Girija Mangrulkar

Capital One- Airline Data Challenge

This data challenge gave me an opportunity to showcase my skills and abilities that align with how the Data Analyst at Capital one work. It helped me to upscale myself in the following areas:

1. Building Mindset: I built the complex queries for data transformation and manipulation which helped me to think in broader perspective when it comes to vast and extensive data.

2. Data Management: As the data was given in csv format, it was easy to manage using it in python, creating data frames and perform data and business analysis on it.

3. Business Intelligence: As it was suggested to do data visualizations in Python, I have created histogram for continuous data and bar graphs which gives answers to the questions asked. Also, I have performed Exploratory Data Analysis in Tableau to showcase my expertise without doing extensive data manipulation as it was mentioned before. I have built 3 dashboards for exploratory data analysis on 3 different datasets given.

Metadata

1. **flights_airport_codes:** This is created when flights and airport_codes dataset are merged on Origin from flights and IATA_CODE from airport_codes dataset.

2. **updated_dataset:** The flights_airport_codes data was merged on tickets data using - 'FL_DATE', 'ORIGIN', 'DESTINATION', 'OP_CARRIER' from flights_airport_codes data and 'YEAR', 'ORIGIN', 'DESTINATION', 'REPORTING_CARRIER' - from tickets data using inner join.

3. **dest_orig:** This is a combination of the origin and destination from flights dataset.

4. **clean_data:** This is created after dropping columns which has maximum null values.

5. **size_filter:** This is created to filter out large airport and medium airport as it is asked in the question.

6. **co**mbined_data: This is created to group the airport type and the distance by a flight.
7. **ROUND_TRIP:** This is created by concatenating ‘ORIGIN’ and ‘DESTINATION’ columns in the flight dataset. This column will show the round-trip which is 1 from one point to other point.

8. **updated_origin_dest:** This is created to store the result of the round trip of medium and large airports.

9. **TOTAL_COST:** This is the total cost incurred by the airline. This was obtained using max_passengers, delay_cost, baggage_check, baggage_fee.

10. **PROFIT:** This is (total revenue - total cost).

Data Cleaning and pre-processing stepwise procedure:

1. **Data merging:**
   flights_airport_codes: This is created when flights and airport_codes dataset are merged on Origin from flights and IATA_CODE from airport_codes dataset.

2. **Datatype Change:** I have changed the datatype of FL_DATE and YEAR into date data type.

3. **Creation of new column:**
   dest_orig: This is a combination of the origin and destination from flights dataset.

4. **Perform join operations:** To merge the data with ticket dataset I have used inner join.

5. **Conversion of all wrong numeric columns into correct format and display using histograms such as DISTANCE, AIR_TIME, ITIN_FARE.**

