A Work Project presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics

Scraping the Web for evidence of Price Discrimination

Tomás Henrique Rijo Mendes

Work project carried out under the supervision of:

Qiwei Han

Fabrizio Esposito

13-12-2022

Acknowledgments: Thank you to Maria for her tireless help and contribution. Also, thank you to Mariana and Pizzi for their support and comfort. To professor Qiwei Han for all the continuous support and availability throughout the whole thesis and to professor Fabrizio Esposito for the valuable input.

Abstract: This paper exhibits a study about the practice of price personalisation on e-commerce websites. With this goal, eleven websites were analysed from three different industries: Fashion retail, General Retail and Travel. Moreover, to find evidence of this practice, the differentiated features tested were the device, operating system (OS), browser, and geolocation of the visiting user. Furthermore, two dimensions of price personalisation were conducted in-depth: price discrimination and price steering. Finally, summarised research about the existing laws and their implications is also presented in this report.

Keywords: Price Personalisation, Price Discrimination, Price Steering, Data Science, Web Scraping, Data analysis, GDPR

Table of Contents

1. **Structure of the paper** – Group Part 3
2. **Introduction** – Group Part 3
3. **Price personalisation** – Group Part 5
   - **3.1 Pricing algorithms** 5
   - **3.2 Price discrimination** 6
   - **3.3 Price steering** 6
4. **Legality behind price personalisation** – Group Part 7
   - **4.1 Brief notes about the history of price personalisation** 7
   - **4.2 Welfare economic effects** 7
   - **4.3 Personal data** 8
   - **4.4 People's behaviour facing price personalisation** 10
   - **4.5 Data protection and transparency** 10
   - **4.6 Unfair Contract Terms Directive** 11
   - **4.7 Portuguese Law** 11
   - **4.8 Overall notes** 11
5. **Methodology for data collection** – Group Part 12
   - **5.1 E-commerce Websites** 13
   - **5.2 User profiles** 14
   - **5.3 Web scraping** 16
   - **5.4 Web scraping for price personalisation** 19
   - **5.5 General methodology** 22
6. **Price Discrimination** 23
   - **6.1 Data cleaning** 23
   - **6.2 Data Analysis and results** 24
   - **6.3 Recommendations for future steps** 34
   - **6.4 Conclusion** 35
7. **Challenges and limitations** 36
8. **References** 37
10. **Annex** 39

1. Structure of the paper

With the goal of searching for evidence of price personalisation on e-commerce websites, the analysis and, subsequently, this paper is presented in the following structure:

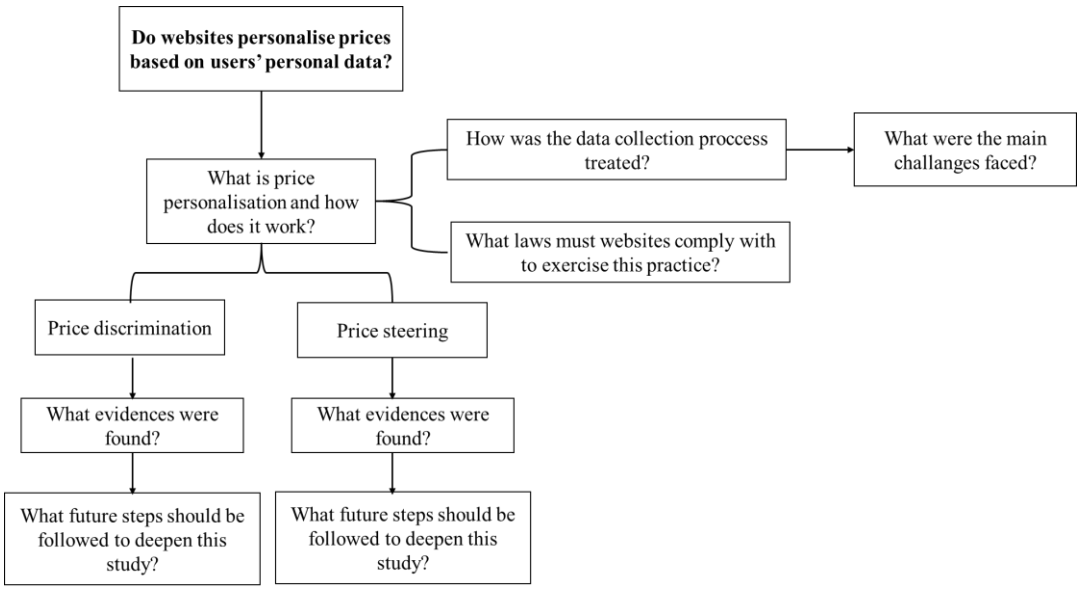

FIGURE 1 - PAPER STRUCTURE

2. Introduction

A good salesperson will say that understanding the customer is one of the most critical steps in a sales strategy. Who they are, what they like, and what drives them. As data has become an increasingly abundant resource, businesses have been developing ways to know their online customers. Tailoring their services to each customer, today's internet is personalised (OECD 2021, 7-9). Personalisation can take many different forms. Spotify, for instance, creates custom weekly playlists with music that matches users' listening habits. Netflix changes the cover art of the entertainment to drive a specific user to a click. The use cases are almost endless.

This paper focuses on a specific type of personalisation, price personalisation, on e-commerce websites. A custom pricing algorithm can be implemented using customer data (e.g. personal

information, search history and browser from which they access a retailer's website) together with big data and artificial intelligence tools. The great push for online personalisation indicates a potential for more extensive adoption of these types of algorithms (OECD 2021, 7-9). Currently, the General Data Protection Regulation (GDPR) clearly states when it is legal to process personal data and the necessity for consent from a data subject (Wolford n.d.). As such, the effectiveness of these rules is also a relevant topic for this study.

This research aims to understand how broad of an issue price personalisation is in the current e-commerce European market. Because websites don't publicly disclose their code, custom scripts that scrape the pricing information for each specific website were set up to look for evidence of personalised pricing. In each website, these scripts mimic different user settings (devices or locations) and store which products were returned and the displayed price. Results obtained for different user profiles are then compared.

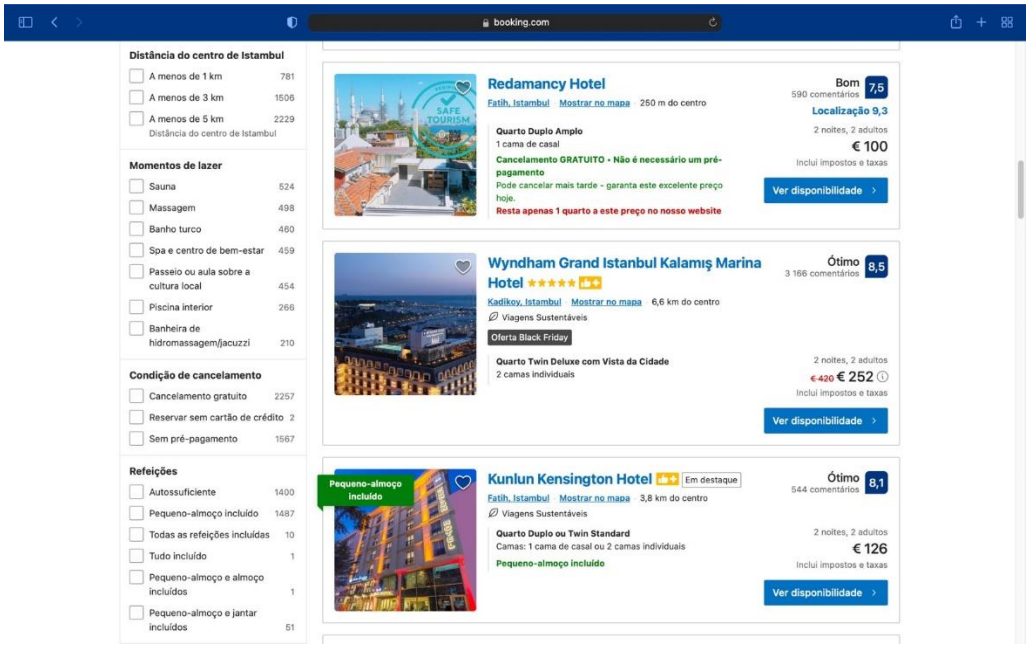

FIGURE 2 - EXAMPLE OF PAGE SCRAPED

This study considers the pricing data for eleven major e-commerce websites from three industries (travelling, apparel, and general retail). Moreover, two different perspectives were investigated – price discrimination and price steering (ranking of offers). Evidence of both was positively identified in some of the explored businesses.

3. Price personalisation

Online personalisation can be defined as the practice of dynamic web experiences that are achieved using data collected from an individual's online activity (European Comission 2018, 30-32). Under the umbrella of which is personalised pricing. A practice where computer algorithms automatically price goods depending on one or more features (UK government 2018) that they were programmed to take as input.

3.1 Pricing algorithms

Dynamic pricing strategies are not a novelty. In the 1990s, Coca-Cola experimented with temperature-sensitive vending machines that would increase products' prices depending on how hot the day was. Online custom pricing can be achieved by implementing pricing algorithms - algorithms that use price as an input (together with other features) and uses a computational procedure to determine the price as an output. These algorithms have been in use for some time with varying complexity (e.g. algorithms that estimate demand or "Match Low Price"). However, there has been a significant increase in data collected by businesses. From 2016 to 2018, over 90% of all data available until that point was created in those two years (UK government 2018). This fact makes the implementation of even more complex pricing algorithms a possible reality. "It is not the algorithmic practice that has significantly changed, although new and more efficient algorithms are invented all the time. It is the data that made the major impact here" (Gal 2017, 3).

