Consent For Sale

Determining the characteristics and price for a 'pay-or-okay' cookiewall in the European market.

Rogier Fransen - 528852

MSc Business Analytics & Management

Rotterdam School of Management [1]

BMMTIBAM Master Thesis

Thesis Coach: Gillian Ponte

Co-reader: David Jakob Kusterer

June 15, 2024

Abstract

This research examines the emerging ”pay-or-okay” cookie walls in Europe, where users are presented with the option to either pay a fee to not be tracked or consent in exchange for free access. The study analyses the characteristics of websites implementing this model, focusing on pricing, country of highest traffic, website popularity, category, and the number of hosts, trackers and associated companies. Findings reveal two distinct pricing clusters, a popularity of pay-or-okay models among German websites, and a tendency for these models to be adopted by websites with lower popularity scores. Additionally, the research concludes a differences between characteristics of websites using independent pay-or-okay models and those affiliated with Subscription Management Platforms (SMPs). The study highlights the need for managers to carefully consider website characteristics and market trends when deciding whether to implement a pay-or-okay model and determining an appropriate pricing strategy. A solution was created as an interactive Shiny application to help managers decide whether to implement a pay-or-okay model and, if so, whether to use a SMP and what competitive price to set.

KEYWORDS: Pay-or-okay, Cookie paywalls, Cookie banners, GDPR, Website tracking, Data privacy, Pricing privacy, Subscription Management Platforms (SMPs), Contentpass, Freechoice, Predictive Modeling

Acknowledgement

I would like to thank Gillian Ponte for his role as thesis coach. He helped me get started using the WhoTracks.Me dataset and supported my novel topic. Moreover, I want to thanks Cristiana Santos for taking her time and doing a short interview with me about her research on this topic. These insight gave me new perspectives on the current pay-or-okay landscape and gave me guidance for writing this thesis. Moreover, my colleague and other thesis interns at KPMG helped me stay concentrated during the thesis period. Finally, to all the pop-up ads that tried to distract me during my research. It was a nice try, but I made it though the end. Your persistence was truly... ad-mirable.

Enjoy my master thesis,

Rogier Fransen
2024

Contents

1 Introduction

1.1 Introduction and problem background

In November last year all the Instagram and Facebook users got two options; (1) Share their data with Meta and therefore receive personalised advertisement or (2) pay €9,95 per month to make use of the service without advertisement and data tracking (only necessary cookies) (Spinner, 2023; Folks, 2024; Ernst 2023. This new policy came into place after the recent court ruling from the The Court of Justice of the European Union (2023). This made it clear that Meta cannot rely on contract necessity or legitimate interest for processing data to personalize advertisement. In order to prevent the mass data processing opt-out in Europe Meta announced 10 days before the European ban on data processing the so-called pay-or-okay model, as described before [1]. By doing this, Meta could exploit their behavioural advertisement services like they used to, but critics, like Tobias Judin, would say this change has opened two classes of web users: Those that can afford to not be tracked, and those who need to pay for services with their data (IAPP, 2024).

The pay-or-okay, also called cookie paywalls or accept-or-pay cookie banners, is not a new phenomenal. It has been common practise on News and Media websites that heavily relies on the advertisement income (Rasaii, Gosain, & Gasser, 2023), (Morel, Santos, Lintao, & Human, 2022). The idea is simple, because of the opt-in on data processing with the GDPR, the value of advertisement has decrease for people that did not opt-in for data processing (also known as clicking on 'reject' button of a cookie banner). A study that looked at the effect of GDPR implementation on users found that due to the decrease in visits and usage intensity, 2.5 million euro could be lost for an ad-based website in the 18 months after GDPR implementation (Schmitt, Miller, & Skiera, 2020). Moreover, a DiD study by Wang, Jiang, and Yang (2024) suggest that after the implementation of the GDPR the revenue per click declined and that consent (opt-in) for data processing has a positive impact on the revenue per click suggesting that people

[1] For a comprehensive timeline of the Meta case see appendix B

that decline cookies are less valuable for the publishers. This is also found in research on users perceptions of cookie banners (Kulyk, Hilt, Gerber, & Volkamer, 2018),(Hofstad & Lundqvist, 2021).

When taking into account the increasing awareness of consumer privacy [2] and in response to the European Privacy Laws and stricter rulings from The Court of Justice of the European Union (2023) companies have adopted (BVDW, 2023). The declining revenue could become a problem in the next few years.[3] Eventually companies choose the last option that was left asking a appropriate compensation.

Recent research on the cookie paywall shows that pay-or-okay websites are most prevalent in Germany, where 8.5% of the top 1k websites has a cookie paywall(Rasaii et al., 2023). The prices of these cookies walls also varies when compared with each other. some websites, like the Austrian DerStandard.de ask for 75 euros per year to not being tracked and removing advertisement. Other companies go as low as 35.88 euros per year (see appendix A, from Morel et al. (2022)). However this policy has just been accepted, in 2023, by the German authorities and is expected that this rise will also blow over to other European Countries (Datanschutzenkonferenz, 2023) (NOYB, 2024). Given this new regime in the personalized advertisement, it is not known how the price of such service should be determined. It is not known how many websites already have implemented a cookie paywall and what are the characteristics of these websites.

The vault of the pay-or-okay model in Europe has opened with the courts' decision and the DPAs opinions, this in turn makes it more important to have data-driven managerial insights before it is widely debated in the marketing industry. This is supported by the regulatory question that has been asked for by the IAB Europe (2023), an association for digital marketing and advertising ecosystem.

[2]For example, the increase in GDPR awareness from 28% to 66% between 2018 and 2022 in the Netherlands(Global Data and Marketing Alliance, 2022).

[3]First, by arguing that data processing for personalized advertisement is within the contractual rights, after this was denied by the court for not being necessary for the services (see also consideration. 102 of The Court of Justice of the European Union (2023), and the now outdated article by White (2023)), companies argued that the data was needed for legitimate interests, which was also not accepted, because the service still needed to ask permission from the user (see also consideration. 117 of The Court of Justice of the European Union (2023)).

1.2 Reading Guide

This thesis begins with Chapter 2, which provides a theoretical background on cookie banners, tracking, and the pay-or-okay models. This chapter also includes the problem statement, research question, and relevance. Chapter 3 details the methodology and dataset used for the analysis, presenting a schematic overview of the steps taken. Chapter 4 presents the results along with some preliminary conclusions. Finally, Chapter 5 offers the research and management conclusions, discusses the limitations of the study, recommendations and provides insights for future research.

2 Theoretical Background

In this chapter, a theoretical overview of the research that has been conducted on this subject will be given. This overview will serve as a foundation for the research conducted. First, a brief background on website tracking and GDPR will be presented, followed by a discussion on the evolution of cookie banners. Then, the main concept of the "Pay-or-okay" in the EU is explained. Lastly, an new perspective on privacy will be elaborated on.

2.1 Background on website tracking and GPDR

On May 25, 2018, the GDPR came into effect. This regulation is meant to protect the fundamental right of processing personal data. One of the important aspect of this law, was the freely given consent when personal data is being processed. In the early days of the internet it was normal to collect data of online users without them knowing which data was collected or a way for a user to repent collecting personal data. By introducing this law (and its predecessors), It became important for every online website data processor to obtain the consent of their users.

As a result, many companies introduced a cookie notices on their websites. In this cookie notice, data processors had to inform the user about the ground on which they want to process the data and the user could also decline the acceptance of non-essential cookies. After the introduction of this regulation, Dutch websites experienced an 17% increase of cookie notices (Degeling et al., 2019). At first, every website employed their own tool and interpretation of the GDPR, which resulted in banners that where not complacent with the terms of the GDPR. A dance with the regulator advanced in the marketing field.

2.2 Cookiebanners evolution

Since the cookie notice was a novel phenomenon, various implementations of this tool were deployed. This resulted in a revolution, where 'Dark Patterns' [1] were getting increasingly popular. For advertisement purposes, having data on the demographics and interests of your target group is essential for targeted ads. By finding ways to optimize the number of users willing to accept cookies for targeted advertisement, ads would be, on average, 18% more valuable (Laub, Miller, & Skiera, 2023). However, regulators have attempted to combat these practices by making the rules less open to interpretation and easier to implement. Nevertheless, Sanchez-Rola et al. (2019) estimate that people are still tracked in 90% of cases. Research from Nouwens, Liccardi, Veale, Karger, and Kagal (2020) concluded that empirical surveys of consent management platforms turn a blind eye or even incentives illegal configurations of the design for accepting consent.

This cat-and-mouse game has been ongoing for a few years. The regulator introduces new guidelines for enforcement, but the websites try to find ways to circumvent it. Leiser and Santos (2023) suggest that regulators should clarify which dark patterns are illegal and also be clearer in enforcing this. Since the beginning of 2024, the Dutch Privacy Authority (DPA) has made a statement that they will start enforcing against misleading cookie banners, making it clear that users should have a clear and "freely given" choice in their tracking options (Autoriteit Persoonsgegevens, 2024).

As a result of the new practise, a decline in online tracking possibilities can be found from research. Utz, Degeling, Fahl, Schaub, and Holz (2019) found right after the implementation of the GDPR on average between 41,0% to 45,6% of the users accept a binary cookie banner on their phone. This trend can also be found in studies from the industry, where the consent rate was 55% in 2021, but declined to 45% in 2022 [2] (CommandersAct, 2022). The majority of users did not response to the banner and thus did not give consent for being tracked. To combat this problem, more deceptive design elements have been used in recent years, especially with the use of legitimate interest and the IAB' TCF framework (Morel, Santos, Fredholm, & Thunberg, 2023). The effects of these design can be of great impact. Research from Utz et al.

[1] The so-called "dark patterns" are methods used by websites to encourage more consumers to accept cookies, enabling more personalized advertisement.

[2] Where consent is the number of people who have consented to at least one purpose / the number of banners exposed to the Consent Management Platform (CMP)

(2019) suggest that small implementation decisions can substantially impact whether and how people interact with consent notices.

2.3 Rise of Pay-or-okay in the EU

From an advertisement and a publisher perspective, the shift to less consent given by users will decrease the value of the advertisement or their space on the website, respectively. Research has calculated that due to an inability to track users, there would be a decrease of 18% in ad revenue (Laub et al., 2023). For a website that only has advertisement as a business model, this decrease in revenue would mean that it is not possible to make a profit or even sustain the website. However, marketers has found that after implementing a pay-or-okay model, visitors accepted tracking in 99,9% of the cases Müller-Tribbensee, Miller, and Skiera (2024). Thus, A new cookie banner was formed.

A pay-or-okay cookie paywall is defined as a paywall where if a user refuses tracking, she is then obliged to provide a sum of money to access that website (Morel et al., 2023) (CNIL, 2022). When implemented, users generally have three options: 1) accept personal data processing, 2) pay a small fee (usually monthly) to enter the website, or 3) leave the website altogether. The first website that introduced this concept was the Austrian publisher DerStandard.de. After that, some variations have emerged [3]. For example, most publishers, like Meta, choose to add an advertisement-free version of the website as a supplement in the monthly fee subscription. Moreover, Subscription Management Platforms (SMPs) [4] like *Contentpass* provide a monthly description for over 450 websites (Tavan, 2018). For a fee of 3.99 a month, users don't have to worry about being tracked and can experience the website advertisement-free. In Figure 2.1 an overview of the growth of pay-or-okay is given.

When looking at Figure 2.1, it can be noticed that some EU countries are not included in the overview. In an 1-1 interview with Cristiana Santos [5], a legal professor at University of Utrecht, it was explained that the stances of the different European Data Protection Authorities are not clear and therefore websites in those countries fear a fine. An overview of the different views by

[3] see for an overview IAB Europe (2023) & Müller-Tribbensee (2023)

[4] This term was first coined by Morel et al. (2023)

[5] Assistant Professor in Law and Technology at University of Utrecht, see https://www.uu.nl/medewerkers/CTeixeiraSantos

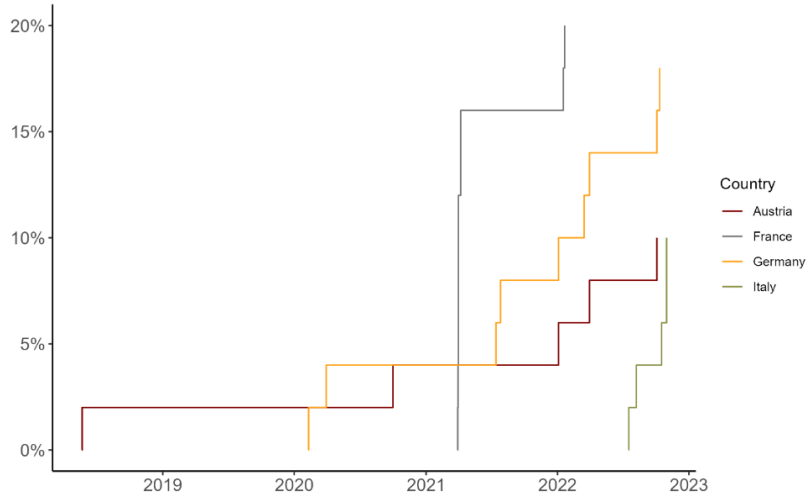


Figure 2.1: Share of pay-or-okay cookiewalls among top 50 publishers per country over time, from Müller-Tribbensee et al. (2024)

the DPAs is given in Morel et al. (2022).

