Documentation

CAPITAL-ONE

(DATA CHALLENGE)

TABLE OF CONTENTS

1. DATA IMPORTS AND QUALITY CHECKS

We have been given three data files and the metadata for each file. The files provide to me are as follows: Flights.csv, Airport_codes.csv, Tickets.csv. After importing the files there are a couple of Quality checks that I performed. Let's start by the first one. Technology used is **PYTHON** for visualization and as well as data analysis. I will be sharing the code in Jupyter notebook as well.

1) **Check for Duplicates**

From this we found out that there are no duplicates in the Airport_codes.csv but there were duplicates in the other two files. Flights.csv had 4410 duplicates and Tickets.csv had 47564 duplicates.

2) **Removing Unnecessary Columns to have a good analysis**

To have the best analysis we need to keep only the important pieces and discard other pieces of information. I am not keeping the reporting_carrier as well because our goal is not to find out the competition. I removed year and quarter as we only had one quarter data.

FILE	**COLUMNS REMOVED**
Flights.csv	'FL_DATE','OP_CARRIER_FL_NUM','TAIL_NUM','ORIGIN_AIRPORT_ID','DEST_AIRPORT_ID'
Tickets.csv	'ITIN_ID','YEAR','QUARTER','ORIGIN_COUNTRY','ORIGIN_STATE_ABR','ROUNDTRIP'
Airplane_codes.csv	'NAME','CONTINENT','ISO_COUNTRY','MUNICIPALITY','COORDINATES','ELEVATION_FT'

3) **Type Casting some columns with object values to int or float**

There were some inconsistencies in data. ITIN_FARE column had somewhere $$ symbol attached to some of its values. Removed $$ symbols that otherwise the columns don't convert to int. Similarly, AIR_TIME and DISTANCE had some string values in its rows such as TWENTY or HUNDRED which does not make sense. Removed them and replaced them with null. Later we will deal with null values.

4) **Making Changes to columns ARR_DELAY, DEP_DELAY and OCCUPANCY_RATE**

We are negating 15 minutes from Dep_Delay and Arr_delay column in flights data. In case subtraction is negative replacing it with 0. This would be very useful in later calculations. Similarly multiplying occupancy_rate by 200 as all the flights are of 200 capacities. We don't want to do aggregation of data and then do the steps.

5) **Dealing with Outliers**

We are dealing with outliers first, then we would deal with null value imputation. The order does not matter much as we would impute nulls with medians later. Here are the screenshots attached for the boxplots of different data files.

