University of North Texas

ADTA 5130 Data Analytics

Flight Price Prediction

By: Biniam Abebe

January 2024

Content

I. 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 price of a flight depends on several parameters, such as its duration, destination, source, and arrival time.

II. 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.

III. 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, 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.

IV. Exploratory Data Analysis (EDA)

1. Research Questions

To further analyze the Spring 2024 Airline Dataset, a series of questions based on provided statements have been formulated to understand the dynamics of flight pricing.

a. Does airport traffic impact flight prices? This question examines the relationship between the Traffic level at the origin airport and the corresponding flight prices, i.e., whether higher traffic levels are associated with higher or lower flight prices due to operational efficiencies.

b. Do flight prices vary by time of day? Investigate whether there are patterns in flight pricing based on the departure time of day.

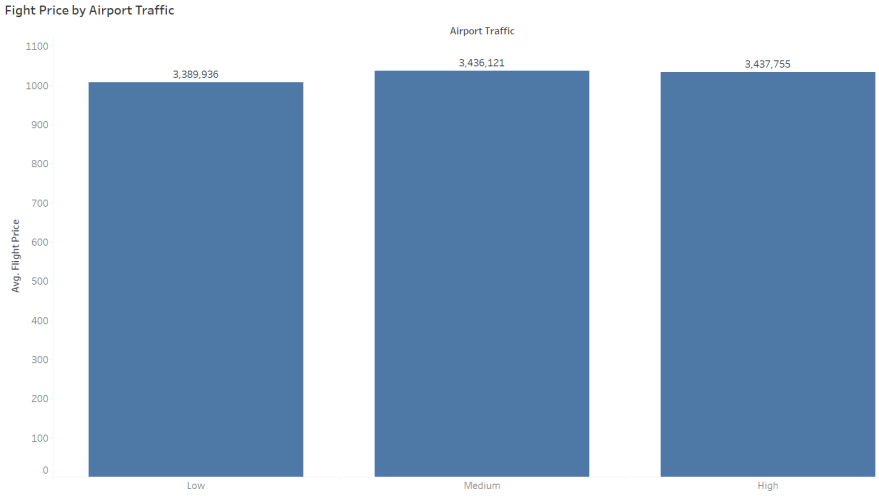
c. Do flight prices change by day of the week? Check flight prices throughout the week to see if certain days are cheaper or more expensive.

d. 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.

e. 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.

