Master Thesis – MSc in Data Science and Marketing Analytics 2019/2020

Beyond the Dashboard

Customer Profiling on the basis of Google Analytics Data using Machine Learning Clustering Algorithms in R

Name Student: Kubilay Ozan Tuerker
Student ID number: 411337ot

Supervisor: Dr. Michel van de Velden
Second Assessor: Dr. Vardan Avagyan

Date final version: 15.08.2020

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of Contents

1. Introduction: "Data Driven", a misunderstood term in the industry?

The inspiration for the topic of my thesis came from my experience working in the marketing industry as an Online Marketing Manager for an agency in Berlin, Germany. During my tenure, I was intensively engaged in visiting marketing fairs and meet-ups, where I was constantly in contact with marketers from other agencies having many insightful conversations about offered services, used tools and industry opportunities & problems. Next to these conversations, which let me dive deeper into the common practices of the industry, I also gained a lot of insight into unsatisfied client needs working with a broad range of online business. Reflecting on my experience infield and contrasting it to the knowledge I have gained during my graduate studies in Data Science and Marketing Analytics, I came to the following realization: The vast majority of agencies I was in contact with utilize buzz words like "performance-oriented", "data agile", and "data driven" in their website copies. Nonetheless, I hardly came across any agency whose data-supported marketing practices exceeded the mere analysis of pre-processed metrics in analytics tools like the Google Analytics Dashboard, with the main purpose to assess the performance of online ads or marketing funnels.

The Google Analytics dashboard offers a tremendous ease of use as marketing insights can be swiftly accessed and interpreted without data wrangling and coding skills. But especially this ease of use and simplicity are hindering advanced data analytical applications. Users are limited to the statistics offered in the dashboard and cannot create their own, data sets within the dashboard can hardly be combined or manipulated and there is no way to apply machine learning algorithms.

In this paper I address how Google Analytics data can be downloaded into the statistical programming software R, where it can be freely manipulated and made accessible to machine learning algorithms. Using theory from Web Usage Mining, I demonstrate which shop-browsing variables can be used to derive meaningful behavioral attributes of website visitors. Based on these browsing variables, I demonstrate how actionable and targetable customer profiles can be built with the help of clustering algorithms, which can be used for high performing personalized advertising. In a case study, which incorporates data of a real Ecommerce business, I demonstrate how discussed methods can be applied. Taking everything into account, the following research shall be answered:

"How can targetable customer profiles be built based on Google Analytics data using machine learning clustering algorithms in R?"

Lastly, in an additional section, I discuss how a typical approach used in the industry to target cold audiences can be improved with the help of cluster results and how the resulting customer profiles of the case company can be applied in marketing practice. More specifically, I show how data can be used as a guide in the creation process of personalized ads and not merely as a performance assessment tool.

Literature addressing how data imported from Google Analytics can be utilized to build customer profiles with clustering algorithms is almost non-existent in the marketing literature landscape - a simple search in Google Scholar or at ResearchGate suffices to verify this rarity. In fact, querying "Google Analytics"+ "customer profiles"+ "machine learning" in Google Scholar delivers merely 71 search result. Furthermore, the resulting papers do not provide an explanation on how used Google Analytics data was downloaded and prepared for data analysis - the data is introduced as simply given. The described data query method in this thesis builds partly on the eBook "Using Google Analytics with R" by Michal Brys. In this book Brys gives an introduction into the connection between Google Analytics and R, explains the R package "googleAnalyticsR" that is used to query data and discusses broadly certain analytical application that can be performed. Nonetheless, the eBook does not delve deeply in into machine learning applications, introduced examples are very brief and not tied to a case study. In my thesis I try to overcome these shortcomings by providing a more machine learning focused and practically oriented research.

Google analytics is a highly cost-effective way for ecommerce companies to track user behavior without having an expensive custom tracking system in place, as it is for free. This is the reason why the vast majority of ecommerce businesses is using this platform. Veritably, "Google Analytics is being used by 52.9 percent of all websites on the internet, more than 10 times the next most popular analytics option, Yandex Metrics." **(W3Techs, 2015)** This high dependence of ecommerce businesses on Google Analytics and the earlier discussed lack of scientific literature, makes this research highly industry relevant.

My goal is to provide tech-oriented marketers with a methodology they can use to "free" themselves from the limiting interfaces of analytics tools and cultivate data-driven practices with Google Analytics that go beyond the dashboard.

2. Thesis Structure

The paper starts with two building blocks in order to provide the reader with 1) the right domain knowledge and 2) necessary technical methods, which are both crucial to understand the applications demonstrated in a case study in a later part of this paper

In the first building block I discuss marketing theory and analyze why specifically profiling and personalization offer tremendous advantages in the age of information overload. In fact, it is very well documented how personalization increases the effectivity of display ads by boosting clickthrough rates. Next, I discuss Web Usage Mining and introduce a broad range of scientific research that addresses different metrics to track and evaluate user behavior.

In the second building block, I elaborate the functions of Goggle Analytics and its integration with the statistical computing software R using the Google Application Programming Interface. Furthermore, I discuss the theory of customer segmentation and how machine learning can be used to create customer segment profiles based on web usage metrics. More specifically, I explain one of the most popular clustering methods, K-means, and an alternate version, Reduced K-means. Lastly, I describe diagnostics which can be used to asses cluster quality and stability.

The case study in the third part of the paper, presents how discussed theory and methods are applied on a real-life ecommerce business example. This business is a German wholefood store that specializes in the production of various chokeberry products like juices and jams. I demonstrate how clustering is performed and how results can be evaluated with the help of diagnostics.

In the fourth and last part, I answer the research question by demonstrating how cluster results can be used to build customer profiles. Based on these established profiles, in an additional section, I create sample campaigns to indicate how personalized advertising can look like to target col audiences with the help of Facebook's Lookalike targeting tool. Lastly, I talk about the limitations of the proposed profiling method, things marketers should be aware of and how further research can look like.

3. Marketing Theory Building Block

3.1. Importance of Personalization in the Era of Information Overload

They golden days of TV advertising and multipage magazine Ads, where long attention spans of customers were a given, have past long ago. In the era of non-stop information transactions, social media noise and unlimited video streaming, user attention prevails only a mere second before it transcends to the next exciting thing available.

These substantial changes over the years in attention span dynamics can be explained by the finding that visual attention of individuals decreases substantially, once they are exposed to multiple visual objects at the same time. This is mainly due to two reasons: 1) At any given time, only a limited spectrum of information can be recognized by the human eye and processed by the brain to act upon. In other words, paying attention to one given target object diminishes an individual's capacity to recognize and process others. 2) Visual objects compete for an individual's processing capacity – meaning, the human eye filters out weak contestants and emphasizes strong salient visual targets. The salience of objects might be due to something as trivial as a strong radiant color, intriguing text or sensational imagery or human faces. **(Desimone & Duncan, 1995)** Due to the information overload on internet platforms, an individual's eye to brain processing capacity is highly overburdened and he/she manages to recognize only a small fraction of the information available. Of this small percentage, individuals automatically filter those visual objects that are not salient enough. Personalization is a tool that aims to increase object salience, by making Ads personally relevant and tailored to a given individual. The effectiveness of personalized Ads in increasing visual salience has been studied by an abundance of researchers over the years and has been proven to be highly significant.

The effects of personalization of display advertising on attention spans of individuals has been examined with the help of eye-tracking data. It was found that the gazing time of individuals on personalized ads is significantly longer. In other words, it was concluded that *"personalized advertisements attract significantly longer and more attention than non-*

personalized ads, indicating the strong attention-grabbing effect of personalization.” (**Bang & Wojdynski, 2016)** Improved engagement levels induced by personalized content can also be verified by other metrics. For instance, ad clicking intentions of users on Facebook - merely a slight degree of personalization is enough to evoke a significant increase of ad clicking intention. **(De Keyzer, Dens, & De Pelsmacker, 2015)** This increase in clicking intention does not only hold true for social media ads but can also be observed for banner display ads within online stores. It was shown that click-through-rates of banner ads increase significantly even with mild degrees of personalization. **(Bleier & Eisenbeiss, 2015)**

Personalization of online advertisement can be implemented for instance on the basis of individuals’ demographic information. In recent years, researchers have found certain behavioral differences in internet surfing and online shopping behavior among individuals with varying demographic backgrounds. For instance, while men have the tendency to surf the web with more functional and entertainment intentions, women are more likely to surf the internet for shopping reasons. (**Wolin & Korgaonkar, 2005)** In terms of age, younger online shoppers tend to be indulged in exploring more products before they commit to a purchase than older generations. Meaning, younger shoppers have a tendency to be more exploratory than their older counterparts, while both generations consume similarly in terms of total purchases. **(Sorce, Perotti & Widrick, 2005)**

Although personalization based on demographics can be successful, it generalizes customers quite intensively. Meaning, just because a customer is a woman, does not mean that she cannot have a tendency to be highly utilitarian and functional in her browsing behavior. Likewise, just because a customer has been identified to be above 50, does not mean that he is not interested in exploring multiple products before a purchase like younger customers would do. Data tracking allows to document the entire shopping journey of each session of each individual customer. The next section will discuss how we can unveil behavioral tendencies, psychology and preferences of website visitors on the basis of real empirical insights rather than generalized assumptions. These data based behavioral tendencies shall later be used to build customer profiles, that yield more specific and reliable personalized advertising.

3.2. Web Usage Mining: Learning from Customer Behavior

Web usage mining is the process of identifying patterns and deriving useful information from browsing data about behavioral attributes of website visitors. **(Neelima & Rodda, 2015)** These patterns and behavioral information are hidden and are extracted on the basis of quantitative website usage metrics - this extraction process is also referred to as mining, hence web usage mining. **(Patel & Patel, 2012)**

Behavioral information gained about users can be used to asses not only psychological tendencies but also engagement levels and attention spans of customers. These insights can be essential in designing personalized advertising campaigns. For instance, customers who are identified to be heavily goal oriented, efficient in their browsing behavior and who extensively read product details, might have higher click-through-rates on ads that are concise, offer clear facts and rational buying arguments. On the other hand, exploratory visitors, who roam without intention but rather for fun through a multitude of product categories, who are more engaged with the website and have longer session durations might be better entertained with longer story-telling copies and emotionally appealing creatives.

Scientific literature has explored these web usage metrics on the basis of experiments with real customers of ecommerce stores. We can differentiate between several different usage measures which will be discussed next.

3.2.1. Hedonic and Utilitarian Usage Measures

Hedonic shoppers lay high emphasis on enjoying the purchasing experience and have a strong focus on exploration. Utilitarian shoppers on the other hand are heavily intent-driven individuals with the main concern of making their shopping experience as efficient as possible. **(Babin, Darden, & Griffin,1994)**

A hedonic user's browsing session is characterized by more visits to category-related pages than specific product-pages. This can be explained by the curious, exploratory motives of hedonic visitors, as they tend to search on a 'broader' level instead of rather specific product pages. The *"product-to-category ratio"* is a metric which can be used to measure this exploration affinity of a user – if the user has a hedonic/exploratory browsing tendency, we can expect the "*product-to-category ratio"* to be low as he/she visits more category pages than product pages. Shoppers with utilitarian tendencies on the other hand visit more product related pages and have thus a very high "*product-to-category ratio"*. They also tend to frequently come back to already visited product pages, measured by unique product revisits. Lastly, utilitarian users are more prone to use the search option in a shop, as they tend to have a clear purchasing goal when they visit a shop and like to find the product fast and efficiently. **(Moe, 2003)**

3.2.2. Engagement Measures

Engagement measures capture how involved, occupied and interested a consumer is in the visited website or shop. Highly engaged users tend to have longer sessions as they are more inclined to invest time in the exploration of a website. While "*session duratio*n" gives an absolute measure of engagement, "*page duration"* can be used as a measure for a visitor's attention span. **(Raphaeli, Goldstein & Fink, 2017)**

"Dwell time" is a more sophisticated measure of user attention spans than "*page duration"*, since it takes into account that users might open a page and leave the computer to engage in other activities. Merely using "*page duration"* as a metric to assess user attention spans could hence be misleading. This is why *"dwell time"* incorporates scrolling trigger events - meaning, the metric tracks only the time in-between the first scrolling activity on a given page and the last. **(Yi et al., 2014)**

"Average unique pages visited per session" is a further metric that assesses the complexity and depth of a user's session. More involved users will tend to have longer surfing paths and hence will visit more website pages. *"Visiting frequency"* on the other hand assesses a user's long-term engagement with a given website or shop. **(Huang, 2009)**

3.2.3. Content Interaction Measures

Content Interaction Measures assess the depth of interactions and engagement level users have with specific content targets on a website or shop. More specifically, *"page views"*, *"clicks"* and *"scroll depth"* in relation to items on a website are measured in order to reveal more about a user's behavioral tendencies. (**Mobasher et al., 2001)** For example, product pages of most online shops have three highly significant content targets, namely: Product images, product details/descriptions and reviews. A visitor's interaction depth with these content targets can give a marketer valuable insight into how a given user evaluates products. While the *"number of images viewed per product"* gives insights on a customer's visual evaluation tendency, the *"scrolling depth of product details"* might be an indication in how far a user is interested in utilitarian information to evaluate a product. Lastly, the *"scrolling depth of product reviews"* is a measure for a user's dependence on social proof and opinion of other customers to evaluate a product.