As pricing algorithms become more prevalent in the market it is expected that, potentialized by the growing amounts of data, more and more cases of price personalisation will emerge (OECD 2021, 7-9). This paper explores two distinct angles: price discrimination and price steering.

3.2 Price discrimination

Price discrimination can be defined as a customized pricing strategy that charges customers different prices for the same product or service based on what the seller believes the customer will accept (Twin 2022). Online price discrimination is possible since e-commerce websites can leverage the available large size of consumers' data to estimate their willingness to pay (OECD 2021, 7-9).

Discriminatory pricing has been implemented across many industries and can take different forms. Some cinemas offer senior discounts for movie tickets. Spotify has a student plan that cuts the cost of a monthly subscription in half. In these cases, the customer decides freely and out of their conscience to knowingly exchange a limited part of their personal information for different, better pricing. The seniors share their ages with the cinema, and the students share their occupations (and academic information).

3.3 Price steering

Price steering, or personalised ranking of offers, refers to changing the order of search results for the same product based on processing data collected from a customer's online activity.

It is a subject to which online consumers are no strangers. When googling "restaurants", we receive, at the top results, options that are near us. We acknowledge that Google's algorithm uses our personal information (location) to return a list of results that will increase our likelihood of clicking. In many cases, personalised offers are not so "customer friendly". In 2011, The Wall Street Journal

found that Orbitz (the American travelling aggregator) steered customers who use apple products towards different and costlier options (Mattioli 2012).

4. Legality behind price personalisation

4.1 Brief notes about the history of price personalisation

The personalisation of prices by sellers has always been present in the economy. The goal of a seller is to extract the maximum willingness to pay from each customer. Therefore, it is in their interest that the price of a good may differ from customer to customer.

A prominent example of price personalisation is in fairs, where different people may buy the same product for a different price. This means that the seller tries to evaluate and takes advantage of the maximum price a buyer is willing to pay. For this, aspects like the image of a person and the way they talk may have an impact.

Finally, as internet use has been increasing at lightning speed, with close to 5 billion users (Pasquali 2022), this concept of price personalisation has also been adopted there. This represents an advantage for sellers with the automated collection of users' data but also presents a challenge regarding its protection and ethical use.

4.2 Welfare economic effects

In general terms, price personalisation benefits the economy as a whole. On the one hand, sellers can extract the whole consumer surplus by charging each customer or group of customers their maximum willingness to pay. On the other hand, it can also benefit customers since a person willing to pay lower than the uniform price may buy the good at a personalised price, lower than the uniform one and equal to their willingness to pay.

Therefore, price personalisation reduces deadweight losses that exist when all customers are charged the same price. However, it cannot be neglected that some customers, the ones willing to pay higher than the uniform price, turn out to pay a higher price, becoming individually harmed (Zuiderveen Borgesius, Poort 2017).

Finally, price personalisation encourages market power. This happens because a monopolist may capture the whole market by being able to charge different prices. This, consequently, makes the entrance unattractive for new competitors.

4.3 Personal data

For websites to be able to personalise prices for a specific user, it has to have access to information about the same, i.e., personal data.

According to the GDPR, Chapter 1, article 4, personal data shall mean "any information relating to an identified or identifiable natural person", and this data shall include information such as "name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (European Union 2016, L119, p. 35).

Firstly, the data must be collected by the website in question or by a third party, then it is processed, and finally, it is ready to be used for price personalisation. With this being said, it can be concluded that online users are constantly being monitored and investigated and, consequently, treated differently (Christl 2017).

4.3.1 Sources of data:

Different sources can collect personal data, such as voluntarily and involuntarily, knowingly and unknowingly, and, finally, by a third party (Zuiderveen Borgesius, Poort 2017).

When a customer creates an account on a website, they know that whenever they log on, the website has access to their past purchases and the personal data they provided when creating the account. In this case, the customer voluntarily provides and knowingly their data.

Alternatively, a customer can provide information involuntarily and unknowingly. This happens, for instance, when a website takes advantage of the IP address, browser, or cookie with a unique identifier of a specific user.

At last, personal data can be obtained by a third party. By the third party, it is meant another website or entity that is not the one in the matter. This can be an advertising network, for example. Using cookies, these entities can create a profile for a specific user, permitting a website to personalise prices for them. Fundamentally, cookies are files which record information such as personal settings of users and interactions with a specific website. Then it can be used by a website to identify a user and provide a customized experience.

4.3.2 Processing personal data

Data protection law ensures that the processing of personal data happens transparently, i.e. with customers' consent. According to the GDPR, personal data must be "processed lawfully, fairly and in a transparent manner about the data subject" (European Union 2016, L119, p. 37). Additionally, in Chapter 2, Article 7, it is reinforced that this process "is based on consent" and "the controller shall be able to demonstrate that the data subject has consented to the processing of his or her personal data" (European Union 2016, L119, p. 39).

Therefore, data protection law grants both rights to users, those from whom the data is collected, and obligations to the websites (controller), those that collect the data.

4.4 People's behaviour facing price personalisation

According to a survey conducted in the US in 2016, 72% of the participants considered that practices such as price discrimination should be prohibited (Poort 2019). Additionally, more than 80% think that it is unfair and unacceptable.

Finally, about 80% of the inquired agree that they must be informed if they are on a website that personalises prices (Poort 2019).

Furthermore, in 2000, Amazon experimented with price discrimination. It turned out that different loyal users realised that once they removed any signs of being regular Amazon customers from their computers, they were given better deals. Consequently, the practice stopped as soon as the complaints began (Streitfeld 2000).

It can be concluded that people generally do not feel comfortable or enthusiastic about buying on a website that states the practice of price personalisation.

4.5 Data protection and transparency

Article 13 of the GDPR presents a list of the information the website must provide. This includes, for instance, "the purposes of the processing for which the personal data are intended and the period for which the personal data will be stored" (European Union 2016, L119, p. 43). When a website practices price personalisation, it must inform each customer explicitly.

Moreover, to comply with Data Protection law, it is not enough for websites to have statements like "we use personal data to offer our customers better-personalized services". This would benefit the website since the user would not have complete knowledge and may end up staying on the website besides the practice of price personalisation.

Overall, transparency could mitigate the information asymmetry that exists. Customers could choose a website that does not personalise prices or, if it benefits them, could delete cookies (Zuiderveen Borgesius, Poort 2017).

4.6 Unfair Contract Terms Directive

The Unfair Contract Terms Directive "protects consumers against unfair terms in all types of business-to-consumer contracts" (European Union 2019, C232, p.7). This shall be applied to all contracts for purchasing goods or services, including financial services and e-commerce or offline commerce. In addition, it states the conditions that make an agreement not fair.

Regarding this topic, Article 4(2) of the Directive mentioned above must be considered. This Article states, in simple words, that if a contract operates with transparency, it complies with the requirements to not be regarded as unfair. In other words, for a contract to not fall into unfair conditions, it must present the price practices in a "plain intelligible language" (European Union 2019, C232, p. 21).

4.7 Portuguese Law

The framework in which this topic falls in Portuguese law is known as *Regime das Cláusulas Contratuais Gerais.* This, combined with the *Lei da Defesa do Consumidor*, define and governs the fairness of contracts and the transparency in the seller-buyer relationship (PGDL n.d.).

4.8 Overall notes

To be by the existing law, websites must inform their customers if they collect and use their personal data. Moreover, they must explicitly notify the purpose of it. This means that a website that personalises prices must unequivocally inform the buyer that it is presenting prices that were explicitly personalised for that user based on their data.

All these legal requirements may not be attractive for sellers since customers are uncomfortable and do not find this practice acceptable. They even may end up purchasing on another website because they do not think this practice is fair and feel disadvantaged.

5. Methodology for data collection

As mentioned before, this paper aims to understand how standard pricing personalisation practices in the European market are. To do so, it is necessary to analyse the results that users with different settings receive, on an e-commerce website, when searching for the same product. Several web scraping techniques were relied upon to get to this quantitative data. Web scraping is automatically extracting data available on the internet (European Commission 2018). Most commonly, web scrapers download the code that composes the visited page which in turn, contains the embedded data. Many services rely on web scrapers. Sporting results pages, stock evaluation services, and price comparisons between competitors are examples, and web scraping can be used in several more use cases.

In this paper, for each product search in each website scraped, different iterations of the same query were run, changing the settings of the user who made the request. Settings that mimic other user profiles that were previously set up. Before getting into the approaches taken and how these different settings were implemented, exploring the design choices taken in this research is crucial.

