Airline Ticket Price Prediction

Capstone Two

Morgan Snellgrove

Problem Statement

The goal of this project was to use Flight Price Prediction dataset to predict the price of airline tickets for various airlines in India. The dataset contains 300,261 observations with the following features:

1. Date of Flight
2. Name of Airline Company
3. The Tail-number of Plane used for Flight
4. Name of Source City
5. Name of Destination City
6. Departure Time
7. Arrival Time
8. Number of Stops
9. Ticket Class
10. Duration of Flight
11. Number days in advance ticket was purchased
12. Price of ticket (target)

Data Wrangling

At the link above we are provided with three different csv files: Clean_dataset.csv, business.csv, and economy.csv. Under the discussion tab, I found the notebook where the author shows his steps for creating Clean_dataset. I went through his steps and decided to make some changes. I started with the business and economy csv files. I followed his steps except for the following changes:

- I kept the date column
- I did not bin departure or arrival times
- I kept stops as a numeric value not a string
- I kept duration as two columns: duration hours and duration minutes
- I kept 108 observations that he dropped
- I corrected an error in handling duration that resulted in extremely high values
- I converted the price to USD and saved this as a new column named price_usd.

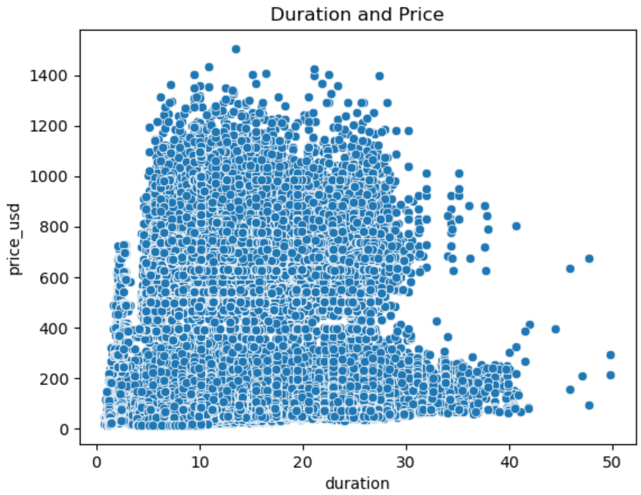
After these steps, the resulting dataset is saved as final_clean_flight_dataset.csv.

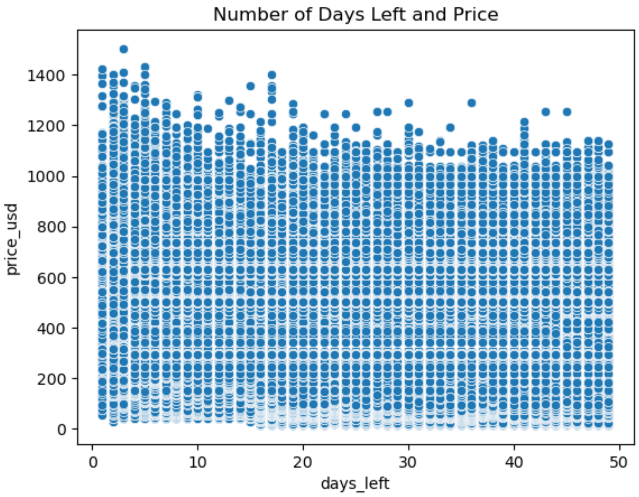
Exploratory Data Analysis

I had several questions that I wanted to explore. The main ones were:

1. Do any of the numeric variables have correlation with 'price_usd'?

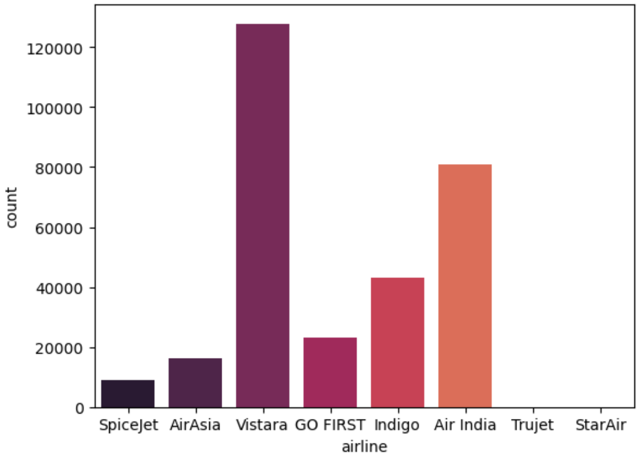
I looked at the pairplot for the dataset, and found that it was difficult to see correlation due to the number of observations. However, you could see that flights with longer durations tended to have lower prices, and tickets that were purchased further in advance also had lower price. One thing that surprised me was in the stop column. I expected non-stop flights to be more expensive, but it was the exact opposite.


