# *EaseMyTrip.com* Data Prediction Final Report

Capstone Project 2

**Adam Reichenbach**

June 2024 Cohort
Springboard Data Science Career Track

## Introduction

Booking tickets for a flight can be a hassle whether getting tickets directly from the airline's website or through another 3rd party site. One such site is ***EaseMyTrip.com*** that compares different prices for flights in India. There are many different factors that can go into creating a price for a listed trip whether that is which class of ticket you buy, the airline the ticket is from, and which city you are going to. My project aims to create a regression machine learning model that can help predict ticket prices based on the different features.

## Data Wrangling

- The data wrangling process can be seen here.
- The data for the project was available as a downloadable CSV file from kaggle.com. The data provided was already mostly cleaned and contained no missing values.
- After getting rid of the 'Unnamed: 0' column which matches the index, I checked to be sure that none of the rows had any missing data and there was still none.
- I found that there were over 300,000 instances with 11 features to work with where all but 3 of the features were categorical and those remaining 3 being continuous numerical features: *duration, days_left,* and the target feature *price*.
- I dropped the *flight* feature as it was the flight number of each instance. It was a categorical feature with over 1,500 unique values, but didn't seem to bear any relevance to the price as each airline has its own unique system on assigning flight numbers.

## Exploratory Data Analysis

- After importing the data I checked for duplicate instances and then dropped them accordingly which brought down the data to 297,940 instances, so only 0.7% of the data was dropped.
- I made a couple of functions that would remove outliers from data of certain categorical features.
    - First function removed outliers that were outside 1.5*IQR of the data. This was for any data that may be skewed.
    - Second function removed outliers that were outside 3 standard deviations of the data. This was for any data that was more normally distributed.
- The following questions I looked to answer during this step of the project:
    - Does the ticket price vary by airline?
    - Do the cities have any impact?
    - How does the ticket price vary between economy and business class?
    - Does departure and arrival time affect ticket price?
    - How are the prices affected when bookings are made 1 or 2 days before departure?