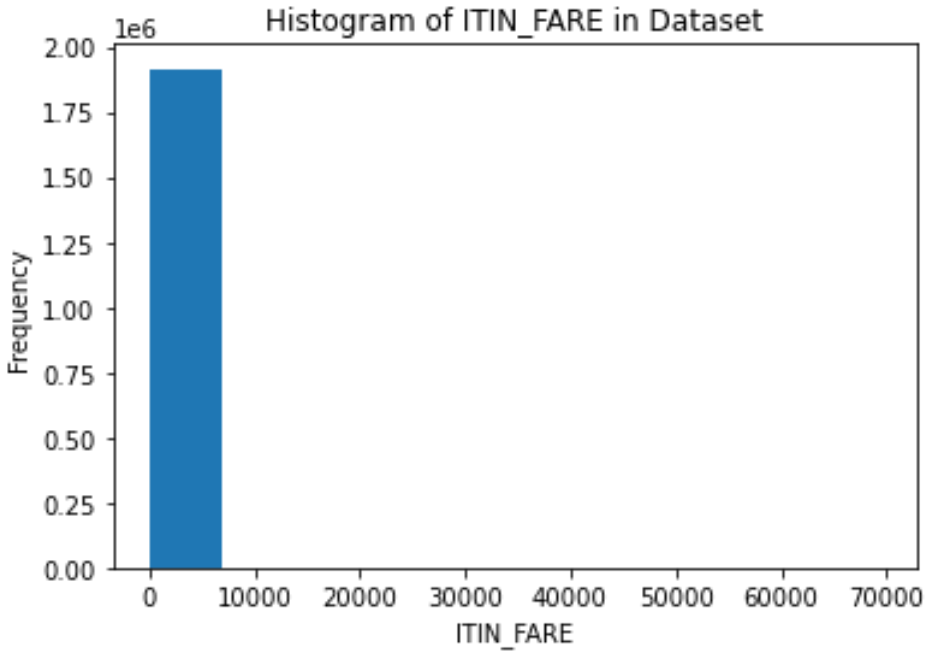

Fig 1. Itinerary fare histogram

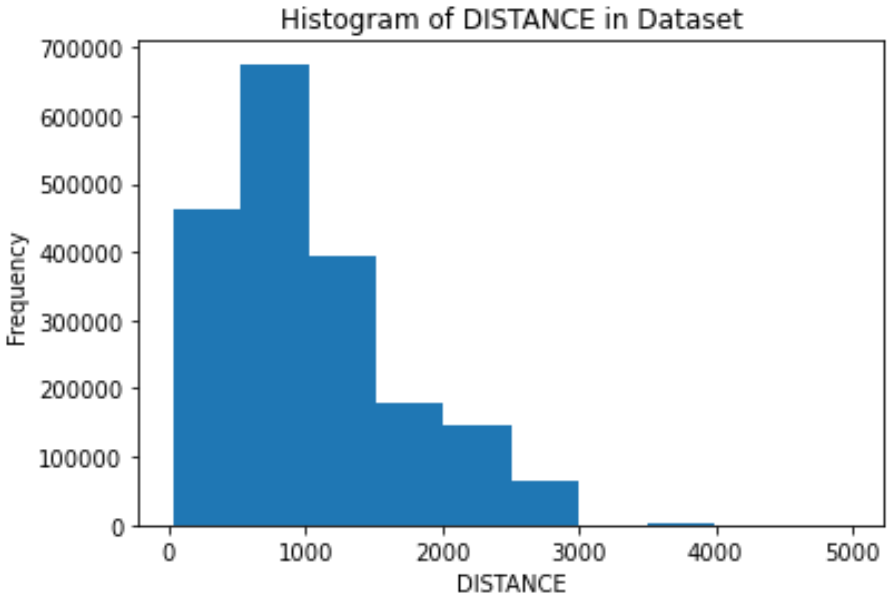

Fig 2. Distance histogram.

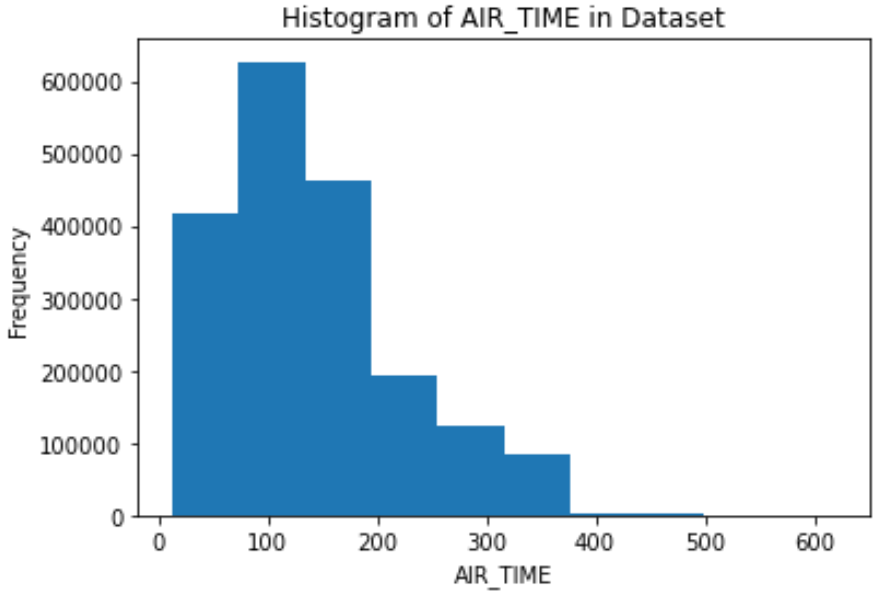

Fig 3. Airtime histogram

6. **Check outliers:** I have checked the outliers for the DEP_DELAY and ARR_DELAY field and displayed it using boxplot.

