Capital One Airline Data Challenge.

by Sossou Simplice Adjisse

February 26, 2024

Context

A new airline company is looking to enter the US domestic market and punctuality is a big part of the company's brand image. I am asked to:

1. find out the 10 busiest round trip routes in terms of a number of round trip flights in the quarter without the canceled flights.
2. find out the 10 most profitable round trip routes (without considering the upfront airplane cost) in the quarter and show profit, 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 without canceled flights.
3. recommend 5 round trip routes to invest in
4. calculate the number of round trip flights it will take to breakeven on the upfront airplane cost for each my recommended 5 round trip routes and print key summary components for them
5. recommend Key Performance Indicators (KPI's) to track in the future to measure the success of the recommended round trip routes.

Datasets : Quality, Cleaning, Restrictions, and Limitation

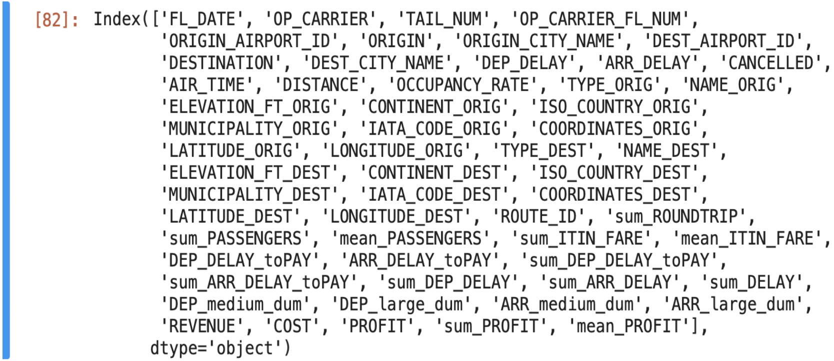
- Datasets $\Longrightarrow$ flights, tickects, airport_codes
- All the columns have the intended values
- Some columns have the wrong format $\Longrightarrow$ I converted to right format.
- Outliers in some of the numeric columns $\Longrightarrow$ I kept them
- Numeric columns with missing values $\Longrightarrow$ I replaced them with medians
- Invalid strings (e.g., airport codes) $\Longrightarrow$ I dropped them
- Duplicates with respect to all the columns $\Longrightarrow$ I dropped them
- Restrictions $\Longrightarrow$ dropped cancelled flights, non-roundtrips, non-US airports
- Limitation $\Longrightarrow$ lack of information to link tickets to their flights

Data: Aggregating and Merging

After cleaning and putting the necessary restrictions on the data:

- merge the flights and airport codes on ORIGIN and IATA_CODES as keys, respectively and inner join ⟹ ensured origin of flights are within the US
- merge the flights and airport codes on DESTINATION and IATA_CODES as keys respectively an inner join ⟹ ensured destination of flights are within the US
- create ROUTE_ID for the routes ⟹ routes' unique identifier in flights dataset
- merge the tickets and airport codes on DESTINATION and IATA_CODES as keys respectively and inner join ⟹ ensured destination of tickets are within the US
- create ROUTE_ID for the routes ⟹ routes' unique identifier in tickets dataset
- aggregate PASSENGER, ITIN_FARE , and ROUTRIP at ROUTE_ID level
- merge back to the previous combined flights and airport codes on ROUTE_ID as key and inner join ⟹ final combined data

Final Data



```
[82]: Index(['FL_DATE', 'OP_CARRIER', 'TAIL_NUM', 'OP_CARRIER_FL_NUM',
             'ORIGIN_AIRPORT_ID', 'ORIGIN', 'ORIGIN_CITY_NAME', 'DEST_AIRPORT_ID',
             'DESTINATION', 'DEST_CITY_NAME', 'DEP_DELAY', 'ARR_DELAY', 'CANCELLED',
             'AIR_TIME', 'DISTANCE', 'OCCUPANCY_RATE', 'TYPE_ORIG', 'NAME_ORIG',
             'ELEVATION_FT_ORIG', 'CONTINENT_ORIG', 'ISO_COUNTRY_ORIG',
             'MUNICIPALITY_ORIG', 'IATA_CODE_ORIG', 'COORDINATES_ORIG',
             'LATITUDE_ORIG', 'LONGITUDE_ORIG', 'TYPE_DEST', 'NAME_DEST',
             'ELEVATION_FT_DEST', 'CONTINENT_DEST', 'ISO_COUNTRY_DEST',
             'MUNICIPALITY_DEST', 'IATA_CODE_DEST', 'COORDINATES_DEST',
             'LATITUDE_DEST', 'LONGITUDE_DEST', 'ROUTE_ID', 'sum_ROUNDTRIP',
             'sum_PASSENGERS', 'mean_PASSENGERS', 'sum_ITIN_FARE', 'mean_ITIN_FARE',
             'DEP_DELAY_toPAY', 'ARR_DELAY_toPAY', 'sum_DEP_DELAY_toPAY',
             'sum_ARR_DELAY_toPAY', 'sum_DEP_DELAY', 'sum_ARR_DELAY', 'sum_DELAY',
             'DEP_medium_dum', 'DEP_large_dum', 'ARR_medium_dum', 'ARR_large_dum',
             'REVENUE', 'COST', 'PROFIT', 'sum_PROFIT', 'mean_PROFIT'],
            dtype='object')
```

