# Analyzing the airfare trends in Indian Airlines

Table of Contents

## Dataset Introduction and Hypothesis

The dataset examines how airlines price their flight tickets depending on a myriad of factors such as the flight classes, duration, number of stops, etc. This report aims to figure out how these distinct explanatory variables contribute to the overall pricing of flight tickets.

We hypothesize that a few variables such as stops, class, and airlines significantly affect flight prices compared to all the other variables. For example, we expect to find substantial differences in pricing between a business class and an economy class seat or between a budget airliner and a state airliner compared to the differences in pricing between a morning departure and a night departure.

## Description of Dataset

| Variable Name | Unit of Measurement | Discrete or Continuous | Categorical/Numerical |
|---|---|---|---|
| Airline | Name of airline company | Discrete | Categorical |
| Source City | Departure city | Discrete | Categorical |
| Departure Time | Designated time intervals | Discrete | Categorical |
| Stops | Number of stops<br>(0 = zero, 1 = one or more) | Discrete | Binary |
| Arrival Time | Designated time intervals | Discrete | Categorical |
| Destination City | Arrival city | Discrete | Categorical |
| Class | Flight class | Discrete | Categorical |
| Duration | Total time between departure and arrival (hours) | Continuous | Numerical |
| Days Left | Number of days left from booking date to trip date | Discrete | Numerical |
| Price | Indian rupees (INR) | Continuous | Numerical |

The dataset contains flight information regarding flights in India from 11th Feb to March 31st in 2022 and consists of six airlines including the pricing for economy class tickets and business class tickets. It consists of all relevant information that goes into pricing such as departure time, arrival time, number of stops, ticket class (economy or business), total duration of the flight, number of days left before the flight, departure city, and arrival city. We trimmed the data down from 300,152 to 30,098 entries, focusing entirely on Delhi and Mumbai as they were the busiest airports in India for 2021 (Sun 2021). There is a tendency for non-significant differences within large sample sizes to be statistically significant even though the differences are not significant as a result trimming around 90% of the data reduces the possibility for small differences to contribute to a statistically significant result that cannot be explained.

From the table above, we see that there are 10 variables with price being the response variable and the other 9 variables as the explanatory variables. Due to the presence of many categorical explanatory variables and numerical responses we think that a Multiple Linear Regression model will be the best fit for the data.

## **Exploratory Data Analysis**

Before fitting a regression model we decided to produce a few graphs to understand how the data is structured and to identify which variables would play a crucial role in determining flight ticket pricing. Looking at airline pricing by itself, without separating economy and business class, was an unhelpful endeavor as Air India and Vistara airlines, being the only airlines in our dataset that offered business class, had significantly higher average prices than the other four airlines. Exhibit 1 shows how flight prices are distributed across different airlines.

The distribution of pricing for flight tickets for business and economy class is shown in Exhibit 2 which confirms that business class tickets are approximately 3-5 times higher in price than economy class tickets. Therefore, we thought it would be better to visualize the distribution of ticket prices across various airlines, specifically for economy class. We plotted a violin plot to analyze the distribution of pricing for economy class tickets across different airlines which can be seen in Exhibit 3. The plot shown in Exhibit 3 suggests that not all airlines offer the same prices despite offering the same class of fare. Therefore, there is a high possibility that airlines do not follow price-match.

We wanted to analyze how the number of days left before the flight influences the pricing of the flight ticket. Since business and economy class tickets have a huge difference in their pricing, we created subsets of the data - one for economy class and one for business class. The distribution of pricing based on the number of days left, specifically for business class tickets can be seen in Exhibit 4. However, for business class tickets a clear change in price for last-minute bookings cannot be seen. A similar visualization for economy class tickets is shown in Exhibit 5. For economy class tickets a clear rise in price for last minute booking can be seen.

Producing descriptive statistics of the data, we found that the frequency of occurrence for the category for the number of stops representing two or more stops represents only 0.46 percent of the data. Since this is a very small percentage we decided to re-code the data such that the number of stops is converted to a binary variable with 0 representing direct flights and 1 representing flights with one or more stops. A similar observation was made for departure time as the frequency of occurrence of late night departures was found to be 0.79 percent of the observations. Therefore, we combined late-night flights into the night category. In addition, early morning arrivals and late-night arrivals had a frequency of 2.93 percent and 5.03 percent respectively. Therefore, we integrated them into morning and night respectively.

