# A Data-Driven Flight Recommendation System

Shihabur Rahman Samrat

**Abstract**

In this project, we are incorporating web-scraping techniques to gather data regarding flights from different websites. This data is then used for an exploratory data analysis (EDA) to find the significant attributes and relationships. This allows us to understand which features might be more important to the users. Finally, a simplified interface is programmed, which allows the users to input certain filters, and based on that, a list of suitable flights are displayed to the user. This is a mini project; however, this approach can be scaled into a larger system, which can become useful in real-life scenarios.

## 1 Introduction

With the increasing availability of flights and airlines, travelers often face the challenge of choosing the most ideal flight for their needs. This project aims to address this issue by leveraging web scraping and data analysis techniques to provide users with personalized flight recommendations.

The objective of this mini project is to develop a flight recommendation system for travelers departing from **Helsinki, Finland (HEL) to Malaga, Spain (AGP) on the 20th of December 2023**. This system will allow users to tailor their flight preferences based on criteria such as price range, trip duration, number of stops, and airlines. The system will then provide a list of ideal flight options based on these user-defined criteria. The data is collected from multiple websites, meaning that the users have a better chance of obtaining the flights that are suitable for them in comparison to just using a single website.

## 2 Data Collection

For this project, we collected flight data from three popular booking websites:

1. Kayak
2. Momondo
3. Agoda

These websites were chosen for their comprehensive flight listings and user-friendly interfaces. Besides this, the scraping complexity of these websites was reasonable. To extract flight data, Python alongside **Selenium and BeautifulSoup** was used. The data extracted includes the following information:

1. Flight price (price)
2. Airline (airline)
3. Number of layovers (no_of_layovers)
4. Total layover duration (layover_duration)
5. Flight duration (duration)
6. Departure time (departure_time)
7. Arrival time (arrival_time)
8. Website name (website)

After scraping, the data is stored in a CSV file for further use and analysis.

## 3 Challenges

While web scraping is a powerful tool for data collection, it comes with its own set of challenges. Some common challenges encountered during scraping included website structure changes, timeouts, and handling CAPTCHA. Also, some websites required Selenium to click on certain elements to load more data.

Firstly, some attributes were very difficult to capture, like "layover_duration." This required me to click through all the elements to obtain this data, which is very time-consuming when done with Selenium. Also, finding the three websites to scrape was challenging since not all websites will allow you to scrape, and some have extreme security measures in place that will disrupt the scraping process.

To address these challenges, we implemented error handling and went with websites that do not have CAPTCHA and other complicated security measures.

The raw data obtained from web scraping may contain inconsistencies and missing values. This is because we used three different websites, and each website has a different way of storing the same sort of value. For example, Momondo might have their departure time in a 24-hour format, whereas Kayak might have it in a 12-hour format. We performed data cleaning by removing duplicates and null values and ensuring data uniformity by converting all values to a specific format.

Much of the data cleaning was performed during the scraping itself. For instance, the flight prices had symbols like '$', and ',' which we removed during the scraping. Also, some of the data had whitespaces for no reason, which were removed. Besides this, after reading the CSV files, we checked and removed the duplicates as well as the null values.

To make the data suitable for analysis, we transformed it into a structured format. For instance, the departure_time and arrival_time were converted to appropriate datetime formats so that they are easier to work with. Also, while scraping, I converted the duration, which was in hours:minutes, to just hours, to ensure uniformity.

## 4 Exploratory Data Analysis (EDA)

To gain insights into the trends in our data, we conducted various data visualizations and analyses. These include heatmaps, bar charts, histograms, line plots, etc.