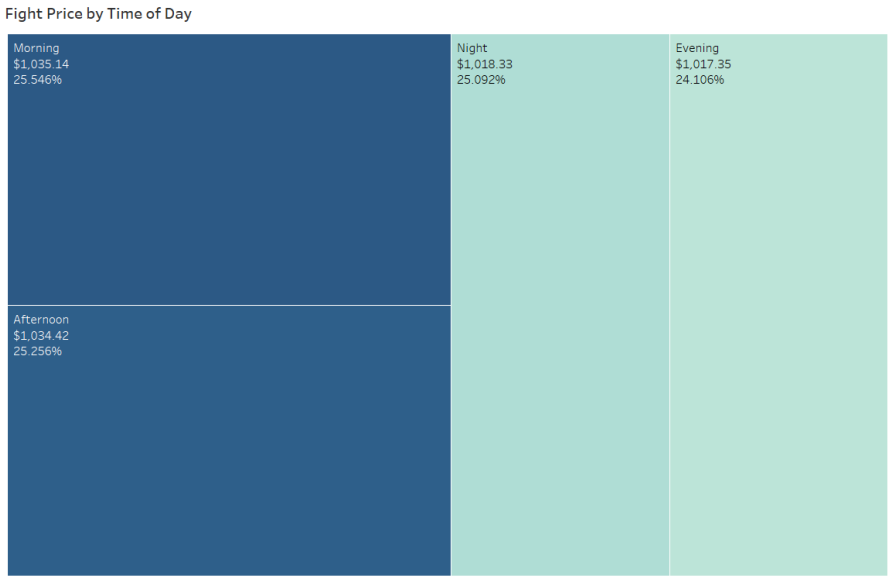
2. Analysis and Observation

a) Flight price by airport traffic



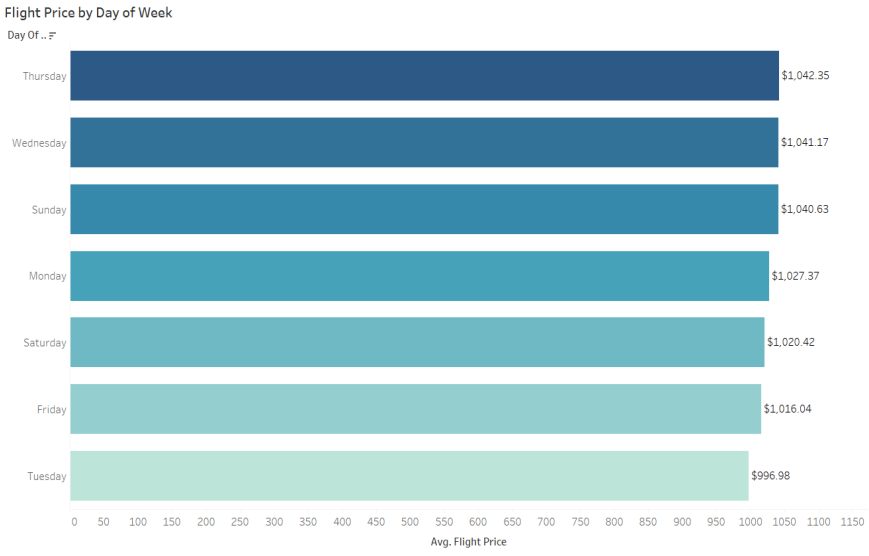
Key Observations: The bar graph shows a slight price difference between the airport traffic volumes (low, medium, and high). The lower the airport traffic, the lower the price, and vice versa. The higher the airport traffic, the more elevated the flight price. So, from these numbers, we can say that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

b) How do flight prices vary by time of day?



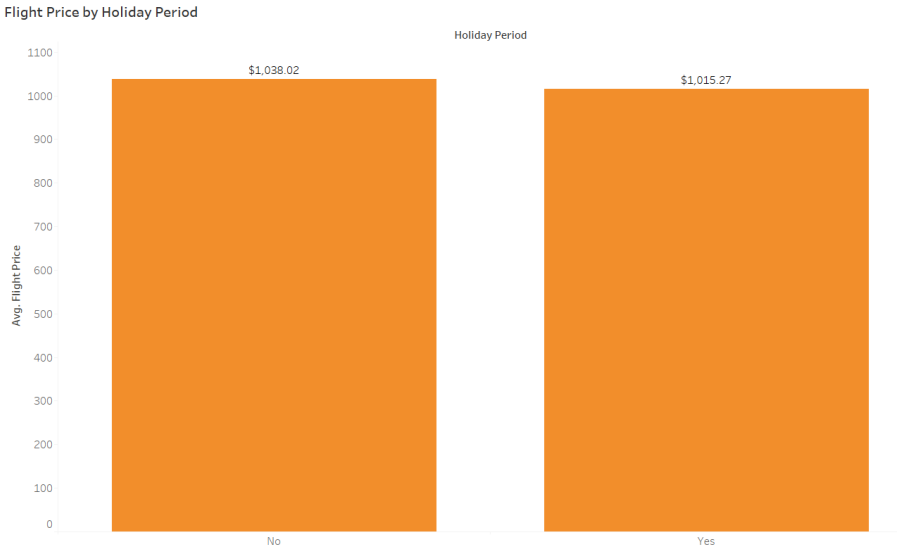
Key Observations: There is a price difference between the time of the flight and the day of the flight. Flight prices are higher in the early morning and early afternoon than in the night and evening.

c) How do flight prices change by day of the week?