Finally, we looked at the distribution of pricing for the number of stops stratified by class of ticket in Exhibit 6. It appears for both economy and business class tickets, direct flights are cheaper than flights with one or more stops. This runs contrary to our expectations as typically one would expect a direct flight to cost more while a flight with stops adds needless transit.

## Hypothesis

Based on these factors, we predict that some variation of airline class, airline flying the flight, number of stops in the flight, and number of days left before the flight when the ticket was bought would explain most of the variation in airline ticket prices.

## Regression Analysis

### Checking for multicollinearity

Due to the way our data was trimmed, multicollinearity between certain variables was certain to happen. One such example was the multicollinearity value between the source city and destination city, a correlation test for this generated a value of 0.99. However, given that there were only two cities involved in this study, this isn't surprising. If the source city was Mumbai, its destination city would be Delhi. Therefore due to their perfect collinearity, we decided to exclude the destination city from our model fit.

Other variables that we explored for multicollinearity are departure time and arrival time and due to the presence of layovers which contributed to increased durations of flights, the observed correlation value was low at 0.24. Another round of collinearity tests between the number of stops and the duration of the flight yielded a collinearity value of 0.60. With this in mind, source city and destination city's impact on airline flight pricing is limited. Other variables such as duration and number of stops will have a more significant representation of their effect on airline pricing.

### Variable Selection

We use stepwise forward selection and backward elimination to find the significant variables that affect how airlines price their ticket fares. We don't use LASSO due to the presence of categorical data. For categorical data, a linear regression model automatically creates binary variables for each category. For example, for the variable airline, six different binary variables are taken each representing an airline company for the model fit. Therefore, when we try to fit LASSO regression some binary variables are excluded from the chosen best model fit. For example, the binary variable for airline Air India might be included in the best chosen model, however, the binary variable for airlines SpiceJet and Indigo might be excluded from the best-chosen model. Since we are interested in discussing which factors are important for flight prices, the best model using LASSO does not work for us

The best model chosen using stepwise forward selection or backward elimination had all the explanatory variables; airlines, source city, departure time, number of stops, arrival city, duration, days left, and price.

Likelihood Ratio Test

Even though we had a good adjusted R-squared value for the forward selection, the AIC values for some of the models were similar therefore we took the last four models from stepwise forward selection and tested them against each other using likelihood ratio tests for model comparison.

The four models with similar AIC values that we tested are listed as follows:

- *Model 1* - Full model with all explanatory variables that are class, stops, days_left, airline, arrival_time, departure_time, duration, and source_city included. This is the best-chosen model according to variable selection using stepwise forward.
- *Model 2* - Model with all the variables included except source_city.
- *Model 3* - Model with all the variables included except source_city and duration
- *Model 4* - Model with all the variables included except source_city, duration and depature_time.

We excluded the intercept from all four models since the intercept was representing the coefficient for airline Air India. Hence, removing the intercept did not change the model fit. For the likelihood ratio test, we chose a simple random sample of 500 observations since we have a large dataset and so statistically non-significant differences might be considered significant due to the large sample size.

Based on the likelihood ratio test results, we found that model 3 is the best model for our analysis. It contains the airline, number of stops, numbers of days left, arrival time, and departure time as the variables while ignoring source city, destination city, and duration. This is to be expected as these variables were previously explained to have a limited scope of impact in our model due to how we trimmed the data.

Pairwise Multiple Comparison Test

We decided to do a pairwise multiple comparison using Tukey's Honestly Significant Difference for all the categorical variables to determine whether the mean pricings between pairs are significantly different from one another. Since pairwise multiple comparisons create a lot of results, an underline diagram is useful to display which pairs have similar effects. This is represented by the corresponding p-values being greater than 0.05. We produced multiple pairwise comparison tests for variables airline, departure time, and arrival time since they are the only categorical variables with more than two categories.

An underline diagram for airlines is presented below:


The airlines are written in increasing order of mean prices and airlines which do not differ significantly in terms of mean pricing are underlined. Therefore, results show that Air Asia, indigo, and SpiceJet do not differ significantly in terms of flight pricing. Similarly, SpiceJet and GoFirst do not differ with respect to flight prices. However, Air India and Vistara differ significantly in pricing when compared to any other airline. Since Air India and Vistara are the only airlines that offer business class tickets this outcome was expected.

