Mini Project 1 Kevin Koljonen

Introduction

The project goal was to make a program that could scrape data from three different booking websites for one-way flights to a certain location on a certain date with the starting location being helsinki-vantaa airport. The location that I've chosen is AGP, the airport in Málaga, Spain and the starting date being the 13th of november. The websites that I've chosen are Kayak, Momondo and Expedia.

Scraping

The data being scraped from the different websites consist of:

- Price of the flight
- The airline providing the flight
- Duration of the flight
- The number of layovers
- In what airports the layovers take place in
- Departure time
- Arrival time

The scraping is done mostly with Selenium but also uses Beautifulsoup to find the correct elements within the website. In this project I was only successful in scraping two of the sites mentioned, Kayak and Momondo, while the scraping of expedia proved to be more difficult than anticipated. Some of the scraping from Expedia works but most of it doesn't so I have chosen to leave it out of the data.

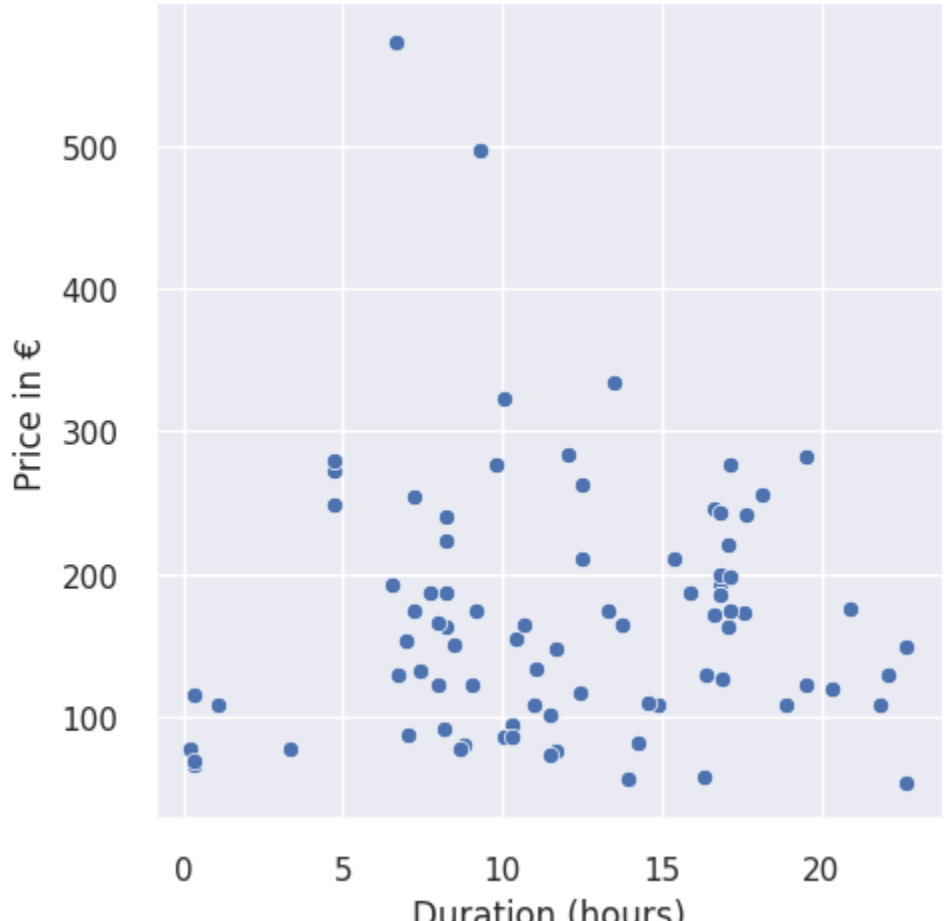
Data Processing

Some of the scraped data is immediately processed by removing unwanted characters or by changing the received data from a string format to a float value so it is possible to work with the data in an easier way. For instance the duration value is calculated into the amount of hours it would take eg. 4.5 hours, and the price has the € sign removed and converted into a float. The data from each website is put into its own data frame which is then combined into the final dataframe. From the data frame the user can see a certain flight's price, airline, duration, arrival time, etc. Because there might be multiple airlines that provide flights in travels with layovers, the column is expanded in the EDA phase to better show what airlines provide the most flights.