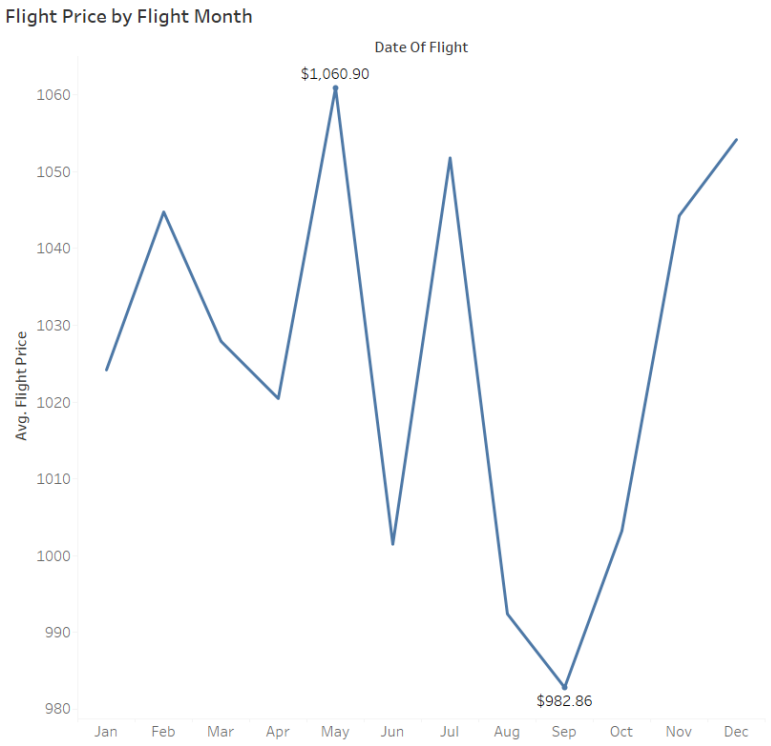
Key Observations: Flight prices will likely be higher in the middle of the week, followed by the weekend and early weekdays. Tuesday and Friday are the two days when flights are at their lowest. Therefore, the day of the week also influences flight prices.

d) What is the impact of the holiday season on flight prices?



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.

e) How does the month of the flight affect its price?



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.

V. 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

H_0: The mean flight prices do not differ significantly across airlines.

H_1: At least one airline has a significantly different mean of flight price compared to others.

```
> aov_result <- aov(FlightPrice ~ Airline, data=airline_data)
> summary(aov_result)
             Df    Sum Sq Mean Sq F value Pr(>F)
Airline       9 5.415e+05   60167   0.192  0.995
Residuals  9990 3.134e+09  313690
>
```

Reading the Result:

Assumption Check

Because n=10000 > 30, we can employ the Central Limit Theorem, which states that the distribution $\overline{x}$ is approximately normal.

Decision Rule

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

Descriptive Statistics:

$\mathrm{F-statistic:} = 0.192$

$p - value = 0.995$

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:

H_0: There is no significant difference in flight prices based on the day of the week for Class Type.

H_1: There is a significant difference in flight prices on certain days of the week and Class Type.

```
> aov_result2 <- aov(FlightPrice ~ ClassType * DayOfWeek, data = airline_data)
> print(tidy(aov_result2))
# A tibble: 4 × 6
  term                  df       sumsq  meansq statistic p.value
  <chr>              <dbl>       <dbl>   <dbl>     <dbl>   <dbl>
1 ClassType              2     848820. 424410.      1.35    0.258
2 DayOfWeek              6    2421594. 403599.      1.29    0.259
3 ClassType:DayOfWeek   12    1932872. 161073.      0.514   0.908
4 Residuals           9979 3129099280. 313568.     NA      NA
```

Reading the Result:

Assumption Check

Because n=10000 > 30, we can employ the Central Limit Theorem that says the distribution $\bar{x}$ is approximately normal**.**

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 Class Type and Day Of Week 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 Class Type and Day Of Week 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

I. Converted all character columns to factors.

```
airline_data[, sapply(airline_data
    # Select all character columns
    , is.character)] <- lapply(airline_data[
    # Select all character columns
    , sapply(airline_data, is.character)]
    # Convert to factors
    , as.factor)
```

II. Created dummy variables for all factor columns.

```
# Create dummy variables for all factor columns
dummy_vars <- model.matrix(~ . - 1, airline_data)
```

III. Created a new data frame with the dummy variables.

```
# Create a new data frame with the dummy variables
dummy_vars <- as.data.frame(dummy_vars)
```

