STAT 306 - Finding Relationships in Data Group Project

Flight Price Prediction

Project Group # C3
Members: Sadia Khan Durani, Richard Shen, Jiacan Deng, Naman Uttamchandani

Introduction:

The question of getting the best-priced flights for every person looking to travel for vacation, work, or any other reason is an issue everyone experiences. The determination of flight prices is influenced by a complex interaction of several factors. This study delves into a dataset from "Ease My Trip," a prominent online platform for booking flight tickets, to explore the relationship between various flight-related variables and their impact on ticket pricing. The original source utilized the Octoparse scraping tool to collect data from the "Ease My Trip" website over 50 days, from February 11th to March 31st, 2022. This dataset contains 300,153 distinct flight booking option tickets across India's top 6 metro cities, providing a comprehensive view of the current flight booking scene in India.

The primary objective of this analysis is to break down these complexities into actionable insights, specifically focusing on the prediction of flight prices. We aim to identify the variables that most significantly influence flight prices, thereby providing a predictive model that can offer passengers a deeper understanding of pricing strategies.

Our investigation will look into the following research question:

- *Among the variables, which will most accurately predict the price of an economy-class ticket in India?*

In conducting our analysis, multiple linear regression models are initially fitted. Subsequently, Backward Selection, a model selection method, is employed to identify and retain the most significant variables. Validation of the selected model is then achieved through the application of the Train/Test Split method, which serves to confirm our results.

Our dataset consists of the following features:

- **Airline:** categorical variable including 6 different airlines
    - SpiceJet, AirAsia, Vistara, GO_FIRST, Indigo, Air_India
- **Source City:** a categorical variable including 6 unique cities
    - Delhi, Mumbai, Bangalore, Kolkata, Hyderabad, Chennai
- **Departure Time:** a categorical variable using 6 time labels
    - Evening, Early_Morning, Morning, Afternoon, Night, Late_Night
- **Stops:** a categorical variable with 3 distinct labels
    - zero, one, two_or_more

- **Arrival time**: a categorical variable using 6 time labels
  - Night, Morning, Early_Morning, Afternoon, Evening, Late_Night
- **Destination City**: a categorical variable including 6 unique cities
  - Mumbai, Bangalore, Kolkata, Hyderabad, Chennai, Delhi
- **Ticket Class:** a categorical variable including 2 unique classes
  - Business, Economy
- **Duration:** a continuous variable which is the time taken to travel in hours
- **Days Left:** a continuous variable which is the number of days between 1 to 49 remaining for flight

Response Variable:

- **Ticket Price:** a continuous variable that doesn't have units listed, but we make the assumption it's in rupees, based on the location and cost of tickets

Analysis:

EDA/Visualizations:

To begin, we perform some exploratory data analysis to learn more about the distribution of our response variable, and how it depends on various other explanatory variables. Considering the information provided and the typical behavior of flight pricing, we plot the following factors: Stops, Class, and Duration.

