Data Science - Fall 2022 - Mini Project I

Alex Montoya Franco
alex.montoyafranco@abo.fi
EDISS Master's Programme - Åbo Akademi University, Turku, Finland

1 Introduction

In 2020, online bookings were estimated to be worth $817 billion dollars, representing one of the largest market shares in the tourism industry [1].

Current statistics indicate that over 90% of travelers will do their research online, and 82% will end up making their booking online as well. Additionally, when it comes to research, many people will spend time on aggregate sites reading reviews of hotels, as well as looking for any travel information which can make their vacation better [2].

Taking this into consideration, online booking websites and OTAs (Online travel Agents) are not only showing people basic information about available accommodations, but are also becoming a hub of traveling information. People can now look in almost any booking site how close a hotel is to touristic places in a city, as well as reviews given by other users, and hotel's ratings, among other data.

Given all this information we might assume a person have all they want to book an accommodation. Nonetheless, as in any other consumer-centric industry, the price is one of the key factors when deciding whether to buy a product/service or not.

However, **what drives the price of an accommodation?** This might seem like an obvious question, and answers tend to point to the "market", inflation and additional services. And although these clearly have an impact on the final price of a booking, this report tries to demonstrate, after analyzing booking related data, that more factors are driving the price of accommodations in online booking websites.

With the discoveries presented by this report, we hope people are able to identify additional factors concerning why a certain booking offer might be more expensive in one website over others.

Data has been collected from three booking websites: Orbitz [3], Hotels.com [4], and Agoda [5] and represents the available booking offers for a trip to Rotterdam, Netherlands between the 24th and 27th of November, 2022 (1 adult).

A data integration process is developed in order to standardize data into a single format. Furthermore, Exploratory Data Analysis (EDA) tasks such as, identifying outliers and visualizing significant relations in the data are performed.

[1] https://www.stratosjets.com/blog/online-travel-statistics/
[2] https://www.stratosjets.com/blog/online-travel-statistics/
[3] https://www.orbitz.com/
[4] https://www.hotels.com/
[5] https://www.agoda.com/en-gb/

2 Data Collection

Data is collected from Orbitz, Hotels.com and Agoda via **Web Scraping**, a method to automatically obtain large amounts of data from websites which do not usually provide specialized APIs for that. Most of this data is unstructured data in an HTML format, which is then converted into structured data so that it can be used in various applications [6].

Initial scrapers built for this report are Selenium WebDrivers [7] to load data from dynamic javascript sites. These sites don't usually have all their data available for automatic scraping, but require some kind of interaction with the website to load content. For this report a simple interaction is implemented, which is using selenium to automatically scroll down until the end of a page so that all content is loaded and available for scraping.

The Selenium scrapers obtain the required HTML code and provide this data to BeautifulSoup [8] which offers an easy-to-use interface to capture elements within the HTML content.

Data collected:

- Hotel name
- Review score [0..10]
- Accommodation price [$USD]
- Hotel rating [0..5 stars]
- Hotel address
- Listing url
- Source [Orbitz, Hotels.com, Agoda]

3 Data Analysis

First of all, an exploration of prices per source is consider and we try to answer the question: **Are accommodation offers generally cheaper in any of the booking websites scraped?**

[6] https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/
[7] https://www.selenium.dev/
[8] https://pypi.org/project/beautifulsoup4/

Listing 1: Average accommodation price per source

```
rotterdam_eda_dataset
.groupby('source')['accommodation_price ($USD)']
.mean()

source
Agoda         92.808511
Hotels.com    95.640000
Orbitz        96.120000
Name: accommodation_price ($USD), dtype: float64

```

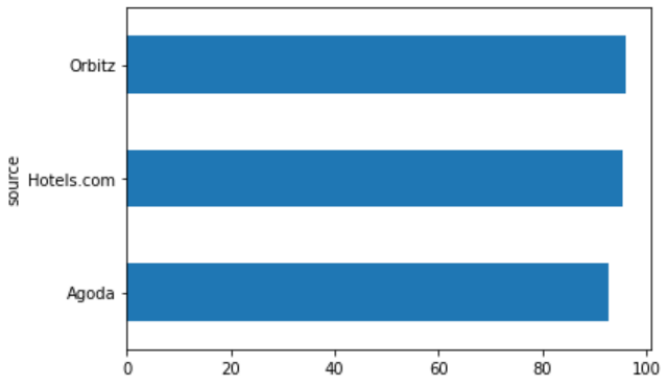

Figure 1: **Mean accommodation price per source**

Orbitz seems to be the most expensive out of the three sources. However, Hotels.com is close to that value, while Agoda seems to be a better option for users looking for a cheaper accommodation. This and further conclusions are merely based on the data available for this report, knowledge of room types, special offers, discounts and more would clearly affect these results.