An underline diagram for departure time is presented below:


The underline diagram suggests that afternoon, evening, and night departures do not have a significant difference in terms of pricing. Furthermore, afternoon and early morning departures do not have a significant difference as well as early morning and morning flights do not have a significant difference in terms of pricing. Results show that there isn't a specific departure time in the data that shows a significant difference in the price of tickets. This also matches our

hypothesis that departure time would not influence flight prices as compared to some of the other variables in the data.

Lastly, an underline diagram for arrival time is presented below:


The underline diagram for arrival time shows that afternoon and evening arrivals do not differ significantly in terms of flight pricing. However, morning and night arrivals have a significant difference in price as compared to any other arrival time.

Main Model Fit

Fitting model 3 with multiple linear regression generated an R squared value of 0.9438 and an adjusted R squared value of 0.9437. Much higher than the R squared value of 0.8912 we obtained using the forward selection method. All airlines were statistically significant with p values at 2e-16, significantly below the 0.05 threshold. The same goes for stops, numbers of days left, and class of the fare. Furthermore, we made residual diagnostic plots for our main model fit, and the plots suggest that our data is approximately normal with small deviations. The price differences between business and economy class cause the residual v/s fitted plot to have two scattered areas of prices. However, other than the class pricing, the data does not seem to have any significant pattern.

However, for categorical variables such as arrival time and departure time, certain factors were not significant. Morning arrival times and evening departure times were both statistically non-significant with p values at 0.5395 and 0.0845 respectively.

Following this, we decided to also fit a log model instead of a multiple linear regression. With it, all of the variables were statistically significant and best of all, our R squared value increased to 0.9987, and adjusted R squared value increased to 0.9987. With such a high R squared value, our model is pretty much an exact one-to-one fit with the data and explains everything well.

## Conclusion

### Limitations of the Model

Another variable that would be underrepresented, due to how the data was filtered, was duration. As there were only two cities, the duration of the flight isn't fully represented in our analysis. With longer and short flights across various cities, duration was expected to play a larger role in pricing.

Once again due to how our data was trimmed to have only two major cities in India, there's a maximum ceiling each flight can have even if the flight contained one or more stops.

### Future Models and Proper Implementation

Despite the high R squared value obtained from our dataset, it still isn't a proper representation of how airlines price their tickets. As we were limited to two cities in India and out of the six airlines selected, only two offered business class. Our data is representational of the airline industry within major cities in India but isn't representational of the overall worldwide airline industry.

However, it provides a good starting point to explore pricing in the airline industry further. As it did disprove our assumption of direct flights being cheaper than flights with one or more stops.

Future models could be improved with more data across different airlines, routes, and classes of fares. But we have to keep note that with larger data sets, small deviances, that can be explained by randomness, can easily be construed as statistically significant despite that not being the case. Stricter cleaning protocols would have to be implemented to ensure the integrity of our analysis doesn't contain any false positives.

Final Remarks

The dataset allowed us to predict the airline ticket pricing quite well. The models we created had extremely high R squared values at 0.9987 with adjusted R squared values at 0.9987, pretty much a one-to-one fit that explains all of the variation seen in the dataset.

In regards to our hypothesis, the airline, number of stops, number of days lift, and airline class all played a role in explaining airline ticket pricing. However, arrival time and departure time played a larger role than anticipated. Despite a collinearity value of 0.24, they were different enough to be considered separate variables and for the model not to be too severely impacted by their collinearity,

To improve our model further, more observations of economy and business class flights would help provide a more holistic view of the airline industry. As well as exploring other cities in India to truly gauge the impact of the source city and arrival city on airline ticket pricing.

