**University of North Texas**

**ADTA 5130 Data Analytics**

Flight Price Prediction

**By:** Biniam Abebe

January 2024

**Content**

1. Introduction----------------------------------------------------------------------------------------
2. Data Description--------------------------------------------------------------------------------
3. Exploratory Data Analysis (EDA)-----------------------------------------------------------  
   1. Research Questions ------------------------------------------------------------------
   2. Analysis and Observation -----------------------------------------------------------
4. Hypothesis Testing-----------------------------------------------------------------------------
   1. ANOVA Test---------------------------------------------------------------------------
   2. Regression Analyses-------------------------------------------------------------------
5. **Introduction**

Many passengers who wish to book a ticket wonder which airline or when to choose. Determining the best time to buy a plane ticket can be difficult when little information about future pricing fluctuations is available. (Abdella and others, 375–391). Forecasts of demand determine how much airlines should charge. Airlines need to efficiently manage demand because there are only so many seats on an airplane. For instance, airlines may increase costs to reduce seat sales when demand outpaces capacity. Passengers can use these dynamics to forecast future airfare patterns and make well-informed choices. Predicting the cost of a flight depends on several parameters, such as its duration, destination, source, and arrival time.

1. **Objective**

Flight price prediction is critical for the agency and customers to purchase tickets with optimized prices and at the right time. We will use the dataset provided for the spring 2024 Airline with flight prices for various airlines between different destinations. To Solve the problem, one will do exploratory data analysis based on the EDA and questions explored. Hypotheses will be developed to create predictive models.

1. **Data Description**

First, look at the overall dataset. It has 10,000 samples/records 22 features/variables, eight numerical features, and 14 categorical features.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Distance | 10000.0 | 4002.7 | 2290.7 | 100.5 | 1994.9 | 3977.8 | 5960.4 | 7999.6 |
| FlightDuration | 10000.0 | 8.0 | 4.0 | 1.0 | 4.5 | 8.0 | 11.6 | 15.0 |
| AdvanceBookingDays | 10000.0 | 182.1 | 105.7 | 0.0 | 90.0 | 181.5 | 274.0 | 364.0 |
| LuggageAllowance | 10000.0 | 21.9 | 4.3 | 15.0 | 18.0 | 22.0 | 26.0 | 29.0 |
| FuelSurcharge | 10000.0 | 54.5 | 25.8 | 10.0 | 32.1 | 54.5 | 76.8 | 100.0 |
| FlightPrice | 10000.0 | 1026.4 | 559.9 | 50.2 | 549.0 | 1030.8 | 1499.9 | 1999.9 |

Table 1 Statistical summary of the numerical features data

Table 1 shows the dataset's information on numerical features such as distance, flight duration, advance booking days, and flight price. The statistical summary shows that the minimum price is $50.20, the maximum price is $1999.88, and the average cost is $1026.38.

1. **Exploratory Data Analysis (EDA)**
2. Research Questions  
   To further analyze the Spring 2024 Airline Dataset, a series of questions have been formulated based on provided statements to understand the dynamics of flight pricing.
3. Does airport traffic have an impact on flight prices?  
   This question examines the relationship between the Traffic level at the origin airport and the corresponding flight prices, i.e., if higher traffic levels are associated with higher or lower flight prices due to operational efficiencies.
4. Do flight prices vary by time of day?  
   Investigate whether there are patterns in flight pricing based on the departure time of day.
5. Do flight prices change by day of the week?  
   Check flight prices throughout the week to see if certain days are cheaper or more expensive.
6. What is the impact of the holiday season on flight prices?  
   Analyze how flight prices change during holiday seasons compared to non-holiday periods. This question captures the effects of increased travel demand during holidays on pricing.
7. Does the month of the flight affect its price?  
   Determine if there are monthly trends in flight pricing, which could be influenced by factors such as seasonal travel demand or holiday periods.
8. Analysis and Observation
9. Flight price by airport traffic

A blue rectangles with white text

Description automatically generated

Key Observations: The bar graph shows a slight price difference between the volume (low, medium, and high) of the airport traffic. The lesser the airport traffic, the lesser the price, and vice versa, the higher the airport traffic, the more elevated the flight price. So, by seeing these numbers, we can say that airport traffic impacts flight prices.

1. How do flight prices vary by time of day?

A screenshot of a computer screen

Description automatically generated

Key Observations: There is a price difference between the time of day of the flight. Flight prices are higher in the early morning and early afternoon than in the night and evening.

1. How do flight prices change by day of the week?

A blue and green stripes

Description automatically generated

Key Observations: The middle of the week will likely have higher flight prices, followed by the weekend and early weekdays. Tuesday and Friday are the two days when flights are at their lowest. Therefore, the day of the week also influences flight prices.

1. What is the impact of the holiday season on flight prices?

A screenshot of a graph

Description automatically generated

Key Observations: The holiday season is a peak travel period in which flight demand tends to be higher than other seasons. Unfortunately, the data doesn’t show that trend; the graph shows that the flight price during the holiday period is lower than on regular days. One reason could be early flight booking; people know that last-minute bookings during peak holiday times are likely to be more expensive, so they tend to plan and book their flights early.

1. How does the month of the flight affect its price?

A graph of a flight

Description automatically generated

Key Observations: As can be seen in the chart, flight prices peak in April, reaching about $1,060.90 and falling to approximately $982.86 in September. Also, it shows that flight prices can fluctuate throughout the year, with several peaks and troughs indicating seasonal changes. In February, April, June, and November, the average prices are higher than those of the months immediately preceding them. The chart shows various factors, including holiday periods, seasonal demand, and airline pricing strategies, that may affect flight prices.