3.2.4. Shopping Cart Usage Measures

Online shoppers quite frequently add products to their electronic cart for other reasons than purchasing and abandon them at the end of their session. There are three main behavioral incentives for "*shopping cart abandonment*". 1) Adding items to the shopping cart without indulging in a purchase is for certain visitors entertainment and a boredom release. "These shoppers may get the thrill of enacting shopping rituals and satisfying impulses to shop without necessarily buying and spending money." (**Kukar-Kinney & Close, 2009)** 2) Visitors utilize the shopping cart as an organizational tool to create an inventory of things they are interested in. In fact, it was shown that shopping cart abandonment is directly and positively influenced by organization and research within the cart. **(Xu et al., 2015)** In other words, the shopping cart is used to track prices, order items and park products of interest in order to indulge in a purchase later down the road or in the next session. 3) Customers realize during their shopping process that the total price of the basket is too high, so they decide to shop at a later point in time when discounts are available and abandon their basket. In order to have a better understanding which of the three incentives apply to a particular customer, other metrics and information should be taken into account. For example, if the customer's purchase history reveals that he/she uses promotion coupons frequently and purchases many products on discount, there is an indication that high shopping cart abandonment might be due to cost sensitivity (3).

In addition to "*shopping cart abandonment*", the variety of products added to a cart could give indications about a customer's interest range. **(Grivia et al., 2018)** For example, marketers could track *"basket variety"*, which is an indication about the breadth of unique products that have been added to a shopping cart by a given customer. *"Basket variance"* on the other hand, measures if the breadth of products added to the cart tend to be the same or if they change consistently from session to session. High basket variety and high basket variance might be a strong indication for a customer's exploratory nature and a rather non-goal-oriented hedonic shopping attitude given his/her appetite for a broad range of products.

3.3. Theory Building Block Conclusion

In conclusions, personalization is a powerful tool which increases salience of advertising in an information overloaded web environment. This induced increase of user attention spans is shown not only by eye movement scans but also increased clickthrough rates of individuals on social media and banner ads. Although personalization based on demographics can be a successful strategy, it tends to generalize customers in terms of their traits and interests. In order to have a more empirical personalization methodology, tracked browsing data can be used to analyze and understand customer behavior. Web usage mining is the process of uncovering behavioral traits and browsing patterns of website visitors based on specific metrics. These metrics can be used to build customer profiles and personalize advertising accordingly.

In the following section the technical process of creating customer profiles using Google Analytics data in R shall be discussed.

4. Technical Building Block

4.1. Google Analytics

Website data tracking is essential for every E-commerce company that is serious about long term sustainable growth. Tracking software installed on a website is capable to basically record every move and click of a visitor. Insights gained through browsing data can be used for the improvement of a multitude of business assets and processes - for instance, website layout optimization (How can we transform our landing page to decrease bounce rates?), product recommendation (What did the customer put in his basket but did not buy?) and most importantly advertising personalization (What behavioral insights can help us to create ads that improve click-through rates?).

As mentioned earlier, Google analytics as a data tracking option is a very popular choice among ecommerce businesses. The system is easily installed on websites and online shops within a few clicks. The entire software is available without any charges, which is a tremendous opportunity for smaller online businesses, who do not have the financial and technical capacity to build their own custom data tracking software. Lastly and most importantly, the tracking

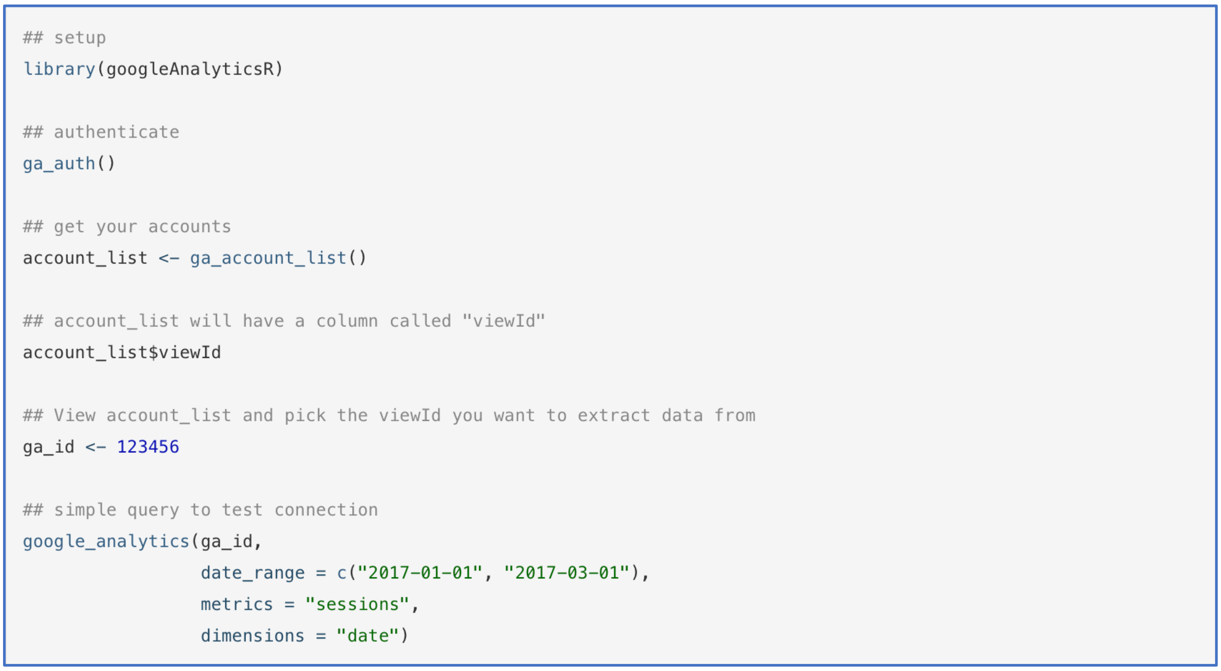
software comes with a dashboard tool that aggregates important performance measures and makes data easily interpretable even for non-technical employees. **(Plaza, 2011)**
Ironically, the tool that was intended to empower through simplification, limits users to unleash the full potential of the system. In fact, the dashboard's capabilities set the boundaries for data processing and analysis. Some advanced users might integrate further third-party add-ons to boost their dashboard's capabilities, but these add-on boosts have rather minimal impact and do not manage to overcome the main issues. Namely: A user is limited to the visualization and statistics offered within the dashboard. She cannot simply decide to calculate a statistic of interest on the basis of multiple tracked metrics for a subgroup of customers and visualize results in a diagram of choice. Also, advanced data wrangling and manipulation practices such as grouping observations on the basis of specific column entries, importing any type of data set and combining it with tracked data in the dashboard based on linking key variables or simply changing the unit of a variable is not possible within the tool. Lastly and most importantly, machine Learning algorithms cannot be coded and applied within the dashboard. Data analytical insights are limited to averages, percentages and rankings - predictions or clustering are out of reach.

4.2. R & The Google Analytics Reporting API

Aforementioned limitations cannot be solved within the dashboard's interface. The only way to overcome these problems is to abandon the dashboard all together and relocate data to a more flexible platform.

The statistical programming software R is one of the most popular tools used by statisticians and data scientists to explore data. Once Google Analytics data is imported into R, the software can be used to reshape the data in all possible formats and to add information from other data bases. R is capable to apply any kind of statistical analysis to the data including the application of cutting-edge machine learning algorithms.

The Google Analytics Reporting Application Programming Interface is a tool that can be used to directly access Google Analytics data and import it into other platforms. **(Weber, 2015)** There are at least six different packages for R that can be used to access the API and import data: `rga`, `RGA`, `RGoogleAnalytics`, `ganalytics`, `GAR` and `googleAnalyticsR`. They all offer roughly the same applications with mild differences in commands and code structure. For the sake of simplicity, only the code of the package “googleAnalyticsR”, shown in Figure 1, will be elaborated. The following code is used to tap into Google Analytics and import data directly into R.



```
## setup
library(googleAnalyticsR)

## authenticate
ga_auth()

## get your accounts
account_list <- ga_account_list()

## account_list will have a column called "viewId"
account_list$viewId

## View account_list and pick the viewId you want to extract data from
ga_id <- 123456

## simple query to test connection
google_analytics(ga_id,
                 date_range = c("2017-01-01", "2017-03-01"),
                 metrics = "sessions",
                 dimensions = "date")
```

Figure 1: Google Analytics Reporting API in R

Once the package has been loaded, the first step in the data downloading process is to authenticate R and allow the software to access the marketer’s Google Analytics user account. Once the authentication was successful, one can load a table into R with information about all business accounts the user has access to. The column “viewID” of the table provides IDs of each company account. Based on this ID the API can identify from which particular business account data shall be downloaded. The last and most important step is the data import itself. The marketer needs to provide the query code with the account viewID, a date range of interest and lastly with preferred “dimensions” and “metrics”. (Edmondson et al., 2019; Brys, 2017)

While *“dimensions”* refer to qualitative attributes like the date of a session, *“metrics”* refer to quantitative measures like the total number of sessions in Google Analytics. In order to get a data frame that shows dimensions and metrics of choice for each individual visitor, the dimension “ClientID” should be added to the query code. The “ClientID” dimension assigns every unique user an ID based on his/her cookies. Once the data is loaded in the desired format other important data bases can be added such as customer email lists from the CRM data base of a company. These email lists can be used to identify company customers and deanonymize Google Analytics browsing data. There is one dimension that can be used to connect client IDs to respective customer emails, the `transactionID`. The `transactionID` provided by Google Analytics is exactly the same as the one listed for each purchase in a CRM data base. Both data bases can be easily connected in R with a few lines of code.

4.3. Google Tag Manager: Tracking Anything

Google Analytics comes with valuable out of the box metrics like session browsing paths, visit durations and basket insights. Nonetheless, out of the box metrics do not track everything on a website. In fact, many metrics discussed in the first building block which are essential for powerful Web Usage Mining are missing. For instance, in case of content usage mining important content targets, such as product details, are specifically tracked with scrolling depths measures - Google Analytics does not come with such a metric out of the box.

Custom Metrics can be created and added with the help of the Google Tag Manager, which is a further marketing analytics tool that can be linked to Google Analytics. The Tag Manager's "trigger option" allows the user to create custom metrics which can track a visitor's interaction with basically anything on a website. The trigger type can be chosen from an abundance of options ranging from a simple click or scroll depth to more complicated types, such as the element visibility trigger, which fires when a selected element becomes visible in the web browser's viewport. **(Silverbauer, 2017)**

Measures added to Google Analytics with the help of the Tag Manager do not need to be limited to those metrics discussed in the theory building block. It is advisable that marketers carefully asses the website of their client, understand the layout infrastructure and identify valuable targets themselves from which they believe deep insights about users can be gained. Based on this assessment and industry knowledge, novel custom metrics are created that can be of high value.

4.4. Customer Segmentation: Identifying Similar Customers

4.4.1. Customer Segmentation Theory

Customers are heterogenous, not only in demographical but also in preferential terms with various behavioral tendencies and psychological attributes. Grouping customers with similar characteristics is essential for personalization applications. Customer segmentation is a method used to segregate a customer base into homogenous groups. Based on the features of these resulting clusters, customer profiles can be created, marketing operations can be tailored, products can be developed, and advertising can be personalized.

Customer segmentation resolves around two sets of variables, namely base and descriptor variables. Base variables are the main characteristics used to segregate customers into clusters. Base variables are hence the foundational parameters upon which heterogeneity is measured and similar customers are identified. Descriptor variables on the other hand are used to gain more informational insights into formed clusters. In other words, these variables describe the segments - hence the term "descriptor". **(Lilien, Rangasswamy & De Bruyn, 2017)**

The concept of customer segmentation is over half a century old **(Smith, 1956)**, nonetheless, the technical methods used to build customer segments are ever changing and improving. With the rise of big data, segmentation methods have become highly dependent on machine learning applications. **(Verdenhofs & Tambovceva, 2019)** The next section discusses clustering techniques that relie on the power of machine learning and several clustering quality diagnostics.

4.4.2. K-means Clustering

Machine learning algorithms can be divided into two families, namely supervised and unsupervised machine learning algorithms. If a researcher's goal is to predict a particular information for observations of a data set on the basis of other variables provided in that very data set, a supervised machine learning algorithm is needed. The researcher trains the algorithm with a labeled data set until it has learned to predict the information of interest in an unlabeled data set. **(Chourasiya & Jain 2019)** For example, if a marketer wants to predict if a customer is going to purchase a product in the third session based on browsing behavior of the first two sessions, she would specify the learning goal (predict purchase) and feed the algorithm with training data of various customers. This training data set would entail information about the first two sessions as well as the purchase information of the third session (labeled data set). Along the way she would observe and optimize the algorithm's prediction accuracy by tweaking algorithmic parameters until the performance is satisfactory. Then the marketer would feed the algorithm data of a new set of customers where only browsing behavior of the first two sessions is known (unlabeled data set) in order to predict if they will indulge in a purchase in their third session. In case of customer segmentation, the goal is not to predict any particular information. Instead, the marketer wants to divide observation into homogenous subgroups on the basis of variables and measures provided in the data set. For these kind of research goals unsupervised machine learning algorithms are ideal as their capability is to find underlying patterns in data without being told what these particular patterns supposed to look like. In segmentation these patterns are clusters of similar observations. One of the most widely used unsupervised clustering algorithms in machine learning is referred to as K-means clustering.

