Airline_Data_Challenge_Report

Business Problem

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 has to 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.

Data Description (Metadata)

There were three data sets namely Airport_codes, Flights and Tickets in csv format to determine the given question and recommendation in the business problem.

Airport_codes : The table of airport_codes metadata has plenty of attributes to show complete details about each airport. The TYPE field displays the kind of airport, such as small_airport , medium_airport or heliport. The NAME field gives us the name of the airport and ELEVATION_FT is for telling its height from sea level. The field CONTINENT informs us about the continent in which the airport is situated, and ISO_COUNTRY indicates the code of the country. The MUNICIPALITY field is used for stating the city or town where the airport lies. IATA_CODE provides a unique three-letter code assigned by the International Air Transport Association that has global recognition as an identifier for airports all around the world. Lastly, the field of COORDINATES gives the airport's geographic coordinates (latitude and longitude). This data is useful to precisely locate and group different airports across our planet.

Flights : The flight table metadata has complete information about every flight. The FL_DATE field is for recording the date of a flight in format of YYYY-MM-DD. OP_CARRIER shows the operating commercial carrier's code and OP_CARRIER_FL_NUM indicates its flight number. TAIL_NUM represents the registration number of an aircraft. ORIGIN_AIRPORT_ID and DEST_AIRPORT_ID are unique identification numbers for the origin and destination airports, given by US Department of Transportation (DOT). The IATA codes for these airports can be found in ORIGIN and DESTINATION, while the respective city names are specified in ORIGIN_CITY_NAME and DEST_CITY_NAME. DEP_DELAY and ARR_DELAY show the variations in minutes between scheduled and factual times of departure or arrival, having negative numbers for flights that depart or arrive early. If a flight is canceled, it will have 1 in the CANCELLED field. The duration of flight is given by AIR_TIME in minutes and DISTANCE measures how far apart origin airport from destination airport are - both these measurements are shown as miles. At last, OCCUPANCY_RATE represents the occupancy rate of flight. This metadata gives all required details for studying performance and trends of flight.

Tickets : For the tickets table metadata, it shows us that each itinerary has a unique identifier called ITIN_ID. The year of the trip is represented by YEAR and QUARTER tells us which quarter it belongs to in that particular year. In ORIGIN, we find IATA airport code for the starting airport. ORIGIN_COUNTRY and ORIGIN_STATE_ABR show the country and state abbreviation of this airport. The full name of origin airport's state is given by ORIGIN_STATE_NM. ROUNDTRIP is a binary field (1 for round trip, 0 for one-way), with emphasis placed on round trips during analysis. REPORTING_CARRIER has the two-character code for the reporting airline carrier. PASSENGERS tells us how many passengers are on this itinerary, and ITIN_FARE stands for the fare per person. When ROUNDTRIP is 1, it signifies that we are dealing with a round trip and thus this applies to represent all of the full travel expenses for both ways; when ROUNDTRIP takes value 0 (zero), then it means we have a one-way journey so only basic cost without any return trip calculations will be shown by this parameter. DESTINATION includes the IATA airport code for the airport where one is going. This information lets us do detailed study on travel habits, cost arrangements and quantity of passengers depending upon itineraries.

Data Cleaning

The process of data cleaning has three main steps. In the flights dataset, all flights that got canceled are taken out by filtering records where the CANCELED column is showing 1 to make sure only finished flights are being analyzed. After this, tickets dataset gets filtered for round trips and we use the ROUNDTRIP column for picking records having a value of 1 which matches with the analysis needed to only consider round-trip route planning. The airport_codes dataset is finalized by filtering the TYPE column to keep rows where type is 'medium_airport' or 'large_airport'. This helps to concentrate the examination on airports that have a moderate or substantial size, ensuring data relevance for later analysis.

Data Merging (joining the tables)

The processes of combining data and removing columns are necessary to bring together and improve the datasets for analysis. At first, the flight dataset is merged with the airport_codes dataset using a left join on origin airport code. This operation occurs twice to make sure all relevant origin airport details are attached to the flights dataset, any existing columns from airport_codes will be added with the suffix '_ORIGIN'. Then, the flight dataset is merged one more time with the airport_codes dataset based on destination airport code. It appends the destination airport details with the suffix '_DESTINATION'. For drop operations, a set of columns to be removed is created. This includes duplicate or unneeded information like names of airports, their coordinates and country data from both origin and destination merges. These selected columns are then conditionally deleted from the flights dataset so only columns present in the dataframe are left behind. This method is concentrated on making the flight data more complete by adding important airport details and cutting extra information, which makes sure that we have a simplified set of data for future study. This merging strategy stands out because it's smoother and simpler than combining with the tickets dataset, which can be very big and complicated.

