

Library user segmentation by transactional log data using k-means

Yvonne Maria van de Glind

ANR: u570071

SNR: 2021449

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE SCHOOL OF
HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Supervisor

dr. ir. Martijn van Otterlo

Second Reader

prof. dr. Eric Postma

Tilburg University
School of Humanities and Digital Sciences Department of Cognitive Science &
Artificial Intelligence Tilburg, The Netherlands
June 2019

Preface

The thesis presented here is written within the framework of the Master of Science in Data Science and Society. This thesis was carried out in cooperation with Cubiss and Tilburg University. For the completion of my thesis I would like to thank the following people. First of all, I would like to thank Martijn van Otterlo for supervising my thesis and for all of the feedback and consultation moments he provided along the way. I would also like to thank Martijn van Otterlo and Eric Postma for evaluating my thesis.

Second, I would like to thank Cubiss and in particular Frederik Theuwis for partially coordinating this project and for the suggestions and advices he provided. I also would like to thank the library involved for making their data available for this project, for the consultation moments they provided and for always answering my questions related to the datasets.

Finally, I would like to thank my parents and brother for their support and all others who have provided me with feedback and revised my texts.

Abstract

This thesis aims to find user segments for Dutch public libraries based on behavioral log data. These can mainly be used for marketing implementations, but might be useful to prove libraries' value as well. A literature review revealed that there has been done data science research in libraries for a variety of areas, among which segmenting users by using a clustering algorithm. This research uses Dutch transactional library data to find user segments that depend on user behavior, which is captured in the main research question: *“To what extent can meaningful user segments be found in behavioral log data in terms of moment, place and nature of the users' actions?”* The dataset used in this study was provided by a Dutch public library. After cleaning the dataset, the data from the year 2018 was used to investigate further. From this data 44 features expressing user behavior were created, in terms of; visits per day of the week, total number of actions, number of different locations visited, actions per moment of the day, average actions per visit, actions per material type/genre, and actions occurring online or offline. After applying the k-means algorithm, the six customer segments “passive users”, “average users”, “Tuesday visitors”, “users with children”, “Wednesday visitors”, and “active users” were found. It was found that these segments are mostly characterized by the amount of user actions, the day of the week on which user actions occur, and actions concerning the borrowed material type and genre, in case the material type is a book.

Table of contents

1. Introduction	5
2. Related Work.....	7
2.1 Data science in libraries	7
2.2 User segmentation and clustering in libraries	8
2.3 User segments in Dutch public libraries.....	10
2.4 Cluster tasks with similar data in retail settings	12
2.5 Conclusion.....	13
3. Methods	13
3.1 Data	13
3.1.1 Data description.....	13
3.1.2 Preprocessing and data cleaning.....	14
3.1.2.1 Missing values.....	14
3.1.2.2 Outliers	15
3.1.3 Data analysis.....	15
3.1.4 Feature engineering	18
3.2 Algorithm	20
3.3 Software.....	21
4. Results	22
4.1 Feature matrix with unnormalized values	22
4.2 Feature matrix with normalized values	23
5. Discussion	27
5.1 Research questions and results	27
5.2 Limitations and future work	30
6. Conclusion.....	31
References	32
Appendix A: Original columns of the transactional log dataset	34
Appendix B: Original columns of the collection dataset	36
Appendix C: Allocation of codes in the column ‘bron’	37
Appendix D: Allocation of codes in the column ‘ama’	38
Appendix E: Allocation of genre codes (originally from the column ‘scatnr’)	40
Appendix F: Cluster sizes and distributions for each number of k for the unnormalized feature matrix	42
Appendix G: Cluster sizes and distributions for each number of k for the normalized feature matrix	46
Appendix H: Mean and standard deviation per feature per cluster for $k=6$ with the feature matrix with normalized values	50

1. Introduction

Public libraries worldwide have always been adapting to changes in society. The growing reach of the internet is such a change, which has influenced libraries over the last couple of decades. Nowadays, information is easily accessible and omnipresent due to the progress of internet and technology. This has led books and libraries as a source of information to become more and more unused. In Dutch public libraries, this is noticeable in the number of books (or other material types) borrowed, which decreased with 46.11% from 2005 to 2017. The number of library members decreased as well, but only with 8.22% (Central Bureau for Statistics, 2018).