The fundamental concept behind K-means clustering can be explained as follows. Given a data set consists of n observations, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, where each observation is a vector of d dimensions, K-means clustering partitions the observations into $k \leq n$ clusters $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$. The goal is to cluster the data set in such a way that the distance of observations to their cluster center, measured as within-cluster sum of squares (WCSS), is minimized. Hence the aim is $\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$, where $\boldsymbol{\mu}_i$ is the mean of points in S_i. This minimization results in maximizing the distance between cluster centers, which is considered a desired solution for clustering observations. **(Forgy, 1965)**

The K-means algorithm operates on the basis of the above concept in a sense that cluster center positions are iteratively optimized until the minimization of WCSS is achieved. Based on a chosen value for k, the algorithm creates centroid vectors, $\mathbf{c}_1, \mathbf{c}_2, \dots \mathbf{c}_k$ of d dimensions. The algorithm first places the centroids at random locations into a d dimensional data cloud consisting of n observations, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$,where each observation is a vector of d dimensions. Then it repeats the following steps until convergence: For each observation, denoted by the vector $\mathbf{x}_i$ where $i = 1,2, \dots, n$, the algorithm finds the closest centroid $\mathbf{c}_j$, where $j = 1, 2, \dots, k$ by calculating the Euclidean distance, $\underset{j}{\arg\min} \|\mathbf{x}_i - \mathbf{c}_j\|$. Once distances are calculated, each point $\boldsymbol{x}_i$ is assigned to one of the j clusters. Then, for each S_j a new centroid position is calculated. This new centroid vector, $\tilde{\boldsymbol{c}}_j$, is determined by calculating the arithmetic mean of all the points, S_j, that were assigned to the cluster: $\tilde{\boldsymbol{c}}_j = \frac{1}{|S_j|}\sum_{\mathbf{x}_i \epsilon S_j} \mathbf{x}_i$. Once the new centroid positions are created, the algorithm repeats itself. Naturally some points that were earlier assigned to one cluster, might be now closer to the new position of a centroid of another cluster, hence these points change cluster membership. Convergence is achieved once no point changes cluster membership. **(Lloyd, 1982)**

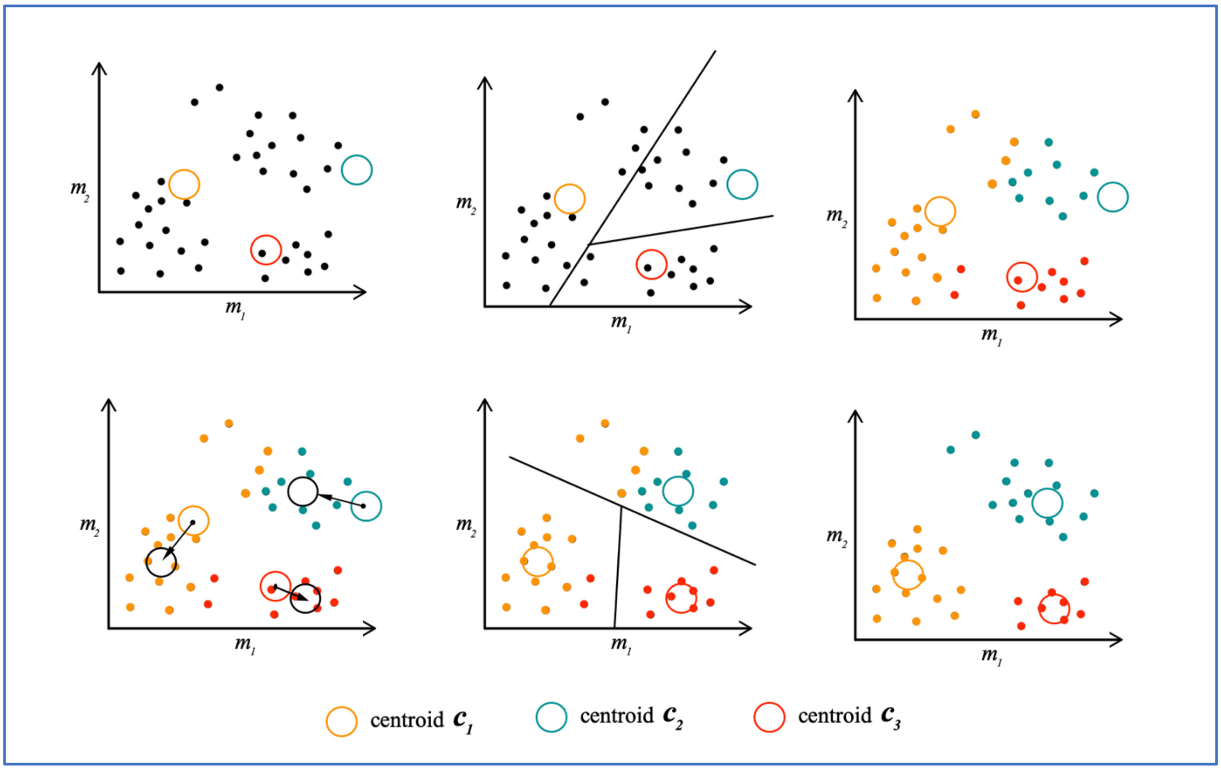


Figure 2: The K-means Algorithm

Figure 2 depicts the repetition phases of the algorithm and convergence for observations in two dimensions. Since K-means measures Euclidean distances, it is advised to use only numerical data as input. Furthermore, variables that shall be used to cluster should be on commensurable scales.

4.4.3. Reduced K-means Clustering

Not all variables in a data set are equally valuable to cluster observations. In fact, in some cases certain variables add noisy data dimensions, which could distort the quality of cluster results. One way to account for this, is to reduce the dimensionality of a data set first before applying any cluster algorithms. There are several approaches to reduce the dimensionality of a data set, one very common approach is referred to as Principal Component Analysis (PCA). This particular approach aims to determine a set of new dimensions, one smaller than the number of original variables, that retain as much variance of the data set as possible. **(Pearson, 1901)** These new dimensions are also referred to as principal components and each can be expressed by a loadings vector. Each element of a loadings vector corresponds to one of the original variables and can be interpreted as a weight whose magnitude indicates how much variance of a particular variable is expressed by the given principal component. While the first principal component accounts for the most variance in a data set, the second principal component accounts for the second, and the third principal component for the third most variance in the data set (and so on). If for instance 90% of the variance in a data set with six variables can be retained by the first three principal components, the remaining two principal components can be disregarded. In that case, the data set is transformed to a lower dimensionality, from six to three dimensions, without losing much information - only 10% of its variance to be exact.

The practice of applying dimension reduction first and cluster analysis after is known as the "tandem" approach. **(Arabie & Hubert, 1994)** The problem with this approach is that resulting cluster solutions might not be ideal as the methods, dimension reduction and clustering, optimize for different goals. While dimension reduction methods aim to find new fewer dimensions that retain as much variance of the original data set as possible, clustering aims to find observations in a data set that are similar/ dissimilar to each other. **(De Soete & Carroll, 1994)** Reduced K-means is a joint method, that combines both Principal Component Analysis and K-means clustering in one optimization function.

Similar to K-means, Reduced K-means aims to minimize the within-cluster sum of squares between observations and cluster centroids. The fundamental difference is, that cluster allocation and dimension reduction is performed simultaneously in such a way that the WCSS is minimized based on the distance between the original observations and the projected centroids in reduced dimension. **(Markos, Iodice D'Enza, Van de Velden, 2019)** Let $\mathbf{X}$ be a data matrix of dimensions $n \times Q$, n denoting number of observations and Q denoting the number of variables. Then the minimizing loss function for Reduced K-means is $\min \phi_{\mathrm{RKM}}(\mathbf{B}, \mathbf{Z}_K, \mathbf{G}) = \|\mathbf{X} - \mathbf{Z}_K \mathbf{G} \mathbf{B}^{\mathrm{T}}\|_F^2$, where $\mathbf{Z}_K$ is a $n \times K$ binary matrix indicating for each observation membership for each cluster K; $\mathbf{G}$ is a $K \times d$ matrix that holds cluster centroids for each principal component d and $\mathbf{B}$ is a $Q \times d$ matrix which holds the loadings for each principal component. (De Soete & Carroll, 1994; Bock, 1987)

4.4.4. The Silhouette Score: Quality Assessment and Choosing K

The quality of clusters can be determined on the basis of several diagnostics tools. One of these diagnostic tools that can be used to validate the quality of clusters by assessing how well datapoints were assigned to the clusters is called the Silhouette Coefficient. This coefficient is based on two measures, namely cohesion and separation. Cohesion is determined by measuring the Euclidean distances between each point and their corresponding cluster members and calculating the average. The final cohesion score is denoted as a_i where i is the cluster of interest. Separation on the other hand measures the Euclidean distance between each point of one cluster to the points of the closest cluster neighbor. Again, once all distances are measured an average is determined – the final separation score is denoted as b_i. The silhouette coefficient s_i for cluster i is nothing more than the normalized difference between the cluster's separation and cohesion score, $s_i = \frac{a_i - b_i}{\max(a_i, b_i)}$. (Hadi, Kaufmn & Rousseeuw, 1990) The resulting silhouette coefficient ranges from -1 to $+1$, where a value close to -1 denotes that data points have been assigned to the wrong cluster. A value around 0 means that the data points are very close to the neighboring cluster and a value close to $+1$ indicates that the data points have been assigned with a high certainty to the right cluster as they are far away from the neighboring cluster. In order to determine a value for K, silhouette coefficients for a range of K clusters can be calculated. Then, the K that yields the highest silhouette coefficient is chosen.

4.4.5. Cluster Stability

Clustering algorithms come with a caveat marketer should be aware of - namely, that cluster results might not depict the "real" underlying structure of a data set. Cluster algorithms are exploratory in nature and find clusters even if there is no meaningful way to segment a data set in reality. For instance, if a data set is highly homogenous it might not need further clustering. In such a case, one can expect that cluster results will vary extensively if the algorithm is applied to a new sample set drawn from the same population. In terms of customer profiling, this is critical since if marketers are not careful enough, they might target a customer segment that does not really exist wasting advertising budget on an artificially created cluster construct.

A quality assessment that evaluates how well resulting clusters are separated from each other, like the silhouette score, fails to give us an indication about this issue. What needs to be assessed is if detected data structures actually re-appear or remain "stable" when the algorithm is applied to a new data sample drawn from the same population of interest. One way to measure cluster stability is to use the so called Jaccard Index. **(Christian Henning, 2007)** The Jaccard Index is a statistic used to assess the similarity between two data sample sets. Given two sample sets A and B, the jaccard index $\gamma(A, B)$ is calculated by dividing the size of the intersection of both data sets by the size of the union of both data sets. **(Jaccard, 1901)**

Jaccard Index:

$$\gamma(A, B) = \frac{|A \cap B|}{|A \cup B|}, \text{ where } 0 \leq \gamma(A, B) \geq 1$$

The fpc package incorporates this concept in an algorithm called clusterboot. (Christian Henning, 2007) The method is based on the idea to generate new data sets from the original data set using different resampling techniques. Once new samples have been created, the clustering method of interest is applied to each new sample. Since a certain amount of variation is introduced by the resampling method, the originally detected cluster structure is put to the test. In other words, it is tested if a similar structure can be identified in the new sample set despite the introduced variation. This similarity assessment between the original data set clustering results and the resampled data set clustering results is done using the Jaccard Index.

One option is to resample data by generating B bootstrapped samples from the original data set $\boldsymbol{x}_n$. The algorithm repeats the following scheme for $i = 1, \dots, B$:

1) The bootstrap sample B_i with n points is drawn with replacement from the original data set and is referred to as $\boldsymbol{x}_n^i$.

2) Using the bootstrapped sample, a clustering is computed, $E_n(\boldsymbol{x}_n^i)$, where E_n is a sequence of mappings that indicates for each point of $\boldsymbol{x}_n^i$ its corresponding cluster membership.
3) The data points that are in the original data set as well as the bootstrapped sample are $\boldsymbol{x}_*^i = \boldsymbol{x}_n \cap \boldsymbol{x}_n^i$. The algorithm determines the intersection between the points mapped to cluster C based on the original clustering $E_n(\boldsymbol{x})$ and the points that are in the original as well as in the bootstrapped data set, namely $C_*^i = C \cap \boldsymbol{x}_*^i$. Then it determines the intersection between the points of each cluster of the new clustering $E_n(\boldsymbol{x}_n^i)$ and the points that are in the original as well as in the bootstrapped data set, namely $\Delta = E_n(\boldsymbol{x}_n^i) \cap \boldsymbol{x}_*^i$.
4) The maximum Jaccard Index $\gamma(C_*^i, D)$is computed, where D is an element of Δ and corresponds to the cluster of the new clustering that maximizes the similarity measure.

