Flight Price Predictions

This manuscript (permalink) was automatically generated from uiceds/project-triples@e4623ab on October 8, 2024.

Authors

- **Shayan Bafandkar**
  0009-0009-8172-5751 · sbafan
  Department of Civil & Environmental Engineering, University of Illinois Urbana-Champaign

- **Sofia Frenk** ✉
  0009-0001-8099-4900 · sofia-frenk
  Department of Civil & Environmental Engineering, University of Illinois Urbana-Champaign

- **Supreme Pandey**
  0000-0003-0775-6313 · supremepandey
  Department of Civil and Environmental Engineering, University of illinois Urbana-Champaign

- **Brandy Diggs-McGee**
  0000-0003-2052-0946 · iloveheat
  Department of Civil & Environmental Engineering, University of Illinois Urbana-Champaign; USACE ERDC CERL, Illinois

✉ — Correspondence possible via GitHub Issues or email to Sofia Frenk <sofiaf6@illinois.edu>.

Abstract

The primary goal of our project is to build a machine learning model that can estimate changes in future flight prices based on historical data by using regression techniques. We will investigate how factors such as time of departure, number of stops, and the choice of airline influence flight prices. The secondary objective is to analyze if certain trends can be linked to broader environmental, economic and/or policy factors. The dataset includes columns for departure and destination locations, total stops, travel duration, and price information. The model will be trained using machine learning techniques, with a focus on determining which features contribute most to price variations. The aviation industry is a critical component of the global transportation network, impacting not only the economy but also the environment due to its significant carbon footprint. By developing accurate flight price prediction models, we can contribute to better planning and optimization of air travel routes, which is essential for both transportation engineering and environmental sustainability. If airlines and passengers can anticipate future price trends, it enables more efficient scheduling, potentially increasing the efficiency of flight operation and possibly minimizing unnecessary emissions.

Proposal

Our team plans to use a Kaggle flight prediction dataset to develop a machine learning model in Julia that predicts future flight prices for domestic routes in India. The primary goal is to build a machine learning model that can estimate changes in future flight prices based on historical data, using regression techniques. We will investigate how factors such as time of departure, number of stops, the choice of airline, among others, influence flight prices. The secondary objective is to analyze if certain trends can be linked to broader environmental or policy factors. While predicting flight prices may seem primarily economic, it intersects with transport engineering by optimizing air traffic and

scheduling. Additionally, these predictions could indirectly inform decisions aimed at reducing the environmental impact of flights. By better understanding pricing trends, stakeholders can implement dynamic pricing strategies that encourage sustainable travel, such as offering discounts for off-peak flights or promoting direct routes to cut down on fuel consumption.

Dataset Description

- Source: The dataset used for this project can be found on Kaggle, at this link: https://www.kaggle.com/datasets/viveksharmar/flight-price-data It was used to help build a predictive model for flight price prediction using the data that will be explained below.
- Format: The dataset is in CSV format, which is commonly used for tabular data storage. Each row represents a specific data point, with columns detailing various features that might impact flight prices.
- Contents: The dataset serves as a basis for training machine learning models for predicion of flight costs. More specifically, the dataset includes the following columns:
  1. Airline: A String value representing the name of the Indian airline company included in the study
  2. Source: Another String value representing the city from which the airline departs
  3. Destination: Yet another String value representing the arrival city
  4. Total_Stops: a ternary integer variable between 0 and 2 that represents the number of of stopd from the city of departure to the arrival
  5. Price: An integer variable presententing the cost, in rupees, for each ticket
  6. Day/Month/Year: Three columns containing integer variables representing the date when the flight took place. Note that the year column contains only the year 2019, so we may remove this column
  7. Dep_hours/Dep_min: Two columns containing integer numbers representing the hour, in military time, and minute at which the flight departed
  8. Arrival_hours/Arrival_min: Similar to the Dep_hours/Dep_min columns, but for the the arrival time of the flight
  9. Duration_hours/Duration_min: Two columns with integer values representing the number of hours and minuted a flight lasted

Exploratory Data Analysis of Indian Domestic Flights (March - June 2019)

The dataset includes domestic flights of Indian airlines from March 2019 to June 2019, and is derived from Kaggle. Each column in the dataset corresponds to a specific variable, and each row represents an observation. The dataset is clean, with consistent measurement units and no missing values.

Dataset Variables:

- **Airlines**: The name of the airline operating the flight.
- **Source and Destination**: Cities where the flights originate and land.
- **Total Stops**: The number of stops made by the flight.
- **Price**: The ticket price for the respective flight.
- **Date, Month, and Year**: The specific date on which the flight is scheduled.
- **Departure and Arrival Times**: Detailed departure and arrival hours and minutes.
- **Duration**: The total duration of the flight in hours and minutes.

Correlation Analysis:

We explored possible correlations between variables in the dataset. One expected correlation is between flight price and flight duration. Using the `cor` function in Julia, we found a positive correlation of **0.51** between these two variables. Similarly, the correlation between the number of stops and price is **0.60**. It makes sense that as the number of stops increases, the flight distance and, consequently, the price also increase.

