Full name: Halidu Abdulai

Project: “Mini Project 1 – Where do I fly next?”

Introduction: In today's fast-paced world, finding the best flight options can be a time-consuming and complex task. The abundance of airlines, routes, fares, and ever-changing schedules can overwhelm even the most experienced travelers. For this reason, the aim of this mini project is to leverage the capabilities of web scraping, a technique that allows us to automatically extract valuable data from various sources on the internet, to aggregate vast amounts of information into a single, easy-to-navigate platform. In this project, we will be utilizing the power of web scraping to make flight selection easier for users. Users can input their travel preferences, such as departure time, flight duration, price range, and any specific airline preferences to help find the best possible flight options effortlessly.

How can web scraping help make finding the best possible flight options easy? This can be done by scraping different flight information from various different websites, comparing the different available options and then deciding on the best possible option.

Data collection: In this project, I chose Stockholm as the destination for flights from Helsinki and set the flight date to 27th November, 2023. It is also worth noting that the cabin class option was set to Economy since it is the most widely chosen option. Different flight information such as airline, time of departure, arrival time, number of stops, duration of the flight, layover time, and the ticket price was gathered using python libraries such as ‘selenium’ and ‘beautifulsoup’. The flight data was collected from three different websites: Kayak.com, Momondu.com, and Hotwire.com. Over 200 data samples was collected from kayak.com and momondu.com each, while around 90 data samples was collected from hotwire.com.

Data preprocessing: Because the data was collected from three different websites, the data format varied a bit. For instance, the airline feature had different values for the same airline on different websites (‘Norwegian’ on kayak.com, ‘Norwegian Air International Ltd’ on hotwire.com). For this reason, the data from the three different websites was first concatenated into one single data and the necessary data preprocessing such as ensuring that each feature column as the same format for each data entry, and converting each feature to the appropriate data type was performed.

Data analysis: After the preprocessing phase, the data was explored for insights. I first explored the distribution of the airlines to understand which airlines provide services from Helsinki to Stockholm. If there was any flight route that had multiple airlines, then they were treated as

though they were one flight. The assumption made was that, since all flights in the route contributed to conveying passages from the take-off point (Helsinki) to the destination (Stockholm), and then treating them as though they were separate flights would mean they all have different take-off points and destinations, which shouldn't be the case. For this reason, some of the labels in the figure below includes more than one airline names, and the explanation is that those airlines combine together to convey passengers from Helsinki to Stockholm. It was discovered that Scandinavian Airlines provided the most of flight options from Helsinki to Stockholm on the given day (27th November 2023) with about 28% overall of all available flight options. AirBatic in conjunction with Scandinavian airlines also constitutes about 20% of the available flight options. What this means is that, the two airline companies combine to convey passengers from Helsinki to Stockholm – which implies there is a stop point between departure and destination points. Passengers are transferred from one flight to another at the layover point. Norwegian Air Shuttle and Braathens Regional Aviation constitutes the least with about 0.18% each of the overall number of flight options available.

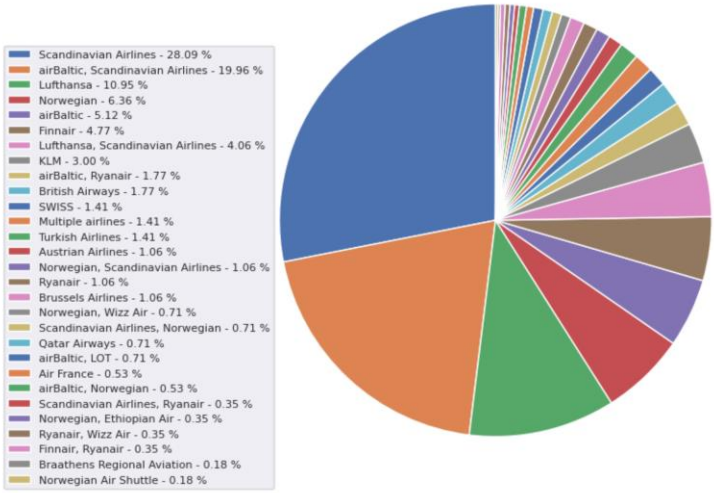

Figure 1: Pie Chart of airline distribution

Next, I explored the stop count of the flights. Zero means the flight is direct, one (1) means the flight has one stop point, etc. It could be observed that most of the flights had 1 stop, which was closely followed by the number of flights that has two stops. Few of the flights were direct flights.

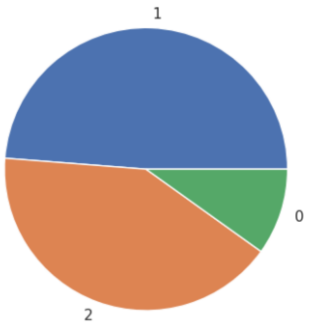

Figure 2: Pie chart of the number of stops of the air flights.

One other feature that of course, is of utmost importance to anyone buying a flight ticket is the price. For this reason, a histogram plot was made to explore the price range distribution in steps of 50 dollars (the bin width is 50). The resulting histogram plot is shown in the figure below.

