**Metadata Information**

|  |  |
| --- | --- |
| Table Names | Description |
| **Airport Subset** | Filter out the large and medium airport types from the airports data frame |
| **US\_airport\_subset** | Filter only for airports with ISO Country as US from the airports\_subset data frame |
| **Tickets (1)** | Filter only for Roundtrips which is merged with the US\_airport\_subset on the IATA Codes for both origin and destination airport city. |
| **Avg\_ticket\_price** | Used to calculate the average ITIN\_FARE for each trip then renaming column ITIN\_FARE to AVG\_PRICE. |
| **Tickets (2)** | Merged with Avg\_ticket\_price on the Trip column to add the new column AVG PRICE for each Trip. |
| **Flights (1)** | Merged with the US\_airport\_subset data to include all IATA CODES for both ORIGIN and DESTINATION in the flights data frame. |
| **Avg\_flight\_stat** | Used to group the flights data frame by each trip and imputing with the median for the ARR\_DELAY, AIR\_TIME, DISTANCE and OCCUPANCY\_RATE column. |
| **Flights (2)** | Merged with avg\_flight\_stat data frame on the Trip col. |
| **Flights\_agg** | A data frame from flights dataset grouped by each trip with the respective columns. |
| **Round\_Trip\_flights** | A data frame which is a self-join on the flights\_agg dataframe to take care of both the onward trip and the return trip(both the legs of the roundtrip) |
| **Trip\_agg** | A data frame grouped by each Round\_TRIP with the ITIN FARE CLEAN imputed with the mean. |
| **Trip\_data** | Merge of round\_trips and tickets\_agg |
| **Late\_pct** | Grouping the flights data frame by each Trip which is aggregated by the mean of late fees |
| **trip\_data** | The merge of trip\_data and the late\_pct on ROUND\_TRIP column |

|  |  |
| --- | --- |
| Column Names | Description |
| IS\_CLEAN\_FARE | New column in the tickets data frame to check whether the is\_clean function takes care of cleaning the ITIN\_FARE column. |
| TRIP | New column in the tickets data frame and the flights dataset which is combination of the origin and destination airport city abbreviations. |
| MEDIAN\_ARR\_DELAY | Renamed from ARR\_DELAY after median imputation |
| MEDIAN\_AIR\_TIME | Renamed from AIR\_TIME after median imputation |
| MEDIAN\_DISTANCE | Renamed from DISTANCE after median imputation |
| MEDIAN\_OCCUPANCY\_RATE | Renamed from OCCUPANCY\_RATE after median imputation |
| COST | US\_airport\_subset column is used to add the airport maintenance cost for medium airport which is $5000 and for large airports is $10000 that is renamed to Origin Cost and Destination Cost for the respective airport city in the roundtrip. |
| LAT | A column in US\_airport\_subset data frame which is used to store the Latitude from the original coordinate’s column |
| LONG | A column in US\_airport\_subset data frame which is used to store the Longitude from the original coordinate’s column |
| MISC COST | A column in the flights data frame which is used to calculate the miscellaneous costs which are 8 USD per mile (Fuel, Oil, Maintenance, Crew) + 1.18 USD per mile (Depreciation, Insurance, Other) |
| DEP\_LATE\_FEES | A column in the flights data frame that is used to calculate Late Fees charged to the airline for delay in departure. First 15 minutes delay are free. Post that 75 USD charges on each minute delay |
| ARR\_LATE\_FEES | A column in the flights data frame that is used to calculate the Late Fees charged to the airline for delay in arrival. First 15 minutes delay are free. Post that 75 USD charges on each minute delay |
| LATE\_FEES | A column in the flights data frame which is the sum of DEP\_LATE\_FEES and ARR\_LATE\_FEES |
| IS\_Late | A column in the flights data frame to calculate the statistics for flights that are late. |
| KEY | A column to swap the Trip column (SFO-LAX to LAX-SFO) |
| AVG\_ARR\_DELAY | A column in the round\_trip\_flights data frame which is the average of (ONWARD\_ARR\_DELAY + RETURN\_ARR\_DELAY) /2 for a roundtrip |
| AVG\_DEP\_DELAY | A column in the round\_trip\_flights data frame which is the average of (ONWARD\_DEP\_DELAY+ RETURN\_DEP\_DELAY)/2 for a roundtrip |
| AVG\_OCCUPANCY\_RATE | A column in the round\_trip\_flights data frame which is the average of (ONWARD\_OCCUPANCY\_RATE + RETURN\_OCCUPANCY\_RATE)/2 for a roundtrip |
| DISTANCE | A column in the round\_trip\_flights data frame which is the average of (ONWARD\_DISTANCE + RETURN\_DISTANCE)/2 for a roundtrip |
| AIRTIME | A column in the round\_trip\_flights data frame which is the average of is the average of (ONWARD\_AIRTIME+ RETURN\_AIRTIME)/2 for a roundtrip |
| NUM\_AIRLINES | A column in the round\_trip\_flights data frame which is the average of (ONWARD\_NUM\_AIRLINES+ RETURN\_NUM\_AIRLINES)/2 |
| FLIGHT\_COUNT | A column in the round\_trip\_flights data frame which is the ONWARD\_COUNT and RETURN\_COUNT for a roundtrip |
| LATE\_FEES | A column in the round\_trip\_flights data frame which is the sum of ONWARD\_LATE\_FEES and RETURN\_LATE\_FEES for a roundtrip |
| AIRPORT\_COST | A column in the round\_trip\_flights data frame which is sum of ONWARD\_DESTINATION\_COST and RETURN\_DESTINATION\_COST for a roundtrip |
| MISC\_COST | A column in the round\_trip\_flights data frame which is sum of ONWARD\_MISC\_COST and RETURN\_MISC\_COST for a roundtrip |
| TRIP\_COST | A column in the round\_trip\_flights data frame which is sum of LATE\_FEES, AIRPORT\_COST and MISC\_COST for a roundtrip |
| ROUND\_TRIP | A column in the round\_trip\_flights data frame which is a swap of the TRIP col |