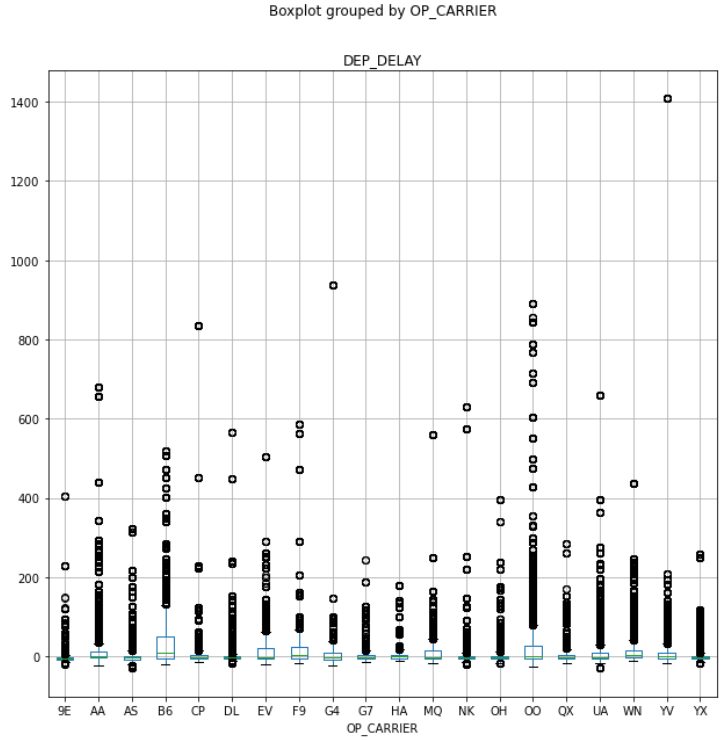

Fig 4. Outliers for departure delay

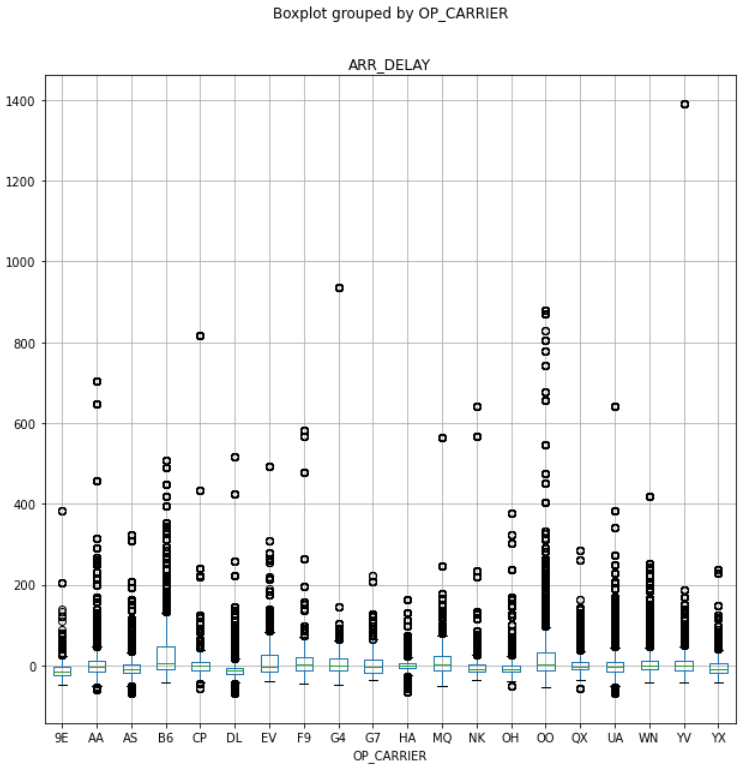

Fig 5. Outliers for arrival delay

7. There are significant number of outliers in the ARR_DELAY and DEP_DELAY columns. I have excluded the values that fall outside the 1st to 99th percentile range during our data analysis and I have replaced those outliers with the median value. For the tickets dataset, I have used a slightly narrower population range, specifically the 1st to 95th percentile.

```
In [102]: #removing outliers and replace them with median
          remove_outliers(updated_dataset, ['ARR_DELAY', 'DEP_DELAY'], 0.01, 0.99)

          #removing outliers and replace them with median
          remove_outliers(updated_dataset, ['ITIN_FARE'], 0.01, 0.95)
```