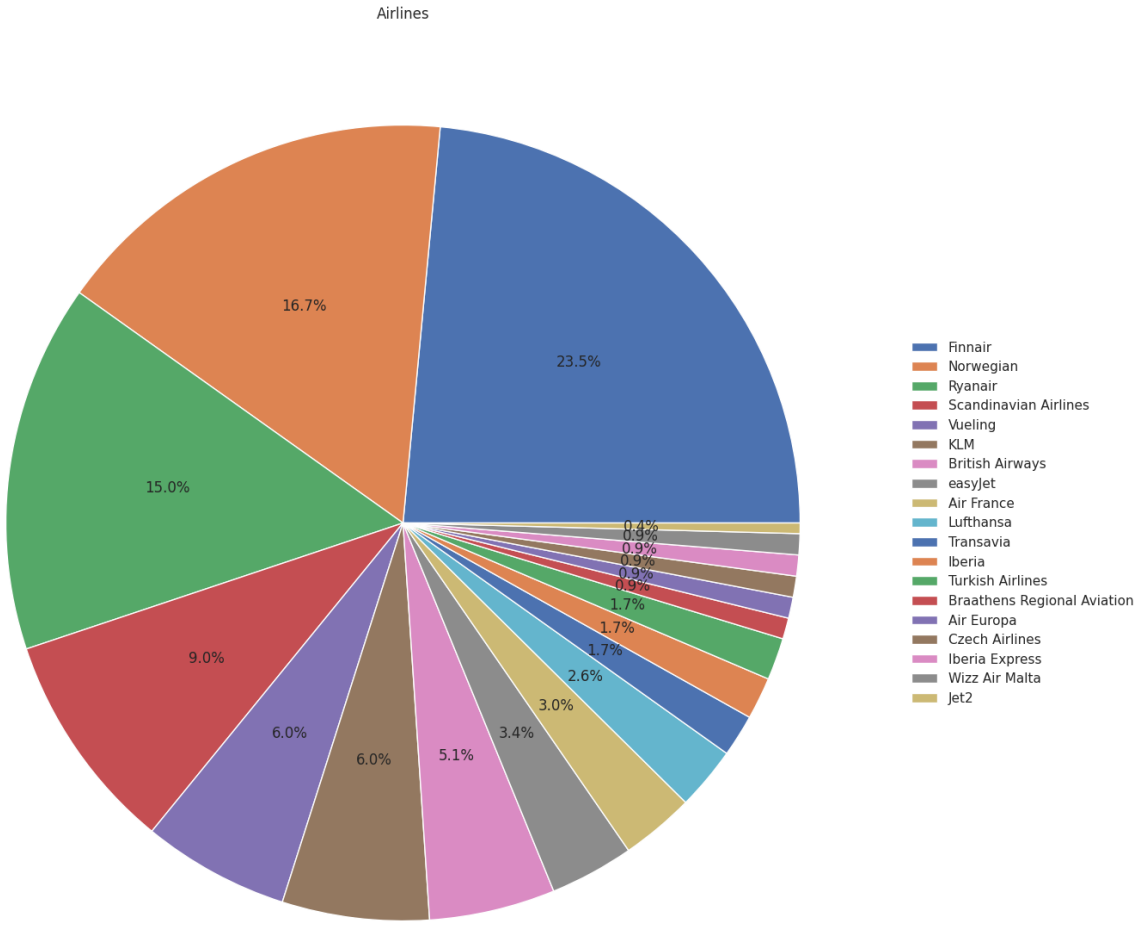
	Price	Airline	Duration	Stops	Layover_place	Layover_time	Departure_time	Arrival_time	Website
0	95€	Norwegian	10t 20m	1	OSL	4t 30m	14:10	23:30	Kayak
1	87€	Norwegian, Ryanair	7t 05m	1	ARN	1t 40m	09:15	15:20	Kayak
2	54€	Ryanair	22t 40m	1	STN	16t 35	22:10	19:50+1	Kayak
3	81€	Scandinavian Airlines	8t 50m	1	CPH	3t 15m	10:05	17:55	Kayak
4	248€	Norwegian	4t 45m	0		None	12:00	15:45	Kayak
...	...	...	...	...	...	...	...	...	...
148	116 €	Scandinavian Airlines	24t 20m	1	CPH	18t 45	18:20	17:40+1	Momondo
149	497 €	Lufthansa	9t 20m	1	MUC	3t 30m	07:00	15:20	Momondo
150	282 €	Air France	19t 30m	1	CDG	13t 40	17:40	12:10+1	Momondo
151	109 €	Scandinavian Airlines	25t 05m	1	ARN	19t 40	10:15	10:20+1	Momondo
152	78 €	Norwegian	27t 20m	1	ARN	22t 00	16:40	19:00+1	Momondo