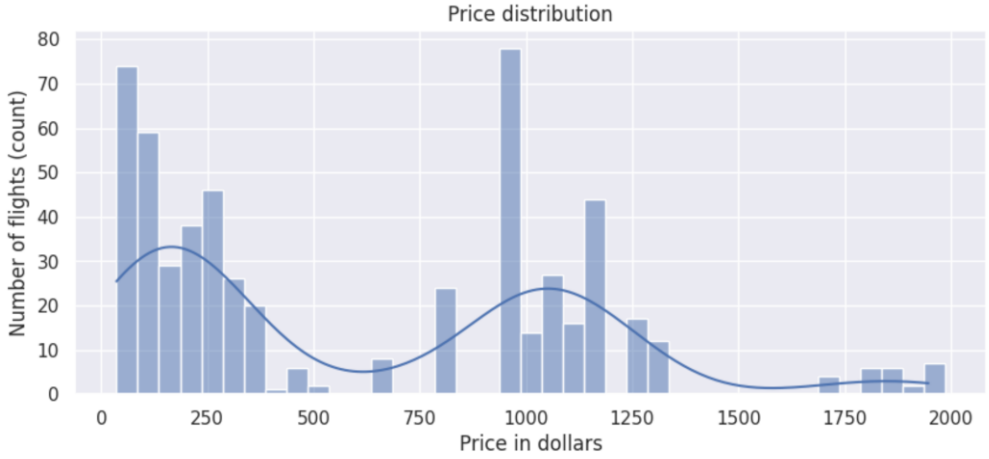

Figure 3: Price distribution in dollars

From the figure above, it can be observed that over 70 flight tickets are priced at or below 50 dollars, which is impressive, very cheap! Also, close to 80 flight tickets are price between 900-950 dollars, which is expensive (the reason could be that these flights have layovers). Very few flight tickets are priced over 1700 dollars. Observing the figure above reveals that, most of the prices are within 500 dollars. Also, most of the flights have prices between 900-1200 dollars.

In addition to these plots, I also explored other features such as the flight duration distribution, and departure times. The bar chart showing the departure times is shown below.

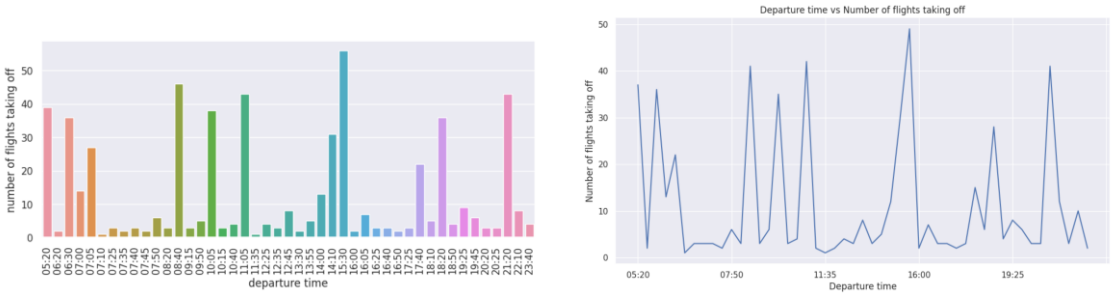

Figure 4: Bar and line charts of departure time

Looking at the figure above reveals that most of the flights take off between 05:00 - 07:05, 08:40 - 11:05, 14:00 - 15:30 and 17:40 - 22:00. In other times, the number of flights taking off is not as many as the time intervals specified. It seems that the time intervals, for which the number of flights taking off is less, are the times that most flights are arriving because there seems to be a trend in the data represented in the figure above. The number of flights taking off goes up and then down, at regular intervals, hence the assumption made.

Interaction with the user: The program allows users to customize their flight search experience, ensuring they find the best, cheapest, and fastest flight options tailored to their preferences and schedules.

Here is how it works:

1. Set their budget: The user starts by specifying their maximum budget for the flight. This ensures that they will only see options that fall within their price range, helping them make the most cost-effective choice.
2. Define their travel time: Next, the user has the flexibility to specify the maximum duration of their flight. Whether they prefer a quick getaway or don't mind a longer journey, it's all about finding flights that match their schedule.
3. Customize layover preferences: If the user wants to avoid lengthy layovers or prefer to have some extra time to explore during a stopover, they can set their maximum layover time, ensuring their journey aligns with their comfort level.
4. Choose direct or stopover flights: The user can select whether they are looking for direct flights or open to those with stops. This choice caters to their travel style and potential destinations.
5. Pick their preferred airline: If they have a favorite airline company they trust and love, they can filter the results to include only flights operated by their preferred airline.
6. Sort their results: The users can customize how their flight data is presented by selecting sorting criteria. Sort by price, duration, or other factors to easily compare options.

Based on these inputs, the program goes to work, filtering and sorting through the flight data to present the user with a tailored selection of flight options that meet their criteria.

But here's where it gets even better: The user can also specify their earliest convenient departure time, and the program will print the cheapest and fastest options among the filtered results that align with their schedule. This way, the user can make an informed decision that suits her time constraints and preferences perfectly. And the best part? It's a flexible process! The user can

always go back, adjust her filtering preferences, and input her departure time again to explore different flight options until she finds the ideal one for her journey..

Conclusion: In this project, flight data scraping was illustrated. It was shown how flight details can be scrapped from different websites on the internet using python libraries such as selenium and Beautifulsoup. The data scrapped, was preprocessed and then visualized for insights.

One of the major challenges in this project was how to scrap data from the internet. Some of the websites had policies that prohibit data scraping, as such data wasn't scrapped from such websites. Some of the websites also had very complex web structure which made getting the individual elements cumbersome and tedious. For instance, it took me two working days to write the code for scrapping data from hotwire.com.

Another challenge that surfaced worth mentioning is that the data format was different for the three websites, so I had to inspect the data scrapped from each website separately to understand the format, write helper functions to help make the data format for each feature column the same for all the separate datasets, before being able to concatenate the datasets together. A lot was leant during this project. I learnt techniques that helped me conveniently scrap data from complex websites. Also, different data visualization techniques for different data types were explored during this project. In conclusion, This project was a great starting point for data science students since it introduced some of the most difficult tasks in data analysis.