**Airline Data Challenge**

**GOAL:** To recommend 5 round trip routes between medium and large US airports for an airline company looking to enter US domestic market

**CLIENT's Moto:** "On time for you"

**Dataset for analysis:** 2019 Q1 flights data

**Quality Check:**

|  |  |  |
| --- | --- | --- |
| **Quality Insights** | **% Data** | **Assumptions and Actions** |
| Both airtime and arrival delays are null for non-cancelled flights | 0.2% | Probably cancelled, excluded |
| Departure delays of greater than 24 hours | 0.002% | Probably cancelled, excluded |
| Pre-departures <15 mins to 1hour from scheduled departure | 0.75% | Less likely, excluded |
| Roundtrip fare extremities (<50 USD and >10000 USD) as per BTS US domestic fare standards | 5.7% | Retained - Nullified fares & averaged by roundtrip route (Unknown if passenger applied coupons/ bulk booking/ on travel class preference) |
| Abnormal for airtime to be negative | 0.005% | Excluded |
| Distance for each leg had non constants, negatives, invalids and nulls | 2% | Retained by Mode level imputation |
| IATA code nulls for filtered medium and large US domestic airports | 4% | Excluded, as limited not to use extra dataset |

**AirportCode.csv**

**Fields of Interest** **–** Type, Country, IATA Code

The country column had 247 nulls but all of them belonged to Africa continent as known by the continent column. I went ahead and filtered medium and large US domestic airports using type and country fields in which I filtered duplicates. In the filtered list, 37 entries had IATA code as null. I explored if I could fill the nulls based on municipality column. We have municipalities in airports file and cities in flights file but same city/municipality can have multiple airports. E.g., New York can have 2 airports - JFK and LGA. Since we are limited not to use extra datasets, we will drop the nulls in IATA code. So, excluding that I finalized the filtered records with Type, Country, IATA Code columns as final preprocessed data frame used for analysis.

**Tickets.csv**

**Fields of Interest** **–** Origin, Destination, ITIN Fare, Roundtrip

I first filtered only the round-trip tickets for my analysis and noticed that all the data was pertaining to 2019 Q1 as known by quarter and year columns. Invalid fare details were there for ~1200 records or ~0.17% which was handled during numeric conversion by null substitution. I noticed that the fare column had maximum fare of around ~40K USD and 25% of data was less than 280 USD. As per Bureau of Transportation Statistics (BTS) for 2019Q1, fares above 10K USD seems too expensive for roundtrip fare for US domestic market. It’s unknown from the data if the passenger had applied coupons or which travel classes (first class or not) they travelled or if bulk booking was done. So as per BTS 2019Q1 and owing to above assumption, excluded <$50 round trip fare that is too less for US domestic flights and >$10000 round trip fare that is too expensive for US domestic flights, hence have nullified such abnormal fares. After above steps %nulls for average round trip fare column is around ~5%, but we will be further grouping by ORIGIN-DEST pairs to get fair enough average fare, so we will handle those nulls in next steps.

**Flights.csv**

**Fields of Interest** **–** Origin, Destination, Dep\_Delay, Arr\_Delay, Distance, Occupancy\_Rate

Flights file has only got 2019-1Q data in which I excluded cancelled flights for our analysis. I checked for invalid non-numeric values by calling custom function in the following columns of interest which is required for aggregation. I converted required columns to numeric which got rid of above invalid values. Distance field ranged from -2000 until 10000 which is unusual and 2680 are nulls. Since we have 1.8 million records, I grouped by origin-destination pair to identify mode and perform mode level imputation for distance which fixed above issues (nulls, outliers, missing values, negatives, and skewness) as well as skewness in distance field.

Even after imputation, there was just one origin-destination pair i.e., MIA-ABQ which is said to have a distance of 10000 miles way beyond 6000 miles. MIA-ABQ is ~2000 miles and here its 5X of its original value which is clearly an outlier. Hence, I removed those 10 records. Post imputation for distance, invalids, nulls were handled, and the distribution has become left skewed. Since as we know, Boston to Honolulu is the longest domestic flight in US - 11 hours - 5095 miles.

I noticed that some records had both the airtime AND arrival delay fields as null, which is equivalent to flight getting cancelled. So, 4377 records of ~1.8Million are probably cancelled flights which were marked instead as not cancelled. So, I dropped those probably cancelled flights from my analysis. Also, its abnormal for airtime to be negative which is 0.005% of dataset and that too for 1 route, JFK-ORD and, hence removed them. There were records with <10 min airtime but it’s still feasible. So, left them as such since actual airtime for Alaska-Petersburg is just 8 mins. Certain extremities have airtime of 18 hrs. which is huge for US domestic flights. But by arrival delay we can say that majority of flights arrived within more or less 5 mins of expected arrival, so these can be considered as expected air times.

I analyzed the relationship between airtime and arrival delay and got to know that airtime range is between 0-800 predominantly with variation delay between 0 and 200. There is not much variation in arrival delay with airtime. Flights can depart few mins early, but otherwise very early departure is very less likely to happen. Hence, I excluded. Almost 44 flights have departed well later than 24 hrs/ 1440 mins which is less likely to happen without being cancelled. Since we don’t have delay reasons excluding such data (0.002%) from analysis.

**Meta data for new fields in flights file:**

I created new fields in flights file as follows.

**ON\_TIME:** Created a new column as ON\_TIME where, if ARR\_DELAY is less than 15 mins its on-time else as delayed (as per Bureau of Transportation Stats standards)

**DELAY\_GRT\_30MINS:** Created a new column as DELAY\_GRT\_30MINS to identify flights with longer delays of more than 30 minutes of arrival delay.

**Data Munging:**

**Merge 1:** I joined the filtered airport data frame with that of tickets data frame by merging by origin and destination fields in tickets to that of IATA code in airport which gives us the fare details for passengers travelling between medium and large US domestic airports. For the round-trip fare tickets, I further grouped at origin-destination pair level for each leg of travel to find out average round trip fare. Since LAS – LAX and LAX-LAS correspond to single round trip, we have averaged the fares to find global average for each unique round trip.

**Merge 2:** I grouped the flights data frame at origin- destination pair level i.e., for each travel leg to find out total flights, average occupancy rate, average arrival departure delay, average distance. We are considering only the legs which has more than 50 flights in Q1 2019. I merged this data frame with that of output from “Merge 1” on origin and destination. I merged in such a way that each round trip of tickets data in “Merge 1” data frame is joined with corresponding two legs of that round trip in grouped flights data frame. I further grouped the flights data frame at each round-trip level to find the average flights, average occupancy rate, average arrival departure delay, total distance for both the legs of that round trip.

I calculated the total profit, revenue, cost and breakeven based on the provided instructions.

**Data frame Metadata:**

**flights\_rtr\_grp\_top10:** contains busiest roundtrip routes

**flights\_rtr\_profit\_top10:** contains profitable roundtrip routes with breakeven value

**recommended\_routes:** contains recommend route details

roundtrip\_route: roundtrip routes

avg\_flights: average roundtrip flights

avg\_carriers: average number of carriers in round trip route

avg\_dep\_delay: average departure delay of round-trip route in mins

avg\_arr\_delay: average arrival delay of round-trip route in mins

avg\_rtr\_fare: average roundtrip fare in USD

total\_revenue\_usd: total revenue for each roundtrip route in USD

total\_cost\_usd: total revenue for each roundtrip route in USD

total\_profit\_usd: total profit for each roundtrip route in USD

total\_profit\_per\_carrier\_usd: total profit per carrier for each roundtrip route in USD

breakeven: breakeven of upfront airline cost in terms of round-trip flight frequency