as it is of less use. Then, we extracted the cleaned data into CSV and imported it in Tableau.

Visualisation

Purpose

We created a dashboard for Airlines delays and cancellations
- To show the comparison between arrival delay and departure delay in a month.
- To show the Arrival Delay of Airlines
- To show which airlines has most delays in 4 years

Design Choices

Software Used – Tableau

Tableau is used as our primary visualisation tool as it can handle large volumes of data with ease. Seamless connection between Jupyter notebook, dataset and Tableau make it an exciting prospect. The practical and easy to use aspect of Tableau makes the visualisation part of our report more appealing.
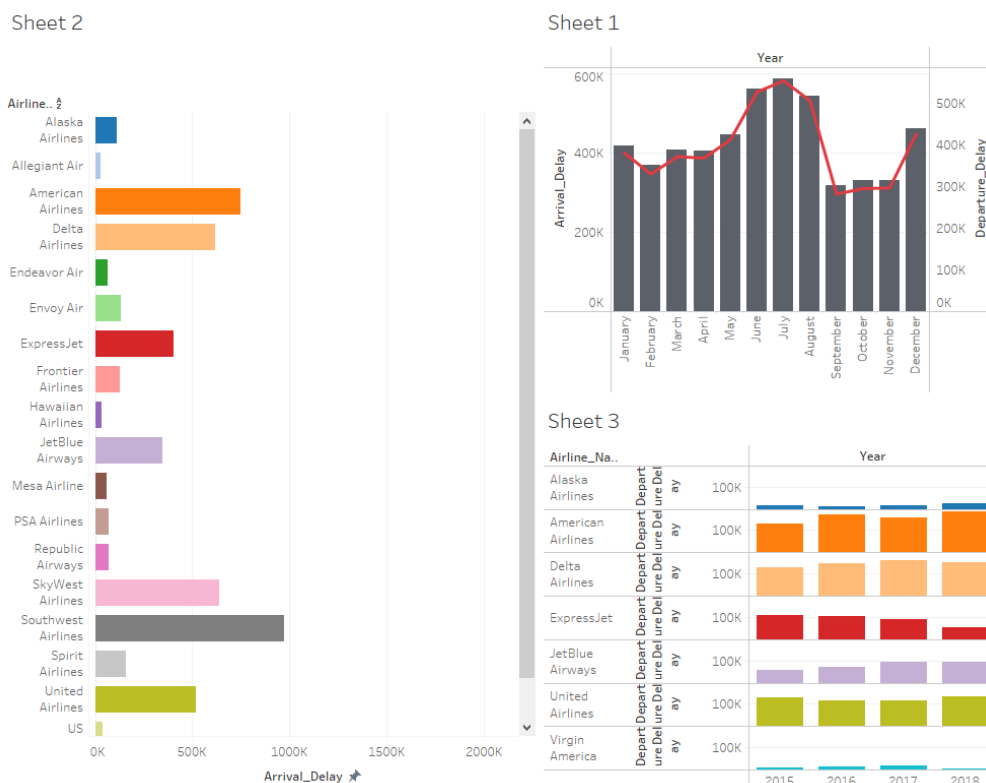
Colour Choices

We decided on a black and white colour scheme as the primary theme for our dashboard. We utilized contrasting colour to highlight different delays in the first graph. For arrival delays, we went with Dark Grey, while for departure delays, we went with Red. By selecting the Tableau Classic 20 colour palette for the airline names in the second graph, we enhanced the dashboard to stand out and draw in more viewers. In the third graph, we used the orange-gold colour palate to gather insights on airlines which have consistently poor service in terms of number of delayed departures and airlines which have improved over a period of 4 years in terms of their service.

Choosing Chart Type

1. Bar Graph and Line Graph – To show the flight's arrivals and departures, we used a bar and line chart. The arrival and departure delays are shown by the grey bars and red line, respectively. The graph also contains the months, showing which ones have the most arrival and departure delays. This graph is thorough and simple to comprehend, making it possible to identify the months with the highest number of departure and arrival delays.

2. Horizontal Bar Graph – The names of the airlines and the lengths of their arrival delays are represented by the horizontal bar graph. Those with the most delays include American Airlines and United Airlines.

Tools and Libraries Used

Data Pre-processing and exploration - We have used Pandas and Numpy libraries of Python and Jupyter Notebook (Software) to perform pre-processing and exploration.

Data Cleaning and Transformation – Again, we have used Python libraries for cleaning and transformation of data.

Visualisation – Tableau 2022.3 was used for creating dashboard and graphs.

The whole dashboard is designed in Tableau Desktop and published to tableau server.

Conclusion

The major goal of this analysis is to assist customers in selecting reliable airlines and the most efficient times to travel. After a thorough analysis and cleaning of our data, a dashboard was created for aircraft delays and cancellations which further aided in making an informed decision about airlines which consistently provided great service. We obtained the US aviation industry data from the public dataset website Kaggle. The dataset includes information about the time of departure and arrival, delays and cancellations, weather information, flight duration, and other columns. The cleaned data was then extracted, exported to CSV, and imported into Tableau. In order to compare arrival delay and departure delay over the course of a month, we designed a dashboard for airline delays and cancellations. We created a bar and line chart to display the arrivals and departures of the flight. The dashboard provides a clear and concise understanding and comparison between the airlines which make it more visually appealing and informative. Through our meticulous analysis we have found that Virgin America airline is the most improved and our top choice for the customers as it has less delay over the years.

References

https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

https://www.tableau.com/blog/bring-your-data-life-viz-animations