Duration and Price

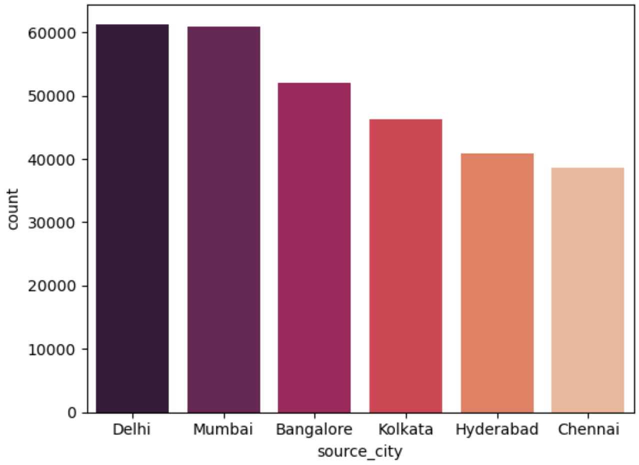

Number of Days Left and Price

2. For each categorical variable, what is the most common/popular value?

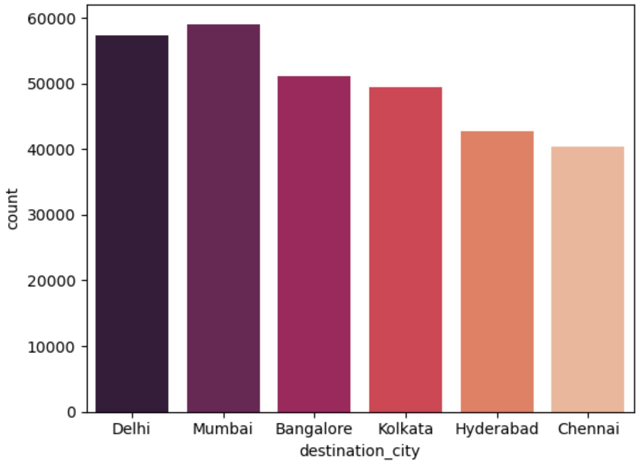
- Most popular airline: Vistara



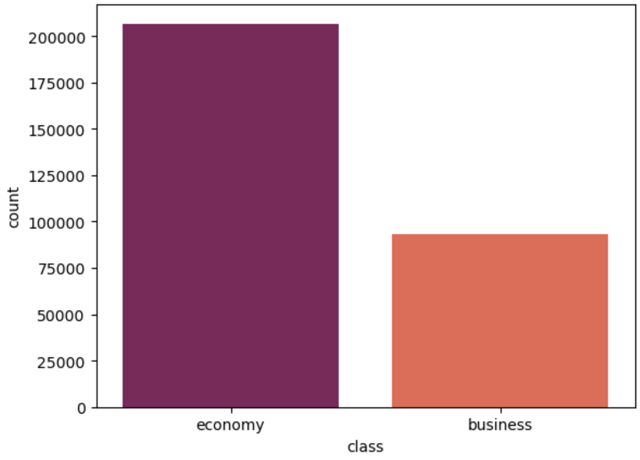
- Most popular source city: Delhi



- Most popular destination city: Mumbai

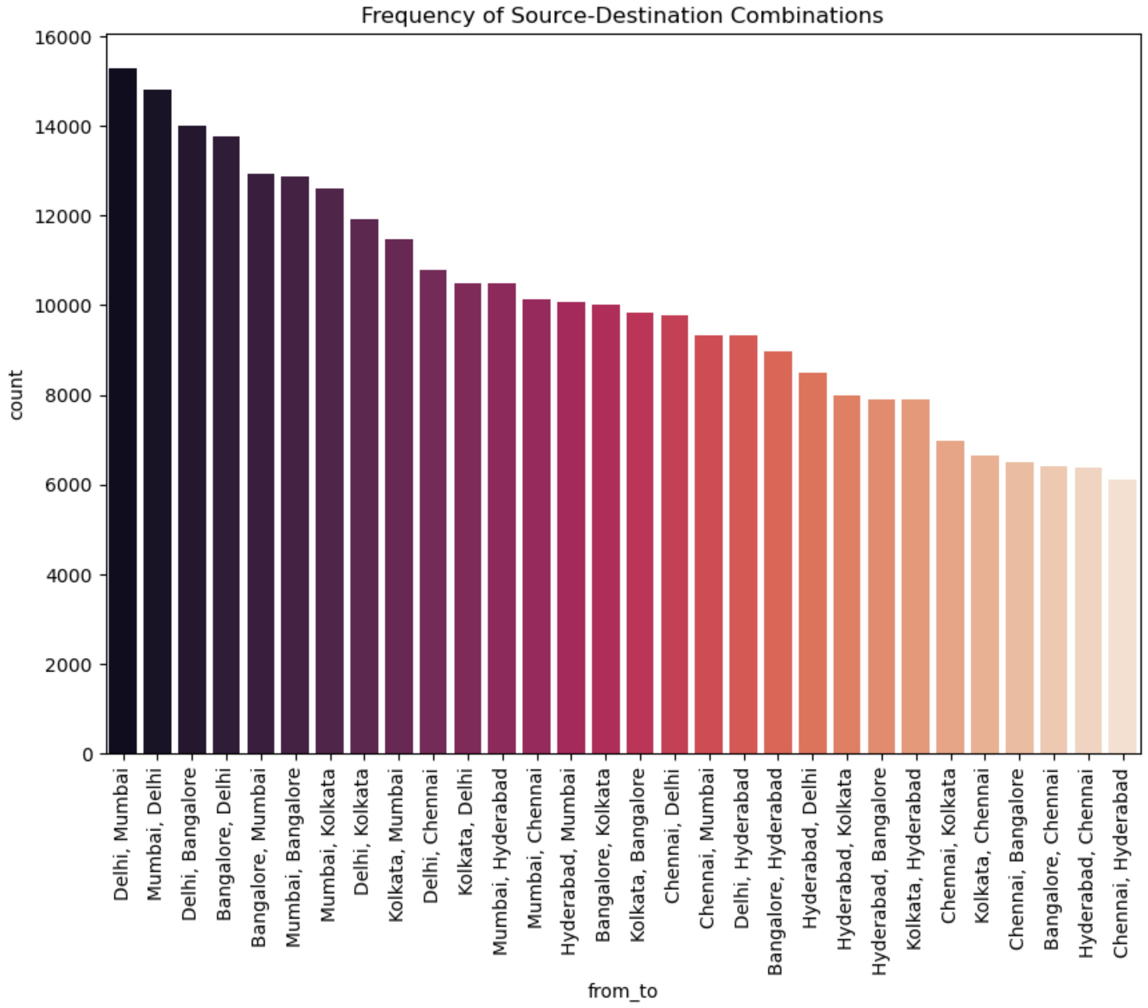