2.4 Publishers view

In this section, a brief overview of the effects for publishers will be given. First of all, the use of a pay-or-okay model is inherently useful for a publisher. From research it is clear that 99,9% of users will consent when presented with a pay-or-okay cookie wall (see, Müller-Tribbensee (2023) and Tavan (2018)). Therefore, a pay-or-okay model will give publishers more quantity of data to process. From the IAB Europe [6], the objective of this model is to ensure compliance while also securing the sustainable funding of their content and services. The IAB Europe even sees a connection between the recent enforcement of the 'decline all' button on the top page and the introduction of the pay-or-okay banner, suggesting that given a clear option, users will decline their consent. The allowance of this model is, therefore, important for publishers. However, most things still remain unclear, which is why the IAB Europe wants to know specifically "What is the price and what criteria are used to determine it for the PUR subscription?" (BVDW, 2023).

[6]IAB Europe is the European organisation for digital marketing and advertising ecosystem. They are known for their industry standard Cookie banners, that a website can customise.

2.5 Pricing privacy

In her essay, Tucker (2023) provides a roadmap for young researchers concerning the economics of privacy. One crucial aspect she addresses is the pricing of privacy and the inherent inequalities associated with it. She posits that understanding how to assess the value of privacy, and how this valuation may vary among users, is increasingly pertinent. In the realm of pay-or-okay cookie banners, this raises the fundamental question: how should we evaluate the tracking of customers, and what constitutes a fair price for websites to track them?

Research conducted by Müller-Tribbensee et al. (2024) indicates that, on average, a publisher generates 0.24 EUR per user in monthly revenue. However, websites employing a Pay-or-Okay model typically charge an average of 3.24 EUR per month, suggesting a 3 EUR surplus under this revenue model. This prompts an inquiry into how we economically assess privacy's value, the elasticity of pricing, and what users are willing to pay to maintain the same ad revenue. For instance, if a decline button is utilized by 20% of users while a pay button is only used by 0.1%, the economic value of declining should theoretically be higher to achieve equivalent revenue [7].

The introduction of new policies that make user tracking more challenging, such as the GDPR, typically results in a decrease in advertisement prices, as demonstrated by Wang et al. (2024). For example, a 3% decrease in prices can be observed following the GDPR's implementation due to reduced tracking possibilities. If the current GDPR policy is explained more strictly, a higher price decrease is likely. Conversely, a positive explanation of the GDPR may increase advertisement prices. However, there is still limited understanding of how to value privacy within a pay-or-okay model.

2.6 Problem statement and research questions

For a lot of publisher the decision to implement a pay-or-okay model and in what form is not easy. Market research suggest that on an economic and data protection reasons implementing it is a sensible thing to do (BVDW, 2023). In Germany more than 280 online publishers have already joined this trend. However, not much is known about the price and what the characteristics are

[7] Müller-Tribbensee et al. (2024) conducted a back-of-the-envelope calculation on precisely this thought experiment.

of these website. Moreover, when a publisher does decide to implement a cookie paywall, there is no information on if they should go with a SMP like *Contentpass* and what the right price is. In other words how much is data tracking worth for that website, and are there indicators that would give you some insights from competitors.

Therefore, the main research question in this paper will be as followed: *How can marketing managers of websites utilize site characteristics to decide on implementing a pay-or-okay model and determine the appropriate pricing strategy?*. For the model building, the characteristics from the WhoTracks.Me sites.csv dataset. However when looking at the sites characteristics the focus will be on the following characteristics: (i) country; (ii) category; (iii) popularity; (iv) hosts/trackers/companies. All of these characteristics are based on the WhoTracks.Me database (Karaj, Macbeth, Berson, & Pujol, 2019).

This main question is divided into the following four sub-question:

1. *What are the characteristics of the current pay-or-okay landscape in Europe?* In this question it is important to further analyse the finding of the European website that have implemented pay-or-okay models. This will be a descriptive analysis using the current datasets of website that have implemented pay-or-okay models in addition to the characteristics that is gathered using the WhoTracks.Me dataset from Ghostery.

2. *What are the key differences in characteristics between websites that implement their own pay-or-okay model and those that use a SMP?* This subquestion explains the the main difference of a website that choose to have their own pay-or-okay model and website that are affliated with a Subscription Management Platforms (SMP) like, *Contentpass* or *Freechoice*. Based on current research it is hypothesised that both group will have a significant difference in characteristics.

3. *What site characteristics influence the pricing of a cookie paywall?* This analysis will be an extension on the analysis done by Rasaii et al. (2023), where the correlation between the price and the amount of tracking cookies was done. The following effects are hypothesised when focusing on the main characteristics:

   (a) H1: Higher popularity of the website (based on traffic rankings) is expected to correlate with a higher cookie paywall price, as more popular websites can demand a higher premium for access to user data.

(b) H2: Websites in categories News & Entertainment have higher cookie paywall prices than others, as user data in these categories could be more valuable for advertising and tracking purposes.

(c) H3: A higher number of hosts on a website lead to a difference in cookie paywall price.

(d) H4: The higher number of tracking companies influence the cookie paywall price, as their tracking capabilities and data may be more valuable and sell to more companies.

(e) H5: The price of website with traffic from other countries than Germany would be lower, as in these countries the pay-or-okay model still has to be developed.

4. *How can marketing managers use these insights to decide whether to implement a pay-or-okay cookie paywall model and set an appropriate price?* This question focuses on developing a practical model to assist future managers in making decisions about implementing and pricing pay-or-okay models. Initially, managers may struggle to decide whether to adopt such a model, and may look to industry trends for guidance. They must then consider whether to partner with a SMP like *Contentpass* or proceed independently. Lastly, based on characteristics of various sites, managers can determine a competitive price. Answering this subquestions will provide managers with valuable insights for implementing data cookie paywalls effectively.

2.7 Managerial relevance

The emergence of cookie paywalls, like the Meta case, presents a new strategic consideration for managers. It challenges the conventional revenue models of behavioural advertisement. The decision to implement a cookie paywall is a strategic move with implications for user trust, market positioning, and financial outcomes. Moreover, determining the right price for a cookie paywall involves understanding the market's valuation of data privacy, by looking at website that has implemented the cookie paywall.

This research provides managers with insights into the current landscape of cookie paywalls. It aims to identify factors influencing paywall pricing and adoption, thereby guiding managers on whether and how to implement such a model effectively.

2.8 Academic relevance

Research has already been done on the European perspective. Some basic insight could be gathered from that. However, not much is know about the pricing strategy of the cookie paywall. Some preliminary research is done by Rasaii et al. (2023), but this only suggest that there should be more research to be done, especially with addition data as supplementary.

In (Papadopoulos, Snyder, Athanasakis, & Livshits, 2020), authors performed an empirical study of the paywall prevalence on the web. Rasaii et al.(2023) also did a small experiment on the price and the amount of tracking cookies and found no meaningful linear correlation between the number of tracking cookies set by websites when accepting tracking and the subscription price. However, this small experiment misses a lot of other variables, like which cookies is found on the website and normalized privacy scores.

This research will bridge the gap in the current studies by adding extra characteristics to the before-done analysis. Moreover, a better understanding of the pricing strategies for website would become invaluable when an European ruling is expected to come in the next half a year. In the end, it could be also for added value as a proceeding for a conference as this topic is now highly debated in the field.

3 Methodology (and Data)

For answering my research question, multiple datasets will be used for the analysis. Fortunately, substantial research has already identified multiple websites with a pay-or-okay model, providing a foundational dataset for this study. Additional data on website tracking characteristics will be collected from WhoTracksMe by Ghostery. This section discusses the choice of datasets and provides some basic statistics. First, the datasets are highlighted, followed by an explanation of how they were used to create the dataset for this analysis. Lastly, the specific methods chosen for the analysis are explained. The repository of the analysis can be found on Github and in Appendix H.

3.1 Datasets

For this research, three datasets are important: (1) websites with a pay-or-okay model from other research, (2) website tracking characteristics, and (3) websites that use Subscription Management Platforms (SMPs) like *Freechoice* or *Contentpass*.

3.1.1 Websites that have a pay-or-okay model from other research

To start, there is already a database that has crawled over 45 thousand websites in the EU from different servers from which 285 website have been identified as having cookie paywalls. This dataset has been used for research on cookie paywalls (Rasaii et al., 2023). The dataset entails 285 URLs of website that have pay-or-okay model. The prices of site that are in the WhoTracks.Me dataset were manually checked and added. Since Rasaii's dataset is limited (285 websites), additional research was considered. Morel et al. (2023) identified over 431 websites with pay-or-okay models by crawling over 1 million sites. This dataset includes price, category, and country, but only URL and price were used for this analysis. Additionally, three more websites were found by the researcher through forums like Reddit and Tweakers. These datasets provided a quick and efficient way to obtain URLs compared to using a scraper to scrape website

cookie banners to find pay-or-okay models.

3.1.2 Website information

To gather comprehensive information about each website, focusing on key characteristics such as (1) country, (2) category, (3) popularity, (4) hosts/trackers/companies, additional data was obtained from the WhoTracks.Me database. This database, maintained by Ghostery, compiles tracking data collected from multiple users and ensures anonymous availability. All the characteristics from the dataset can be found in appendix C and on the WhoTracksMe Github page. The data can be accessed through the GitHub repository or via an MS Azure connection.

The dataset was downloaded from MS Azure in April 2024. It contains extensive information. The most important variable is summarized in Table 3.1. However, it is important to note that the dataset only covers the top 10,000 websites within a given region (e.g., ”de”, ”eu”, ”fr”, ”global”, ”us”). For this study, data from all regions were utilized. The implications of this limitation on the research are discussed in the relevant section.

An attempt was made to supplement this information by using the Ghostery browser extension to gather data on additional websites. However, several challenges were encountered, also see the chapter 'Discussion'.

3.1.3 Websites that has Subscription Management Platforms (SMPs)

An additional layer of analysis involves determining whether a website has an SMP for users to avoid tracking. These services charge a monthly fee, allowing customers to access multiple sites without consenting to tracking. The two primary competitors in this space are *Contentpass* and *Freechoice*.

Contentpass, which charges €3.99 per month, eliminates banner ads, video ads, and personalized tracking for its subscribers. Users subscribing to *Contentpass* gain access to approximately 450 websites without the need to consent to tracking. *Freechoice*, costing €2.99 per month, allows users to avoid tracking while still being exposed to ads. This service covers around 234 websites. an overview of each SMP and their characteristics can be found in Table 3.2

To understand the full landscape and address the research question, it was necessary to gather the URLs of websites that use a SMP. This was done by using their respectively marketing sites

and scraping all partnered website that they have published on their site. By doing this an dataset was created that contained the URL of site that was associated with the *Freechoice* or *Contentpass*.

3.2 Making the dataset

Before starting the analysis and answering the research question, a dataset needs to be created from the above mentioned different sources. This is done by combining the three different sources into one dataset. The steps involved are: (1) data cleaning, (2) adding more data on website using SMPs, (3) combining the URLs with WhoTracksMe data and (4) Last cleaning and outliers. An overview of all the steps can be found in Figure 3.1.

3.2.1 Data Cleaning

To combine the websites found by different researchers, data cleaning was necessary to create a list of all the currently known websites with pay-or-okay data. Three datasets were used for this: (1) Rasaii et al. (2023) who found 285 websites with cookie banners. (2) Morel et al. (2023) Who found 431 websites. & (3) Three websites found manually by the researcher on the internet.

These three data sources were combined, and duplicates were removed. Eventually, 574 websites with pay-or-okay tracking walls were identified.

3.2.2 SMPs

After the initial identified website with pay-or-okay models, additional website where it is known that they use a Subscription Management Platforms (SMP), like *Freechoice* or *Contentpass* is added to the site. The duplicated website is removed and in the end 968 website has been found using a form of pay-or-okay.

To identify if a website uses a SMP the binary variable ('1') was added to the column of a specific SMP (i.e. "Freechoice") if a match was found. Additionally, a variable termed 'pass' was introduced to denote whether a website uses a SMP.

3.2.3 WhoTracks.Me data

Subsequently, additional information was appended to the list from the WhoTracks.Me database. Consequently, only 338 of the 968 identified websites could be cross-referenced with the WhoTracks.Me database. Variables of interest are summarized in Table 3.1, see for information on all variables appendix C.

3.2.4 Last cleaning & outliers

Post final data cleaning, four websites were identified that no longer displayed pay-or-okay cookie banners was removed, and one outlier with a price of €49.00 per month was being considered for removal. This outlier was excluded only for subquestion three to maintain the integrity of the analysis. For subquestions one, two and four, the dataset including the €49.00 outlier was retained.

A final overview of the dataset steps is conceptualized in a simplified Figure 3.1, and the summary statistics are presented in Table 3.3. This dataset will serve as the basis for answering the specified subquestions.

Variable	**Description**	**Type**
site	The domain of the website.	Character
price	The cost associated with the pay-or-okay no-tracking subscription.	Numeric
category	The site's category according to Ghostery.	Nominal
popularity	Relative traffic amount compared to the most popular site (0-1).	Numeric
country	The main region where the traffic originates (e.g., global, US, EU).	Nominal
hosts	Average number of tracker's domains on the site.	Numeric
trackers	Average number of trackers on the site.	Numeric
companies	Average number of companies on the site.	Numeric
freechoice	*Freechoice* option on their website. 1 if 'yes' or 0 if 'no'.	Boolean
contentpass	*Contentpass* option on their website. 1 if 'yes' or 0 if 'no'.	Boolean
pass	indicator is a website is using a SMP	Boolean

Table 3.1: Overview of most important variables, with the name, descriptions and type from left to right.