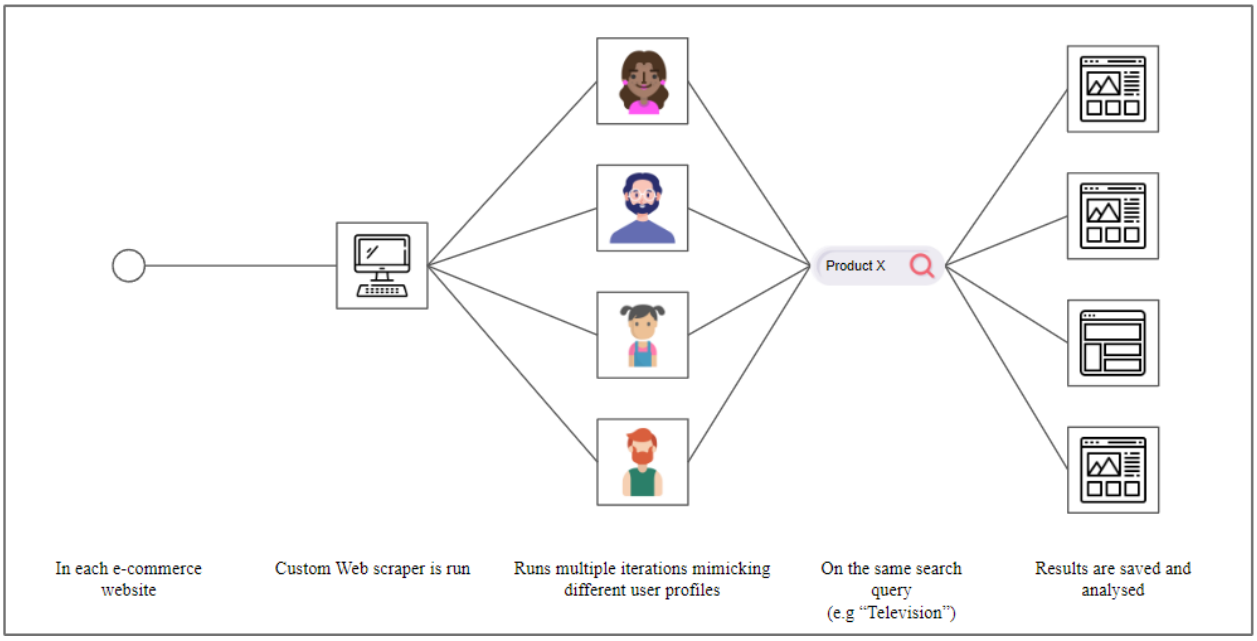

FIGURE 3 - OVERVIEW OF THE DATA COLLECTION PROCESS

5.1 E-commerce Websites

As objects for this study, eleven e-commerce websites were chosen. The criteria for the selection of these websites differ. Firstly, general retailers and fashion retailers specifically. These two industries were identified as highly detrimental to the customer (European Commission 2017). Also, the travelling industry has been found to have performed price personalisation in other related studies (Hannak, Mislove, Wilson, Lazer, and Soeller 2014). Then, the businesses selected must be of considerable size in the European/Portuguese market. Larger companies have access to more data and the infrastructure to implement complex algorithms. These are the e-commerce websites that are going to be investigated when looking for price personalisation.

			Tested characteristics			
Industry	**Website name**	**Website URL**	**Device**	**OS**	**Browser**	**Location**
Fashion Retail	Nike	https://www.nike.com/pt/	X	X	X	X
	Adidas	https://www.adidas.pt/	X	X	X	X
	New Balance	https://www.newbalance.pt/	X	X	X	X
	Farfetch	https://www.farfetch.com/pt	X	X	X	X
General Retail	Worten	https://www.worten.pt/	X	X	X	X
	Decathlon	https://www.decathlon.pt/	X	X	X	X
	Pingo Doce	https://mercadao.pt/store/pingo-doce	X	X	X	
Travel	Booking	https://www.booking.com/	X	X	X	
	Expedia	https://www.expedia.com	X	X	X	X
	Tripadvisor	https://www.tripadvisor.com/	X	X	X	
	Airbnb	https://www.airbnb.com	X	X	X	

TABLE 1 - WEBSITES INVESTIGATED AND USER CHARACTERISTICS TESTED

5.2 User profiles

The selection of the user settings that this paper aims to test is a crucial step in this process. These features will be what the existing (or not) personalisation algorithm receives as input. If no evidence for personalisation were to be found, it only means that no pricing algorithm is based on the selected features. Personalisation could still exist but be based on characteristics not tested in this paper.

Similar research on the effects of online price personalisation has taken place, and although the "evidence of the prevalence of online personalised pricing remains sparse" (OECD 2021, 7) when pricing variations exist, they are mostly found when changing location or browser (European Comission 2018, 40-46). Based on the results from those previous studies, the selected user settings tested in this paper are device, operating system (OS), browser, and geolocation.

5.2.1 Device, OS, and browser profiles

In the context of this paper, it is essential to differentiate mobile from non-mobile devices. Mobile devices are portable computing devices (i.e. smartphones, tablets) and non-mobile devices (or desktop devices) are customer-owned personal computers (i.e. laptops, desktop computers). Moreover, an Operating system is a software that acts as an intermediator between the user and the device hardware, for example, Windows (University of Wollongong n.d). Finally, the browser is a piece of software that "retrieves information from other parts of the web and displays it on your desktop or mobile device" (Mozilla 2022). It connects the user to the internet, for example, Firefox.

The choice for the device-browser combinations is based on what are the most used in today's market. Using standard devices, operating systems and browser profiles increases the chances that an e-commerce website has tailored a pricing algorithm based on those (or part of those) specific characteristics. Nowadays, online consumers use mobile devices as much as desktop machines when visiting online e-commerce websites (Smith 2022). However, the conversion rate for mobile visitors is almost half the rate of desktop users (Jehanne 2022). As such, online businesses could see custom pricing models targeted at mobile customers as an opportunity to increase their likelihood of buying.

Seven different device-OS-browser combinations were selected, which cover both mobile and non-mobile devices. For desktop devices, 90% of the non-mobile devices used today are running either Windows or macOS. Android and IOS cover more than 98% of the operating systems market for mobile devices. The browsers chosen represent over 95% of users (gs.statcounter 2022). Besides the individual market share, the combinations' commonness was considered. For example, Firefox on mobile devices has a market share of less than 1% in 2022. As such, it is not tested in this research.

	Operating System	Browser
Mobile devices	Android	Google Chrome
	IOS	Safari
		Google Chrome
Desktop devices	macOS	Safari
	Windows	Google Chrome
		Microsoft Edge
		Firefox

TABLE 2 - DEVICE, OS AND BROWSER USER PROFILES TESTED

The selected profiles allow for the exploration of possible personalisation focused on specific characteristics like the browser used or on a more general approach, for example, comparing mobile to non-mobile devices.

5.2.2 Geolocation

The user's geolocation while navigating e-commerce sites has been found to be a factor that some businesses use for price personalisation (OECD 2021, 7-9). As such, it is also a user setting tested on this paper. Since this research focuses on the European market, four geolocations and European countries have been selected. These are **Portugal, Spain, the Netherlands, and the United Kingdom** region. The selected regions cover a sparse area of Europe and allow for the exploration of price personalisation practices in regions with varying purchasing power.

5.3 Web scraping

Understanding how web scraping works is essential to acknowledge the approaches taken to answer the proposed challenge. Web scrapers are versatile and can take many forms, but their ultimate function is clear: To extract content from internet pages. Before comprehending web

scrapers, another step back must be taken. Because before pulling the content from a web page, one must understand how that content is retrieved in the first place.

The web has immense amounts of websites and information. Websites are hosted on servers containing the data and information seen when visiting an internet location. When accessing a website, on the back end, the client (in this case, the scraper/ browser) is requesting that the host (website server) returns, as a response, the resources needed to build the content up. The request-response dynamic occurs under the Transmission Control Protocol (TCP), which permits the exchange of resources between the two. HTTP – Hypertext Transfer Protocol – is how the server and the client communicate with each other when these transactions of information occur (sematext n.d). The requests are known as HTTP requests. After an HTTP request, the server will return an HTTP response. If the request is successfully validated, the response will return, as a result, the resources requested by the client (IBM 2021). The resources are then translated by the browser to a visual form and presented to the user as we are familiarised. When web scraping, these resources aren't usually transformed into visual assets. The focus is mainly on the data embedded into the page's response.

5.3.1 HTTP requests

HTTP requests play an essential role in this paper. They allow for defining different user characteristics for the user profiles described previously. As such, it is crucial to understand how that can be achieved and, consequently, what composes an HTTP Request. An HTTP request comprises three elements: A request line, HTTP headers, and a message body (if needed).

The request line is composed of three other elements – an HTTP method, URL, and the HTTP version number being used. The HTTP method indicates to the server the action to be taken. For

instance, the GET method is used to retrieve data, and DELETE is used to delete a particular resource on the server. The URL serves as a path to the resources the client wishes to access.

The header provides the server with information about the message and the client itself. It permits adding information about the client such as cookies, and user agent string. The user agent string contains details about the client, allowing the server to identify **the device, operating system, and browser when making a request**. Lastly, the server uses the message body to return the response back to the client. In cases where the client wants to use methods that add to the server's resources, also described in the message body.

This paper uses the header to mimic different user profiles when scraping e-commerce websites. The illusion that a user with a specific, pre-chosen characteristic is visiting the website is enabled by the customisation of the header before making the request. This works for both different geolocations and different device-OS-browser combinations.

Different geolocations are accomplished using IP addresses from other countries. IP address stands for Internet Protocol Address. It is associated with a specific computer network that allows for transferring information by connecting the client to the server when connected to the internet. When making a request, the server receives the client's IP address so both devices can communicate. The server can then receive the request and send back the appropriate response. Once the IP is shared, the e-commerce website can trace the user's location in real-time.

The understanding of whether price personalisation exists will be made based on whether the response from the website is different for requests that only differ in the user-agent string or IP address passed.

5.3.2 Creating the user profiles

To mimic the different device-OS-browser combinations, different header settings are used. For each combination selected, a profile was created. In each, three user agent elements are changed.

- Sec-ch-ua-mobile: A flag which is either “1” if the device is mobile or “0” if non-mobile.
- Sec-ch-ua-platform: Represents the OS. It can be, for example, “Windows” or “macOS”.
- User-agent: Represents the Device/Os and browser used.

To simulate different geolocations, a service called smartproxy was used. Smartproxy provides IP addresses from other locations.

5.4 Web scraping for price personalisation

It is necessary to extract the pricing data for different products to understand if personalised pricing practices are in motion on each website. The data necessary is a product identifier and product price for different user profiles on the same search query.