Figure 1

Top 10 Busiest Routes in Terms of Total Round Trip

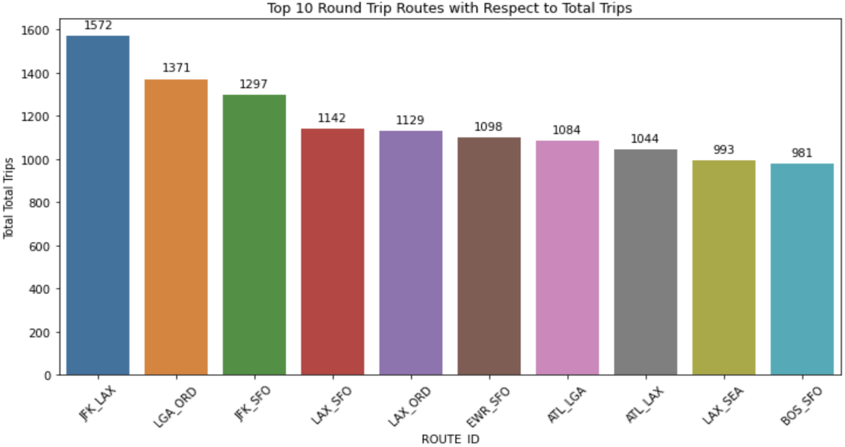


Figure 2

Top 10 Profitable Routes

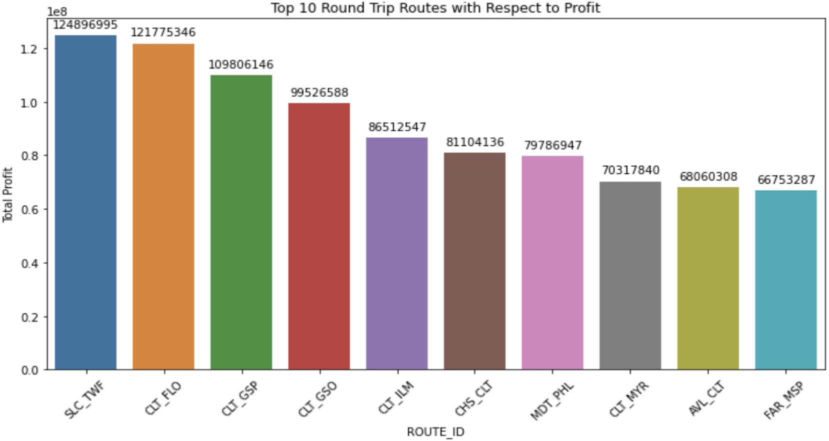


Figure 3

Top 5 Routes I Recommend the Company Invest in

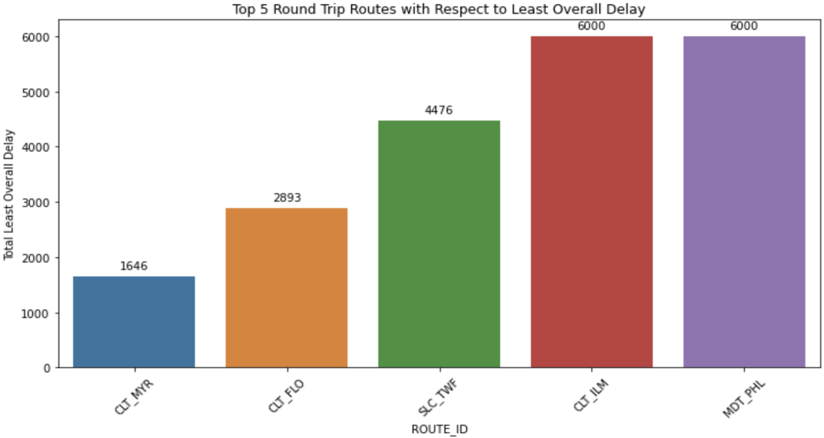
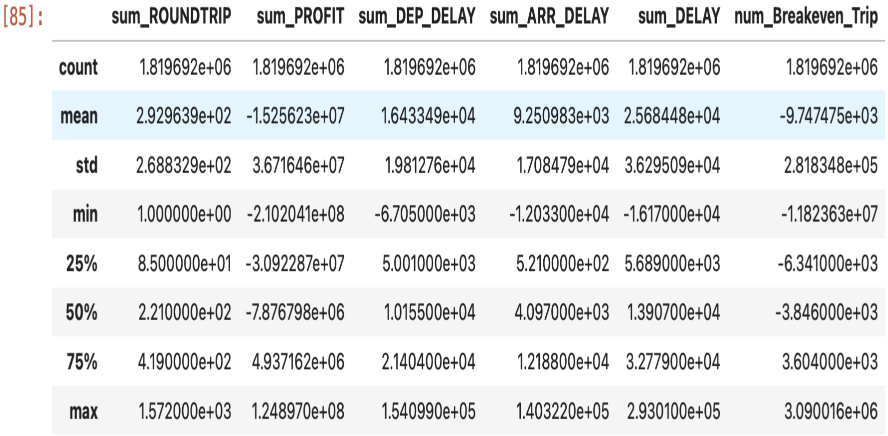