Fig 6. Removing outliers and replacing with median

8. Creating a boxplot after removing outliers for checking

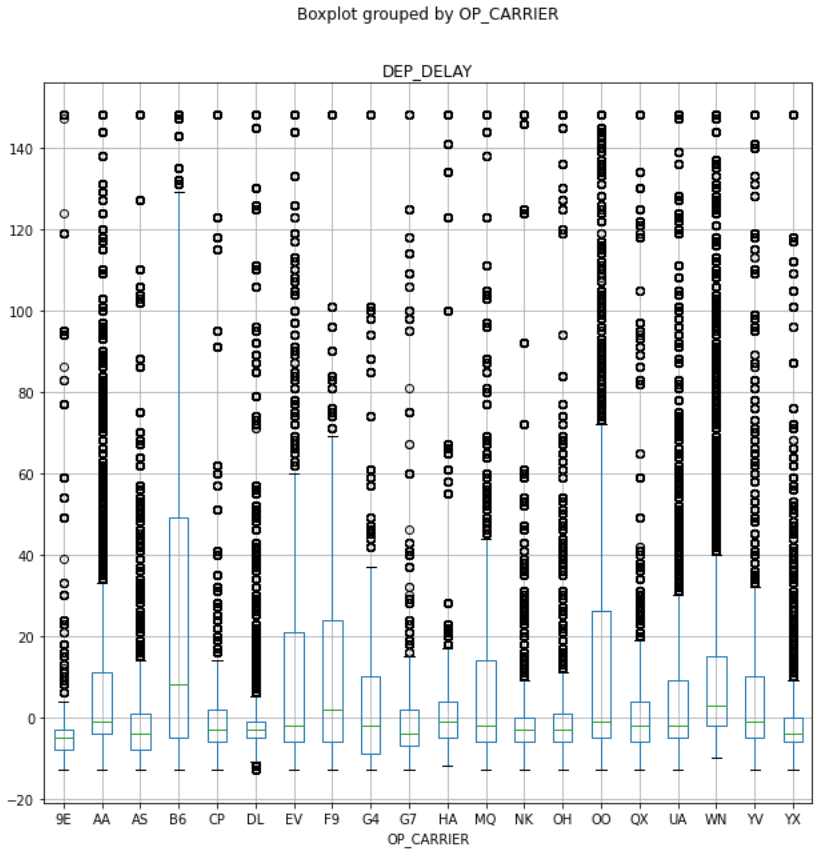

Fig 7. Outliers for departure delay

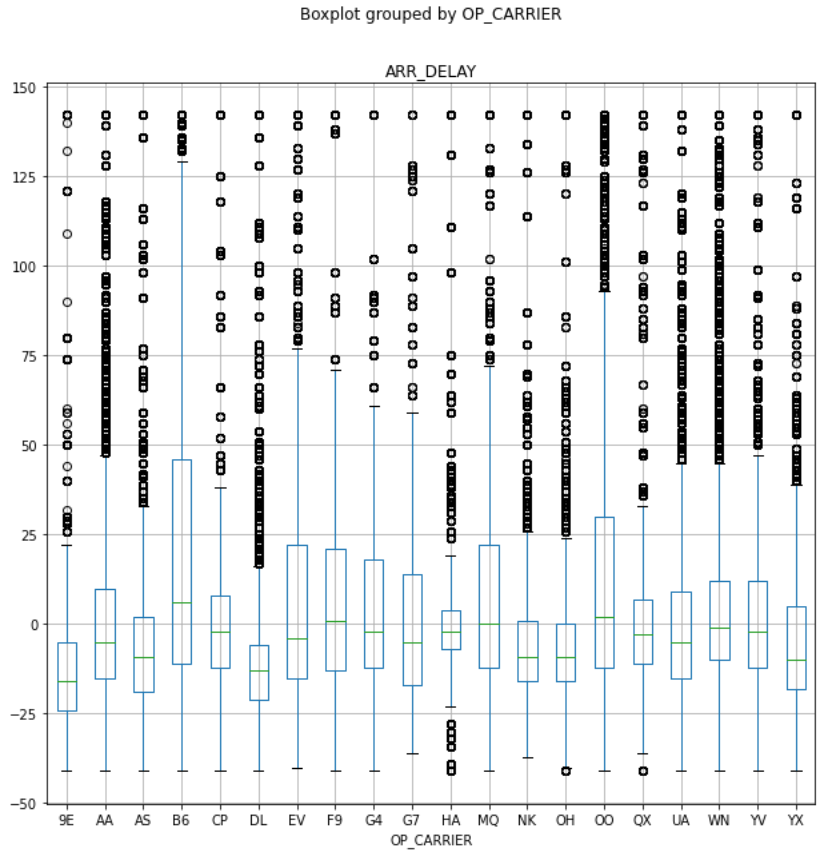

Fig 8. Outliers for arrival delay

9. Checking for null values and duplicate values and dropping them

```
In [106]: #Checking for Null Values
          updated_dataset.isnull().sum()
```

```
Out[106]: FL_DATE                 0
          OP_CARRIER              0
          TAIL_NUM              148
          OP_CARRIER_FL_NUM       0
          ORIGIN_AIRPORT_ID       0
          ORIGIN                  0
          ORIGIN_CITY_NAME        0
          DEST_AIRPORT_ID         0
          DESTINATION             0
          DEST_CITY_NAME          0
          DEP_DELAY               0
          ARR_DELAY               0
          CANCELLED               0
          AIR_TIME            11970
          DISTANCE                0
          OCCUPANCY_RATE          0
          TYPE                    0
          NAME                    0
          ELEVATION_FT        15477
          CONTINENT         1923238
          ISO_COUNTRY             0
          MUNICIPALITY        15477
          IATA_CODE               0
          COORDINATES             0
          dest_orig               0
          ITIN_ID                 0
          YEAR                    0
          QUARTER                 0
          ORIGIN_COUNTRY          0