## Does ticket price vary by airline?

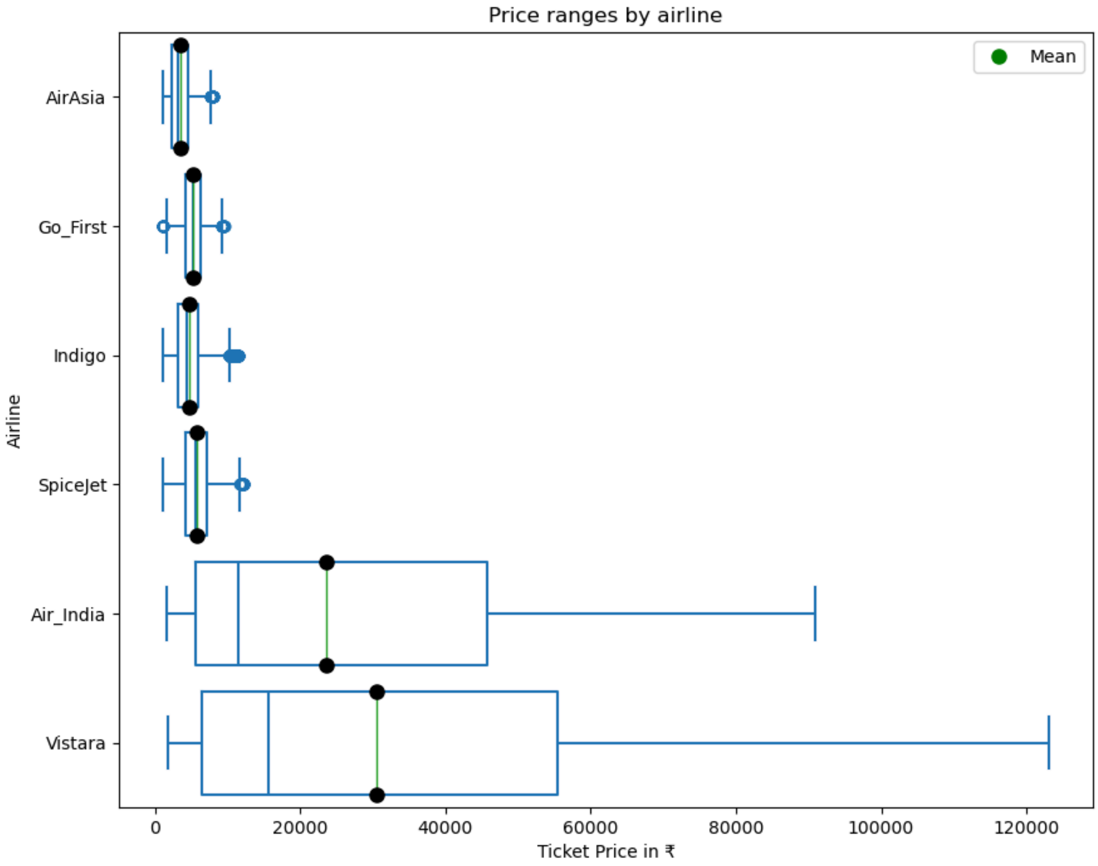

**Fig. 1**

After removing the outliers with the IQR method and plotting boxplots of the prices sorted by the different airlines, we see that Air India and Vistara have the biggest price variations, highest means, medians and max values.

The remaining 4 airlines have pretty similar price ranges as well as means and medians. AirAsia generally has the lowest mean, median, and max prices.

## Do cities have any impact?

I listed down where each of the airline's respective HQ/main hub is located as well as which airline each city serves as a hub for. I've noticed that neither Kolkata or Chennai serve as a major hub for any of the airlines featured in the data. I suspected that this could impact ticket pricing.

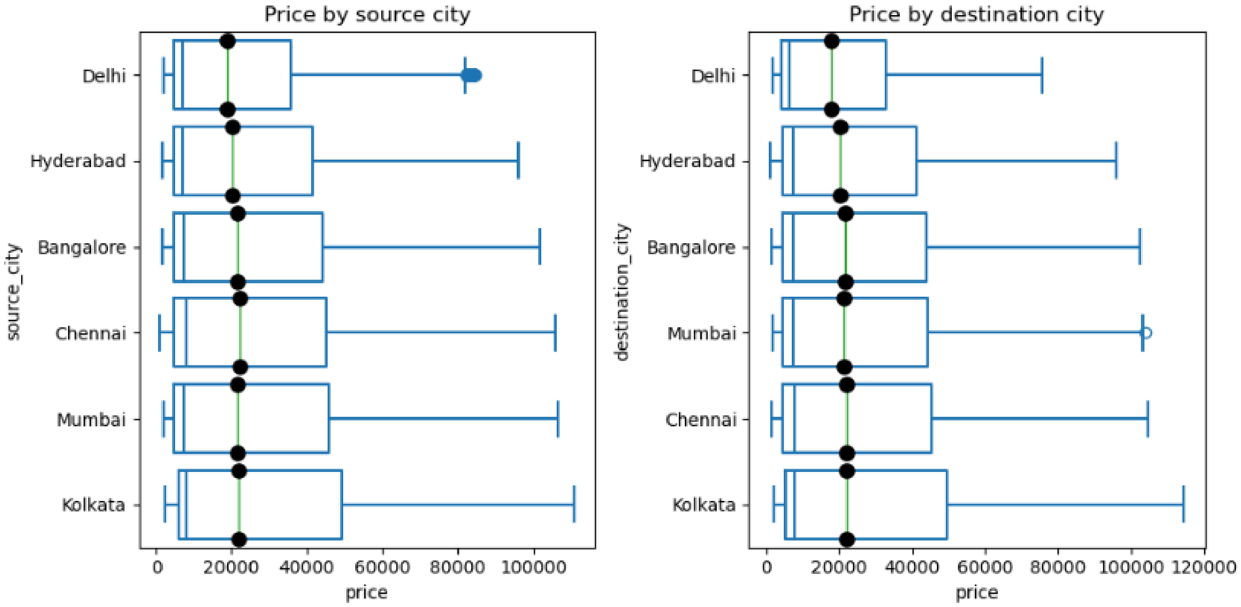

**Fig. 2**

After sorting the data into prices by city and removing the outliers I noticed that all the airlines have very similar mean and median prices, but Delhi had the lowest price range variation as well as the lowest max price for both as a source and destination city. Kolkata has the biggest variance as well as the highest max price. This seems to reinforce the idea that since Kolkata isn't a major hub for any of the featured airlines, that it will have higher max prices since there aren't as many available flights serving the city.