Data Preparation and Calculated Fields

The following process does a complete study to check the profit of different flight paths, using data from flights and tickets. At first, it includes all costs for every flight like fuel, maintenance charges, depreciation amount as well as insurance coverage along with airport operational expenses and delays. All these costs are added up to give calculated_total_cost for each flight. Next, the flights dataset is sorted based on origin and destination airports. This helps to find the average total cost, rate of occupancy, arrival delay and total number of trips for each route. The combined metrics are put together in one dataframe for a complete overview of every flight route's operational efficiency.

Now, we handle the tickets dataset to find out total and average revenue for each passenger on a round trip. We group data by starting and ending airports, this gives us the sum of round trip money from all flights along with how many passengers there are. Using these figures helps in finding average income per person. The two-way journey data is combined by joining details of outbound and return flights, adding up the typical total costs, averaging occupancy rates as well as arrival delays plus figuring out the minimum number of total trips for every route that makes a full circle. In the last stage, this processed round journey flight data is combined with ticket revenue data. Profitability can be found by using the average round trip revenue for every traveler, mean occupancy rate and more money from increased occupancy - total average cost against total revenue. The resulting data frame gives a thorough profitability analysis for every route to identify which routes are most profitable and least.

Afterwards, the process performs route normalization. This means organizing the origin and destination pair for every row in the dataframe. It gives a fresh column named 'Normalized_Route'. The main aim of this stage is to make the routes uniform, even if they have different orderings for starting and ending points. After sorting, DataFrame becomes grouped by normalized routes and many summary statistics are calculated for every group like mean total average cost, mean occupancy rate etc. Aggregation procedure gathers all the separate data and makes it smaller in size; it gives us information about typical performance measurements for each distinct route.

Solution (Analysis)

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.*

Answer : Following the data processing steps mentioned above, we sorted the DataFrame according to profitability. The goal of this sorting was to find out which routes are most profitable within our dataset by making the number of round trips per quarter the main sorting factor. Then, we extracted the top ten results for understanding which are ten busiest routes from these based on their profitability.

This method helps us not just to measure how much travel happens on different routes, but also understand if these routes are making profit or not. This is very useful for making smart decisions about strategy.

Result

LAX - SFO
LGA - ORD
LAS - LAX
JFK - LAX
LAX - SEA
BOS - LGA
HNL - OGG
PDX - SEA
ATL-MCO
ATL-LGA

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.*

Answer : Secondly, while we were doing the data processing steps that I mentioned earlier, our DataFrame also went through a sorting process. This was useful for identifying the top ten routes which were most profitable. When studying the first ten records from this sorting process, we obtained significant understandings about those routes that are crucial in determining overall profitability of business – it showed us essential money-making avenues within our study area.

Result :

Origin	Destination	Total Profit per round trip($)	Total Revenue per round trip($)
SLC	TWF	633358.645	669283
CLT	FLO	230222.750	263822
EGE	JFK	116062.230	186284
MDT	PHL	88996.020	122543
EYW	ORD	79559.365	134785
DEN	SUN	71905.380	114225
ISP	PHL	63361.395	97823
DEN	MOT	60265.755	106205

CLT	GSP	58183.865	101561
DFW	KOA	57505.325	102751

Question 3 : The 5 round trip routes that you recommend to invest in based on any factors that you choose.
Answer : These suggestions were found after filtering and sorting the data. Looking at routes with least average arrival delays shows which ones are most reliable for customers, possibly increasing their satisfaction and likelihood to return. The arrangement by profit per round trip helps in finding routes that bring back more money, guiding decisions towards making strategies for maximum income and profitability.
Result :

Route	Average_occupancy_rate	Avg_arrival_time_delay
SLC - TWF	0.6686	0.870780
FLO - CLT	0.6485	1.319914
EYW - ORD	0.6490	3.947585
HNL - LAS	0.6524	2.589789
KOA - DFW	0.6588	-0.543562