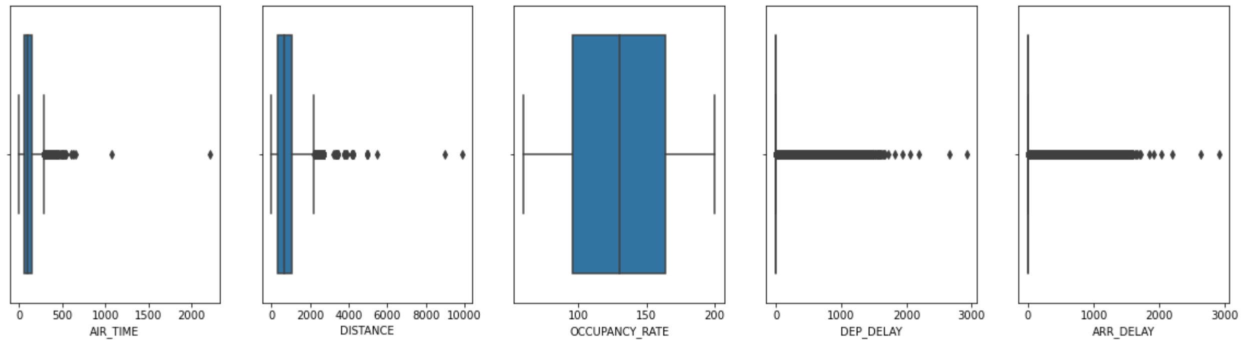


Figure 1 Boxplots for flight data

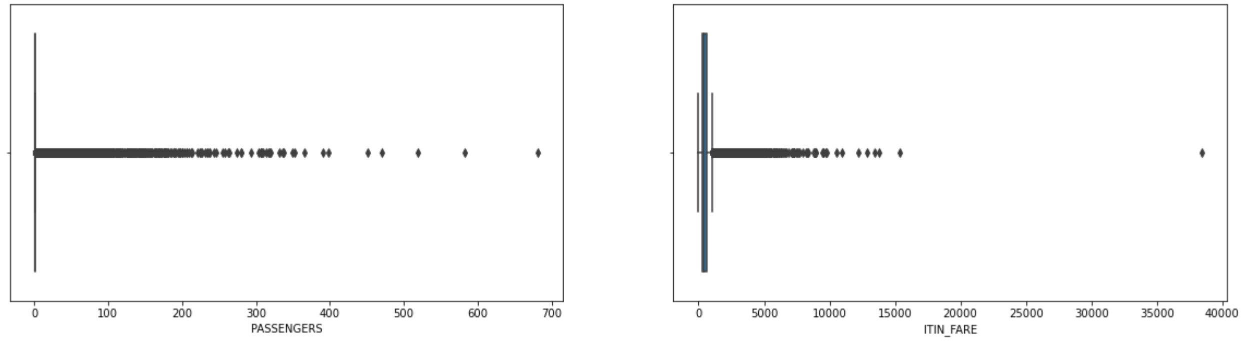
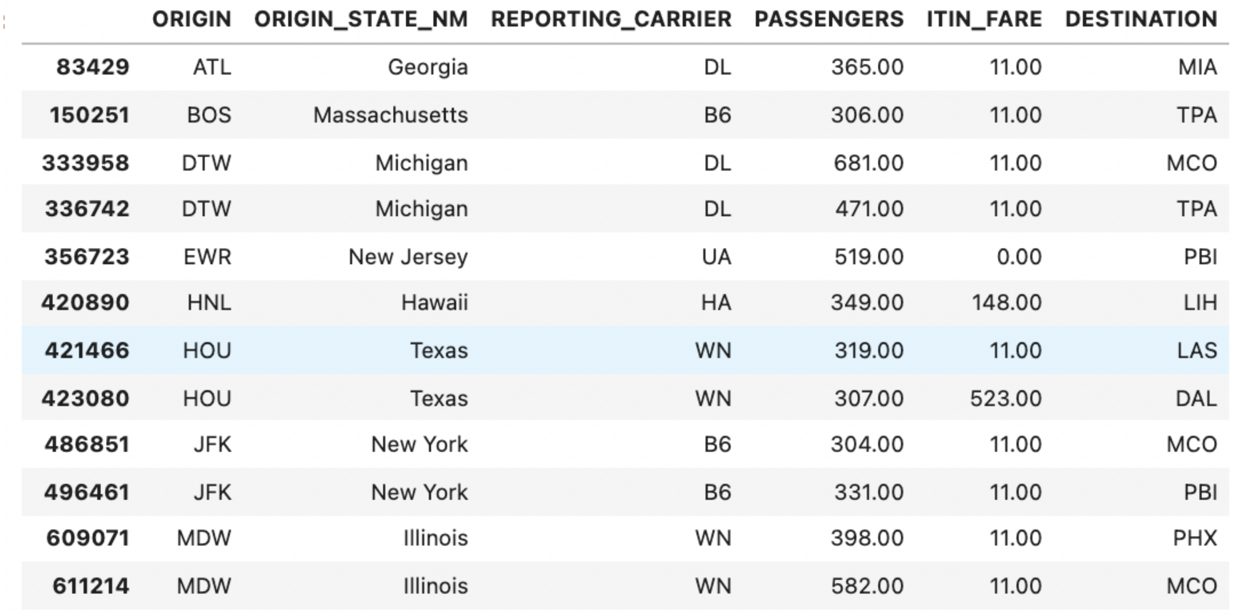


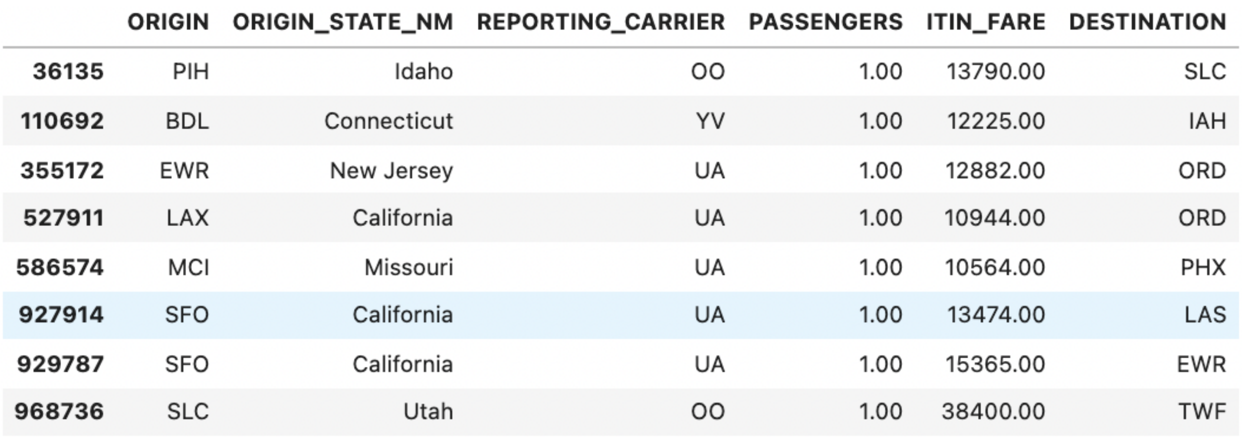
Figure 2 Boxplots for ticket data

Anomalies that I found out. The first screenshot below shows us that there are tickets which have more than 200 passengers. It is told in the file that ideal capacity is 200 passengers for a plane, and it is told to use occupancy rate for baggage calculation. And the ITIN_FARE is also very low per passenger