I explored the question if city population had any sort of affect on the prices. After gathering looking up what the estimated populations of each city would be around February 2022, I created

5 bar plots to see if there was any noticeable correlation between the populations, max prices by source & destination city as well as the demand count for the source and destination city.

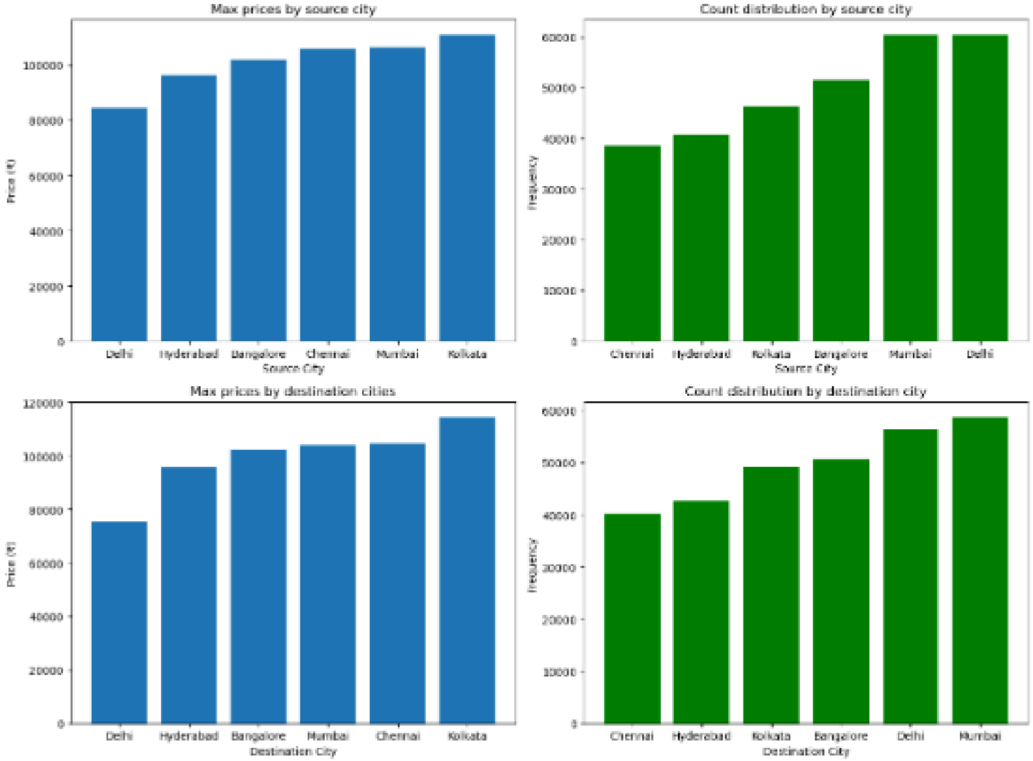

**Fig. 3**

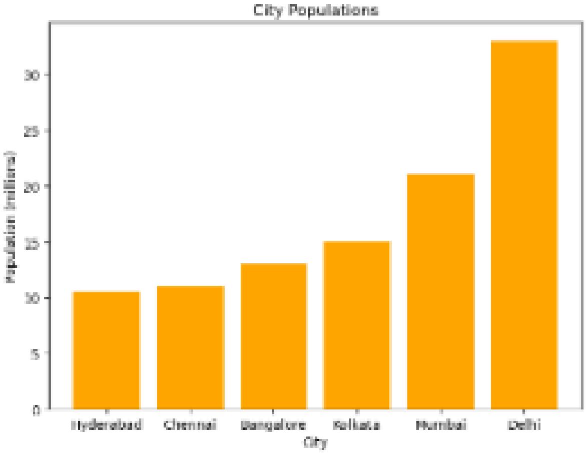

**Fig. 4**

As seen in **fig. 4**, Delhi has the highest population of the featured cities. It is the highest population of India and it serves as the country's capital. Looking at **fig. 3** it seems that there isn't really any noticeable correlation between the city population with either the demand count or ticket pricing as a whole. There may be some inverse correlation between Delhi's population and demand count with having the lowest ticket price range and max price. This could be due to the fact that the city has high political and administrative importance to the country that there are just more available flights for customers and thus lower prices in general compared to other cities in the country.