The algorithm results in a sequence of i Jaccard Indices for cluster C. The mean of this sequence is calculated, namely $\bar{\gamma}_C = \frac{1}{B^*}\sum_{i=1}^{B}\gamma_{C,i}$, where $\bar{\gamma}_C \geq 0.75$ is interpreted as an indication for a stable cluster. **(Christian Henning, 2007)**

A further resampling technique offered by the algorithm clusterboot can be utilized to put cluster stability to a more rigorous test. Instead of creating a bootstrapped sample, a new sample is created by the algorithm drawing m points from the original data set and replacing them by random values from a noise distribution. A recommended value for m is 0.2, meaning 20% of the original observations are replaced by the algorithm with random noise. **(Christian Henning, 2007)** This new data set is essentially a contaminated version of the original data set. Now it can be tested if original data structures can be recovered despite introduced noise. The scheme for this variant of the algorithm is exactly the same as described above, only that $\boldsymbol{x}_n^i$ is the newly created contaminated data set and not a bootstrapped sample. Now B refers to the number of contaminated versions of the original data set that have been created.

4.5. Technical Building Block Conclusion

Customer segmentation is a well-established method in marketing theory used to group homogenous costumers. This grouping is performed using base variables, which are the main characteristics used to assess customer homogeneity. Descriptor variables on the other hand give more information about each built cluster. Unsupervised machine learning can be used to segment customers – one very popular algorithm is K-means clustering. Not all base variables are equally valuable - in fact, some variables might add unnecessary noise distorting cluster results. Reduced K-means Clustering is an alternative algorithm that addresses this issue of excess variables by clustering observations in reduced dimensionality. Lastly, while the silhouette score can be used to assess how many clusters should be build, the Jaccard Index can be used to assess stability of each cluster.

In the next section I use knowledge from both building blocks and demonstrate how proposed methodologies can be applied in a real case example.

5. Case Study: Putting Theory to Practice

5.1. Case Study: Business Profile

The family business "Aronia vom Langlebenhof" is a middle-sized company based in Passau (Germany), which specializes in the cultivation, harvesting and production of chokeberries. Their products are diversified in ten different categories, namely: regular juice (their flagship product category), herbal juice, tea, dried chokeberries, vinegar, chokeberry powder, lemonade, cookies and a company magazine. Next to these product categories they also offer mixed product bundles, which include a combination of items from a broad variety of product categories. All their products are certified with Germany's oldest and most renowned organic seal, the Demeter seal. Furthermore, a certain percentage of their profits is directly donated to the Langlebenhof which is a care home for physically and mentally disabled individuals, which is located directly next to the chokeberry fields of the company.

The business used to be exclusively dependent on traditional retail channels until they launched their ecommerce store in 2019. Since then they successfully scaled their online business to such an extent that the majority share of their profits is generated online. Their business attracts up 4,700 users every week to the online shop resulting in roughly 7,200 Euros in revenue (per week).

The shop has the out of the box Google Analytics version installed - unfortunately, custom tracking events like product detail scroll depth have not been installed using the Google Tag manager. On the one hand, this reduces the amount of insights that can be gained applying Web Usage Mining. On the other hand, using a standard Google Analytics set up as case basis can be quite valuable for Ecommerce businesses who have never used the Google Tag manager to ad custom metrics to their tracking system.

5.2. Case Study: Data

5.2.1. Data Description

After querying and downloading data using the Google Analytics Reporting API, several data sets have been created. Each of these data set have been linked to contact information of clients on the basis of the transactionID variable as explained in the methods section. The first data set, “Base Variable Data Set”, holds all the browsing variables that will used in the Reduced K-means algorithm. The data set consists of 10 variables and 4836 rows - each row corresponds to a single client. The values captured by each variable are averages per session or per transaction for the time frame of an entire year (start date: 2019-06-07; end date: 2020-06-07, 366 days). The below table provides descriptions for each of the 10 variables of the mentioned data frame.

Base Variable Data Set	
ClientName	Used as an aggregation variable, will be excluded from the cluster algorithm.
Total.sessions	Total number of sessions a given client had within the set time frame (used to filter clients with fewer than three sessions, **not used as cluster variable**)
avgSessionDuration	Average length of a user’s session in seconds
pageviewsPerSession	Average number of pages viewed per session, including repeated views of a single page
Avg.ProdViewVariety	Average number of unique product categories viewed per session. Example: If a client viewed a “Nike Basketball”, a “Adidas Basketball” and “Jordan Sneakers” he/she viewed 3 unique product categories in that session.
Avg.QuantityAddedToCart	Average number of product units client adds to his/her cart per session
Avg.PurchaseValue	Average purchase value (in Euro) per transaction.
Avg.PurchaseQuantity	Average quantity of product units purchased per transaction.

Avg.ProductPurchaseVariety	Average number of unique product categories purchased per transaction.
ConversionProbability	Total Number of Transactions / Total Number of Sessions

Table 1 – Base Variables for Clustering

The following data sets all hold variables that shall be used to interpret cluster results. These data sets have in common that their variables are aggregated means per cluster. Since they are cluster aggregated, the data sets are created after final cluster results have been obtained.

Descriptor 1: Browsing Behavior	
Cluster	Integer that labels a given cluster
sessions	Average number of sessions members of a given cluster had within the set time frame.
Avg.SearchToolUsage	Average number of times per session members of a given cluster used the search box on the website.
Avg.CartAbandonmentRate	Average number of product units members of a given cluster added to their basket per session but did not buy.
Avg.CouponUsage	Average number of coupons members of a given cluster used per transaction.
Avg.quantityAddedInCheckout	Average number of product units members of a given cluster added to their cart in the checkout section per session.

Table 2.1: Variables for Cluster Description

Descriptor 2: Product Category Preferences	
Cluster	Integer that labels a given cluster
ProductCategory	Name of the product category
Avg.UnitsPurchasedPerCategory	Average number of units members of a given cluster have purchased of a particular product category

Table 2.2: Variables for Cluster Description

Descriptor 3: Gender Distribution

Cluster	Integer that labels a given cluster
Male	Percentage of males for a given cluster
Female	Percentage of females for a given cluster

Table 2.3: Variables for Cluster Description

Descriptor 4: Device Usage	
Cluster	Integer that labels a given cluster
Desktop	Average number of times members of a given cluster visited the shop from their desktop
Mobile	Average number of times members of a given cluster visited the shop from their phone
Tablet	Average number of times members of a given cluster visited the shop from their tablet

Table 2.4: Variables for Cluster Description

Descriptor 5: Top 5 Visiting Sources	
Cluster	Integer that labels a given cluster
direct	Average number of times members of a given cluster visited the shop by directly typing in the shop URL in the browser search field
facebook.paid	Average number of times members of a given cluster visited the shop after clicking on a Facebook/Instagram ad
organic.search	Average number of times members of a given cluster visited the shop after searching a shop/product related term in Google (or other search machines)
referrel	Average number of times members of a given cluster visited the shop after clicking on a link on any other website.
google.search.ad	Average number of times members of a given cluster visited the shop after searching a shop/product related term in Google and clicked on a search ad of the company

Table 2.5: Variables for Cluster Description

Descriptor 6: Time Spent Non-Product Pages	
Cluster	Integer that labels a given cluster
About Us	Average duration (in seconds) members of a cluster spent on the "About Us" page. This page holds information about the company and its staff
Blog	Average duration (in seconds) members of a cluster spent on the company's blog
Juice Info	Average duration (in seconds) members of a cluster spent on a page that provides more details on benefits of the company's flagship product – the chokeberry juice
FAQ	Average duration (in seconds) members of a cluster spent on a page that provides answers on questions relating chokeberry consumption and health benefits
Organic Certificate Info	Average duration (in seconds) members of a cluster spent on a page that provides information about the company's green farming practices and their handcrafted production process

Table 2.6: Variables for Cluster Description

Descriptor 7: Topics of Search Queries on Google	
Cluster	Integer that labels a given cluster
Company.Topic	Average number of times members of a cluster had a Google query with a search term relating directly to the company, such as "Langlebenhof" or "Passau" (location of the company).
Halth.Topic	Average number of times members of a cluster had a Google query with a search term relating directly to health topics, such as "immune system" or "cardiovascular system".
Products.Topic	Average number of times members of a cluster had a Google query with a search term relating directly to the different products of the shop, such as "powder" or "vinegar".

Broad.Topic	Average number of times members of a cluster had a Google query with broad search terms, such as “chokeberry” or “Aronia” (German for chokeberry).

Table 2.7: Variables for Cluster Description

5.2.2. Data Exploration and Pre-Processing

Since the goal is to create profiles that capture general behavior of clients, a certain amount of sessions has to be captured per client in order to derive generalizable assumptions about a client’s surfing behavior. In other words, one session alone might not be sufficient enough to create a reliable behavioral profile of an individual. A cutoff value of 3 has been used as a minimum number of sessions a client has to have in order to be “eligible” for profiling - meaning, clients who had only 1 or 2 sessions were excluded from the “Base Variable Data Set”, reducing its dimensionality to 2083 observations.

As discussed in the technical building block, K-means’ performance hinges on the fact that variables are on commensurable scales. In order to make sure that this is the case, I standardized the data. Additionally, in order to make data more symmetrical a logarithmic transformation will be applied before variables are set to commensurable scales. A logarithmic transformation is effective in eliminating the skewness of a data distribution. First, a 1 is added to each numeric value of the data set to eliminate zeros - this is necessary to take logarithms. Then, once the logarithmic transformation has been performed the data is scaled and centered.

5.3. Case Study: Clustering

5.3.1. Reduced K-means Results

The clustrd package **(Markos et al., 2019)** is used to perform Reduced K-means Clustering. The function tuneclus provides the average silhouette width **(Rousseeuw, 1987)** for a range of k values and d dimensions. Once the algorithm has run through all combinations of k and d, it delivers the “best” result based on the highest average silhouette width. The best clustering results is achieved for 3 clusters in 2 dimensions.

Results for 4 Clusters in 2 dimensions

average silhouette score	0.213
Cluster 1 (silhouette score)	0.260
Cluster 2 (silhouette score)	0.220
Cluster 3 (silhouette score)	0.130

Table 3.1: Custer Results

Figure 3 provides a visual representation of the three clusters in the first two dimensions and shows how clusters differentiate along the eight clustering variables (depicted by the colored arrows).

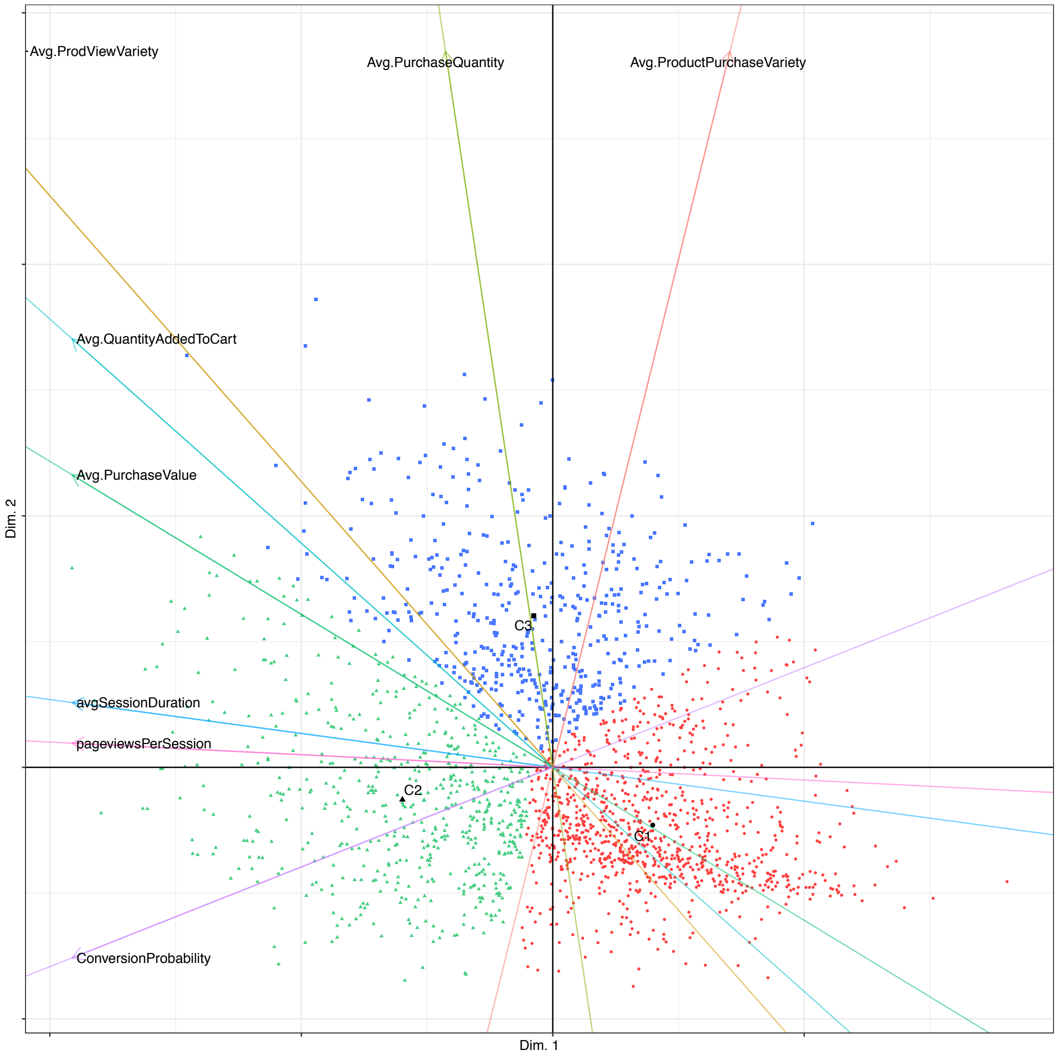


Figure 3: Biplot of Reduced K-means Cluster Results in 2 Dimensions
Cluster 1 is marked red, cluster 2 is marked green and cluster 3 is marked blue

In the next section it will be assessed if the resulting three clusters represent true customer segments which can actually be targeted in the real world. The section 5.4.1. Cluster Results Dashboard provides a broad range of diagrams and graphs that can be used to gain further insight into each cluster’s characteristics. Once cluster stability has been assed, the dashboard shall be used to describe in full detail each customer segment.

