Round Trip Route Recommendation Report and Visualization

Name - Tanmay Rajesh Bhardwaj

Email - tbhardwa@syr.edu

Table of Contents

1. PROBLEM STATEMENT

You are working for an airline company looking to enter the United States domestic market. Specifically, the company has decided to start with 5 round trip routes between medium and large US airports. An example of a round-trip route is the combination of JFK to ORD and ORD to JFK. The airline company must acquire 5 new airplanes (one per round trip route) and the upfront cost for each airplane is $90 million. The company's motto is "On time, for you", so punctuality is a big part of its brand image.

2. OBJECTIVE

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.

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.

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

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.

3. ASSUMPTIONS

1. Each airplane is dedicated to one round trip route between the 2 airports
2. **Costs**:
    - Fuel, Oil, Maintenance, Crew - $8 per mile total
    - Depreciation, Insurance, Other - $1.18 per mile total
    - Airport operational costs for the right to use the airports and related services are fixed at $5,000 for medium airports and $10,000 for large airports. There is one charge for each airport where a flight lands. Thus, a round-trip flight has a total of two airport charges.
    - For each individual departure, the first 15 minutes of delays are free, otherwise each minute costs the airline $75 in added operational costs.
    - For each individual arrival, the first 15 minutes of delays are free, otherwise each minute costs the airline $75 in added operational costs.
3. **Revenue**:
    - Each plane can accommodate up to 200 passengers and each flight has an associated occupancy rate provided in the Flights data set. Do not use the Tickets data set to determine occupancy.

- 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.
- Disregard seasonal effects on ticket prices (i.e. ticket prices are the same in April as they are on Memorial Day or in December)

4. METADATA CREATED

1. **fare_confidence**: This column indicates how confident we are in the average fare for a particular route. More on this in **section 2.2**

2. **profit_confidence**: This column is the product of fare confidence and average profit between a route.

3. **delay**: This column is the average of arrival and departure delay for a route

4. **profit_delay_ratio** : This column is the ratio of profit_confidence and delay

5. EXPLORATORY DATA ANALYSIS

1. airport_codes.csv
   - Rows 55369
   - Features 8

2. Flights.csv
   - Rows 1915886
   - Features 16

3. Tickets.csv
   - Rows 1167285
   - Features 12

6. DATA ANALYSIS

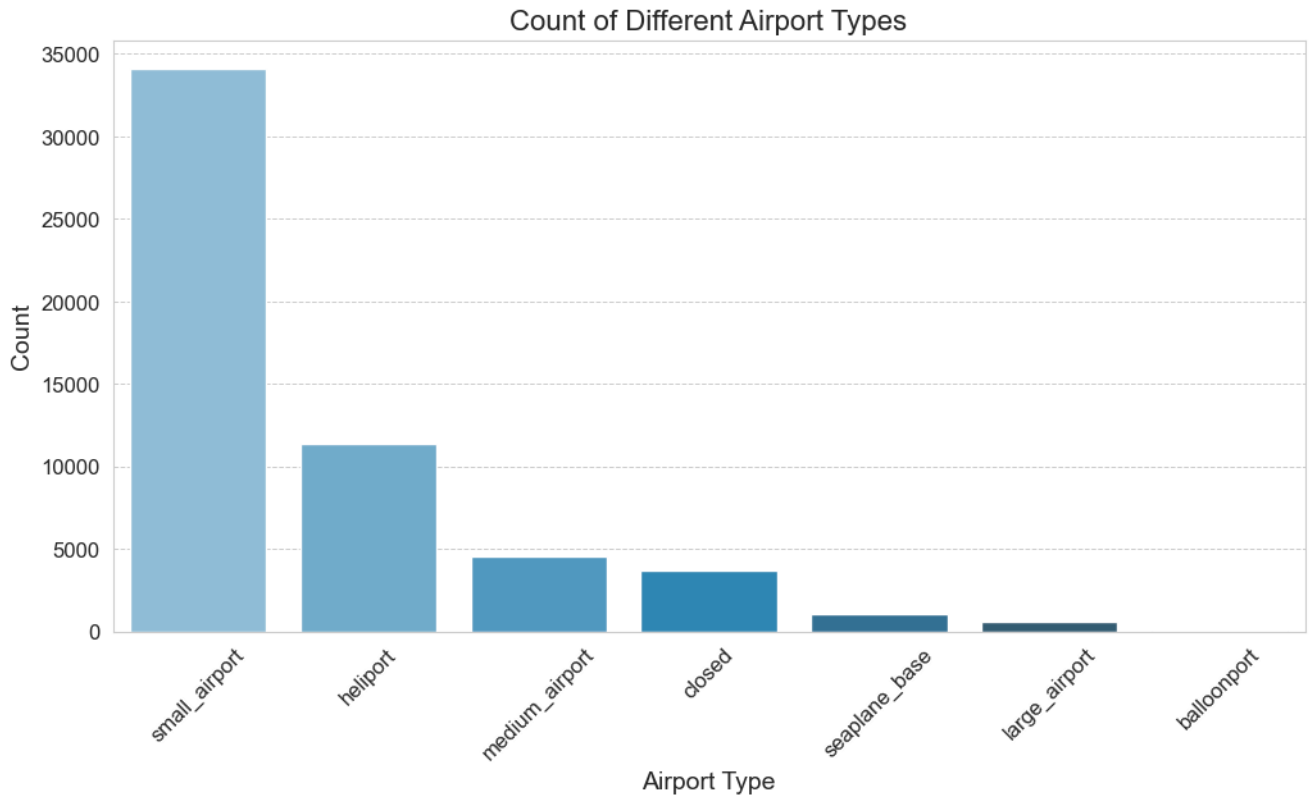
6.1 ANALYSING AIRPORT_CODES DATA:

- Below image shows the columns in the airport data

Field Name	Description
TYPE	The type of the airport, valid value like: small_airport, medium_airport, heliport, etc.
NAME	The name of the airport
ELEVATION_FT	Elevation of the airport from the sea level
CONTINENT	The continent airport belongs to
ISO_COUNTRY	The country of the airport
MUNICIPALITY	The city or town of the airport
IATA_CODE	An airport code is a three-letter geocode designating many airports and metropolitan areas around the world, defined by the International Air Transport
COORDINATES	Longitude and latitude coordinates of the airport