## How do the prices vary by economy vs business class?

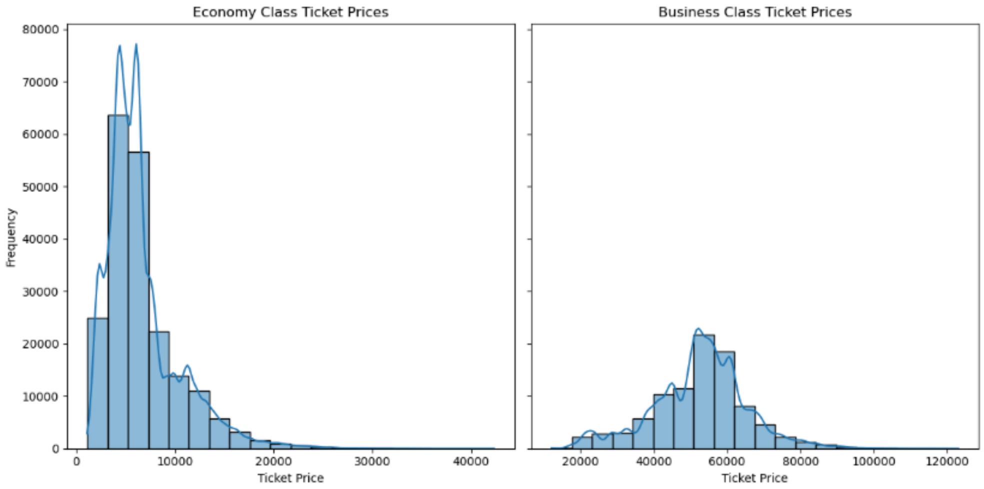

**Fig. 5**

Plotting the distribution of the economy and business class ticket prices (**fig. 5**), it looks as if there is a left-skewed distribution of economy class tickets that can reach up to about 40,000 Indian Rupees whereas the business class tickets are more normally distributed just centering roughly around 60,000 rupees. Business class generally has much higher prices and is likely why there is much more demand for the economy class tickets.

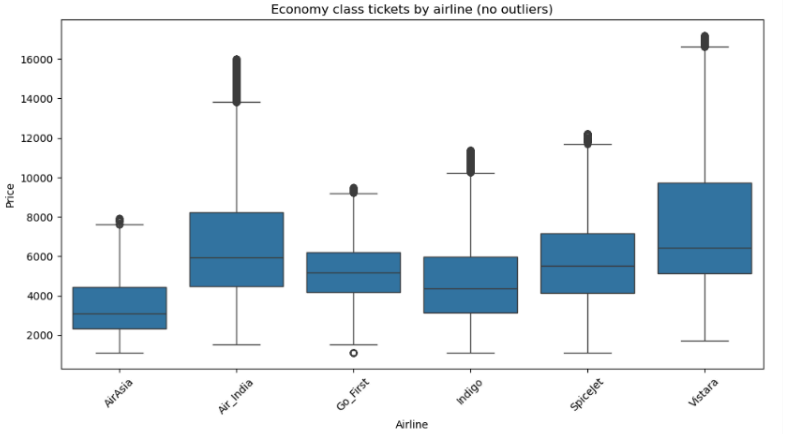

**Fig. 6**

I looked at exploring the price ranges of economy class tickets by airline (without outliers, see **fig. 6**) and unsurprisingly Vistara and Air India have the largest spreads as well as the highest max prices with the highest price reaching 16,000 rupees. AirAsia has the smallest spread and max price.

