Programming Lab 1 Group Project:

Flight Fares Price Variations

Alma Mater Studiorum Università di Bologna

2023/2024 Academic Year

Professor M. Novelli

Francesco Grandi 0001071804

Tommaso Maccaferri 0001071630

Alessio Serraglio 0001070669

Cosimo Zatti 001068010

Abstract

The following study has the objective of analyzing, evaluating, and discussing the fluctuation in the prices that flight fares adopt in response to different variables. The study has been conducted using a sample of 300,000 flights collected between February 11th 2022 and March 31st 2022, by the website Easemytrip.com, a well-known Asian private company specialized in the arrangement and selling of flight trips all over the eastern continent. The company's goal is to propose the most convenient flights to its customers and the dataset collected from its website is a reflection of the tickets purchased on that platform in the mentioned span of time, for the within-India airplane movements. The dataset takes into consideration nine variables for each flight: airline company, flight code, source city, departure time, stops, arrival time, destination city, class, and duration. The subsequent quantitative analysis, done through the statistical software R, has imposed as its objective the discovery of the variations in the price range according to the different variables. The following, and last, step of this project has been the qualitative evaluation of the results produced, starting from the correlation obtained and aiming at figuring out the potential reasons behind them, questioning thus, if the patterns observed can in any form predict any future observation.

Introduction

Before proceeding to the analysis, it would be helpful to define how is the dataset composed, providing a background to the reader, and preventing misunderstandings. The nine variables analyzed, and their distributions, present an overall sight of the airplane movements within India, helping the reader to understand the logic behind the quantitative research developed in this project. The dataset shows that the majority of these travels are made between Delhi and Mumbai, the most chosen destination and source cities, summarizing an overall 40% of all the flying made within the Asian country. The average length of the flights is 12.22 hours, this variable is strictly associated with the number of stops that the travelers have to do before arriving at their destination, and it is expected to have a high coefficient of variation over the fare. 84% of the movements are subjected to one single stop, measures that vary according to the source and destination city as well as the airline that operates the flight. The airplane companies have also an influence over the departure and the arrival time, variables that are expected to have an impact on the prices outlining the peak and bottom hours.

The nine companies analyzed are the most dominant low-cost carriers of the Indian peninsula, accruing a market share of 82% over all the sold flight tickets in India, as per statista.com. These airlines are the biggest provider of short and cheap flights, with few stops and, on average, low-quality service. Since the low price and the cut of all comforts, high price variation is expected whenever an unnecessary and unplanned convenience is added, such as in-flight meals, extra luggage, or preferential lines. In our dataset, the variables that most prominently represent an increase in the client's comfort are the departure and arrival time and, even more importantly, the class in which the customer decides to fly in. It is thus of great interest to the analysis relative to these variables and their impact on the price when compared to more economically tangible variables such as the overall flight duration.

Another helpful and relevant analysis would have been the price comparison across time, this type of research would guarantee future travelers a panoramic picture of the oscillation of the prices throughout the year, suggesting the best months in which to travel. Unfortunately, due to the dataset limitation in time, this kind of analysis could not be conducted, but qualitative research has been made to help travelers over this topic. According to data gathered from statista.com and skyscanner.com, for in-country movements in India, the peak season is between December and February, coinciding with Indian main holidays, while the cheapest season is positioned in the summer months. The period chosen by the dataset is considered a 'shoulder-period' for within-country movements and can therefore be considered as an approximate average of the prices between the top and the bottom seasons.

A major objective of this project, as stated in the abstract, is to be able to have a general view over the trends that flight fares go through in order to explain present variations and understand future patterns. Even though the variables this project relied upon are solid and applicable for every travel across time, there may be others that can unexpectedly change the prices airlines charge, such as the cost of fuels or passenger demand. So, since prices are difficult to predict in an exact way because they depend on a large number of unpredictable variables, the relevance of this project must be found among the correlations it presents, thought to be a valid representation of the mutation of prices for past, current and future purchases. That is to say that the aim of this work is not to crystal-ball future flight fares but to bring to the attention of the reader how the enlisted variables might have an impact on future purchases.

In the script part, we used the software R to conduct the quantitative analysis aimed at correlating all the variables listed in the dataset to the fare. To do so, we have most of the time used the linear regression model,

considering the price as the independent variable and all the categories as regressors, and observed the produced correlation. For a complete understanding and interpretation of the statistical results, we considered the introduction of graphs as an essential, complementary part of the mathematical results displayed.

A last fundamental point to discuss before starting the quantitative analysis is an acknowledgment of the possible biases. Since the data has been extracted from a website of a private company, and not from a public governmental agency, it is necessary to say that the data may be in some ways biased by the records of Easemytrip.com.

The Study

1) How does the ticket price vary between Economy and Business class?