- Most popular ticket class: economy



3. What source/destination combos are most popular?

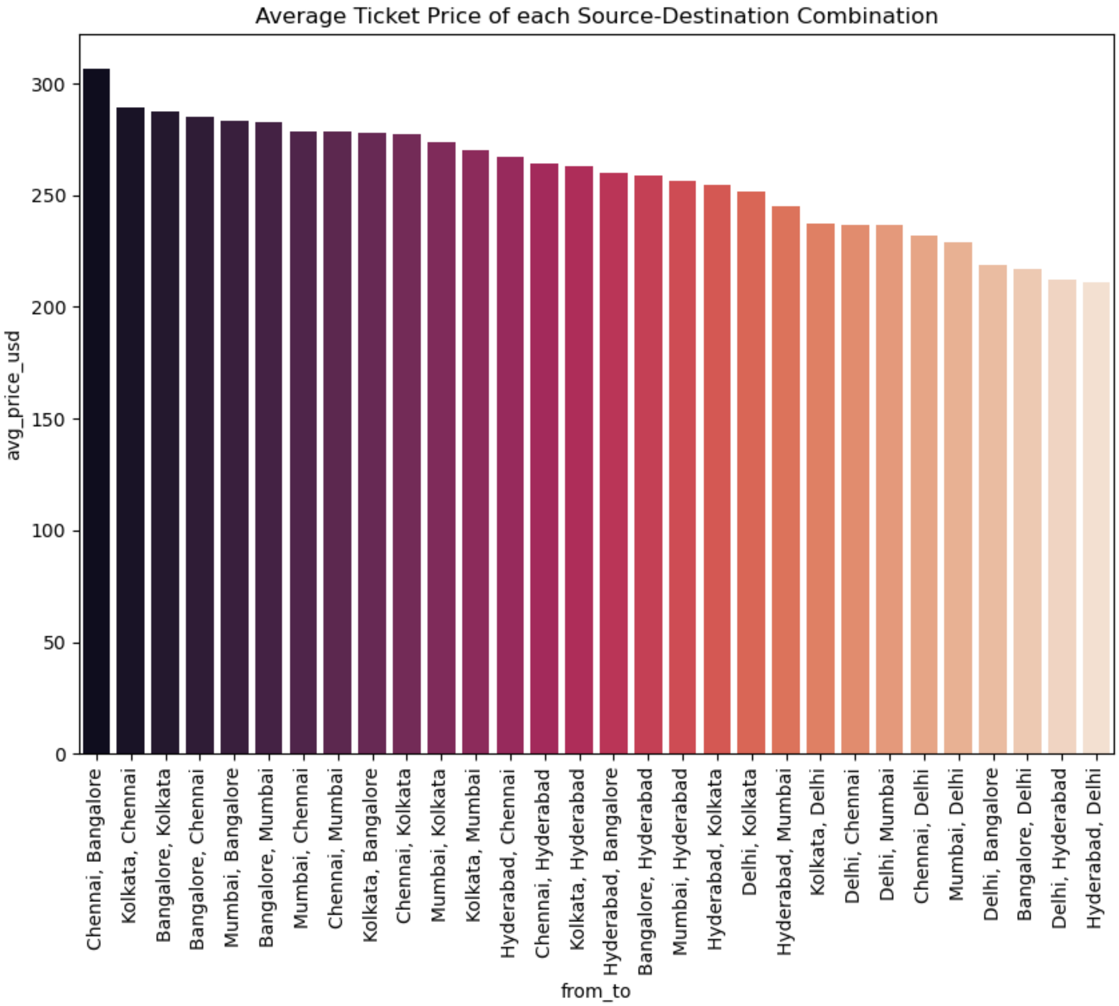
To look at the combinations of source and destination cities I created a new column that zipped those two columns together and then used value counts. I found that most passengers were flying between Delhi and Mumbai, Delhi and Bangalore, and Mumbai and Bangalore.