1. Does flight duration impact flight prices?

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.

```
> lm_result <- lm(FlightPrice ~ FlightDuration, data = airline_data)
> summary(lm_result)

Call:
lm(formula = FlightPrice ~ FlightDuration, data = airline_data)

Residuals:
    Min      1Q  Median      3Q     Max
-981.95 -478.44    4.66  473.00  983.56

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1014.510     12.433  81.598   <2e-16 ***
FlightDuration    1.480      1.384   1.069    0.285
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 559.9 on 9998 degrees of freedom
Multiple R-squared:  0.0001144, Adjusted R-squared:  1.436e-05
F-statistic: 1.144 on 1 and 9998 DF,  p-value: 0.2849
```

Reading the Result:

Looking at the Residuals, the range of residuals from -981.95 to 983.56 suggests a wide variance in the flight prices that the model does not capture.

The coefficients table shows the estimated slope coefficient of b_0 (Intercept) = 1014.510, which suggests a positive relation, and b_1 = 1.480. Thus, the sample regression equation in $\widehat{FlightPrice} = 1014.51 + 1.480 FlightDuration$. Even though an analysis for Flight Price vs. Flight Duration reveals that the intercept coefficient is significantly positive (Estimate = 1014.510, $p < 0.001$). However, Flight Duration's effect on price is not statistically significant (Estimate = 1.480, $p = 0.285$). This indicates that flight prices do not significantly change as flight duration increases. The model explains a negligible portion of the variance in flight prices (R-squared = 0.0001144), suggesting other factors may better predict flight prices.

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

The impact of holidays on flight prices. By comparing price spikes during holiday seasons against non-holiday, we can understand the influence of demand surges on pricing.

Hypothesis:

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

H1: Flight prices significantly vary during holiday seasons.

```
> lm_result3 <- lm(FlightPrice ~ HolidayPeriodYes , data = dummy_vars)
> summary(lm_result3)

Call:
lm(formula = FlightPrice ~ HolidayPeriodYes, data = dummy_vars)

Residuals:
    Min      1Q  Median      3Q     Max
-987.05 -479.59    3.69  473.02  984.54

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1038.022      8.011 129.576   <2e-16 ***
HolidayPeriodYes  -22.749     11.199  -2.031   0.0422 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 559.8 on 9998 degrees of freedom
Multiple R-squared:  0.0004126, Adjusted R-squared:  0.0003126
F-statistic: 4.126 on 1 and 9998 DF,  p-value: 0.04224
```

Reading the Result:

Observing the Residuals, the range of residuals from -987.05 to 984.54 suggests a wide variance in the flight prices the model does not capture.

The coefficients table shows the estimated slope coefficient of b_0 (Intercept) = 1038.022, which suggests a positive relation, and b1 = 1.470. Thus, the regression equation in $\widehat{FlightPrice} = 1038.022 - 22.749 HolidayPeriod$. Even though an analysis for FlightPrice vs. Holiday Period reveals that the intercept coefficient is significantly positive (Estimate = 1038.022, $p < 0.05$). However, the effect of the holiday period on price is not statistically significant (Estimate = -

22.749, p = 0.0422). This indicates that flight prices do significantly change as the Holiday Period decreases. The model explains a negligible portion of the variance in flight prices (R-squared = 0.0004126), suggesting other factors may better predict flight prices.

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

We will conduct various regression analyses using flight price as the dependent variable and several independent variables. Here, we try to identify which factors have the most significant impact on pricing, offering insights into pricing strategies and customer preferences.

Hypothesis:

H_0: The considered factors (flight duration, distance, time of day, day of the week) do not significantly predict flight prices.

H_1: The considered factors significantly predict flight prices.