Initially, we were interested in understanding whether the ticket price varied between Economy and Business classes. By running the regression model with price as the independent variable and class as the regressor we found that the average price in business class (β0) was equal to 577,94 €, the average price in economy class (β0 + β1) was equal to 72,3 €, and that the average difference between business and economy (β1) was equal to -505.64 €. Looking at the p-value, we can say that β0 and β1 differ from 0 at every significance level. Moreover, the R squared is 0,8796, therefore the class counts for approximately 88% of the price variance.

To better illustrate the distribution of the variable class, a categorical variable, we chose to use a barplot:

2) Does the price change with the airline?

We dug further into the analysis investigating whether the fluctuations in price were due to the different characteristics of the companies operating these flights. In fact, without accounting for classes, it seemed like there was a big difference between airline prices.

```
Call:
lm(formula = flightfare$price.eur ~ flightfare$airline)

Residuals:
    Min      1Q  Median      3Q     Max
-315.51 -202.00  -15.02  168.87 1019.42

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                258.577      0.774  334.08   <2e-16 ***
flightfare$airlineAirAsia  -213.575      1.900 -112.42   <2e-16 ***
flightfare$airlineGO_FIRST -196.405      1.640 -119.75   <2e-16 ***
flightfare$airlineIndigo   -200.011      1.313 -152.38   <2e-16 ***
flightfare$airlineSpiceJet -190.605      2.445  -77.97   <2e-16 ***
flightfare$airlineVistara    75.785      0.989   76.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 220.1 on 300147 degrees of freedom
Multiple R-squared:  0.2227,    Adjusted R-squared:  0.2226
F-statistic: 1.719e+04 on 5 and 300147 DF,  p-value: < 2.2e-16

> |
```

Looking at the coefficient we can see there is a difference in price mean among the flight companies different from 0 at every significance level.

To go deeper into the analysis, we run a multiple regression line keeping price as the independent variable and class and companies as regressors. By running a multiple regression line accounting for both class and airline, however, the difference between airline prices, keeping the class constant is very little.

```
Call:
lm(formula = flightfare$price.eur ~ flightfare$airline + flightfare$class)

Residuals:
    Min      1Q  Median      3Q     Max
-418.23  -30.38   -5.23   29.03  760.79

Coefficients:
                          Estimate Std. Error   t value Pr(>|t|)
(Intercept)               550.22830    0.36884  1491.772  < 2e-16 ***
flightfare$airlineAirAsia -13.66006    0.74178   -18.415  < 2e-16 ***
flightfare$airlineGO_FIRST  3.51023    0.64492     5.443 5.25e-08 ***
flightfare$airlineIndigo   -0.09548    0.52398    -0.182    0.855
flightfare$airlineSpiceJet  9.31021    0.94668     9.835  < 2e-16 ***
flightfare$airlineVistara  42.75970    0.37888   112.858  < 2e-16 ***
flightfare$classEconomy  -491.56644    0.37119 -1324.295  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.15 on 300146 degrees of freedom
Multiple R-squared:  0.8864,    Adjusted R-squared:  0.8864
F-statistic: 3.903e+05 on 6 and 300146 DF,  p-value: < 2.2e-16

> |
```

We can notice that the R-squared of the model that only accounted for classes and the one accounting for classes and airlines are similar. We therefore concluded that airline differences do not influence the price keeping class constant. This is because there is a huge price gap between the economy and the business classes, as shown in model 3:

```
lm(formula = flightfare$price.eur ~ flightfare$class)

Residuals:
    Min      1Q  Median      3Q     Max
-445.94  -28.98   -7.02   31.21  775.84

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               577.9409     0.2834    2040   <2e-16 **
flightfare$classEconomy -505.6451     0.3415   -1481   <2e-16 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.64 on 300151 degrees of freedom
Multiple R-squared:  0.8796,    Adjusted R-squared:  0.8796
F-statistic: 2.192e+06 on 1 and 300151 DF,  p-value: < 2.2e-16
```

Looking at our observations, we noticed that some companies operated only economy, thus we decided to run two different analyses by creating two subsets: one for the economy class and one for the business class, plotting the respective barplots.

Economy class by airline:


```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   80.4505     0.1774  453.47   <2e-16 ***
data.economy$airlineAirAsia  -35.4487     0.3540 -100.14   <2e-16 ***
data.economy$airlineGO_FIRST -18.2784     0.3109  -58.79   <2e-16 ***
data.economy$airlineIndigo   -21.8841     0.2579  -84.86   <2e-16 ***
data.economy$airlineSpiceJet -12.4784     0.4462  -27.96   <2e-16 ***
data.economy$airlineVistara    5.4259     0.2322   23.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.87 on 206660 degrees of freedom
Multiple R-squared:  0.1092,    Adjusted R-squared:  0.1092
F-statistic:  5066 on 5 and 206660 DF,  p-value: < 2.2e-16
```

As we can see, the companies that include the business class in their offer, have a higher mean price, mainly because are not "low cost" companies as the others.