4. What is the average price for each source/destination combo?

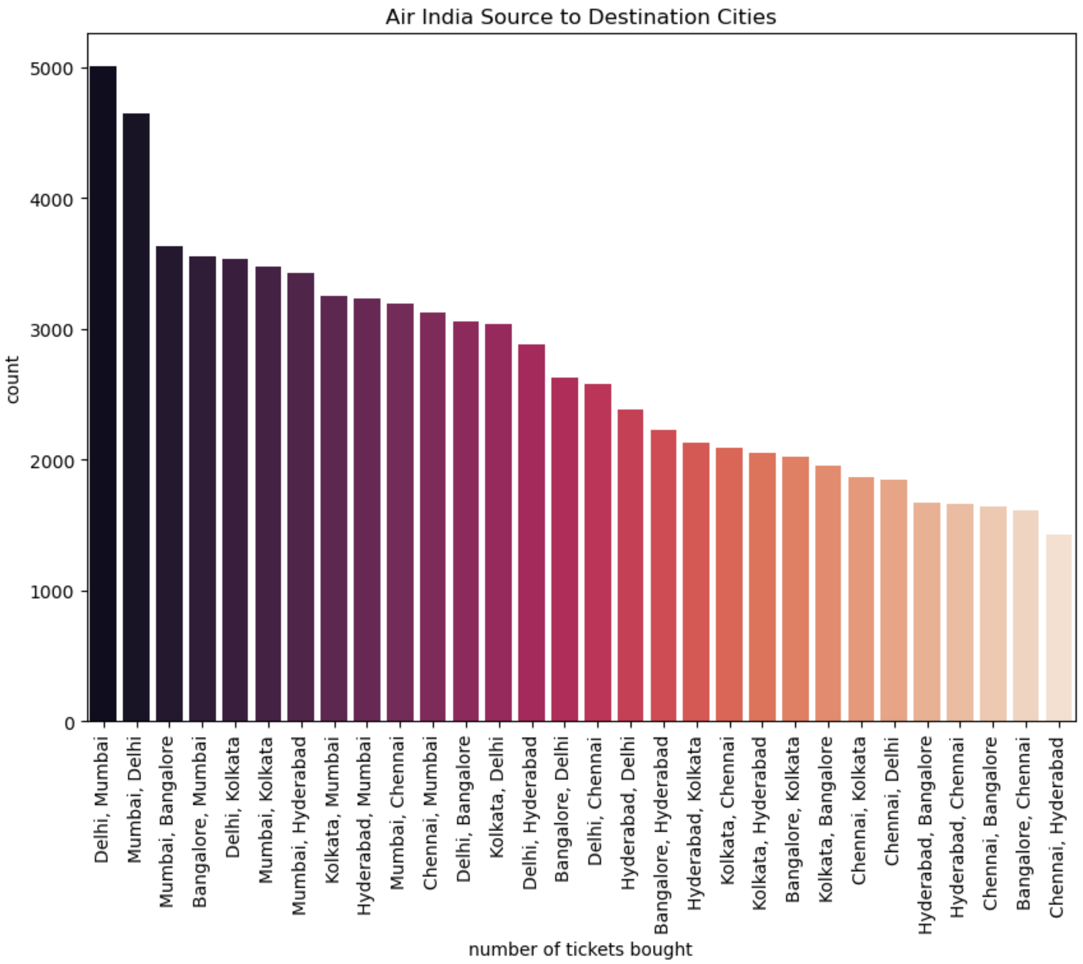
Next, I wondered if these popular trips would tend to have cheaper tickets. To find out I grouped by the from_to zipped column of source and destination cities, then calculated average ticket prices. The cheapest average ticket price was $210.85 for flights between Delhi and Hyderabad. The most expensive tickets were $306.71 for flights from Chennai to Bangalore. The popular from-to combos mentioned above had average ticket prices of:

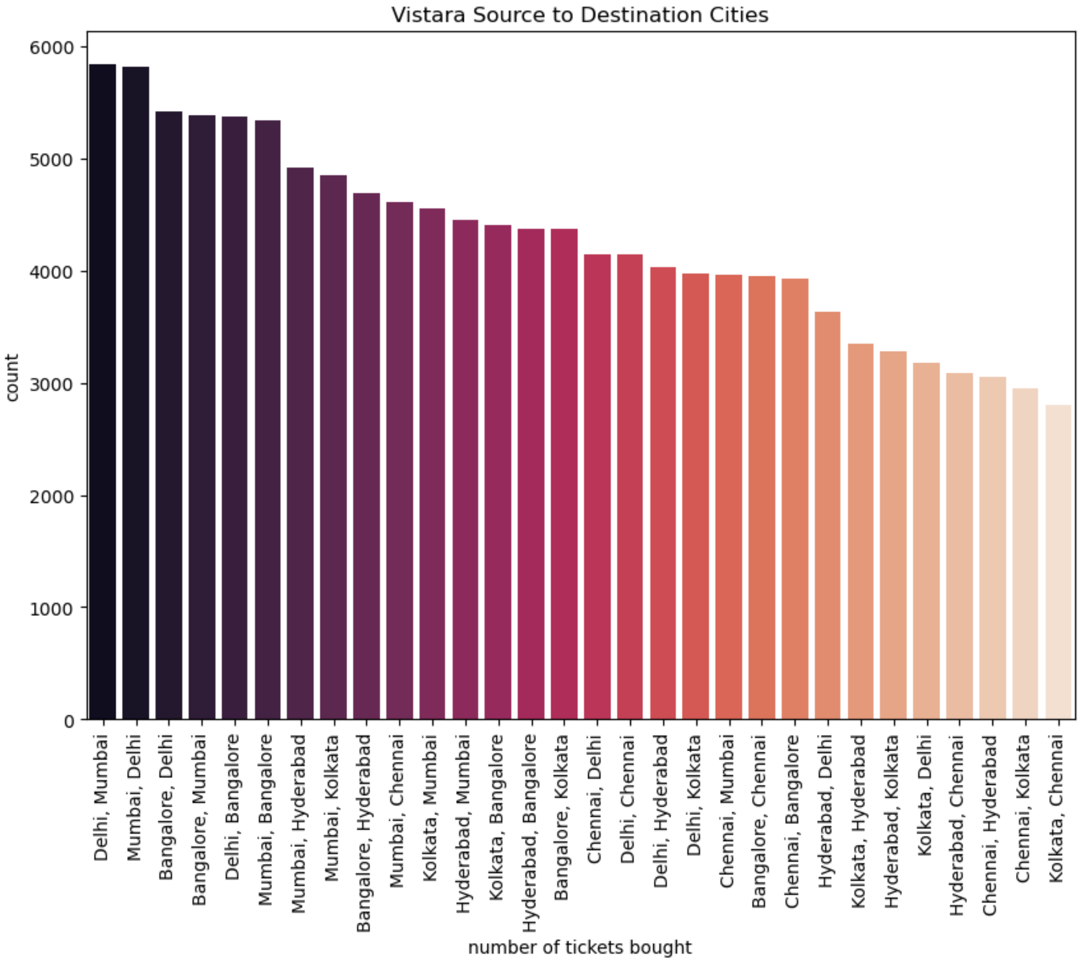
Delhi to Mumbai	$236.67
Mumbai to Delhi	$228.98
Delhi to Bangalore	$218.65
Bangalore to Delhi	$216.73
Mumbai to Bangalore	$283.06
Bangalore to Mumbai	$282.81

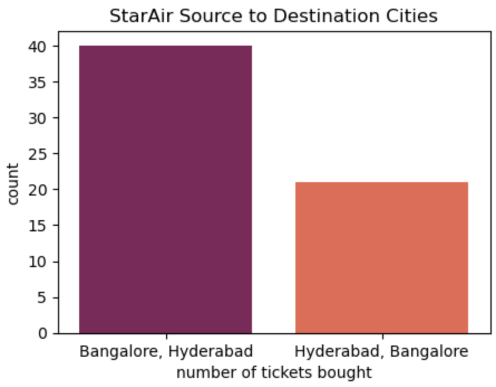

Average Ticket Price of each Source-Destination Combination

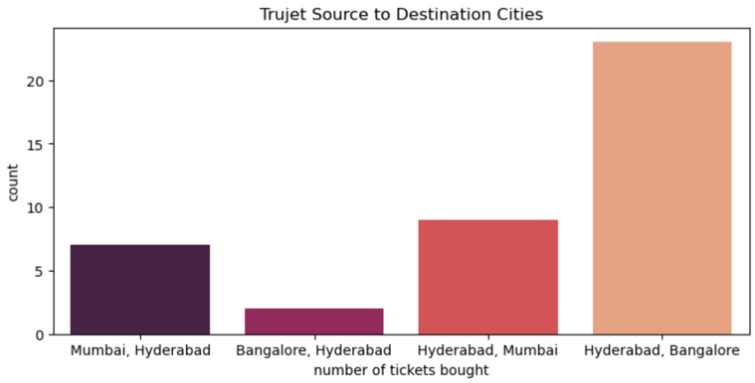
5. For each airline, what source/destination combos do they offer?