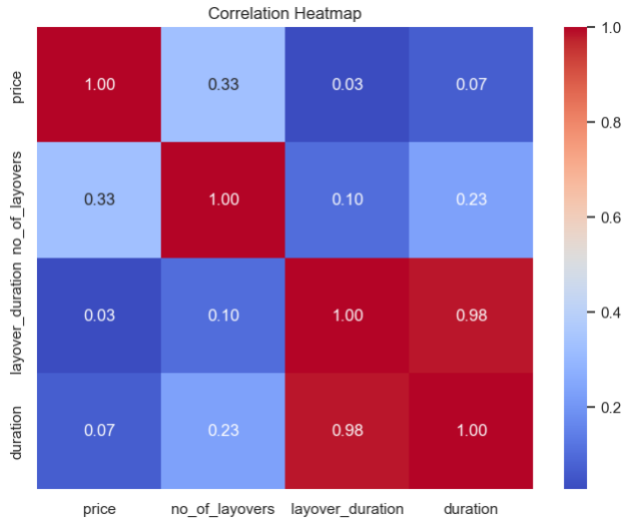


*Figure 1 - Correlation Heatmap for Numerical Columns*

The heatmap in Figure 1 shows a strong positive relationship between the layover duration and the flight duration, which makes sense since the layover duration is also part of the flight duration. A negative relationship exists between price and layover duration since higher prices usually mean lower layovers.

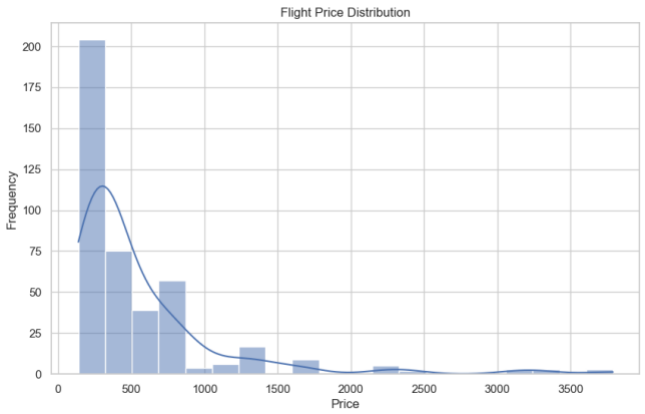


*Figure 2 - Flight Price Distribution*

From the histogram above in Figure 2, it can be observed that the majority of the flights have a price range under 1000 USD, with the highest number of flights being even below 500 USD. Since the flight is within Europe, this makes sense, as many of the flights are cheap.

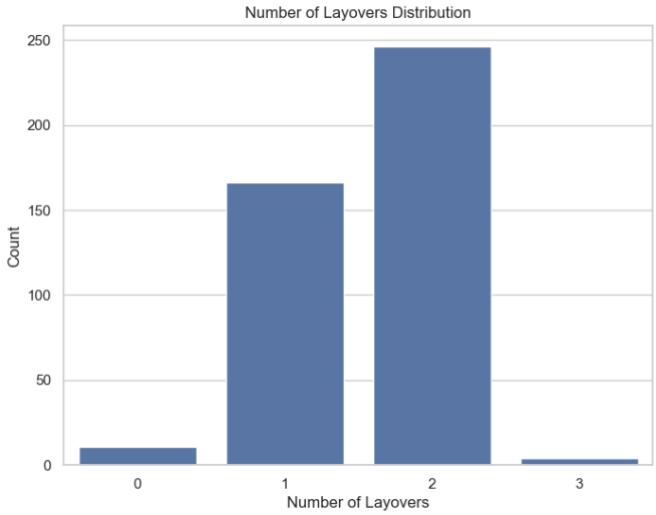


*Figure 3 - Number of Layovers Distribution*

The histogram in Figure 3 clearly shows that the majority of the flights from HEL-AGP have two layovers, with one layover being the second highest. This means that the number of direct flights is low. There is at least one layover in most circumstances.

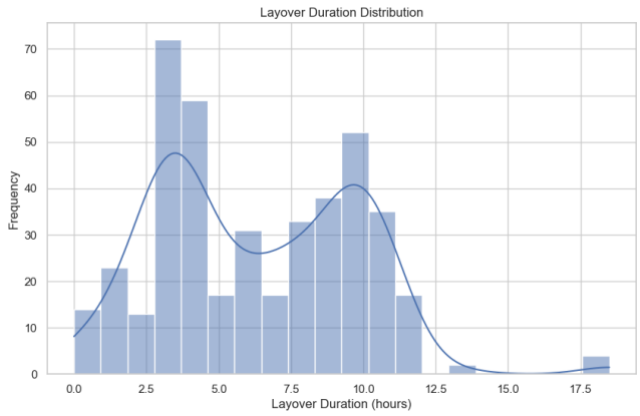


*Figure 4 - Layover Duration Distribution*

This histogram in Figure 4 is interesting since it shows us how long the layovers are. Most of the flights have their layover duration on the lower end, between 2.5 to 5 hours, which is not that bad. Very few flights have a layover duration of over 12 hours.