	ORIGIN	ORIGIN_STATE_NM	REPORTING_CARRIER	PASSENGERS	ITIN_FARE	DESTINATION
83429	ATL	Georgia	DL	365.00	11.00	MIA
150251	BOS	Massachusetts	B6	306.00	11.00	TPA
333958	DTW	Michigan	DL	681.00	11.00	MCO
336742	DTW	Michigan	DL	471.00	11.00	TPA
356723	EWR	New Jersey	UA	519.00	0.00	PBI
420890	HNL	Hawaii	HA	349.00	148.00	LIH
421466	HOU	Texas	WN	319.00	11.00	LAS
423080	HOU	Texas	WN	307.00	523.00	DAL
486851	JFK	New York	B6	304.00	11.00	MCO
496461	JFK	New York	B6	331.00	11.00	PBI
609071	MDW	Illinois	WN	398.00	11.00	PHX
611214	MDW	Illinois	WN	582.00	11.00	MCO

Another Anomaly is that there are some occurrences where the ticket price is very high. Here is the screenshot below



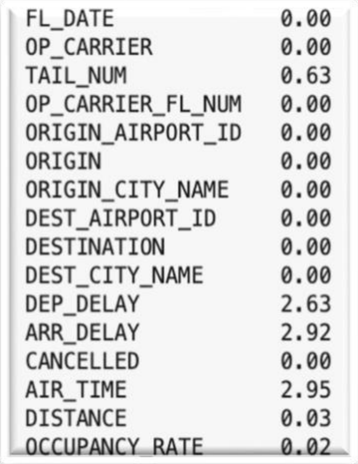
	ORIGIN	ORIGIN_STATE_NM	REPORTING_CARRIER	PASSENGERS	ITIN_FARE	DESTINATION
36135	PIH	Idaho	OO	1.00	13790.00	SLC
110692	BDL	Connecticut	YV	1.00	12225.00	IAH
355172	EWR	New Jersey	UA	1.00	12882.00	ORD
527911	LAX	California	UA	1.00	10944.00	ORD
586574	MCI	Missouri	UA	1.00	10564.00	PHX
927914	SFO	California	UA	1.00	13474.00	LAS
929787	SFO	California	UA	1.00	15365.00	EWR
968736	SLC	Utah	OO	1.00	38400.00	TWF

These anomalies could be treated as outliers. The boxplot also says the same and by looking at these I determined the following range for the columns.

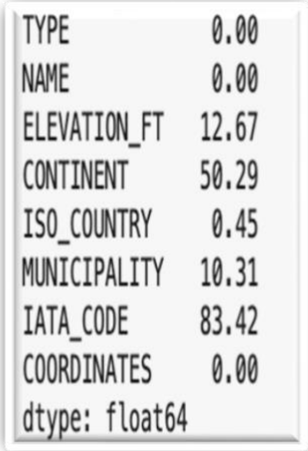
replacing_outliers_with_median(flights,'AIR_TIME',50,1000)
replacing_outliers_with_median(flights,'DISTANCE',50,6000)
replacing_outliers_with_median(flights,'DEP_DELAY’, False,1750)
replacing_outliers_with_median(flights,'ARR_DELAY’, False,2000)
replacing_outliers_with_median(tickets,'PASSENGERS',0,300)
replacing_outliers_with_median(tickets,'ITIN_FARE',20,15000)

We can interpret the function as func (Datafile,column, lower_cut , upper cut). If the value does not lie in between these range, we are replacing it with median. For more info on implementation please check out the jupyter notebook.

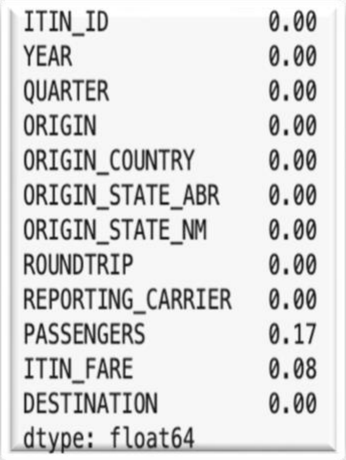
6) Dealing with Null Values



```
FL_DATE              0.00
OP_CARRIER           0.00
TAIL_NUM             0.63
OP_CARRIER_FL_NUM    0.00
ORIGIN_AIRPORT_ID    0.00
ORIGIN               0.00
ORIGIN_CITY_NAME     0.00
DEST_AIRPORT_ID      0.00
DESTINATION          0.00
DEST_CITY_NAME       0.00
DEP_DELAY            2.63
ARR_DELAY            2.92
CANCELLED            0.00
AIR_TIME             2.95
DISTANCE             0.03
OCCUPANCY_RATE       0.02
```



```
TYPE            0.00
NAME            0.00
ELEVATION_FT   12.67
CONTINENT      50.29
ISO_COUNTRY     0.45
MUNICIPALITY   10.31
IATA_CODE      83.42
COORDINATES     0.00
dtype: float64
```



