Flight Data Analysis

Project proposal by:

Asmaa Ismail 201600780
Anhar Hassan 201601171
Ahmed elgabry

I. Dataset description:

The Reporting Carrier On-Time Performance Dataset contains 200 million flight records from domestic flights in the US. The dataset contains some information about the flight's date and time, source, destination, and estimated delay with its reason if applicable. The complete dataset is stored as one CSV file with 200 million records and 218 feature columns, the detailed names and descriptions of which are listed in the dataset glossary here.

II. Basic analysis questions:

1- When is the best time of hour/day of week/month of year to fly to minimize delays?

For each time category; day, day of week, or time of year, we will group by each instance in that category, then calculate the average delay. We will then have the delays for each hour, day of week, and month.

2- Compare travel rates and average delays in 2020 during the pandemic against prior years.

We will start by counting the number of flights and calculating the average delay for every year. By plotting these statistics in a histogram against the year we can compare the rates before and after 2020 lockdown.

3- compare the number of flights and average delays before and after september 11, 2001 (the twin tower attack).

This requirement may be branched into three main sub-requirements:

- Analysis and comparison of flights number, average delay, and most frequent destinations/origins before and after 9/11.
- Analysis for the flights cancellation during the day, “there was a documentary saying that the air management officer cancelled ALL of them”.
- Deeper analysis at the states of the attack the days following it.
  All of these are straightforward, by counting the corresponding columns, and focusing on or grouping by duration of interest”days, hours” when needed.

4-How does the number of people flying between different locations change over time?

This requirement can be approached by counting the number of flights between each pair of states, and tracking how much it changes over time “over years for example” by aggregating by the time period.

5-Trying to detect cascading failures as delays in one airport create delays in others.

This can be approached by grouping flights by day, then choosing an airport, and making an analysis over the delay of flights in this airport vs. the delays of the flights after these flights in all other airports, in other words, we will create a new column only for the delays of the flights of the airport we choose, then do our analysis, this would show if there are any correlation between the delay in one airport and all the following flights in other airports.

6-The average delay across different seasons (eg, summer, winter) and which season is associated with the most delayed flights?

For this requirement we will first need to assign a season tag to each flight record (winter, summer, fall, spring) which we will infer from the month and day information. Then we will aggregate all the records over each season while averaging the entries in the delay columns. We will end up with a simpler dataframe with 4 records and two columns. Each record is a season and the average delay associated with it across all flights over all years.

7-The most popular destination.

This can simply be found by counting the unique occurrences of each city in the destination city column.

8- The busiest airport across all years.

We define the 'busiest' airport as the one that received, and dispatched the most flights for each given year. We start by grouping by the year to have a yearly record of all flight information. For each year we count the unique occurrences of each airport both in the source and destination columns. We then add the source and destination counts for each airport for every year to get a new value which we will call 'airport_traffic'. We then filter for the maximum traffic airport in every year and visualize the busiest airport and its number of flights against the years in a histogram.

9- ML Analysis: A regression model to estimate expected delay time given the time of month, week, day, source airport and destination airport.

The goal of this analysis step is to answer the question: 'If I book a flight from source X to destination Y on date dd/mm/yy and time H:M, how much estimated delay can I expect to encounter?'. This will be a simple regression model that takes these features into account to estimate an expected delay value.