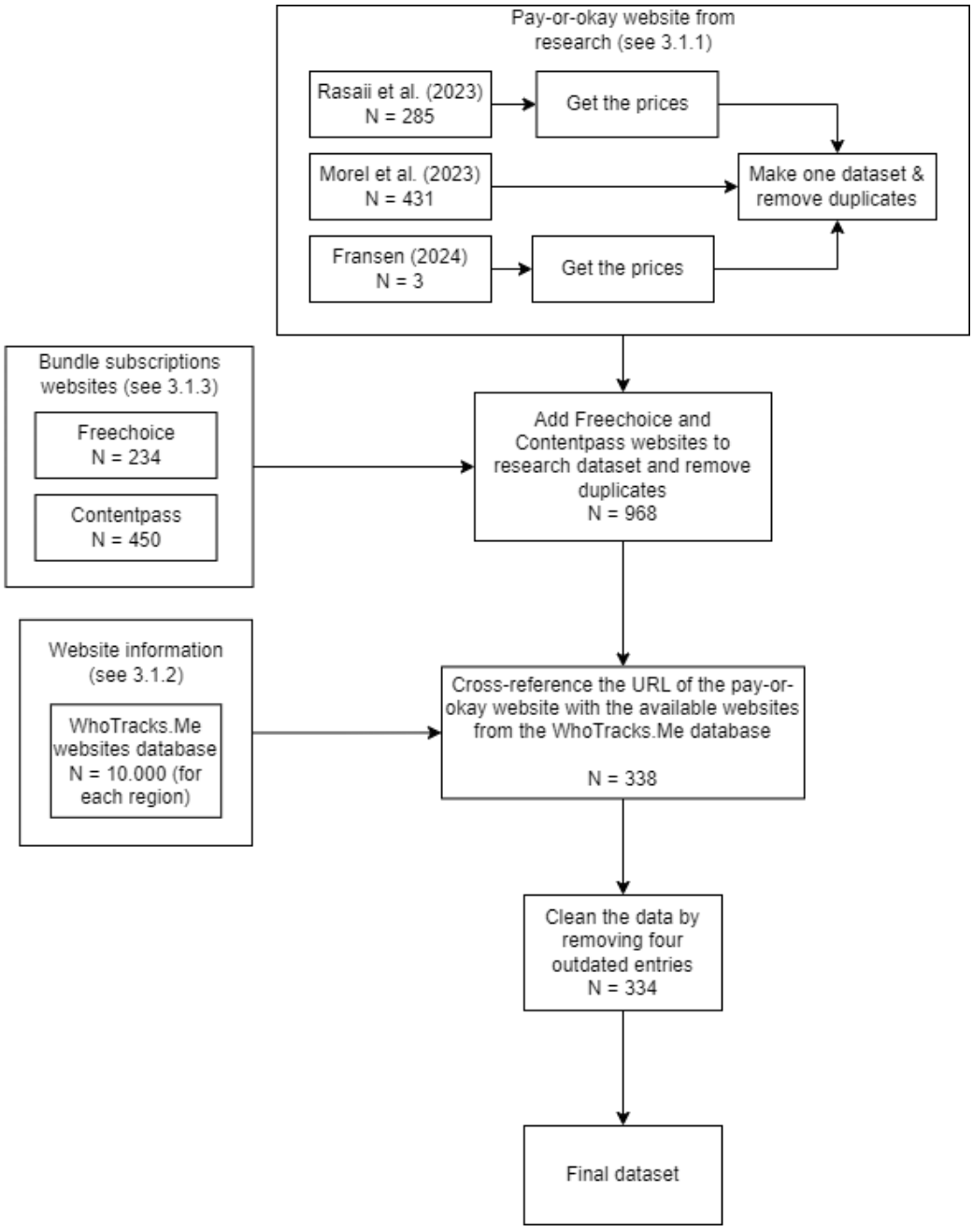


Figure 3.1: Simplified overview of the proces of making the final dataset for the analysis.

Feature	Contentpass	Freechoice
Monthly Cost	€3.99	€2.99
No Banner Ads	Yes	No
No Video Ads	Yes	No
No Personalized Tracking	Yes	Yes
Number of Websites	450	234

Table 3.2: Comparison of characteristics between *Contentpass* and *Freechoice*.

Statistic	N	Mean	St. Dev.	Min	Max
Price	334	3.769	2.98	0.750	49.000
Popularity	334	0.004302	0.0157	0.000057	0.149838
Hosts	334	10.435	6.03	1.210	26.921
Trackers	334	7.857	4.48	1.173	21.704
Companies	334	6.547	3.57	1.066	18.293
Pass (1 if SMP)	334	0.494	0.501	0.000	1.000

Table 3.3: Summary statistics from the final dataset

3.3 Subquestion 1

To answer this question, a descriptive analysis of the current dataset is provided. This includes an overview of the following variables: (1) price, (2) popularity, (3) category and (4) hosts/-trackers/companies. These variables are the most relevant from the WhoTracks.Me dataset and help in understanding the landscape of pay-or-okay models in Europe.

Graphs will be created to show the distribution of price, popularity, categories, and the presence of hosts/trackers/companies.

It is important to note that analyzing the price of pay-or-okay banners for websites with SMPs is challenging. Paying the price of these bundles gives a customer access to multiple sites without being tracked, making it difficult to determine the value for a specific website. Dividing the pass price among all websites in the network does not accurately reflect its true value, as users are unlikely to fully utilize all sites. Therefore, the prices of these passes do not provide meaningful insights and are excluded in the analysis of price.

3.4 Subquestion 2

In addressing Subquestion 2, which examines the key differences in characteristics between websites implementing their own pay-or-okay model and those using a SMP, the following steps are undertaken:

Firstly, an assessment of the structure and variance of characteristics between the two groups is conducted to determine if they follow a normal distribution. This evaluation is used to informs the selection of appropriate statistical tests.[1] It has been established that no characteristic is normally distributed and no characteristic has a low variance.

Consequently, for the comparison between websites implementing their own pay-or-okay model and those utilizing a SMP, the Mann-Whitney U test will be employed. The Mann-Whitney U test, a non-parametric test, differs in its hypotheses from the t-test commonly used for normally distributed data. Unlike the t-test, which assesses differences in means between two groups, the Mann-Whitney U test focuses on rank differences.

Specifically, it evaluates whether the distributions of ranks between the two groups differ significantly. Therefore, the hypotheses formulated for the Mann-Whitney U test will center around determining whether there are significant differences in the rank order of characteristics between websites with their own pay-or-okay model and those with a SMP. The hypothesis for this test could thus be written as follows:

H1: The rank distribution of the focus characteristics between websites with their own pay-or-okay model differs significantly from those with a SMP.

This methods ensures a reliable way to understand the presence of difference between the two group, especially when the normality assumption is not met.

3.5 Subquestion 3

For the third sub-question the effects of website characteristics on the price of a cookie paywall is analysed. This analysis excludes websites that use SMPs, such as *Contentpass* or *Freechoice*. The exclusion is necessary because bundled pricing complicates the interpretation of individual

[1] Research from Havlicek and Peterson (1974) suggest that when the distribution is skewed but the variance is low then it is no problem, however if both is the case, then another test is needed and the t-test should not be used.

website prices.

an simple Ordinary Least Squares (OLS) regression is used for this analysis. The OLS regression model is chosen due to its interpretability compared to other models like regression trees or K-Nearest Neighbors (KNN). This method gives insights into the relationship between website characteristics and the price of a cookie paywall and the proposed hypotheses.

To address the issue of non-normality in the residuals, we transform the dependent variable (price) using its natural logarithm. This transformation helps mitigate the skewness of the residuals, resulting in a distribution that is closer to normal. However, despite this transformation, the residuals are still not perfectly normally distributed, though they are less skewed than before.

Three models were tested to determine the best fit. In the non-log model, the normality assumption, like noted above, does not hold. This model has an adjusted R^2 of 0.100. The log model, while still not fully meeting the normality assumption, exhibits residuals with a more bell-shaped distribution. However, this model, like the non-log model, still has heteroscedasticity. The adjusted R^2 improves to 0.196. The unrestricted model, which includes additional website characteristics as control variables and after removing the highly correlated ones, improves explanatory power, results in an adjusted R^2 of 0.307. This model meets the homoscedasticity assumption and provides the best fit among the tested models. A comparison of these models is presented in Table 3.4

The specific models are defined as below. For the non-log and log model the focus variable trackers is removed, because it created multicollinearity with the variables hosts and companies. This by itself is not suprising, the hosts (url of the tracker), is part of a certain tracker, which is therefore associated with a company:

Model 1 (non-log):

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{country} + \beta_2 \cdot \text{category} + \beta_3 \cdot \text{popularity} + \beta_4 \cdot \text{hosts} + \beta_5 \cdot \text{companies} + \epsilon$$

Model 2 (log):

$$\ln(\text{Price}) = \beta_0 + \beta_1 \cdot \text{country} + \beta_2 \cdot \text{category} + \beta_3 \cdot \text{popularity} + \beta_4 \cdot \text{hosts} + \beta_5 \cdot \text{companies} + \epsilon$$

Model 3 (expended):

$$\begin{aligned}\ln(\text{Price}) =& \beta_0 + \beta_1 \cdot \text{country} + \beta_2 \cdot \text{category} + \beta_3 \cdot \text{popularity} + \beta_4 \cdot \text{bad\_qs} + \beta_5 \cdot \text{tracked} \\ &+ \beta_6 \cdot \text{https} + \beta_7 \cdot \text{requests\_tracking} + \beta_8 \cdot \text{content\_length} + \beta_9 \cdot \text{requests\_failed} \\ &+ \beta_{10} \cdot \text{has\_blocking} + \beta_{11} \cdot \text{script} + \beta_{12} \cdot \text{iframe} + \beta_{13} \cdot \text{beacon} \\ &+ \beta_{14} \cdot \text{image} + \beta_{15} \cdot \text{stylesheet} + \beta_{16} \cdot \text{font} + \beta_{17} \cdot \text{xhr} \\ &+ \beta_{18} \cdot \text{plugin} + \beta_{19} \cdot \text{media} + \beta_{20} \cdot \text{referer\_leaked} + \beta_{21} \cdot \text{referer\_leaked\_header} \\ &+ \beta_{22} \cdot \text{referer\_leaked\_url} + \beta_{23} \cdot \text{cookie\_samesite\_none} + \beta_{24} \cdot \text{t\_active} \\ &+ \beta_{25} \cdot \text{hosts} + \beta_{26} \cdot \text{companies} + \epsilon\end{aligned}$$

In the end, the expended model is chosen because it has the highest explanatory power. In this model the main characteristics that effect the price of the pay-or-okay paywall can be found and will be discussed in the results section.

Assumption	Non-log model	Log model	Expanded model
Adjusted R^2	0.100	0.196	0.307
Heteroskedasticity (BP Test)	Present	Present	Not Present
Normality of Residuals	Not Met	Not Met	Not Met
Linearity	Met	Met	Met
Independence of Errors	Met	Met	Met
No Multicollinearity (VIF)	Met	Met	Met

Table 3.4: Comparison of linear model assumptions for the non-log, log and expended model.

3.6 Subquestion 4

In this question a solution will be presented to a problem manager could face in their daily life. The methods for the solution is creating an predictive model for (1) classifying sites with pay-or-okay model, (2) classifying whether the website should implement a SMP and if no bundle services is advised (3) predict a competitive pricing strategy for the website. A simplified schematic of this approach is given in Figure 3.2. In this section each predictive model will be elaborated on. All models have a train, validation & test split of 50%/20%/30%. This split is chosen to have enough data for training (50%), still some validation (20%) and still having data where the results can be tested (30%). Other approaches with percentages has been considered, however, in

this approach the trad-off between a good model and enough validating & testing data is taken into account.

3.6.1 Step 1: Pay-or-okay model

The first step for managers is to determine if their website characteristics are close to the characteristics of websites with a pay-or-okay model. Three prediction model has been made. To make this prediction model, the sites characteristics dataset from pay-or-okay model websites (N = 334) is used in combination to a balanced unstratified sample (N= 500) of the sites that where available though the WhoTracksMe database. A bigger sample of the available sites could have been used, however this would create an unbalanced dataset where the minority group (pay-or-okay) would not get the right weights over the majority group (other sites). Therefore, the majority class was down-sampled to get a balanced dataset for this analysis. The downsample of the majority class was not stratified on characteristics, because this would not create a difference in results. The random downsampled model is a big enough sample, and there were no characteristics where a stratification would be necessary.

For the predictors of the model, all the variable of the WhoTracks.Me dataset is used expect the site, month and country information. These variable were removed because the site and month is metadata about the dataset or row and country would not be a good predictor, as this will skew the model to more German sites. However, for this model the goal is to also help managers in other upcoming regions do decide whether they should implement the model based on website characteristics. Moreover, all the numeric predictors were normalised and for all nominal predictors dummies were formed.

Three models are deployed and compared to each other. A lasso model, k-NN and Boosting model. For the lasso, the tuning parameters for the penalty laid between 10^{-2} & 10^{2} with 100 steps in between -2 and 2. Moreover, for the k-NN model the tuning parameters where between 1 to 30 neighbours. For the boosting model the tuning happened with the following value for the parameters *trees*: 500 to 10000 with steps of 500, *learn rate*: 0.1, 0.01 and 0.001 and *tree depth*: 1,3,7. After tuning and choosing the best model the following formula is used

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

This metric looks at the amount of true positive that has been correctly identified. In this case the amount of true pay-or-okay cookiewalls that has been identified. This metric is most important when compared to specificity, because we want our model to most accurately reflect the characteristics of the current website as opposed to the wrongly negatively predict the pay-or-okay wall. Moreover, for managers discovering if their website might be a good fit the specificity $\frac{TN}{TN+FP}$ should actually be lower, because then more website would be identified as possible candidates. From the three model the best model is then used for deployment.

3.6.2 Step 2: Bundle service

In this step, the aim is to predict whether a website has characteristics associated with a SMP website. The dataset used includes all pay-or-okay websites ($N = 334$), which is divided into two equal groups: SMP websites ($N = 165$) and websites with their own pay-or-okay model ($N = 169$, one more as in subquestion 3, because of the outlier). This results in a balanced dataset.