The most popular airline, Vistara, offered the most destinations. Air India and Spice Jet also flew to almost every destination. On the other end of the spectrum, TruJet and Star Air had the fewest options. TruJet only flies between Bangalore, Hyderabad, and Mumbai and Star Air only flies between Hyderabad and Banglore.


Air India Source to Destination Cities

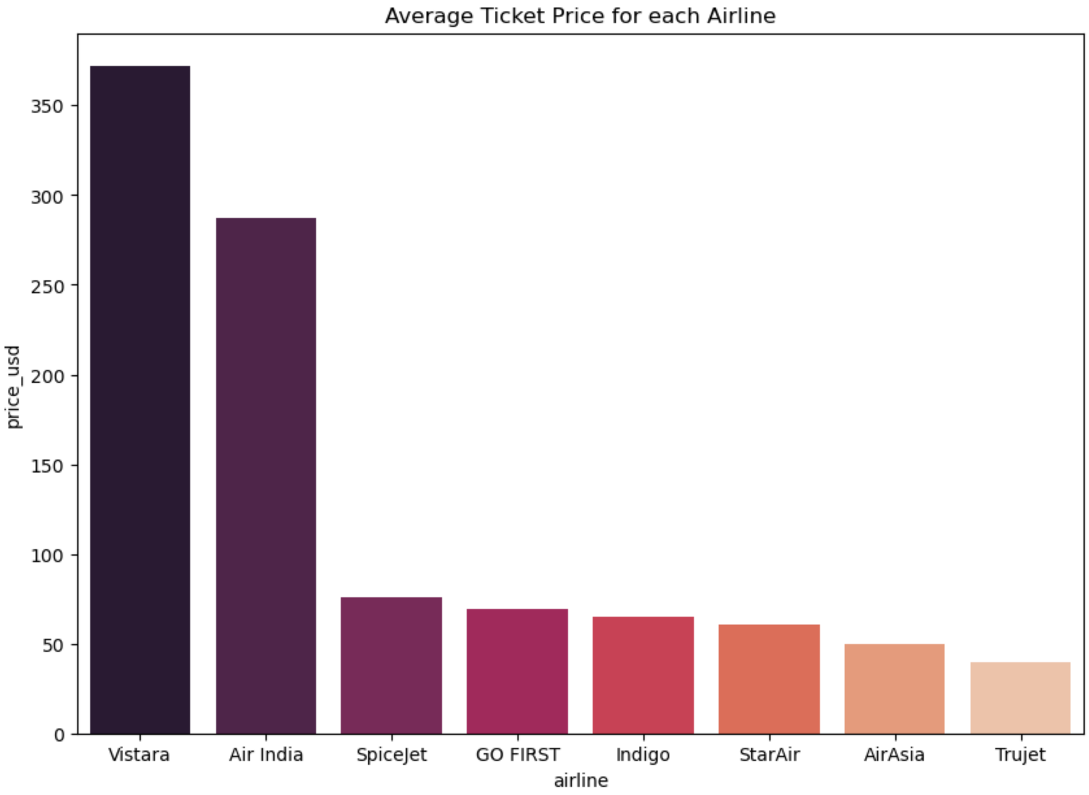

Vistara Source to Destination Cities





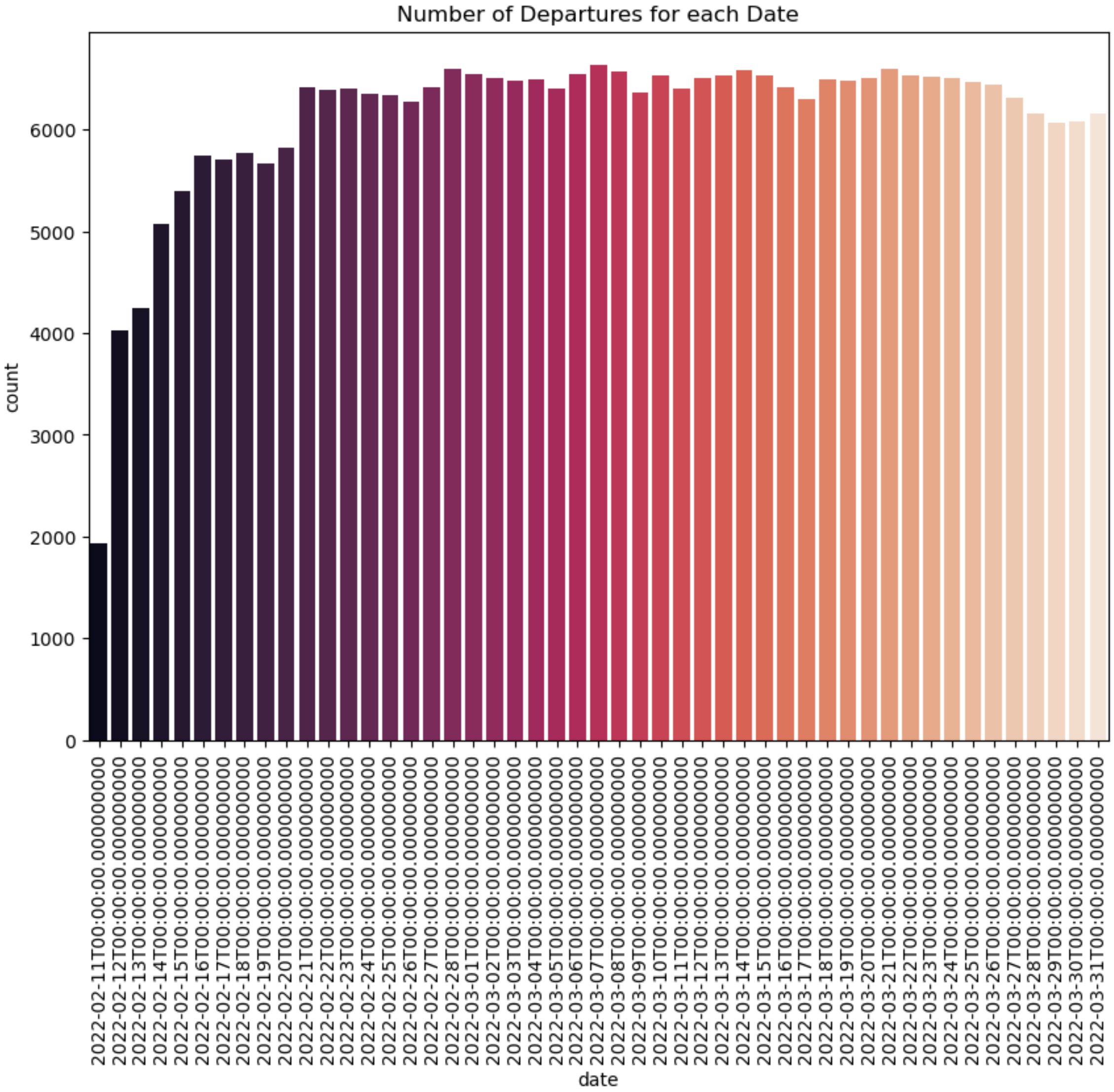