- A lot of the features are irrelevant for our analysis, hence we only chose the following feature
  - TYPE
  - NAME
  - IATA_CODE

- We are only interested in considering medium and large size airports



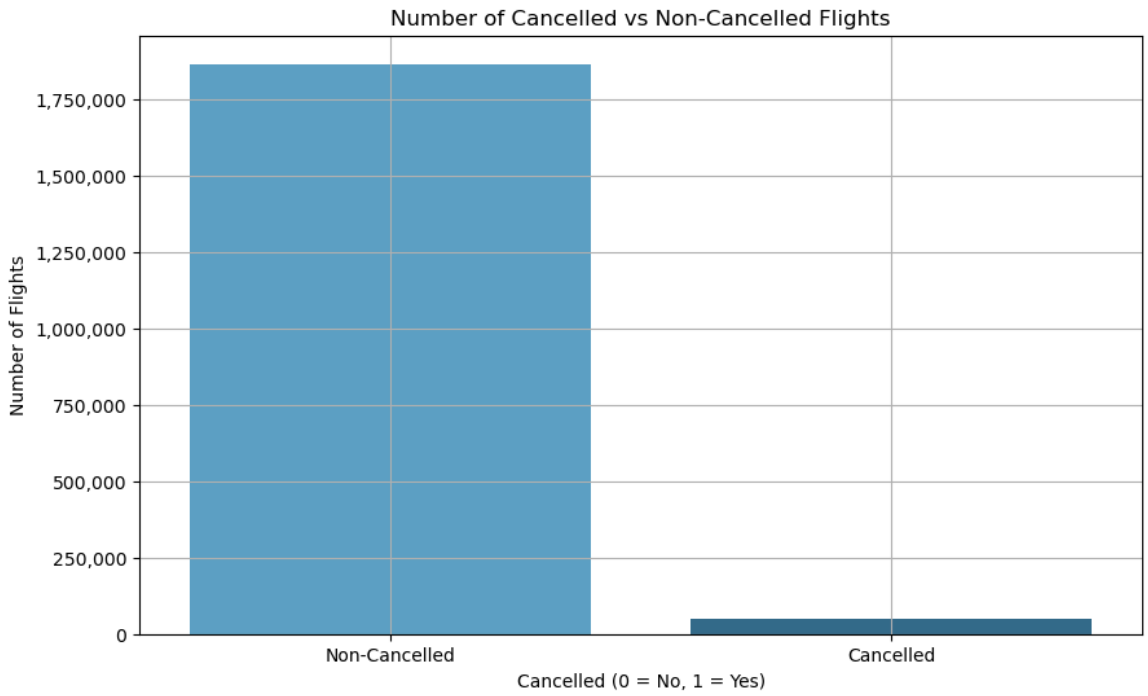
- Total number of **medium and large airports: 4459**

6.2 ANALYSING FLIGHTS DATA:

- Below image shows the columns in flights data:

Field Name	Description
FL_DATE	Flight Date (yyyy-mm-dd)
OP_CARRIER	Operating commercial carrier Flight code
TAIL_NUM	Tail Number is the aircraft registration number for the aircraft used (similar to VIN number for cars).
OP_CARRIER_FL_NUM	Operating commercial carrier Flight number
ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
ORIGIN	Origin Airport, International Air Transport Association (IATA) Airport Code
ORIGIN_CITY_NAME	Origin Airport, City Name
DEST_AIRPORT_ID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
DESTINATION	Destination Airport, Operating commercial carrier Flight code (IATA)Airport Code
DEST_CITY_NAME	Destination Airport, City Name
DEP_DELAY	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
CANCELLED	Cancelled Flight Indicator (1=Flight is cancelled). Cancelled flights should be excluded
AIR_TIME	Flight Time, in Minutes
DISTANCE	Distance between Origin and Destination Airports in Miles
OCCUPANCY_RATE	Occupancy rate of the flight

- Only choosing below relevant features and dropping the rest:
    - ORIGIN
    - DESTINATION
    - DEP_DELAY
    - ARR_DELAY
    - CANCELLED
    - DISTANCE
    - OCCUPANCY_RATE



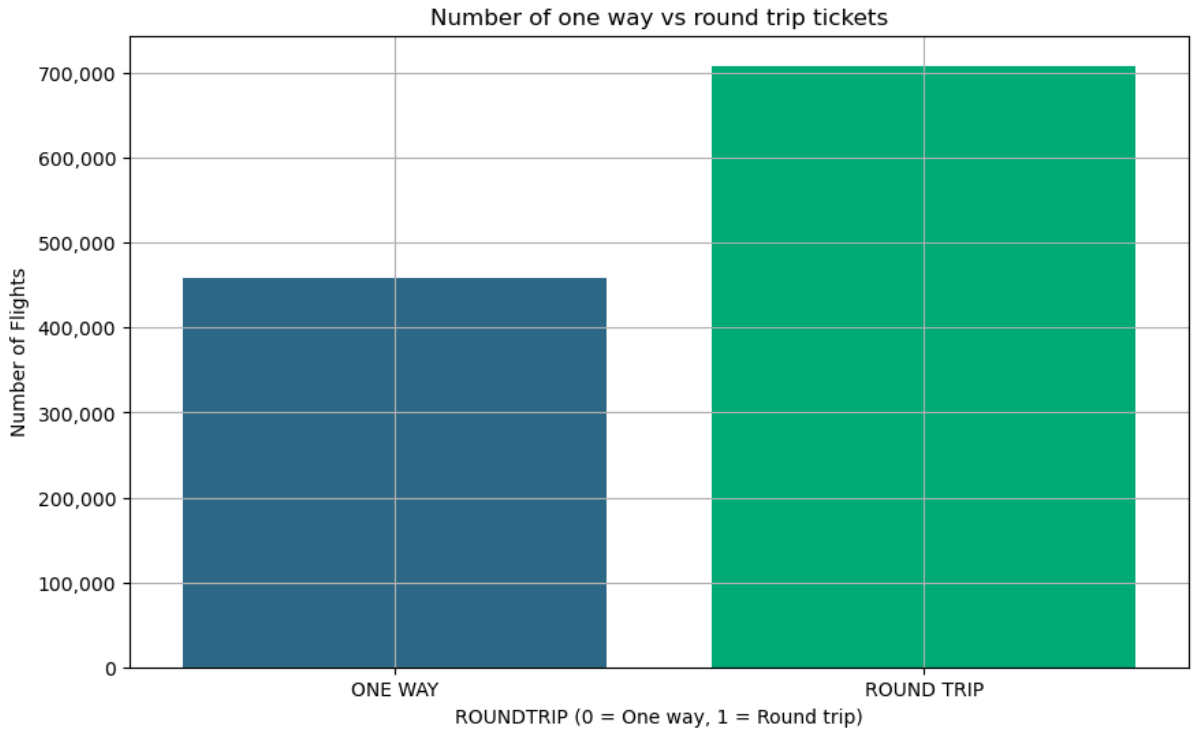
- We are only interested in considering non-cancelled flights. Above is the split in data of cancelled vs non cancelled flights