          ORIGIN_STATE_ABR        0
          ORIGIN_STATE_NM         0
          ROUNDTRIP               0
          REPORTING_CARRIER       0
          PASSENGERS           3832
          ITIN_FARE               0
          dtype: int64
```

```
In [107]: #Drop the columns which has maximum null values and are not required for further analysis
          drop = ['ELEVATION_FT', 'CONTINENT', 'MUNICIPALITY']
          clean_data = updated_dataset.drop(columns=drop)
```

```
In [108]: clean_data=clean_data.dropna()
          clean_data=clean_data.drop_duplicates()
```

Fig 9. Null values and duplicate

Problem Statement

Q1. The 10 busiest round-trip routes in terms of number of round-trip flights in the quarter.

To solve this question, I have performed group by operation on 'ROUND_TRIP' and stored in a new data frame. After that I have counted the total number and arranged them in descending order to get the top 10 busiest routes.
Bar graph is used to visualize the data.

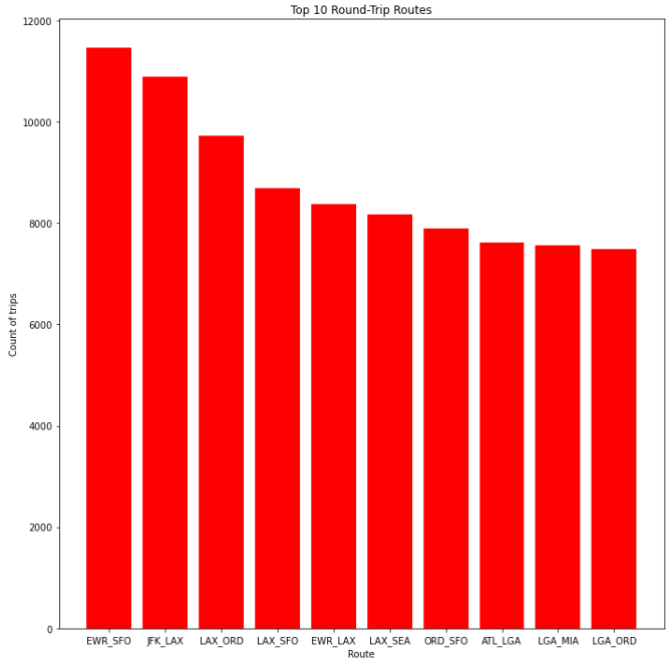

Fig10. Q1: Top 10 busiest routes

From the above data we can see that EWR_SFO has the maximum round trips.

Q2. The 10 most profitable round-trip routes (without considering the upfront airplane cost) in the quarter. Along with the profit, show total revenue, total cost, summary values of other key components and total round-trip flights in the quarter for the top 10 most profitable routes. Exclude canceled flights from these calculations.