6. Do all airlines have both economy and business tickets?

I looked at average ticket price grouped by airline and saw a large jump between the top two most expensive and the rest of the airlines. I wondered if this came down to ticket class. The top two most expensive airlines, Vistara and Air India, were the only airlines that offered business class tickets.



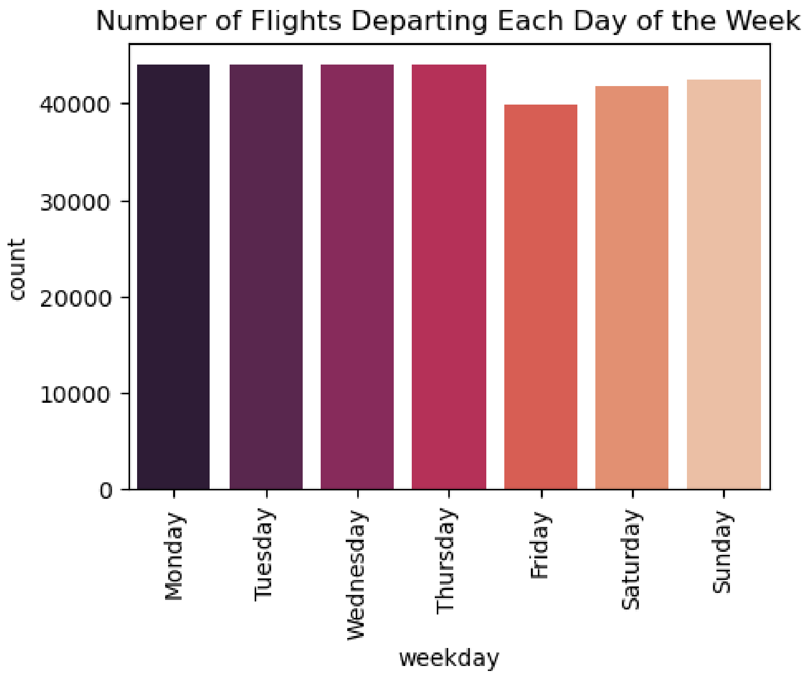
7. Are the dates evenly distributed? Are there any holidays during the time period represented?