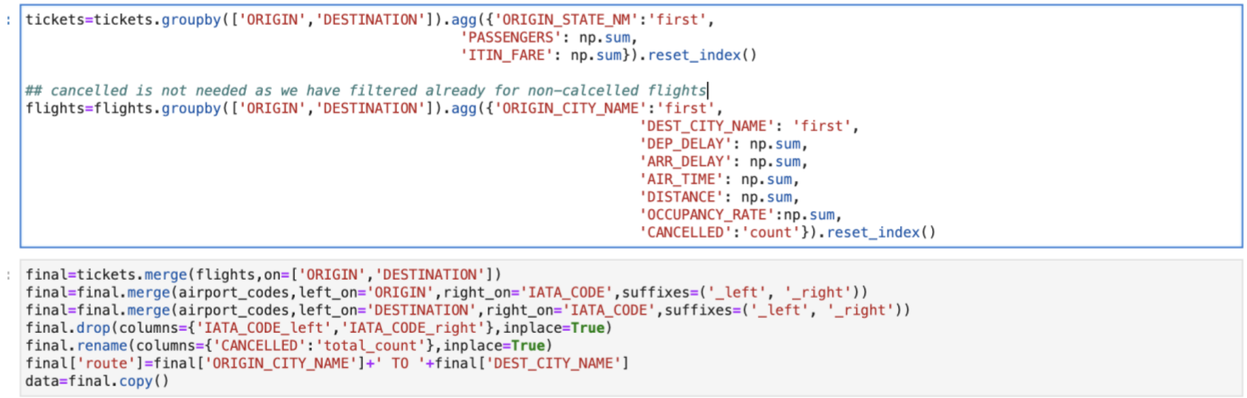
```
ITIN_ID              0.00
YEAR                 0.00
QUARTER              0.00
ORIGIN               0.00
ORIGIN_COUNTRY       0.00
ORIGIN_STATE_ABR     0.00
ORIGIN_STATE_NM      0.00
ROUNDTRIP            0.00
REPORTING_CARRIER    0.00
PASSENGERS           0.17
ITIN_FARE            0.08
DESTINATION          0.00
dtype: float64
```

Flights data   Airport codes data.   Ticket data

2. DATA TRANSFORMATIONS AND JOINS

After dealing with the quality of data we would now try to merge all the 3 datasets. Before merging all the files, we would try to aggregate the data as we are focusing on finding most busy and profitable routes. So, merging before aggregation will have a big-time complexity and we want to reduce on that. Hence, I grouped **FLIGHTS** and **TICKETS** dataset on **ORIGIN** and **DESTINATION**. We take the sum of all the important columns. After that we join all the tables. We do the inner join on tickets and flights and then join on **AIRPORT_CODES** to know if an airport is large or medium.

Code Snippet



```
tickets=tickets.groupby(['ORIGIN','DESTINATION']).agg({'ORIGIN_STATE_NM':'first',
                                                  'PASSENGERS': np.sum,
                                                  'ITIN_FARE': np.sum}).reset_index()

## cancelled is not needed as we have filtered already for non-calcelled flights
flights=flights.groupby(['ORIGIN','DESTINATION']).agg({'ORIGIN_CITY_NAME':'first',
                                                                       'DEST_CITY_NAME': 'first',
                                                                       'DEP_DELAY': np.sum,
                                                                       'ARR_DELAY': np.sum,
                                                                       'AIR_TIME': np.sum,
                                                                       'DISTANCE': np.sum,
                                                                       'OCCUPANCY_RATE':np.sum,
                                                                       'CANCELLED':'count'}).reset_index()
```

```
final=tickets.merge(flights,on=['ORIGIN','DESTINATION'])
final=final.merge(airport_codes,left_on='ORIGIN',right_on='IATA_CODE',suffixes=('_left', '_right'))
final=final.merge(airport_codes,left_on='DESTINATION',right_on='IATA_CODE',suffixes=('_left', '_right'))
final.drop(columns={'IATA_CODE_left','IATA_CODE_right'},inplace=True)
final.rename(columns={'CANCELLED':'total_count'},inplace=True)
final['route']=final['ORIGIN_CITY_NAME']+' TO '+final['DEST_CITY_NAME']
data=final.copy()
```

3. DATA VISUALIZATIONS AND FINAL RECOMMENDATIONS TO ANSWERS

Question-1

The 10 busiest round-trip routes in terms of number of round-trip flights in the quarter. Exclude canceled flights when performing the calculation.

We could find this out by sorting the data with respect to occupancy rate. I have already aggregated all the occupancy rate by a route and multiplied by 200. In the document it was shared that each plane could accommodate up to 200 passengers. Here are my top 10 busiest routes.