1. **Hypothesis Testing**

**Limitation**

The data set we use has many categorical variables for the Regression test. We will be converting all categorical values to Dummy Variables (1,0). We will be using the function dummy\_cols(). (Kaplan ).

Research Questions and Result  
ANOVA Testing  
The following Tests are conducted with the following assumptions to be true.

* The population is Normally distributed.
* The population standard deviations are unknown but assumed equal.
* The samples are selected independently.

1. Does the meaning of flight price differ significantly across airlines?

The flight prices are the dependent variable, and the airlines act as the independent, categorical variable. The purpose is to uncover if there are statistically significant differences in average flight prices among different.

**Hypotheses**

H0: The Mean of flight prices do not differ significantly across airlines.

H1: At least one airline has a significantly different mean of flight price compared to others.  
A black screen with white text

Description automatically generated

Reading the Result:  
**Assumption Check**

Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribution is approximately standard.  
**Decision Rule**We will reject the null hypothesis if the p-value is less than the pre-determined alpha 0.05.  
**Descriptive Statistics:**

**Decision**Because the p-value of 0.995 is greater than 0.05, we fail to reject H0.

**Conclusion:**

We cannot conclude that the mean number of the average flight price in at least one population is different due to the differences in flight prices among airlines.

2. Are flight prices significantly different based on the day of the week and Class Type?

This question aims to explore how pricing strategies might vary depending on the class type and weekly travel patterns if there are significant differences in flight prices on certain days of the week when comparing different classes of airline tickets.

**Hypothesis:**

H0: There is no significant difference in flight prices based on the day of the week for Class Type.

H1: There is a significant difference in flight prices on certain days of the week and Class Type.

A screen shot of a computer

Description automatically generated

Reading the Result:  
**Assumption Check**Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribut is approximately standard.

**Decision Rule**We will reject the null hypothesis if the p-value is less than the pre-determined alpha 0.05.  
**Descriptive Statistics::**

* **ClassType:**
* F-statistic: 1.35
* P-value : 0.258 indicates that ClassType does not significantly affect the dependent variable since p > 0.05.
* **DayOfWeek:**
* F-statistic: 1.29
* P-value: 0.259, indicates that DayOfWeek does not significantly affect the dependent variable since p > 0.05.
* **ClassType:DayOfWeek (Interaction):**
* F-statistic: 0.514
* P-value: 0.908, very high, strongly indicating that the interaction between ClassType and DayOfWeek does not significantly affect the dependent variable since p > 0.05.

**Decision**

p-values are above the typical threshold of 0.05, we fail to reject H0.  
**Conclusion:**

At the 5% significance level, we fail to reject H0 and conclude that there is no interaction between the ClassType and DayOfWeek or that No interaction significantly contributes to the change in flight prices based on their p-value.

**Regression Testing**

Steps taken to optimize the data for Regression Testing

1. Converted all character columns to factors.

A screen shot of a computer code

Description automatically generated

1. Created dummy variables for all factor columns.  
   A black background with green text

   Description automatically generated
2. Created a new data frame with the dummy variables.  
   A black screen with text

   Description automatically generated

3. Does flight duration impact flight prices?

Method Regression Testing: Utilize regression analysis to explore the relationship between flight duration (independent variable) and flight prices (dependent variable).

**Hypotheses:**

H0: Flight duration does not significantly predict flight prices.

H1: Flight duration significantly predicts flight prices.  
A computer screen with white text

Description automatically generated

Reading the Result:  
**Assumption Check**

Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribut is approximately standard.

**Decision Rule**

We will reject the null hypothesis if the p-value is less than the pre-determined alpha 0.05.  
Descriptive Statistics:

Decision

In the regression summary, examine the p-values for the flight duration and airline variables (including their interaction term, if applicable). Significant p-values (typically <0.05) suggest that these factors significantly predict flight prices.

4. Do holiday seasons lead to a significant increase in flight prices compared to non-holiday?

Multiple Regression Testing: The effect of holiday seasons on flight prices. By comparing prices during holiday seasons against non-holiday periods for various destinations, we can understand the influence of demand surges on pricing.

Hypothesis:

H0: Flight prices do not significantly vary during holiday seasons across different destinations.

H1: Flight prices significantly vary during holiday seasons across different destinations.

A screenshot of a computer

Description automatically generated

Reading the Result:  
**Assumption Check**

Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribut is approximately standard.

**Decision Rule**

We will reject the null hypothesis if the p-value is less than the pre-determined alpha 0.05.  
Descriptive Statistics:

Decision

Look at the p-value for the interaction effect between holiday seasons and destinations. A significant p-value suggests that holiday seasons impact flight prices, and this impact varies by destination.

5. What factors best predict the price of a flight?

Multiple Regression Testing: Conduct various regression analysis using flight price as the dependent variable and several independent variables such as flight duration, distance, time of day, and day of the week. This comprehensive approach can identify which factors have the most significant impact on pricing, offering insights into pricing strategies and customer preferences.

Hypothesis:

Null Hypothesis (H0): The considered factors (flight duration, distance, time of day, day of the week) do not significantly predict flight prices.

Alternative Hypothesis (H1): The considered factors significantly predict flight prices.

Reading the Result:  
**Assumption Check**

Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribut is approximately standard.

**Decision Rule**

We will reject the null hypothesis if the p-value is less than the pre-determined alpha 0.05.

Descriptive Statistics:

Decision

Review the overall model significance (F-statistic and p-value) and the p-values for individual predictors. Significant results indicate that these factors collectively (and possibly individually) predict flight prices.

**Reference**

Juhar Ahmed Abdella, et al. "Airline ticket price and demand prediction: A survey." *Journal of King Saud University - Computer and Information Sciences* 33.4 (2021): 375-91. Web.

Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Version 1.7.1. URL: https://github.com/jacobkap/ fastDummies.