The predictor variables used are the same as in the Pay-or-Okay prediction model. All numeric predictors were normalized, and categorical predictors were converted into dummy variables. The same tuning metrics as step 1 is used. For selecting this model, accuracy is the most important metric, unlike in Step 1 where sensitivity was prioritized. Accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is prioritized in this step because both groups (SMP and pay-or-okay) are equally important. A false positive (incorrectly predicting an SMP website) is not less costly than a false negative (missing an SMP website). This contrasts with the pay-or-okay model, where minimizing false negatives was more crucial to ensure correct identification of pay-or-okay websites. Therefore, in the context of SMPs, both types of errors are equally significant, making accuracy the appropriate metric for evaluating the model's performance.

In the end the best performing model is chosen to be deployed in the solution.

3.6.3 Step 3: Price

The last step happens when the SMP model predicts deploying your own pay-or-okay model. The prediction from this model will be the competitive pricing strategy that a manager could use. to make this solution three models are used: (1) the best permorming model that is made in subquestion two, (2) a lasso regression, with all the variables & (3) a boosting model (XGBoost). Unlike the other steps KNN is not used for a continuous variable like price. The outcome variable is also changed to the natural logaritm of the price, to compare the outcome between the model, especially with the model made in subquestion two. The same tuning parameter for the lasso and the boosting models is used like the before mentioned steps.

For evaluating the model the metric RMSE is used. this can be calculated like $\sqrt{MSE}$, where MSE is the mean squared error. This metric give the average squared difference between the estimated values and the actual value. This error needs to be the lowest for the best model.

After training and tuning, the performance of all three models is compared based on their RMSE values. The model with the lowest RMSE is selected as the final model for predicting competitive pricing strategies.

By incorporating these steps, it is ensured that the chosen pricing strategy model is both accurate and reliable, providing valuable insights for managers deploying their own pay-or-okay models.

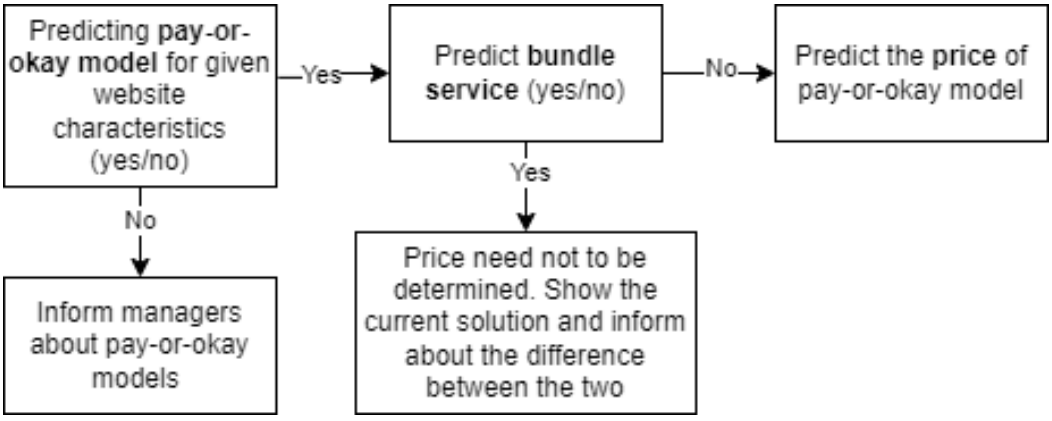


Figure 3.2: Simplified schematic of the approach used for research question four.

4 Results

This section provides a overview of the key outcomes from the data analysis, offering insights into the characteristics of websites implementing pay-or-okay models and those utilizing SMPs and making the best prediction model for managers.

4.1 Subquestion 1

To address this subquestion, the five most important variables is examined. These variables are: (1) price, (2) country, (3) popularity, (4) category, and (5) hosts/trackers/companies. The graphs illustrates the differences between websites with and without SMP. A closer examination of these differences will be discussed in subquestion two. An overview of the key statistics of the dataset can be found in Table 4.4. It should be noted that for this subquestion, the analysis is based solely on the sample, so conclusions about the entire population are limited. However, some insights into the current small population of websites with pay-or-okay model may be inferred.

When looking at the price it has an average of 3.76 euro. Moreover, two distinct groups can be observed in Figure 4.1. One large group clusters around 3-4 euros, and a smaller group is around 10 euros. Due to this bimodal distribution, using the average price as a representative metric is inappropriate. Additionally, an outlier at 49 euros is noted.

As shown in Figure 4.2, the main traffic for websites with a pay-or-okay model predominantly comes from Germany. France is the second-largest source of traffic, followed by the rest of the EU. Notably, some sites also receive significant traffic from the US.

Websites with a pay-or-okay model tend to have low popularity scores, see Figure 4.3. The distribution of popularity is left-skewed with a long tail, especially for websites without SMP.

In Figure 4.4, it can be seen that most websites do not have a specified category, possibly due to their generally low popularity. This might also be attributed to Ghostery's categorization method. Among categorized websites, the highest category is News and Portals, followed by Entertainment and Reference.

The distribution of the number of hosts, trackers, and companies appears bell-curved with a left skew in Figure 4.5. Pair plots reveal that the number of hosts, trackers, and companies are interrelated. This is expected since trackers are identified from host names using the Ghostery database, and trackers belong to companies, resulting in correlated variables.

	Total		SMP		No SMP	
	Mean	Std	Mean	Std	Mean	Std
Price	3.769	2.976	3.675	0.466	3.860	4.162
Popularity	0.004	0.016	0.002	0.006	0.007	0.021
Hosts	10.435	6.032	8.659	5.090	12.170	6.381
Trackers	7.857	4.478	6.473	3.680	9.208	4.776
Companies	6.547	3.569	5.441	2.881	7.627	3.844

Table 4.1: Mean and standard deviation for the numeric and boolean focus variables of the total, SMP and no SMP data, $N = 334$.

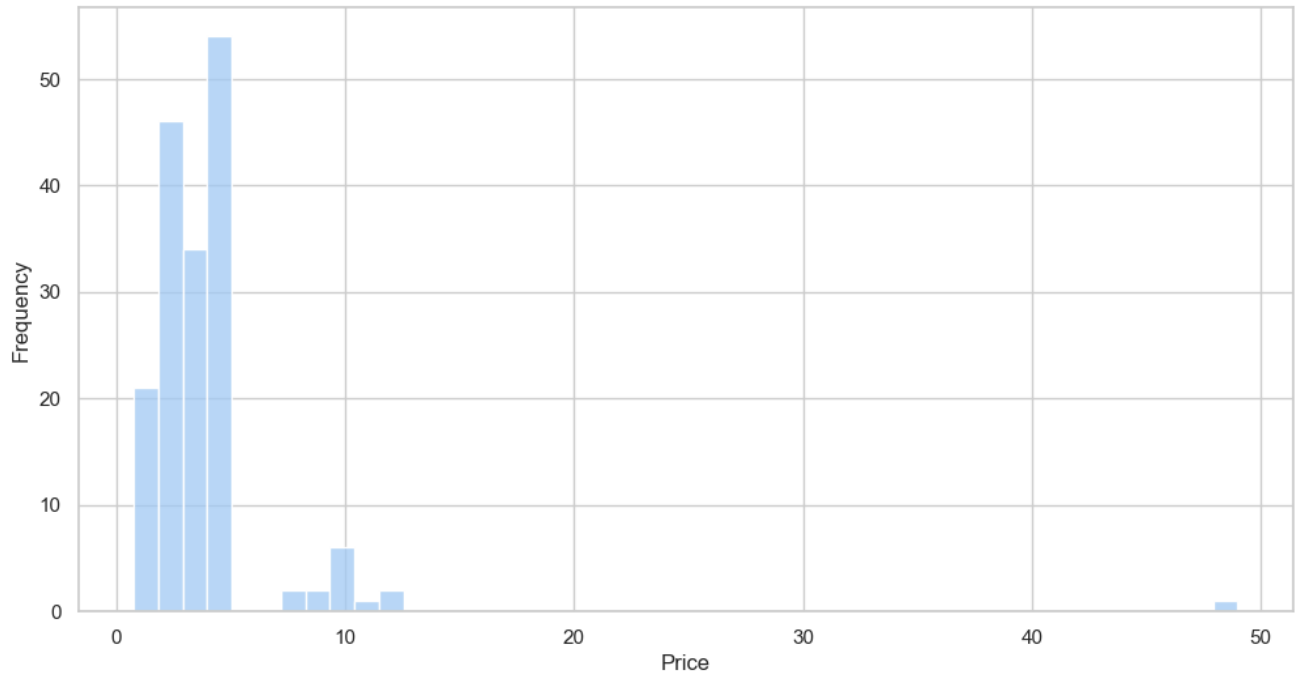


Figure 4.1: Distribution of prices for sites without a SMP, $N = 169$

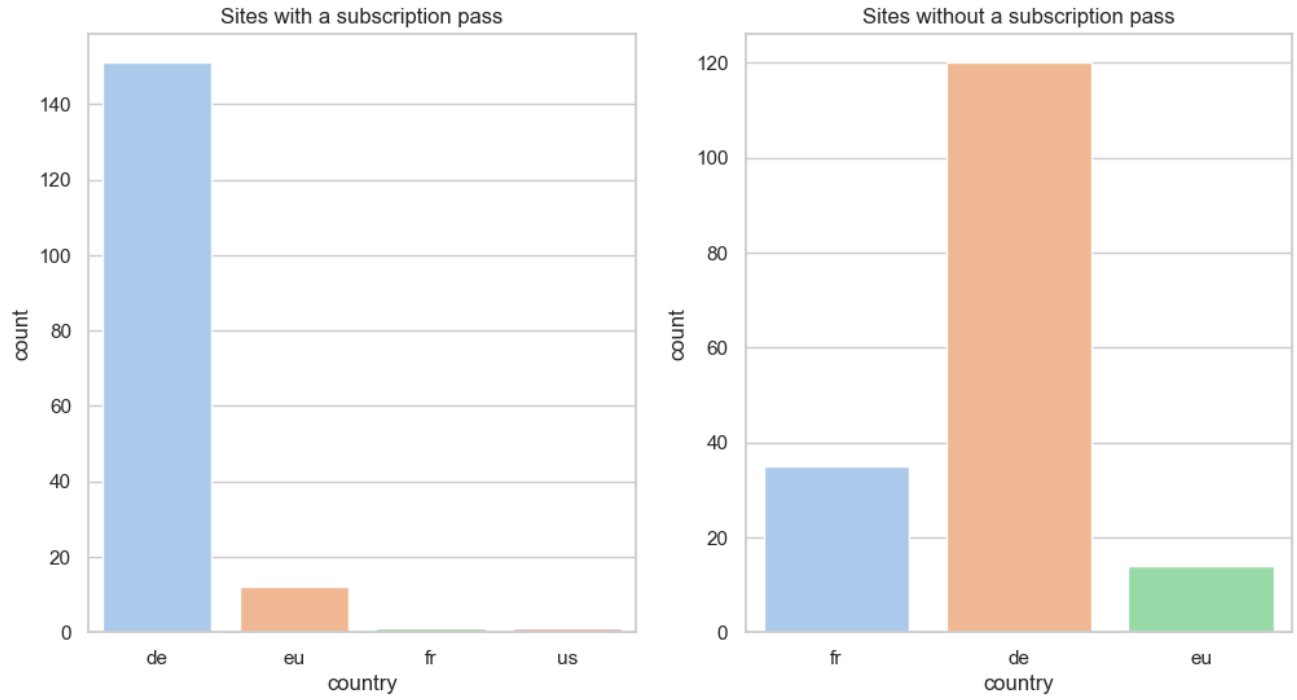


Figure 4.2: Distribution of country for sites with SMP (left) and without SMP (right).

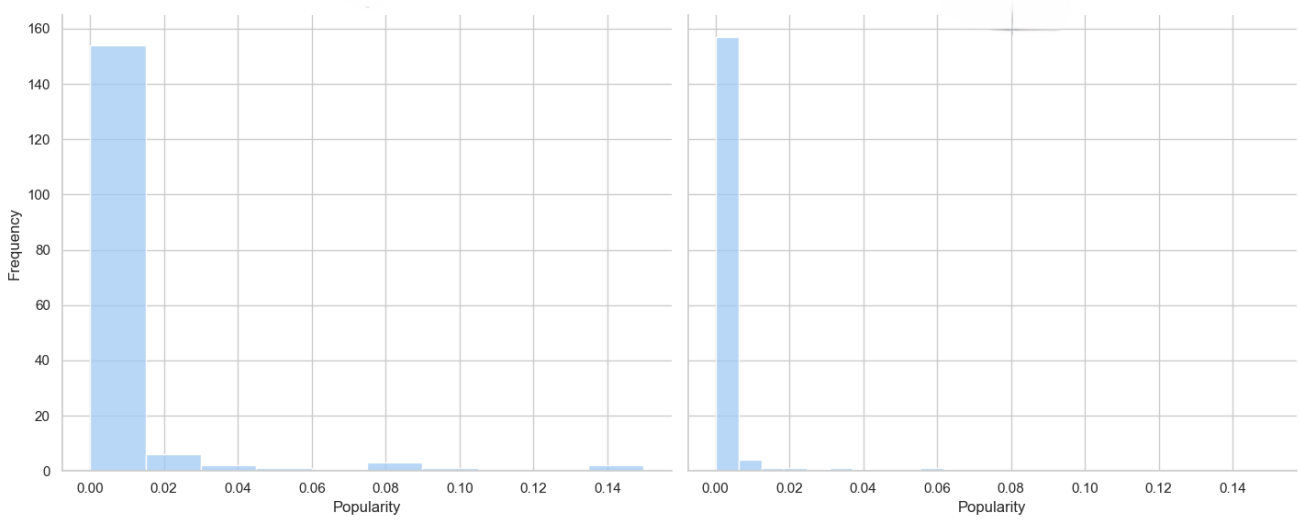


Figure 4.3: Distribution of popularity for sites with SMP (left) and without SMP (right).

