**Exploratory Data Analysis**

**PROJECT TITLE**

**ESTIMATING THE FLIGHT PRICE**

**By**

**Manasa Dontireddy -11810613**

**Context of Problem Statement:**

The airline business is so competitive; airlines need to comprehend price dynamics to maximize profits and raise customer satisfaction for an Agency. This dataset includes details about several flights, including their costs, which might vary depending on several criteria like:

1. The Airline dataset consist of various attributes of different airline service providers and their associated pricing strategies.
2. Travel characteristics: Airline prices are competitive and sensitive towards various characteristics such as on-spot booking versus advance booking, duration of light, class type of airline, holiday season versus non-holiday seasons etc.,
3. In-flight amenities: Customers are highly sensitive to airline prices in making their decision either to book it or not. In-general, amenities like meals, WIFI, checked bags, and in-flight entertainment plays a crucial role in booking status.
4. External events: External events such as weather, traffic have certain degree of variations on the flight prices.

**Objectives of Exploratory Data Analysis:**

The following are the main goals of the EDA on this dataset:

1. Recognize Pricing Variability: If helps to analyze how airline, class type, day of the week, time of day, and length of flight affect cost of flight.
2. Find Trends and Patterns: Keep an eye out for pricing trends related to peak travel times (holidays or weekends) or reservation practices.
3. Find Outliers: Check to see if any prices stand out that might point to incorrect pricing or premium services.

**Step-1: Data Dictionary:**

The **“Spring 2024 Airline Dataset”** includes different types of airline information related to prices and its metadata. The dataset covers 10000 records and has distinct features (22). Each feature consists of binaries, numerical and categorical values.

The below [**Table\_1**](#Table_1) data dictionary gives an overview of various attributes, types of data, description and sample values of the Airplane Prices

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Type of Data** | **Description** | **Sample Value** |
| Flight ID | Numerical | Unique identifier for each flight. | 1,2,3 |
| Airline | Categorical | Airline operating the flight. | American Airlines, Delta, United |
| Origin Airport | Categorical | Airport from where the flight departs. | HND, ORD, DFW |
| Destination Airport | Categorical | Destination airport. | LHR, DFW, JFK |
| Distance | Numerical | Distance of the flight in miles. | 339,322,431,245.60 |
| Date Of Flight | Date | Date of the flight. | 2/20/2021,3/6/2022 |
| Day Of Week | Categorical | Day of the week of the flight | Saturday, Monday |
| Time Of Day | Categorical | Time of day of the week | Morning, Afternoon, Evening |
| Flight Duration | Numerical | Duration of the flight in hours. | 1.87,2.24 |
| Aircraft Type | Categorical | Type of aircraft used for flight. | Airbus A380, Boeing 777 |
| Class Type | Categorical | Class of the ticket | Economy, Business, First Class |
| Seat Availability | Categorical | Seat availability in the flight | High, Medium, Low |
| Advance Booking Days | Numerical | Number of days the ticket was booked in advance. | 234, 56,78 |
| Holiday Period | Categorical | Whether the flight is in a holiday period | Yes, no |
| Luggage Allowance | Numerical | Luggage allowance in kilograms. | 18, 34 |
| Meal Included | Categorical | Whether a meal is included | Yes, no |
| WIFI Available | Categorical  Categorical | Whether WIFI is available | Yes, no |
| In Flight Entertainment | Categorical | Availability of in-flight entertainment | Yes, no |
| Fuel Surcharge | Numerical | Fuel surcharge (in USD). | 70,85, 23,45 |
| Airport Traffic | Categorical | Traffic level at the origin airport | High, Medium, Low |
| Weather Conditions | Categorical | Weather conditions during the flight | Clear, Rainy, Snowy |
| Flight Price | Numerical | Price of the flight ticket in USD (Dependent Variable). | 592.34, 123,45,678 |

**Table\_1. Data dictionary and its sample values**.

**Step-2: Data Inspection**

As a second step, need to check for any kind of missing values or NULLs in the dataset. Also, if any outliers are present, we can highlight those to find out the deviations. This helps to get accurate information. The list of outliers, null values and total filed count is displayed in [Table\_2](#Table_2_DI)

