Final Project

Lauren Walker and Shanaya Piramal

Introduction

Project Motivation and Revelance

The data set we selected for our final project is the "Flights" data set. We got this data set from Kaggle, and it contains 3000153 unique observations, each representing a flight from one of India's 6 metro cities. In order to collect the data, researchers made use of the Octoparse scraping tool to extract data from the website "Ease My Trip". Data was collected in two parts: one for economy class tickets and another for business class tickets. The data was collected for 50 days, from February 11th to March 31st, 2022.

While Covid-19 took a toll on a lot of industries, one of the industries that it took a major toll on was the Airline industry. As a result, flight prices have significantly increased (https://www.nerdwallet.com/article/travel/travel-price-tracker). Both of us, have had a passion for travelling and seeing new places. However, as college students, we understand that travelling is a big cost. Hence, exploring how to strategically book tickets, depending on the number of stops, arrival/destination time, and airline is a topic that interested us.

As travel becomes a significant part of our lives again, this project has practical applications for individuals, businesses, and policymakers and can contribute to a better understanding of air travel pricing dynamics.

Data

Our data has 300153 observations and 12 variables. Each row in our dataset represents a particular flight. Each of the variables. The variables are explained below:

Table 1: Variables Explained

Variable	Explanation
Airline	The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
Flight	Flight stores information regarding the plane's flight code.
Source City	City from which the flight takes off.
Departure Time	This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.
Stops	A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
Arrival Time	This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
Dest City	City where the flight will land. It is a categorical feature having 6 unique cities.
Class	A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
Duration	A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
Days Left	This is a derived characteristic that is calculated by subtracting the trip date by the booking date.
Price	Target variable stores information of the ticket price in INR.

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

Research Question and Hypothesis

After doing some research, we found out that the typical price of a domestic flight in India is ~ 7,000 INR (As per SkyScanner, in May 2023. Keeping in mind our motivation for this project, the main research question that we will explore through our final project,is: "For flights that are departing out of cities, to what extent do the source destination, arrival time, departure time,and number of days before booking the flight influence the predictive odds of a flight being priced above 7,000 INR?"

Our formal null hypothesis and alternative hypothesis based on this research question are as follows:

Null Hypothesis H_0: None of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR.

Alternative Hypothesis H_1: At least one of the aforementioned factors will be predictive of the odds that the flight price is above 7,000 INR.

Data Cleaning

We first checked our data for missing values and duplicates. After looking through our data, we realized that we had none of the aforementioned problems. However, we still felt the need to modify certain variables to make our modelling process smoother and conduct an effective EDA.

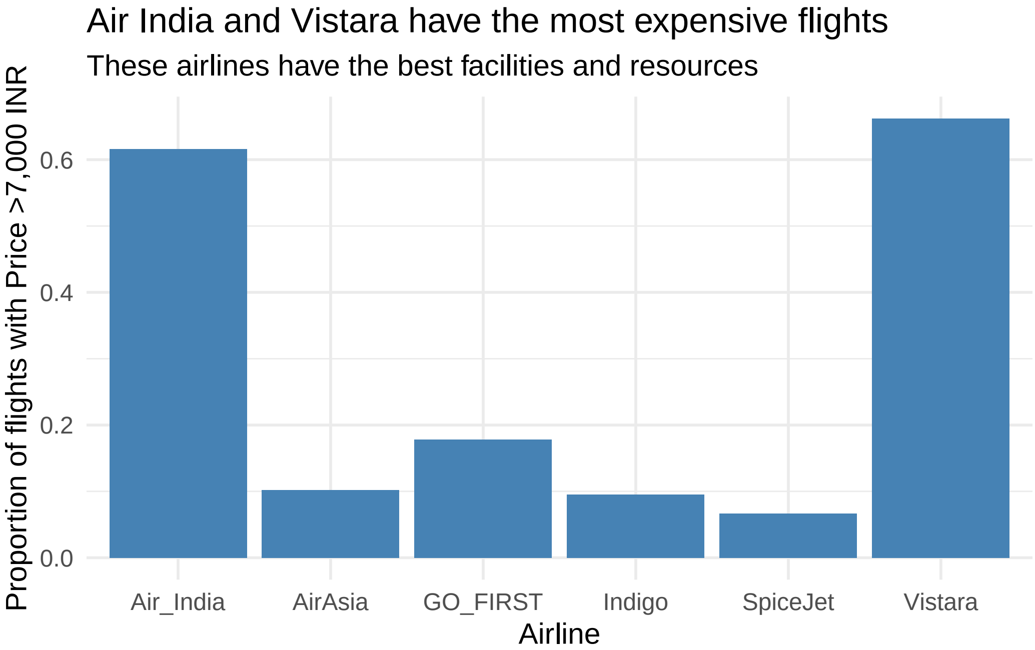
Since Mumbai and Delhi are the two biggest Hubs in India, we decided to focus our research project with these two cities as our primary source and destination.

This then reduces our number of observations to 30,098 flights.

Our main change was to our main variable of interest. We converted it to a binary variable, 1 for if it was greater than 7,000 INR and 0 if it was less than 7,000 INR.

Exploratory Data Analysis

Before we explore our research question, we are going to analyze summary statistics and other relevant relationships between our main response variable of interest Ticket Price and other variables in the data set such as duration, airline, class and days left. By understanding the patterns behind these trends, we will be able to effectively explore our research question.



National airlines such as Vistara and Air India, have the highest proportion of flights that we are considering expensive. The more commercial airlines have cheaper options. I found this quite interesting, because I would assume the national airlines would be cheaper and more affordable to all. We also see a great difference between the proportion of expensive flights within the airlines. This suggests that the airline chosen might be predictive of the odds of a flight being overpriced.

Another important factor to consider is the duration of the flights.