EDA

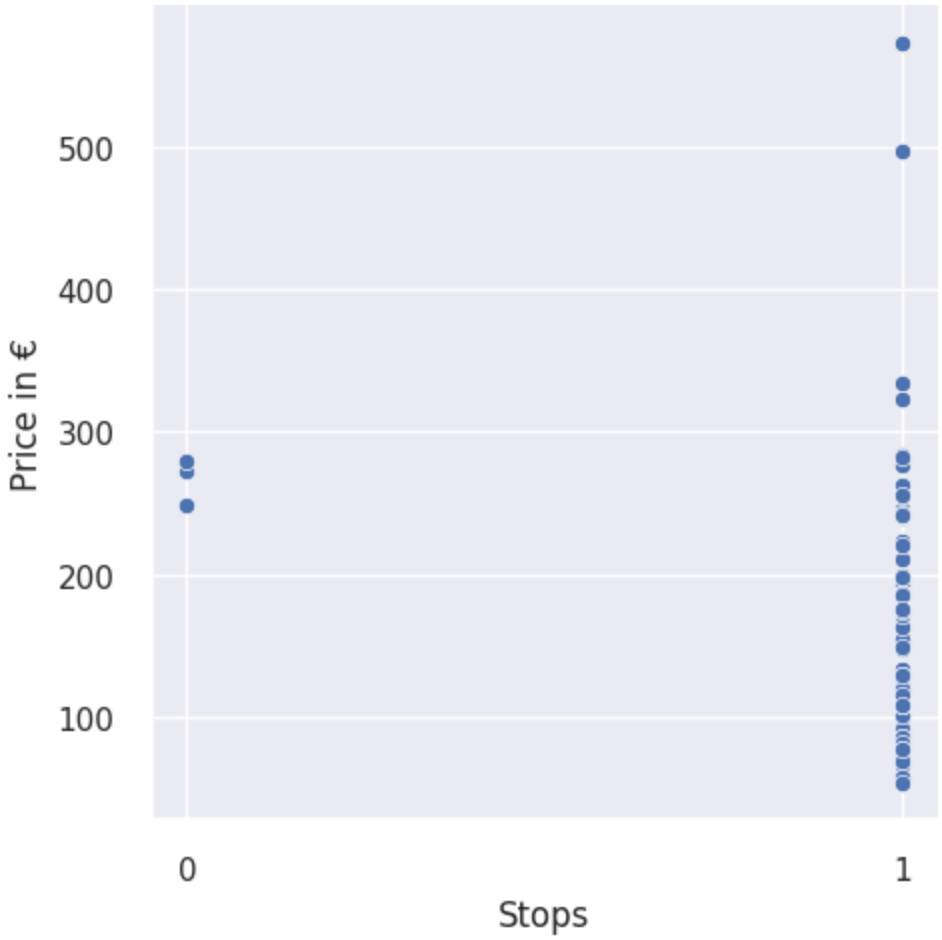
Some visualization has been done concerning the price, duration, airlines, number of flights with layovers, the amount of flights and pricing of them on the different websites.



