**Documentation**CAPITAL ONE - Airline Data Challenge

By  
Nadakkannu Kuthalaraja

1. [***Data Imports and Quality Checks 3***](#_bookmark0)

**TABLE OF CONTENTS**

1. [***Data Transformations and joins 6***](#_bookmark1)
2. [***Data Visualizations and final recommendations to answers 6***](#_bookmark2)
3. **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.

* 1. **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' |

* 1. **Type Casting some columns with object values to int or float**

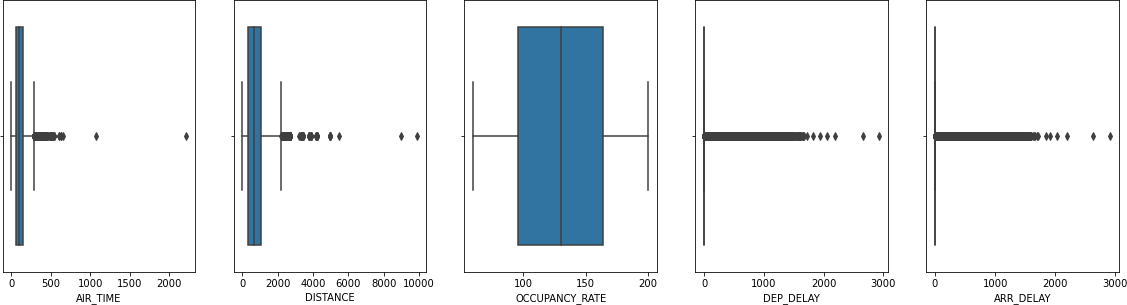
There were some inconsistencies in data. ITIN\_FARE column had somewhere $$ symbol attached to some of its values. Removed $$ symbols that otherwise the columns don’t convert to int. Similarly, AIR\_TIME and DISTANCE had some string values in its rows such as TWENTY or HUNDRED which does not make sense.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

* 1. **Making Changes to columns ARR\_DELAY, DEP\_DELAY and OCCUPANCY\_RATE**

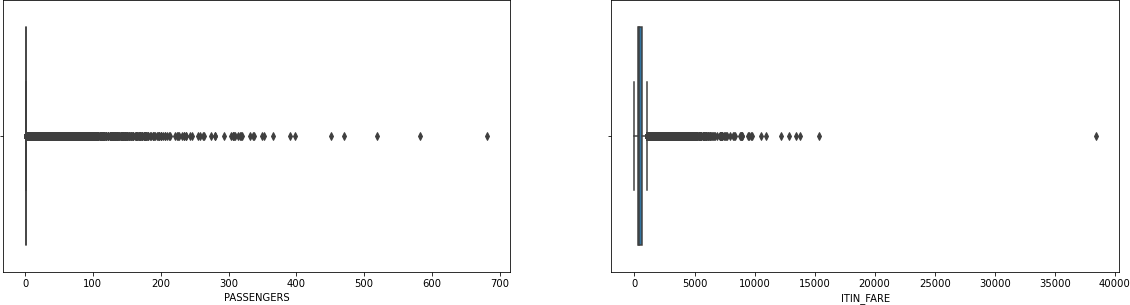
We are negating 15 minutes from Dep\_Delay and Arr\_delay column in flights data. In case subtraction is negative replacing it with 0. This would be very useful in later calculations. Similarly multiplying occupancy\_rate by 200 as all the flights are of 200 capacities. We don’t want to do aggregation of data and then do the steps.

* 1. **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.