Because libraries as only a place to borrow books have become of insufficient interest, libraries worldwide had to reconsider their purpose in order to stay valuable (Field & Tran, 2018). One way in which libraries change their identity to stay relevant for their users, is by becoming a venue for discussion and education (Field & Tran, 2018). However, many library users are unaware of these new services that libraries provide (Zickuhr, Rainie, & Purcell, 2013).

In the Netherlands the identity of public libraries is categorized in five requirements, which include both traditional and relatively new services. These requirements are provided by the Association of Public Libraries in the Netherlands (Vereniging Openbare Bibliotheken, henceforth VOB). They state that public libraries are a place for (1) the collection of knowledge and information, (2) development and education, (3) art and culture, (4) inspiration for reading and literature, and (5) meetings and debates (VOB, 2015, for more information see section 2.2).

Another consequence of the rise of the internet, is that public libraries are not as taken for granted as they once were. This has led libraries to increasingly having to prove their value to their investors (Vakkari, Aabø, Audunson, Huysmans, & Oomes, 2014) based on evidence (Siguenza-Guzman, Saquicela, Avila-Ordóñez, Vandewalle, & Cattrysse, 2015).

Ball (2019) notices that data science can directly help libraries, as they could use tools to analyze their already collected data in order to better understand their users. One way to accomplish this is by creating customer segments. Customer (or market) segments are groups of consumers with comparable desires and needs. This study aims to find customer segments for a number of Dutch public libraries, using data science methods. These segments can help libraries with the developments mentioned above. First, customer segments can be used as a basis for carrying out marketing efforts, which can be aimed to attract new library members and stimulate existing members to visit the library more often. Second, another aim of marketing based on customer segments could be to make patrons (library users) aware of the new services provided by libraries. Third, insight into library users and their behaviors might help libraries provide evidence-based arguments explaining their value.

The need of marketing for libraries is also described by Garoufallou, Siatri, Zafeiriou, and Balampanidou (2013). They underline that it has become essential for libraries to implement marketing strategies. In addition, they argue that customer segmentation is an important and powerful marketing tool which can help identify user needs. Furthermore, knowledge about customer segments improves marketing strategy decision making (Khajvand, Zolfaghar, Ashoori, & Alizadeh, 2011; Chiu, Chen, Kuo, & Ku, 2009). The need for more insights in patron behavior and desires is also noticed by the Dutch Royal Library (Koninklijke Bibliotheek, henceforth KB) (KB, 2018). They anticipate that new techniques increasingly will be used in order to accomplish this.

Not only has the expansion of the internet changed the library's purpose, it has changed users' expectations of libraries as well. Library users have become used to the systems of companies such as Google and Amazon, and they now expect similar systems from libraries as well (Knott, Bradley, DeGeorge, & Ottaviani 2007; Zickuhr et al., 2013). A recommendation system is a much-discussed suggestion for such a system. Customer segments could contribute to recommendation systems by making recommendations based on segments.

User segments of public libraries have been created before (see section 2.2). One Dutch example is 'De klant is koningin' (The customer is queen) (VOB, 2008). This project found seven customer segments of Dutch public libraries (see section 2.3). However, in this project only adult library users were considered. In the last 12 years the composition of Dutch library users has changed, such that there are more library members aged younger than 17 than older than 17. In 2017, the composition of members even consisted of almost twice as many members under 17 years than over 17 years (Central Bureau for Statistics, 2018). Because of the shift in composition of Dutch library members, it would be valuable to investigate the customer segments of Dutch public libraries again. Apart from that, it would be interesting to see if some characteristics of the segments have changed in other ways as well. However, the segments would not be completely comparable, because of difference in input data.

The dataset that will be used for this research consists of behavioral library user data. It is provided by a Dutch public library consisting of multiple locations. The dataset consists of behavioral data of the library users over the period of January 1, 2014 until January 1, 2019. In the data, users' actions can be distinguished, which are annotated with user id, time and date, location, subscription type, the type and genre of the borrowed material, and more (see section 3).

The aim of this research is to find user segments in the dataset by clustering the data. This leads to the following research question and sub questions:

1. To what extent can meaningful user segments be found in behavioral log data in terms of moment, place and nature of the users' actions?

1.1 Which features should be entered into the model?

1.2 Which typical characteristic features can be distinguished within the segments?

Besides the goal of this research to discover user segments for public libraries, it has a broader goal as well. This research originates from a collaboration between Tilburg University and Cubiss. Cubiss is a service organization for organizations that deal with the themes reading, learning and informing, among which public libraries. Cubiss' goal for this collaboration is to discover the potential value of data science for public libraries and to make public libraries aware of this. Hence, the broader goal of this research is to explore the possibilities of data science for public libraries.