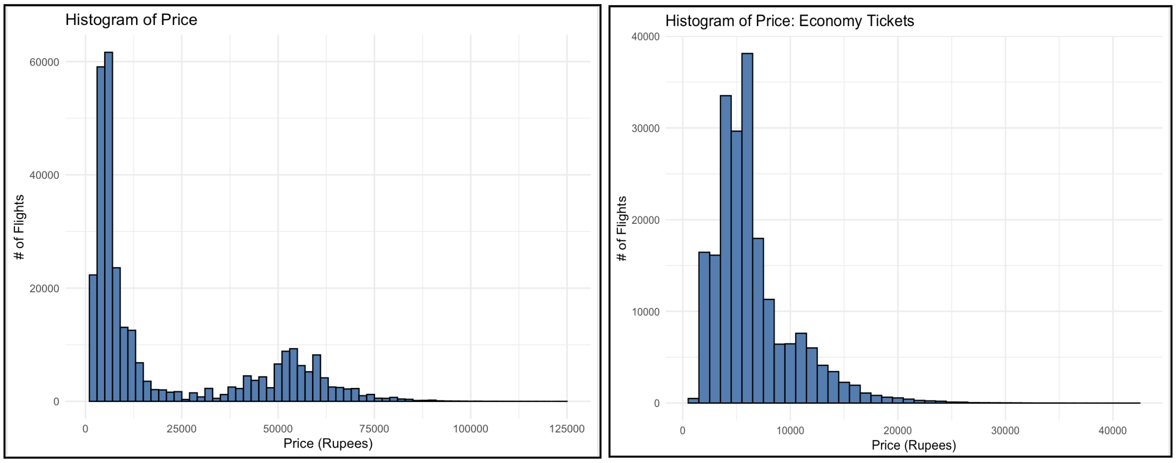


Figure 1: Histogram of Price for Both Classes vs Economy

In the first histogram of Figure 1, it is clear that there are more low-cost flights than high-cost flights. The histogram includes a second peak at the price range of 50,000 to 70,000 rupees representing the Business class tickets. The distribution suggests a market that favors affordability where consumers are more inclined to purchase lower-priced options when considering domestic travel.

In the second histogram of Figure 1, the distribution of economy-class tickets is plotted. This right-skewed distribution indicates that a log transformation may be useful to make the data more symmetric.

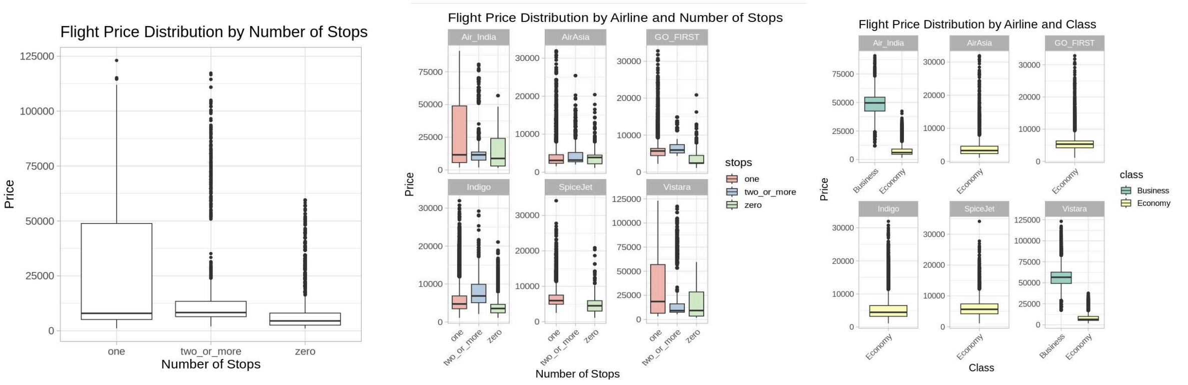


Figure 2: Price depending on various other explanatory variables

In the first chart in Figure 2, we plot price across the different levels of the variable STOP as a side-by-side boxplot. From this, we see that tickets with a one-stop have a higher price range whereas the ones with zero and two or more stops have a smaller range. However, the median price across all 3 is similar.

In the second chart in Figure 2, the boxplot illustrates how flight prices vary by airline and number of stops, with generally higher medians and wider ranges for flights with more stops. Vistara's non-stop flights notably have higher median prices, while AirAsia maintains a lower and tighter price distribution across all categories.

In the third chart in Figure 2, the boxplots represent flight price distribution by airline and class. We choose the side-by-side boxplots that create a separate plot for each airline with boxplots for price based on the business and economy class. There are a total of 6 airlines which fly domestically. The airlines, Air India and Vistara both offer economy and business class flights, whereas other airlines only offer economy class flights. The flights of Air India and Vistara may be more expensive than the other carriers due to this observation. However, given our research motivation, we focus only on the economy tickets.