Business class by airline:


```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 518.4414     0.7485  692.65   <2e-16 ***
data.business$airlineVistara  91.8059     0.9297   98.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.8 on 93485 degrees of freedom
Multiple R-squared:  0.09445,   Adjusted R-squared:  0.09444
F-statistic:  9750 on 1 and 93485 DF,  p-value: < 2.2e-16
```

By comparing the R-squared of the model with class= economy and class=business we can conclude how the difference in Airline counts more on the variance of the price when flighting economy.

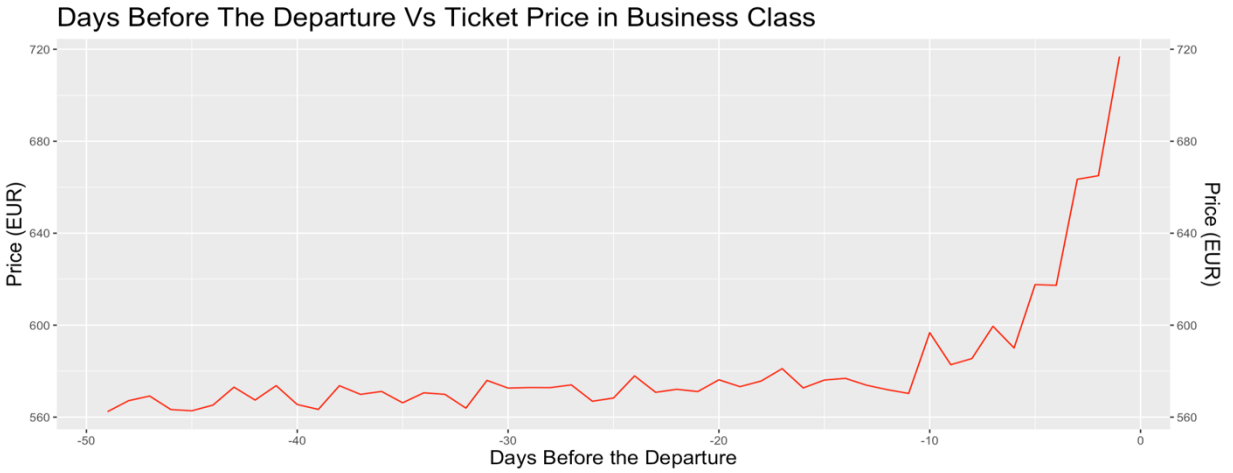
3) How is the price affected when tickets are bought X days before departure?

After these results, we focused on whether the price was affected when tickets were bought X days before departure.

Therefore, we added a new variable to the data set and multiplied by -1 the days left so that now the x-axis shows how many days before the day of the departure (day n.0) to simplify the graph understanding.

To simplify even further the visualization and to enhance the description of our data set, we created two vectors with the mean price for a ticket for each day before the departure instead of plotting each observation separately.

As we can see, the price fluctuates very little until around two weeks before departure when most travellers start buying their tickets. Then around the 6th day before departure, the ticket price in economy class diminishes a little, just to soar again around the 4th day.



On the other hand, the business class does not exhibit the high fluctuations in price that characterize the tickets in economy class. Between the 50th and the 11th day before departure ticket prices in business class increased relatively little and started soaring exponentially only by the 10th day before departure.

With these data in hand, we can safely discard, at least in this data set, the fake myth of "last-minute deals". In fact, tickets for both economy and business class increase in their daily mean price going toward the day of the departure.