6.3 ANALYSING TICKETS DATA:

- Below image shows the columns in tickets data:

Field Name	Description
ITIN_ID	Unique identifier for the itinerary
YEAR	Year of the Itinerary
QUARTER	Quarter Number (1-4) for the Itinerary
ORIGIN	Origin Airport Code, International Air Transport Association Airport Code (IATA) which is unique for each airport
ORIGIN_COUNTRY	Country of the Origin Airport
ORIGIN_STATE_ABR	Origin Airport, State abbrevations.
ORIGIN_STATE_NM	Origin Airport, State Full Name
ROUNDTRIP	Round Trip Indicator (1= Round Trip and 0 = One Way). **Consider only round trips for your analysis.**
REPORTING_CARRIER	2 character Reporting Airline Carrier codes
PASSENGERS	Number of Passengers on the itinerary
ITIN_FARE	Itinerary Fare Per Person. Itinerary fare represents the whole round trip fare if ROUNDTRIP = 1. If ROUNDTRIP = 0, then the itinerary
DESTINATION	Destination Airport Code, International Air Transport Association Airport Code (IATA) which is unique for each airport

- In our analysis we are only considering roundtrip tickets and dropping the rest.



7. DATA QUALITY INSIGHTS

7.1 Airports_codes.csv

- **IATA_CODE**: there are 687 airports without IATA_CODE. These specific airports could have been part of potential round-trip routes.

7.2 Tickets.csv

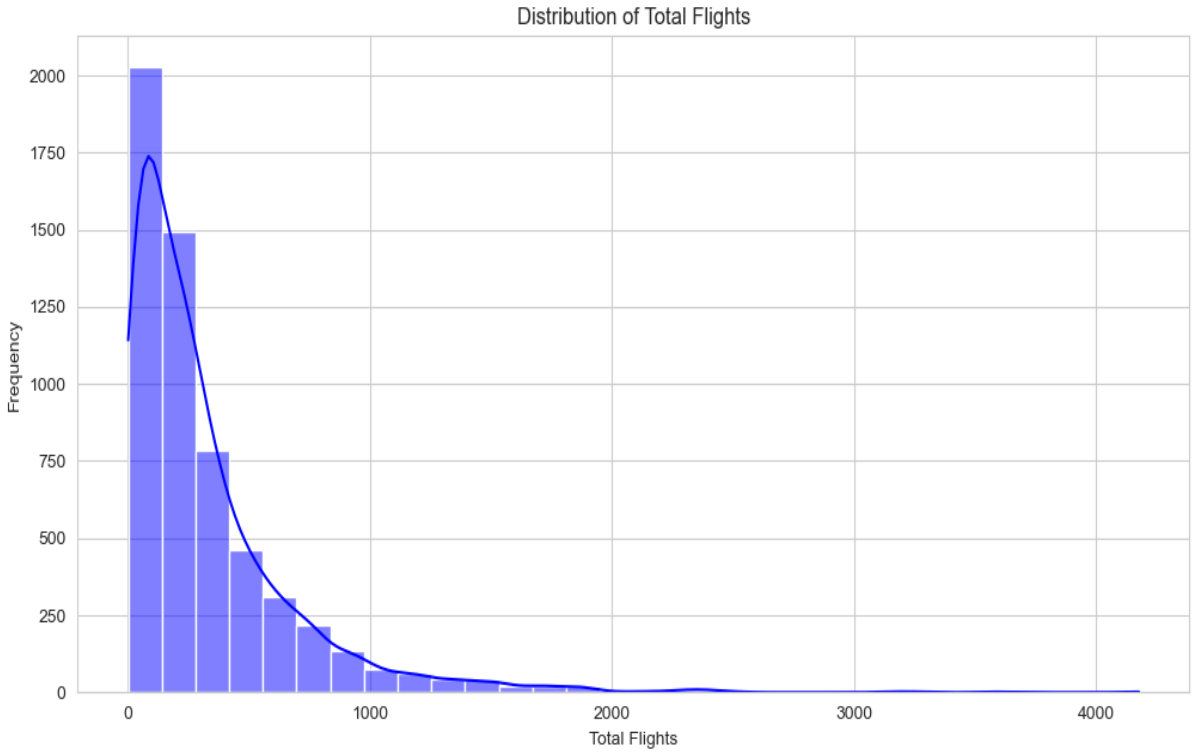
- **ITIN_FARE**: 960 tickets have missing fare information, which results in less accurate calculation of average fare for a particular route

7.3 If there could have been a way to know which tickets belonged to which flight (no common field in flights.csv and tickets.csv) the fare estimate would have been better.

7.4 Also, the tickets data after removing direct flights became very small when compared with the flights data. More tickets data also have resulted in better average fare prices.

8. TASK 1

- Started with choosing flights which are not cancelled and flights whose origin and destination airports belonged to medium/large airports.
- After this filtering, grouped the data on origin and destination to get the number of flights per route.