## Does departure and arrival time have any effect on the prices?

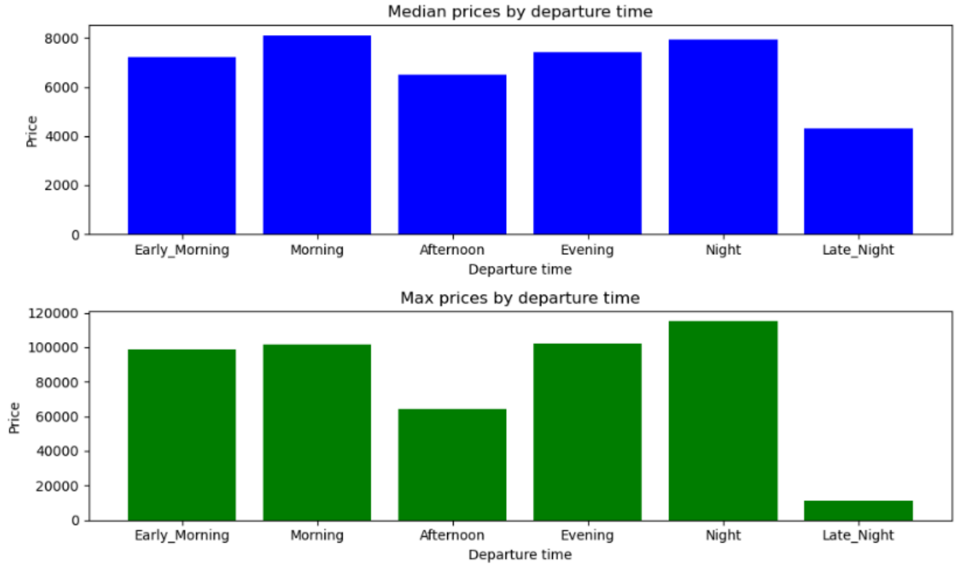

**Fig. 7**

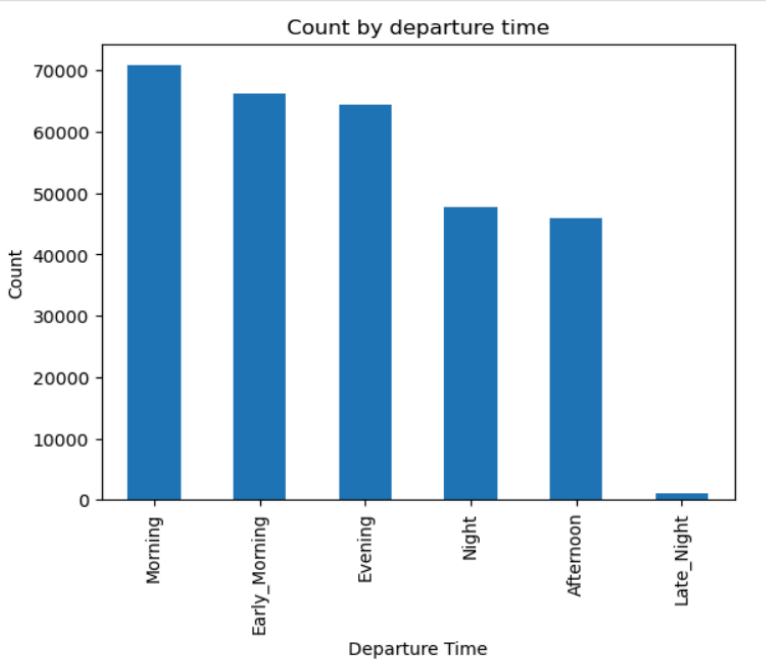

**Fig. 8**

Plotting the median and max prices of the tickets by departure time (**fig. 7**), there does seem to be a strong correlation between the time of day with the prices. The afternoon and late night departure times have the lowest prices, while night and morning prices seem to be the highest. The trend does seem to match up a bit with the demand (**fig. 8**) for tickets at those particular times. It would make sense that more people would want to pay for flights early in the day so that they would want to get to their destination with plenty of time to spare during the day. High evening count and prices could be due to passengers wanting to fly at the end of their day, but not getting to their destination at a super late time and trying to avoid any overnight/red-eye flights.