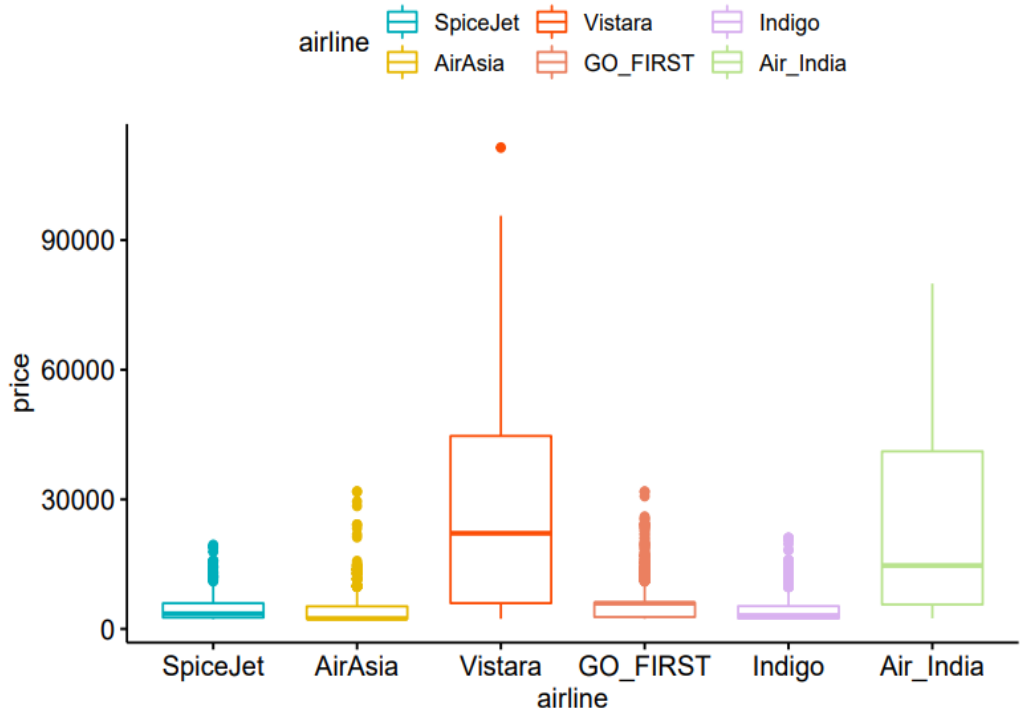
References

Bathwal, S. (2022, February 25). *Flight price prediction*. Kaggle. Retrieved March 03, 2022, from https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Sun, S. (2021, November 9). *India: Leading airports by number of passengers handled 2021*. Statista. Retrieved April 11, 2022, from https://www.statista.com/statistics/589115/indian-airports-passenger-traffic/