5.3.2. Cluster Stability Results

For $k = 3$ and $k = 4$ Reduced K-means Clustering results in $d = 2$ dimensions, cluster stability measures have been calculated as can be seen in Table 3. In the left column one can see the values for mean Jaccard Indices for each cluster based on $B = 100$ bootstrapped samples. The right column contains mean Jaccard Indices for each cluster based on $B = 100$ data sets where $m = 0.2n$ points were substituted by noise - meaning for each of the 100 samples, 20% of the n points of the original data set were replaced by points from a noise distribution.

Cluster	Bootstrap		Noise (20%)	
1	0.97	0.85	0.90	0.81
2	0.95	0.82	0.86	0.76
3	0.93	0.73	0.82	0.63
4	____________	0.65	____________	0.44

Table 3: Cluster Stability Results

Considering the results in Table 3, Jaccard Indices of both resampling methods for $k = 3$ are above the quality cut off value of 0.75. Meaning, both stability assessment methods suggest that the three cluster structures can be recovered even with resampled data sets that are disturbed by noise. This in turn is strong evidence that resulting clusters represent real customer segments and that targeting is feasible. In order to assess if this stability is also given if more clusters are formed, I calculated Jaccard Indices for $k = 4$. It is clearly evident that the fourth cluster is highly unstable and risky to target, hence, I will stick to $k = 3$.

5.4. Case Study: Results

In this part, I interpret cluster results from the case study to demonstrate how customer profiles can be finalized. For the case there are three customer profiles, each representing unique behavioral traits and buying habits. Due to the fact, that user IDs have been linked to their corresponding Email addresses, each customer profile can be targeted. The following "Cluster Results Dashboard" is used as basis for the interpretation of each cluster.

5.4.1. Cluster Results Dashboard

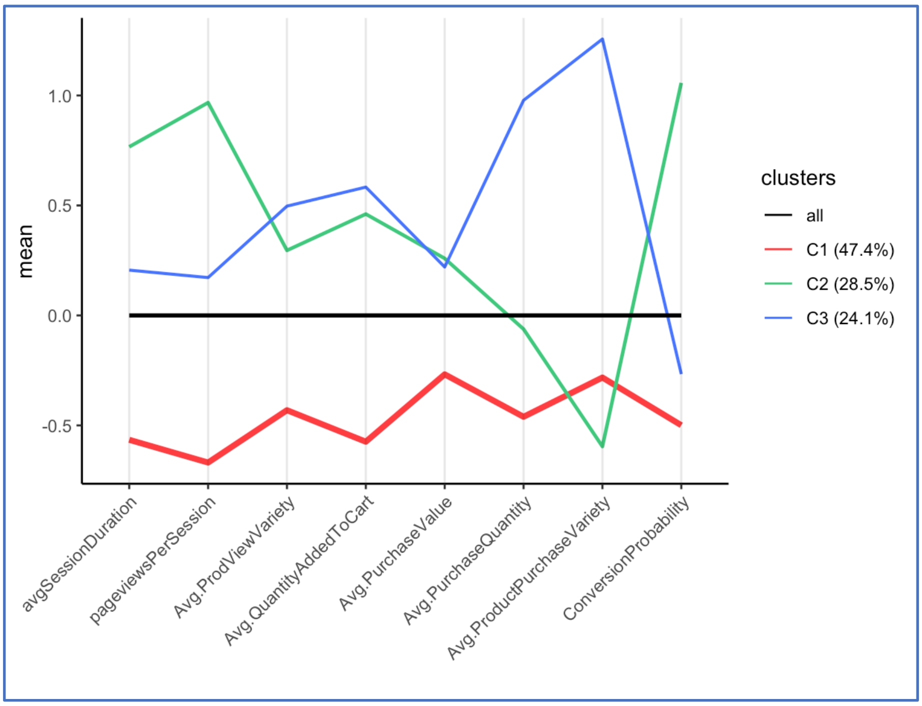


Figure 4: Base Variables

This plot provides insights into how strongly a cluster differentiates along the 8 clustering variables in contrast to other clusters. Means are scaled and centered for better comparison between clusters

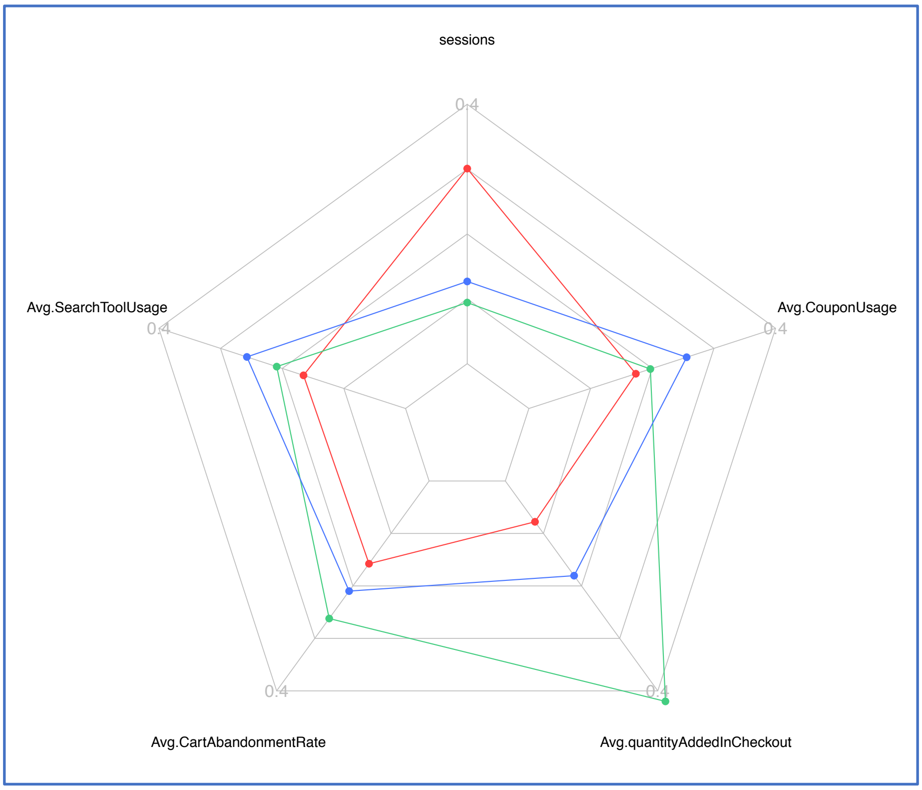


Figure 5: Browsing Behavior Descriptors

This radar chart includes the 5 variables, which were excluded from the cluster algorithm given their skewed distribution. For each variable the means have been calculated per cluster. Given the different scales of each variable, the means have been centered and scaled.

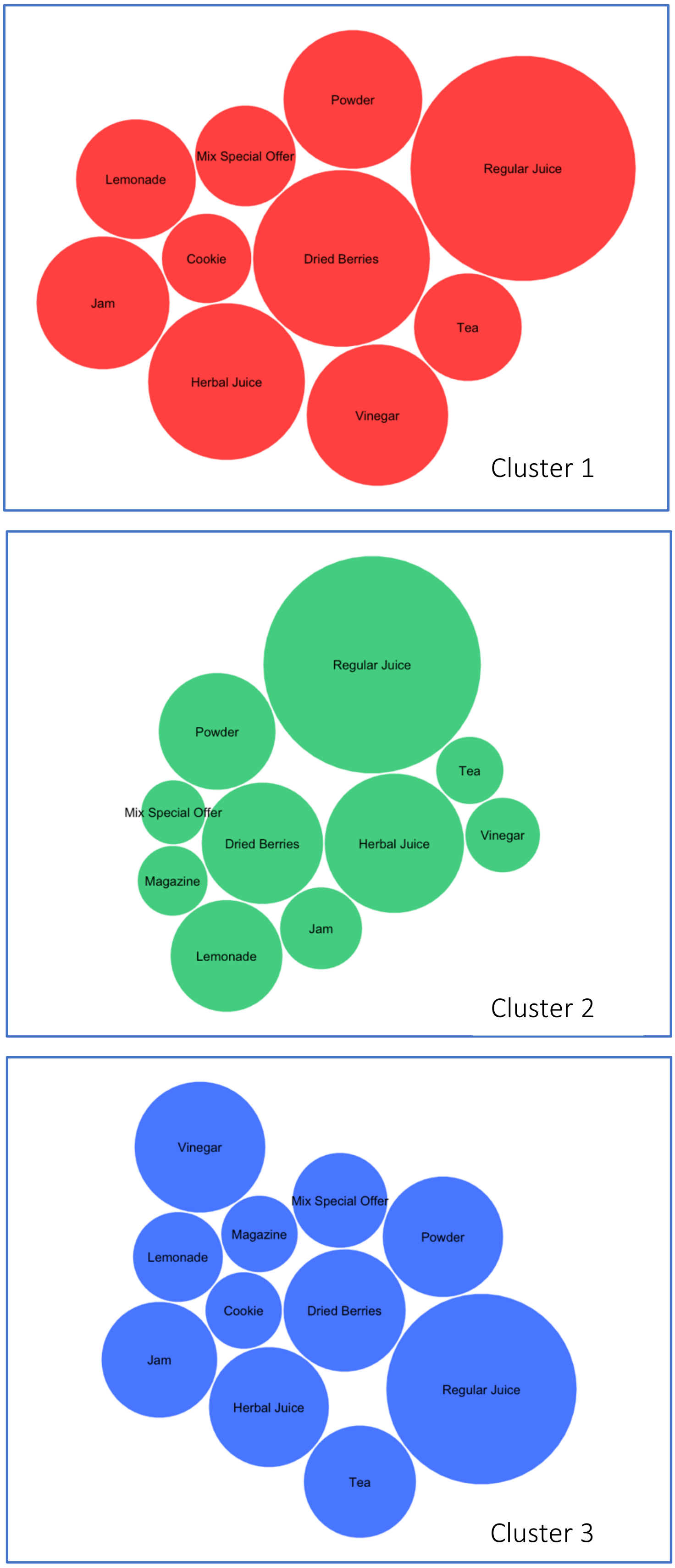
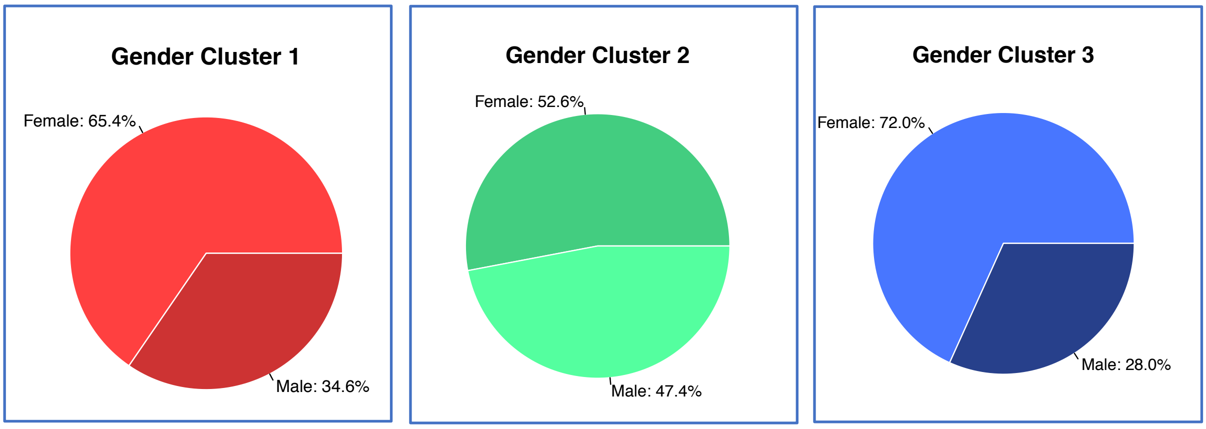


Figure 6: Product Category Bubble Chart - Average Units Purchased

For each product category the mean units purchased has been calculated per cluster. The bigger the size of a bubble, the more popular a product category is for a given cluster.