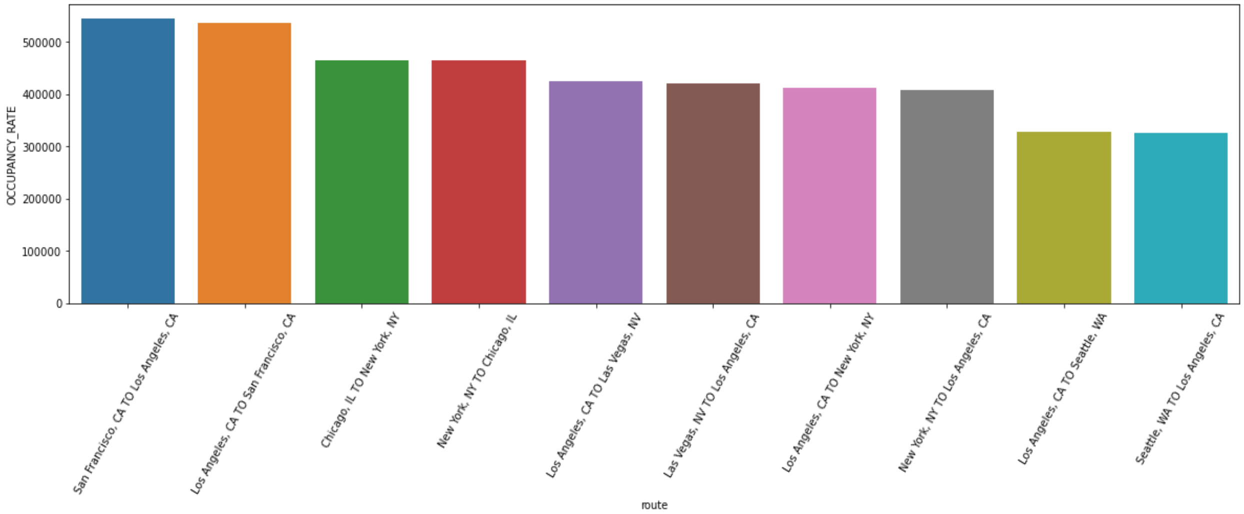


Figure 3: 10 Busiest Routes

Question-2

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.

Inputs from the document.

1) Fuel, Oil, Maintenance, Crew $8 per mile total
2) Depreciation, Insurance, Other $1.18 per mile total
3) Airport cost is 5000 dollars for medium airports and 10,000 dollars for large airports
4) For each individual departure, the first 15 minutes of delays are free, otherwise each minute costs the airline $75 in added operational costs.
5) Baggage fee is $35 for each checked bag per flight. We expect 50% of passengers to check an average of 1 bag per flight. The fee is charged separately for each leg of a round trip flight, thus 50% of passengers will be charged a total of $70 in baggage fees for a round trip flight.

Calculations

1) **Fare charges income** = **PASSENGERS***ITIN_FARE**. We took passengers not occupancy rate because it could be possible that some people had the ticket but did not travel.

2) **Baggage income** = **OCCUPANCY_RATE***70*0.5. Baggage income would be based on customers who are travelling, and we had to assume that half of the customers will have minimum one baggage. So, the above calculation gives us the minimum baggage income

3) **Arrival and Departure Delay cost** = **Arrival delay***75+ **Departure delay***75

4) **Airport cost** = Adding 5000 for medium airport and 10000 for large airport.

5) **Essential's cost** = (8+1.18) * **Distance**

6) **Total Profit** = (**Fare charges** + **Baggage Income**) – (**Arrival and Departure Delay cost** + **Airport cost** + **Essential's cost**)

Sorting the dataset by total profit gives us the following chart.

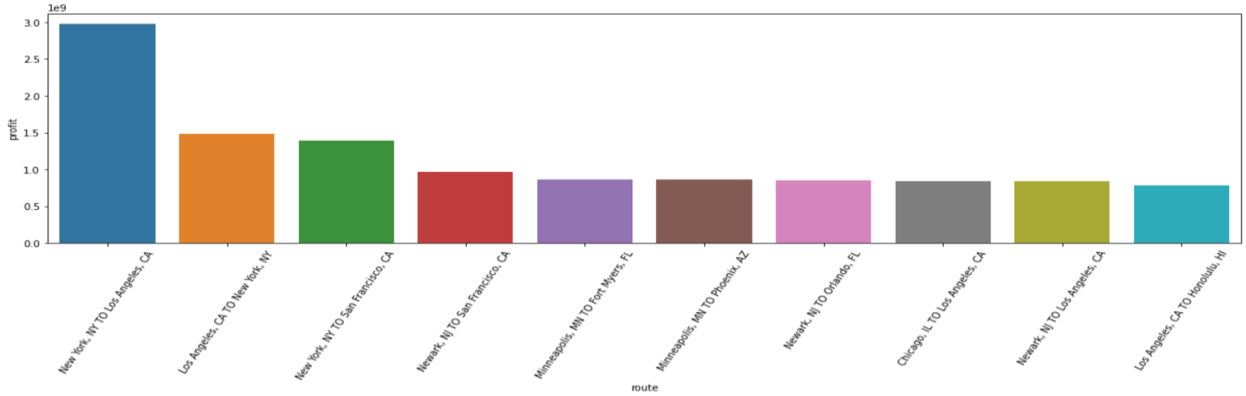


Figure 4: Total Profit in Quarter

This gives us the top 10 routes which gives us the maximum profit. But we also want to look at profit per trip. Because the above graph could be biased because if the route is busier, it will have more income. Hence