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Missing/Null Value** | **Outliers [Yes/No]** | **Total Field count** |
| Flight ID | 0 | No | 10000 |
| Airline | 0 | No | 10000 |
| Origin Airport | 0 | No | 10000 |
| Destination Airport | 0 | No | 10000 |
| Distance | 0 | No | 10000 |
| Date Of Flight | 0 | No | 10000 |
| Day Of Week | 0 | No | 10000 |
| Time Of Day | 0 | No | 10000 |
| Flight Duration | 0 | No | 10000 |
| Aircraft Type | 0 | No | 10000 |
| Class Type | 0 | No | 10000 |
| Seat Availability | 0 | No | 10000 |
| Advance Booking Days | 0 | No | 10000 |
| Holiday Period | 0 | No | 10000 |
| Luggage Allowance | 0 | No | 10000 |
| Meal Included | 0 | No | 10000 |
| WIFI Available | 0 | No | 10000 |
| In Flight Entertainment | 0 | No | 10000 |
| Fuel Surcharge | 0 | No | 10000 |
| Airport Traffic | 0 | No | 10000 |
| Weather Conditions | 0 | No | 10000 |
| Flight Price | 0 | No | 10000 |

**Table\_2: Data Inspection – Finding Nulls or Outliers**

**Step-3: Descriptive Statistics:**

As a third stage calculated the Descriptive Statistics (Mean, Mode, Median, Maximum Value, Minimum value, Standard Deviation, Range, Variance, Skewness, Total Count) for the numerical features and are displayed in [**Table\_3**](#Table_3).  
Descriptive statistics are used to more easily examine, summarize, and comprehend patterns and trends within the dataset.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive**  **Metrics** | **Duration** | **Distance** | |  | | --- | | **Advance**  **Booking Days** | | |  | | --- | | **Luggage**  **Allowance** | | |  | | --- | | **Fuel**  **Surcharge** | | |  | | --- | | **Flight**  **Price** | |
| Average | 8.02245334 | 4002.655 | 182.0954 | 21.947 | 54.49919481 | 1026.381 |
| Median | 8.03019094 | 3977.8289 | 181.5 | 22 | 54.49717445 | 1030.798 |
| Mode | #N/A | #N/A | 123 | #N/A | #N/A | #N/A |
| Standard Deviation | 4.04626371 | 2290.671 | 105.724644 | 4.30342224 | 25.7812542 | 559.8764 |
| Variance | 16.37225 | 5247175 | 11177.7003 | 18.5194429 | 664.673066 | 313461.6 |
| Maximum | 14.9966236 | 7999.6440 | 364 | 29 | 99.99524101 | 1999.882 |
| Minimum | 1.00041090 | 100.51747 | 105.719357 | 15 | 25.77996507 | 559.8484 |
| Range | 13.9962127 | 7899.1266 | 14 | 364 | 89.9852 | 1949.674 |
| Total Count | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| 25th percentile | 4.5365858 | 1994.904 | 90 | 18 | 32.0699185 | 548.9523 |
| 50th percentile | 8.03019094 | 3977.828 | 181.5 | 22 | 54.4971744 | 1030.797 |
| 75th percentile | 11.5521856 | 5960.401 | 274 | 26 | 76.8186323 | 1499.946 |
| Skewness | 0.01205035 | 0.0203444 | 0.00504767 | 0.01555916 | 0.007789497 | -0.01342 |

**Table\_3. Descriptive Statistics for Numerical Overview**

**Step-4: Visualization of the plots**

From the [**Figure\_1**](#Figure_1)and[**Table\_4**](#Table_4) we can infer that the number of trips/flights ran by the United airlines is higher than any other airlines. The minimum number of trips ran by the Lufthansa for the given years.

A table with numbers and names

Description automatically generated

**Table\_4: #flights/trip frequency of the airlines**

**Figure\_1. Bar plot representation of the frequency plots.**

The below [**Figure\_2**](#Figure_2)and[**Table\_5**](#Table_5) we can infer that the maximum number of flights were driven by the United Airlines this is in line with the Table-4 as well. However, the above table provides the detailed view of how United Airlines ran a greater number of flights during the year 2022 and 2023.