Methodology

1. Going to fit a model with all factors with flight price as a target variable and check the model.

```
# Run the regression
model <- lm(FlightPrice ~ ., data = dummy_vars)

# Summary of the model
summary(model)
```

```
efficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
Intercept)                8.015e+02  2.642e+02   3.033  0.00243 **
lightID                  -8.056e-04  1.946e-03  -0.414  0.67885
AirlineAir France`       -3.897e+00  2.454e+01  -0.159  0.87381
AirlineAmerican Airlines` -1.640e+01  2.461e+01  -0.666  0.50527
AirlineBritish Airways`  -4.312e+00  2.465e+01  -0.175  0.86113
AirlineDelta Air Lines`  -1.234e+01  2.496e+01  -0.494  0.62122
irlineEmirates           -2.559e+01  2.460e+01  -1.040  0.29838
irlineLufthansa          -9.083e+00  2.506e+01  -0.362  0.71705
AirlineQatar Airways`    -2.829e+00  2.502e+01  -0.113  0.91000
AirlineSingapore Airlines` -2.900e+00  2.482e+01  -0.117  0.90701
AirlineSouthwest Airlines` -8.289e+00  2.492e+01  -0.333  0.73945
AirlineUnited Airlines`          NA         NA      NA       NA
riginAirportCDG           4.792e+01  2.493e+01   1.922  0.05458 .
riginAirportDFW           3.781e+01  2.477e+01   1.526  0.12695
riginAirportDXB           5.956e+01  2.444e+01   2.437  0.01482 *
riginAirportHND           1.854e+01  2.483e+01   0.746  0.45540
riginAirportJFK           3.794e+01  2.459e+01   1.543  0.12295
```

2. Identify significant Predicators with a threshold of 0.05 significance.

```
# Identify significant predictors
significant_predictors <- summary(model)$coefficients[which(summary(model)$coefficients[, 4] < 0.05), ]

print(significant_predictors)  # Print the significant predictors
```

```
PROBLEMS 64  OUTPUT  TERMINAL  PORTS  DEBUG CONSOLE

> print(significant_predictors)  # Print the significant predictors
                        Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)           801.4660965 264.2417058  3.033079 0.002426964
OriginAirportDXB       59.5620295  24.4396756  2.437104 0.014822843
OriginAirportLAX       54.6893908  24.8038958  2.204871 0.027485965
DestinationAirportDFW -78.7662269  25.2243055 -3.122632 0.001797562
DestinationAirportORD -54.2442480  25.2461032 -2.148619 0.031688724
DestinationAirportSIN -70.5675354  25.4336420 -2.774575 0.005537716
FuelSurcharge          -0.4777976   0.2177923 -2.193822 0.028271267
AirportTrafficLow     -27.3303642  13.7491743 -1.987782 0.046863170
>
```

The result shows Hub airports like DXB, LAX, DFW, etc.; however, these cannot be used as overall price predictors unless all passengers originate from or have their destination in these locations. So, we will be refit only using Fuel Surcharge and Airport Traffic.

3. Run only with the two predictors.

```
model2 <- lm(formula = FlightPrice ~ FuelSurcharge + AirportTrafficLow, data = dummy_vars)
summary(model2)
```

```
> model2 <- lm(formula = FlightPrice ~ FuelSurcharge + AirportTrafficLow, data$
> summary(model2)

Call:
lm(formula = FlightPrice ~ FuelSurcharge + AirportTrafficLow,
    data = dummy_vars)

Residuals:
    Min      1Q  Median      3Q     Max
-999.86 -478.25    4.07  473.38 1006.50

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1061.6032    13.6871  77.562   <2e-16 ***
FuelSurcharge       -0.4752     0.2171  -2.189   0.0286 *
AirportTrafficLow  -27.7283    11.8458  -2.341   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 559.6 on 9997 degrees of freedom
Multiple R-squared:  0.001024,  Adjusted R-squared:  0.0008246
F-statistic: 5.126 on 2 and 9997 DF,  p-value: 0.005957
```

Final Model Readying the Result:

The regression analysis with Fuel Surcharge and Airport Traffic as predictors for Flight Price reveals a wide range in residuals, from -999.86 to 1006.50, indicating substantial price variability not explained by the model. The intercept (b_0 = 1061.6032, $p < 0.001$) suggests a high baseline price for flights. Fuel Surcharge negatively influences flight prices (b1 = -0.4752, p = 0.0286), indicating a slight decrease in price with higher fuel surcharges. Airport traffic has a more substantial negative effect (b2 = -27.7283, p = 0.0193), suggesting lower prices associated with low airport traffic. Thus, the regression equation in $\widehat{FlightPrice} = 1061.6032 - 0.4752FuelSurcharge - 27.7283AirportTraffic$. Despite these significant relationships, the models remain low due to (R-squared = 0.001024), inferring at other significant factors influencing flight prices beyond fuel surcharges and airport traffic levels.

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.