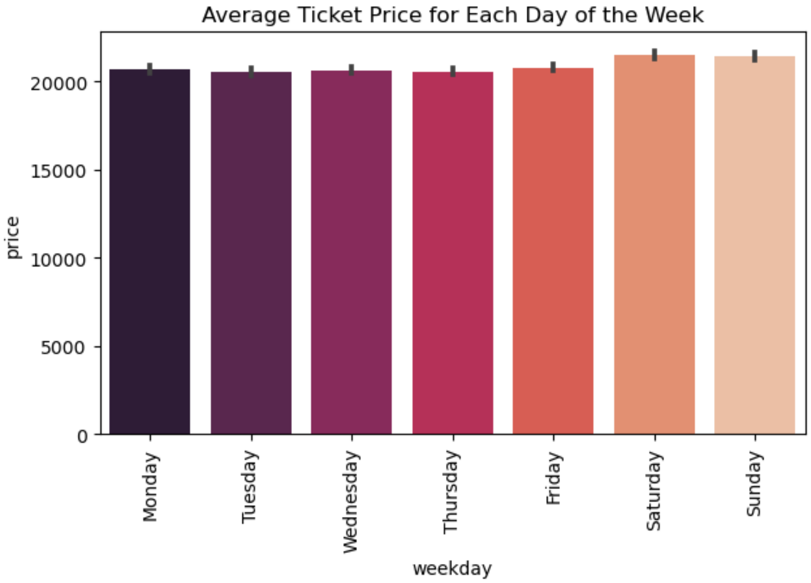
Looking at the min date and max date, I found that all of the observations in this dataset ran from Feb 11, 2022 to March 11, 2022. A quick Google search told me there were not any major holidays in India during that time. The number of departures per date were close to uniformly distributed starting Feb, 21. The departures were lower for Feb 11 -20.



8. Are ticket prices different on Weekdays vs. Weekends?

I extracted the name of the day of the week that the flight was to depart. Looking at the counts for each day, I found that the most popular day to fly was Monday. The least popular day was Friday. I also looked average ticket price grouped by day of the week. The most expensive day to fly was Saturday. The least expensive was Tuesday.





Pre-processing and Training Data Development

To prepare the data for modeling I followed the following steps:

1. I extracted the information from the date column by creating columns for the month and the day of the month. Similarly, I extracted the hour and minute from departure time and arrival time. This created 6 new columns. I then dropped the original columns.
2. I separated out the categorical features and encoded them using Pandas' get dummies. Then I dropped the original columns and combined the new binary columns with the rest of the dataset. Since we have some categorical features with a lot of categories, this increased the data frame size to 300,261 by 1,604.
3. Next, I defined the target (price_usd) and feature data frames. I passed these to train_test_split with the default training set size of 75%.
4. Lastly, I fitted the Standard Scaler to the training features, and transform the training and testing features.

Modeling

Dummy Regressor

As a baseline for comparison, I started by building a Dummy Regressor. This first "model" just guesses the mean value of the target for every observation. Using this to make predictions we get the following metrics for the test set:

• R-Squared	-4.69
• Mean Absolute Error	$241.66
• Root Mean Squared Error	$277.54

Linear Regression Model (OLS)

Next I built a pipeline with the first step being SelectKBest(f_regression) and second step LinearRegression. Fitting our training data and then making predictions on the test set, we got the following scores:

• R-Squared	0.91
• Mean Absolute Error	$55.64
• Root Mean Squared Error	$84.90

Lasso and Ridge Models

For Lasso I tested alpha = 1, 5, 10, and 20. I generated the R-Squared for each. The best R-Squared was 0.92 for alpha=1.

For Ridge I tested alpha = 10, 50, 100, and 1000. The best R-Squared for this model was 0.9248 with alpha=100.

Random Forest Model

Next, I tested out a random forest model. I decided to use the default settings to see how it performed. I fitted the model using the training data and ran predictions on the test set. This was the slowest model as running this code took over 12 minutes. The test set scores were:

• R-Squared	0.9907
• Mean Absolute Error	$10.02
• Root Mean Squared Error	$26.69

XGBoost Models

I made both tree based and linear based XGBoost models. Training these models was much faster than the random forest model. The scores for the tree based XGBoost model were:

• R-Squared	0.9774
• Mean Absolute Error	$24.25
• Root Mean Squared Error	$41.68

The scores for the linear based XGBoost model were:

• R-Squared	0.9245
• Mean Absolute Error	$51.41
• Root Mean Squared Error	$76.24

Selecting the Best Model

Examining these results, the best model is the random forest model. Out-of-the-box settings provided excellent results. The model decreased the mean absolute error by more than 95% and the root mean squared error by more than 90% from our dummy regressor.

Most Important Features

I also decided to look at what features played the most significant role in the random forest model. The top 5 features were:

- Class Economy
- Duration
- Days Left
- Airline Vistara
- Arrival Hour