Note: *The relative positions of the bubbles have no meaning - the arrangement is random.*

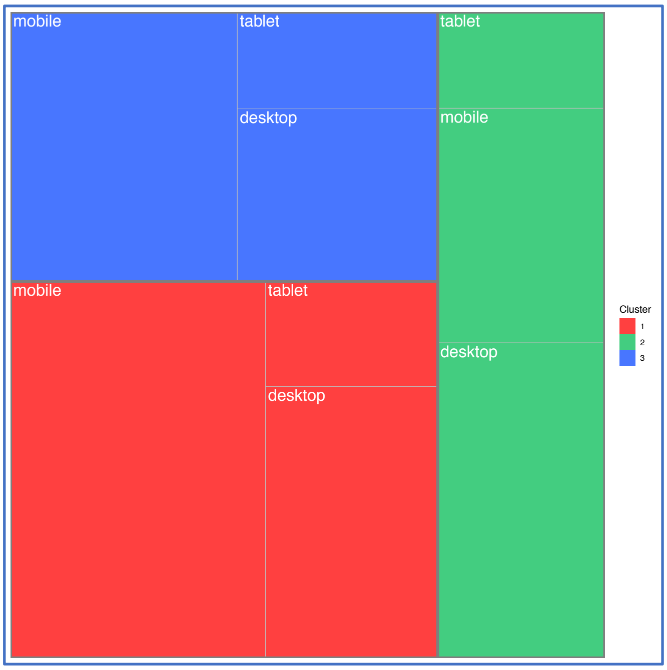


Fgure 7: Gender Distribution per Cluster

Figure 8: Device Usage per Cluster

For each cluster, the average number of sessions per device category have been calculated. The bigger a partition, the more dominantly used is a given device.

.



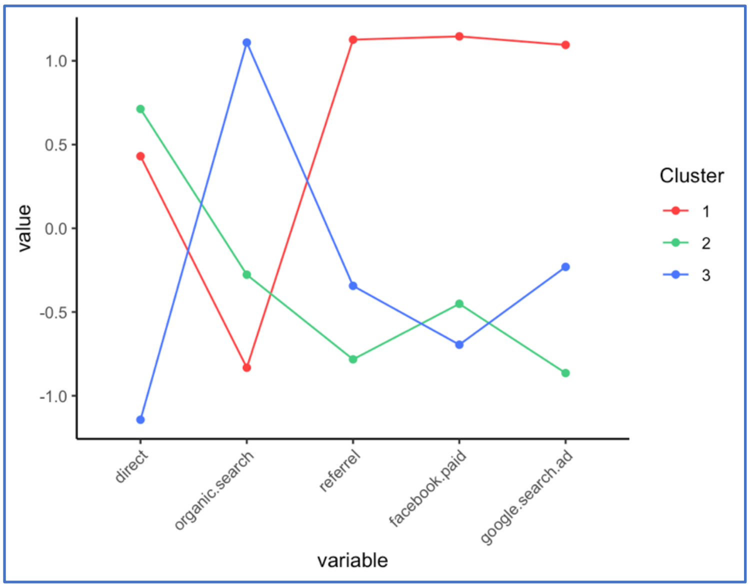


Figure 9: Average number of times members of a given cluster came from a particular source

The y-axis of the graph represents the means. Variable means for each cluster (denoted by dots) were scaled and centered for better comparison between clusters.

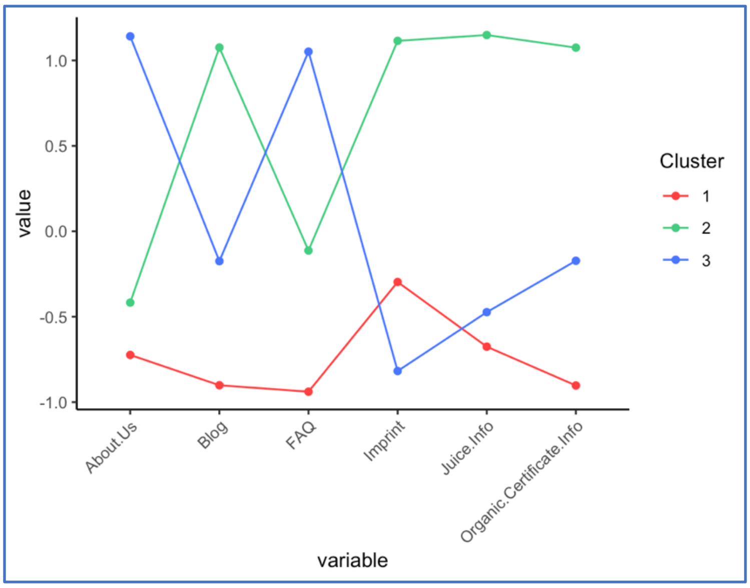


Figure 10: Average Time spent on pages other than product and checkout pages per cluster

The y-axis of the graph represents the means. Variable means for each cluster (denoted by dots) were scaled and centered for better comparison between clusters.

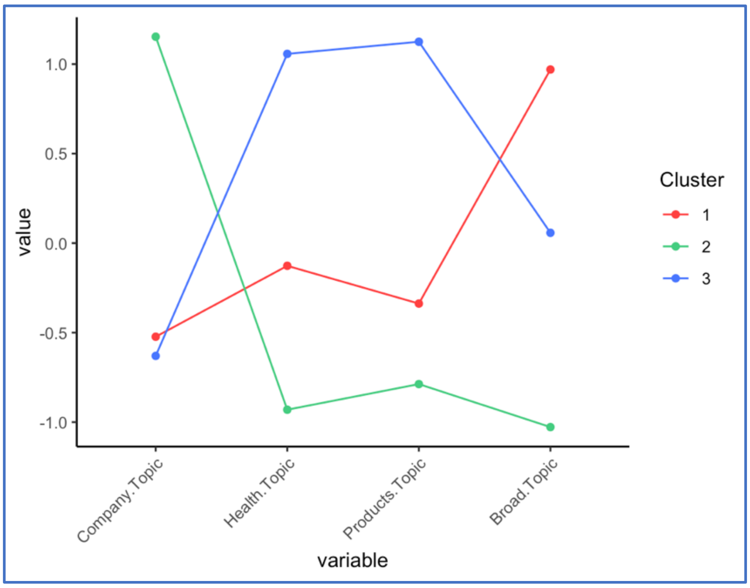


Figure 11: Average number of times a Google query related to one of the four topics per cluster

The y-axis of the graph represents the means. Variable means for each cluster (denoted by dots) were scaled and centered for better comparison between clusters.

5.4.2. Customer Profile 1: Disengaged Frugals (Cluster 1)

This segment is the largest as it constitutes with 987 members almost half (47.4%) of the clustered customers. Members of this segment are clients who show extremely low engagement levels when they visit the shop as indicated by low *avgSessionDuration* and low *pageviewsPerSession* in Figure 4. Their tendency to explore a broad range of product categories (*Avg.ProdViewVariety*) is lowest among all other segments. The probability that a member of this segment actually buys anything in a session is rather unlikely. If they indulge in a purchase, the average transaction value is the lowest among all three clusters, as can be also seen in Figure 4. Surprisingly, Disengaged Frugals visit the shop clearly more frequently than other segments (Figure 5). In light of the extremely low conversion probability, one can make the assumption that these customers are more resistant or even cautious as they need more sessions and thus more persuasion until they purchase compared to the other segments. This cautiousness might also be reflected by the fact that they rarely add products impulsively in the checkout section of the shop, reflected by the lowest means of *QuantityAddedInCheckout* as can be seen in Figure 9. In terms of promotion, their tendency to use coupons is lowest among all other clusters (Figure 5) – this might be due to the fact that coupon codes are only offered by the shop after a certain browsing time and if users put items into their shopping cart. As both surfing time as well as units added to cart are utterly low for Disengaged Frugals, it might be reasonable that this segment has less access to coupons.

Their product-purchase-variety is neither particularly broad nor heavily specialized - this can be seen easily in the bubble chart of Figure 6. The bubbles illustrated for cluster 1 are neither as homogenous as in Cluster 3 nor highly skewed as in Cluster 2. Their top three most frequently purchased product categories are regular juice, dried berries and herbal juice. The majority, roughly 65.4% of customers of this segment are female (Figure 7). In comparison to the other two segments they barely use the desktop to shop but have the strongest mean usage of mobile devices (Figure 8).

Although members of this segment also frequently visit the shop by directly typing in the shop URL into the browser search box, they more frequently than member of other segments visit the shop by means of a marketing stimulus. This is clearly depicted in Figure 9, where it can be seen that members of this segment visit the shop, more than any other segment, by either clicking on Facebook or Instagram ads, querying a term and clicking on a

Google ad, or by browsing through another third party website and clicking on a referral link after reading a blog post. Figure 10 emphasizes once more how disengaged this customer segment as their average time on non-product pages are extremely low in comparison to the other two segments - the only page that has somewhat moderate average visiting times is the imprint page. The fact the most intensively visited non-product page is the imprint, might be a further indication of the segment's cautiousness. One can speculatively interpret this behavior as a "safety check" in which a user reassures herself of the legitimacy of the company before entering into any type of bargain. Lastly, Figure 11 shows that queries of Disengaged Frugals that lead to a Google Ad click, are relatively broad. In other words, their queries barely include the company name, products associated with the company or health topics related to chokeberry consumption. Used query terms are rather unspecific like "chokeberry", "Aronia" (German for chokeberry), or "chokeberry juice".

5.4.3. Customer Profile 2: Loyal Specialists (Cluster 2)

This segment is the second largest with 593 members and constitutes 28.5% of the clustered clients. Customers of this segment are the most engaged based on the highest mean values for *avgSessionDuration* and *pageviewsPerSession* as can be seen in Figure 4. Their *Avg.PurchaseValue* and their *ConversionProbability* is highest among all other segments, making Loyal Specialists a highly profitable and reliable segment. What is clearly evident, not only from the lowest *Avg.ProductPurchaseVariety* among all segments in Figure 4, but also from the skewed bubble chart in Figure 6, is the fact that these customers are mainly interested in a very narrow branch of product categories. The bubble chart shows that especially chokeberry juice is their specialty as the size ratio between the "regular juice" bubble and the other bubbles is the most extreme among all other segments. Members of this shop have the lowest mean for *sessions*, in light of their extremely high *ConversionProbability* and highly specialized interest in product categories one can speculate that member of this segment have a somewhat intent-driven tendency when they indulge in a shopping session on the website. In other words, they appear to visit with a purchase incentive (as they convert on most of their sparse sessions) and given their high product specialization one can assume that they know what they want to buy before they start the session. Figure 5 shows that this segment has the highest *Avg.QuantityAddedInCheckout*, which might be an indication that Loyal Specialists tend to indulge in rather impulsive behavior at the end of their sessions. This impulsiveness might be based on a certain trust and familiarity they associate with the products and the brand. A high cart abandonment rate on the other hand is an indication that this described impulsiveness is somewhat mitigated, perhaps by their intent-driven tendency. In other words, they manage at the end of their shopping session to evaluate which items and unit quantities in their baskets are in line and which exceed their shopping need, which is in itself a rather controlled and reflective behavior. Their tendency to add products to their basket and abandon a certain proportion again, might be an indication that Loyal Specialists use their cart like an organizational tool to assess prices and sort out the best suiting offer before they eventually purchase.

Loyal Specialists have a gender distribution which is almost equal for both females and males (Figure 7). They constitute the segment that has the highest tendency to shop from their desktop as depicted in Figure 8. The segment's loyalty is emphasized amongst other things by the segments high tendency to visit the shop directly by typing the shop URL into in the browser search field. Figure 11 shows that, more than any other segment, loyal specialists use search queries which include directly company related search terms such as the company name and location. This is an indicator that shopping sessions of Loyal Specialists are more frequently initiated with a clear intent to visit the shop than those of the other segments. This “company focused” shopping intent is also reflected by the fact that Loyal Specialists have low averages for paid advertising and referrals as visiting source - meaning they don't just react to marketing stimuli by clicking on an ad in their social media feed or on a link on a referral page, but tend to choose independently their time of visit. Considering Figure 10, the segment’s loyalty to the brand is further emphasized by relatively long times spent on the shop’s blog, which describes events that happen on the farm and news concerning the company. Their time spent on the juice info page is highest among all other segments, highlighting their strong specialized interest in this particular product category. Lastly, high visiting times of the organic certificate information page might be an indication Loyal Specialists value the green farming practices of the company. Interestingly, their average time spent on the “Imprint” page is also the highest among all other segments. Considering their loyalty to the brand and strong interest in its practices (reading blog and showing interest for green farming), their tendency to spend relatively long time on the imprint page might be interpreted as their tendency to gain more familiarity with the brand. In other words, these customers want to be close to the company, get a "behind the scenes" look into its practices and want to know the people behind the brand.

