Where to fly next?

Vinay Sanga

1 Introduction

You are having a long weekend and want to fly to another city. As a smart person, you want to know the cheapest as well as fastest option to help you decide your flight better. However, a single website will not suffice as there may be better deals on other websites. So, you want to aggregate result from three flight aggregators and then choose your flight. To start with this, we will first scrap the data for three websites. After this, we will carry out data processing to clean the data and then Exploratory Data Analysis(EDA) to get interesting facts about the flights. Finally, we will have some user interaction with the data. It will contain the user inputs like price, duration, etc. Finally, we will show the cheapest flight for the as per user's most convenient time, followed by the fastest flight for the same.

2 Data Collection

We collected the data from Momondo, Booking.com, and Kayak for the flights between Helsinki Vantaa (HEL) - Paris Charles de Gaulle (CDG) on 25th October. The tools available for scraping the data are BeautifulSoup and Selenium Web Driver. Since these are dynamic websites, we used the Selenium Web Driver to interact with them and collect the data. The dataset contains the following information:

- Layover - The layover in the journey
- Stops - Total number of stops
- Duration - The flight duration (including layover)
- Departure - The departure time from HEL
- Arrival - The arrival time at CDG
- Carriers - The airline names
- Aircrafts - The aircraft type (if available)
- Price - The cost of the ticket (in USD($))
- Site - The flight aggregator where the data was taken from

3 Data Analysis

We used libraries such as Seaborn and Matplotlib for data visualization to gain insights into the dataset. Various aspects of the data were explored, including flight attributes such as stops, duration, departure/arrival times, and carriers. Here's an interpretation of the data visualizations and interesting observations.

The majority of flights in our dataset have 1 stop. From figure 1 we can imply that, flights with 1 stop are typically priced around $200. This suggests that **non-stop flights may be more expensive**, while flights with one stop tend to be more affordable.

Our analysis also looked at the duration of flights and how it relates to prices. We found that the majority of flights have a duration of **approximately 500 minutes**. Interestingly, flights with this duration are consistently priced around $200.

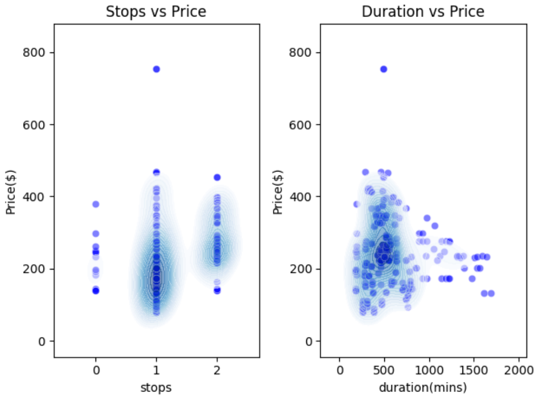

Figure 1: Stops and duration against the price

We investigated the impact of departure times on flight prices as shown in figure 2. The most common departure times are around **06:00 and 15:00**. Flights departing at these times are priced at approximately **$200 and $230**, respectively. Correspondingly, the arrival times for these flights are around 13:00 and 20:00. This shows that the **afternoon flights are faster**. Travelers who prioritize convenience may find these options appealing, but they come at a higher cost.

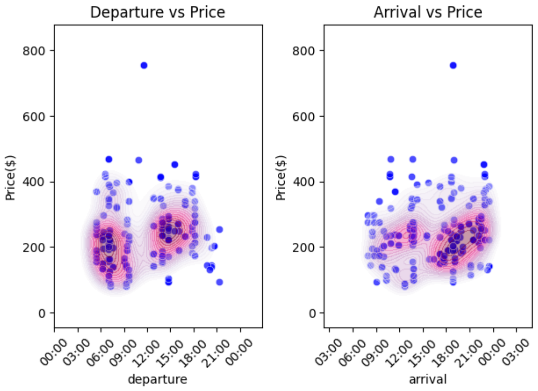

Figure 2: Departure and arrival against the price

To further analyze the relationship between departure times and prices, we created a bar plot as figure 3. This plot reveals interesting insights:

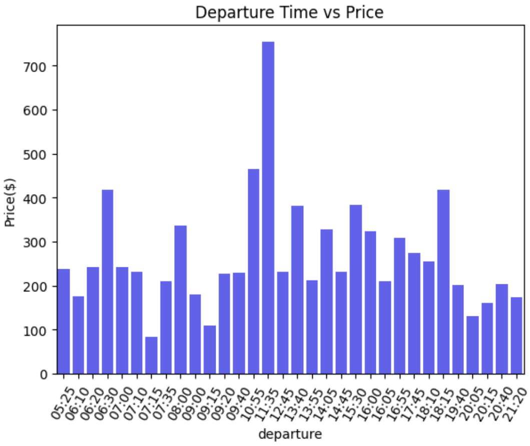

Figure 3: Price vs departure

- Prices are **highest** for flights departing at **11:35**, followed by 10:55 and 18:15.

- Conversely, later departure times, ranging from **20:05 to 06:20**, tend to have **lower** prices. This suggests that flights during peak hours come at a premium, while off-peak flights are more budget-friendly.

- Furthermore, **late-night** and **early morning** flights exhibit **price stability** and do not experience sharp price variations, unlike morning and afternoon departures.

- Finally, we analyzed the relationship between different airline carriers and their respective flight prices which is figure 4. Our findings indicate that **Braathens Regional Aviation** tends to offer the most expensive flights, while **Norwegian** consistently provides more budget-friendly options.

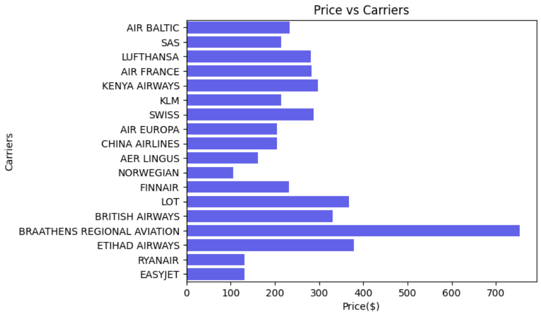

Figure 4: Price vs carriers

4 Conclusion

The scientific bottlenecks in the project were related to the need for comprehensive data collection and data cleaning. The websites were not allowing selenium web driver to access the pages directly. In order to emulate a real user, we used selenium in stealth mode by varying user agents across the webdriver sessions. Also, sometimes there would be pop-ups related to cookies, ratings, etc. which needed to be handled. During the data cleaning, we had to remove the null entries which were introduced due to flights becoming full. Also, as the three sites were different, their data was also formatted in different formats. This included the fields like departure and arrival times , layover locations, etc. These were overcome by converting the formats using regex.

But, in the end we were able to find out the cheapest and the fastest flights based on the criteria provided by the user. The plots and the data not only helped to find 'where to fly' but also 'when to fly'!