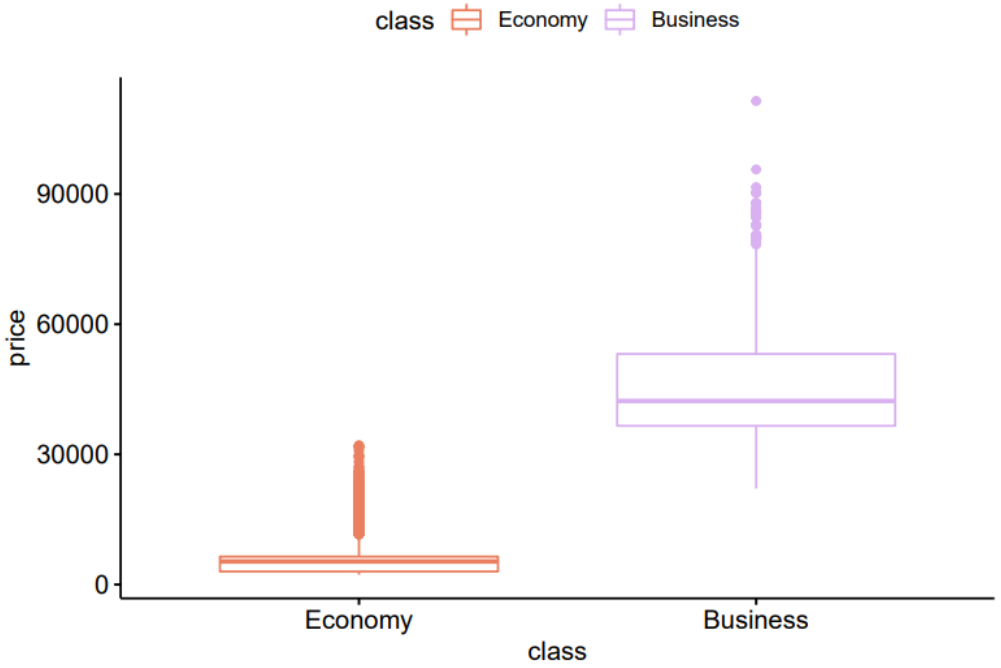
Exhibit

**Exhibit 1:** Distribution of flight ticket prices for different airline companies



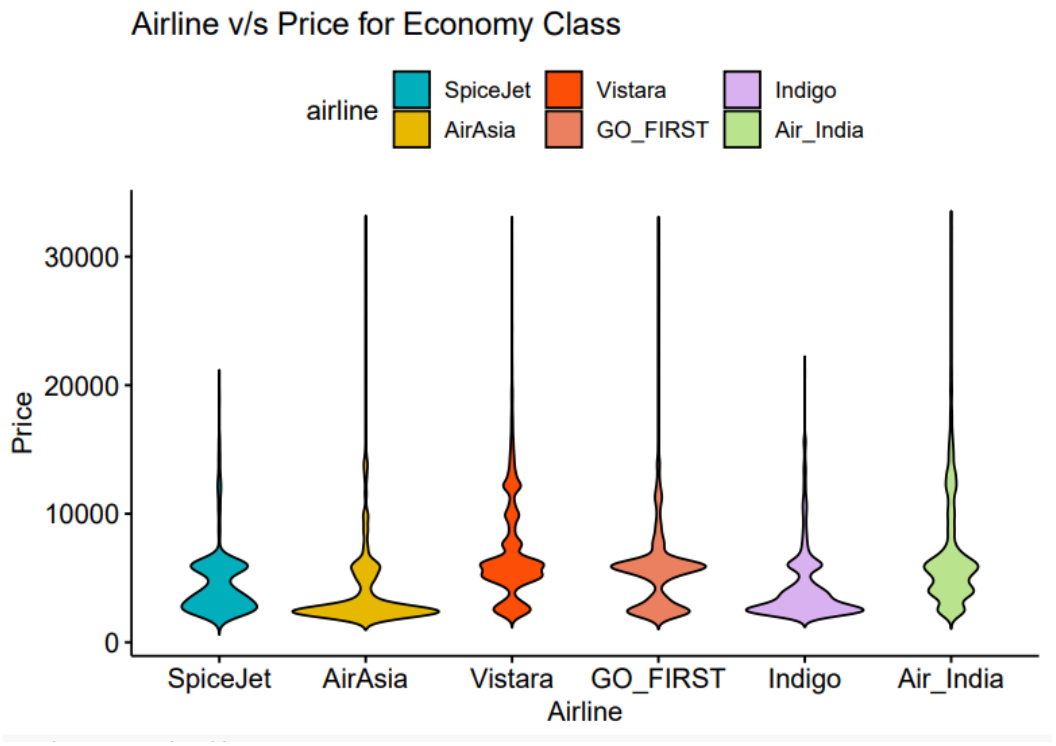

A box plot representation of the pricing of flight tickets for each airline company. This figure includes pricing for both economy and business classes.

**Exhibit 2:** Distribution of flight ticket price based on class



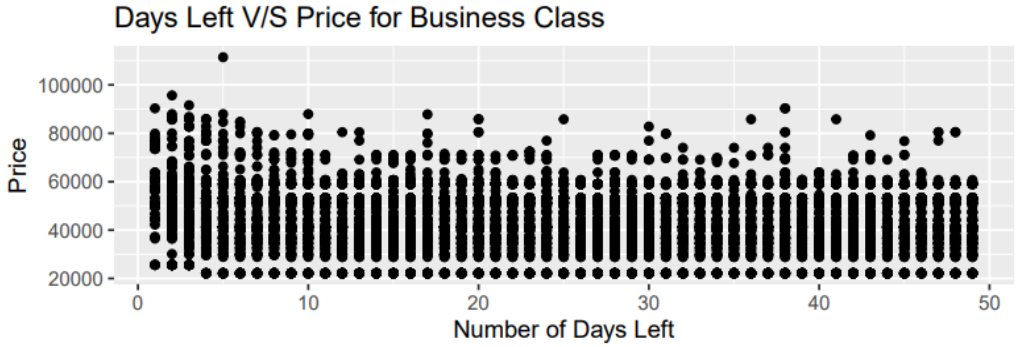

A box plot representation of the pricing differences for economy and business class tickets.