Profit per trip =Profit/Total no of flights in that route

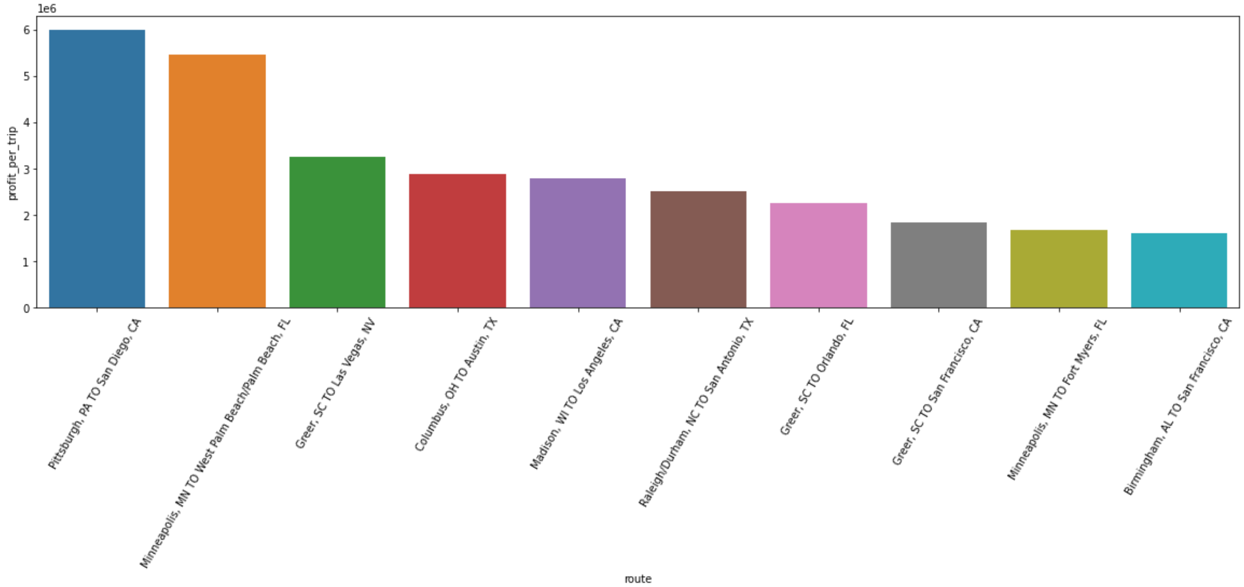


Figure 5: Top 10 profit per trip

This shows a very vivid picture and hence we see the total number of flights. This does not show the right picture as there is not many flights in these routes. We want to gain on customers and revenue therefore we should take only the profit in busy airport.

So, let's now add one more column whether an airport is busy or not.

Calculating busy airports based on our needs.

Method

1. We have 5 planes to operate
2. The total count of routes is for a quarter which is 3 months so approximate 90 days
3. Let's say a plane fly once every day and is then under maintenance.
4. It gives us 90*5 =450. We could fly up to 450 times in a month and we need to make profit in that.
5. Increasing the value by 20% because we don't know about seasonality. So, we would consider all airports as busy if we have above 600 total counts.

So here is the final plot that we get.

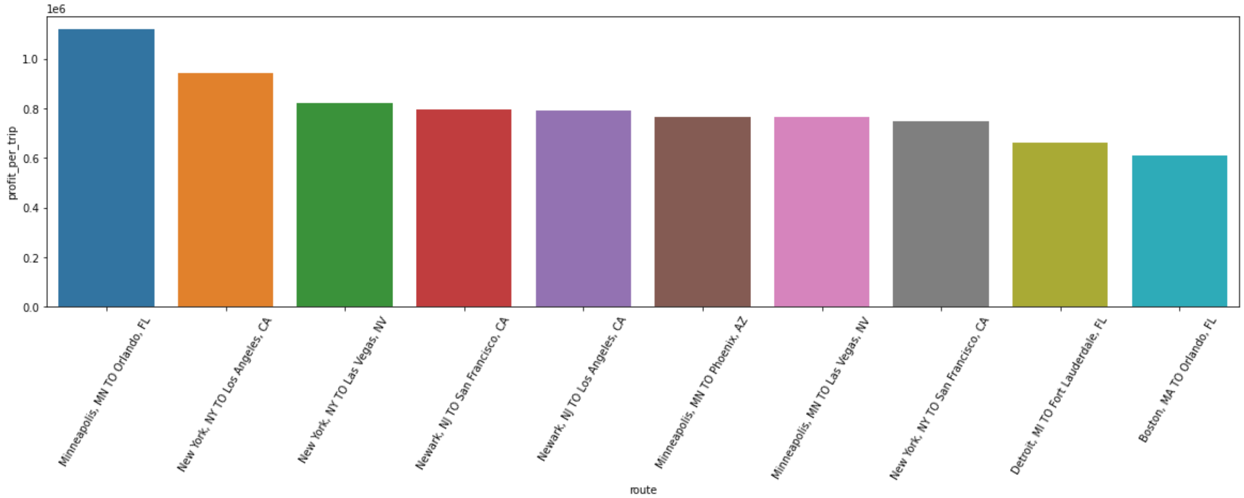


Figure 6: Total Profit generated by a route in 1 round trip in busy routes

Question-3

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.