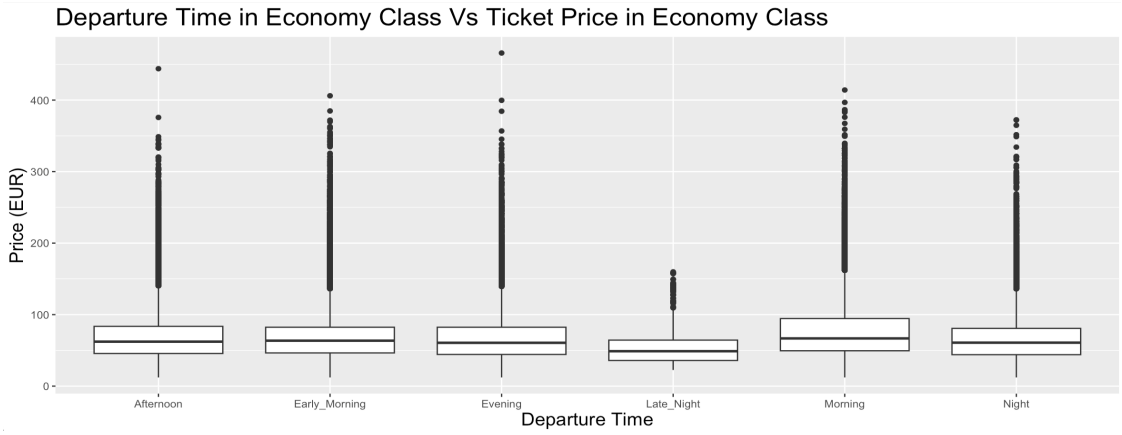
Performing a linear regression for both the business class and the economy class we indeed found that the R-squared is significantly higher for the economy class, therefore how many days in advance a ticket is bought explains a lot of the variability in prices.

```
lm(formula = price_mean_economy$price.eur ~ price_mean_economy$days_before)

Residuals:
    Min      1Q  Median      3Q     Max
-29.118 -14.000   3.507   9.682  40.248

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   122.3588     4.9360   24.79  < 2e-16 ***
price_mean_economy$days_before   1.8616     0.1719   10.83 2.28e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.01 on 47 degrees of freedom
Multiple R-squared:  0.714,     Adjusted R-squared:  0.7079
F-statistic: 117.3 on 1 and 47 DF,  p-value: 2.276e-14
```

```
lm(formula = price_mean_business$price.eur ~ price_mean_business$days_before)

Residuals:
    Min      1Q  Median      3Q     Max
-29.270 -13.374  -3.560   8.096 104.424

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   613.7252     6.5944  93.068  < 2e-16 ***
price_mean_business$days_before   1.2865     0.2296   5.603 1.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.73 on 47 degrees of freedom
Multiple R-squared:  0.4005,    Adjusted R-squared:  0.3877
F-statistic:  31.4 on 1 and 47 DF,  p-value: 1.066e-06
```

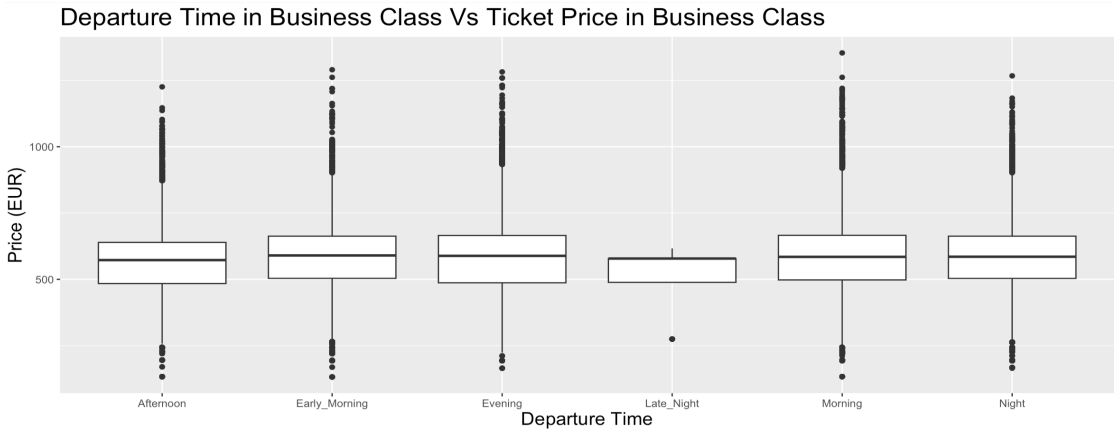
4) Does the departure time and arrival time affect the price?

Next, we asked ourselves whether the departure time and arrival time affected the price. We factored the variables arrival time and departure time and plotted a boxplot to visualize the dataset's central tendency, spread, and skewness.