This result might show a tendency, but it's an unfair comparison since one source might just happen to have some very expensive locations for the city and dates chosen. Therefore, in order to get a better sense of the prices per source, let's analyze the price difference when two or three sources are offering accommodations in the same hotel. Once again, this analysis is just taking into consideration the data available, more information about the specific offer of each source might give a better sense of why an offer is or is not more expensive than others.

Listing 2: Grouping same hotels in different sources

```
aggregated_offers = rotterdam_eda_dataset
.groupby('hotel_name')
['source', 'accommodation_price ($USD)']
.agg(list)

aggregated_offers = aggregated_offers[aggregated_offers[
'accommodation_price ($USD)'].str.len()!=1].copy()

```

In Listing 2 we group together hotels with the same name and display their source and accommodation price. For this analysis, we remove rows for which there were no duplicated value for the hotel name (no offers from other sources).

This is an extract of the results:

hotel name	source	price
171. Urban Design Hotel	O,H,A	92.0, 92.0, 82.0
Art Hotel Rotterdam	O,H	83.0, 83.0
Bilderberg Parkhotel Rotterdam	O,H,A	90.0, 90.0, 95.0

This comparison presents a better picture since it shows a similar distribution of prices between the three sources. When offering rooms in the same hotel Orbitz and Hotels.com would usually have the same price while Agoda would either not offer that room, offer it at the same price, cheaper or more expensive. There is not a specific behavior to look at it here since there is not a clear pattern and there is not enough data to compare.

Since prices between sources are not providing enough insight into what drives the price of an accommodation, let's now compare the data available in each site. For example, prices and hotel ratings. Let's answer the question: **Are higher rated hotels more expensive?**

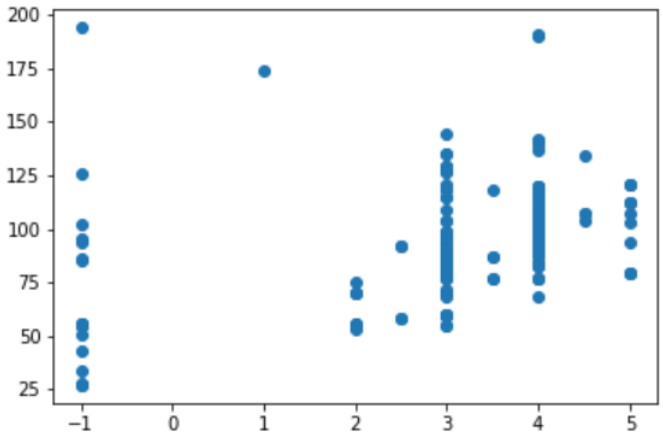

Figure 2: **X: hotel rating, Y: accommodation price**

Hotel ratings with value -1 are missing values which rows were kept in order to use the information for further analysis. The hotel with 1 star and a high price might be consider an outlier. Further exploration of the data might help detect whether this is an outlier or not. It is certainly an odd behavior at least. Apart from that, there seem to be certain tendency of having a more expensive price by having a better rating. However, between 4 and 5 stars there might be outliers or in terms of business context there might not be a big difference between 4 and 5 star hotels, so prices are just similar. This analysis might need a further exploration per source since here we are visualizing all sources together. Is a 5 star accommodation in Agoda cheaper than an 4 star in Hotels.com? That might be the case.

Let's explore prices and reviews. **Are higher reviewed hotels cheaper?** This question differs from the last one since users might prefer cheaper accommodations. Higher rating for a hotel would probably be related to more services and a more luxurious experience.

It seems that higher reviewed hotels are NOT cheaper than those with medium or lower reviews. They actually have a tendency to be more expensive. This plot is clearly skewed by the presence of the "-1" value in the reviews axis. Removing this should give a better representation of the data.