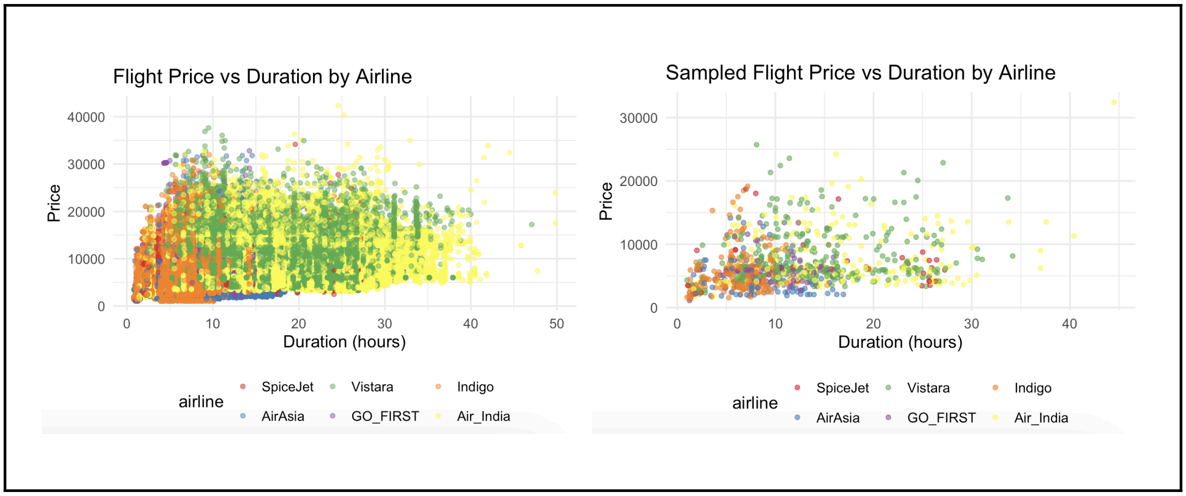


Figure 3: Price vs Duration (Economy Class)

For the third selection as shown in Figure 3, a scatter plot seems reasonable with duration on the x-axis and price on the y-axis to depict the relationship between duration, airline, and price, while differentiating the airlines with color. However, the plot appears quite dense, with many overlapping points which makes it difficult to discern individual data points for each airline, especially in areas of high data concentration.

To improve the plot, resampling can be a good strategy when dealing with dense scatter plots. The second chart in Figure 3 is a random sample of approximately 0.5% of the data. There appears to be a somewhat linear trend with a significant spread of the price of a flight depending on duration. There is a larger concentration of points in the (0, 15) hour interval. That could be because there are much less flights that are greater than 15 hours. This provides intuition on how our response variable depends on the flight duration.

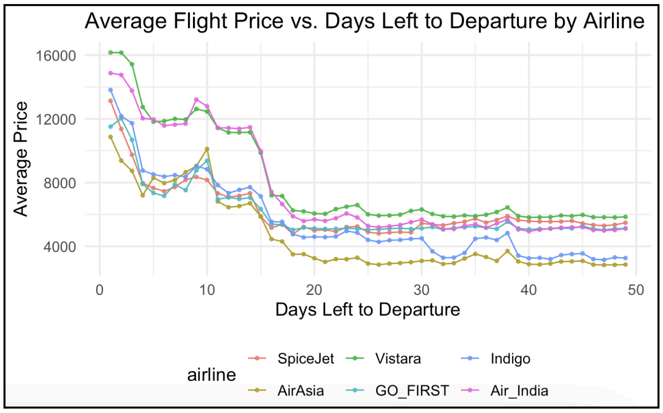
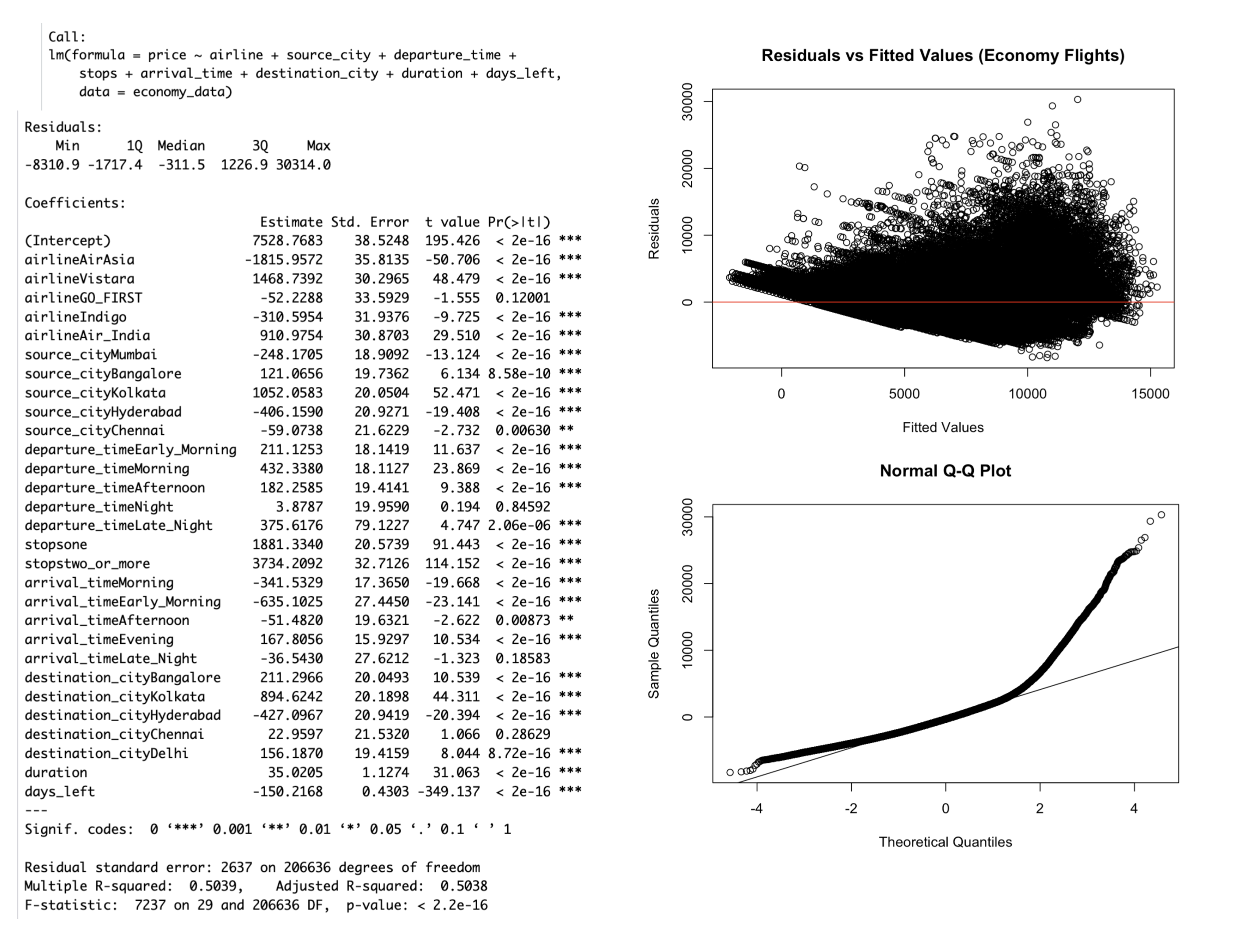