**Exhibit 3:** Price distribution for different airline companies specifically for economy class ticket



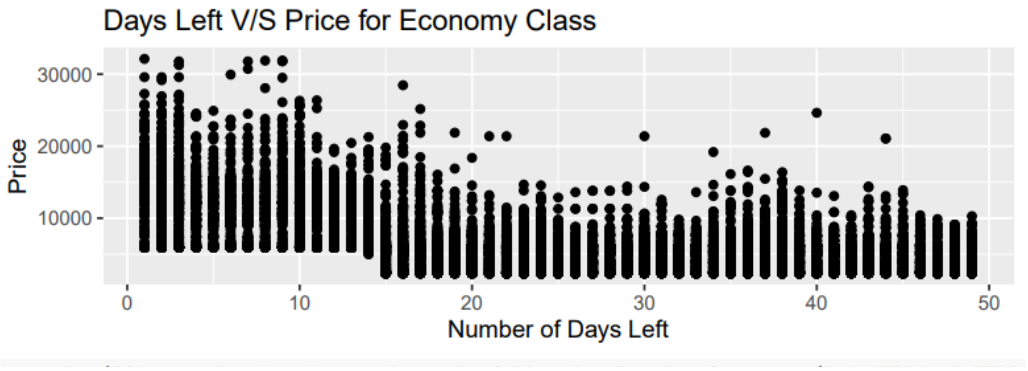

We made a violin plot to analyze the distribution of economy class ticket pricing across different airline companies.

**Exhibit 4:** Distribution of pricing based on the number of days left for business class



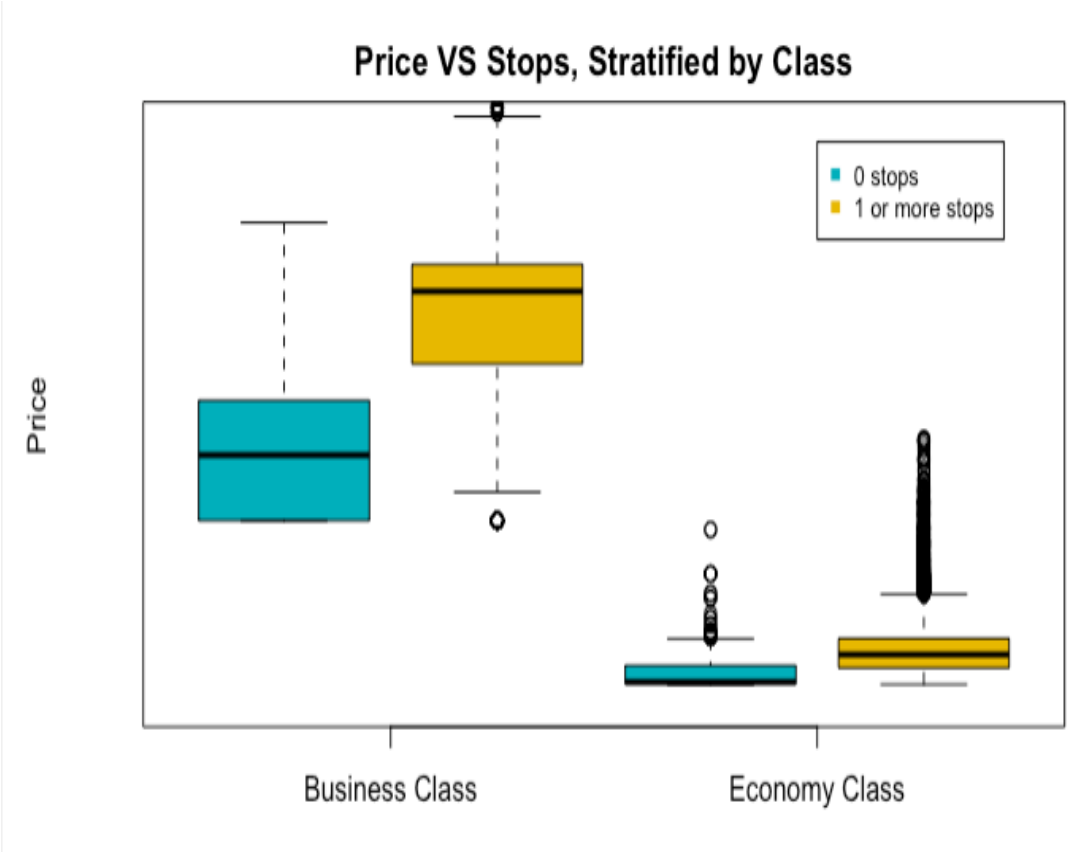

Scatterplot representation of the price of a business class ticket purchased depending on the number of days left until the flight.

**Exhibit 5:** Distribution of pricing based on the number of days left for economy class




Scatterplot representation of the price of an economy class ticket purchased depending on the number of days left until the flight.

**Exhibit 6:** Distribution of price of an airline ticket for the number of stops, stratified by class




Boxplot showing the distribution of price of an airline ticket separated by their class and number of stops in the flight.