A table with numbers and words

Description automatically generated

**Table-5: #flights/trip frequency based on the class Type of a given airlines**

Figure-2: Clustered Column Bar plots for #flights/trip frequency based on the class Type for airlines

The below [**Figure\_3**](#Figure_3)and[**Table\_6**](#Table_6) helps in understanding the average prices on various dimensions of the flight and trip attributes - whether WIFI or meal included or not. Interestingly,

United and Southeast airlines have higher prices even with No Meal included in it. Also, with or without meal/WIFI inclusion, the average price for all airlines is in between 1020 to 1040

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Airlines** | **WIFI-No** | **WIFI-Yes** | **Meal-No** | **Meal-Yes** |
|  | Air France | 1031.008389 | 1030.26517 | 1023.67188 | 1037.30866 |
|  | American Airlines | 1010.081321 | 1030.94705 | 1030.19022 | 1010.91037 |
|  | British Airways | 1048.006525 | 1013.05265 | 1017.07539 | 1043.98862 |
|  | Delta Air Lines | 1008.086614 | 1039.07343 | 1025.18156 | 1021.52672 |
|  | Emirates | 991.6608798 | 1025.77551 | 1023.88264 | 994.720817 |
|  | Lufthansa | 1034.947069 | 1018.09494 | 1027.50109 | 1024.94121 |
|  | Qatar Airways | 1047.579264 | 1013.91717 | 1033.15102 | 1030.5595 |
|  | Singapore Airlines | 1024.687557 | 1033.23409 | 1031.13362 | 1026.81312 |
|  | Southwest Airlines | 1020.513733 | 1033.16834 | 1055.84151 | 998.84219 |
|  | United Airlines | 1045.234253 | 1027.291 | 1050.98289 | 1021.85366 |

Table\_6: Prices of airlines with or without WIFI and Meal Inclusion

A graph of different colored bars

Description automatically generated

Figure-3: Clustered Bar plots for the prices for all the airlines – WIFI and Meal

The below [**Figure\_4**](#Figure_4)and[**Table\_7**](#Table_7) shows the average frequency of flights ran during afternoon, evening and morning times. More number of flights is noticed during Afternoon and morning times.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Time Of Day** | **Sum of mean** |
|  |  | Afternoon | 10348.2318 |
|  |  | Evening | 10175.4233 |
|  |  | Morning | 10346.6533 |
|  |  | Night | 10185.8484 |

**Table\_7: Sum of means based on the Time of Day**

A graph with red lines

Description automatically generated

**Figure\_4: Bar plots of Sum of means based on the Time of Day**

The below [**Figure\_5**](#Figure_5)and[**Table\_8**](#Table_8) shows the sum of frequency of all flights ran during the weekdays. More number of flights is noticed during Monday, Wednesday and Thursday. Less flights seen on Tuesday.

A table with numbers and a few words

Description automatically generated

**Table\_8: Sum of Means based on Week**

**Figure\_5: Bar plots of sum of means based on Weekdays**

The below [**Figure\_6**](#Figure_6) shows the maximum and mean frequency of given flights during the weekdays. From the below figure we can notice that British Airways has the lowest price during Saturdays. Interestingly the maximum price for majority of airlines is either on Monday or Tuesday or Wednesday. Also, the mean price across all the airlines is beyond $1000. However, we can notice that the mean prices for Lufthansa and Southwest airlines during weekends are higher. Similarly, the lowest mean price for the same airlines falls below $1000 during Tuesday.

**Figure\_6: Max and Mean of flight prices over weekdays**

The below [**Figure\_7**](#Figure_7)and[**Table\_9**](#Table_9) illustrates how average flight costs vary depending on the weather. This can assist in determining whether specific weather patterns are linked to increased or decreased pricing.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Weather Labels** | **Average of Flight Price** |
|  |  | Clear | 1028.846603 |
|  |  | Cloudy | 1035.700282 |
|  |  | Rainy | 1021.390667 |
|  |  | Snowy | 1019.721509 |
|  |  | **Grand Total** | **1026.381289** |