I have calculated the Total Revenue, Total cost from the given data and I have grouped that data on ROUND_TRIP. I have calculated Total Profit which is (Revenue - Cost). This data is grouped together and then sorted in descending order to get Top 10 profitable routes based on Total cost, Itinerary Fare and Profit.

Bar graphs are used to show the profitable routes for 3 different key components.

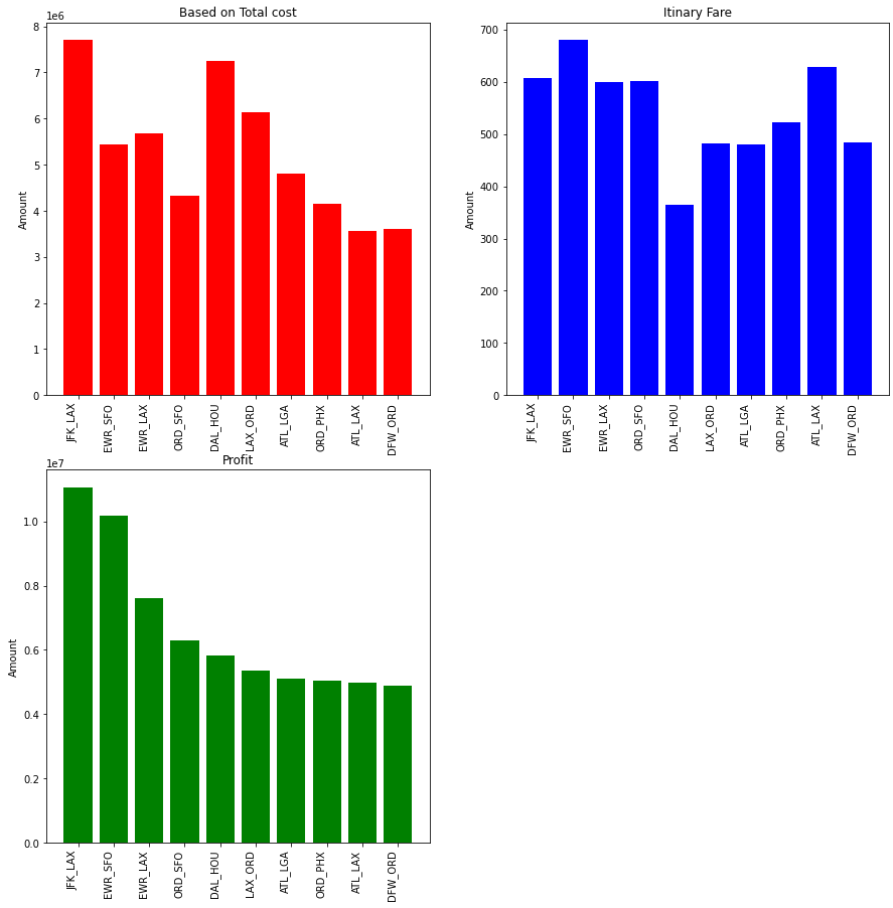

Fig 11. Q2. Top 10 profitable routes

From the above data we can see that

- JFK_LAX has the maximum cost.
- EWR-SFO has maximum itinerary fare.
- JFK_LAX has maximum profit.

Q3. The 5 round trip routes that you recommend to invest in based on any factors that you choose.

I have given the best 5 round trip routes considering Profit, Itinerary Fare, and total cost.

```
Top 5 round trip routes to invest in based on profit:
     ROUND_TRIP      PROFIT
1708    JFK_LAX  11050888.0
1346    EWR_SFO  10179912.0
1314    EWR_LAX   7623084.0
2179    ORD_SFO   6284580.0
868     DAL_HOU   5837256.0

Top 5 round trip routes to invest in based on Itinary Fare:
     ROUND_TRIP   ITIN_FARE
141     ATL_FSD  981.000000
68      ANC_DEN  928.352941
1480    GUM_HNL  918.846154
775     CLT_TRI  895.000000
1065    DEN_SUN  884.000000

Top 5 round trip routes to invest in based on total cost:
     ROUND_TRIP  TOTAL_COST
1501    HNL_OGG  12390560.0
1708    JFK_LAX   7701735.0
1834    LAX_SFO   7318600.0
868     DAL_HOU   7244415.0
1498    HNL_LIH   6843520.0
```

Fig 12. Q3. Top 5 best round trips

From the above data we can see that,

- JFK-LAX is giving maximum profit.
- ATL_FSD is giving best Itinerary.
- HNL-OGG is giving the best cost.