Figure 4: Focus on "Days Left" Variable

Finally, the variable "days left", in Figure 4, which indicates the remaining number of days until a flight's departure, is plotted with the average price. The line graph displays the trend of average flight prices for various airlines as the departure date approaches. As the days left to departure decrease, most airlines show an increase in average ticket prices, with SpiceJet and Indigo maintaining the most consistent pricing throughout. This is in line with the common knowledge that booking a flight in advance tends to be cheaper. Vistara stands out with generally higher prices that don’t fluctuate significantly as the departure date comes up.

MODEL SELECTION:

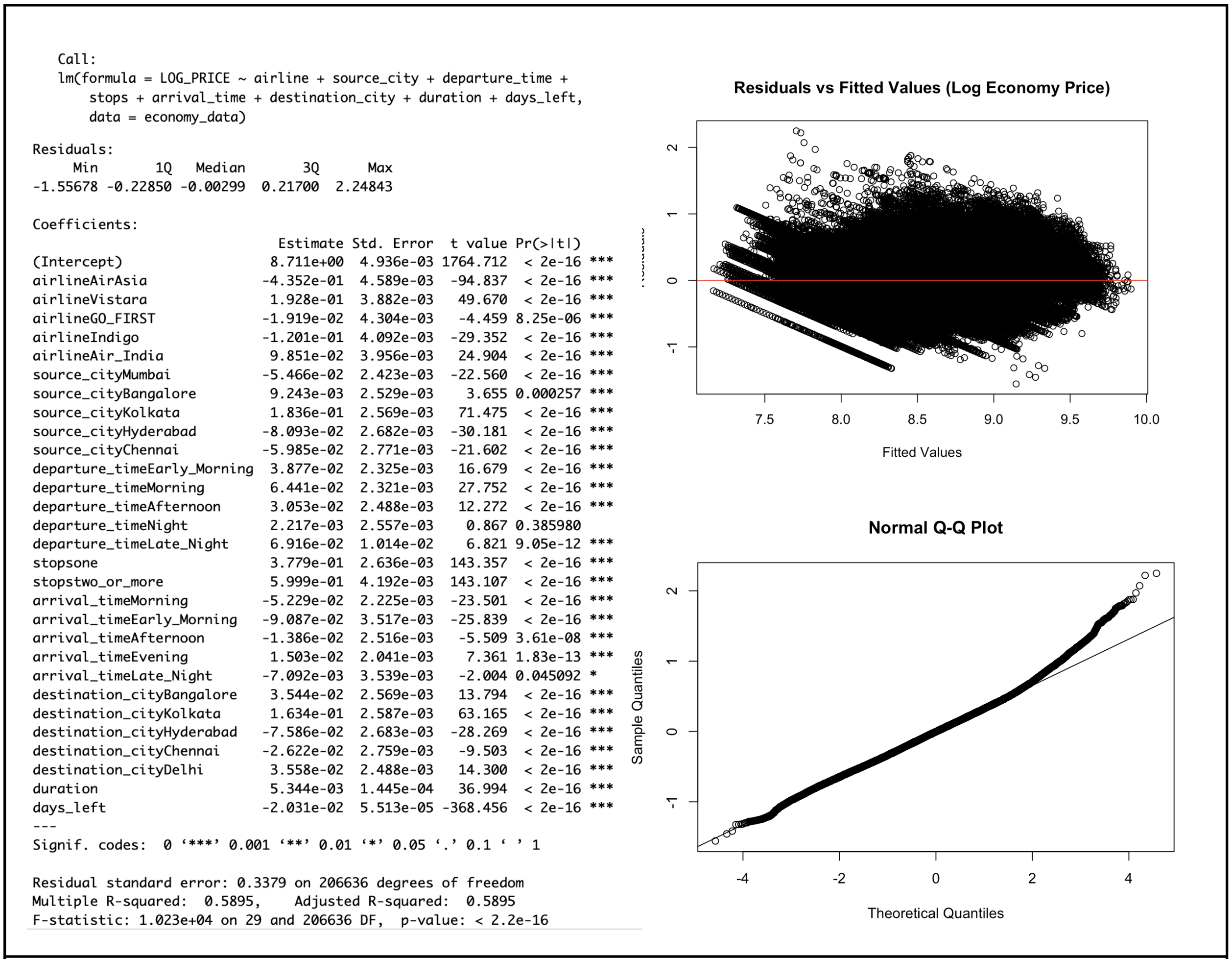
We begin our model fitting process by fitting a linear regression model with price as the response variable and all the remaining variables as the predictor variables.



```
Call:
lm(formula = price ~ airline + source_city + departure_time + 
    stops + arrival_time + destination_city + duration + days_left, 
    data = economy_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-8310.9 -1717.4  -311.5  1226.9 30314.0 

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                  7528.7683    38.5248  195.426  < 2e-16 ***
airlineAirAsia              -1815.9572    35.8135  -50.706  < 2e-16 ***
airlineVistara               1468.7392    30.2965   48.479  < 2e-16 ***
airlineGO_FIRST               -52.2288    33.5929   -1.555  0.12001    
airlineIndigo                -310.5954    31.9376   -9.725  < 2e-16 ***
airlineAir_India              910.9754    30.8703   29.510  < 2e-16 ***
source_cityMumbai            -248.1705    18.9092  -13.124  < 2e-16 ***
source_cityBangalore          121.0656    19.7362    6.134 8.58e-10 ***