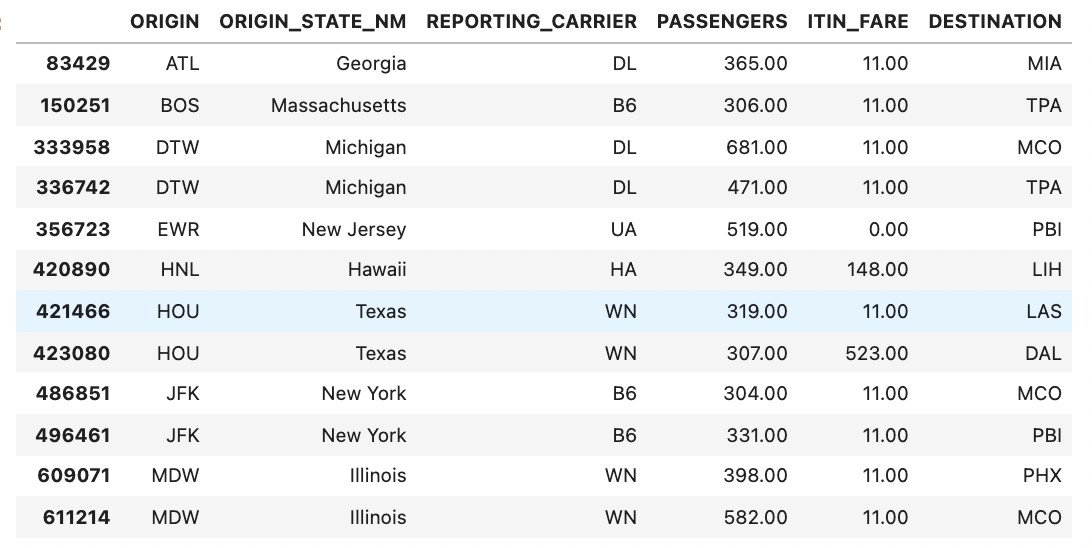


**Figure 1 Boxplots for flight data**

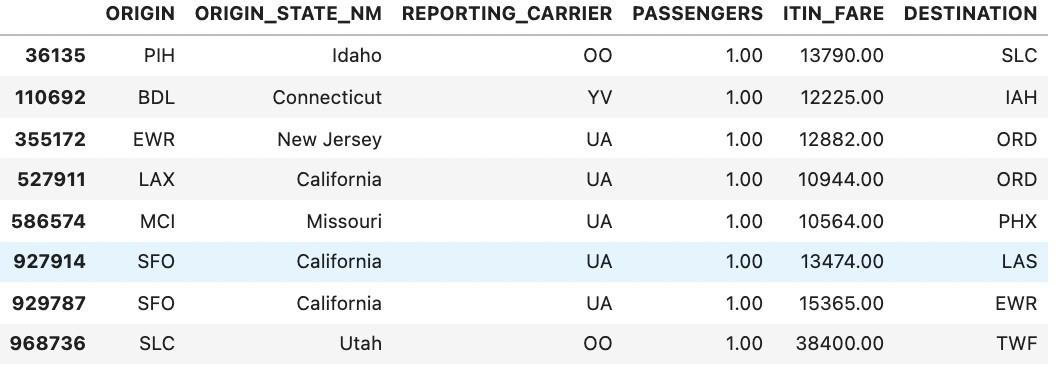
****

**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



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



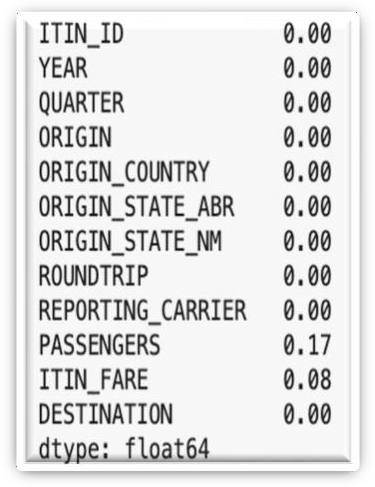
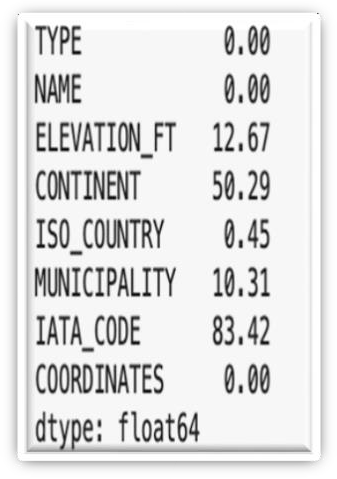
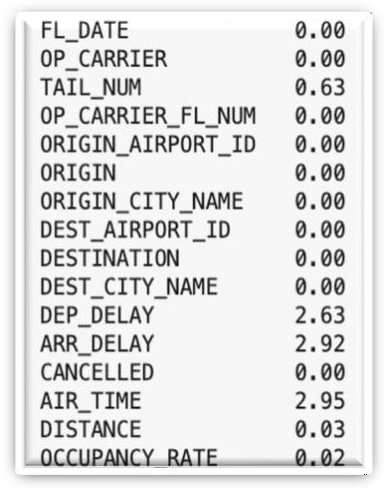
These anomalies could be treated as outliers. The boxplot also says the same and by looking at these I determined the following range for the columns.

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.

* 1. **Dealing with Null Values**

****

Flights data Airport codes data. Ticket data

1. **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.