Our Major Goal is to

1) Increase Profit
2) Increasing Customer Base
3) Making sure that there is minimum delay as our motto is "On time, for you"

Doing the above analysis on total delay as well. We need to make sure it is in busy routes. I have defined the busy metric above in Q2.

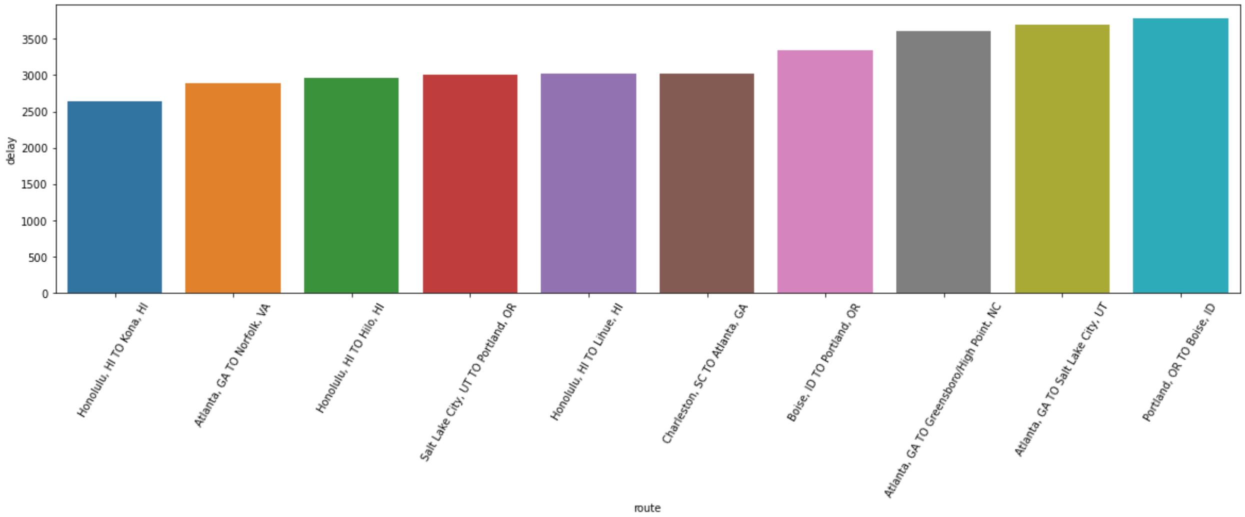


Figure 7: Top 10 Least delay in busy routes

Building our Score metric

First, we will be normalizing Profit, total Flights and Delay metric in between 0,1 and these will be our scores individually.

For profit and total flights increasing scores means good scores but for delay it would be inverse. So, after normalizing we would negate it by 1

Then defining our **score** metric.
It would be our **weighted average**. We will **have 33% weightage on delay and 33% on total flights and 33% on profit.**