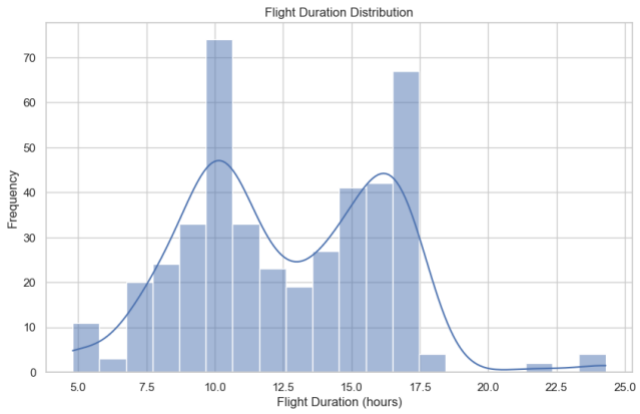


*Figure 5 - Flight Duration Distribution*

Since layover duration and flight duration are strongly related, a similar trend is also seen here in Figure 5. Very few flights are over 17 hours, with many of the flights being around 10 hours.

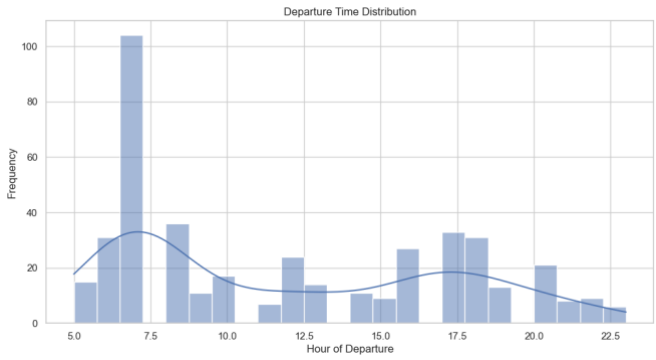


*Figure 6 - Departure Time Distribution*

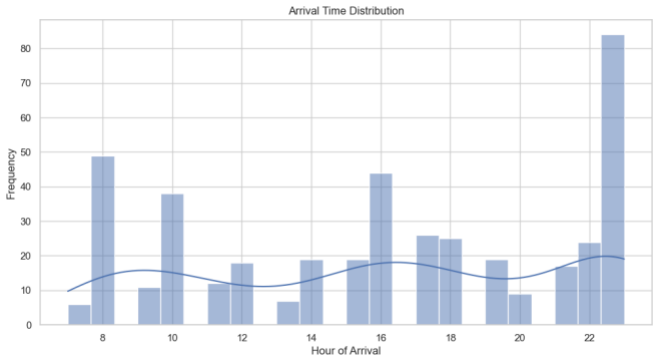


*Figure 7 - Arrival Time Distribution*

Both the histograms in Figures 6 and 7 above show that most of the flights depart in the morning and arrive at the destination right before midnight. There are other flights at different timings throughout the day as well.

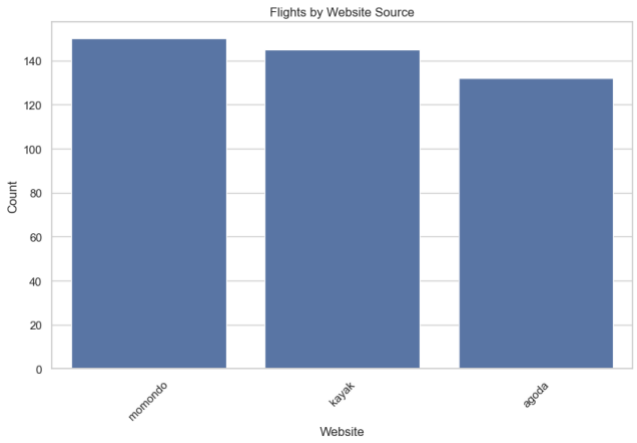


*Figure 8 - Website Source Distribution*

We can observe in Figure 8 that we acquired an almost equal amount of data from the three websites. There is little to no imbalance.