Q4. The number of round-trip flights it will take to breakeven on the upfront airplane cost for each of the 5 round trip routes that you recommend. Print key summary components for these routes.

I have taken the top 5 round trips based on profit and calculated the number of trips to breakeven by total cost of plane divided by profit. I have displayed that data in ascending order.

Bar graph is used to display the breakeven.

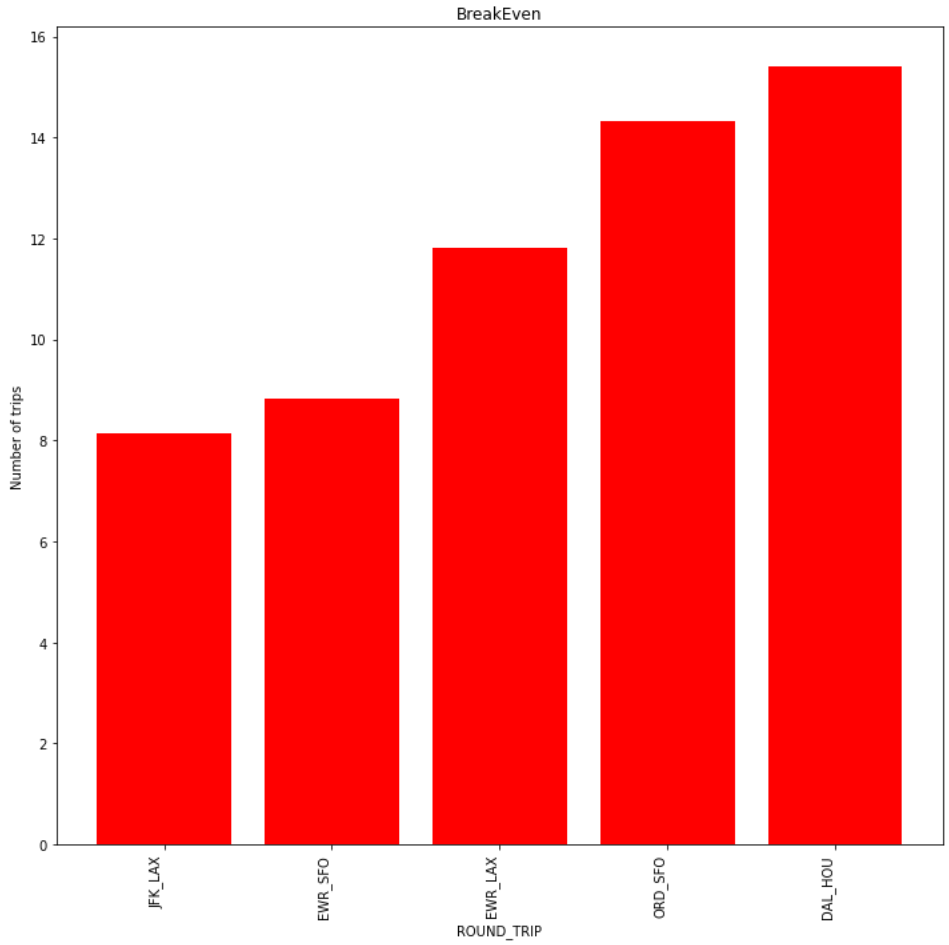

Fig 13. Q4. Round trips with their break evens

From the above data we can see that:

JFK_LAX has minimum break even as it has maximum profit.

Q5. Key Performance Indicators (KPI's) that you recommend tracking in the future to measure the success of the round-trip routes that you recommend.

1. **Profitability:** Monitor the profitability of each round-trip route regularly.
2. **Performance:** Track the performance of each round-trip route, measuring the percentage of flights that depart and arrive on time.
3. **Customer Retention:** Track the percentage of customers who choose to fly the same route again.
4. **Cancellations and Delays:** Track the number of flight cancellations and delays for each route.
5. **Number of Round Trips:** An increase in the number of rounds trips can be an indicator of success due to the revenue received for each passenger for each flight.

Future Scope:

I would have liked to further exam the following more in detail:

- Delayed flights (Departure and arrival): How it can be reduced as much as possible and the level of effect it has on the revenue, cost, and profit.
- Cancelled Flights: How it can be reduced and the level of effect it has on the revenue, cost, and profit.
- Routes with high occupancy rate and low income and how it can be maximized.