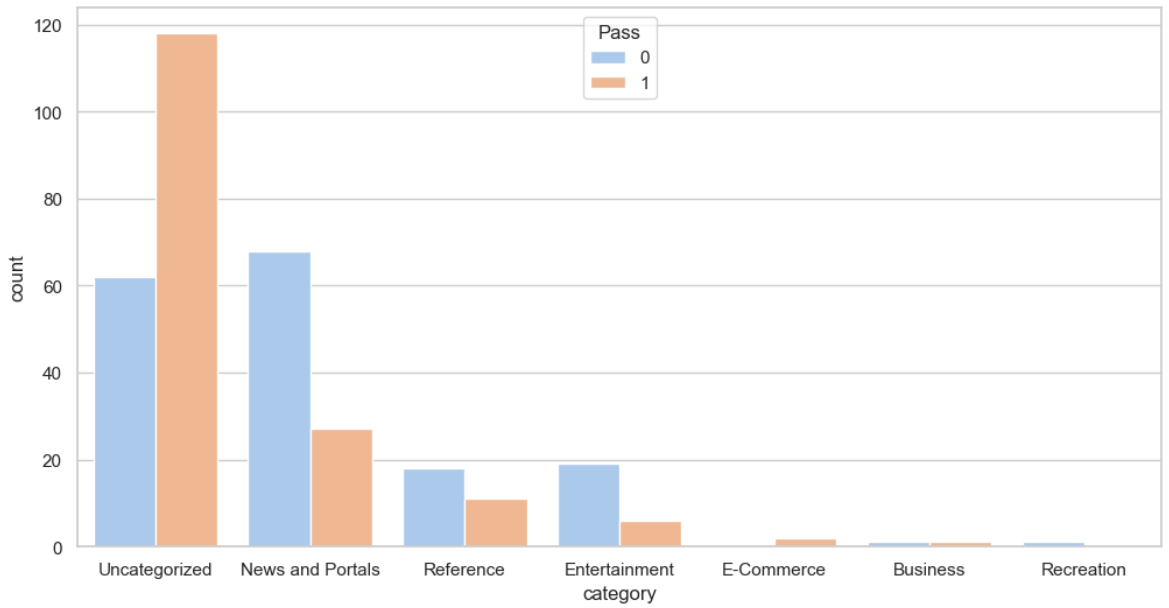


Figure 4.4: Distribution of category for sites with SMP (red) and without SMP (blue).

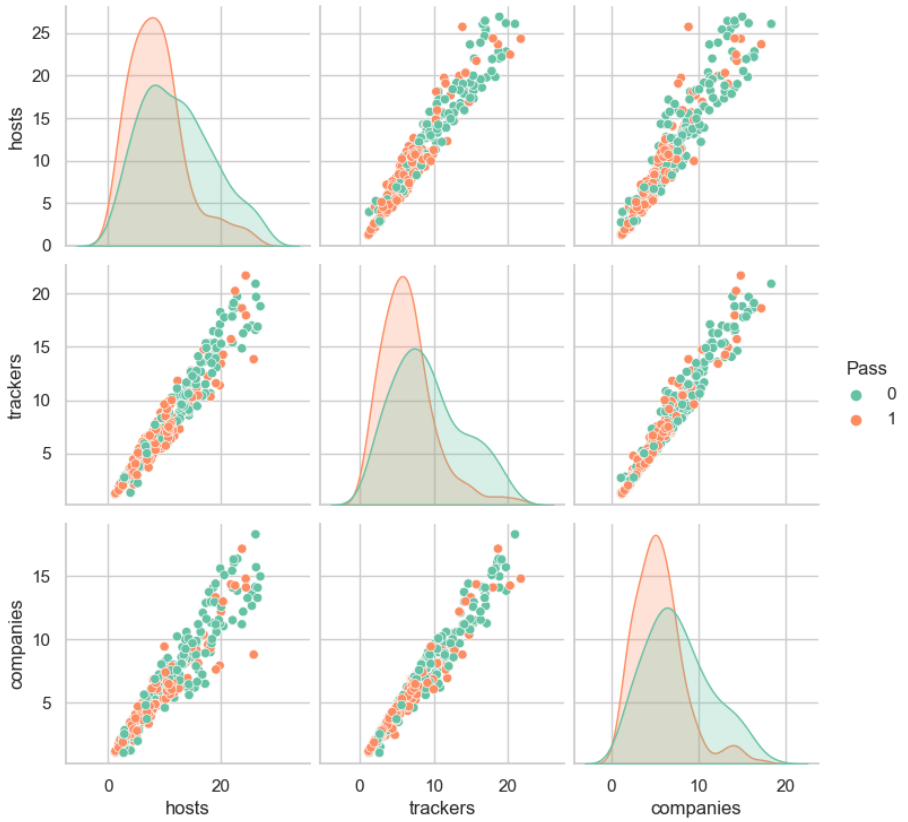


Figure 4.5: Distribution of the host, trackers and companies count plotted against each other, where a site with SMP is noted as red and without as green.

4.2 Subquestion 2

In this subquestion a closer look to the sites where a pay-or-okay subscription bundle is offered. Based of the descriptive analysis, the T-test and the Mann-Whitney U test a few key findings emerge regarding the characteristics of websites using their own pay-or-okay model versus those utilizing a SMP. In appendix D a distribution between the SMPs in the dataset is given.

Websites with their own pay-or-okay model have a slightly lower average price compared to those using SMPs. Websites with a SMP have an average price of 3.67 euro whereas website without SMP have an average price 3.86 However, this difference is not statistically significant (t = -0.465, p = 0.643). Moreover, as we can see in the histogram of price in Figure 4.1 the t-test would not be a good solution, because it would violate the normality assumption. Moreover, the price of the no SMP group has a high standard deviation of 4.16, see Table 4.4. However, the Mann-Whitney U test did show a significant difference in prices between the two groups (U = 11313, p <0.05).

There is a significant difference in popularity scores based of the Mann-Whitney U Test (U = 18643.5, p $<$ 0.05) with websites using a SMP having lower popularity scores ($\bar{x} = 0.00167$) on average compared to those with their own pay-or-okay model ($\bar{x} = 0.006883$). Moreover, the t-test concludes a significant difference in means between the two group. However, we should be careful to make conclusion based on this statistics because the assumption of normality is not met and the standard deviation is high.

Websites with their own pay-or-okay model tend to have more hosts $\bar{x}_{SMP} = 8.66, \bar{x}_{no\_SMP} = 12.2$, trackers $\bar{x}_{SMP} = 6.47, \bar{x}_{no\_SMP} = 9.21$, and companies $\bar{x}_{SMP} = 5.44, \bar{x}_{no\_SMP} = 7.63$ compared to those with a SMP. This difference is statistically significant for the number of hosts (U = 1103.0, p = 0.027), trackers (U = 1004.5, p = 0.006) and for the number of companies (U = 1208.0, p = 0.155). Therefore the H0 of no difference can be rejected in favor of the H1, meaning there is a significant difference in rank. Moreover, the T-test also show a significant difference in the mean of the characteristics between the group, as can be seen in Table 4.2. However, just like noted above, we should be careful with interpreting this results.

In conclusion, the analysis reveals that websites using their own pay-or-okay model tend to have slightly lower average prices compared to those using SMPs, though this difference is not statistically significant. More notably, significant differences were found in popularity scores and

the number of hosts, trackers, and companies, with websites using their own model showing higher values in these categories. These findings, while indicative, should be interpreted cautiously due to the non-normality of the data and high standard deviations. The Mann-Whitney U test results provide a more reliable indication of differences, confirming significant variations in price, popularity, and website characteristics between the two groups.

	T-test		**Mann-Whitney U test**	
Variable	t-value	p-value	W-value	p-value
Popularity	3.074	0.002	18643.5	< 0.05
Price	-0.465	0.643	11313	0.003
Hosts	5.326	< 0.05	18526	< 0.05
Trackers	5.646	< 0.05	18802	< 0.05
Companies	5.631	< 0.05	18837	< 0.05

Table 4.2: Test value and the p-value of the T-test (left) and the Mann-Whitney U test (right)

Gzlideline 6. When sampling from two non-normal distributions of the same shape, i.e., both skewed in the same direction and with eqzl variancer, thme is little disstortion in the obtained t distributions for both equal and unequal size ampje. (from https://journals-sagepub-com.eur.idm.oclc.org/doi/epdf/10.2466/pr0.1974.34.3c.1095)

4.3 Subquestion 3

When looking at the significant results of the logarithmic regression model, the third subquestion can be answered. For the first hypothesis H1, the null hypothesis (H_0) can not be rejected ($p > .05$, $\beta = 0.646$), meaning that popularity does not positively influences the log(Price). For H2, the null hypothesis could not be rejected in all website category ($p > .05$), see also the extended results in appendix E. This meant that no category significantly influences the price of the pay-or-okay model. For H3 and H4 the null hypothesis could be rejected in favor of the alternative hypothesis ($\beta_{hosts} = -0.102$, $p < 0.05$ & $\beta_{companies} = 0.153$, $p < 0.05$). this meant that the number of hosts has a negative effect on the log(Price) and the amount of companies has a positive effect on the log(price). Interestingly, when it comes to the country where the main traffic is coming from France or other part of the EU it negatively affect the log(price). For France it decreases 0.284 ($p < 0.05$) and for other parts of the EU the decrease is higher with

0.490 ($p < 0.05$)

Moreover, Some other effects have been found in our model. When the referer_leaked & referer_leaked_url variable would increase, it respectively positively and negatively effects the log(price). From the documentation of the WhoTracks.Me dataset, the exact definition of referer_leaked, could not have been found.

	Dependent variable:		
	Price	ln(Price)	
	(1)	(2)	(3) Extended
countryeu	0.464	−0.266	−0.490***
	(0.683)	(0.172)	(0.178)
countryfr	0.067	−0.190	−0.284**
	(0.515)	(0.130)	(0.143)
popularity	12.555	2.501	0.646
	(8.573)	(2.159)	(2.166)
referer_leaked			1.437***
			(0.396)
referer_leaked_url			−0.998***
			(0.358)
hosts	−0.265***	−0.077***	−0.102***
	(0.090)	(0.023)	(0.026)
companies	0.404***	0.125***	0.153***
	(0.150)	(0.038)	(0.039)
Observations	168	168	168
R^2	0.154	0.243	0.436
Adjusted R^2	0.100	0.195	0.307
Residual Std. Error	2.153 (df = 157)	0.542 (df = 157)	0.503 (df = 136)
F Statistic	2.858*** (df = 10; 157)	5.048*** (df = 10; 157)	3.391*** (df = 31; 136)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.3: Selection of results for the three models, see appendix E for all the results.

4.4 Subquestion 4

For this subquestion the results of the used models in every step will be given. To compare each model to each other, the best model based on the metrics sensitivity (step 1), accuracy (step 2) and RMSE (step 3) is taken. After choosing the best model, the results of the solution is presented.

4.4.1 Step 1: Pay-or-okay

In step one, we evaluate the models based on their sensitivity to determine which model performs best in identifying true negatives, which is crucial for our goal of minimizing false positives. Table 4.4 shows the metrics for the boosting, lasso, and k-NN models. From the table, we see that the k-NN model has the highest sensitivity, making it the best model for this step. An interesting note is that the sensitivity is also relatively low compared to the other models. For the goal of our model, giving recommendation to managers about a possible use of pay-or-okay model, a lower sensitivity is only better for our goals, have a low amount of False Negative and a higher amount of False Positive.

Name	**Specificity**	**Sensitivity**	**F-measure**
Boosting	0.708	0.882	0.841
Lasso	0.613	0.882	0.814
k-NN	0.774	0.771	0.796

Table 4.4: Specificity, sensitivity and F-measure of the three model for step 1.

4.4.2 Step 2: Pass

In step two, the models are evaluated based on their accuracy, which is the most critical metric for this phase, given the equal importance of true positives and true negatives. Table 4.5 displays the metrics for the boosting (0.693), lasso (0.792), and k-NN models (0.703). From the table, it is evident that the lasso model achieves the highest accuracy with 0.792, making it the best model for this step. The lasso model's performance in specificity is second based behind kNN, but lasso still outperforms both models in terms of sensitivity.

Name	**Accuracy**	**Specificity**	**Sensitivity**
Boosting	0.693	0.653	0.731
Lasso	0.792	0.816	0.769
k-NN	0.703	0.837	0.577

Table 4.5: Accuracy, specificity and sensitivity of the three model for step 2.

4.4.3 Step 3: Price

In step three, the focus is on predicting the competitive price for managers considering deploying their own pay-or-okay model. The models are evaluated using Root Mean Square Error (RMSE), which is a measure that takes the squared root of the average error of the predictions. Moreover, the R-squared and Mean Absolute Error (MAE) are also presented for additional context. Table 4.6 presents the evaluation metrics for the extenden model, lasso and boosting models. The boosting model outperforms the others with the lowest RMSE and MAE, indicating it provides the most accurate and reliable price predictions. The higher R-squared value of the boosting model further indicates the performance in explaining the variance in the pricing data.

Model	**RMSE**	**R-squared**	**MAE**
Extended model	0.747	0.0657	0.512
Lasso	0.577	0.00311	0.451
Boosting	0.536	0.197	0.395

Table 4.6: RMSE, R-squared and MAE of the three model for step 2.