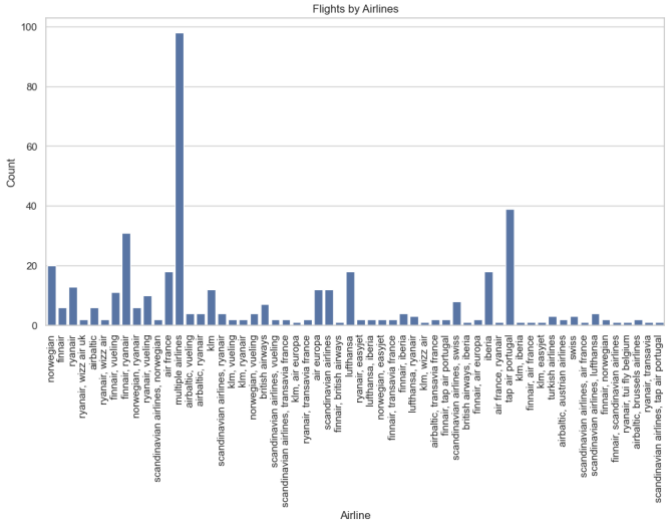


*Figure 9 - Airlines Distribution*

Since most of the flights are offered by multiple airlines, we can see the highest peak in Figure 9. Very few flights are offered by a single airline. Most of the flights are offered by at least two airlines.


*Figure 10 - Average Price by Website*

Figure 10 shows that Agoda has the highest average prices, followed by Momondo and Kayak. However, the difference in prices is not that significant.

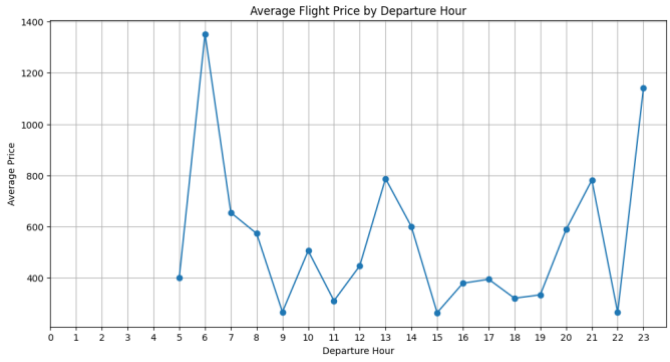


*Figure 11 - Average Price by Departure Hour*

The line plot in Figure 11 shows that the flight prices are the highest at 6 AM and 11 PM. It is better to travel at other times of the day when the prices are lower.

## 5 Observations

Based on the performed EDA, the following assumptions can be made:

- Agoda has the highest prices, followed by Momondo and then Kayak.
- Most flights are under $500, usually around $300.

- It is better to avoid traveling around 6 AM or 11 PM as this is when the flight prices are highest.
- Most flights usually leave in the morning and reach the destination right before midnight.
- It is better to be prepared for a layover since most flights have 1-2 layovers, although they are mostly around 2.5-5 hours long.

## 6 Conclusion

In summary, this project demonstrates the power of web scraping and data analysis for creating a data-driven flight recommendation system. We collected flight information from three major booking websites to offer users comprehensive travel options from Helsinki, Finland, to Malaga, Spain, on a specified date.

We successfully addressed web scraping challenges like changes in website structure, timeouts, and CAPTCHA handling through effective error handling and selecting websites with simpler security measures. Data processing was vital to ensure data quality and consistency, with efforts to handle inconsistencies and missing values, resulting in a structured dataset for analysis.

Exploratory data analysis provided valuable insights into flight attributes such as price distribution, layover duration, and flight duration. It also revealed correlations between factors like layover duration and flight duration. This project lays the groundwork for a more robust flight recommendation system for real-world applications, enhancing travelers' decision-making processes and overall travel experiences.