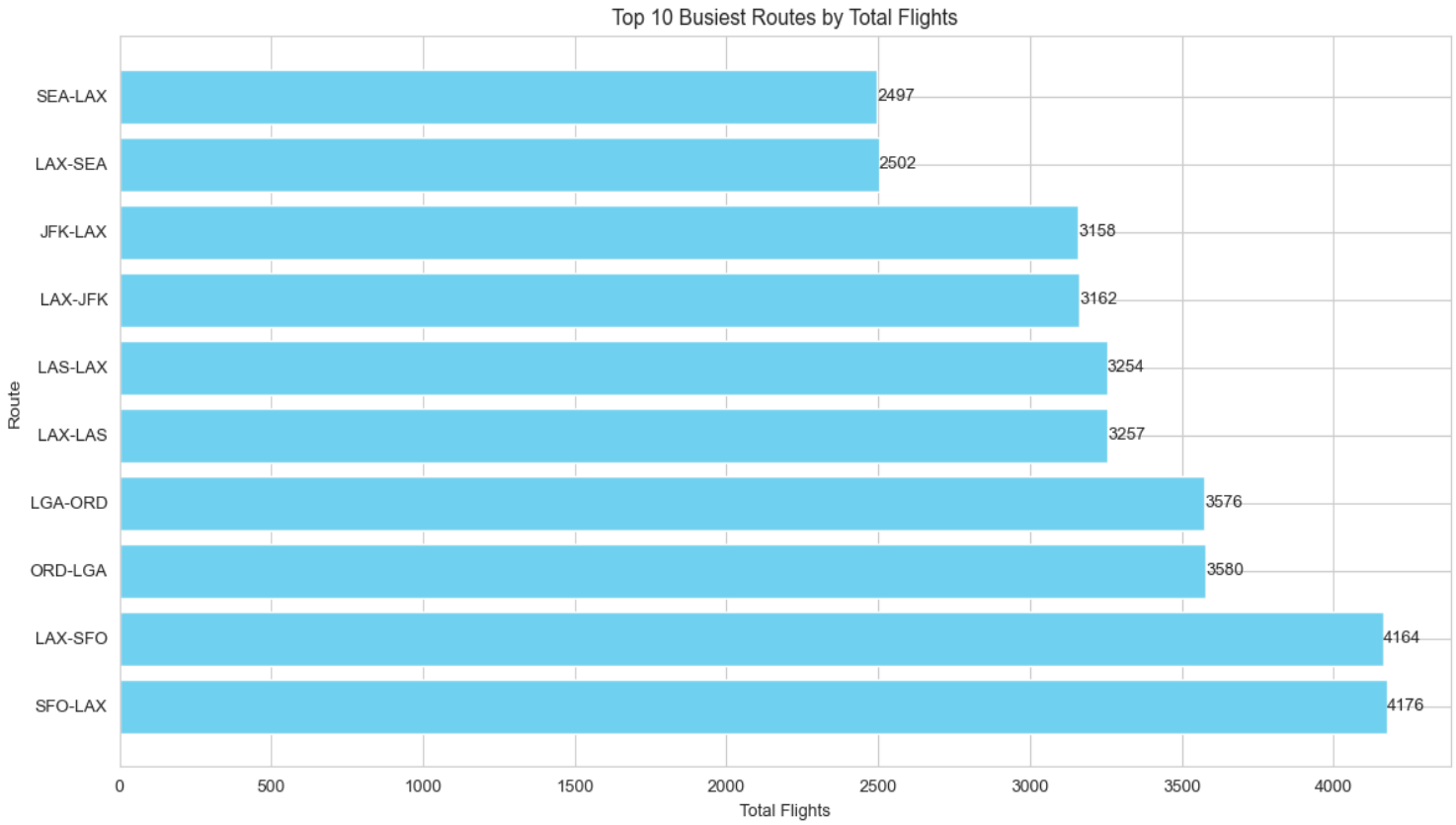


- The Above plot displays the distribution of the total number of flights between various origin and destination pairs. The bars represent the number of routes that fall within specific ranges of total flights. For example, the tallest bar indicates that there are about 2000 routes with a total of 0 to 100 flights in that quarter.
- The plot shows that most routes have a relatively low number of total flights, with the frequency decreasing as the number of flights increases, suggesting that only a few routes have a very high number of flights
- Selected the top 10 roundtrip routes with maximum number of flights between them.

8.1 Top 10 busiest flights

1. Seattle Tacoma International Airport - Los Angeles International Airport
2. Los Angeles International Airport - Seattle Tacoma International Airport
3. John F. Kennedy International Airport - Los Angeles International Airport
4. Los Angeles International Airport - John F. Kennedy International Airport
5. McCarran International Airport - Los Angeles International Airport
6. Los Angeles International Airport - McCarran International Airport

7. LaGuardia Airport - Chicago O'Hare International Airport
8. Chicago O'Hare International Airport - LaGuardia Airport
9. Los Angeles International Airport - San Francisco International Airport
10. San Francisco International Airport - Los Angeles International Airport