5.4.4. Customer Profile 3: Inquisitive Allrounders (Cluster 3)

This segment is the smallest with 502 members and constitutes 24.1% of the clustered clients. Members of this segment show moderate engagement levels as their mean values for *avgSessionDuration* and *pageviewsPerSession* in Figure 4 are clearly higher than those of Disengaged Frugals but slightly lower than those of Loyal Specialists. The average number of sessions is slightly higher than those of Loyal Specialists (Figure 5), but their *ConversionProbability* is clearly lower. Meaning, they quite frequently visit the shop just to brows without indulging in any purchase. This browsing tendency is further depicted by the highest *Avg.ProdViewVariety* meaning they look at a broad variety of products per session, broader than any other customer segment (Figure 4). Inquisitive Allrounders most frequently use promotion coupons (Figure 5), which could be an indication that they are more price sensitive, looking for better bargains before they commit to any purchase. This might be further reflected by the fact that they buy in large quantities as depicted by the highest mean for *Avg.PurchaseQuantity* (Figure 4), since the shop offers free shipping for purchase values above 30,00€ and coupon codes apply to the whole basket price, meaning buying in bulk results in better price deals. In light of the lower *ConversionProbability*, one could speculate that customers of this segment have "browsing-sessions" were they merely browse through products they are interested in (highest *Avg.ProdViewVariety)* and "purchase-sessions" where they overload their shopping cart (highest *Avg.QuantityAddedToCart*) and buy preferably with discounts in bulk - perhaps to profit from free shipping options.

Their inquisitive nature is mainly depicted by their broad interest in products, further emphasized by their utterly high *Avg.ProductPurchaseVariety*. Purchases of Inquisitive All-rounders are on average the most heterogeneous in terms of product categories, this is particularly shown in the bubble chart in Figure 6. The bubbles of each product category are of comparable size and not as skewed as those of Loyal Specialists. The fact that they most frequently use the search tool option (Figure 5) can be seen as a further exploratory indicator, in a sense that they actively examine the shop's product offer with their own search terms. 70% of Inquisitive All-Rounders are female (Figure 7), their average desktop and mobile usage is of comparable size (Figure 8).

The main visiting sources for Inquisitive Allrounders are organic search and google search ads (Figure 9). Compared to other segments, they rarely visit the page by directly typing in the company URL, but also rarely visit after reacting on marketing stimuli like Facebook/Instagram Ads and Referral links (Figure 9). Considering their search queries in Figure 11, one can clearly notice that they frequently include keywords relating to "Health.Topics", more than any other segment. These keywords are for instance terms like "immune system", "cardio vesicular system" and "vitamin C", which are all beneficially associated with the effects of chokeberry consumption. Unsurprisingly, their queries also incorporate more frequently than any other segment terms that relate to the different product categories of the company such as "powder", "dried berries" and vinegar", which is a further supporting indicator for their broad product interest. Lastly, Figure 10 show that Inquisitive Allrounders spend time, more than any other segment, on the About Us page and the FAQ page of the company. The FAQ page of the company includes mainly questions concerning the health and nutritional benefits of chokeberry product consumption such as "What is the scientific proof that chokeberry products induce health benefits?" or "How does choke berry consumption improve my cardio vesicular system?". Considering both, their search queries relating to health topics and their long time spent on the FAQ page (which mainly features answers to health-related questions), one might cautiously make the assumption that Inquisitive All-Rounders, more so than the other two segments, have a strong interest in health and nutritional topic fields.

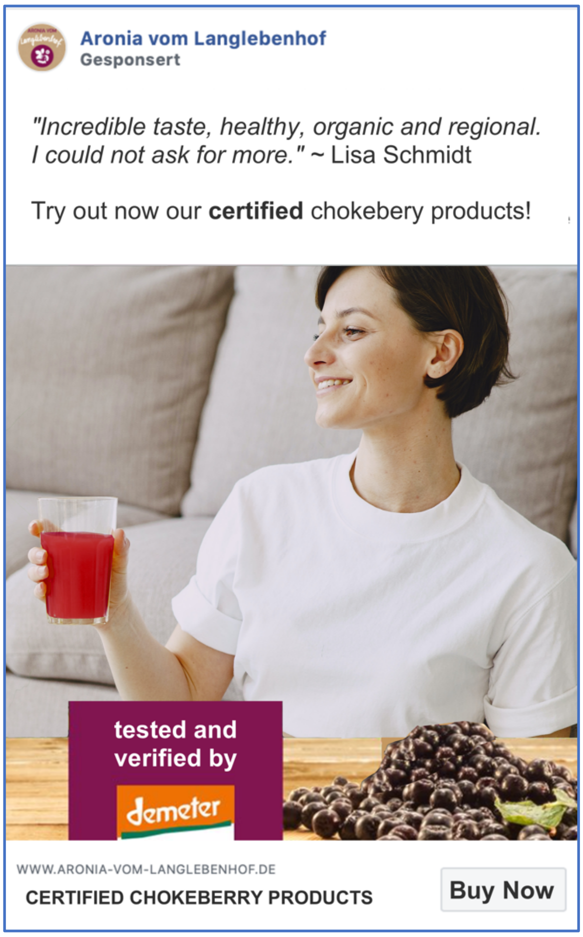
5.5. An Alternative Approach to Cold Advertising

In this additional section, I want to demonstrate how data-driven customer profiles can be used to improve current advertising practices to reach cold audiences. To target new prospective customers the capabilities of Facebook's "Lookalike Audience" or Google's "Similar Audiences" targeting tools can be used. Both tools allow marketers to advertise to new people who share very similar behavioral characteristics and preferences as existing customers. For instance, if a marketer has a set of email addresses of customers and she wants to find new people who share very similar interests and behavioral traits, she can use these email address list, also referred to as "seed", as an input for the Lookalike Audience/ Similar Audience algorithm. Both, the Lookalike Audience as well as Similar Audience targeting tool, use Facebook's or Google's data respectively, in order to find similar new users in their corresponding advertising network. For example: A marketer uses the contact details of customer "John Smith" as input for Facebook's Lookalike Audience targeting tool. Facebook then checks if John Smith is an Instagram or Facebook user – if so, John's Facebook/Instagram data is used to find new users who share very similar demographics, behavioral tendencies, interests and preferences as John.

The current use of the targeting tool in the industry (based on personal experience), is frequently based on unidimensional seeds, meaning individuals that constitute the seed share only one common denominator. One very popular approach is for instance to create Lookalike/Similar audiences of all people who have purchased something in the shop. As can be clearly seen from the cluster analysis, customers of the Langlebenhof shop are extremely heterogenous. Meaning, using the single dimension "purchase" as grouping variable might be rather ill advised and limited since customers have different shopping behaviors and interests, hence campaigns have to look differently for each of the three segments in order to perform very well. Furthermore, the search for new high performing customers is problematic. For instance, almost 50% of the customers of the Langlebenhof shop are members of the Disengaged Frugals segment. A seed based on "purchase" alone might skew lookalike/similar audiences towards this rather underperforming segment as Disengaged Frugals weigh in more than the other two segments. As an alternative and solution to these above discussed challenges, I propose to use the established multivariate clusters as seeds for lookalike audiences. In the next section I demonstrate how insights gained from the cluster analysis can be used to design personalized Facebook advertisements for each customer profile.

5.5.1. Sample Ad1: The Disengaged Frugals Lookalike

As discussed, Disengaged Frugals tend to have the lowest shop engagement rates among all segments. In light of this low engagement levels lookalike audiences that are based on the Disengaged Frugals seed should be targeted with strong stimuli. Meaning, the used creative should be highly sensational – for instance, pictures of humans are shown to be highly effective in grabbing attention. **(Desimone & Duncan, 1995)** Furthermore, the copy of the ad should be short and catchy as it can be presumed that Disengaged Frugals Lookalikes also won't have the lasting engagment to read a long-detailed story or informative product description. Given the tendency of Disengaged Frugals to be cautious and reserved in their consumption, trust arguments like the Demeter certification in the creative, a catchy customer review in the copy as well as the repeated use of the word "certified", might increase click-through-rates especially with this lookalike audience. Two thirds of email addresses used as seed for this Lookalike Audience belong to women. It might increase click-through-rates if the used creative portrays a woman or if the review in the copy belongs to a female customer. Lastly Facebook ads for Disengaged Frugals lookalikes should be mobile optimized given the fact that their seed's device usage is strongly mobile focused. In fact, it might be advisable to advertise only on mobile devices and ignore desktop and tablet all together to be more efficient with the advertising budget spend.



Aronia vom Langlebenhof
Gesponsert

"Incredible taste, healthy, organic and regional. I could not ask for more." ~ Lisa Schmidt

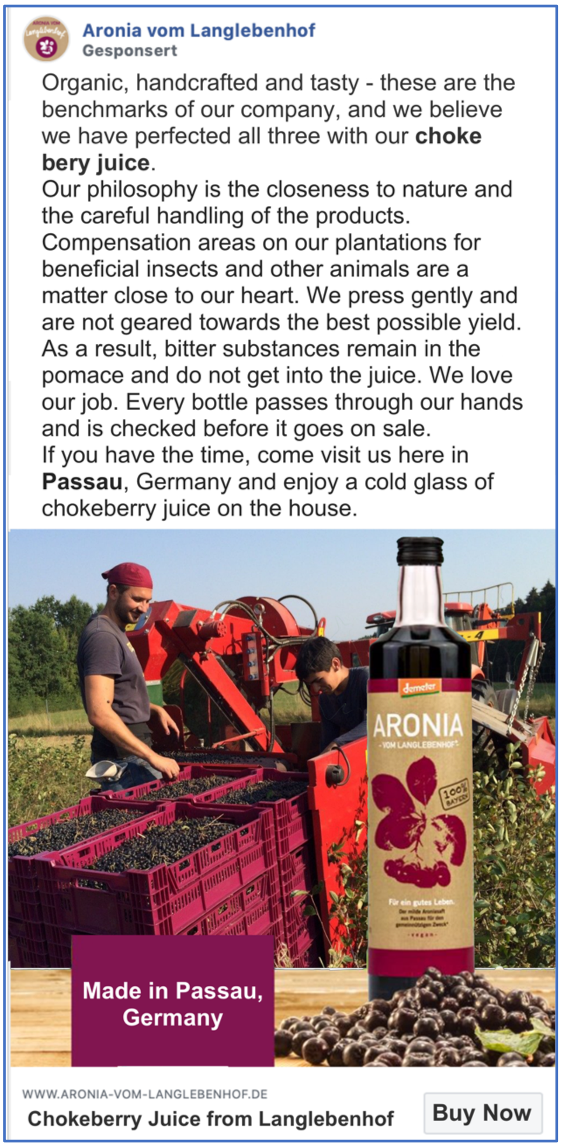
Try out now our **certified** chokebery products!

WWW.ARONIA-VOM-LANGLEBENHOF.DE
CERTIFIED CHOKEBERRY PRODUCTS

Figure 12: Sample Ad1 – Disengaged Frugals Lookalikes

5.5.2. Sample Ad2: Loyal Specialists Lookalikes

Loyal Specialists have the highest engagement with shop content among all other segments. It was established that they spent the most time on the company's blog and show strong interest for the Langlebenhof brand and their all-time favorite product category is choke berry juice. Hence the copy of a Facebook ad that targets Loyal Specialist Lookalikes can be in a long detailed narrative format. It should not only thematize the juice but also give the reader a portrait of the brand, its organic handcrafted practices and the people behind the company, as Loyal Specialists show a stronger interest than other segments in the brand and company practices. The used creative should put a juice bottle in the foreground and incorporate a banner that refers to Passau, Germany to give the reader a feeling of familiarity and closeness to the brand.



Aronia vom Langlebenhof
Gesponsert

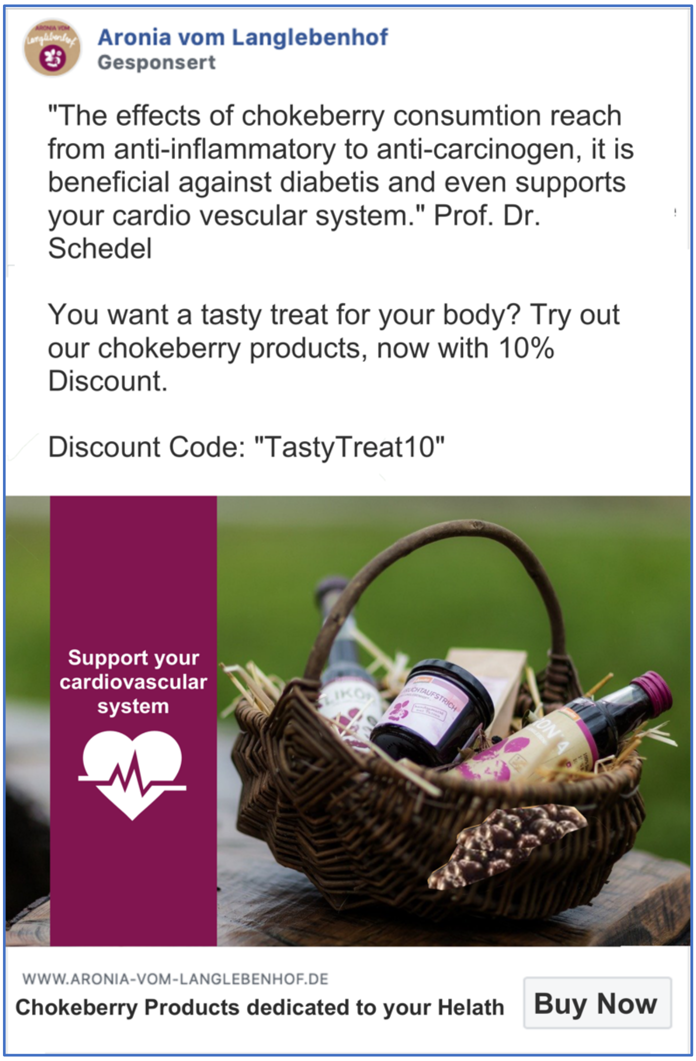
Organic, handcrafted and tasty - these are the benchmarks of our company, and we believe we have perfected all three with our **choke bery juice**.
Our philosophy is the closeness to nature and the careful handling of the products. Compensation areas on our plantations for beneficial insects and other animals are a matter close to our heart. We press gently and are not geared towards the best possible yield. As a result, bitter substances remain in the pomace and do not get into the juice. We love our job. Every bottle passes through our hands and is checked before it goes on sale.
If you have the time, come visit us here in **Passau**, Germany and enjoy a cold glass of chokeberry juice on the house.