4.4.4 Solution

After having selected the top models for every step, a Shiny solution is created where the schematic approach of Figure 3.2 is made into a interactive application. The solution can also be found online via https://bamthesis2024rff.shinyapps.io/Thesis_Rogier/

In this solution an approach is given where the manager of a website put their site in the form. This URL will be converted to the base domain (i.e. maps.google.com to google.com). This base domain is then matched with all the domain in the WhoTracksMe dataset (N = 27860). Based on these characteristics it will give recommendation on if the website looks similar to other pay-or-okay sites. If this is the case, a prediction if a SMP is needed or use their own pay-or-okay cookiewall. If the later is the case the price will be determined.

In the application an treshold slider and a field where the manager could add their e-mail. The treshold slider is to see if the predictor change if the treshold is lowered, this is done for the SMP model, because the KNN model doesn't work with probabilities and therefore tresholds. The e-mail field is so future leads can be collected. Is has to be noted that the current form is not GDPR complaint.

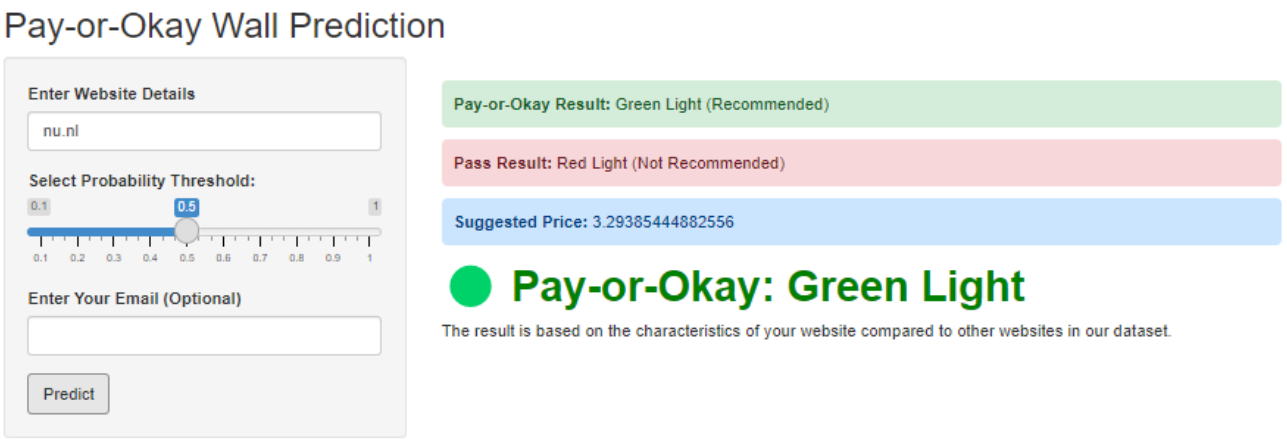


Figure 4.6: Screenshot of how the solution application might look like, with as an example the dutch nu.nl, see for more examples appendix F

This solution will give a starting point from where managers can choose their own next steps. What can be seen from the results in the sections before is that the classification model are not that good with a 77% specificity rate and a 79% accuracy rate. Moreover, predicting the price of a pay-or-okay model has an MAE of 0.395, meaning a mean absolute error of the predicted natural logarithm of prices. Translated to regular prices it would be 1.48 euro. With an average price of 3.86 euro this mean absolute error is high, which could give the manager a under or overestimated price.

5 Discussion

This research extends on the current knowledge on the subject of pay-or-okay. Rasaii et al. (2023), Morel et al. (2022), Morel et al. (2022) & Müller-Tribbensee et al. (2024) Papadopoulos et al. (2020) has already shared some insight into the pay-or-okay landscape, but in this paper the characteristics of website with a pay-or-okay model and even Subscription Management Platforms (SMPs)is further examined.

Rasaii et al. (2023) presented a basic figure that plots the relationship between price and the number of trackers, showing a non-significant downward trend (see appendix G). In this research, we further explore the effect of trackers on the price within the pay-or-okay framework. One key finding from the BVDW, an industry advocacy group, highlighted the need for support in the pay-or-okay practice, particularly in determining pricing and criteria. With the conclusions from our hypotheses and the proposed predictive model, managers can better identify criteria to set a competitive price for a pay-or-okay model.

Additionally, this research proposed a better understanding about the use of SMPs. It can be concluded that the characteristics of these sites significantly differ from other website with pay-or-okay models. Moreover, some descriptive analytics of these website has been given. In research from Morel et al. (2022) and Müller-Tribbensee et al. (2024) some characteristics of the effect of the SMPs is given. However, this research extends this knowledge by showing that the characteristics especially related to tracking of those sites, differ from the characteristics of 'normal' pay-or-okay sites.

In this chapter the research and strategic conclusion of every subquestion and the main research question will be given. These conclusion will be related to the current research. Moreover, the limitation of this study will be elaborated and finally, for recommendation and future research will be presented.

5.1 Research & Strategic conclusion

Based on the first research question the following conclusions can be given. Two price groups can be seen; one group around 3-4 euros and another group at the 10 euros. This shows that the pricing strategy is mostly around 3-4 euro. However if the price is higher it will jump more seldom to around 10 euro and skip over 6 and 7 euros. This division in price is also seen in other research. In Rasaii et al. (2023) and Morel et al. (2023) a slight divide was visible. However in both papers the SMPs were added to the analysis of the price. In this paper these were removed for the price analysis.

The main traffic of the pay-or-okay website are from Germany. This conclusion is in line with the findings in Morel et al. (2023) & Rasaii et al. (2023) that paywalls are prevalent in Germany.

The popularity of websites with pay-or-okay models seems to be on the lower side. This insight is new for the research. To the contrary, in the paper of Rasaii et al. (2023) it is concluded that also popular sites has pay-or-okay model.

The second highest category being News and Portal can be verified in other studies. Morel et al. (2023) Rasaii et al. (2023) both came to the same conclusion. From the left-skewed bell-curved distribution of the host, trackers and companies of those trackers no new information is gathered.

In the second research question the difference between SMP websites and other pay-or-okay sites is analysed. No previous extended research has been done on this subject. It can be concluded that the hypothesis that there is no difference between the two group can be rejected for every characteristic in favor of the alternative hypothesis, however the difference in mean can not be concluded because of the violation of the normality distribution and the high variance. In this research some descriptive analysis with descriptive difference between bundle and no bundle website has been given. From this, four insight can be given. (1) the popularity of SMP websites seems to be lower than other sites, (2) less hosts, trackers and companies are on website with a SMP than no SMP websites. This finding is interesting because other research from Rasaii et al. (2023) & Morel et al. (2023) has found the opposite and has seen that with *Contentpass* more cookies and trackers are send and therefore these websites rely more on advertisement than other sides. From our results the trackers are less on SMP sites compared to other pay-or-okay

websites. (3) Website where the traffic comes from Germany has more SMP sites than no bundle sites, where also traffic comes from EU and FR. This conclusion was also coined in Morel et al. (2023). A reason for this could be the german-based SMPs *Contentpass* and *Freechoice*. (4) More SMP websites has a category unknown than news sites, a reason for this is the lower popularity of SMP sites.

After knowing the characteristics of pay-or-okay website, we examined the effect of them on their price. It has been found that the popularity does not influences the log(price). Meaning that the amount of people that visit the website does not matter to the Price. The price between categories do not have a significant difference. This could be because of the low amount of difference in the dataset or the way the categories are labels by Ghostery. An interesting effect between hosts and tracking companies can be found. the amount of hosts on the website has an negative impact on the price, meaning website with higher hosts has a lower price. This is in line with the plot from Rasaii et al. (2023), where more cookies has a downwards trend on price. However, the amount of companies of the trackers a website has, positively impact the Price. This could mean that the more data broker companies a website is affiliated with the more the Price will increase. This suggest that the value of your data is more worth on those sites because it is sold to other more companies. lastly, the higher price in Germany, compared to other European companies highlights the maturity of pay-or-okay and willingness to pay higher prices. Websites owner in other european countries tends to price their websites lower. A deeper understanding of what effects the characteristics of price is an extension from the current (small scale) research from Rasaii et al. (2023), where the amount of cookies was plotted against the price. In this research it can be concluded that Hosts has a positive effects on the log(price), however the amount of companies has a negative effect on the log(Price). These conflicting conclusion needs to be further investigated.

In the last question, three predictive models are designed for managers to make better decisions when it comes to pay-or-okay models for their sites. Looking at the results of these models it is still hard to accuracly predict is a website would fit in a pay-or-okay environment and if there is a model if it should use a SMP or not. Both the specificity and the accuracy of the two models are relatively low. Moreover, the RMSE of the log(price) is high, so the prediction for a website is still lacking accuracy. Manager could make use of the application to get a really rough estimation, but in the end the landscape needs to mature even more in the future to get

more accurate results. In research by Papadopoulos et al. (2020) a predictive models has already been made. However, unlike this research, where privacy features was used, Papdopoulos' model is made from websites, textual, structural and visual features. The results of this model was 77% on both precision and recall. In this research characteristics of the trackingof a website is used for the model. With a recall of 77% this model performs in line with the model coined in previous research. Moreover, the model predicting an competitive price estimation could be better, however the question that the BVDW has, can be answered by this research.

5.2 Limitations

This study has several limitations that should be considered when interpreting the findings. Firstly, the analysis was constrained to information from the top 10,000 websites. Initially, the dataset 968 websites employing a pay-or-okay model, but this number was reduced to 334 by the time the dataset was finalized. This reduction potentially limits the generalizability of the findings.

Secondly, the recommendations provided in this study are based on the characteristics of current websites with pay-or-okay models. This reliance on existing data could lead to a self-fulfilling prophecy, as the recommendations may not apply to future websites that might adopt this model. Additionally, the predictive models were trained on only 50% of the 334 pay-or-okay websites, which could affect their accuracy and robustness.

Thirdly, there was a selection bias in the prediction models due to the need to match site characteristics with websites that have a pay-or-okay cookie wall. This requirement significantly reduced the number of instances available for analysis, from an initial total of 968 identified pay-or-okay websites.

Moreover, technical issues with the Bannerclicker tool hindered the discovery of new websites for research. Despite efforts to resolve these issues, including making changes to the Bannerclick repository, setting up a Linux container, and testing different Firefox versions, the tool did not function as required.

Additionally, this study focused solely on data from WhoTracks.Me, which primarily includes tracking information. Other data sources, such as WHOIS data and HTML cookie banner data used in previous research were not included. This limitation means that the insights gathered

are restricted to tracking information from WhoTracks.Me, and further research is needed to generalize these findings to broader tracking statistics.

Finally, the analysis primarily focused on five characteristics, suggesting that more research is needed to explore additional variables. The normality assumption for linear models was not met, which could pose a problem for generalizing the results. Furthermore, the high RMSE of the model indicates poor prediction accuracy, highlighting the need for further refinement.

It is important to note that not all websites are up to date, as the data comes from earlier research, but most are double checked. However, this provides a sense of the earlier landscape of pay-or-okay models.

5.3 Recommendation & Further research

Based on the research three recommendation can be done by research. (1) When looking at implementing the pay-or-okay model for your site, think about your characteristics and if it fits into the current competitive landscape. (2) If the popularity of your website seems lower or the amount of trackers you have is lower then consider using a SMP for your pay-or-okay model. (3) Lastly, when choosing a competitive price for your pay-or-okay cookiewall think about the division between the 3-4 euro pricetag and the 10 euro price. Combining the recommendations, a manager could make use of the made Shiny applicationto get a competitive insight of their own site.

However, this research could not answer all the managers question. Future research should aim to address the limitations identified in this study. Expanding the analysis to include additional variables from WhoTracks.Me could provide a more comprehensive understanding of the factors influencing pay-or-okay models.

Research should also explore the characteristics of pay-or-okay models compared to websites without such models. This comparison could offer valuable insights into the unique attributes and effectiveness of pay-or-okay strategies.

Moreover, collecting more data on the number of marketing cookies used by websites could refine the assumptions made in this study. This additional data could enhance the accuracy of the predictive models and provide a clearer picture of the pay-or-okay landscape.

Utilizing the Ghostery extension to scrape more data from websites could also be beneficial.

Although recent updates to Ghostery's tracking abilities and user interface posed challenges during this research, overcoming these obstacles in future studies could yield richer datasets and more robust analyses.

Finally, further efforts should be made to resolve technical issues with tools like Bannerclicker to facilitate the discovery and analysis of new websites. This, combined with an expanded dataset and more sophisticated models, could significantly improve the accuracy and applicability of predictive models for pay-or-okay strategies.

