Flight price prediction

Lujain Bouh d00264191

December 2023

1 Abstract

The flight pricing plan has developed into a sophisticated set of regulations, including numerical simulations that influence airfare marketing methods, since the airline firm was privatised (**Kimbhaune**). Research has shown that these principles are affected by a variety of events, despite the fact that they are still mostly unknown. Even while they are still significant, factors like distance are no longer the only ones that affect pricing structure.

The importance of this analysis is for customers to consider this strategy and gain insight into the pricing patterns and also receive a projected price that they may view before to purchasing airline tickets, resulting in cost savings.

RESEARCH QUESTION: Distance is not the only thing that affect pricing structure , what other factors have an impact on prices of flight in India

2 Motivation

There are gaps/limitation in the current existing flight price prediction model , these are stated in the next section. Therefore, the objective of this research paper is to improve the existing model and increase their accuracy by addressing the research question above

3 Literature review

Anybody who has ever purchased a plane ticket is aware of how quickly the costs may change. Modern methods known as Revenue Management are used by aircraft to conduct a distinct price structure. The most affordable ticket that is available can be depending on the historical period, costly or affordable. The Seasons, like winter and cricket, can also affect pricing. The seller's main goal is to increase its price, yet the purchaser is constantly searching get the best possible deal. Generally speaking, purchasers attempt to get the ticket in advance of the departure date. Given that many people think that tickets would cost more if purchased closer to the date of departure, even though this isn't

always the case. The ultimate objective is to book the flight at the best price at a specific time from a specific departure location to a specific arrival location.

Many studies have been completed using machine learning to try to predict the flight of prices . These include : The survey study by Supriya Rajankar on the use of machine learning models for aviation fare forecasting makes use of a small dataset of flights between Delhi and Bombay. The techniques employed are support vector machine (SVM), linear regression, and K-nearest neighbours (KNN)(**Kuhn**). Another example would be , studies by Santos studied airline routes over several months connecting Madrid with London, Frankfurt, New York in addition to Paris(**Groves**).

Tziridis et al.(**K. Tziridis**) projected ticket prices using eight machine learning algorithms, such as ANNs, SVM, and LR, and then compared how well each performed. With an 88% accuracy rate, the most accurate regression model was found.

There are limitations and gaps in the existing model which can be the main focused of future research in order to improve flight prediction prices. Future models can incorporate data from air tickets transactions such as time or arrival , covered additional information, the placement of the seat etc as they can offer more information about a particular route and airline(**A. Bhatia**). Additionally , the existing models fail to include the sudden influx when there's concerts on , sport matches or breakout of diseases such as COVID which can affect the accuracy of the model.

4 Data description

The dataset chosen that I felt is best to address my research question is called 'flight price prediction'. This dataset consists of eleven parameters with the majority of the columns containing 1083 but some contain 1082 due to missing data. Before data cleaning, the dataset contains the following columns:

- Airline: The column contains the name of the airline firm. It has twelve airlines.
- Date of journey :the date the journey was made with each specific airline, it a numerical variable.
- Source: Lists the city that the flight leaves from. Categorical variable , having 5 unique cities.
- Destination: City where the flight will land. It is a categorical feature having 6 unique cities.
- Route : shows the route the flight takes. It is categorical variable.
- Dep. time is the departure time which is the time the flight leaves the source city. Numerical variable .

- Arrival time is the time that the flight arrives in the destination city. Numerical variable.
- Duration: is the length of the time so arrival time- departure time. Numerical variable
- Price : Price of the flight for each specific journey.
- Total Stops: is the number of layover/stop in each flight. Can range from 0 to 4 stops. Discrete numerical variable.
- Additional information: provides extra information on the passengers such as if they paid for extra luggage, business class, had to change airports, had meals included etc. it's a categorical variables.

However , these columns are altered after data cleaning to make it look more presentable and tidier , this will be further discussed in the methodology section and in the coding script. Explanatory data analysis was completed as part of the pre-processing step to clean the data and ensure its suitable for data visualization and statistical analysis , these include the following :

1. The first step carried out was to convert date the format to %d/%m/%Y in the date of journey column to make it presentable and suitable for analysis.
2. Looking at the 'arrival time' column and from the dataset , it's visible that some rows have the date written after if the time is after midnight(i.e.00:00) and some rows don't , this would cause a problem later on for data visualization and statistical analysis. To resolve this issue , the string was split after the space to get rid of the date , as it's not needed due to it being obvious what date it is from looking at the arrival time column and the date of journey.
3. The 'total stop' column was converted to float, so the numbers are numerical rather than strings again for the purpose of making it easier for analysis later on.
4. As mentioned above, there was missing values in the dataset in column such as 'Total stop' and 'Route'. Forward fill were used to replace the missing value as the row above it had the same source , destination city and route so it was more appropriate to do that than to remove the row
5. Duration column was then split into 'duration hours' and "duration minutes" . Duration column is then dropped
6. 'Duration hours' and 'Duration minutes' were then converted to integers.
7. 'Duration hours' was then converted into minutes and a new column called 'duration in minutes' was created and it was the sum of 'duration hours' and "duration minutes" . This column is used for data visualisation and statistics.