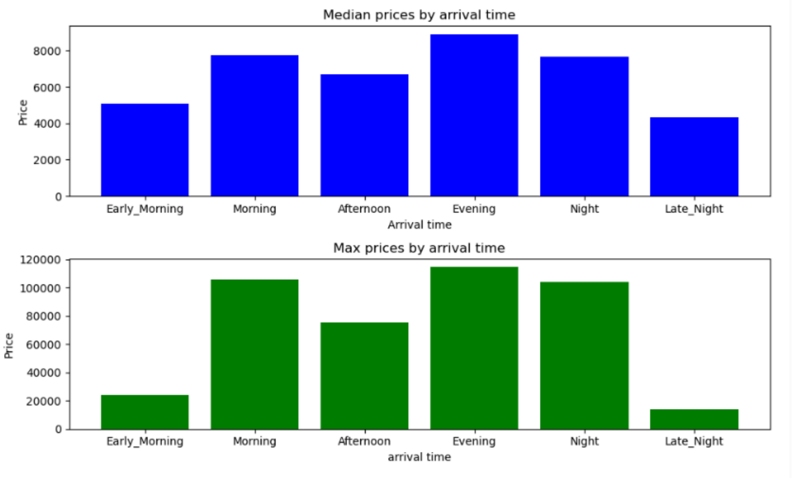

**Fig. 9**

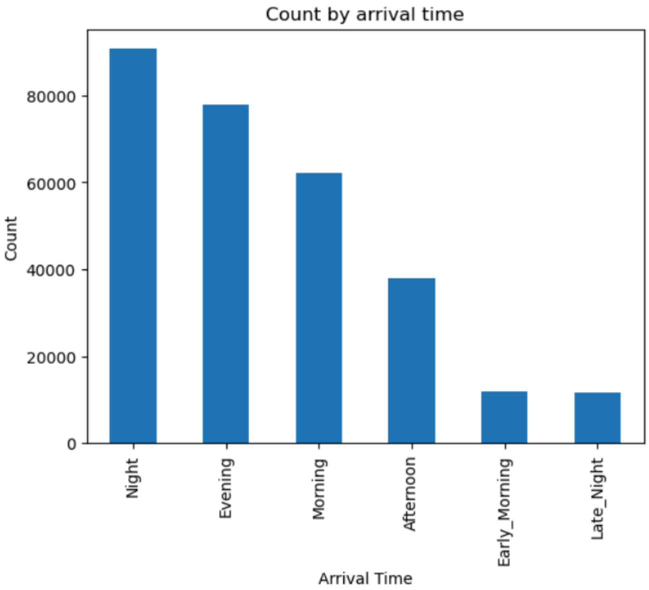

**Fig. 10**

The distribution (**fig. 9**) shows that morning, evening, and night have the highest prices and demand count (**fig. 10**). Late night and early morning have the lowest numbers of each plot. Once again the low demand and thus prices for late night and early morning is likely due to passengers wanting to avoid red-eye flights. The high demand for morning flights as an arrival time is probably due to the high demand from the early morning departure time as passengers likely want to get in at a good time when the day is fairly young and have time to take care of other business during the day. Evening and night high demand/prices could be due to passengers flying as part of the last part of their day while avoiding overnight/red-eye flights.

In general it seems that the highest demand for tickets and thus prices are for flights that take off and land during the morning as well as during the evening and night. Lower prices tend to center around the overnight/red-eye flights.

### How are prices affected when the tickets are brought 1 or 2 days before departure?

I explored what the prices looked like when bought 1-2 days before departure.