The chart depicted in **Figure 1** illustrates that most flights in the dataset have ticket prices below 10,000 Rupees.

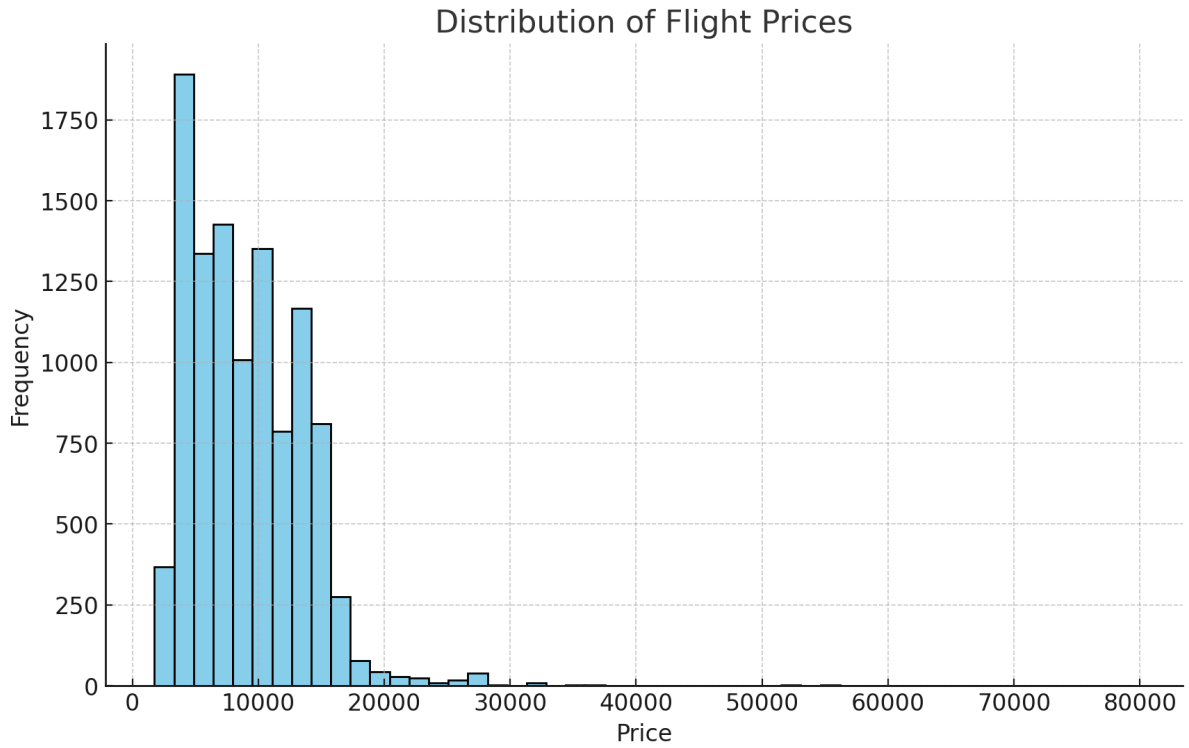

Figure 1: Distribution of Flight Prices (Positive Skew).

Seasonal Price Variations:

To analyze seasonal price variations, we created a new column, `Adjusted-Date`, by combining the values from the `Date`, `Month`, and `Year` columns into a single date format. We then plotted the mean price over time using this adjusted date. As shown in **Figure 2**, flight prices fluctuate significantly over time, with notable peaks around the major Indian holidays.