8. Index is then reconstructed as it got disrupted and jumbled up from dropping the rows with the missing values.
9. Another new column was created called "Class" which was based on the price column. If the price is above 12000 Indian rupee then it's classified as business class and if it's less than 12000 then it's considered to be economy class.
10. Date column is converted to string so it can be split to year and month for analysis and was then dropped.
11. A new column called 'Log price' was created which consisted of finding the log value of the price column, due to the fact that prices are skewed and this would effect my data
12. Column's order was then changed so the dataset can tell a story from looking at the airline column to the additional info column.
13. Column names were also modified to ensure that they all started with capital letters to make the dataset look presentable.
14. new dataset was then stored as csv file

Post the EDA , the new dataset now has 16 columns with all of the columns containing 10682 variables. A CSV file of the updated dataset is also uploaded.

5 Formulation of Hypotheses:

Based on the research question stated above , multiple hypotheses will be tested in order to get an answer for our research question and to determine which factors effect the price of flight to help with cost savings and make travel more affordable.
hypothesis will be tested with significance level of 0.05 The following hypotheses are tested:

1. Is there a relationship between class of the ticket and flight price?
2. Does duration of flight have effect on price?
3. Does the number of stops have effect on price?
4. Does Airline have an impact on price?
5. Which variables can you remove to improve the flight prediction model.

6 Methodology

- **Data visualization**: For the data analysis, univariate and bivariate/multivariate plot were conducted . Count plot or boxplot were plotted for each of the variables individually. Bivariate/multivariate plot were then constructed to study the relationship between each variables, studying the relationship

between the variables determines which statistical tests are required. The bivariate /multivariate plots include:

1. Airline vs Price.

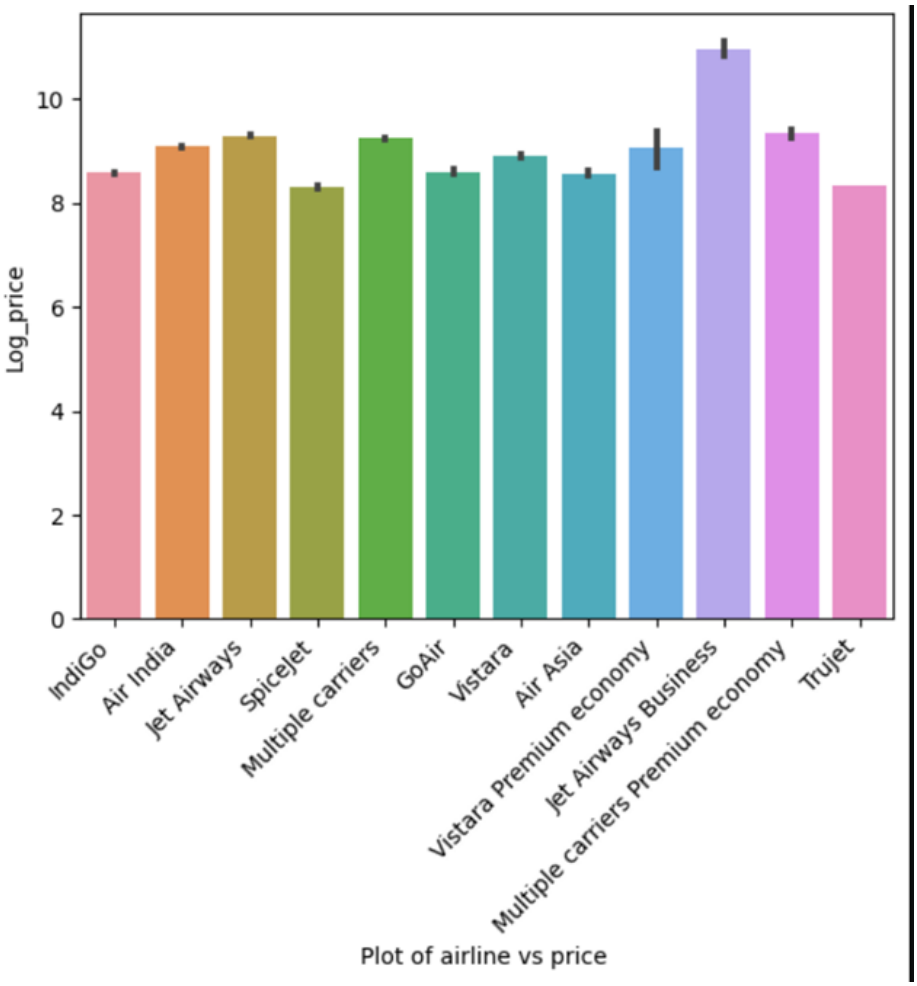

Figure 1: Relationship between number of stops and flight prices

2. Total stops vs Price.

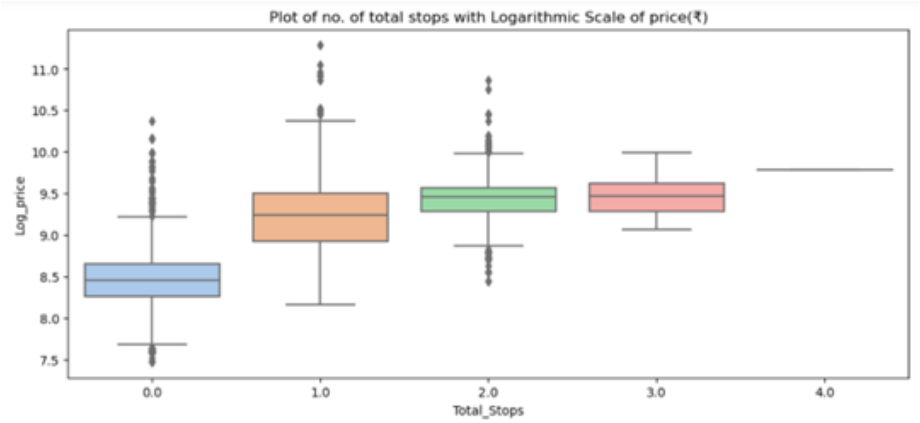

Figure 2: Relationship between number of stops and flight prices

3. Destination vs Price.

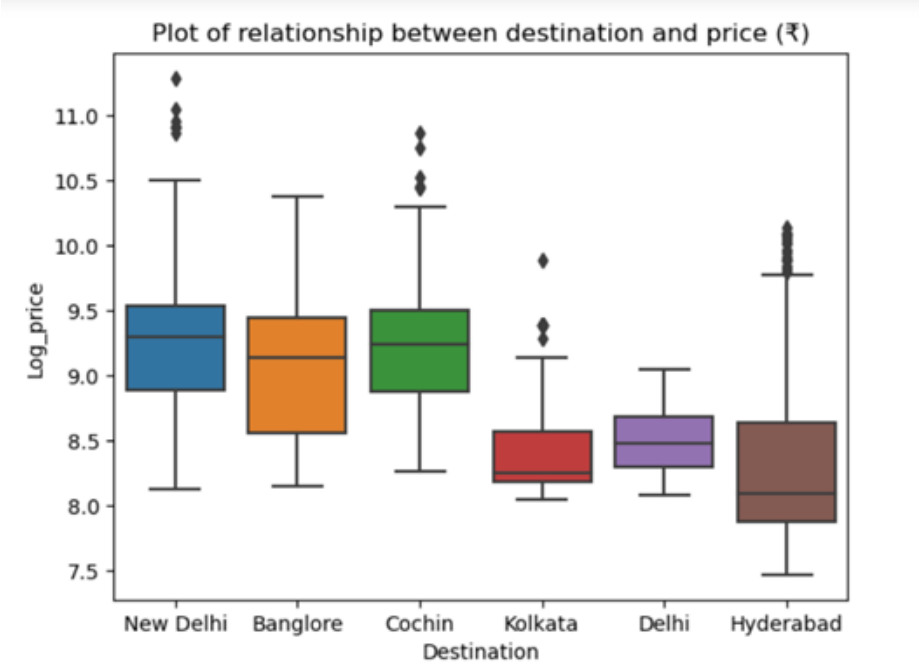

Figure 3: Relationship between destination and flight prices

4. Source vs Price.

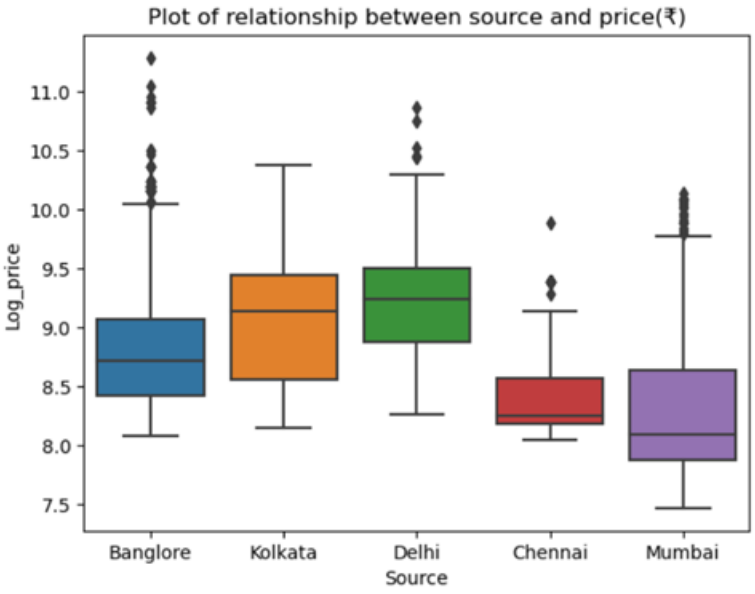

Figure 4: Relationship between Source and flight prices

5. item Airline vs log price vs class.