There are many options for web scrapers. Pre-built scrapers or browser extensions are available online that help the average user extract online content. For this paper, specialised tools were needed. The websites where personalised pricing practices were investigated belonged to different industries and were built using diverse techniques. Also, there was a need for scrapers that allow for the passing of custom HTTP request headers. With these requisites in mind, custom web scrapers were created for each website.

For this search, the scrapers were developed in Python. Python is a high-level programming language widely used for data science, machine learning and data analysis. It has grown in popularity over the last few years and is now the second-most-used programming language in the

world. Many Python libraries have been made available by the community. These libraries, collections of existing code, can be leveraged when implementing new web scraping solutions.

Because the chosen websites are built in different ways, no solution fits all. Therefore, three different scraping techniques were used.

5.4.1 Direct request methods

Using the *requests* Python library, it is possible to directly request the website for the resources available on a page using a specific URL. Because the URLs often have information on the query made by the user (for example, "*https://www.worten.pt/search?query=iphone*"), this approach allows for the retrieval of information using different URLs for different products. Such a method works for websites where the data is embedded directly into the page's HTML code – static websites. HTML – Hypertext Markup Language – is a markup language that tells the browser how to logically translate the HTTP responses' content into being displayed in a web browser.

The output of scraping a website using this method is the HTML code for the page with the specified URL.

5.4.2 Direct request to the API

Not all websites, especially in 2022, have the data directly embedded into the HTML code. Most complex websites have functions that perform different tasks on the website itself. These dynamic websites have, commonly, JavaScript code. JavaScript is a programming language used by most websites to handle a page's behaviour on the client's side (for example, on what clicking a button does). Dynamic websites cannot be scraped using direct requests. The returned HTML code does not contain data but the JavaScript code on the page. This happens because the page hasn't had time to load its content.

On dynamic websites, the approach followed was to take advantage of the page's API use. APIs - Application Programming Interface - allow two software components to communicate with each other using a specific set of settings. Moreover, they perform tasks between the client and the server. The APIs leveraged were the ones that are responsible for the fetching of data. A dynamic e-commerce website where a user searches for "iPhone" will call for the API responsible for retrieving the information on "iPhone", and the API shall return the data that is then displayed.

The network activity (where the Requests activity is logged in a browser) was analysed to find the API responsible for returning the data. Once the appropriate API was found, using software called *Insomnia* allowed for a better understanding of the API and how to manipulate it. For example, to return data equivalent to two pages of results, not just one. *Insomnia* also facilitates the translation of the API request into Python code.

Using this translation, for most dynamic websites, the approach was to make requests directly to the API. Dodging the need for loading the page and going directly to the software responsible for fetching the data. The output of this approach was the HTML code for some pages with the data embedded in them and, for others, the data in an already structured format ready for analysis.

5.4.3 Puppeteering a browser

On some dynamic websites making the request directly to the API does not work. Either it can't be found, or extra layers of security exist, such as not allowing similar requests to be made from the same client side. Or, in the travelling industry, it is common for websites to redirect the user to another page with the requested information.

In these cases, the *Playwright* library was used. *Playwright*, initially designed for testing feature testing, allows for the launching of browser instances that will follow specific instructions on what steps to take next using HTML selectors. In a website built using HTML, each element will have

a unique identifier - element selectors. Examples of instructions given are clicking a specific button, scrolling a particular number of pixels, or filling an input box.

To retrieve pricing data, the browsers were instructed to search for a specific product and then extract the HTML content present in the newly loaded page, which is the final output for this approach.

The service used for achieving different geolocations was not compatible with *Playwright*. As such, Booking.com, Adidas, TripAdvisor and Airbnb were not tested on whether user location impacts price personalisation.

5.4.4 Extracting the content from the HTML code

After scraping the HTML code, its content is extracted using the *beautifulsoup4* library. It facilitates the navigation of HTML code and easier retrieval of the desired data. For this paper, the product identifier and product name were required. This data was extracted using *beautifulsoup4* and the HTML selector for where this data is embedded on the pages HTML code.

5.5 General methodology

For all websites, the general methodology is the same. A page is to be scrapped on a website with no purchase history, no user logged in and complete acceptance of cookies permission requests. All iterations for the same website and product were done in the same conditions, apart from the variable that is to be tested.

A minimum of four product queries searched was established per website. The pricing data extracted is restricted to the first and second page of the results. When scraping a website, both variables are used at different times. The website is first scraped using the different device-OS-browser profiles and then the different geolocations, keeping the other variable fixed throughout

the process. When the content is scraped, in case it is HTML code, it is parsed to extract a product identifier and the lowest shown product price.

For each website and product searched, iterations of the same request are made with the different profiles. The result of each iteration is then saved on a CSV file for future analysis.

6. Price Discrimination

Once the collection of data is complete for every website, it is possible to analyse that data and search for evidence of price personalisation. This part of the paper focuses specifically on Price Discrimination.

Price discrimination refers to changing prices according to a user's characteristics. The tested features are device, operating system, browser, and user geolocation when searching for a product. To understand if e-commerce websites have discriminatory pricing practices, the same exact search was done at different times, changing only the characteristics of the user. With the goal of quantifying price discrimination on e-commerce websites, many angles are investigated. Specifically, how standard price discrimination is on each website and how the different variables impact pricing.

6.1 Data cleaning

Before the analysis, data must be cleaned. Removing faulty entries, guaranteeing consistency, and preparing the data to be used. The data obtained from web scraping comes in one file per search iteration – each file contains the results (identifier and price) from each request. After retrieving the data from the web pages, an extra step was taken. Both product identifiers and prices were formatted using uniform criteria. For identifiers, double spaces were removed. For product prices,

spaces and any currency indicators were erased. The product cost was also saved as numeric variables, allowing statistical metrics to be calculated.

Two different tables were created for each website. One for the user-agent strings (with information on the device, operating system, and browser) results and another for the different proxy geolocations. Each row represents a unique product and the prices found on each characteristic.

search_term	Product_name	Product_price_Android-Chrome	Product_price_Ios-Chrome	Product_price_Ios-Safari	Product_price_Mac-Safari	Product_price_Windows-Chrome	Product_price_Windows-Edge	Product_price_Windows-Firefox
iphone	iPhone 14 APPLE (6.1" - 128 GB - Luz das estr...	1032	1032	1032	1032	1032	1032	1032
iphone	iPhone 13 APPLE (6.1" - 128 GB - Meia-noite)	929	929	929	929	929	929	929
iphone	iPhone 11 APPLE (6.1" - 64 GB - Preto)	533	533	533	533	533	533	533
iphone	iPhone 14 APPLE (6.1" - 128 GB - Meia-noite)	1039	1039	1039	1039	1039	1039	1039
iphone	iPhone 11 APPLE (6.1" - 128 GB - Preto)	579	579	579	579	579	579	579
iphone	iPhone 11 APPLE (6.1" - 128 GB - Branco)	588	588	588	588	588	588	588
iphone	iPhone 14 Pro APPLE (6.1" - 256 GB - Roxo Esc...	1399	1399	1399	1399	1399	1399	1399
iphone	iPhone 14 Pro APPLE (6.1" - 128 GB - Roxo Esc...	1299	1299	1299	1299	1299	1299	1299
iphone	iPhone 13 Pro APPLE (6.1" - 256 GB - Azul Sie...	1148	1148	1148	1148	1148	1148	1148
iphone	iPhone 13 APPLE (6.1" - 128 GB - Luz das estr...	839	839	839	839	839	839	839

Figure 4 - Data structure example (Worten)

In these tables, not all rows are complete. Many websites show different results to different users - for similar searches, the website returns a different set of products. There is no price for some products on specific variables in these cases. For a product to be available for conclusions on price discrimination, a minimum of two available prices must exist for two sets of user settings.

6.2 Data Analysis and results

6.2.1 Products with price differences

Once the data is clean and ready for analysis, the first question is how standard different prices for the same product are. The standard deviation on the prices for each product was calculated to answer it. The standard deviation is a measure that represents the amount of variation in comparison to the mean for a given sample.

$$S_N = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})}$$

Where N is the total number of observations, $\{x_0, x_1, \ldots x_n\}$ the observed values and $\bar{x}$ the mean value for the sample. The standard deviation is the appropriate measure to check for price differences. It detects any discrepancy between prices, and for cases where the price is only available for one iteration, the *pandas* (Python library used) standard deviation function returns a null value. This way, not only are differences captured, but at the same time, products without enough information to make conclusions are filtered out. Products with a standard deviation different from zero (with different prices) are flagged as products where price discrimination is found. To quantify the prevalence of price discrimination on the various websites, the percentage of products with a standard deviation different than zero was calculated concerning the total number of products where the standard deviation was quantifiable (other than null). A website is considered to show evidence of price discrimination if over 5% of products have any differences in prices across the varying characteristics. This threshold guarantees that noise values are not significant enough to mislabel a website as being responsible for discriminatory pricing.