**Revenue Metrics**

* *Number of passengers = average occupancy rate \* 400*
* *Base Revenue = number of passengers \* fare per passenger*
* *Luggage Revenue = (number pf passengers \* 70)/2*
* *Trip Revenue = Base Revenue + Luggage Revenue*
* *Trip Profit = Trip Revenue – Trip Cost*

|  |  |
| --- | --- |
| Column Names | Description |
| NUM\_PASSENGERS | A column in the trip\_data data frame which is the AVG\_OCCUPANCY\_RATE \* 400(200 passengers for onward trip and 200 passengers for the return trip) |
| BASE\_REVENUE | A column in the trip\_data data frame which is NUM\_PASSENGERS \* FARE\_PER\_PASSENGER (renamed from ITIN\_FARE\_CLEAN |
| LUGGAGE\_REVENUE | A column in the trip\_data data frame which is NUM\_PASSENGERS \* 70/2 ( 50% of passengers have baggage) |
| TRIP\_REVENUE | A column in the trip\_data data frame which is the sum of BASE\_REVENUE and LUGGAGE\_REVENUE |
| TRIP\_PROFIT | A column in the trip\_data data frame which is the difference between TRIP\_REVENUE and TRIP\_COST |

**Profit Metrics**

* *Total Profit = Flight Count \* Trip Profit*
* *Total Cost = Flight Count \* Trip Cost*
* *Total Revenue = Flight Count \* Trip Revenue*

|  |  |
| --- | --- |
| Column Name | Description |
| TOTAL\_PROFIT | A column in the trip\_data data frame which is FLIGHT\_COUNT \* TRIP\_PROFIT |
| TOTAL\_COST | A column in the trip\_data data frame which is FLIGHT\_COUNT \* TRIP\_COST |
| TOTAL\_REVENUE | A column in the trip\_data data frame which is FLIGHT\_COUNT\* TRIP\_REVENUE |

**Recommendations for:**

**10 busiest routes**

* *Busiest Routes = Display top 10 of sort values (by='Flight Count', descending=True)*

|  |  |
| --- | --- |
| Column Name | Description |
| Busiest routes | A column in the trip\_data data frame which is sorting trip\_data by flight count in the descending order |

**10 profitable routes**

* *Profitable Routes = Display top 10 of sort\_values (by='Average Profit', descending=True)*

|  |  |
| --- | --- |
| Column Name | Description |
| Profitable Routes | A column in the trip\_data data frame which is sorting by avg profit in the trip data that is converted to int in the descending order |

**Investment on the Top 5 Routes**

* *Filter List = Busiest routes + Profitable Routes*
* *Recommended Trip = Display top 5 of sort\_values (by='Late Percentage') in the 'Roundtrip' column*

|  |  |
| --- | --- |
| Column/Table Name | Description |
| Filter list | A data frame with a combination of the busiest and profitable list |
| Trips\_subset | A data frame that identifies only ROUNDTRIP in the trip data which is in the filter list with a condition that the Num of airlines being > 3 |
| Recommended\_trip | A column in the trip subset sorted in the ascending order by Late PCT |

\*\* *Select roundtrips that fall under the Filter List with a condition that the Number of Airlines are greater than 3*

**Flights To Breakeven the Upfront Airplane cost ($ 90 million)**

* *Flights to breakeven = 90\*1000000/ Trip Revenue*
* *Average Flights = Flight Count/ Number of airlines*
* *Number of Quarters taken to breakeven = Flights to breakeven/ Average Flights*

|  |  |
| --- | --- |
| Column Name | Description |
| FLIGHTS\_TO\_BREAKEVEN | A column in the recommended\_trip data frame which is 90\*1000000/ TRIP\_REVENUE |
| AVG\_FLIGHTS | A column in the recommended\_trip data frame which is the FLIGHT\_COUNT/ NUM\_AIRLINES |
| NUM\_QUARTER\_TO\_BREAKEVEN | A column in the recommended\_trip data frame which is FLIGHTS\_TO\_BREAKEVEN/ AVG\_FLIGHTS |