Mini Project 1 - Where do I fly next?

Thinakone Louangdy

1. Introduction

Given a destination, flight date, departure location, and arrival destination, we want to get the best possible flight for a trip. Websites such as Kayak.com, Momondo.com, and Booking.kayak.com offer the chance to look into the different flights available for this proposed journey with information about the price, flight duration, layover time, transit location, and others. However, it would be better if we had a process to compare the flight deals offered by the different sites.

To solve this issue, we can get the data from these websites and apply data processing steps to get the desired results. The data from these websites will be obtained through web scraping, which is when we extract the data from a website using either the HTML structures of the website or the APIs used by the website. Through this, we can get all the data in one place to compare and get the best deal possible.

2. Data Collection

I collected the flight data from three websites, namely:

i. Kayak.com
ii. Momondo.com
iii. Booking.kayak.com

First, we set the following to predefined values for the process.

Departure: Helsinki
Destination: Vienna
Flight date: 23rd November 2023

Using these values, we scraped the available flight data from the three websites with main details such as the **name of the airline, departure time, arrival time, flight duration, whether it is a direct flight or not, layover time, transit location/airport name, price, and website**. These details were collected for more than 50 flights presented in the search results for each website.

The tools I used for scraping the data were

- BeautifulSoup
- Selenium

- Selenium-stealth

I checked the class names and CSS selectors for each required element to get the necessary information. To get some of the additional details of each flight, it was also necessary to perform a button click on the specific HTML element. For example, the flight layover time was unavailable in all three scraped websites with the main search results. Since websites nowadays are dynamic, which made it hard to scrape data, we also had to perform a click with the selenium library on the HTML element that responds for load more flight deal. The data collected were saved as one CSV file.

3. Data Processing

The collected data had to be cleaned before it could be used. I have to compare the data from the three websites. It was important that the data would be performable with data visualization. For example, we must ensure that price values are computable with pandas functions, e.g., the price data includes a dollar sign; this has been handled in the data cleansing process. Departure and arrival times are most likely in the DateTime type, which we need to convert to a digit (integer or float). However, the layover time and flight duration are in hh:mm:ss format, and we have to make it become a digit for easy visualization.

I checked the dataset for duplicated values, dropped it, and checked the data types for of all the columns.

As mentioned above, I updated the price, departure, arrival time, layover time, and flight duration to maintain only numeric values. The airline name column values were processed by removing an unused data row, e.g., Ads that come into our data when performing the scrape.

4. Data Analysis

For data analysis, I first investigated the description of the numerical data. The mean price of the flight was found to be $255, with a minimum of $82 and a maximum of $761. On average, the flight duration was 3.1 hours (3h 06mins), and the mean layover time was 8.4 hours. I then looked at the distribution of the numerical values by bar chart of price and flight.

Looking at the price distribution, most of the flight's prices were in the $200-$300 range, followed by $100-$200. There is a gradual decrease in the number of flights in each higher price range.

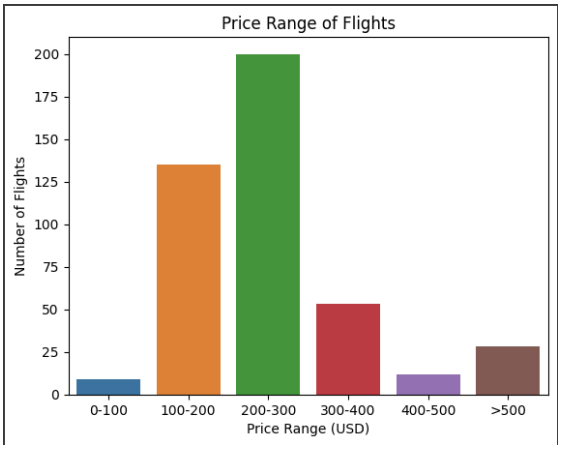

Fig 1. Countplot Showing how many flights belong to each price range

From the distribution of airlines and price, it can show that the airlines which offered the most of their flight in the price range of $200-$300 were Scandinavian Airlines and Lufthansa, while Finnair, Ryanair, and airBaltic were offering the flights in $0-200 price range.

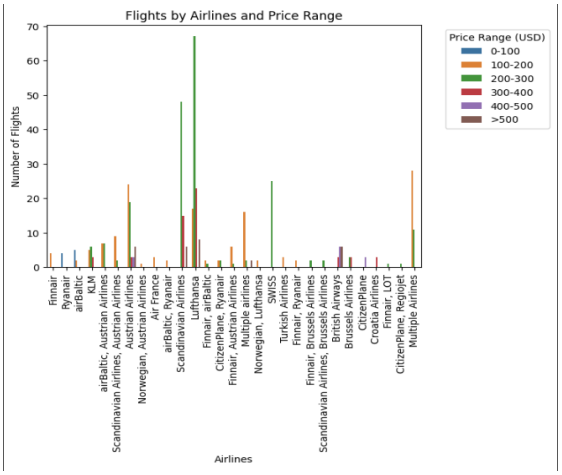

Fig 2. Bar chart to show the number of flights by airline and price

I used a histogram to plot the distribution of flight duration by direct or non-direct flights. The flight duration was in hours. The plot shows that direct flights were typically shorter than non-direct flights. The majority of direct flights are between 2 and 4 hours long, while the majority of non-direct flights are between 5 and 11 hours long. We can see a broader range of flight durations for non-direct flights than for direct flights because non-direct flights can have one or more layovers, adding to the overall flight duration.

