Airline Data Challenge

Documentation

• Key Points:

The Data Analysis of this challenge has been done in Jupyter notebook with use of different libraries like Pandas, NumPy and Matplotlib and Plotly libraries. Please install required libraries on your local system in command prompt or in the Jupyter notebook using following command **"! pip install library name"**

The order of the project is done in Single Notebook:

- Capital_one_Data_challenge.ipynb

Order followed:

1. Performed the Data Analysis individually for Airport codes, Flights and Tickets dataset.
2. Merging all the datasets into one using selected columns.
3. Performing the Visualizations for given tasks to identify insights and trends.

The notebook contains documentation for each tab and insights in the analysis. Please reach out directly at **nagarapu37@gmail.com** for any concerns.

Assumptions and Considerations:

- I have filtered the data to medium and large airports and considered only US country round trip flights.
- Making assumptions on cost of fuel, oil, maintenance, crew which costs 8$ per mile and depreciation, insurance and other costs 1.18$/mile. And also made airport operational cost for medium and large is 5000$ and 1000$ and considering round trips as double charged.

- And for arrival and delay arrival the first 15 minutes are free and then 75$ per minute. And considered the flight can accommodate maximum of 200 passengers.
- Considered the baggage fees is 35$ per bag. Assuming 50 % of passengers to check an average of 1 bag per flight and charged 70$ per round trip flight. Disregarded the seasonal effects on ticket prices.
- Considering the five recommended round trip routes based on the profit, revenue, fuel, insurance, depreciation, baggage, ticket cost and occupancy rate.
- Considering the number of round trips required to breakeven the upfront costs for each of the recommended route using the revenue column.

Metrics Created:

- **Latitude and Longitude:** Created these columns by splitting the Coordinates column into two for map visualizations.

- **Airport operational costs**: This column is used to calculating the operational costs for medium and large airports which is 5000$ and 10000$ for the airports.

- **Persons travelled:** This column is used to find the number of persons travelled in each flight by using Occupancy column. The occupancy is number of for each flight the maximum number of passengers is 200.
  Number of persons travelled= (occupancy rate*maximum number of person)/100

- **Other Cost:** This other cost sums all of the costs Fuel, Oil, Maintenance, Crew which is 8$/mile and also Depreciation, cost, insurance which costs 1.18$/mile.
- **Arrival delays operational costs:** Calculates the arrival delays if it is delayed by 15 minutes then no cost or more then per each minute it costs around 75$/minute. Calculated using if else statement.

- **Departure delay Operational Costs:** Calculates the departure delays if it is delayed by 15 minutes then no cost or more then per each minute it costs around 75$/minute.

- **Baggage Cost: (Number of persons travelled in each flight/2) *35** $ assuming 50 % of people pays the average of 1 bag for one trip which is 35 $.

- **Itinerary fare:** calculated the cost per each passenger.

- **Ticket cost:** (Number of passengers travelled * itinerary fare) to get the total cost for each flight

- **Total expenses:** Calculated the total expenses calculated for each flight by sum of other cost, departure delay operational cost, arrival delay operational cost and Airport operational cost.
  Expenses= other cost+ departure delay operational cost+ arrival delay operational cost+ airport operational costs.
- **Total Revenue:** Calculated the revenue by sum of Baggage cost and Tickets cost.
  Revenue=Baggage+ Tickets
- **Final profit:** Calculated the Final profit by difference of revenue and expenses
  Profit= Revenue-Expenses

- **Single trip Revenue:** Calculating the single trip Revenue for each flight.

  Single trip Revenue= Total Revenue/Total number of flights
- **Number of roundtrips:** It calculates the number of round trips required to breakeven the upfront costs.

Usefulness of these metrics:

- **Ticket cost:** This tells us how much we gain from the revenue from the tickets of each flight.
- **Baggage cost:** This tells us how much revenue we gain from the baggage of flights.
- **Total expenses:** This tells us the how much expenses are happening on the airport operations cost, arrival delay operation costs, departure delay cost and other costs which include fees, insurance oil. And can evaluate and analyze to reduce the total expenses
- **Total Revenue:** This metric gives the assess to how much revenue is gained from baggage and tickets costs for each flight.
- **Total Profit:** This metric is a good metric gives the total overall profit for each flight and can recommend more number of flights can be proposed in these routes based on the profits.

Data Quality issues and Processing:

After performing the data analysis, I feel that the quality issues that I have faced is

- **Accuracy**: The information is not accurate in every detail they are many outliers in the data.
- **Completeness**- This data has the missing values.
- **Consistency -**The data should have data format as expected and can be cross referenceable with the same results
- **Conformity-** some of the columns of integers are in string format.
- **Hidden data-** It has more hidden data where we need to extract it from other columns

Key Points:

- Select the different columns for better interesting analysis and recommendation.
- Handling the missing values in data and doing the necessary imputations.
- Handling the outlier in each column and removing the null values using Inter Quartile Range
- Merging the data using different columns.
- Making the data as per needs of visualizations.

Methodology:

Airport codes data:

The concern with this data it has outliers in the elevation feet column. By using the box plot and finding the Inter Quartile range (IQR) and dropping all the outliers present in the column and lot of null values and missing column values of IATA code and country code. So dropped those columns. Filling the missing values in elevation feet using median value of elevation column. And for coordinate column splitting the latitude and longitude to separate columns for making the map visualizations. And filter the type of airport based on the medium and large airports.

Flights data:

In this data it has a lot of null missing values and most concerned with a lot of outliers in many columns such as Arrival delay and departure delay. And also missing values in distance columns and dropping duplicates from the dataset. Calculating the number of the passengers travelled in the flight based on the occupancy rate and dropping sum of null values which are not required and finding the other cost by summing up of cost of fuel, oil, maintenance, crew depreciation, insurance which cost total of 9.18$ per mile and calculating the operational costs for arrival delay and departure delay if greater than 15 minutes it charges the 75$ per minute. And finally remove the columns that are not required for the analysis

Tickets data:

At first filtering this data on 1 because we need only for round trip. And splitting the some of the columns to remove 0 after point and itinerary columns replaced space with dollar and replacing the null values and converting it into int and found the outliers using boxplot on itinerary column and filter using the lower bound and upper bound value

Merging data:

First filtering the data on the country US because need to perform analysis on US round trip flights. And finding the flight number based on the last four digits of itinerary number and joined the tickets and flights data using the fields on left is origin, flight number and destination and right is origin, op carrier flight number and destination from flights data using inner join. And merged the airport codes and new data frame based on municipality which is cities present in airport codes data and origin city name in new data frame and dropped the duplicates. And this is the final dataset to be used for the problem statement.