source_cityKolkata           1052.0583    20.0504   52.471  < 2e-16 ***
source_cityHyderabad         -406.1590    20.9271  -19.408  < 2e-16 ***
source_cityChennai            -59.0738    21.6229   -2.732  0.00630 ** 
departure_timeEarly_Morning   211.1253    18.1419   11.637  < 2e-16 ***
departure_timeMorning         432.3380    18.1127   23.869  < 2e-16 ***
departure_timeAfternoon       182.2585    19.4141    9.388  < 2e-16 ***
departure_timeNight             3.8787    19.9590    0.194  0.84592    
departure_timeLate_Night      375.6176    79.1227    4.747 2.06e-06 ***
stopsone                     1881.3340    20.5739   91.443  < 2e-16 ***
stopstwo_or_more             3734.2092    32.7126  114.152  < 2e-16 ***
arrival_timeMorning          -341.5329    17.3650  -19.668  < 2e-16 ***
arrival_timeEarly_Morning    -635.1025    27.4450  -23.141  < 2e-16 ***
arrival_timeAfternoon         -51.4820    19.6321   -2.622  0.00873 ** 
arrival_timeEvening           167.8056    15.9297   10.534  < 2e-16 ***
arrival_timeLate_Night        -36.5430    27.6212   -1.323  0.18583    
destination_cityBangalore     211.2966    20.0493   10.539  < 2e-16 ***
destination_cityKolkata       894.6242    20.1898   44.311  < 2e-16 ***
destination_cityHyderabad    -427.0967    20.9419  -20.394  < 2e-16 ***
destination_cityChennai        22.9597    21.5320    1.066  0.28629    
destination_cityDelhi         156.1870    19.4159    8.044 8.72e-16 ***
duration                       35.0205     1.1274   31.063  < 2e-16 ***
days_left                    -150.2168     0.4303 -349.137  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2637 on 206636 degrees of freedom
Multiple R-squared:  0.5039,	Adjusted R-squared:  0.5038 
F-statistic:  7237 on 29 and 206636 DF,  p-value: < 2.2e-16
```