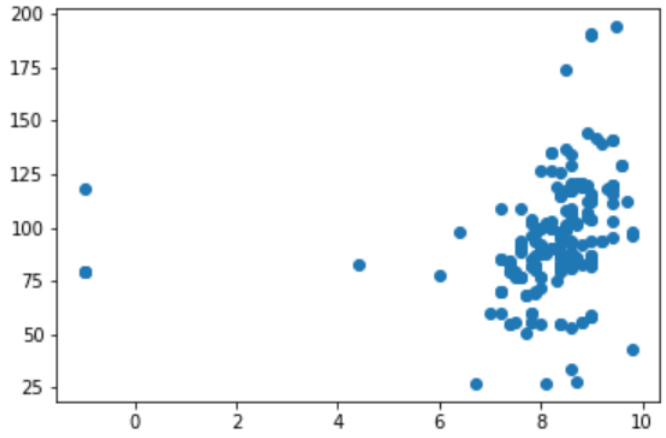

Figure 3: **X: review score, Y: accommodation price**

Finally, let's explore the correlation matrix between accommodation price, hotel rating and review score. This should give us a better idea into what might have a bigger impact or relation with the prices.

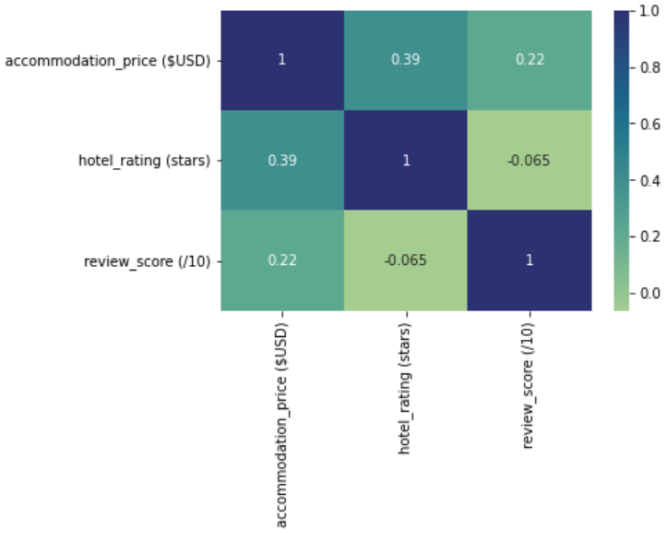

Figure 4: **Correlation matrix** - Price, Rating, Review Score

This correlation matrix shows that there is not a high correlation between hotel rating and review score. This is in part surprising since you would expect these two values to be aligned. On the other hand, the highest correlation seems to be between hotel rating and accommodation price. However, as well known all across the industry "correlation does not imply causation" so with the data available we cannot guarantee that a higher rating leads to a higher price or vice versa.

4 Conclusion

Scraping the web requires understanding how the web works. At first, I was unable to scrap any data from the booking websites, but then I discover my requests were lacking an 'user-agent', which is basically telling a site who is making a request e.g., a browser. When an user-agent is not explicitly set for a scraping task, websites recognize the source of the request as a python script and then they can proceed to block the request.

After solving this problem and being able to send requests to the different websites, a new set of challenges appeared. Each website has its own stack of technologies and its own way to load content. Some are static and some are dynamic. Thus, there is not a single scraping platform or scraping strategy that meets all needs. Simple requests are not enough.

These challenges were solved by focusing on one website at a time and understanding their behaviors. For instance, two of the websites in this report just needed a simple scroll until the end of the page (performed by Selenium) to load all the content while the third website needed an "slower" scroll down so that some elements could be loaded before making a scraping request.

While processing the collected data, integrating the information from the different websites was a must. And although the task did not represent a major obstacle for the development of the report, it took its time, and it goes to show how pre-processing and cleaning data is one of the most time consuming tasks in the data life-cycle.

Finally, Exploratory Data Analysis (EDA) is a tricky, but very interesting process. Some initial visualizations give you insights to formulate some questions and while looking the answers to those questions, new questions appear, some are refined, some are answered, and some are not. However, what was probably the most important part of this process, while also being the most challenging part of it, was to keep in mind and keep on track this question: **What story am I telling?**