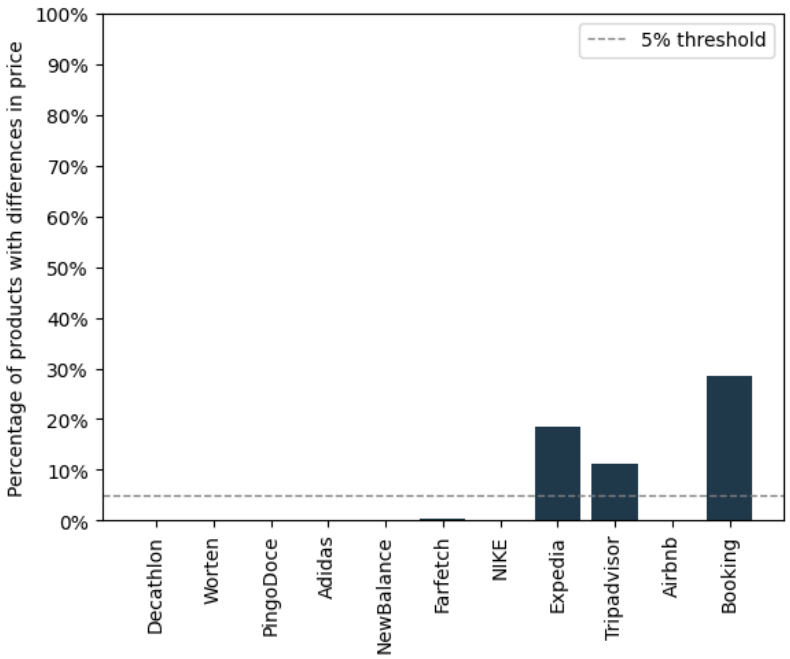

Figure 5 - Proportion of products with differences in price based on users device-OS-browser, per website

Only three websites show price differences when comparing prices received from varying user-agent strings. Expedia, TripAdvisor, and Booking.com have more than the defined 5% threshold for the number of products shown at different price points. TripAdvisor shows the smallest percentage of products with differences, with nearly 12% of products. Expedia, with almost 19% and Booking.com, with about 30% of products showing a discrepancy between prices. All three businesses are in the travelling industry. Industry in which only one out of the four investigated e-commerce websites sees no evidence of personalised pricing - Airbnb.

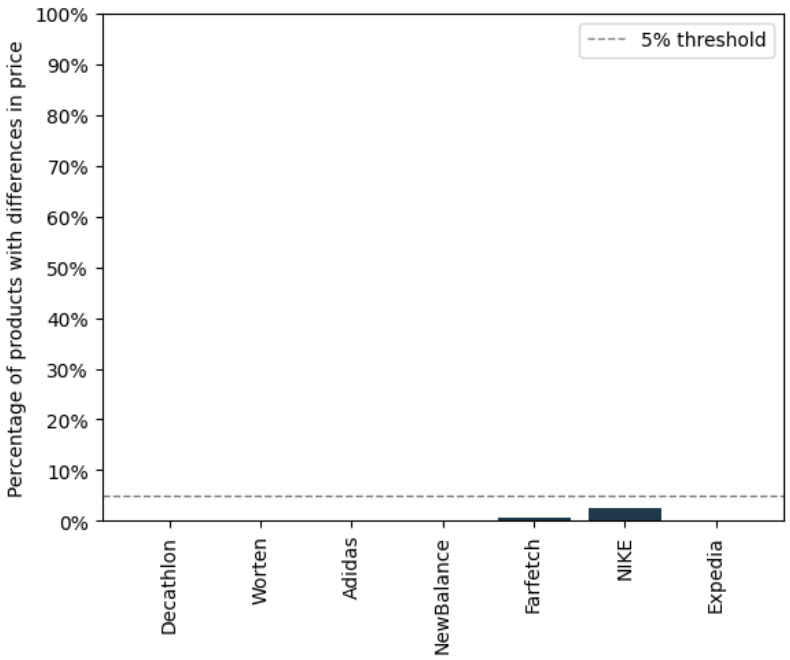

FIGURE 6 - PROPORTION OF PRODUCTS WITH DIFFERENCES IN PRICES LOCATION, PER WEBSITE

When the variable is geo-location, not one investigated website shows different prices to users in Portugal, Spain, the Netherlands, or countries in the United Kingdom. As such, no evidence for the existence of pricing algorithms that receive as input the geolocation of a user when searching for a product is found. Because no price discrimination based on location is seen, no further analysis is done related to that variable. All analysis from now on is based on the data obtained when iterating over the Device - Operating system - Browser combinations.

6.2.2 Quantifying price differences

Even though the e-commerce websites that were found to have pricing differences belong to the same industry, their websites may operate differently. When trying to understand how the different pricing algorithms work, an individual conclusion for each website must be drawn. The same input can have diverging levels of impact on the output prices of each website.

Price differences were found in three out of the eleven investigated websites. Once it was quantifiable how many products had different prices according to various search settings, the next goal was to understand how large of an impact the difference is. How spread is the variance between prices?

As an initial step, the average price for the products in each user profile was calculated. The average price is calculated using all products with an available price. This view allows for an initial overview of the behaviour of the pricing algorithm.

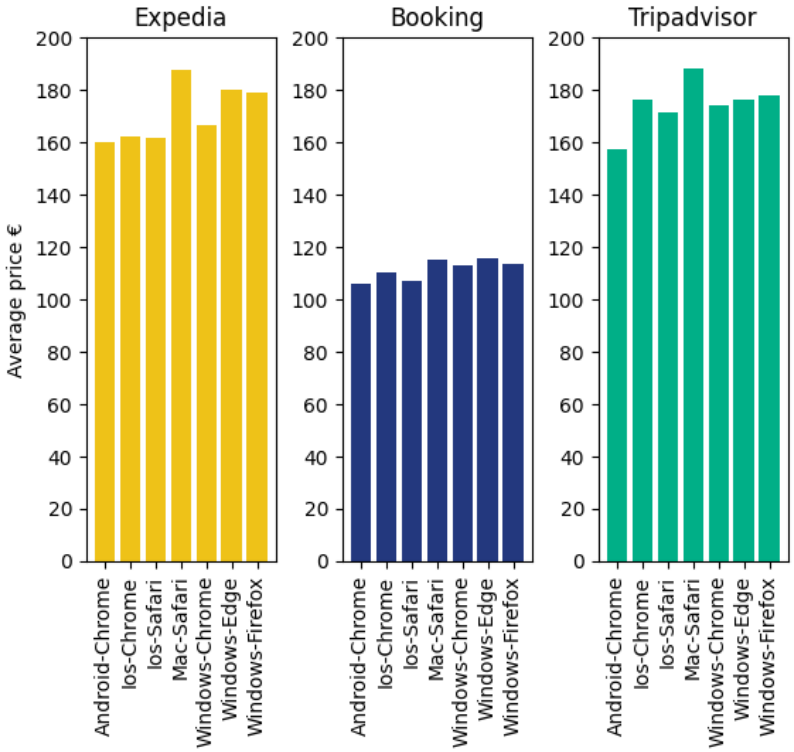

Figure 7 – Average product price found for each user profile, per website

A first look at the results shows that websites behave differently. Booking.com, although having the most significant number of products with different prices, shows the smallest amount of price variation. On average users with desktop devices seem to receive results with higher prices, but no clear pattern can be observed. However, on TripAdvisor and Expedia, users with desktop devices that run macOS and use Safari seem to be driven to higher costing options. TripAdvisor shows no apparent pattern. Expedia also appears to display more expensive options to desktop users overall. Both show higher price variations between the different user settings compared to Booking.com.

The drawback from a view like this is that the differences seen may not be caused by different prices for the same product but by showing other products altogether (price steering). To exclude the latter, it is crucial to complement this perspective with another view, one that focuses exclusively on the variance caused by custom pricing of the same products.

Since the objective is to understand how dispersed the differences in each website are, a comparison must be drawn between the different variants. This comparison is made against a baseline previously defined. The baseline is constructed in the following way:

- For each product, the initial comparison is between each user variable and the price seen on the most common user profile settings. That is desktop devices running Windows and using Google Chrome
- If a product is not shown on the first pages for searches done using the Windows (desktop device) - Google Chrome combination, the baseline considered is the price seen when searching as a mobile, Android device using Google Chrome
- If a product is not part of the results for either Windows - Google Chrome or Android – Google Chrome users, the baseline takes the value of the mean for the available prices

The baseline is calculated for all products previously flagged with signs of price variance (using the standard deviation). For each user combination, the cost seen is subtracted from the baseline, and the absolute value from that difference is taken. To understand if a website promotes a more aggressive approach to discriminatory pricing, the maximum variance found between settings for each product is normalised by the baseline value and saved.

Normalising the variation helps reduce the noise caused by products with naturally different prices. A more expensive hotel can have a smaller price difference proportionally that will not be considered as such unless normalised. Also, as seen in figure 7, each website promotes products of varying price ranges. Booking shows products in the 100 - 120 monetary units' range, while Expedia and TripAdvisor products are in the 140 – 160 monetary units price range. Normalising the differences enables a fairer comparison between websites as well.

A viable way to understand how severe the discrimination is and have a comparison between websites is to analyse the distribution of the normalised maximum difference for each product.