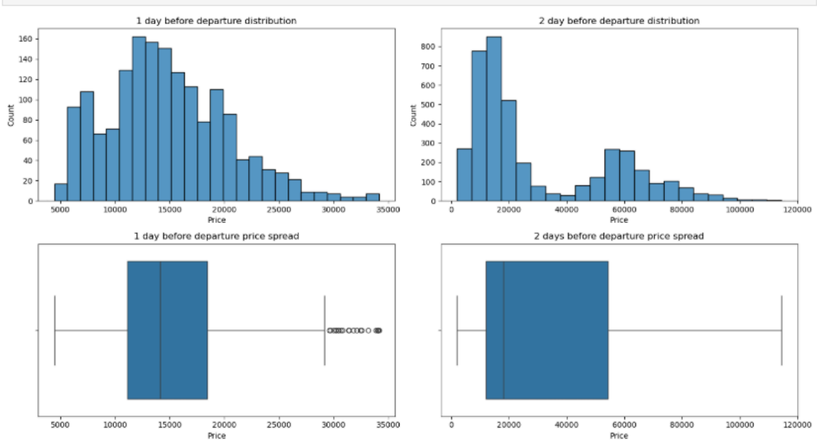

**Fig. 11**

After removing outliers, it looks from **fig. 11** that there seems to be a lower price spread and median for tickets bought the day before compared to buying 2 days before.

## Pre-processing and training

The machine learning algorithm for finding our continuous target variable requires that the other features be numeric and thus the categorical features need to be encoded. I decided to manually encode the categorical features as there aren't very many values of each categorical feature. This is also to get a correlation heatmap to decide which features should be focused on for the model.

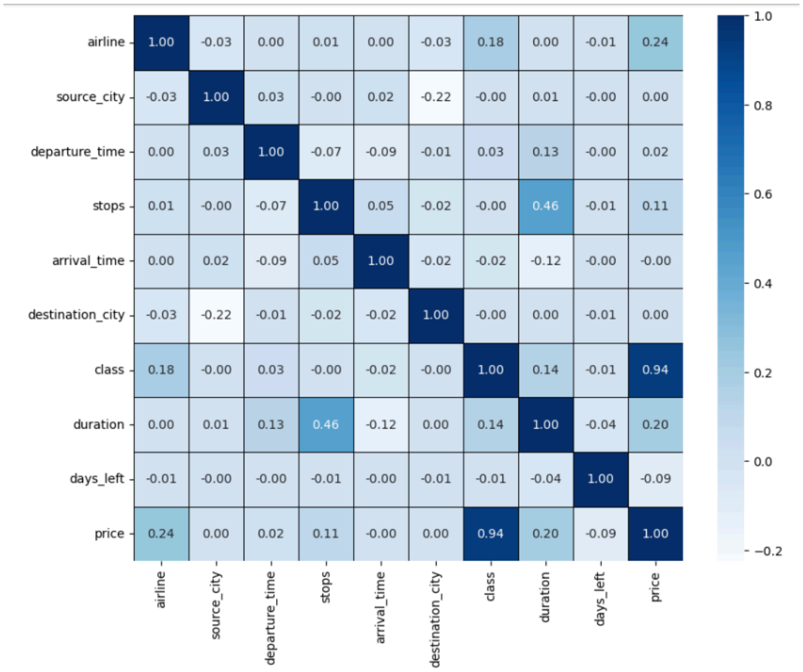

**Fig. 11**

The heatmap shows that the flight class has the biggest correlation with ticket prices. Next highest is airline and this is likely due to Vistara and Air India having the highest overall prices while the other 4 airlines are very close to each other as seen in the EDA step.

Final step for the preprocessing was splitting the data randomly into 80% for training, 10% for testing, and 10% for validation.

## Modeling

The 3 models that I decided to try out were linear regression, lasso regression, and gradient boosting.

### Linear Regression

For linear regression I decided that the ticket class feature was the best one to use and see what other features would work best with it.