References

Autoriteit Persoonsgegevens. (2024, June). *AP pakt misleidende cookiebanners aan.* Retrieved 2024-04-04, from https://www.autoriteitpersoonsgegevens.nl/actueel/ap-pakt-misleidende-cookiebanners-aan

BVDW. (2023, October). *PUR Models, Status quo on the European Market* (Tech. Rep.). BVDW. Retrieved 2024-02-24, from https://iabeurope.eu/knowledge_hub/bvdws-comprehensive-market-overview-pur-models-in-europe-legal-framework-and-future-prospects-in-english/

CNIL. (2022, May). *Cookie walls : la CNIL publie des premiers critères d'évaluation.* Retrieved 2024-02-20, from https://www.cnil.fr/fr/cookie-walls-la-cnil-publie-des-premiers-criteres-devaluation

CommandersAct. (2022, May). *Privacy Barometer 2022* (Tech. Rep.). CommandersAct. Retrieved 2024-04-25, from https://commandersact.com/en/the-quantum-lounge/enterprise-tag-manager/consent-management/privacy-barometer-2022/ (Section: Consent Management)

Datanschutzenkonferenz. (2023). *Beschluss der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder vom 22. März 2023.* Retrieved 2024-02-20, from https://datenschutzkonferenz-online.de/media/pm/DSK_Beschluss_Bewertung_von_Pur-Abo-Modellen_auf_Websites.pdf

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2019). We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings 2019 Network and Distributed System Security Symposium.* Retrieved 2024-01-15, from http://arxiv.org/abs/1808.05096 (arXiv:1808.05096 [cs]) doi: 10.14722/ndss.2019.23378

Ernst, A. (2023, December). *Ist Meta´s „Pay or Okay"-Methode rechtswidrig?* Retrieved 2024-02-24, from https://www.datenschutz-notizen.de/ist-metas-pay-or-okay-methode-rechtswidrig-1345660/ (Section: Allgemein)

Folks, A. (2024, September). *Meta's new digs: A deep dive into practical considerations of consent.* Retrieved 2024-01-15, from https://iapp.org/news/a/metas-new-digs-a-deep-dive-into-practical-considerations-of-consent/

Global Data and Marketing Alliance. (2022). *Global Data Privacy: What The Customer Really Thinks* (Tech. Rep.). Global Data and Marketing Alliance. Retrieved 2024-03-05, from file:///C:/Users/megar/Zotero/storage/L357WYKS/global-privacy-monitor-2022-consumers-more-data-conscious-privacy-concerns-decline.html

Havlicek, L. L., & Peterson, N. L. (1974, June). Robustness of the T Test: A Guide for Researchers on Effect of Violations of Assumptions. *Psychological Reports*, *34*(3_suppl), 1095–1114. Retrieved 2024-06-15, from https://doi.org/10.2466/pr0.1974.34.3c.1095 (Publisher: SAGE Publications Inc) doi: 10.2466/pr0.1974.34.3c.1095

Hofstad, N., & Lundqvist, A. (2021). *Cookies, cookies everywhere! : A qualitative interview study about how internet users interact with cookie consent notices.* Retrieved 2024-02-18, from https://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-84344

IAB Europe. (2023, November). *BVDW's Comprehensive market overview: PUR models in Europe – legal framework and future prospects in English - IAB Europe.* Retrieved 2024-02-29, from https://iabeurope.eu/knowledge_hub/bvdws-comprehensive-market-overview-pur-models-in-europe-legal-framework-and-future-prospects-in-english/

IAPP. (2024, February). *Pay or OK: Practical considerations for adtech and beyond.* Retrieved 2024-03-05, from https://www.linkedin.com/company/iapp---international-association-of-privacy-professionals/posts/?feedView=all

Karaj, A., Macbeth, S., Berson, R., & Pujol, J. M. (2019, April). *WhoTracks .Me: Shedding light on the opaque world of online tracking.* arXiv. Retrieved 2024-03-04, from http://arxiv.org/abs/1804.08959 (arXiv:1804.08959 [cs]) doi: 10.48550/arXiv.1804.08959

Kulyk, O., Hilt, A., Gerber, N., & Volkamer, M. (2018). "This Website Uses Cookies": Users' Perceptions and Reactions to the Cookie Disclaimer. In *Proceedings 3rd European Workshop on Usable Security.* London, England: Internet Society. Retrieved 2024-02-18, from https://www.ndss-symposium.org/wp-content/uploads/2018/06/eurousec2018_12_Kulyk_paper.pdf doi: 10.14722/eurousec.2018.23012

Laub, R., Miller, K. M., & Skiera, B. (2023, March). *The Economic Value of User Tracking for Publishers.* arXiv. Retrieved 2024-03-28, from http://arxiv.org/abs/2303.10906

(arXiv:2303.10906 [econ, q-fin]) doi: 10.48550/arXiv.2303.10906

Leiser, M. R., & Santos, C. (2023, April). *Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2024-03-07, from https://papers.ssrn.com/abstract=4431048

Morel, V., Santos, C., Fredholm, V., & Thunberg, A. (2023, November). Legitimate Interest is the New Consent - Large-Scale Measurement and Legal Compliance of IAB Europe TCF Paywalls. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society* (pp. 153–158). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-02-20, from https://dl.acm.org/doi/10.1145/3603216.3624966 doi: 10.1145/3603216.3624966

Morel, V., Santos, C., Lintao, Y., & Human, S. (2022, November). Your Consent Is Worth 75 Euros A Year - Measurement and Lawfulness of Cookie Paywalls. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society* (pp. 213–218). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-02-13, from https://dl.acm.org/doi/10.1145/3559613.3563205 doi: 10.1145/3559613.3563205

Müller-Tribbensee, T. (2023). Bezahlen für mehr Privatsphäre: Verbreitung und Ausgestaltung des Pur-Modells bei deutschen Medien. In Deutscher Dialogmarketing Verband e.V. (Ed.), *Dialogmarketing Perspektiven 2022/2023: Tagungsband 15. wissenschaftlicher interdisziplinärer Kongress für Dialogmarketing* (pp. 211–228). Wiesbaden: Springer Fachmedien. Retrieved 2024-02-29, from https://doi.org/10.1007/978-3-658-40753-7_10 doi: 10.1007/978-3-658-40753-7_10

Müller-Tribbensee, T., Miller, K. M., & Skiera, B. (2024, March). *Paying for Privacy: Pay-or-Tracking Walls* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2024-03-07, from https://papers.ssrn.com/abstract=4749217

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020, April). Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-04-25, from https://doi.org/10.1145/3313831.3376321 doi: 10.1145/3313831.3376321

NOYB. (2024, February). *28 NGOs urge EU DPAs to reject "Pay or Okay" on Meta.* Retrieved 2024-02-29, from https://noyb.eu/en/28-ngos-urge-eu-dpas-reject-pay-or

-okay-meta

Papadopoulos, P., Snyder, P., Athanasakis, D., & Livshits, B. (2020, May). *Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls.* arXiv. Retrieved 2024-02-20, from http://arxiv.org/abs/1903.01406 (arXiv:1903.01406 [cs]) doi: 10.48550/arXiv.1903.01406

Rasaii, A., Gosain, D., & Gasser, O. (2023, October). Thou Shalt Not Reject: Analyzing Accept-Or-Pay Cookie Banners on the Web. In *Proceedings of the 2023 ACM on Internet Measurement Conference* (pp. 154–161). Montreal QC Canada: ACM. Retrieved 2024-01-15, from https://dl.acm.org/doi/10.1145/3618257.3624846 doi: 10.1145/3618257.3624846

Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P.-A., & Santos, I. (2019, July). Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* (pp. 340–351). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-02-18, from https://doi.org/10.1145/3321705.3329806 doi: 10.1145/3321705.3329806

Schmitt, J., Miller, K., & Skiera, B. (2020). The Impact of Privacy Laws on Online User Behavior. *SSRN Electronic Journal.* Retrieved 2024-02-13, from https://www.ssrn.com/abstract=3774110 doi: 10.2139/ssrn.3774110

Spinner, Y. (2023, October). *Meta kondigt abonnement voor Facebook en Instagram zonder reclame aan.* Retrieved 2023-11-30, from https://tweakers.net/nieuws/215092/meta-kondigt-abonnement-voor-facebook-en-instagram-zonder-reclame-aan.html

Tavan, C. (2018, January). *How Contentpass Handles User Data.* Retrieved 2024-02-25, from https://medium.com/contentpass/how-contentpass-handles-user-data-40985bb6051a

The Court of Justice of the European Union. (2023, July). *Case C-252/21.*

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019, November). (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (pp. 973–990). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-04-04, from https://dl.acm.org/doi/10.1145/3319535.3354212 doi: 10.1145/3319535.3354212

Wang, P., Jiang, L., & Yang, J. (2024, February). The Early Impact of GDPR Compliance on Display Advertising: The Case of an Ad Publisher. *Journal of Marketing Research*, *61*(1), 70–91. Retrieved 2024-02-29, from https://doi.org/10.1177/00222437231171848 (Publisher: SAGE Publications Inc) doi: 10.1177/00222437231171848

White, M. (2023, April). *2023, the year of the registration wall: how publishers can succeed in their registration strategy.* Retrieved 2024-02-24, from https://thefix.media/2023/3/21/2023-the-year-of-the-registration-wall-how-publishers-can-succeed-in-their-registration-strategy

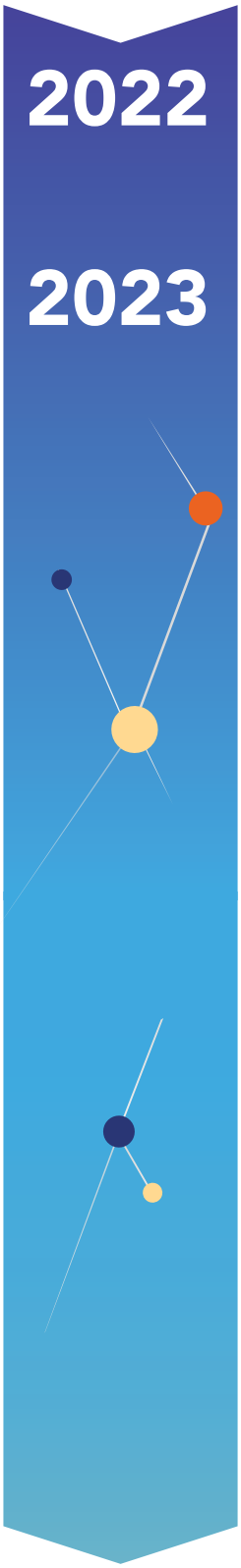
Appendix

Appendix A: Prices from (Morel et al., 2022)

Table 3: Prices of cookie paywalls

	DerStandard	Krone	Kurier	Vol	Vienna	Spiegel	Bild	T-online	Welt	Zeit	Heise	Stern	Rp-online
Price in € (normalized over a year)	75	59.88	43.2	62.4	62.4	59.88	47.88	35.88	47.88	62.4	59.4	59.88	52
Price in € (normalized over a month)	6.25	4.99	3.6	5.2	5.2	4.99	3.99	2.99	3.99	5.2	4.95	4.99	4.33
Notes	1€/month for 3 months then 8€/month	4.99€/month	3.6€/month	1,2€/week	1,2€/week	4,99€/month	3,99€/month	2.99€/month	3,99€/month	1,2€/week	4,95€/month	4,99€/month	1€/week

Appendix B: Timeline Meta Case



5 December

EDPB Binding Decisions concerning Meta services Facebook and Instagram. The EDPB considers that contract is not an appropriate legal basis for processing of personal data for behavioural advertising purposes.

31 December

Irish DPA Final Decisions adopted on the basis of the EDPB Binding Decisions. They included orders to bring processing into compliance within three months.

April

Meta informs Irish DPA that it is changing the legal basis for behavioural advertising from contract to legitimate interest; DPAs begin assessing compliance of this change.

May

Norwegian DPA sends request for mutual assistance to the Irish DPA, asking it to issue a temporary ban on the processing of personal data for behavioural advertising based on legitimate interest and to share a timeline on how the Irish DPA will ensure that Meta complies with the GDPR.

June

Irish DPA informs other DPAs that it is awaiting the judgment of the Court of Justice of the European Union in the Bundeskartellamt case and that it will finalise its assessment of Meta's compliance with its final decisions afterwards.

4 July

CJEU Buderskartellamt Judgement relating to Meta, including the legal basis used for their processing carried out for advertising purposes.

11 July

Irish DPA informs other DPAs of its preliminary conclusion that Meta has not complied with its decisions. No corrective measures to stop the processing were envisaged.

14 July

Norwegian DPA imposes temporary ban on Meta's processing. In accordance with Art. 66 (1) GDPR, the Norwegian DPA takes urgent measures and bans Meta from processing the personal data of Norwegians for the purposes of behavioural advertising on the legal bases of contract and legitimate interest. This ban takes effect on August 4th, is limited to three months and only applies in Norway.

26 September

Norwegian DPA sends request for final measures to the EDPB. The Norwegian DPA requests the EDPB to adopt an urgent binding decision pursuant to Art. 66 (2) GDPR in order to implement final measures that would apply to the whole EEA. The EDPB Secretariat starts the assessment of the completeness of the file.

13 October

The file is complete. The two weeks for the EDPB to adopt its urgent binding decision start.

27 October

EDPB urgent binding decision: EDPB considers that there is a need for urgent measures and instructs Irish DPA as lead supervisory authority (LSA) to issue a ban on Meta's processing of personal data for behavioural advertising purposes on the legal bases of contract and legitimate interest across the EEA.

10 November

Irish DPA adopts the decision imposing a ban on Meta's processing. Meta is given one week to implement the ban, as of the notification of the national decision to them.