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.
Answer : When finding the break-even round trips for best suggested routes, you divide upfront cost by profit per round trip of each route. The total upfront cost is given as $90,000,000 which means this much money should be invested at start to run these routes. If we divide this cost by profit per round trip then it tells how many round trips are needed for covering initial investment and reaching break-even point. This calculation supposes that all profits made from every round trip help to balance out the initial expense until reaching the break-even point. The outcome is rounded off to the nearest whole number because it shows how many complete trips are needed for regaining back what was initially invested. The output, which is shown in DataFrame top_routes_summary has important parts like route's starting point and ending point along with profit per round trip as well as its equivalent break-even round trips. This conclusion gives useful understanding about the financial possibility and investment future of the suggested routes. It helps in making decisions related to resources distribution and operation planning.
Result :

Route	Breakeven Routes
SLC - TWF	143
FLO - CLT	391
EYW - ORD	776
HNL - LAS	1012
KOA - DFW	1132

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.

Answer :

Future KPI’s(Recommended):

1) *Market Expansion and tread:* To grow the customer base, this business might extend its route network by including more domestic and worldwide places to fly. Creating Strategic Alliances and Partnerships can assist in improving market scope and interconnectivity for attracting a bigger portion of customers. For fulfilling the need for Low-Cost Travel, it is crucial to provide competitive pricing combined with simplified services. Furthermore, if we observe the revival of business travel post the pandemic, it will make certain our company adjusts to alterations in requirements by corporate travelers - this encourages adaptability while ensuring growth endurance within aviation sector. Our company can increase their customer base by adding more domestic and international destinations to its route network. This is accomplished through forming Strategic Alliances and Partnerships, which assist in enhancing market reach and connectivity for attracting a larger share of customers. Regarding the demand for Low-Cost Travel, it is crucial to provide prices that are competitive and services which are streamlined. Furthermore, observing how business travel rebounds following the pandemic aids in securing adaptability - this ensures our company can effectively respond to alterations in needs from corporate travelers while guaranteeing sustainable growth within the aviation sector.
2) *Customer Experience:*Personal touch is very important for improving customer loyalty and satisfaction. It's about making services and experiences fit to each person by using data from every individual customer. When we use knowledge gained from customer data, like preferences or history, airlines can provide solutions that are customized for the passenger. This helps build a closer relationship between them and it makes the passenger more likely to stick with this brand; they also become loyal customers (Rahman & Islam, 2021). Furthermore, if we improve the In-Flight Experience by providing better

amenities such as food choices or entertainment options along with improved connectivity features on board flights - all these things help us to stand out in this competitive market (Aviation Planning Outlook 2020-2039 , 2022).

3) *Regulatory and Compliance:* Adherence to FAA Regulations is crucial. This means that we have to follow the rules set by Federal Aviation Administration (FAA) and other related authorities for safety and operational stability. The ability to change or flexibility holds great importance because new regulations about emissions, levels of safety and passenger's rights can affect operations significantly. Airlines need to keep pace with changes in law, adjusting their actions as required fast so as not only maintaining procedures for safety but also adapting within an evolving regulatory environment where they have responsibility towards standards set by different governing bodies.
4) *Technological Advancements:* In entering into Digital Transformation, one must involve the use of digital technologies within booking, customer service and operational areas. This is beneficial for boosting system efficiency as well as enhancing the customer experience. Utilizing Data Analytics with big data and analytics tools can assist airlines in comprehending what customers appreciate more, adjusting prices with increased accuracy and enhancing work process smoothness. The third method, grounded in data, permits airlines to make decisions dependent on facts and figures. This approach aids them in efficiently handling their resources while staying up-to-date with alterations within the field of aviation.

Assumption

1) Average Revenue was calculated by mean of occupancy rate by flight data and average ticket price per passenger by ticket data set grouping every column with origin and destination
2) Average cost operational cost of total flight per round trip was calculated by taking the mean of both average cost of trip from origin to destination and average cost of trio from destination to origin by grouping every column with origin and destination.
3) As there was no information given about the brand awareness and how much company is willing to spend to start in US market so recommendation was only made with respect to companies motto that is ‘’On time ,for you” and profitability of each round trip.