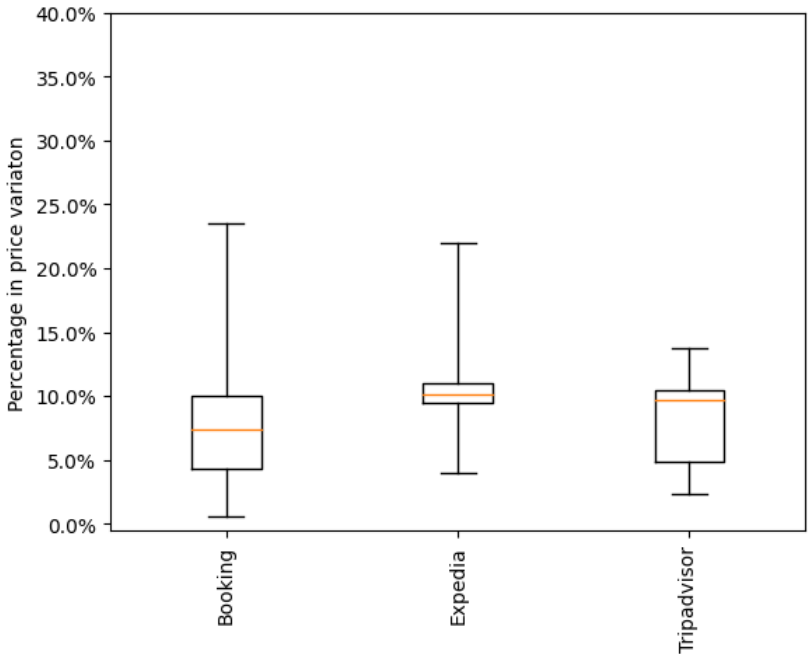

FIGURE 8 - PRICE VARIATION ACROSS USER-PROFILES, PER WEBSITE

Booking.com has the most disperse distribution. With a median value of 7.5%, the middle 50% of values are in the 5-10% range. The products between the 75th quantile and the 95th show a much more comprehensive range of variation values, between 10% and 24%. Expedia shows the highest percentage of price variation with a median value of 10%. Also, it offers little dispersion. The middle 50% of values are in the interval from 9% to 12%. TripAdvisor also shows a higher median percentage of variation, around the 10% mark. It is the website with the most compact range of values. The 5th quantile and the 95th are between 2.5% and 14%. Comparing the three websites, Expedia shows that its custom pricing algorithm is the most aggressive – more focused with the highest median variation value. Booking.com, on the other hand, offers the lowest median price variation with the most extensive range. Variation values can be found in both the lowest and highest sides of the spectrum.

6.2.3 Quantifying differences between variables

The last overview allowed an understanding of how deep the algorithms can go and quantify the differences in the price displayed. The next question to be asked is what causes these differences. Are websites discriminating based on specific user profiles or broader characteristics like mobile or desktop devices? To answer this question, three views were created for each website.

- **Mobile profiles**: The prices for mobile user profiles are compared to those seen in the mobile market leader profile – Android devices with Google Chrome browser are the baseline.
- **Desktop profiles**: The prices for desktop user profiles are compared to those seen in the desktop market leader profile – a Windows device with Google Chrome browser is the baseline.

- **Mobile vs desktop profiles**: The average prices for mobile user profiles are compared to the average prices seen in the Desktop devices – A product must have available prices in both device types to be considered.

Variation is calculated as the normalised difference between the prices seen.

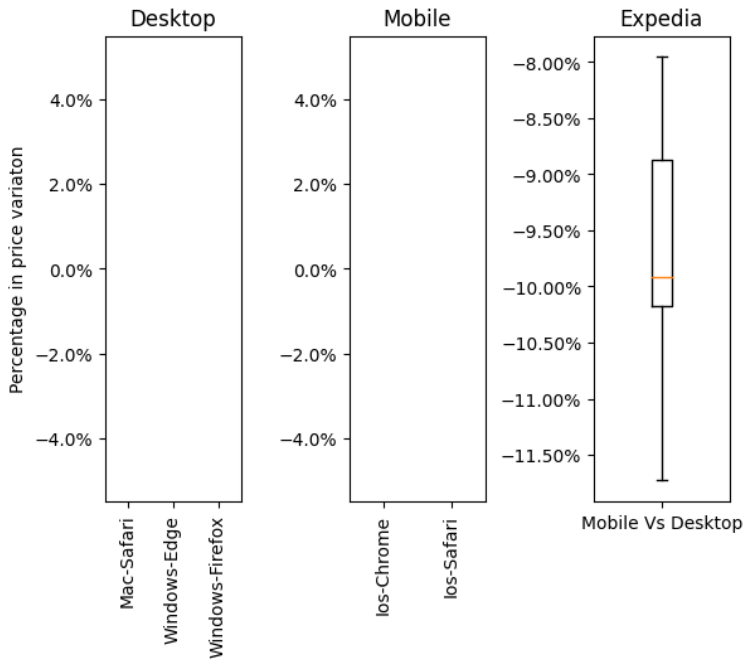

FIGURE 9 - VARIATION ON EACH USER PROFILE, EXPEDIA

In **Expedia**'s results, no significant difference is found between the profiles on both desktop and mobile. However, mobile prices are lower when comparing the average mobile price with the average desktop price, with a median value of 10% lower. Expedia performs discrimination based on the type of device being used - with desktop devices being shown higher prices for the same product.

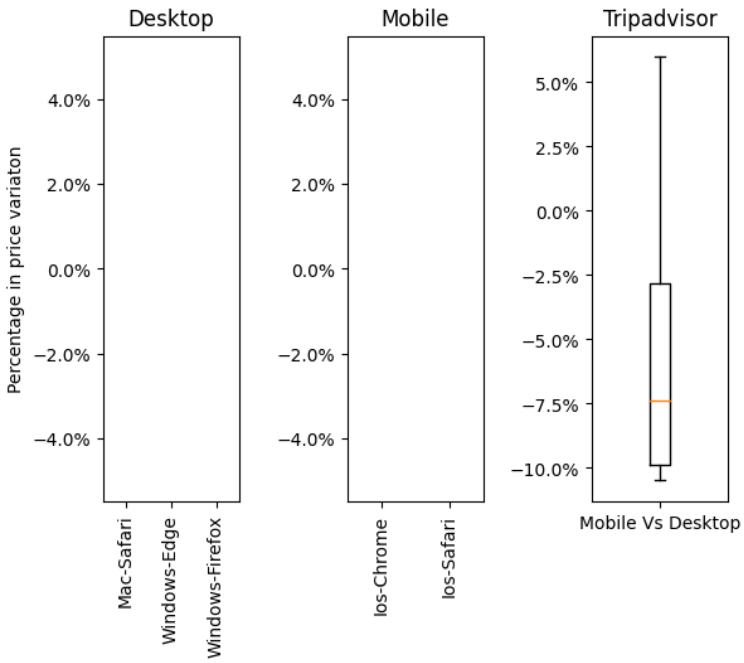

FIGURE 10 - VARIATION ON EACH USER PROFILE, TRIPADVISOR

Also, no significant difference is found in the **TripAdvisor** comparison between Desktop and Mobile profiles. When comparing the average prices in each segment of devices, mobile users show that they receive lower prices for the same products - with a median price difference of 7.5% smaller. In some, less common cases, desktop devices see lower prices than mobile. TripAdvisor displays evidence of discriminating prices based on the device type being used by its users.

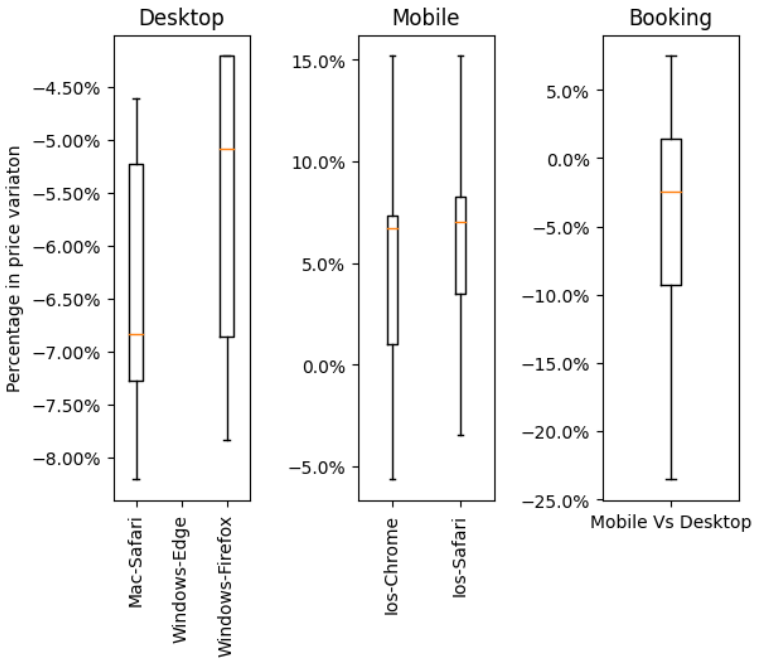

FIGURE 11 - VARIATION ON EACH USER PROFILE, BOOKING.COM

Booking.com shows that there are differences when comparing profiles on desktop and mobile. For desktop devices, Mac-Safari users and Windows-Firefox receive smaller prices for the same product than a user using a Windows device with the Google Chrome browser. For mobile users, iPhone sees higher prices than Androids with Google Chrome as a browser – with prices seeing a median variation of 7%. Comparing the average product cost found for mobile vs desktop devices, Booking.com shows that mobile users usually see smaller prices when compared to desktop users. In Booking's case, although mobile finds a median of -2.5% on price many products find differences close to zero and even higher. Meaning that a significant number of users see small differences in price.

The last question to be answered is whether the differences found are dependent on the product being searched for. If searching for hotels in Lisbon has more or fewer cases of price discrimination than searching for hotels in London. To understand the differences between each searched location, the number of products showing evidence of price discrimination (standard deviation different from

zero) is divided by the total number of results gathered. Hotels for ten major European cities were searched for.