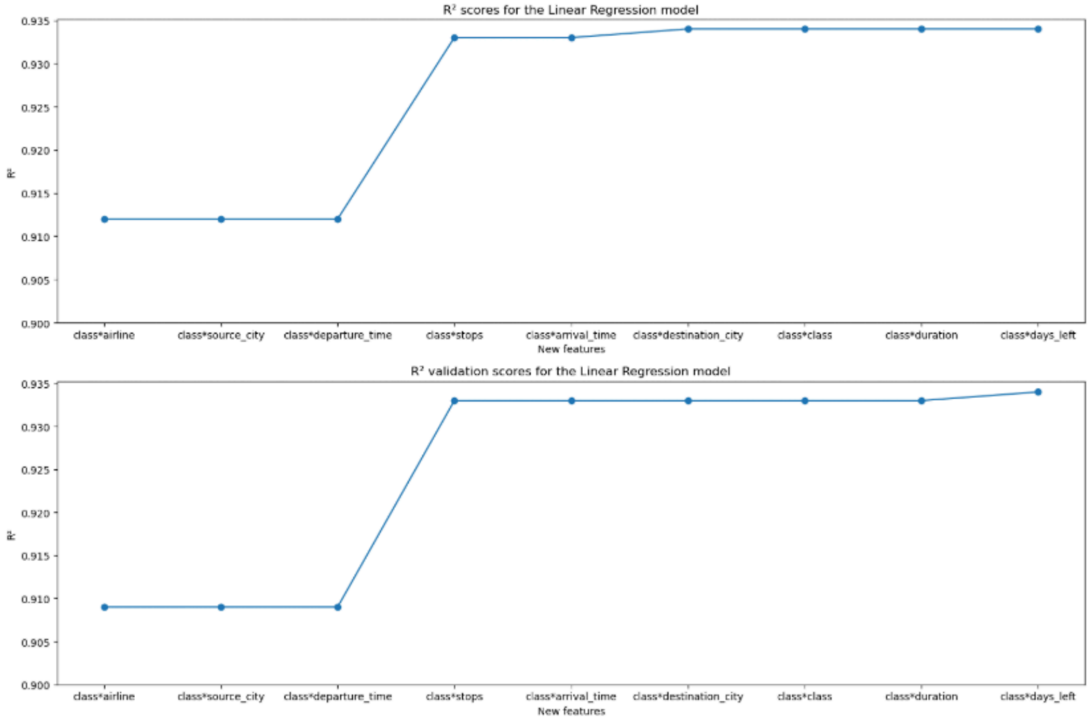

**Fig. 12**

After plotting the $R^2$s it looks as if days left and destination city with class have the highest scores at 0.934 each.

### Lasso Regression

Training the data with a lasso regression I found that the best parameter for the model was for the alpha to be 0.001. The $R^2$ for both the training and validation data came out to 0.934 with similar RMSE (5812.9 for training data, 5846.4 for validation). The $R^2$ looks to be the same as the linear regression model, so either model could be good final choices.

### Gradient Boosting

After training and fitting the data for a gradient boosting model the $R^2$s and RMSE come out to 0.956, 4761.8 and 0.955, 4807.5 for the testing and validation data respectively. So far this model has the best $R^2$ and RMSE of all the models tried.

### Model of choice

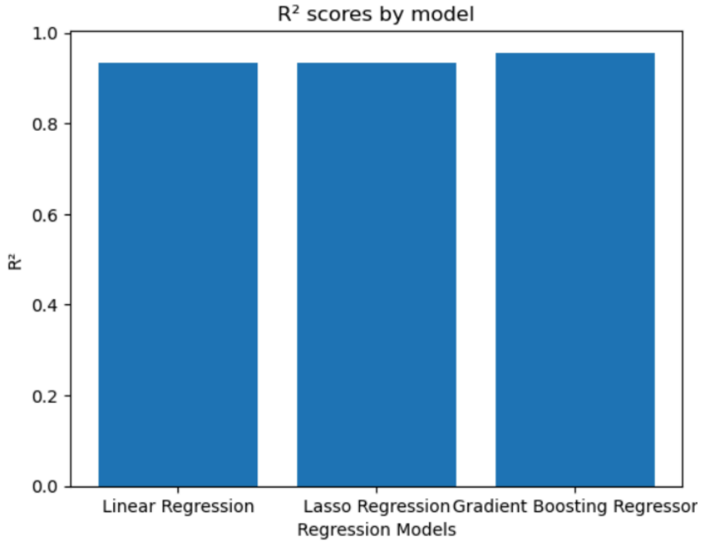

**Fig. 13**

As seen here in **fig. 13** it looks like the gradient boosting model just narrowly beats out the other two models. It will be our model of choice to use for predicting ticket prices.