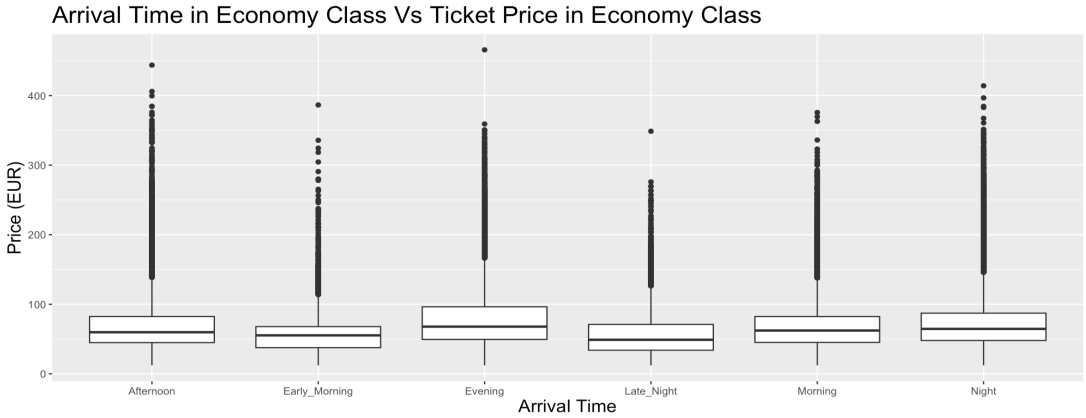
```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                           71.2041     0.2184 326.016  < 2e-16 ***
data.economy$departure_timeEarly_Morning   0.9594     0.2893   3.316 0.000913 ***
data.economy$departure_timeEvening        -1.2358     0.2928  -4.221 2.44e-05 ***
data.economy$departure_timeLate_Night    -18.5724     1.2195 -15.230  < 2e-16 ***
data.economy$departure_timeMorning         7.1052     0.2866  24.791  < 2e-16 ***
data.economy$departure_timeNight          -2.9384     0.3206  -9.167  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41 on 206660 degrees of freedom
Multiple R-squared:  0.008547,  Adjusted R-squared:  0.008524
F-statistic: 356.3 on 5 and 206660 DF,  p-value: < 2.2e-16
```

For economy class, the tickets, on average, cost more if the departure time is morning, while cost significantly less during late night and a bit less during early morning, afternoon, evening, and night.




```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                           561.625      1.272 441.653  < 2e-16 ***
data.business$departure_timeEarly_Morning   15.281      1.621   9.429  < 2e-16 ***
data.business$departure_timeEvening         18.029      1.609  11.206  < 2e-16 ***
data.business$departure_timeLate_Night     -39.434     12.193  -3.234  0.00122 **
data.business$departure_timeMorning         25.395      1.589  15.978  < 2e-16 ***
data.business$departure_timeNight           16.014      1.667   9.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 142.5 on 93481 degrees of freedom
Multiple R-squared:  0.002993,  Adjusted R-squared:  0.002939
F-statistic: 56.12 on 5 and 93481 DF,  p-value: < 2.2e-16
```

Regarding the price variations during the day for business class tickets, we found that tickets on average cost more if the departure time is in the morning, in the evening, at night, and in the early morning. On the other hand, the cost significantly decreases during the afternoon, reaching on average the lowest prices late at night.




```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        70.4773     0.2468 285.590   <2e-16 ***
data.economy$arrival_timeEarly_Morning -11.7741     0.4473 -26.320   <2e-16 ***
data.economy$arrival_timeEvening         8.7972     0.3046  28.876   <2e-16 ***
data.economy$arrival_timeLate_Night    -13.5734     0.4439 -30.575   <2e-16 ***
data.economy$arrival_timeMorning         0.1039     0.3184   0.326    0.744
data.economy$arrival_timeNight           3.5206     0.2956  11.908   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.71 on 206660 degrees of freedom
Multiple R-squared:  0.02283,   Adjusted R-squared:  0.02281
F-statistic: 965.9 on 5 and 206660 DF,  p-value: < 2.2e-16
```


```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        534.438      1.349 396.276  < 2e-16 ***
data.business$arrival_timeEarly_Morning   -9.709      2.734  -3.551 0.000384 ***
data.business$arrival_timeEvening         61.710      1.604  38.479  < 2e-16 ***
data.business$arrival_timeLate_Night      35.989      3.559  10.113  < 2e-16 ***
data.business$arrival_timeMorning         37.081      1.653  22.437  < 2e-16 ***
data.business$arrival_timeNight           55.063      1.582  34.799  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141 on 93481 degrees of freedom
Multiple R-squared:  0.02327,   Adjusted R-squared:  0.02321
F-statistic: 445.3 on 5 and 93481 DF,  p-value: < 2.2e-16
```

For economy class, the tickets cost more if the arrival time is evening, while cost significantly less during early morning and late night and a bit less during afternoon, evening, and night.

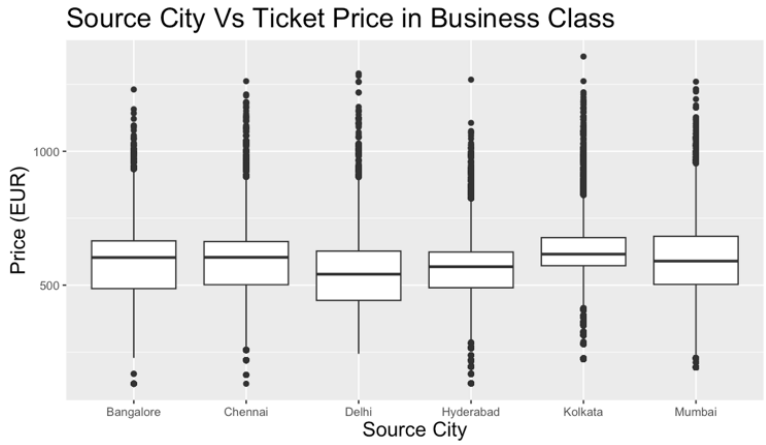
For business class, the tickets cost more if the arrival time is morning, evening, and night, while cost significantly less during early morning and a bit less during the afternoon and late night.