**Table\_9: Average Flight Price varies on Weather conditions**

A diagram of weather conditions

Description automatically generated

**Figure\_7: Pie chart of average prices based on weather**

The next [**Figure\_8**](#Figure_8)and[**Table\_10**](#Table_10) displays the average distance that flight travelled to the destination port irrespective of origin port. The port “ATL” has travelled the highest among other and “HND” stands at the lowest rate.

|  |  |  |
| --- | --- | --- |
|  | **Destination Port** | **Average of Distance** |
|  | ATL | 4111.710031 |
|  | CDG | 3980.984454 |
|  | DFW | 3925.735108 |
|  | DXB | 3952.126209 |
|  | HND | 3909.728864 |
|  | JFK | 4009.117512 |
|  | LAX | 4046.108959 |
|  | LHR | 3986.231909 |
|  | ORD | 4070.860478 |
|  | SIN | 4041.223725 |
|  | **Grand Total** | **4002.655497** |

**Table\_10: Average flight distance to the destination port**

A graph with a line

Description automatically generated

**Figue\_8: Line chart of avg flight distance to destination port**

The next [**Table\_11**](#Table_11)and[**Figure\_9**](#Figure_9) shows the total number of flights per each aircraft type. Almost all flights fall under same aircraft type. The type “Boeing 737” falls highest among other types while “Airbus A380” is at lowest.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **AirCraft Type** | **Count of FlightID** |
|  |  | Airbus A320 | 2481 |
|  |  | Airbus A380 | 2479 |
|  |  | Boeing 737 | 2526 |
|  |  | Boeing 777 | 2514 |
|  |  | **Grand Total** | **10000** |

**Table\_11: count of Flight Ids per aircraft type**

A graph of flight statistics

Description automatically generated with medium confidence

**Figure\_9: Bar chart for total flights over aircraft type**

The next [**Figure\_10**](#Figure_10)and[**Table\_12**](#Table_12) illustrates total number of flights/trips based on years for each airline. All years has almost equal count. Emirates has slightly low count rate in the year 2024 and turned highest in the year 2021.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Airlines** | **2021** | **2022** | **2023** | **2024** |
|  | Air France | 265 | 253 | 266 | 250 |
|  | American Airlines | 272 | 260 | 228 | 264 |
|  | Delta Air Lines | 227 | 249 | 236 | 256 |
|  | Southwest Airlines | 238 | 252 | 231 | 253 |
|  | United Airlines | 251 | 273 | 282 | 251 |
|  | British Airways | 227 | 261 | 261 | 269 |
|  | Emirates | 279 | 250 | 276 | 219 |
|  | Lufthansa | 241 | 238 | 233 | 242 |
|  | Qatar Airways | 251 | 234 | 231 | 243 |
|  | Singapore Airlines | 255 | 264 | 248 | 221 |
|  |  |  |  |  |  |
|  |  |  | Min |  |  |
|  |  |  | Max |  |  |

**Table\_12: #flights/trip frequency based on the year of a given airlines**

A graph of different colored bars

Description automatically generated

**Figure\_10: Clustered Bar chart for flight count based on year**

The below [**Figure\_11**](#Figure_11) **and** [**Table\_13**](#Table_13) shows the total count of the inflight entertainment provided by each airline throughout different dates. The United Airlines has the highest count among other airlines.

|  |  |  |
| --- | --- | --- |
|  | **Airlines** | **Count of Inflight Entertainment** |
|  | Air France | 1034 |
|  | American Airlines | 1024 |
|  | British Airways | 1018 |
|  | Delta Air Lines | 968 |
|  | Emirates | 1024 |
|  | Lufthansa | 954 |
|  | Qatar Airways | 959 |
|  | Singapore | 988 |
|  | Southwest | 974 |
|  | United Airlines | 1057 |
|  | **Grand Total** | **10000** |