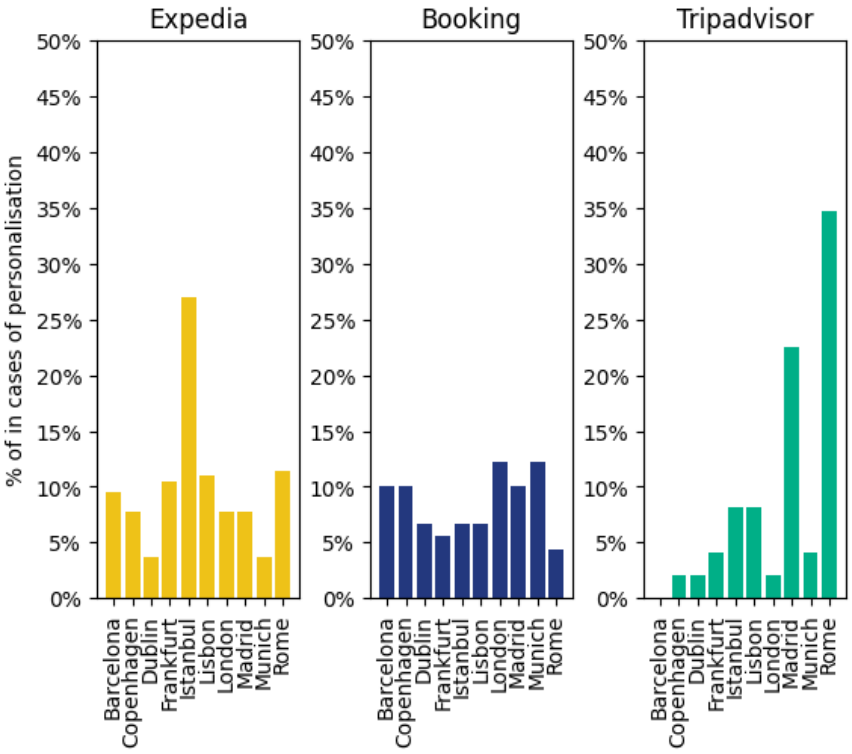

FIGURE 12 - PROPORTION OF PRODUCTS WITH PRICE DIFFERENCES, PER PRODUCT SEARCHED

In all three websites, the differences are not uniform. Booking.com shows the most consistency across all ten cities. Expedia and TripAdvisor show cities with the most discrimination cases, proportionally. Expedia presents the most discrimination issues when searching for hotels in Istanbul, with more than 25% of products showing price differences across multiple user profiles. When searching for hotels in Rome, TripAdvisor shows that 35% of results have different prices for different user profiles. Overall, no clear pattern is found but it can be concluded that the product being searched impacts the price shown to the user.

6.3 Recommendations for future steps

A natural next step is to increase the number of websites being scraped. Analysing more e-commerce businesses would promote a better overview of the price discrimination issue. Also,

performing similar analysis during a discounted season (cyber-Monday, Black Friday, Christmas) to understand if the user's profile is impactful on the level of discount applied.

Another important step would be to increase the variables used - measuring price discrepancies between users who are logged in vs not, or users who have a previous buying history on a website or similar websites, etc.

6.4 Conclusion

Three of the eleven investigated websites show evidence of price discrimination based on what device, operating system and browser is used when searching for products. Websites are Booking.com, TripAdvisor and Expedia. On all three, mobile users see, on average lower prices when compared to desktop users for the same products. No pattern was found on how a specific product searched impacts the price on each website.

The typical small banner with cookie information is displayed when visiting these websites. Booking.com asks permission to analyse a user's data for analytic and monitoring purposes. TripAdvisor's banner says that besides geolocation, they "Actively scan device characteristics for identification". Expedia says it stores data and cookies essential for marketing, personalisation, and analytics. No information is clearly shared with every visiting user alerting for the possibility that the prices seen might change based on their device characteristics. Even in cases where an alert is given, it is not clearly displayed to every user. Booking.com shows an alert tag "Price exclusive for mobile devices" but the user on a desktop device is not aware that is being charged a different price for the same hotel.

No evidence was found when searching for price discrimination based on the user's location.

7. Challenges and limitations

The main challenge in developing this thesis was the data collection step. More websites were investigated, but some constraints were found. The scraping was sometimes very limited. Websites like FNAC would ban the exact search from being done multiple times in a limited period. Other e-commerce platforms, Trivago, for example, have in their terms and conditions that web scrapping is not allowed on their websites. Also, when the scrapping was possible, extracting the data embedded into the page's code was only sometimes possible. In some websites, it did not follow a standard pattern which made the collection of data impossible.

Another limitation mentioned before is that the scrapping accomplished using Playwright could not consider the geolocation variable. Since it was mainly on the websites of the travelling industry where price discrimination was found, important insights might have been lost.

8. References

OECD. 2021. "The effects of online disclosure about personalised pricing on consumers: Results from a lab experiment in Ireland and Chile.". Paris.
https://doi.org/10.1787/1ce1de63-en.

European Commission. 2018. “Consumer market study on online market segmentation through personalised pricing/offers in the European Union.” London.
https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_rights/aid_and_development_by_topic/documents/synthesis_report_online_personalisation_study_final_0.pdf

UK government. 2018.“Pricing algorithms Economic working paper on the use of algorithms to facilitate collusion and personalised pricing”. London.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746353/Algorithms_econ_report.pdf

Wolford, Ben. n.d. “What is GDPR, the EU’s new data protection law?”. GDPR.EU
https://gdpr.eu/what-is-gdpr/

Twin, Alexandra. 2022.“What Is Price Discrimination, and How Does It Work?” Investopedia.
https://www.investopedia.com/terms/p/price_discrimination.asp

Mattioli, Dana. 2012. “On Orbitz, Mac Users Steered to Pricier Hotels.” Wall Street Journal.
https://www.wsj.com/articles/SB10001424052702304458604577488822667325882

Bertini, Marco, Koenigsberg, Oded. 2021. “The Pitfalls of Pricing Algorithms.” Harvard Business Review.
https://hbr.org/2021/09/the-pitfalls-of-pricing-algorithms

Gal, Avigdor. 2017.“It's a Feature, not a Bug: On Learning Algorithms and what they teach us - Note by Avigdor Gal”
https://one.oecd.org/document/DAF/COMP/WD(2017)50/en/pdf

Ho, Shuk Ying. 2010. “The Attraction of Internet Personalization to Web Users.”
http://www.electronicmarkets.org/fileadmin/user_upload/doc/Issues/Volume_16/Issue_01/V16I1_The_Attraction_of_Internet_Personalization_to_Web_Users.pdf

Pasquali, Marina. 2022. “E-commerce worldwide – statistics & facts”.
https://www.statista.com/topics/871/online-shopping/#topicHeader__wrapper

Zuiderveen Borgesius, Frederik, & Poort, Joost. 2017. “Online Price Discrimination and EU Data Privacy Law”. Journal of Consumer Policy, 40(3), 347-366.

Christl, Wolfie. 2017. “Corporate surveillance in everyday life: How companies collect, combine, analyse, trade, and use personal data on billions”.
http://crackedlabs.org/dl/CrackedLabs_Christl_CorporateSurveillance.pdf

Poort, Joost. 2019. “Does everyone have a price? Understanding people’s attitude towards online and offline price discrimination.” https://policyreview.info/articles/analysis/does-everyone-have-price-understanding-peoples-attitude-towards-online-and-offline

Streitfeld, David. 2000. “On the Web, Price Tags Blur”. The Washing Post. https://www.washingtonpost.com/archive/politics/2000/09/27/on-the-web-price-tags-blur/14daea51-3a64-488f-8e6b-c1a3654773da/

European Union. 2016. L119. Official Journal of the European Union. 59. 35-43.

European Union. 2019. C232. Official Journal of the European Union. 62. 7-21.

European Commission. 2020. “Consumer market study on online market segmentation through personalised pricing/offers in the European Union.” https://ec.europa.eu/eurostat/cros/system/files/04_-_web_scraping_policy.doc

European Commission. 2017. “STUDY ON MEASURING CONSUMER DETRIMENT IN THEEUROPEAN UNION.” https://ec.europa.eu/info/sites/default/files/consumer-detriment-study-executive-summary_en.pdf

Hannak, Mislove, Wilson, Lazer, and Soeller. 2014. “Measuring Price Discrimination and Steering on E-commerce Web Sites.” https://mislove.org/publications/Ecommerce-IMC.pdf

n.d. “Understanding Operating Systems – University of Wollongong.” UOW. https://www.uow.edu.au/student/learning-co-op/technology-and-software/operating-systems/

“What is a web browser?”, Mozilla, n.d. https://www.mozilla.org/en-US/firefox/browsers/what-is-a-browser/

Smith, Justin. 2022.“Mobile eCommerce Stats in 2022 and the Future Online Shopping Trends of mCommerce.” outerboxdesign. https://www.outerboxdesign.com/web-design-articles/mobile-ecommerce-statistics

Jehanne, Marie. 2022. “Mobile vs Desktop Usage: What Are the 2022 Consumer Trends?” contentsquare. https://contentsquare.com/blog/mobile-vs-desktop/

gs.statcounter. 2022. “Desktop Operating System Market Share Worldwide.” https://gs.statcounter.com/os-market-share/desktop/worldwide

gs.statcounter. 2022. “Mobile Operating System Market Share Worldwide.” https://gs.statcounter.com/os-market-share/mobile/worldwide

gs.statcounter. 2022. “Browser Market Share Worldwide.” https://gs.statcounter.com/browser-market-share

sematext. n.d. “HTTP Requests.”
https://sematext.com/glossary/http-requests/#what-is-an-http-request

IBM. 2021. “HTTP responses.”
https://www.ibm.com/docs/en/cics-ts/5.2?topic=protocol-http-responses

Hicks, Kristen. 2020. “How IP Addresses Are Tracked.” hostgator.
https://www.hostgator.com/blog/how-ip-addresses-are-tracked/

n.d. 2022. “Cookie: Definition”. Google Ads Help.
https://support.google.com/google-ads/answer/2407785?hl=en