Figure 5: Relationship between airline, class and flight prices

6. Additional info vs log price.

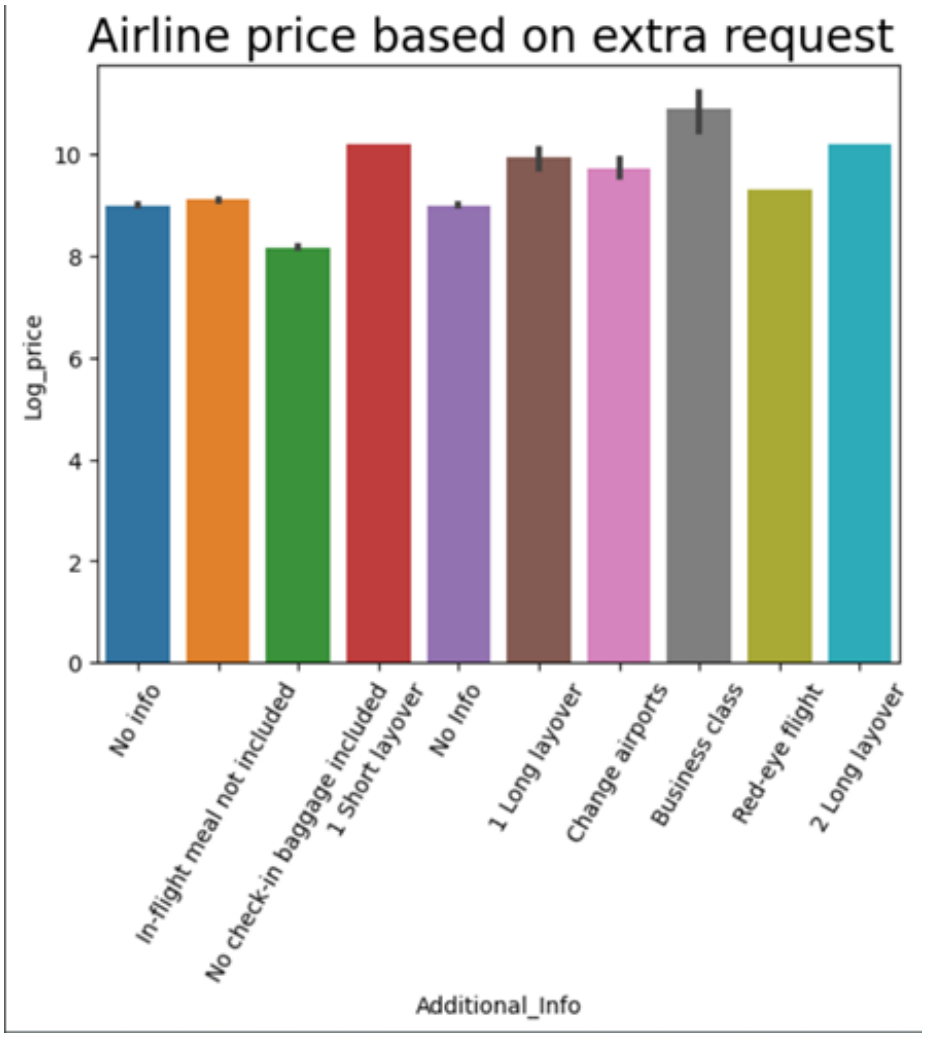

Figure 6: Airline price based on extra request

7. Class vs log price

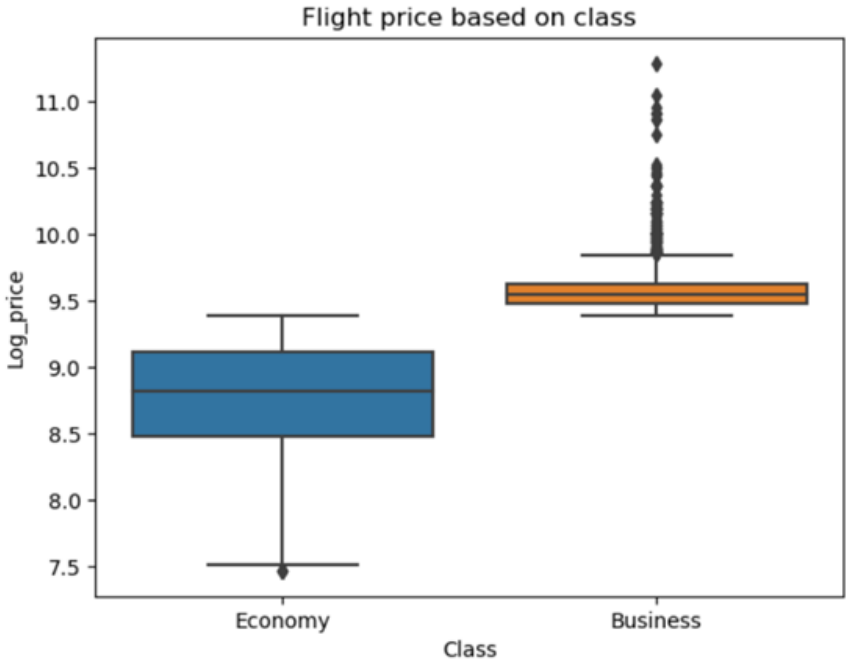

Figure 7: Airline price based on class

8. Duration__Minutes vs Price.

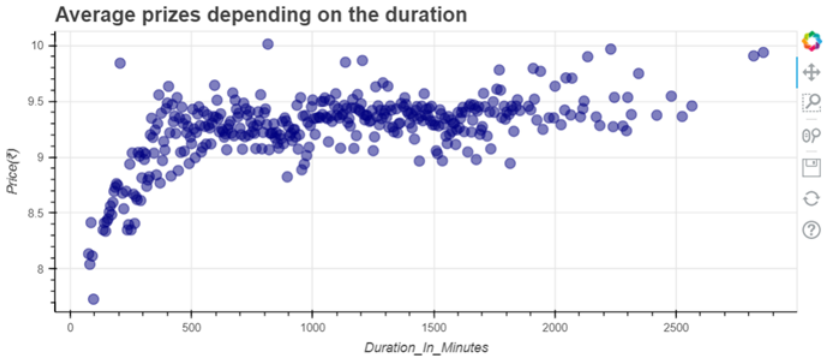

Figure 8: Airline price based on duration of flight

9. Duration__Minutes vs Price vs Class.

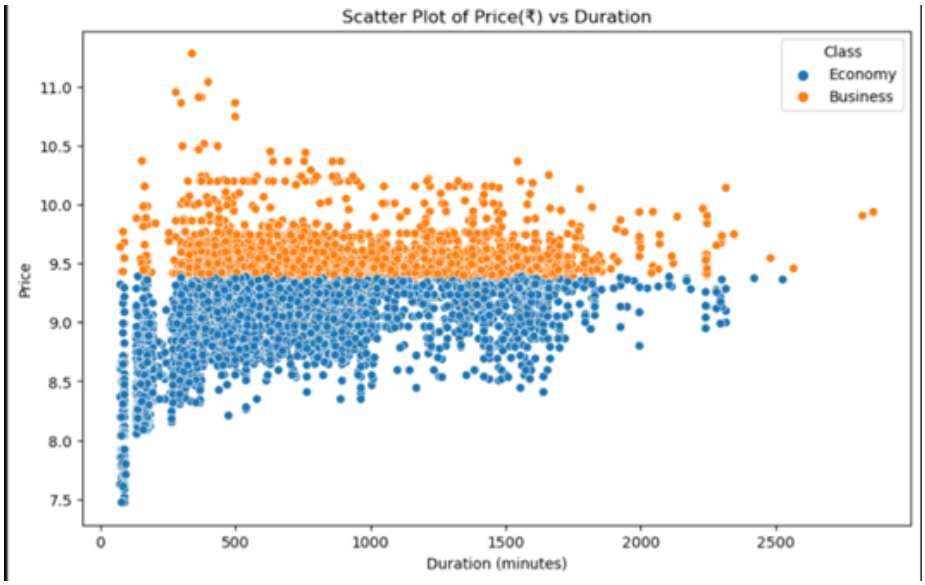

Figure 9: Airline price based on duration of flight and class

10. Duration_Minutes vs Price vs Total stops

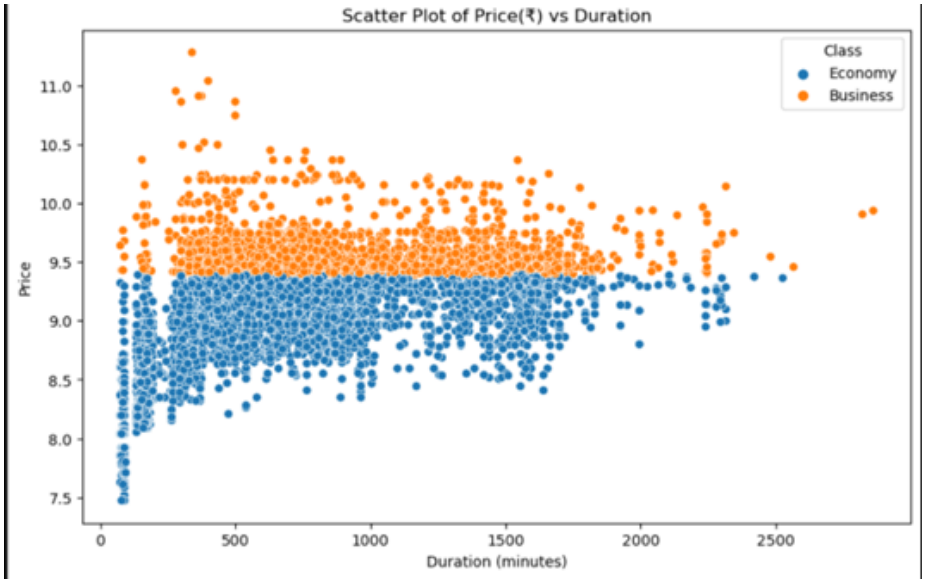

Figure 10: Airline price based on duration of flight and number of stops

11. Journey month vs Price

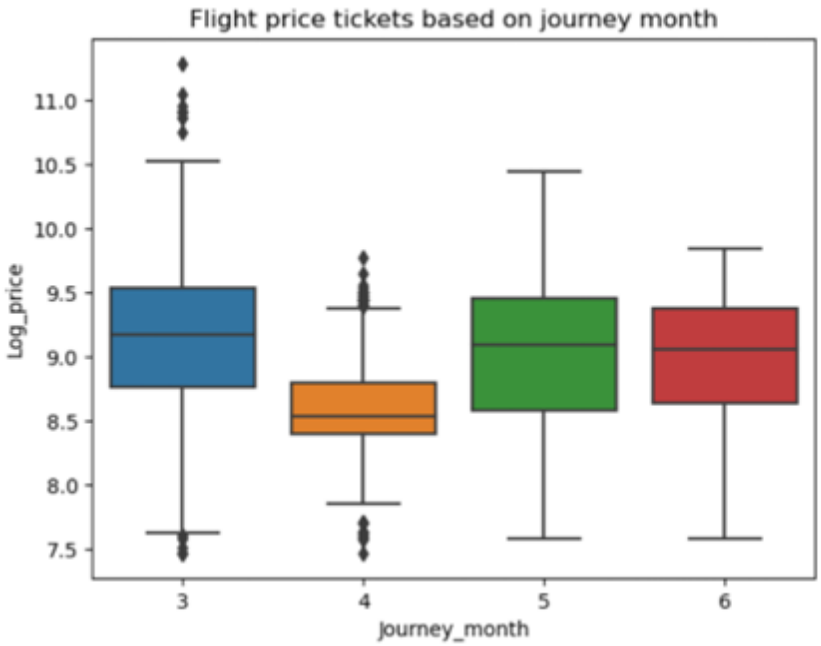

Figure 11: Airline price based on month of journey

Data visualisation techniques, such as charts, diagrams, maps, and more, to translate and show complicated data and relations in an understandable manner. Making data comprehensible typically requires selecting the appropriate technique and configuring it properly. The data visualization tools were used are Matplotlib , Seaborn and Bookeh.

- **Statistical analysis:** Data exploration: Descriptive Statistics was completed for each variable:

    1. Duration_In_Minutes: Mean of Duration_In_Minutes is 643 and the mean is 520 which shows that it could be roughly symmetrical, but the difference is due to the outliers are shown above. The IQR is : 760. Duration values range from 75(lowest minute) to 2860(highest minute). Greater data variability is indicated by a higher standard deviation which is 507.85 in this variable.
    2. Log_price. Average price of flight is 8.98rupee. Median is 9.3 , shows that the distribution is roughly symmetrical but however , presence of outlier must be noted. Price of flight ranges from approx. 7.47 to 11.28 . Greater data variability is indicated by a higher standard deviation which is 4610.89 in this variable. Interquartile range(IQR): 9.4-88.57 : 0.33.

3. Total stops. Average number of stops is 0.82 . Median is 1 , difference between median and mean show that the distribution is roughly skewed. Number of stops per flight ranges from approx. 0 to 4. Greater data variability is indicated by a higher standard deviation which is 0.675 in this variable. Interquartile range(IQR): 1-0 : 1
4. Airline. There are 12 unique airlines used in the dataset . Most common used airline is jet airways . Destination. There are 6 unique destinations used in this dataset. Cochin is the most popular destination.
5. Source: There are 5 unique sources in this dataset. Delhi is the most common city that the airlines depart from

In order to answer the research question stated above which is ' what factors/variables in the data impact the price of flight tickets in India?' , multiple statistical analyses have been completed.

Hyptoheses:

- First hypothesis : How does class impact the price of flight tickets. First variable in the dataset we're going to look at is type/class of ticket so Business or Economy, the test was conducted is sample hypothesis tests to for mean. H0 and H1 were clearly stated.

  H0 mean_class_price =mean_business_price (i.e. no relationship).

  H1: mean_economy_price!=mean_business_price (i.e. relationship).

  Class vs price plot was performed using sns.countplot to visualise the distribution and therefore determine if it's appropriate to complete T test. Once that's established and assumptions are met such as normal distribution and large data size, homogeneity , P-value is then examined to decide whether to reject or fail to reject the null hypothesis and that will provide information on weather class is significant and help us predict future flight prices. If the assumptions are not met then ranksum test is used alternatively. Analysis and interpretation will be explained in the next section

- Second hypothesis:Does duration of the flight have an effect on price of flights.