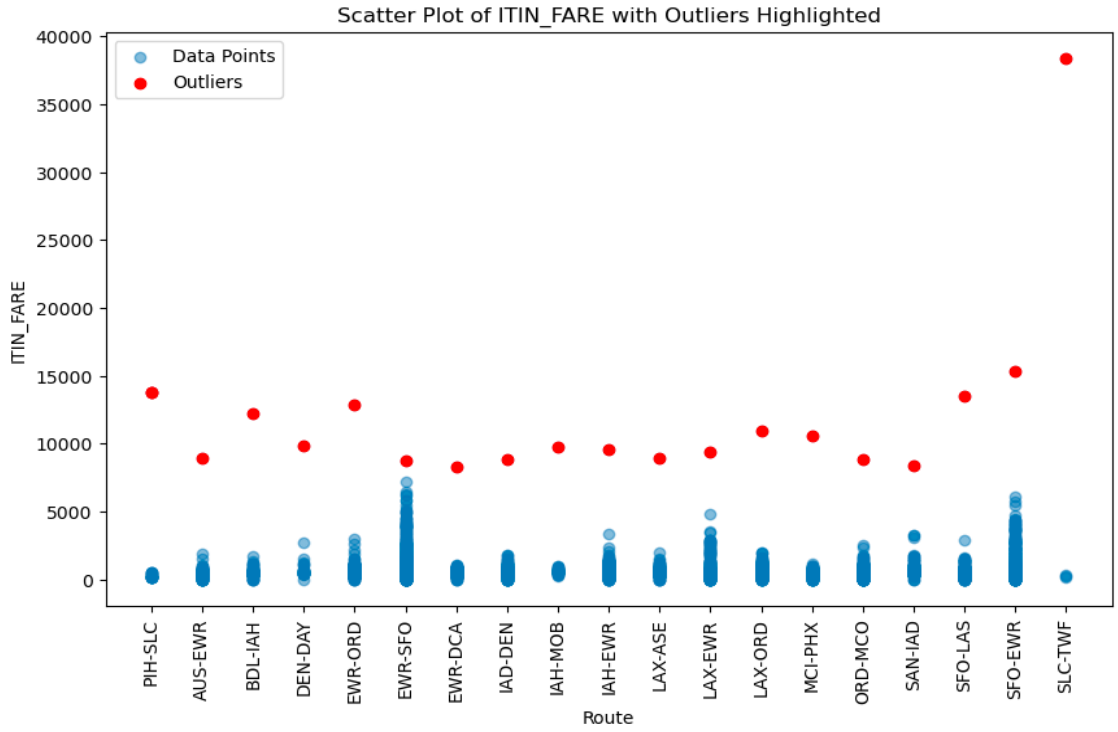
9. TASK 2

- The 10 most profitable round-trip routes (without considering the upfront airplane cost) in the quarter.
- To start with the calculation of which round trip routes were most profitable in the quarter, we need an average fare between those routes.
- We have an ITIN_FARE column in the tickets data which represents Itinerary Fare Per Person. Since we already filtered the data for round trip tickets, we can group the tickets based on origin and destination to get an average fare between one route.
- $Average\ Fare\ = \frac{sum\ of\ different\ fare\ between\ one\ route}{number\ of\ tickets\ between\ that\ route}$

9.1 Identifying Outliers

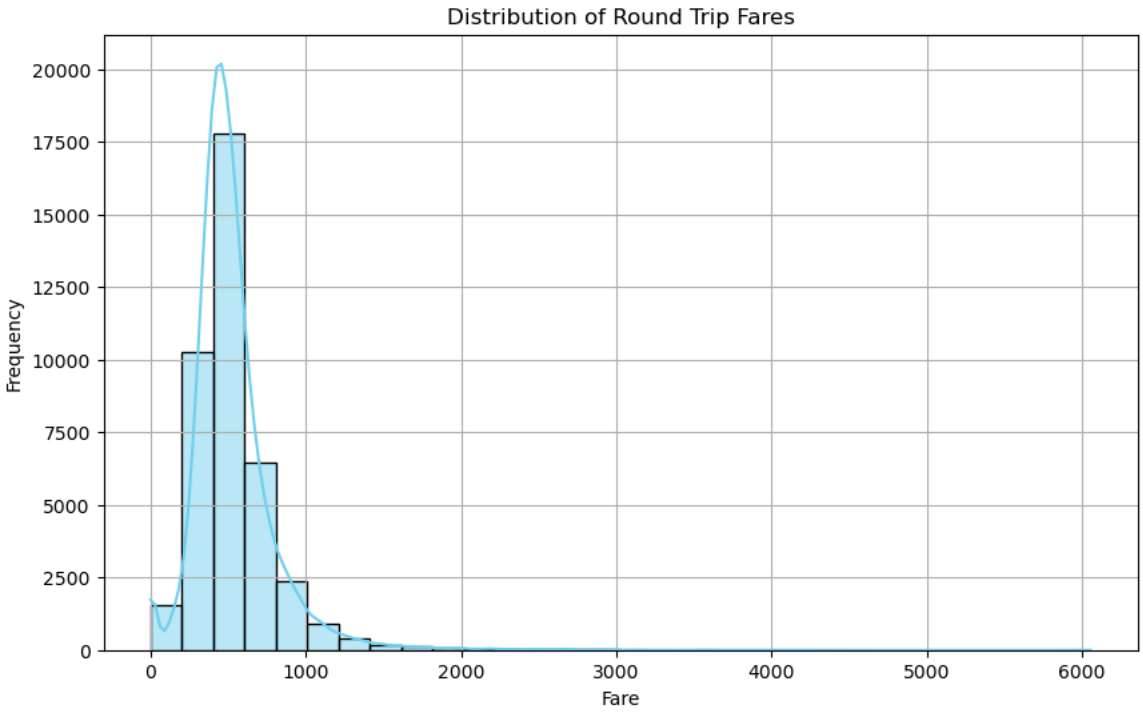
- There are a couple of outliers in the ITIN_FARE which would give us the wrong average fare for one route.