Regression 1 With 8 Explanatory Variables

The model produced suggests that most of the parameters are significant due to their low p values. We get an adjusted R squared of 0.5038 which is not satisfactory. The Residuals vs Fitted Values plot shows a spread of residuals which is more concentrated to the bottom possibly indicating heteroscedasticity. The Normal QQ plot says that the residuals are right-skewed, violating the normal assumption. To resolve this and aim for better results, we perform a log transformation on the response variable and refit the model.



```
  Call:
  lm(formula = LOG_PRICE ~ airline + source_city + departure_time +
      stops + arrival_time + destination_city + duration + days_left,
      data = economy_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55678 -0.22850 -0.00299  0.21700  2.24843

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 8.711e+00  4.936e-03 1764.712  < 2e-16 ***
airlineAirAsia             -4.352e-01  4.589e-03  -94.837  < 2e-16 ***
airlineVistara              1.928e-01  3.882e-03   49.670  < 2e-16 ***
airlineGO_FIRST            -1.919e-02  4.304e-03   -4.459 8.25e-06 ***
airlineIndigo              -1.201e-01  4.092e-03  -29.352  < 2e-16 ***
airlineAir_India            9.851e-02  3.956e-03   24.904  < 2e-16 ***
source_cityMumbai          -5.466e-02  2.423e-03  -22.560  < 2e-16 ***
source_cityBangalore        9.243e-03  2.529e-03    3.655 0.000257 ***
source_cityKolkata          1.836e-01  2.569e-03   71.475  < 2e-16 ***
source_cityHyderabad       -8.093e-02  2.682e-03  -30.181  < 2e-16 ***
source_cityChennai         -5.985e-02  2.771e-03  -21.602  < 2e-16 ***
departure_timeEarly_Morning 3.877e-02  2.325e-03   16.679  < 2e-16 ***
departure_timeMorning       6.441e-02  2.321e-03   27.752  < 2e-16 ***
departure_timeAfternoon     3.053e-02  2.488e-03   12.272  < 2e-16 ***
departure_timeNight         2.217e-03  2.557e-03    0.867 0.385980
departure_timeLate_Night    6.916e-02  1.014e-02    6.821 9.05e-12 ***
stopsone                    3.779e-01  2.636e-03  143.357  < 2e-16 ***
stopstwo_or_more            5.999e-01  4.192e-03  143.107  < 2e-16 ***
arrival_timeMorning        -5.229e-02  2.225e-03  -23.501  < 2e-16 ***
arrival_timeEarly_Morning  -9.087e-02  3.517e-03  -25.839  < 2e-16 ***
arrival_timeAfternoon      -1.386e-02  2.516e-03   -5.509 3.61e-08 ***
arrival_timeEvening         1.503e-02  2.041e-03    7.361 1.83e-13 ***
arrival_timeLate_Night     -7.092e-03  3.539e-03   -2.004 0.045092 *
destination_cityBangalore   3.544e-02  2.569e-03   13.794  < 2e-16 ***
destination_cityKolkata     1.634e-01  2.587e-03   63.165  < 2e-16 ***
destination_cityHyderabad  -7.586e-02  2.683e-03  -28.269  < 2e-16 ***
destination_cityChennai    -2.622e-02  2.759e-03   -9.503  < 2e-16 ***
destination_cityDelhi       3.558e-02  2.488e-03   14.300  < 2e-16 ***
duration                    5.344e-03  1.445e-04   36.994  < 2e-16 ***
days_left                  -2.031e-02  5.513e-05 -368.456  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3379 on 206636 degrees of freedom
Multiple R-squared:  0.5895,    Adjusted R-squared:  0.5895
F-statistic: 1.023e+04 on 29 and 206636 DF,  p-value: < 2.2e-16
```