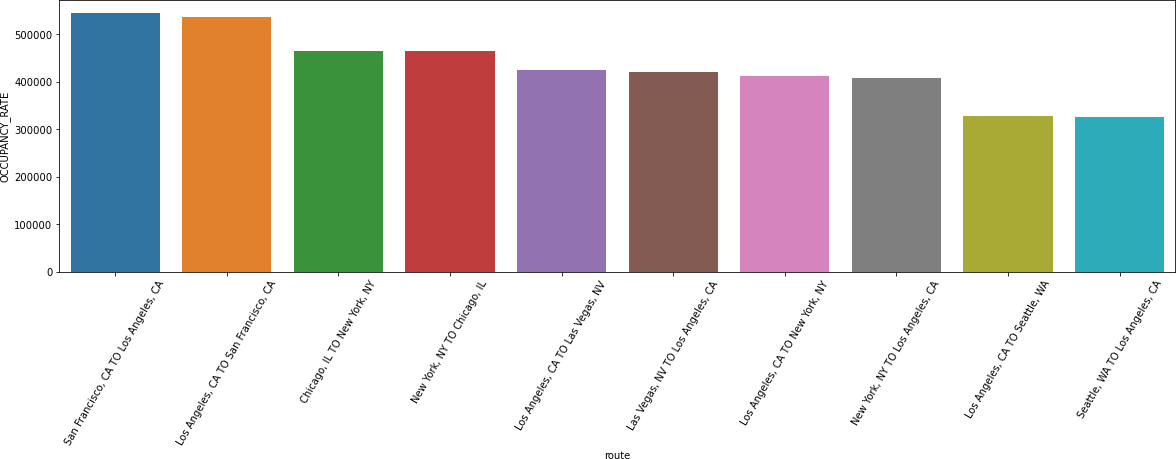
**Code Snippet**



1. **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**

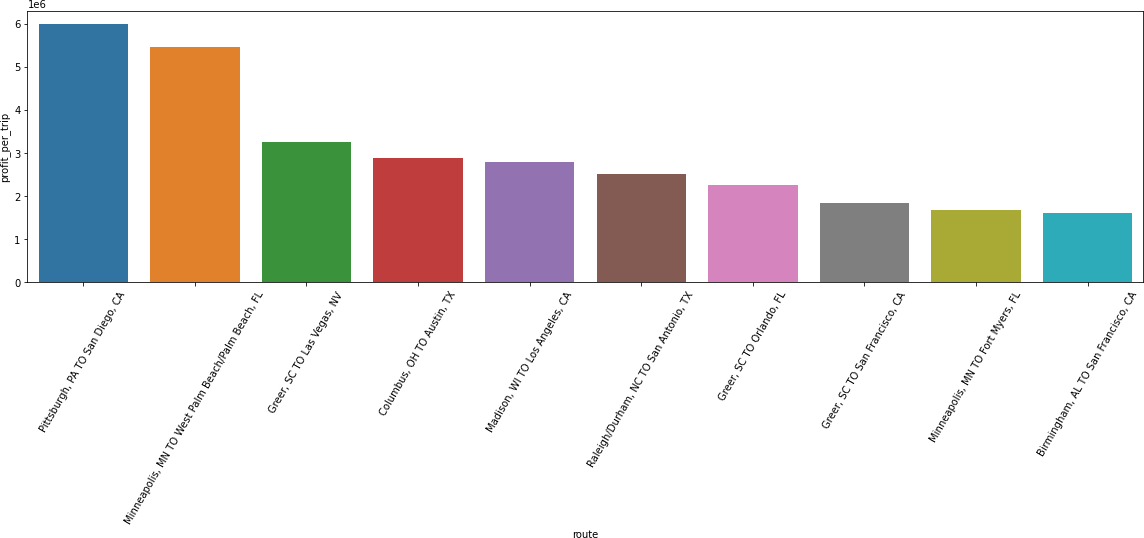
1. **Fare charges income** = **PASSENGERS**\***ITIN\_FARE**. We took passengers not occupancy rate because it could be possible that some people had the ticket but did not travel.
2. **Baggage income** = **OCCUPANCY**\_**RATE**\*70\*0.5. Baggage income would be based on customers who are travelling, 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**