5) Does the source city and destination city affect the price?

Let us factorize the variables source city and destination city and let us see whether, for the economy class, the source city of the flight affects the price.


```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      72.4026     0.2169  333.82   <2e-16 ***
data.economy$source_cityChennai   0.2664     0.3333    0.80    0.424
data.economy$source_cityDelhi    -3.2282     0.2932  -11.01   <2e-16 ***
data.economy$source_cityHyderabad -3.8092    0.3271  -11.64   <2e-16 ***
data.economy$source_cityKolkata   9.6423     0.3131   30.80   <2e-16 ***
data.economy$source_cityMumbai   -2.4545     0.2964   -8.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.93 on 206660 degrees of freedom
Multiple R-squared:  0.01186,   Adjusted R-squared:  0.01184
F-statistic: 496.2 on 5 and 206660 DF,  p-value: < 2.2e-16
```

We can see that Kolkata has the highest effect on price benchmarked against the reference category, in this case, Bangalore. Whereas Delhi, Hyderabad, and Mumbai have negative effects, which may be due to the higher availability of flights in these large cities. It must be pointed out, however, that the R-squared features a low value, and Chennai is not statistically significant because the p-value is greater than 0.05.



```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      590.959      1.091 541.737  < 2e-16 ***
data.business$source_cityChennai   4.587      1.659   2.765  0.00569 **
data.business$source_cityDelhi   -55.281      1.503 -36.788  < 2e-16 ***
data.business$source_cityHyderabad -36.884    1.646 -22.405  < 2e-16 ***
data.business$source_cityKolkata  31.728      1.625  19.519  < 2e-16 ***
data.business$source_cityMumbai  -10.813      1.475  -7.332 2.29e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.9 on 93481 degrees of freedom
Multiple R-squared:  0.03878,   Adjusted R-squared:  0.03873
F-statistic: 754.3 on 5 and 93481 DF,  p-value: < 2.2e-16
```

Here, all the coefficients are statistically significant, with Delhi and Hyderabad as source cities making on average the price decrease the most with respect to the omitted variable Bangalore.


```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       72.8015     0.2197 331.319  < 2e-16 ***
data.economy$destination_cityChennai   0.2562     0.3319   0.772     0.44
data.economy$destination_cityDelhi    -4.0395     0.2995 -13.488  < 2e-16 ***
data.economy$destination_cityHyderabad -3.4405    0.3259 -10.556  < 2e-16 ***
data.economy$destination_cityKolkata   6.4042     0.3110  20.595  < 2e-16 ***
data.economy$destination_cityMumbai   -1.7461     0.3004  -5.812 6.18e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.03 on 206660 degrees of freedom
Multiple R-squared:  0.007147,  Adjusted R-squared:  0.007123
F-statistic: 297.5 on 5 and 206660 DF,  p-value: < 2.2e-16
```

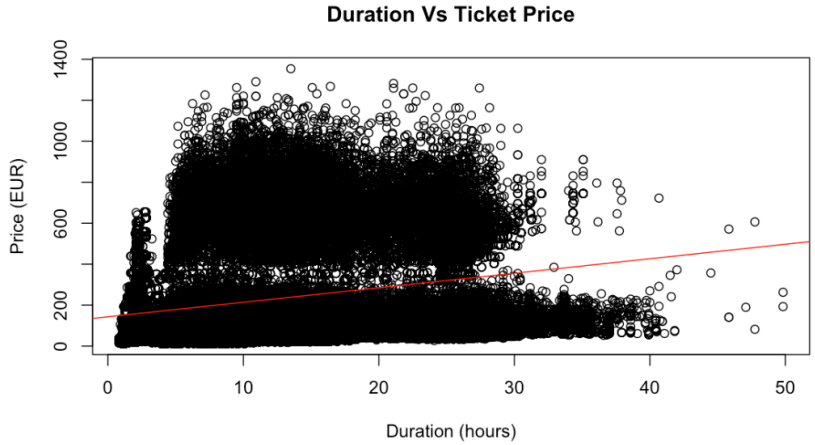
```
Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         592.158      1.096 540.362  < 2e-16 ***
data.business$destination_cityChennai  -2.370      1.637  -1.448    0.148
data.business$destination_cityDelhi   -63.189      1.538 -41.087  < 2e-16 ***
data.business$destination_cityHyderabad -37.444      1.620 -23.118  < 2e-16 ***
data.business$destination_cityKolkata  31.960      1.587  20.137  < 2e-16 ***
data.business$destination_cityMumbai  -10.628      1.492  -7.125 1.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.5 on 93481 degrees of freedom
Multiple R-squared:  0.0443,    Adjusted R-squared:  0.04424
F-statistic: 866.5 on 5 and 93481 DF,  p-value: < 2.2e-16

> |
```

We can see that the coefficient for Chennai in both regressions is not statistically significant at any confidence interval. Regarding the outliers, those data points that fall outside the interquartile range (IQR) and are displayed as individual points outside the "whiskers" lines in the boxplot, it is possible to notice how tickets in the Economy class suffer only from upper outliers whereas tickets in the Business class also from lower outliers.