  H0 and H1 were clearly stated.

  H0: Duration of the flight has no effect on price (i.e. no relationship).

  H1: Duration of the flight has an effect on flight pries(i.e. relationship).

  Model summary using 'ols' was conducted to extract the equation and analyse the R-squared, p-value.

  Just like the T test above , assumptions must be met before we can conduct ANOVA Test . such as variance is equal and normal distribution. Type

of ANOVA is also considered which is determined by whether the data is equal or not (type 1 is equal , type 2 is not).

If the assumptions are not met then alternative test is conducted which is Kruskal-Wallis test is used . Looking at the P-value of either the ANOVA or Wallis test will reveal if the duration of the flight is significant or not and therefore whether the null hypothesis is rejected or fail to be rejected. Interpretation of the result will be explained in the next section. P-value is then compared to the plot 'price vs duration' just to see if they align.

- Third hypothesis: Does number of stops have effect on price?

  H0: median of price = median of total stops (i.e. no relationship).

  H1: median of price != median of total stops (i.e. relationship).

  To consider the relationship between total stop and price , median was tested and compared , this was done using Mann-Whitney test (specifying the alternative is two sided).

  P value was then evaluated to determine if there's a relationship between total stop and price and see if it aligns with the box plot that was plotted in the data visualization section which highlighted that as the stop number increase the price of the flight ticket increase.

  Fourth hypothesis : Does the type of airline have an impact on the price of the flight ticket?

  H0: airline doesn't impact price of flights (i.e. no relationship).

  H1: airline impacts price of flights (i.e. relationship).

  Airline versus price was plotted in the data visualization section , for this statistical analysis , the two most popular airline were picked, and these are : IndiGo and Jet airways.

  The data was then sliced to extract the price from these airlines where they both have the same source and destination (i.e. from Bangalore to New Delhi) . The extracted data were labelled as df_indiGo and df_jet_airways and they were concatenated to form one data frame and then that was used in the analysis to answer the hypothesis. T test was then conducted to inspect the P value and either reject the null hypothesis or fail to reject the hypothesis. As mentioned earlier , assumptions for normality must be considered and tested before we interpret the p value from the t -test.

  For the second part of the experiment for testing the hypothesis , interaction test was conducted to investigate the relationship between price , airline, class of the tickets, total stops.

  H0: no interaction between price , airline, class of the tickets, total stops.

  H1 : there is interaction between price , airline, class of the tickets, total stops.

  Model summary using 'ols' was conducted to extract the equation and analyse the R-squared, p-value. Interaction plot was plotted of the mean

to determine if there's an interaction occurs or not by looking at the lines. ANOVA test is then conducted to analyse the p-value but as mentioned above, assumptions of equal variance and normality must be tested before interpreting the p-value.

- Fifth hypothesis: Which variables can you remove to improve the flight prediction model? H0: Reduced model is preferred.

  H1: full model is preferred.

  Full model includes Log price, duration(min), total stops, source , additional info , destination, journey month, airline, class. Reduced model includes Log price , class, destination, total tops, duration(minutes), journey.

  Assumptions of equal variance and normality are tested for both models to conduct ANOVA test to determine which model is preferred . Cooks distance was also measured for both models to identify outliers in X values.

  Model summary of both full model and reduced model was then compared to test the hypothesis.

  From the model summary , F-statistics(p-value), R2 , BIC and AIC were then analysed for comparison.

7 Data analysis:

- Hypothesis 1 : How does class impact the price of flight tickets?T-TEST was used to conduct this hypothesis.

  H0 mean_class_price =mean_business_price (i.e. no relationship).

  H1: mean_economy_price!=mean_business_price (i.e. relationship).

  The distribution is symmetrical and for the normality is roughly normal but the presence of outliers make it look skewed so we can interpret the p-value of t-test.

  P value is less than 0.05 therefore its significant and we can reject the null hypothesis and conclude there's a relationship between price and class.

- Hypothesis 2: Does duration of the flight have an effect on price of flights?

  H0 and H1 were clearly stated.

  One way ANOVA is used to address the following hypothesis:

  H0: Duration of the flight has no effect on price (i.e. no relationship).

  H1: Duration of the flight has an effect on flight pries(i.e. relationship).

  The plots meet the ANOVA assumptions as the variance/spread is the same throughout the plot, however it's also important to note the presence of outliers.

As for the normality , the plot meets the assumptions so ANOVA can be used.

p-value is less than 0.05 therefore its significant and we can conclude duration of flight has a relationship with price.

Looking the model summary , R2 is 35% which means that 35% of the variation of the model is explained by this model , which is fairly low, so interaction test was completed to look total stops and duration are dependent,t The interaction increased the model R2 to 56% which suggests an interaction occurs and variables are not independent.

- Hypothesis 3: Does number of stops have effect on price?

  T-test was used.