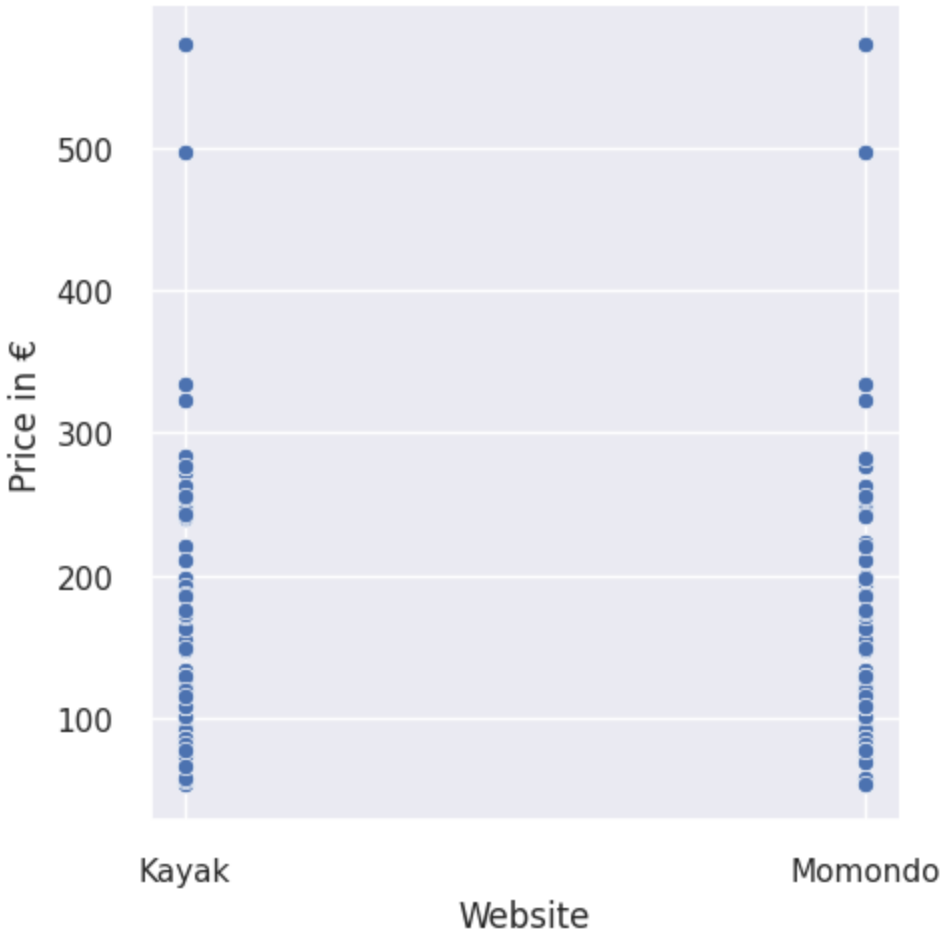
The graph above shows a scatterplot with the duration of the flights as the x axis and the price of the flights on the y axis, and this comparison shows that flights that have a lower price might not necessarily be longer or shorter in duration. This graph also shows some errors with the data processing or the scraping itself because there are some flights shown here that have a duration of almost 0 hours which is highly unlikely.



The next graph shows the distribution of the amount of airlines that provide flights through this route including if the airline changes during a layover. From this it is quite obvious that Finnair provides the most flights along this route even if they are only to a layover airport.



This graph represents the price of flights and the amount of stops the flights take. As the graph shows there aren't many direct flights from Helsinki-Vantaa airport to Málaga, Spain. This graph could've been done better with a histogram in my personal opinion but my inexperience with seaborn made it a difficult process.



The last graph shown in this report is the difference in flight prices between the two websites that I successfully scraped. Because of the similarities between Kayak and Momondo the data visualized isn't very helpful.

Other visualizations that are in the program include a piechart of layover airports, duration compared to stops and price compared by airline.

User interactivity

The user can narrow down the criteria for the flights based on the price, duration and number of stops. The program then outputs the 10 cheapest flights based on those parameters, as shown below.

```
Input the minimum for price range: 90
Input the maximum for the price range: 200
Input the minimum duration range (in hours): 4
Input the maximum duration range (in hours): 8
Input the minimum amount of layovers: 0
Input the maximum amount of layovers: 1
```

	Price in €	Airline	Duration (hours)	Stops	Layover_place	Layover_time	Departure_time	Arrival_time	Website	ID
15	122.0	Finnair, Ryanair	8.00	1	ARN	2t 30m	08:20	15:20	Kayak	15
91	122.0	Finnair, Ryanair	8.00	1	ARN	2t 30m	08:20	15:20	Momondo	91
12	130.0	Finnair, Norwegian	6.75	1	ARN	1t 25m	12:35	18:20	Kayak	12
88	130.0	Finnair, Norwegian	6.75	1	ARN	1t 25m	12:35	18:20	Momondo	88
16	132.0	Finnair, Vueling	7.42	1	BCN	1t 50m	17:05	23:30	Kayak	16
92	132.0	Finnair, Vueling	7.42	1	BCN	1t 50m	17:05	23:30	Momondo	92
13	153.0	Air France	7.00	1	CDG	1t 10m	17:40	23:40	Kayak	13
89	153.0	Air France	7.00	1	CDG	1t 10m	17:40	23:40	Momondo	89
26	166.0	Finnair, Ryanair	8.00	1	DUB	1t 50m	07:45	14:45	Kayak	26
102	166.0	Finnair, Ryanair	8.00	1	DUB	1t 50m	07:45	14:45	Momondo	102

Conclusion

This project has managed to provide and illustrate some of the data that was scraped and through visualization and user interactivity give the user some sort of idea what the flights from Helsinki-Vantaa airport to Málaga airport in Spain are like, price and duration wise. The project has been difficult due to my inexperience and rustiness in coding and many of the aspects of the program can be improved upon, for instance the scraping of the third website, processing the rest of the data, the visualization of the data and the interactivity, but the core goal of the project has been sufficiently fulfilled in my opinion.