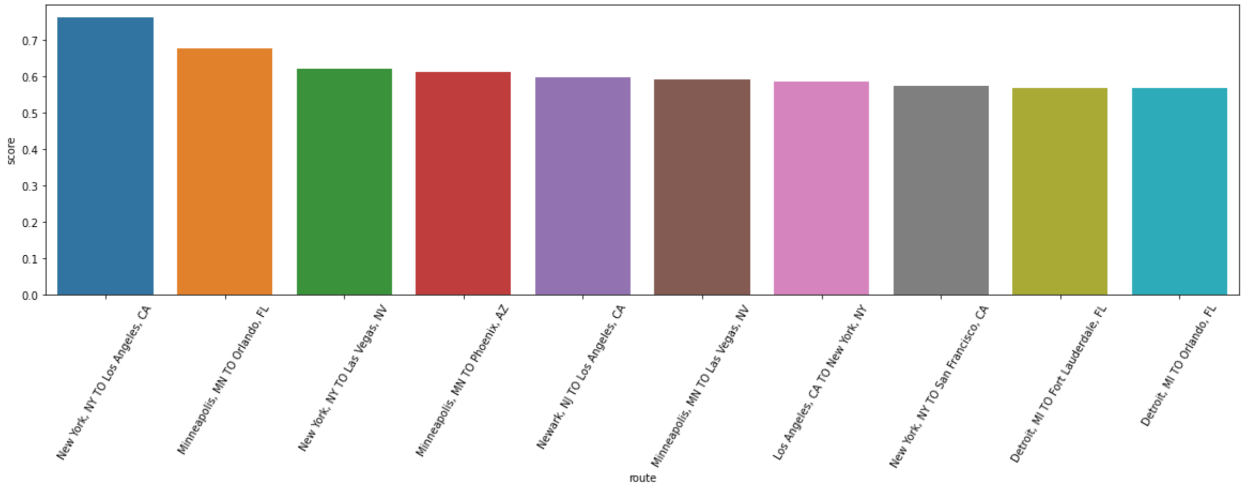


Figure 8: Best Routes to Invest according to our metric

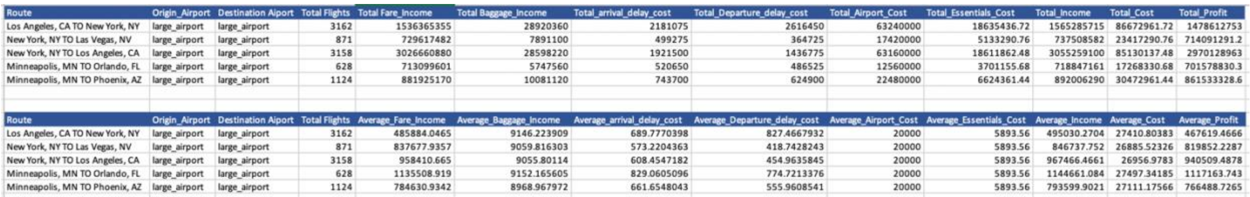
My top choices would be-

1) **New York, NY to Los Angeles, CA**
2) **Minneapolis to Orlando, FL**
3) **New York, NY to Las Vegas, NV**
4) **Los Angeles, CA to New York, NY**
5) **Minneapolis to Phoenix, AZ**

Question-4

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

The summary components for these routes are as follows:



Route	Origin_Airport	Destination Aiport	Total Flights	Total Fare_Income	Total Baggage_Income	Total_arrival_delay_cost	Total_Departure_delay_cost	Total_Airport_Cost	Total_Essentials_Cost	Total_Income	Total_Cost	Total_Profit
Los Angeles, CA TO New York, NY	large_airport	large_airport	3162	1536365355	28920360	2181075	2616450	63240000	18635436.72	1565285715	86672961.72	1478612753
New York, NY TO Las Vegas, NV	large_airport	large_airport	871	729617482	7891100	499275	364725	17420000	5133290.76	737508582	23417290.76	714091291.2
New York, NY TO Los Angeles, CA	large_airport	large_airport	3158	3026660880	28598220	1921500	1436775	63160000	18611862.48	3055259100	85130137.48	2970128963
Minneapolis, MN TO Orlando, FL	large_airport	large_airport	628	713099601	5747560	520650	486525	12560000	3701155.68	718847161	17268330.68	701578830.3
Minneapolis, MN TO Phoenix, AZ	large_airport	large_airport	1124	881925170	10081120	743700	624900	22480000	6624361.44	892006290	30472961.44	861533328.6

Route	Origin_Airport	Destination Aiport	Total Flights	Average_Fare_Income	Average_Baggage_Income	Average_arrival_delay_cost	Average_Departure_delay_cost	Average_Airport_Cost	Average_Essentials_Cost	Average_Income	Average_Cost	Average_Profit
Los Angeles, CA TO New York, NY	large_airport	large_airport	3162	485884.0465	9146.223909	689.7770398	827.4667932	20000	5893.56	495030.2704	27410.80383	467619.4666
New York, NY TO Las Vegas, NV	large_airport	large_airport	871	837677.9357	9059.816303	573.2204363	418.7428243	20000	5893.56	846737.752	26885.52326	819852.2287
New York, NY TO Los Angeles, CA	large_airport	large_airport	3158	958410.665	9055.80114	608.4547182	454.9635845	20000	5893.56	967466.4661	26956.9783	940509.4878
Minneapolis, MN TO Orlando, FL	large_airport	large_airport	628	1135508.919	9152.165605	829.0605096	774.7213376	20000	5893.56	1144661.084	27497.34185	1117163.743
Minneapolis, MN TO Phoenix, AZ	large_airport	large_airport	1124	784630.9342	8968.967972	661.6548043	555.9608541	20000	5893.56	793599.9021	27111.17566	766488.7265

Number of round trips needed to breakeven the cost would be

Airplane Cost	Route	Average Profit	Round Trip Needed
90000000	Los Angeles to New York	467619.4666	192.464186 = 193
90000000	New York to Los Vegas	819852.2287	109.7758801 = 110
90000000	New York to Los Angeles	940509.4878	95.69281455 = 96
90000000	Minneapolis to Orlando	1117163.743	80.56115373 = 81
90000000	Minneapolis to Phoenix	766488.7265	117.4185567 = 118

Question-5

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

Essential KPI's which are already given are

1) Distance
2) Fare
3) Arrival Delay
4) Departure Delay
5) Occupancy Rate

Others which could give us more info are:

1) Distribution of Business and Economy level seats
2) Target audience for the airlines. The scoring metric will change according to that
3) Income distribution of the city. More income means more premium travel requirement
4) Promotion and discounted ticket information was not given. This can also help with the analysis
5) Busy airports. Sometimes load factor on airports are very high so therefore there could be more delays and such airports could be avoided.
6) Weather data. Places with good weather should be prioritized as there would be less last minute cancellations and preventing bad news for customers.
7) Research on places should be carried out where government is supporting tourism or business.
8) Flight change information. Customers are not happy when there are frequent flight changes as it affects their schedule.
9) Income from Pantry inside the airlines should be mentioned. Indirect cost associated to that as well. It will help us know how the profit or loss in long and short travel journey is.
10) More Baggage information could have helped. The current method tells us the minimum baggage income we are making.