Table\_13: Total inflight entertainment count per each airline

A graph of different colored bars

Description automatically generated with medium confidence

**Figure\_11: Bar chart of count of Inflight Entertainment vs airlines**

The below [Figure\_12](#Figure_12) and [Table\_14](#Table_14) shows the total count of flights based on year and the average distance travelled. Highest number of flights travelled on year 2022 with greater distance while 2023 seems to be lowest.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Year** | **Average Distance** | **Count of Flights** |
|  | 2021 | 4017.547511 | 2506 |
|  | 2022 | 4037.698912 | 2534 |
|  | 2023 | 3970.04977 | 2492 |
|  | 2024 | 3984.476433 | 2468 |
|  | **Grand Total** | **4002.655497** | **10000** |

**Table\_14: Distance travelled by flight year wise**

A graph with numbers and a line

Description automatically generated

**Figure\_12: Histogram chart to count flights per year with average distance travelled**

The next [**Figure\_13**](#Figure_13) **and** [**Table\_15**](#Table_15) shows the average fuel surcharge used by each airline over past 4 years. Almost every airline has same average value. while emirates seem to be bit low when compared to other airlines.

|  |  |  |
| --- | --- | --- |
|  | **Airlines** | **Average of Fuel Surcharge** |
|  | Air France | 54.49420173 |
|  | American Airlines | 53.91077887 |
|  | British Airways | 54.98004303 |
|  | Delta Air Lines | 54.50441972 |
|  | Emirates | 55.47744999 |
|  | Lufthansa | 53.97671976 |
|  | Qatar Airways | 54.5478477 |
|  | Singapore. | 54.09582932 |
|  | Southwest | 55.08145598 |
|  | United Airlines | 53.92643418 |

**Table\_15: Average fuel Surcharge per each airline**

A graph of a graph showing the average fuel surcharge by airlines

Description automatically generated

**Figure\_13: Each airlines average Fuel surcharge value**

The below [**Figure\_14**](#Figure_14) **and** [**Table\_16**](#Table_16) shows the average price across each class type for the seat availability of low, medium and high. The seat “low” for the economy class flight price is high when compared to other class types. The first class “Medium” seems to be lowest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Class Type** | **Seat Avail - High** | **Seat Avail - Low** | **Seat Avail - Medium** |
|  | Business | 1015.018949 | 1042.667608 | 1025.80089 |
|  | Economy | 1020.119441 | 1050.573384 | 1041.461616 |
|  | First Class | 1025.861883 | 1013.860402 | 1004.684205 |

**Table\_16: Flight price across seat availability over each class**

A graph of seats available

Description automatically generated

Figure\_14: clustered bar chart for average flight price over each class type seat

The last [**Figure\_15**](#Figure_15) **and** [**Table\_17**](#Table_17) shows the total number of flights count for each airline that goes through the origin port. The origin port “ALT” and “LAX” has highest count for air France and United airlines. And “SIN” and “LHR” seems to be lowest.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Origin Port** | **Air France** | **American** | **British** | **Delta** | **Emirates** | **Lufthansa** | **Qatar** | **Singapore** | **Southwest** | **United Airlines** |
| ATL | 123 | 93 | 115 | 105 | 118 | 107 | 84 | 114 | 98 | 97 |
| CDG | 108 | 100 | 103 | 96 | 103 | 87 | 84 | 96 | 98 | 101 |
| DFW | 96 | 97 | 112 | 105 | 98 | 98 | 106 | 88 | 112 | 92 |
| DXB | 101 | 115 | 108 | 95 | 97 | 105 | 113 | 109 | 103 | 110 |
| HND | 107 | 114 | 100 | 84 | 93 | 97 | 97 | 113 | 86 | 100 |
| JFK | 87 | 105 | 99 | 104 | 114 | 109 | 105 | 95 | 108 | 106 |
| LAX | 102 | 97 | 96 | 88 | 114 | 83 | 105 | 100 | 91 | 121 |
| LHR | 90 | 91 | 94 | 98 | 80 | 76 | 74 | 84 | 103 | 106 |
| ORD | 117 | 115 | 88 | 92 | 107 | 90 | 78 | 88 | 101 | 117 |
| SIN | 103 | 97 | 103 | 101 | 100 | 102 | 113 | 101 | 74 | 107 |

**Table\_17: Total flight count per each airline for origin port**

A graph of different colored lines

Description automatically generated

**Figure\_15: Total flight count per airline vs Origin Port**

s

**Thank You,**

**Manasa Dontireddy**