Regression 2: With Log Transformation

In Regression 2, we take advantage of a new model selection with log transformation, which is a regression model output detailing the relationship between a log-transformed flight price and various predictors like airlines, cities, departure times, and days left before departure. The summary reveals the magnitude and significance of each predictor's effect on flight prices, with many variables showing strong significance. Diagnostic plots to the right display the residual patterns and the normality of data distribution, essential for validating the assumptions of linear regression. The diagnostic plots show a significant improvement after the log transformation, the residual plot shows a more constant spread and the Normal QQ plot shows a great improvement, such that the normal assumption is reasonable.

Based on the p-values, all covariates are statistically significant. Likewise, testing different models and comparing them using the adjusted r-squared statistics, the model fitted with all covariates produces the greatest adjusted r-squared value yielding a value of 0.5895. The table below shows the trend of the adjusted R-Squared decreasing as the number of variables decreases, so the best model according to the adjusted R-Squared statistics is the model containing all explanatory variables.

Model with All Variables:	Adjusted R-Squared: 0.5894876
Model with 7 Variables:	Adjusted R-Squared: 0.5646244
Model with 6 Variables:	Adjusted R-Squared: 0.5535994

Our final statistic for selecting an appropriate model to fit the data is the AIC (Akaike Information Criterion). The step() function in R was used to run backward selection using the AIC criteria to choose the best model as dictated by the AIC. Similarly, the model with all the covariates ended up being the best.

Based on these 3 different statistics, it is reasonable to say that the model that fits the data best is the one with all explanatory variables.

Train/Test Method & RMSE

Finally, the data is split into a test set and a training set. The best model determined from the previous methods is trained on the train set. The RMSE of the model is computed to obtain an estimate of the prediction error of the best model chosen. The RMSE of the model is 0.33. This indicates our final selected model performs well on new data.

Conclusion:

To answer our research question, *Among the variables, which will most accurately predict the price of an economy-class ticket in India?*

This analysis found the variables that most accurately predict the log price of an economy-class ticket in India are all the provided explanatory variables. The log transformation we performed on our data helped solve issues that would've arrived if we continued to analyze our first regression model. With the log transformation, we solve the issues of right-skewed tail residual QQ plots and centering the residual vs fitted value plots. Our analysis tells us that the variation in the log price of an economy ticket is described with all variables: airline, source_city, departure_time, stops, arrival_time, destination_city, class, duration, and days_left.

This was conducted through fitting linear regression models and through multiple indicators such as each p-value being significant, the adjusted R-squared, as well as the comparison of the AIC values made during the backward selection process with the step function in R. While this may seem extreme, the conclusion that all variables are significant is not unreasonable due to the competitive nature of the airline industry. For airlines, the margins for profit are thin (simpleflying.com), and so the price will be reflected closely with each change in the parameters.

Our model may present some issues when determining the price of an airplane ticket regularly. Our data only considers a length of 50 days from February 11th to March 31st. Travel is considered popular and ideal in the months of February and March due to numerous festivals that occur over this period (indiasomeday.com). Prices will reflect this and will generally be higher for flights chosen in this timeframe. There may be some correlation between travel date and price that may not be accurately reflected in our study.

References:

Mitchell, Alexander. "Why Do Airlines Run on Such Wafer Thin Profit Margins?" *Simple Flying*, 10 Feb. 2024, simpleflying.com/airlines-thin-margins-analysis/#:~:text=Summary,contribute%20to%20these%20low%20margins.

Sonawala, Harsh. "Best Time to Visit India - Seasons, Festivals & Expert Tips." *India Someday Travels*, 3 Apr. 2024, indiasomeday.com/en/article/best-time-to-visit-india/.

DATASET:
https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/code?datasetId=1957837&searchQuery=R