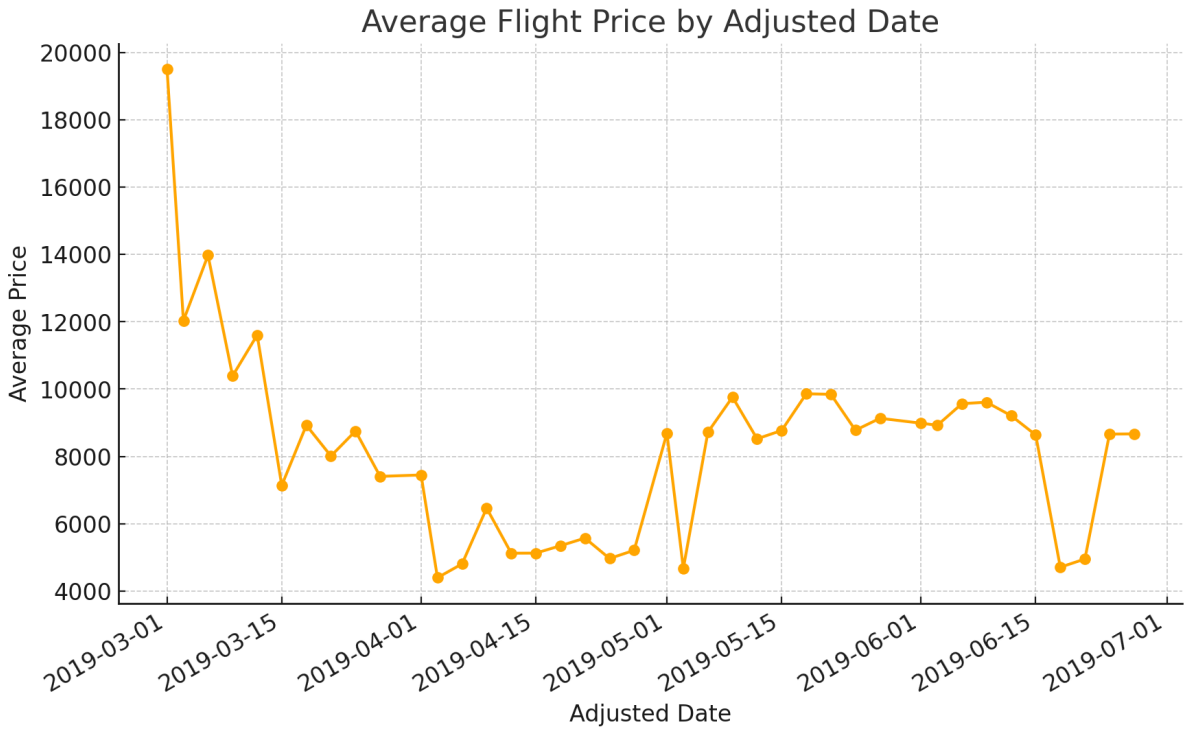

Figure 2: Flight price trends over time.

These price variations can be correlated with the seasonal demand and cultural events during this period. Upon reviewing the price fluctuations, we explored the major holidays in India during this period to identify possible correlations between price peaks and holidays. Interestingly, many of the price peaks align with Indian holidays. For example: - In March, price spikes around March 4th and 21st coincide with **Maha Shivaratri** and **Holi**, respectively. - In April, a price increase occurs around April 13th and 14th, aligning with **Ram Navami**, **Baisakhi**, and **Tamil New Year/Vishu**. - In May, a price increase is observed around May 1st (coinciding with **May Day**) and May 18th (coinciding with **Buddha Purnima**). - High prices persist into early June, corresponding with **Eid-ul-Fitr** (June 4th) and **Ganga Dussehra** (June 12th).

Destination Analysis:

We reviewed **10,684** flights during this period. **Cochin**, **Bangalore**, and **Delhi** were the top destinations, with Cochin being the most attractive, receiving the highest number of flights. The details of the top destinations are shown in **Table 1**.

Table 1: Top Flight Destinations

Rank	Destination	Count
1	Cochin	4,537
2	Bangalore	2,871
3	Delhi	1,265
4	New Delhi	932
5	Hyderabad	697
6	Kolkata	381

Origin-Destination (O/D) Pair Analysis:

We also identified the most frequent origin-destination pairs, as shown in **Table 2**.

Table 2: Most Frequent Origin-Destination Pairs

Rank	Source	Destination	Count
1	Delhi	Cochin	4,537
2	Kolkata	Bangalore	2,871
3	Bangalore	Delhi	1,265
4	Bangalore	New Delhi	932
5	Mumbai	Hyderabad	697
6	Chennai	Kolkata	381

Airline Insights:

Our analysis of the airlines provided the following insights:

1. Mean Price by Airline:

The table below (**Table 3**) shows the mean flight price for each airline, sorted from highest to lowest.

Table 3: Mean Price by Airline

Rank	Airline	Mean Price (INR)
1	Jet Airways Business	58,359
2	Jet Airways	11,644
3	Multiple Carriers Premium	11,419
4	Multiple Carriers	10,903
5	Air India	9,611
6	Vistara Premium Economy	8,962
7	Vistara	7,796
8	GoAir	5,861
9	IndiGo	5,674
10	Air Asia	5,590
11	SpiceJet	4,338
12	Trujet	4,140

2. Airlines with the Most Number of Flights:

The table below (**Table 4**) lists the airlines with the most flights in the dataset.

Table 4: Airlines with the Most Number of Flights

Rank	Airline	Number of Flights

Rank	Airline	Number of Flights
1	Jet Airways	3,849
2	IndiGo	2,053
3	Air India	1,752
4	Multiple Carriers	1,196
5	SpiceJet	818
6	Vistara	479
7	Air Asia	319
8	GoAir	194
9	Multiple Carriers Premium	13
10	Jet Airways Business	6
11	Vistara Premium Economy	3
12	Trujet	1

3. Airlines Frequently Used in Long-Haul Flights:

The table below (**Table 5**) lists the airlines frequently used for long-haul flights (flights with a duration greater than 10 hours).

Table 5: Airlines Frequently Used in Long-Haul Flights

Rank	Airline	Long-Haul Flights
1	Jet Airways	2,395
2	Air India	1,178
3	Multiple Carriers	625
4	IndiGo	231
5	Vistara	197

It is worth noting that there is limited data available for multiple-carrier flights, so further analysis of these flights is not possible.

References