6) Does the duration influence the price?

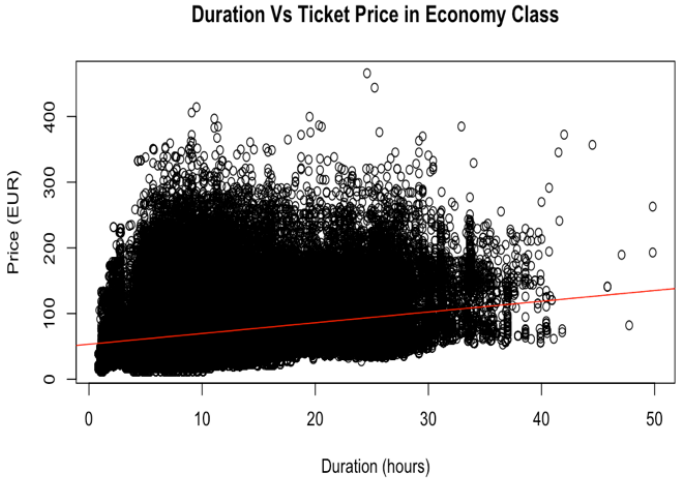
To answer the question we first provided a general analysis without distinguishing between classes

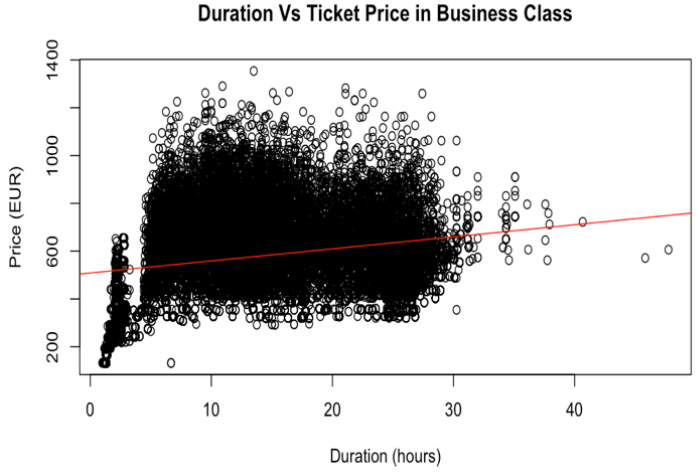
```
lm(formula = flightfare$price.eur ~ flightfare$duration)

Residuals:
   Min     1Q Median     3Q    Max
-399.6 -166.9 -124.3  215.8 1114.9

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         143.14254    0.87961   162.7   <2e-16 ***
flightfare$duration   7.08973    0.06203   114.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 244.4 on 300151 degrees of freedom
Multiple R-squared:  0.04171,   Adjusted R-squared:  0.0417
F-statistic: 1.306e+04 on 1 and 300151 DF,  p-value: < 2.2e-16
```



On average an hour's increase in the duration of the flight makes the price increase by approximately 7€.

To dig further into the analysis, we must be able to differentiate between classes.





In both cases, there is a correlation between duration and price even if it is not that strong.

Moreover, the R-squared is very low in both analyses.

Let us see if conducting the analysis with more regressors helps increase the R-squared.

```
Call:
lm(formula = flightfare$price.eur ~ flightfare$duration + flightfare$stops)

Residuals:
   Min     1Q Median     3Q    Max
-351.6 -180.7 -116.5  221.6 1158.5

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  180.29417    1.08371  166.37   <2e-16 ***
flightfare$duration            5.30631    0.07185   73.85   <2e-16 ***
flightfare$stopstwo_or_more -106.32377    2.16072  -49.21   <2e-16 ***
flightfare$stopszero         -88.78670    1.58871  -55.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 242.3 on 300149 degrees of freedom
Multiple R-squared:  0.05848,   Adjusted R-squared:  0.05847
F-statistic:  6214 on 3 and 300149 DF,  p-value: < 2.2e-16
```