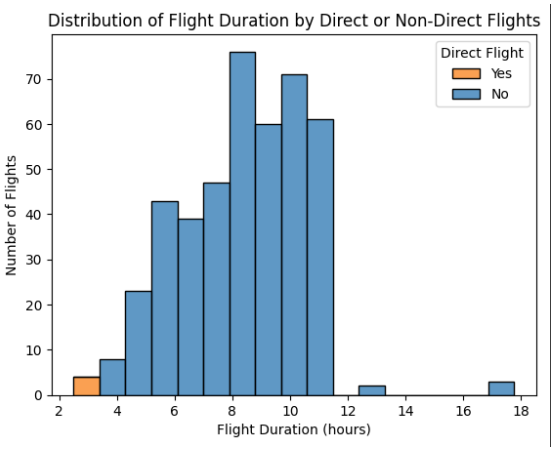

Fig 3. Histogram to show the distribution of flight duration and varies by direct or non-direct flights

I used a scatter chart to show the relationship between departure hour and price, with airlines as the hue. The price of a flight generally increases with departure hour. This is because flights that depart during peak hours (such as early morning or evening) are more popular and in higher demand. There is also some variation in the price of flights by airline. For example, Norwegian and Lufthansa flights are generally more expensive than Croatia Airlines flights, even if they depart simultaneously. The flight price is generally influenced by the departure hours and the airline.

Fig 4. The scatter chart shows how departure time affects the price of the flights and differs by airlines

Again, I used the scatter chart to see the relation between layover time and price, with transit airport as the hue. The price of a flight is generally influenced by the layover time and the transit airport.

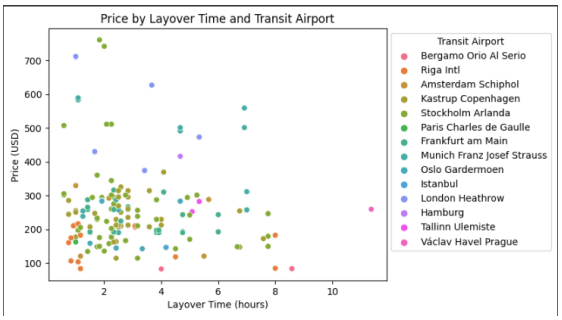

Fig 5. The scatter chart shows that layover time affects the price of the flights and differs by transit location

Then, I used a bar chart to show the average price of flights by airline and website. The price of a flight can vary depending on the specific website where it is booked. For example, airBaltic and Ryanair flights are often cheaper on Booking.com than on Kayak.com.

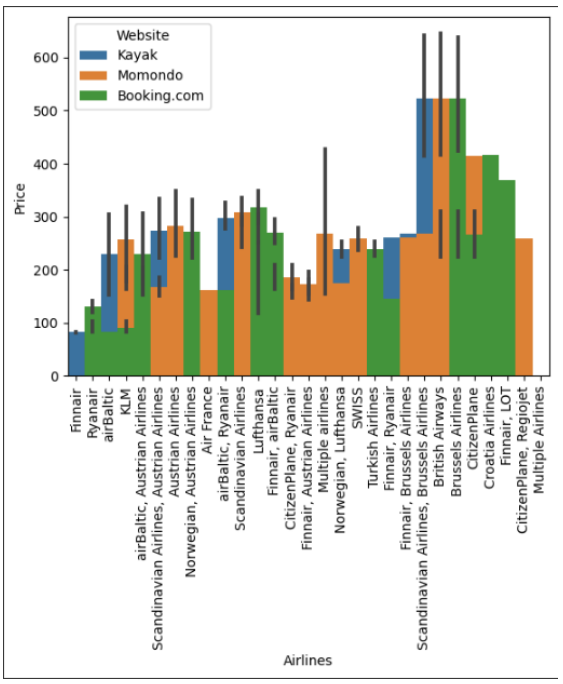

Fig 6. A bar chart to show the price by airline on each website

5. Conclusion

Web scraping aids in the extraction of essential data from many websites. However, because data is a valuable resource and businesses are aware of online scraping activities, websites are now enforcing steps to prevent web scraping.

The websites I scraped above are dynamic websites that require user input to obtain the desired results, such as search results for the departure airport or city and on the specified date of departure. These websites make it more challenging to obtain the necessary data pieces.

All three websites I scraped data from use pagination to keep the website lite, and it is considered a modern website, which means we will only get the data that is first displayed in the list, which is not enough to perform an EDA. It can resolved using

Selenium to verify and wait until the required element has loaded before proceeding. Then, I also used Selenium to perform a click to load more flight data but still faced the more challenging part. Scraping the flight booking agency is quite challenging when we perform the scrape; in our case, we could obtain more data if we scraped it from 9 p.m. to 6 a.m.

This project taught me about many challenges that can arise during web scraping and how to overcome them on real-world data, practice data processing, exploratory data analysis (EDA), and data visualization. Lastly, I created a user interaction for others to filter the cheapest or fastest flight by inputting their price range, favorite airlines, and flight duration range.