10. Annex

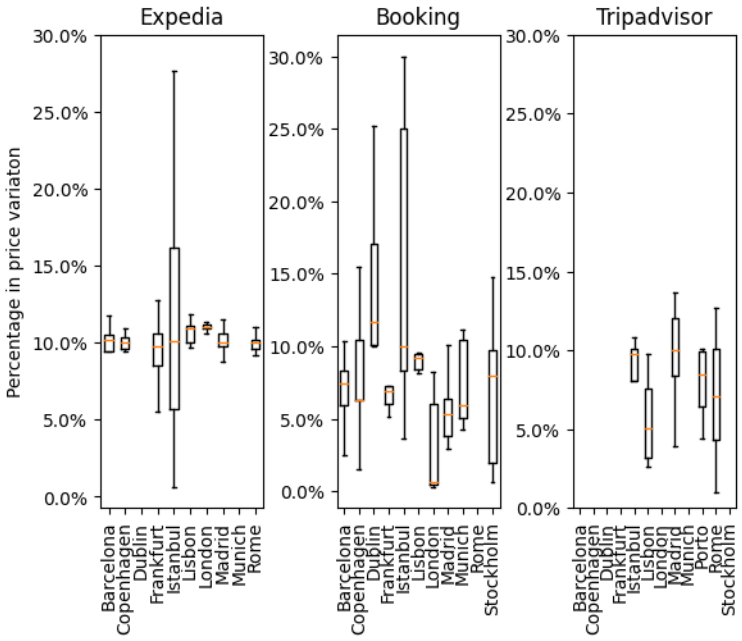

FIGURE 13 - PRICE VARIANCE PER PRODUCT SEARCHED, PER WEBSITE

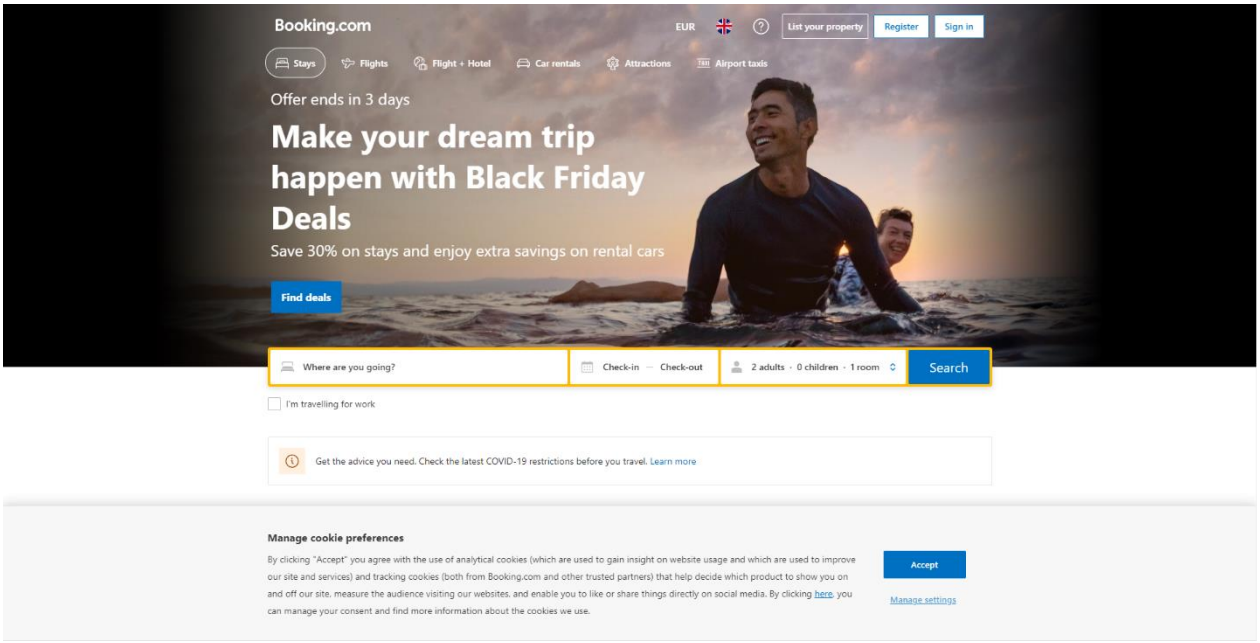

FIGURE 14 - COOKIE ACCEPTANCE BANNER FOR BOOKING.COM

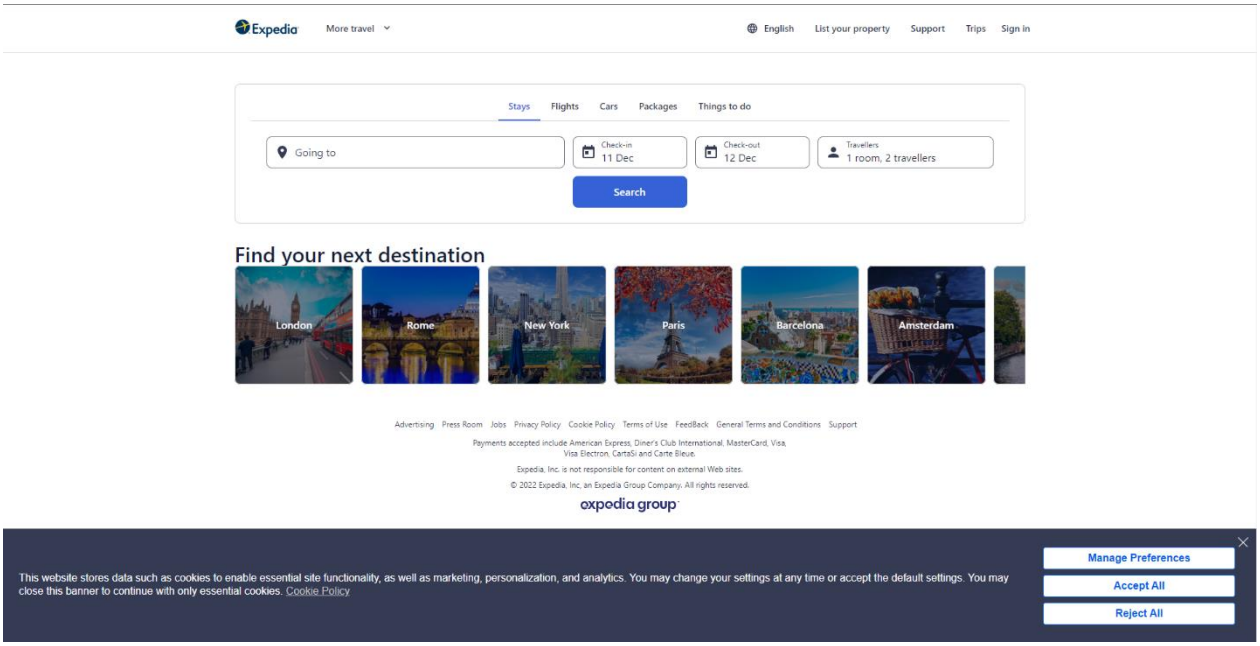

FIGURE 15 - COOKIE ACCEPTANCE BANNER FOR EXPEDIA

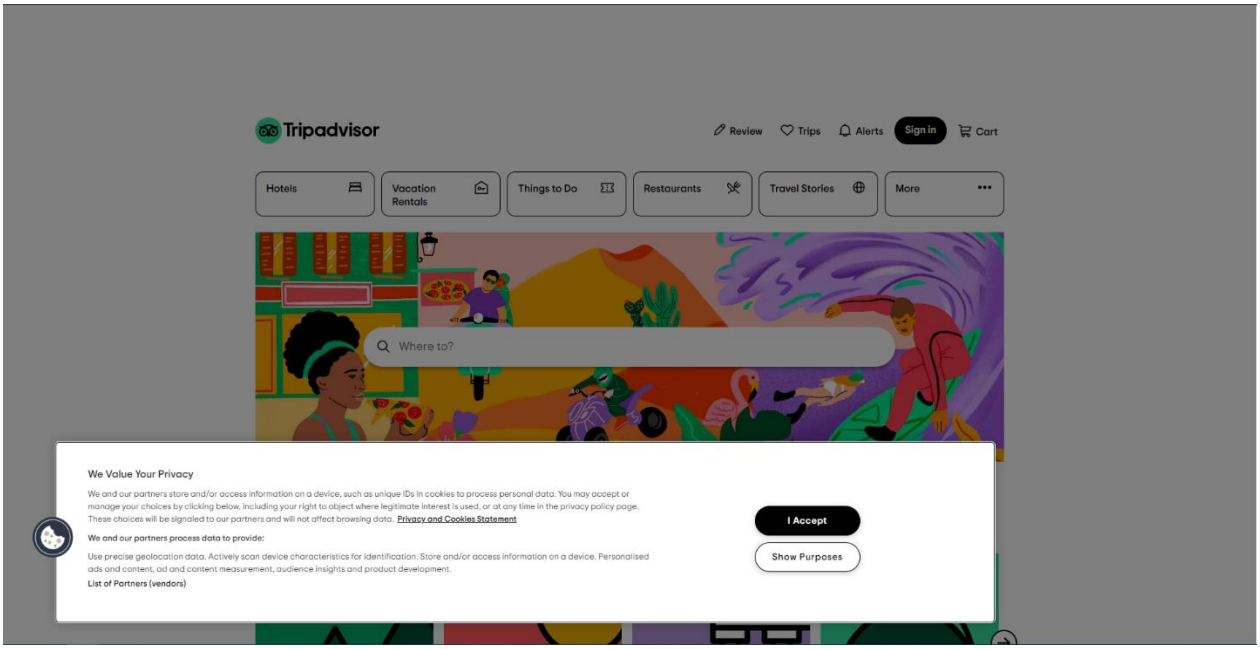

FIGURE 16 - COOKIE ACCEPTANCE BANNER FOR TRIPADVISOR