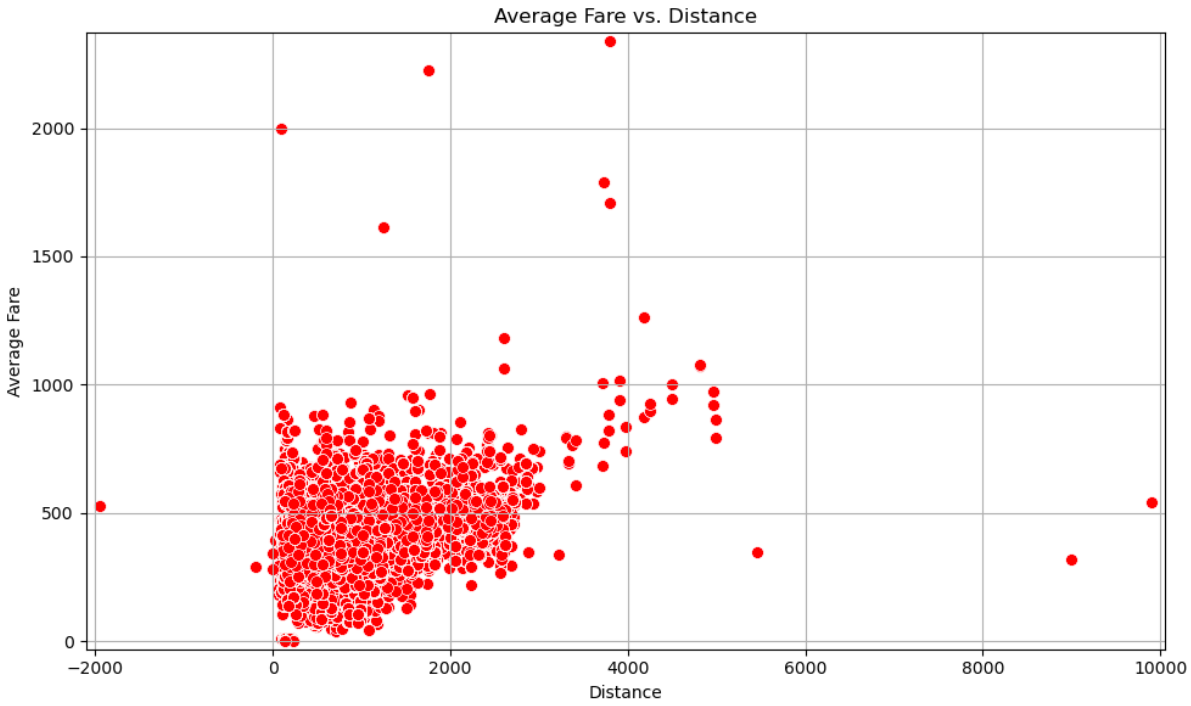
- Below plot displays presence of outliers in the ITIN_FARE



- Used **Interquartile range** method to remove outliers before calculating the average fare between routes.

- The below plot shows the distribution of round-trip fares. The histogram indicates that most fares are concentrated below $1000, with a sharp peak around $200 to $400.





- We can't solely rely on the average fare we've calculated because the number of tickets sold varies significantly between routes. For instance, some routes may only have one ticket sold, making it unreliable to determine the average fare.

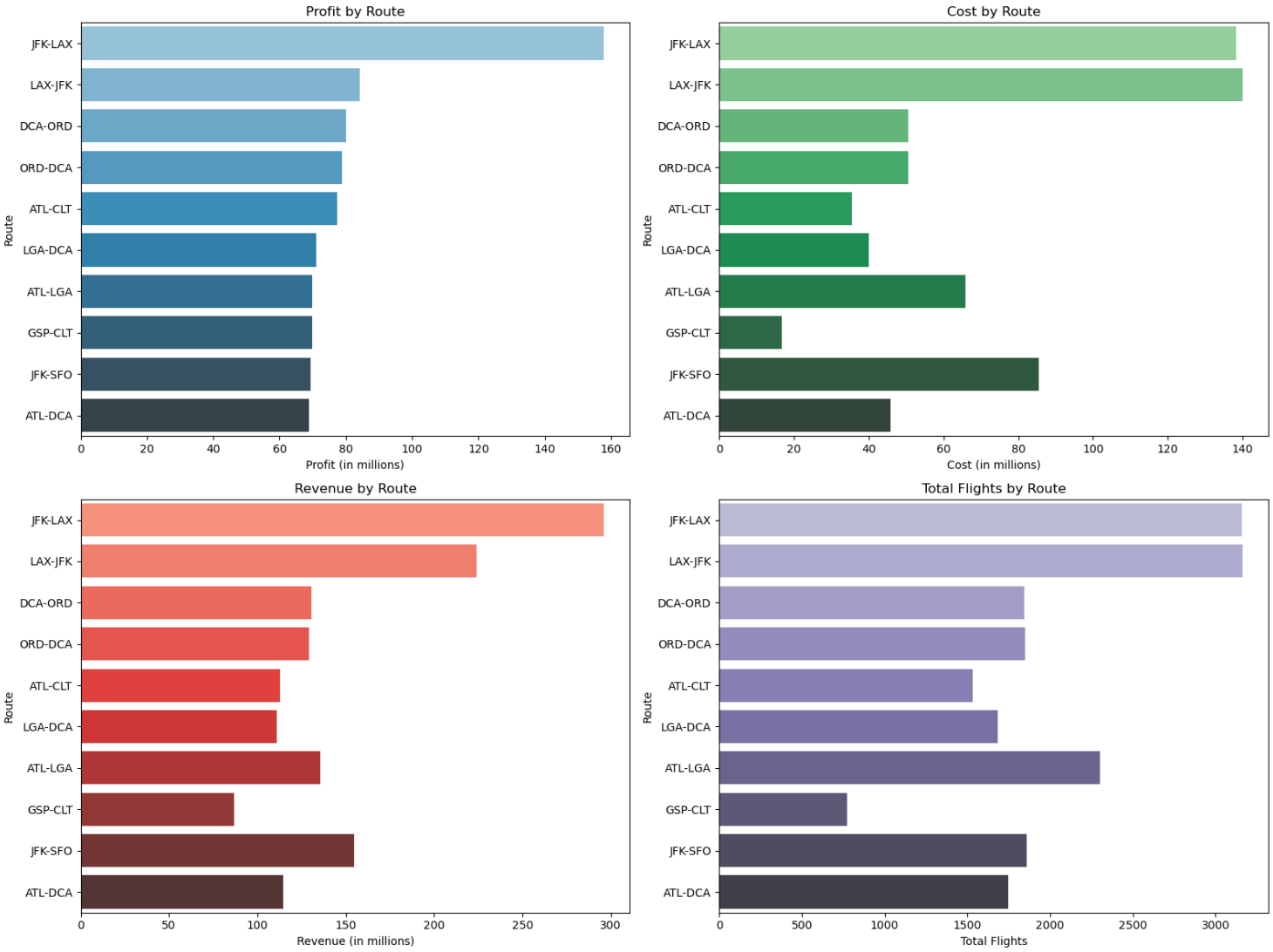
- Conversely, other routes may have 700-800 tickets sold, providing a more accurate estimate of the average fare for those routes.