Appendix C: All variable in the WhoTracksMe dataset

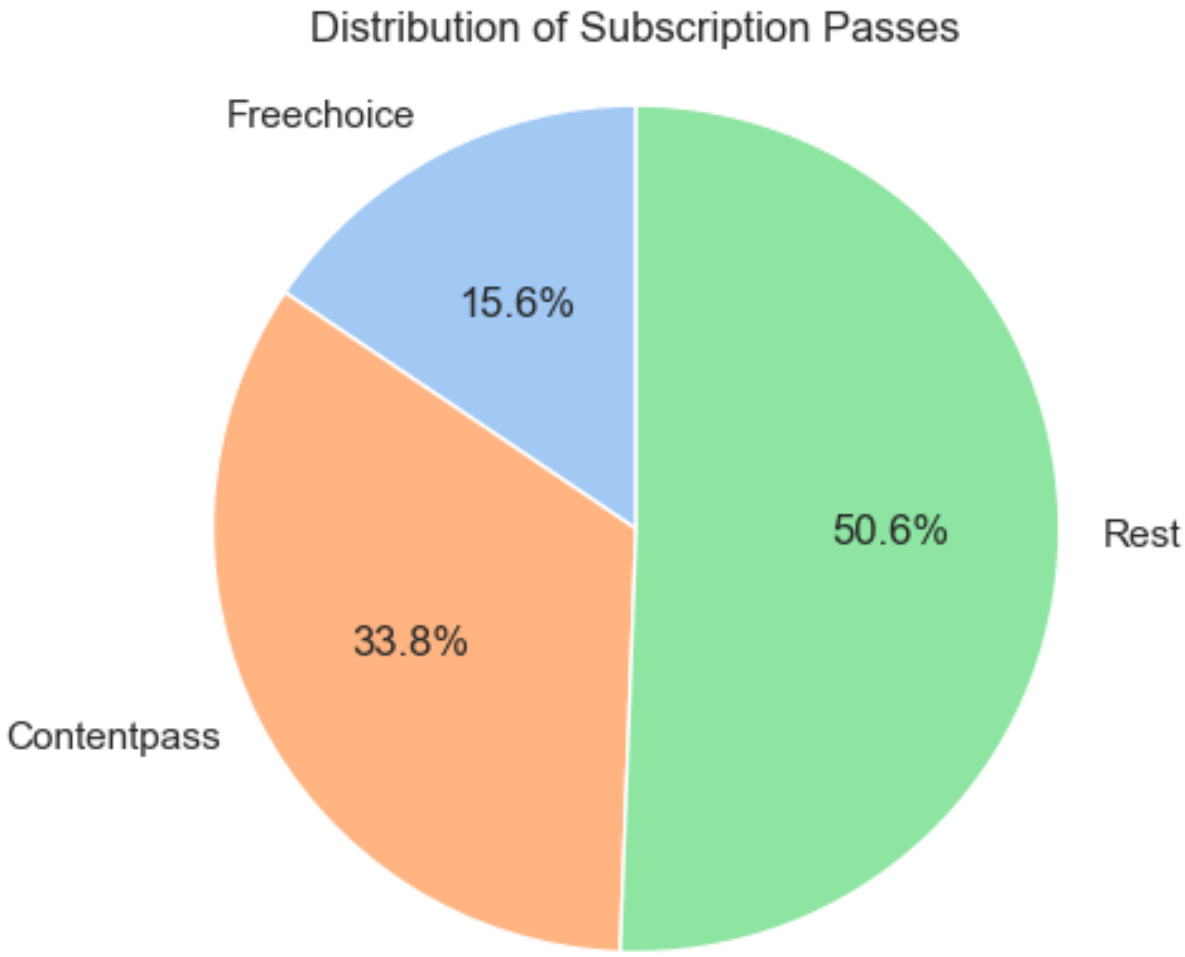
Variable	Description
site	The domain of the website. One of the most frequently visited websites, not generated by Alexa or similar services but by real users. String.
price	The cost associated with the pay-or-okay no-tracking subscription. Float.
category	The site's category according to Ghostery. String.
popularity	The relative amount of traffic compared to the most popular site (google.com) (0-1). Float.
country	The main region where the traffic originates (e.g., global, US, EU, DE, FR). String.
month	Month of observation. Format: mm-yyyy. String/Date.
freechoice	Indicates if the website has a Freechoice option. 1 if 'yes', 0 if 'no'. Binary.
contentpass	Indicates if the website has a Contentpass option. 1 if 'yes', 0 if 'no'. Binary.
pass	Indicator if a website uses a Subscription Management Platform (SMP). Binary.
cookies	Proportion of pages where a cookie was sent by the browser or a Set-Cookie header was returned by the tracker's server. Float between 0 and 1.
bad_qs	Proportion of pages where a unique identifier (UID) was detected in the query string parameters sent with a request to this tracker. Float between 0 and 1.
tracked	Proportion of pages where a UID transmission was detected, either via cookies or bad_qs. Float between 0 and 1.
https	Proportion of pages where the tracker only used HTTPS traffic. Float between 0 and 1.

Table 5.1: Overview of WhoTracksMe variables

Variable	Description
requests	Average number of requests made to the tracker per page. Positive float.
requests_tracking	Average number of requests made to the tracker with tracking (cookie or query string) per page. Positive float.
content_length	Average of Content-Length headers received per page. Expressed in bytes. Positive float.
requests_failed	Average number of requests made to the tracker per page that did not succeed. Positive float.
has_blocking	Proportion of pages where some kind of external blocking of the tracker was detected. Float between 0 and 1.
script	Proportion of page loads where a script was loaded by the tracker.
iframe	Proportion of page loads where an iframe was loaded by the tracker.
beacon	Proportion of page loads where a beacon was used by the tracker.
image	Proportion of page loads where an image was loaded by the tracker.
stylesheet	Proportion of page loads where a stylesheet was loaded by the tracker.
font	Proportion of page loads where a font was loaded by the tracker.
xhr	Proportion of page loads where an XMLHttpRequest or fetch API was used by the tracker.
plugin	Proportion of page loads where a plugin was used by the tracker.
media	Proportion of page loads where media (video or audio) was loaded by the tracker.
hosts	The average number of tracker domains on the site. Positive float.
trackers	The average number of trackers on the site. Trackers are third-party domains present on multiple ($>$ 10) websites with significant combined traffic and use cookies or fingerprinting methods to transmit user identifiers. Positive float.
companies	The average number of companies on the site. Positive float.
referer_leaked_header	Indicates if the full site URL was sent to a tracker via the Referer HTTP header. Average of the binary observations.
referer_leaked_url	Indicates if the full site URL was sent to a tracker via URL parameters. Average of the binary observations.
referer_leaked	Indicates if either referer_leaked_header or referer_leaked_url occurred. Average of the binary observations.
cookie_samesite_none	Indicates if the cookie attribute SameSite: None was sent, allowing cross-site tracking. Binary.
t_active	Measure of time active on the site (with the tab open and focused). Positive float.

Table 5.2: Overview of WhoTracksMe variables

Appendix D: Pie of distribution between the different SMPs and the rest of the sites



Appendix E: Extended results of the extended model

	Dependent variable:		
	Price	log(Price)	
	(1)	(2)	(3)
countryeu	0.464	−0.266	−0.490***
	(0.683)	(0.172)	(0.178)
countryfr	0.067	−0.190	−0.284**
	(0.515)	(0.130)	(0.143)
categoryEntertainment	1.261	0.654	−0.592
	(2.225)	(0.560)	(1.372)
categoryNews and Portals	2.598	0.934*	−0.306
	(2.214)	(0.558)	(1.381)
categoryRecreation	3.244	1.234	0.016
	(3.092)	(0.779)	(1.492)
categoryReference	1.823	0.755	−0.332
	(2.262)	(0.570)	(1.383)
categoryUncategorized	2.623	0.937*	−0.340
	(2.209)	(0.556)	(1.381)
popularity	12.555	2.501	0.646
	(8.573)	(2.159)	(2.166)
bad_qs			0.873
			(0.781)
tracked			0.120
			(0.305)
https			1.429
			(1.407)
requests_tracking			−0.004
			(0.003)
content_length			−0.000
			(0.000)
requests_failed			−0.0001
			(0.001)
has_blocking			0.090
			(0.285)
script			0.118
			(0.501)
iframe			−0.235
			(0.243)
beacon			−0.330
			(0.308)
image			0.089
			(0.188)

	Dependent variable:		
	Price	log(Price)	
	(1)	(2)	(3)
stylesheet			−0.074
			(0.200)
font			0.156
			(0.186)
xhr			0.080
			(0.218)
plugin			33.187
			(79.706)
media			0.904
			(0.738)
referer_leaked			1.437***
			(0.396)
referer_leaked_header			−0.027
			(0.361)
referer_leaked_url			−0.998***
			(0.358)
cookie_samesite_none			0.237
			(0.365)
t_active			0.00000
			(0.00000)
hosts	−0.265***	−0.077***	−0.102***
	(0.090)	(0.023)	(0.026)
companies	0.404***	0.125***	0.153***
	(0.150)	(0.038)	(0.039)
Constant	1.230	0.255	−0.521
	(2.215)	(0.558)	(0.693)
Observations	168	168	168
R^2	0.154	0.243	0.436
Adjusted R^2	0.100	0.195	0.307
Residual Std. Error	2.153 (df = 157)	0.542 (df = 157)	0.503 (df = 136)
F Statistic	2.858*** (df = 10; 157)	5.048*** (df = 10; 157)	3.391*** (df = 31; 136)
Note:			*p<0.1; **p<0.05; ***p<0.01

Appendix F: Screenshot of the Shiny application for managers

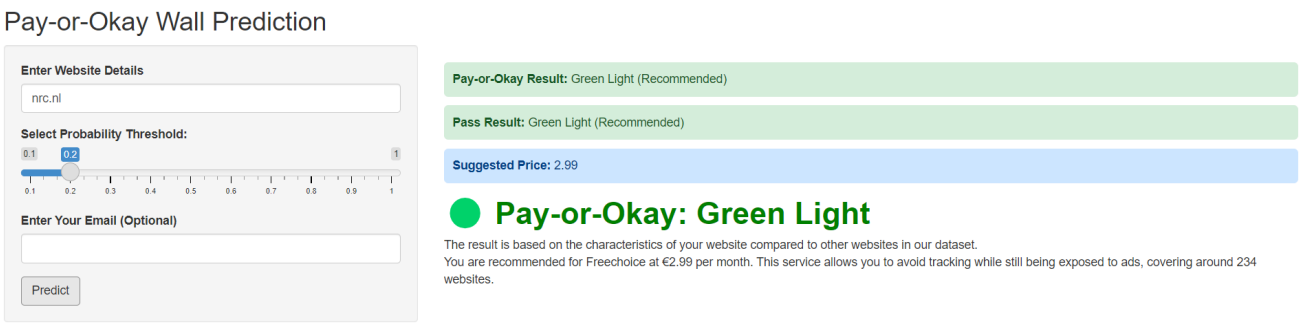
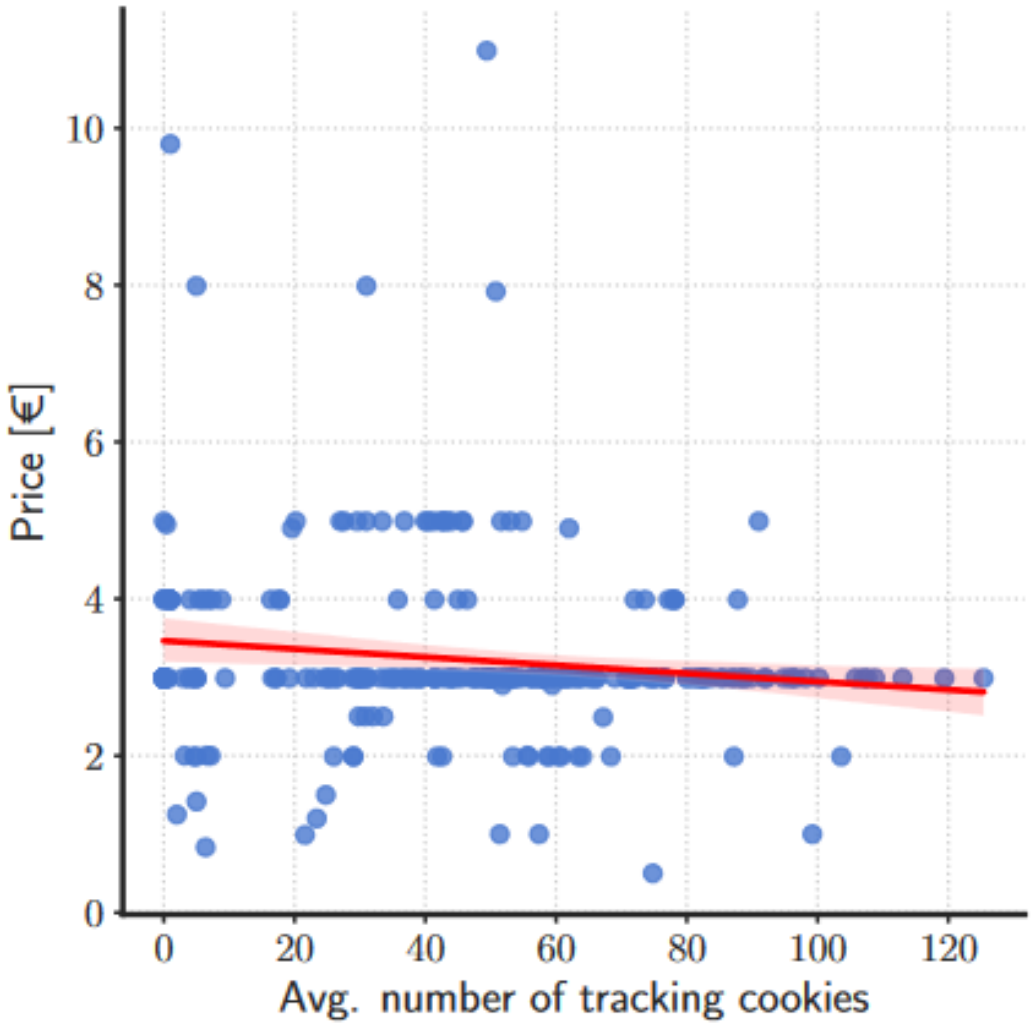


Figure 5.1: Screenshot with NRC.nl in the Shiny application. It is recommended that they use pay-or-okay models, more specifcally SMP

Appendix G: From Rasaii et al. (2023): "Correlation between the number of tracking cookies and price of cookiewall website subscriptions."



Appendix H: Link to the repository of the code and Shiny application from this research

Repository: https://github.com/RogierFrans/Scriptie

Shiny application: https://bamthesis2024rff.shinyapps.io/Thesis_Rogier/