For both Economy and Business class, there is correlation and R-squared is higher using separate analysis for the two classes

```
Call:
lm(formula = data.economy$price.eur ~ data.economy$duration +
    data.economy$stops)

Residuals:
   Min     1Q Median     3Q    Max
-79.58 -24.04  -9.93  13.09 378.30

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   61.16206    0.20002  305.77   <2e-16 ***
data.economy$duration          1.07314    0.01372   78.19   <2e-16 ***
data.economy$stopstwo_or_more 22.87340    0.36391   62.85   <2e-16 ***
data.economy$stopszero       -19.34477    0.28964  -66.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.62 on 206662 degrees of freedom
Multiple R-squared:  0.1203,    Adjusted R-squared:  0.1203
F-statistic:  9421 on 3 and 206662 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = data.business$price.eur ~ data.business$duration +
    data.business$stops)

Residuals:
    Min      1Q  Median      3Q     Max
-651.86  -66.23   -6.11   62.46  750.12

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   625.55844    1.01417  616.82   <2e-16 ***
data.business$duration         -1.62219    0.06324  -25.65   <2e-16 ***
data.business$stopstwo_or_more 169.12605    3.47764   48.63   <2e-16 ***
data.business$stopszero       -315.53770    1.54104 -204.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 113.7 on 93483 degrees of freedom
Multiple R-squared:  0.3647,    Adjusted R-squared:  0.3647
F-statistic: 1.789e+04 on 3 and 93483 DF,  p-value: < 2.2e-16
```

Results Analysis

The findings of this study may have brought to the attention of the readers some unexpected results and some suggestions for future air traveling. The most evident result is the huge price difference that exists between economy and business class, which counts for 88% of the price variance. Surprisingly enough, flying in business class has a much higher degree of relevance over the price than other economically significant factors such as flight duration, which, although its evident impact on the fuel expenses that the airline has to bear, increases the mean price by roughly 4% of its value for each hour added. While it is true that larger business seats avoid the full maximization of the spaces inside the aircraft, the astonishing fact that this kind of transportation costs nearly 8 times the cheaper one must be a symptom of something else. The two factors that have to be taken into consideration are passenger demand and the low-cost effect. Low-cost airlines are deeply concerned with the maximization of space and the provision of the minimum necessary service, the opportunity cost of losing space and adding features to the service to be appropriate for a business class traveler should be thus repaid by a great sum. This great disparity in prices affects the market as a whole and, as demonstrated in the second point of the research, it fades away price differences between airlines, creating two subgroups of them: those who offer business class travel and those who don't.

Of all the study, the most unforeseen evidence that has been highlighted is probably the proof that last-minute booked flights are not the cheapest on the market. This false myth has been proven wrong in the 3rd passage of the study which shows how the flight fares increase sharply between 10 and 15 days before the departure and reach incredible prices that for economy class are, at their peak, three times those of two weeks prior. Things don't change much for the business class which experiences the price growth similarly to the economy one. The reasons behind it have to be found in the willingness of airplane companies to fill their plane up as soon as possible to avoid cancellation and to know ahead of the flow of people expected and the employees to assign.

When considering the variables relative to the time of departure or arrival and the source or the arrival cities, the dominant factor affecting the flight fare is the passengers' demand. As the demand in general decreases for night flights, we weren't surprised by the finding that early-morning arrivals and late-night departures are characterized by lower prices, even though they have higher personal costs. Our expectations suggested that a higher demand and therefore higher prices for business seats during night movements were likely to be found as they provide better sleeping comfort, but this was not the case. A similar influence that the departure and arrival times have over the prices of the two classes was in fact found and, to understand these slight differences, a more specifical study based on the personality traits of these two types of travelers needs to be made. Passengers' demand would be as well at the center of the discussion when confronting different prices depending on the source and the destination city: higher demand involves higher prices, but it also translates into more competition between airlines. The overall picture we found does not assure us an impeccable analysis and comparing the high volume of traffic that all these airports go through (in all of them passenger transit was more than 15 million people in 2022, according to the 'Airport Authority of India' report) and the small statistical differences in prices, is possible to conclude that other factors such as the runway rent or the structural management must be considered.

Conclusion

Through the use of statistical indicators and the support of the R software, we have been able to analyze and study 300 thousand elements, correlating each of them to the nine variables described throughout the pages of this project, and obtaining noticeable results. Some of them were aligned with our expectations, others not. The relevance of the study must be found exactly in those outcomes that we, as well as many travelers, didn't expect at all. That is because they bring to our attention variables that were not thought to have an impact on flight fares or, even worse, were thought to have the opposite effect on the prices of airplane tickets. Although using a sample of Indian flights, the investigation of the causes behind each of these results and the conclusions drawn from them, suggest that these patterns are likely to present themselves in every low-cost company of the world, providing to the study its value and replicability, and achieving the intent of it as stated in the abstract.

Reference

Dataset: https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data