9.2 Fare Confidence

- To address the above issue, I developed a new metric called "fare confidence." This metric indicates how confident we are in the average fare for a particular route.
- Fare confidence for a route is calculated using the following formula:
  - $$\boldsymbol{Fare\ confidence} = \frac{Number\ of\ tickets\ for\ that\ route}{Max\ number\ of\ tickets\ for\ any\ route} \times \mathbf{100}$$
- Its value ranges from 0 to 100%, the less the number the less confidence we have in that fare.
- This metric helps us gauge the reliability of the average fare based on the volume of tickets sold for each route. This metric will play a crucial role when we are working on which top 5 routes the airline company should invest in.
- So now we have an average fare for each route and its fare confidence. We merged these columns with the flight data on origin and destination for the calculation of profit between routes.

9.3 Top 10 most profitable round-trip routes.

1. John F. Kennedy International Airport - Los Angeles International Airport
2. Los Angeles International Airport - John F. Kennedy International Airport
3. Ronald Reagan Washington National Airport - Chicago O'Hare International Airport
4. Chicago O'Hare International Airport - Ronald Reagan Washington National Airport
5. Hartsfield Jackson Atlanta International Airport - Charlotte Douglas International Airport
6. La Guardia Airport - Ronald Reagan Washington National Airport
7. Hartsfield Jackson Atlanta International Airport - La Guardia Airport
8. Greenville Spartanburg International Airport - Charlotte Douglas International Airport
9. John F. Kennedy International Airport - San Francisco International Airport
10. Hartsfield Jackson Atlanta International Airport - Ronald Reagan Washington National Airport



10. TASK 3

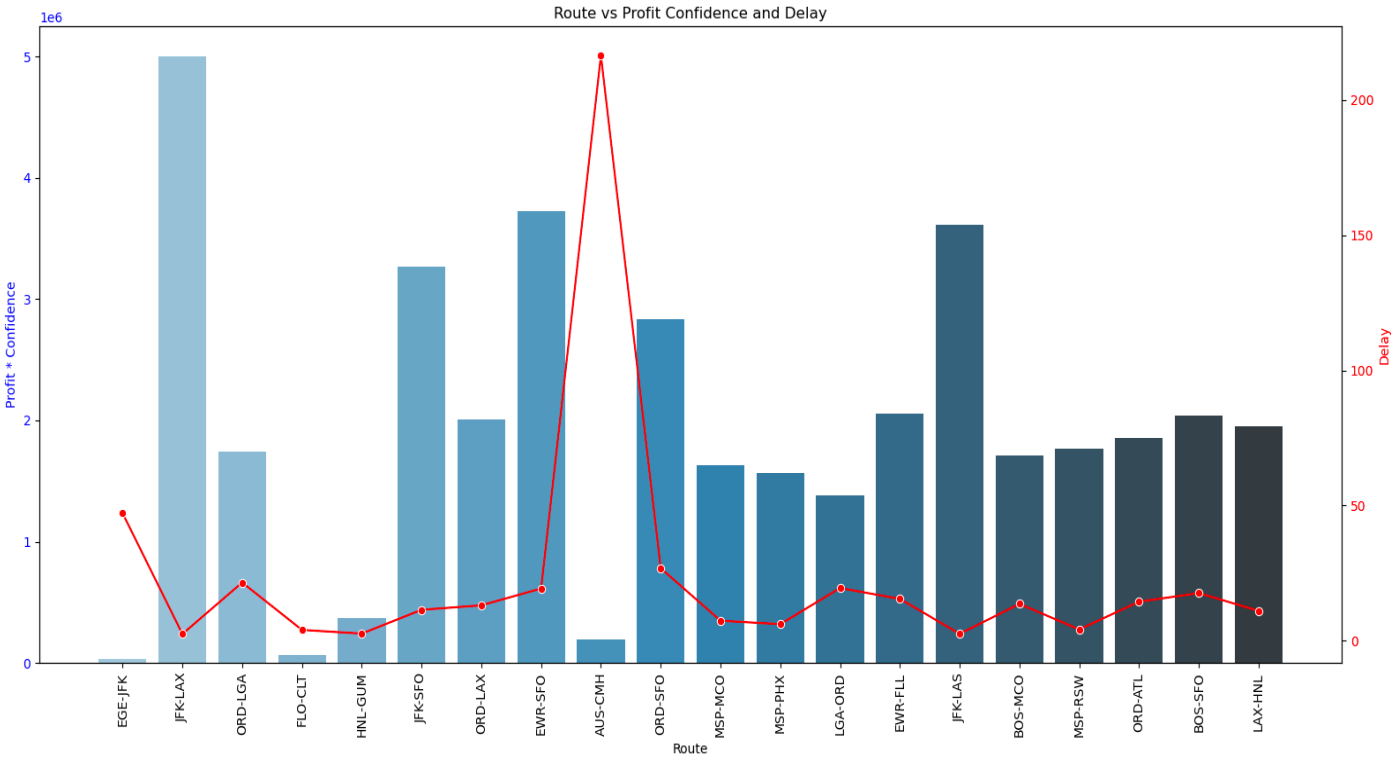
- The 5 round trip routes that you recommend investing in based on any factors that you choose.

10.1(KPI's) Key Performance Indicator for recommended flights

1. **Fare confidence**
2. **Average profit per route**
3. **Departure delay**
4. **Arrival delay**

- The best round-trip routes will be those with maximum fare confidence, average profit and minimum departure and arrival delay.

- Multiplied fare confidence and average profit to create a new value, **profit confidence.**
- Took average of departure and arrival delay to create a new value, **delay**.