Figure 4

Summay Statistics of Key Performane Indicators



[85]:

	sum_ROUNDTRIP	sum_PROFIT	sum_DEP_DELAY	sum_ARR_DELAY	sum_DELAY	num_Breakeven_Trip
count	1.819692e+06	1.819692e+06	1.819692e+06	1.819692e+06	1.819692e+06	1.819692e+06
mean	2.929639e+02	-1.525623e+07	1.643349e+04	9.250983e+03	2.568448e+04	-9.747475e+03
std	2.688329e+02	3.671646e+07	1.981276e+04	1.708479e+04	3.629509e+04	2.818348e+05
min	1.000000e+00	-2.102041e+08	-6.705000e+03	-1.203300e+04	-1.617000e+04	-1.182363e+07
25%	8.500000e+01	-3.092287e+07	5.001000e+03	5.210000e+02	5.689000e+03	-6.341000e+03
50%	2.210000e+02	-7.876798e+06	1.015500e+04	4.097000e+03	1.390700e+04	-3.846000e+03
75%	4.190000e+02	4.937162e+06	2.140400e+04	1.218800e+04	3.277900e+04	3.604000e+03
max	1.572000e+03	1.248970e+08	1.540990e+05	1.403220e+05	2.930100e+05	3.090016e+06

Figure 5

Top 10 Leats

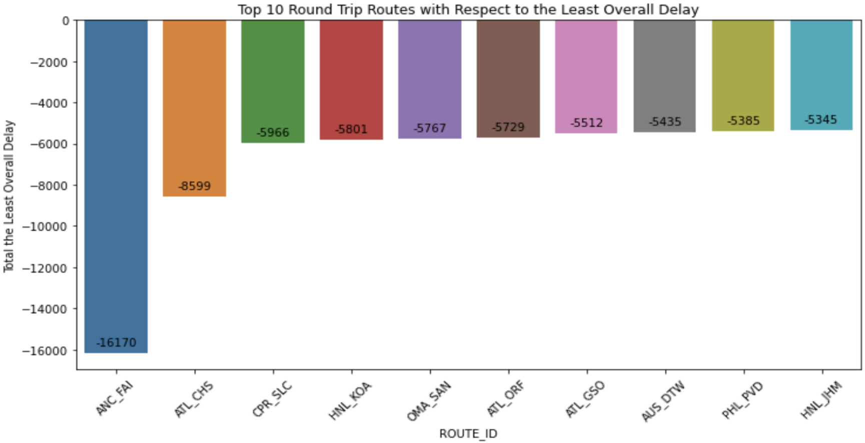


Figure 6

Top 5 Routes I Recommend the Company Invest in

They are from the 10 top most profitable routes in Figure 3

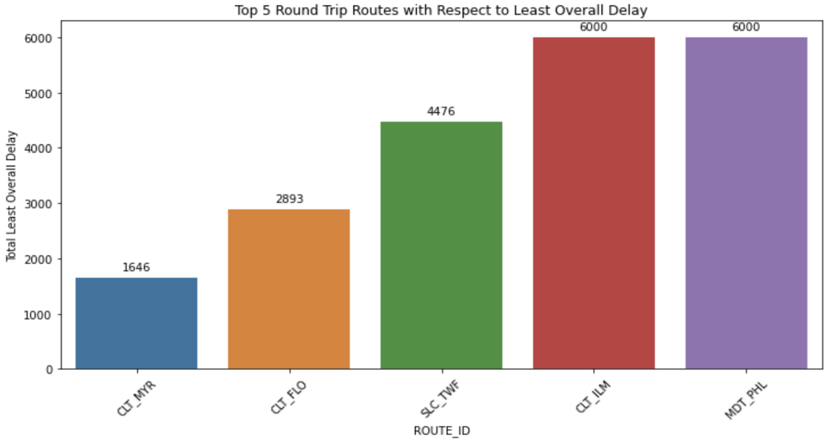


Figure 7

Key things to Track in the Future

- Competitions, entrance cost, opportunity cost, comparative advantages
- Most profitable routes which have the least delay
- Delay metrics:
    - departure delay
    - arrival delay
    - Overall delay