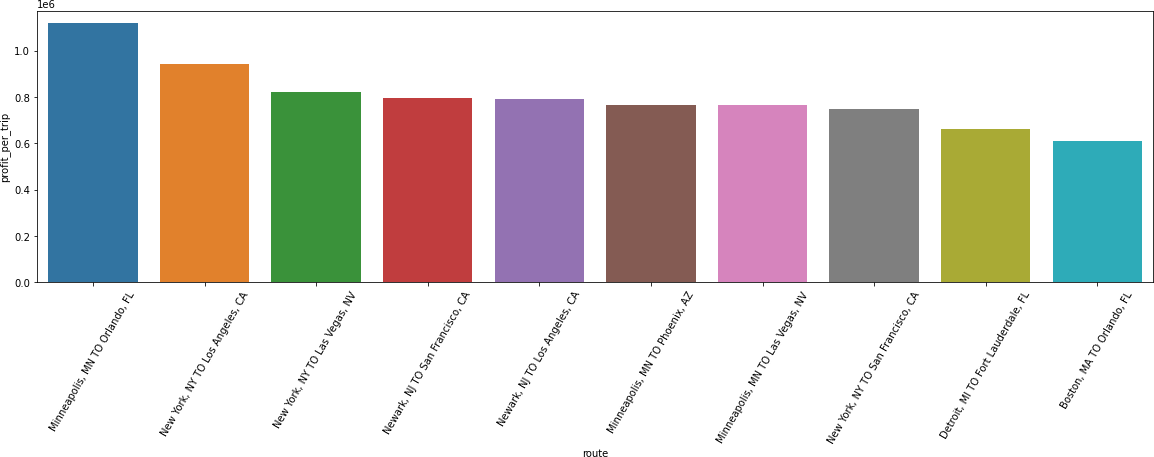
This shows a very vivid picture and hence we see the total number of flights. This does not show the right picture as there is not many flights in these routes. We want to gain on customers and revenue therefore we should take only the profit in busy airport.

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.



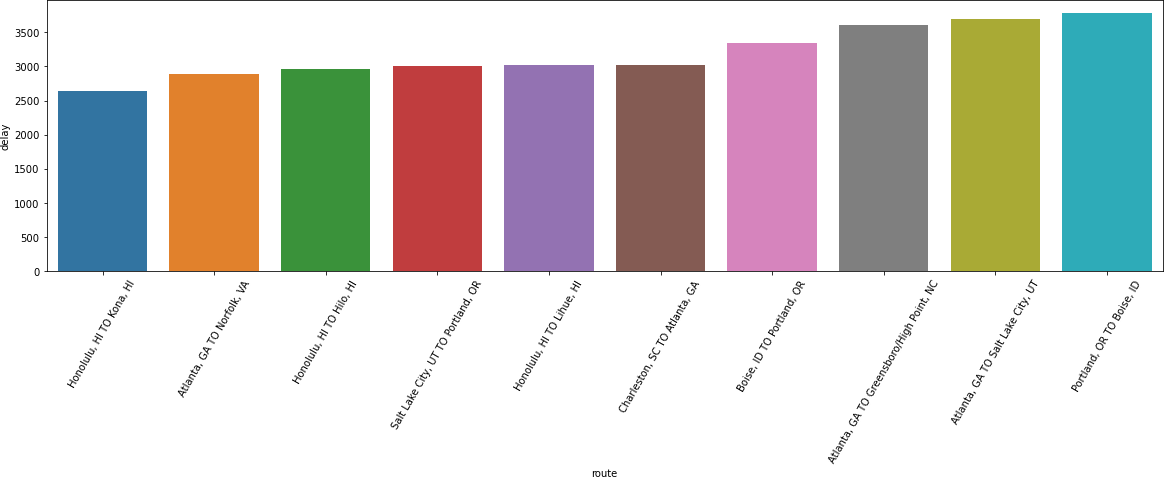
**Figure 6: Total Profit generated by a route in 1 round trip in busy routes**

**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**

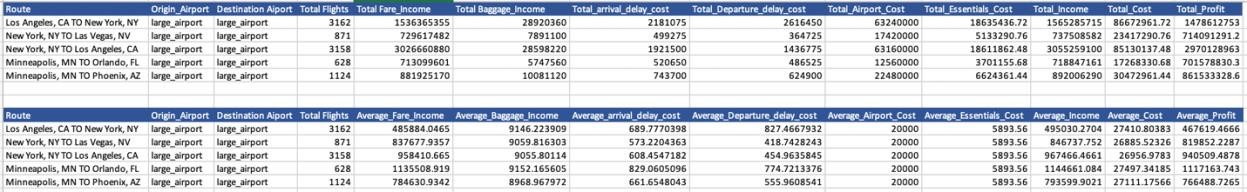
My top choices would be-

* 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:



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

Analyzing factors for optimizing airline operations:

1) Allocating Business and Economy seats strategically.

2) Tailoring scoring metrics based on the target audience.

3) Considering income distribution for premium travel demands.

4) Utilizing promotions and discounts for enhanced analysis.

5) Avoiding busy airports to minimize delays.

6) Prioritizing locations with favourable weather for fewer disruptions.

7) Researching government-supported tourism or business areas.

8) Minimizing frequent flight changes for customer satisfaction.

9) Evaluating pantry revenue and associated costs for profitability insights.

10) Exploring additional baggage data to optimize income streams.