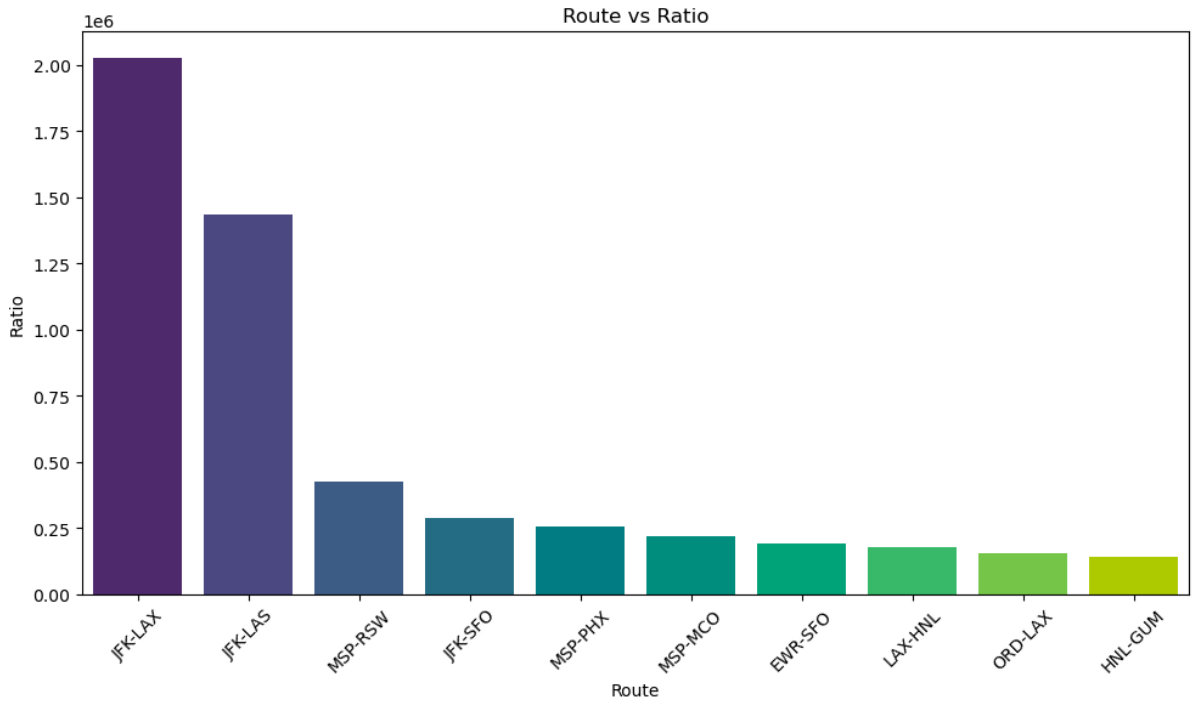


- Above plot shows the profit confidence vs delay for top 20 routes.

- Since best flights value is directly proportional to fare confidence profit and inversely proportional to delay.

- Took a ratio of profit confidence and delay and we will use this ratio to find recommended routes.

- Below plot shows the top 10 routes with the highest profit_delay_ratio.

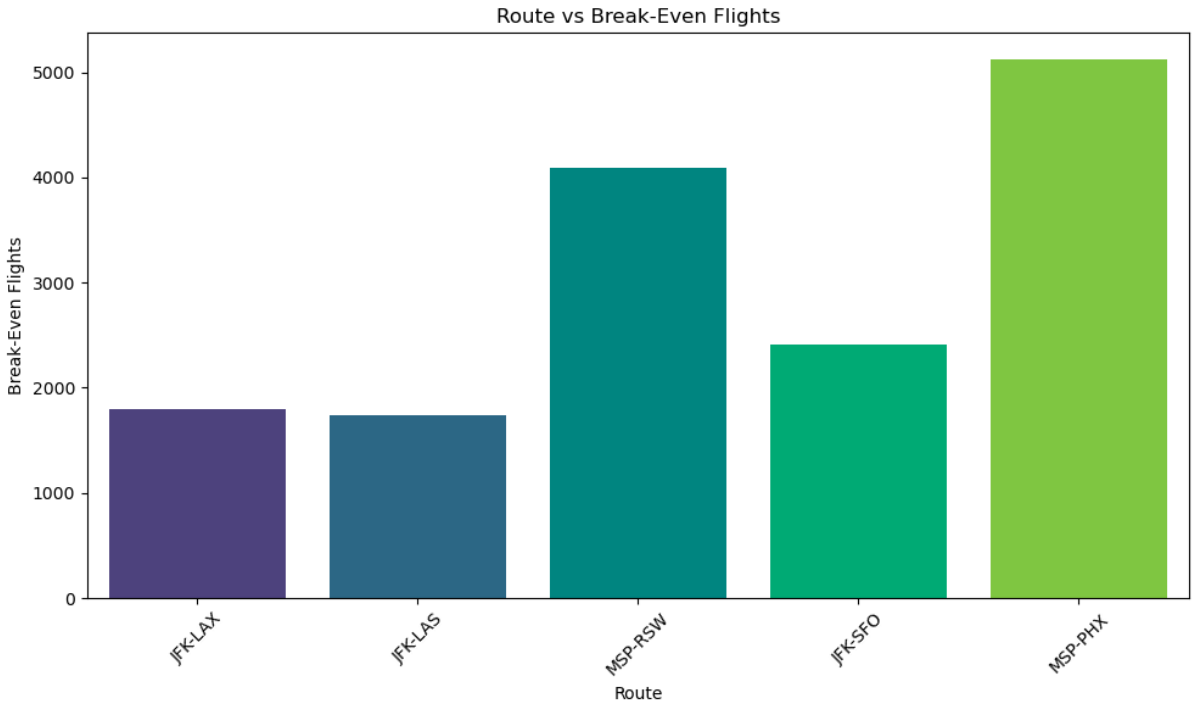


10.2 Top 5 recommended routes the company should invest in

1. John F. Kennedy International Airport - Los Angeles International Airport
2. Los Angeles International Airport - John F. Kennedy International Airport
3. Minneapolis-St Paul International/Wold-Chamberlain Airport - Southwest Florida International Airport
4. John F. Kennedy International Airport - San Francisco International Airport
5. Minneapolis-St Paul International/Wold-Chamberlain Airport - Phoenix Sky Harbor International Airport

10.3 Round trip flights needed to breakeven the cost of airplane

- Below plot shows the number of round-trip flights it will take for these routes to breakeven the upfront cost of buying an airplane.



11. FUTURE STEPS

- We can perform time series analysis to identify trends and seasonal patterns in delays.
- Conduct a detailed profitability analysis for each route by also incorporating fare by airline company.
- Incorporate competitor analysis to identify market trends and opportunities.
- We could have searched for IATA CODES that were missing in the valid airports data. These airports could have been potential routes