WWW.ARONIA-VOM-LANGLEBENHOF.DE
Chokeberry Juice from Langlebenhof

Figure 13: Sample Ad2 – Loyal Specialists Lookalikes

5.5.3. Sample Ad3: Inquisitive Allrounders Lookalikes

The central topics that should be incorporated in an ad that targets Inquisitive Allrounder Lookalikes is 1) the variety of products offered by the brand and 2) the health benefits provided by chokeberry consumption. For instance, the used creative could show a product bundle that incorporates all products on the website. An emblem that is associated with medical care can be incorporated into the creative so visually the health topic is evident to the reader. The copy could incorporate answers of the most relevant health questions on the FAQ page in medium length format given moderate engagement levels of the Inquisitive Allrounder seed. Lastly, given the high usage of discount coupons of Inquisitive Allrounders, a discount code offered in the copy might also increase click-through rates for lookalikes.



Aronia vom Langlebenhof
Gesponsert

"The effects of chokeberry consumtion reach from anti-inflammatory to anti-carcinogen, it is beneficial against diabetis and even supports your cardio vescular system." Prof. Dr. Schedel

You want a tasty treat for your body? Try out our chokeberry products, now with 10% Discount.

Discount Code: "TastyTreat10"

WWW.ARONIA-VOM-LANGLEBENHOF.DE
Chokeberry Products dedicated to your Helath

Figure 14: Sample Ad3 – Inquisitive Allrounders Lookalikes

6. Conclusion: Answering the Research Question

In this concluding part I want to refer back to my originally stated research question: **"How can targetable customer profiles be built based on Google Analytics data using machine learning clustering algorithms in R?"**

Data tracking allows to capture the entire customer journey of each individual shop visitor, which in turn gives marketers the foundation to build more empirical customer profiles.

On the basis of Web Usage Mining behavioral patterns can be unleashed, more specifically, a deeper understanding about customer tendencies and preferences can be established with the help of specific behavioral metrics. Google Analytics is the most popular tracking tool in the ecommerce industry and can be used to build these metrics. Furthermore, the Google Tag Manager can be utilized to track basically anything on a website and create even more customized metrics. In order to manipulate data and apply machine learning algorithms the Google Analytics Dashboard needs to be abandoned, and data should be imported into the statistical programing software R using the Google Analytics Reporting API. Once data is imported and Google Analytics browsing data has been linked to CRM email lists, customers can be grouped based on aforementioned behavioral metrics. The K-means algorithm can be used to build clusters - as an alternative, the Reduced K-means algorithm can be applied if variable selection is of concern. The quality of built clusters should be assessed in terms of how well they are separated from each other or if they overlap, which is depicted by the silhouette score. Next to this quality assessment, cluster stability should be analyzed in order to verify if clusters represent customer segments that actually exist in the real world or if they were artificially created by the algorithm. Based on stable cluster results detailed customer profiles can be established which can be used for the design of personalized advertising. Thanks to the linkage between Google Analytics Data and CRM email lists, clusters are actually targetable, meaning members of a cluster can be reached with new advertising campaigns.

An alternative cold advertising approach is to use email addresses of each cluster as seed to create Lookalike Audiences on Facebook or Google. That way cold audiences, which resemble established customer profiles, can be reached with personalized advertising that can be designed on the basis of cluster results.

7. Limitations & Future Research

The first limitation that needs to be addressed is concerned with data privacy and legality. Deanonymizing customer data and building customer profiles is generally not forbidden. But data regulation laws dictate that website visitors need to be informed about a company's profiling agenda in the privacy policies and visitors need to confirm to these practices. In other words, in order to use data of existing customers, the company needs to inform clients about new data analysis practices and ask for permission. Otherwise the company needs to collect data of new customer, who consent to the renewed privacy policies right away.

One of the key limitations of the profiling methodology discussed in this paper is based on a common problem with data tracking systems that rely on cookies in the identification process of returning customers. This type of identification can be unreliable to a certain extent as many users erase their cookie caches after each session or switch devices and browsers. Furthermore, many browser providers altered their default settings in recent years to automatic cookie deletion, as it is the case with Mozilla Firefox and Safari, where cookies are deleted either every time the browser is closed, or after 24 hours have passed. In context of Google Analytics, this means that new unique ClientIDs are assigned every time cookies are not accessible, meaning the system might give several unique ClientIDs to one particular customer. By connecting Google analytics data to the CRM data base of the company, I managed to identify some of the multiple unique ClientIDs that have been assigned to customer based on the transactionID. Still, it can be expected with certainty that not all assigned unique ClientIDs were fully recovered. This is in turn means that not all sessions of customers were used to create customer profiles, but solely those that were identifiable. One way to solve this problem is to change the tracking system from cookie-based identification to log in based identification. For instance, clients could be motivated to create an account on the shop and encouraged to log in every time they visit the shop or remain logged in. By adopting this tracking methodology, users could be constantly identified even if they delete their cookie cache or change the browser or device. Luckily, Google Analytics offers the option to set up such an identification system, so the introduced profiling method can easily be improved.

A further limitation of this research is based on the profiling methodology itself. The method hinges on the assumption that customers have a certain behavioral tendency that remains somewhat stable across their customer journey. By calculating averages of variables for at least three sessions, it is expected that a generalizable behavioral snapshot of each customer can be captured. But It might be the case that customers actually change their behavior from session to session. For instance, a given client might engage thoroughly with content of a shop and explore a broad range of products in his first three sessions. But once he has figured out what he is actually interested in, he might reduce his visiting times and view less products as he always ends up buying the same old product. Hence, although averages might capture a general behavioral trend of customers, they fail to detect behavioral fluctuations that might occur from session to session, which might offer marketers further valuable information. Future research could explore a method that uses sessions as input for clustering algorithms instead of average session aggregations for each customer to detect aforementioned behavioral fluctuations. Meaning, instead of representing customers with different behavioral trends, resulting clusters would depict shopping states customers have been in, allowing for customers to be in multiple clusters at the same time. For each customer, all shopping states could be summarized in a stream. Based on this resulting "shopping state stream", similar customers could be grouped together yielding customer profiles. Results could be compared to those of the methodology introduced in this paper in order to assess if one approach is superior or if they substitute each other in their different nature.

8. Bibliography

Arabie, P., & Hubert, L.J. (1994). Cluster analysis in marketing research. In R.P. Bagozzi (Ed),*Advanced methods in marketing research* (pp. 160–189). Oxford, England: Blackwell.

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or Fun: Measuring Hedonic and Utilitarian Shopping Value. Journal of Consumer Research, 20(4), 644. doi: 10.1086/209376

Bang, H., & Wojdynski, B. W. (2016). Tracking users visual attention and responses to personalized advertising based on task cognitive demand. Computers in Human Behavior, 55, 867–876. doi: 10.1016/j.chb.2015.10.025

Bleier, A., & Eisenbeiss, M. (2015). Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where. Marketing Science, 34(5), 669–688. doi: 10.1287/mksc.2015.0930

Brys, M. (2017*). Using Google Analytics with R*. Krakow: Michal Brys

Chourasiya, S., & Jain, S. (2019). A Study Review On Supervised Machine Learning Algorithms. *International Journal of Computer Science and Engineering,6*(8), 16-20. doi:10.14445/23488387/ijcse-v6i8p104

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annual Review of Neuroscience, 18(1), 193–222. doi: 10.1146/annurev.ne.18.030195.001205

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. *New Approaches in Classification and Data Analysis Studies in Classification, Data Analysis, and Knowledge Organization,*212-219. doi:10.1007/978-3-642-51175-2_24

Edmondson, M. Kletsov, A., DeBoer, J., Watkins, D. Brode-Roger, O., Sohi, J., Selinger, Z., Corlade, O. (2019). Google Analytics API into R. *CRAN.* Version 0.7.1

Forgy, E.W. (1965) Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. Biometrics, 21, 768-780.

Gentleman, R., & Carey, V. J. (2008). Unsupervised Machine Learning. *Bioconductor Case Studies,*137-157. doi:10.1007/978-0-387-77240-0_10

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How Is the Mobile Internet Different? Search Costs and Local Activities. Information Systems Research, 24(3), 613–631. doi: 10.1287/isre.1120.0453

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. Expert Systems with Applications, 100, 1–16. doi: 10.1016/j.eswa.2018.01.029

Hadi, A. S., Kaufman, L., & Rousseeuw, P. J. (1992). Finding Groups in Data: An Introduction to Cluster Analysis. *Technometrics,34*(1), 111. doi:10.2307/1269576

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis,52*(1), 258-271. doi:10.1016/j.csda.2006.11.025

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods. Journal of Marketing, 73(2), 55–69. doi: 10.1509/jmkg.73.2.55

Jaccard, P. (1901) étude Comparative de la distribuition florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 7, 547-579.

Keyzer, F. D., Dens, N., & Pelsmacker, P. D. (2015). Is this for me? How Consumers Respond to Personalized Advertising on Social Network Sites. Journal of Interactive Advertising, 15(2), 124–134. doi: 10.1080/15252019.2015.1082450

Kukar-Kinney, M., & Close, A. G. (2009). The determinants of consumers' online shopping cart abandonment. Journal of the Academy of Marketing Science, 38(2), 240–250. doi: 10.1007/s11747-009-0141-5

Lilien, G. L., Rangaswamy, A., & De_Bruyn, A. (2017). *Principles of marketing engineering and analytics*. State College, PA: Decisionpro.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory,28*(2), 129-137. doi:10.1109/tit.1982.1056489

Markos, A., Iodice D'enza, A., & Van de Velden, M. (2019). Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. *Journal of Statistical Software,91*(10). doi:10.18637/jss.v091.i10

Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000). Integrating Web Usage and Content Mining for More Effective Personalization. Electronic Commerce and Web Technologies Lecture Notes in Computer Science, 165–176. doi: 10.1007/3-540-44463-7_15

Moe, W. W. (2003). Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream. Journal of Consumer Psychology, 13(1-2), 29–39. doi: 10.1207/s15327663jcp13-1&2_03

Neelima, G., & Rodda, S. (2015). An Overview on Web Usage Mining. Advances in Intelligent Systems and Computing Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2, 647–655. doi: 10.1007/978-3-319-13731-5_70

Patel, K., & Patel, D. A. R. (2012). Process of Web Usage Mining to find Interesting Patterns from Web Usage Data. International Journal Of Computers & Technology, 3(1), 144–148. doi: 10.24297/ijct.v3i1c.2767

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,2*(11), 559-572. doi:10.1080/14786440109462720

Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management,32*(3), 477-481. doi:10.1016/j.tourman.2010.03.015

Raphaeli, O., Goldstein, A., & Fink, L. (2017). Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. Electronic Commerce Research and Applications, 26, 1–12. doi: 10.1016/j.elerap.2017.09.003

Silverbauer, J. (2017). Tracking Website Data with Google Tag Manager. Journal of Brand Strategy, 242-249(8)

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing,21*(1), 3. doi:10.2307/1247695

Sorce, P., Perotti, V., & Widrick, S. (2005). Attitude and age differences in online buying. International Journal of Retail & Distribution Management, 33(2), 122–132. doi: 10.1108/09590550510581458

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology,59*(1), 1-34. doi:10.1348/000711005x48266

Usage statistics of traffic analysis tools for websites. (n.d.). Retrieved August 02, 2020, from https://w3techs.com/technologies/overview/traffic_analysis

Verdenhofs, A., & Tambovceva, T. (2019). Evolution of Customer Segmentation in the Era of Big Data. *Marketing and Management of Innovations,*238-243. doi:10.21272/mmi.2019.1-20

Wang, Q., Wang, C., Feng, Z., Ye, J. (2012). Review of K-means clustering Algorithm. *Electron des. Eng.* , 20, 21-24

Weber, J. (2015). Google Tag Manager and Google Analytics APIs. *Practical Google Analytics and Google Tag Manager for Developers,*257-263. doi:10.1007/978-1-4842-0265-4_16

Wolin, L. D., & Korgaonkar, P. (2003). Web advertising: gender differences in beliefs, attitudes and behavior. Internet Research, 13(5), 375–385. doi: 10.1108/10662240310501658

Xu, Y., & Huang, J.-S. (2015). Factors Influencing Cart Abandonment in the Online Shopping Process. Social Behavior and Personality: an International Journal, 43(10), 1617–1627. doi: 10.2224/sbp.2015.43.10.1617

Yi, X., Hong, L., Zhong, E., Liu, N. N., & Rajan, S. (2014). Beyond clicks. Proceedings of the 8th ACM Conference on Recommender Systems - RecSys 14. doi: 10.1145/2645710.2645724