  H0: Median of price = median of total stops (i.e. no relationship).

  H1: Median of price != median of total stops (i.e. relationship).

  As per results , the p value is less than 0.05 so we can reject the null hypothesis and conclude that the number of total stops have an impact on the price of flights.

- Hypothesis 4: Does the type of airline have an impact on the price of the flight ticket?

  T-test was used.

  H0: Airline does not impact price of flights (i.e. no relationship).

  H1: Airline impacts price of flights (i.e. relationship).

  Distribution is symmetrical and normality assumptions are also met therefore we can interpret the p value from the t-test. p-value less than 0.05 so we can reject the null hypothesis and conclude airline and flight prices have a relationship.

  Second part of the experiment is to test if interaction occurs,

  Interaction plots represents the mean of the variables. From Looking at the plot , we can see that they're not parallel and they're heading in a direction where they will meet at some point which concludes that interaction occurs.

  Analysing the model summary report, we can see that the R2 is 69% when the variables airlines and class aren't treated as independent, we can conclude 69% of the variation in the model is explained by these variables.

  Assumptions of equal variance and normality are met , it's just outliers that is making it look uneven. Therefore , we can conduct an ANOVA test.

  H0: no interaction

  H1 : there is interaction

All p values are less than 0.05 so therefore are significant and we can reject the null hypothesis and conclude there is an interaction

- Hypothesis 5: Which variables can you remove to improve the flight prediction model?

  Based on the results I've achieved above ; we can conclude that the predictors have effect on price of flight as their p value is significant so it's not appropriate to remove a predictor for the reduced so interaction between class and airline is included in the full model.

  Therefore full models , include the predictors(airline, class, duration in minutes, journey month and interaction between class and airline . Whereas , for the reduced model is the full model without the interaction between airline and class.

  In the full model , assumptions of equal variance and normality have to be met before we can interpret the model summary. From the plots we can see that both assumptions are met however it's important to note that the outliers present make the variance look unequal.

  Looking at the model summary , R2 is 80% , therefore 80% of the variation is explained by the full model.

  Most of the p-value are significant except few certain airlines in both full and reduced model

  Assumptions of equal variance and normality have to be met in the reduced model before we can interpret the model summary.

  From the plots we can see that both assumptions are met however it's important to note that the outliers present make the variance look unequal.

  78% of the variation is explained by the model.

  All of the p-value are significant except journey of the month which tells us that the month the ticket is booked doesn't impact the price of the ticket.

  Log-likelihood decreased in the reduced model.

  its visible that the BIC a, AIC decreased in the reduced model but R2 increased in the reduced model.

  ANOVA was conducted to test the following hypothesis:

  H0: Full model is preferred.

  H1: reduced model is preferred. Based on the result of the p-value , we can reject the null hypothesis and conclude that the reduced model is preferred.

8 Data interpretation

Based on the statistical analysis that was completed I can conclude that the following variables have an impact on the price of flight tickets: type

of airlines , ticket class , duration of the flight , number of stops/layovers , journey month. This answers the research question that was stated in the beginning. There's always room for improvement in the current model tested which is listed in the next section.

9 Future work

After completing this project , there are room for improvement to increase the accuracy of flight price prediction model , these include:

- When gathering the data , future models should include the date the customer buy their flight ticket , this would give us a clear idea of how far in advance do you have to book your ticket to save money.
- Another point would be , to compare season prices so instead of just gathering prices for march-June . this would provide information if there's difference between winter and summer flight tickets.
- More work is being done in this field to use different machine learning techniques to improve forecast accuracy. The current approach can be further optimised with the current model.
- Find log of the numbers instead of just count for more accurate plots and statistical analysis.

10 Reflective journal

The purpose of this project is to improve the flight price prediction by revealing the variables that impact the price of flights. I've researched existing literature papers to learn more about the existing models and reveal their gaps that I can use in my research. Next step was to find a dataset that aligns with my research question . During this project , I've found the data cleaning and writing the report relatively simple compared to figuring out which statistic test appropriate to use for the hypothesis I've listed above. I have learned from this project to definitely form your research question and read the literature papers before deciding on a dataset. I've also learned that to not hesitate to transform the variables to make it more suitable for analysis, for example , a new column was created, and it consisted of log of price column price.

11 Reference

- Kimbhaune, V., Donga, H., Trivedi, A., Mahajan, S., Mahajan, V. (2021). Flight Fare Prediction System (No. 5542). EasyChair.
- Kuhn, Nathalie and Navaneeth Jamadagni. "Application of Machine Learning Algorithms to Predict Flight Arrival Delays.

- William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.
- K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine larning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras" Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. pp. 1-6 Oct. 2009.
- A. Bhatia, V. Kedia, A. Shroff, M. Kumar, B. K. Shah and Aryan, "Fake Currency Detection with Machine Learning Algorithm and Image Processing," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 755-760, doi: 10.1109/ICICCS51141.2021.9432274