In this research, the six user segments "passive users", "average users", "Tuesday visitors", "users with children", "Wednesday visitors", and "active users" were found. These segments are mostly characterized the amount of user actions, the day of the week on which user actions occur, and actions concerning the borrowed material type and genre, in case the material type is a book. The segment "passive users" consists of most users (63.82%) which perform relatively few actions. The highest total action average is found in the segment "active users", which is the smallest cluster. "Tuesday visitors" mainly perform actions on Tuesdays, and have a preference for books about action and thrillers, and large lettered books. "Users with children" on average perform by far the most actions per visit, they visit the library mainly on Wednesday and show high averages for a wide variety of books, both for adults and children. "Wednesday visitors" typically visit the library on Wednesday and

borrow books about fairytales, comic strips, and books for the age 6 to 15. Lastly, “Average users”, the second largest user segment, is characterized by the fact that for every feature they neither score a highest or a lowest average.

Following to this introduction, first a section is dedicated on giving an overview of the previous work, related to the context of this research. In particular the topics of data science in public libraries, creating user segments both in library as in retail settings, and user segments specific for Dutch public libraries are addressed. Then follows a section about the dataset and methods involved, where insights into the data, feature engineering, algorithm, and validation methods are given. Thereafter the results are presented, followed by a discussion of those results. Finally, a conclusion is drawn.

2. Related Work

In this chapter, work related to the present study is discussed. First a short overview is given of research on data science in libraries worldwide, after which cluster applications thereof are focused on. Then user segments in Dutch public libraries are discussed, followed by research on retail customer segmentation.

2.1 Data science in libraries

Just as for a broad range of fields, sophisticated data science techniques can be useful for libraries. According to Litsey and Mauldin (2018), librarians already make use of static data and statistics, but they could maximize the benefits of their data by using machine learning and predictive analytics. In addition, Massis (2012, 2016, 2018) discusses the benefits of predictive analytics, data-driven decision making and artificial intelligence for libraries. Ball (2019) also argues that sophisticated tools can help libraries to use their data to better understand their users and improve the way they serve them.

Data mining for libraries can be referred to as bibliomining. Nicholson and Stanton (2003) first introduced this term. Although the term at that time was new, the concept was not (Siguenza-Guzman et al., 2015). This indicates that sophisticated data science techniques for libraries have been explored for some time already. In their literature research, Siguenza-Guzman et al. (2015) reviewed 41 articles that use data mining techniques for academic libraries. The data mining methods could be categorized as association, classification, clustering, and regression. These methods were used for the analysis of the service, collection, usage, and quality of academic libraries. More concrete, bibliomining can for instance have applications for library workflow and foreseeing user needs (Siguenza-Guzman et al., 2015).

Some other examples of practical applications of sophisticated data science techniques in libraries are analysis of user search behavior through library catalogs (Niu & Hemminger, 2015), analysis of libraries’ social media channels (Al-Daihani & Abrahams, 2016, 2018; Palmer, 2014), and even optimizing physical library spaces (van Otterlo & Warnaar, 2017; Silva & Miguéis, 2017). Examples of applications that are suitable to be driven by user loan data are data-driven collection management (Showers, 2015), creating recommendation systems (Tsuji, Yoshikane, Sato, & Itsumura, 2014; Yi, Chen, & Cong, 2017), predicting customer behavior such as churn (Lai & Zeng, 2014), and creating user segments (see section 2.2).

2.2 User segmentation and clustering in libraries

Libraries have much experience with collecting and analyzing data about their users. For example through surveys, observational studies (van Otterlo & Warnaar, 2017), and transactional data (Eden, 2015). Ullah and Ameen (2018) found that the most applied data gathering method in Library and Information Science (LIS) research are surveys (33% usage), followed by theoretical analysis (7.2%), and content or protocol analysis (7%). Transactional log analysis as a method appeared to be a dominant research method, but with only 1.5% usage. Cluster analysis was found to be a rarely used research method. Transactional log analysis is an appealing method because nowadays transactional log data is automatically collected. In addition, as this data is fact-based it has an advantage over surveys which reflect the more subjective user's perception (Boter & Wedel, 2005).

Although clustering is not the most explored method for libraries, libraries have been segmenting their users at least since the late 1970s (Johannsen, 2014). During the period 1980 until 2014 there are five principal approaches used by public libraries to segment their users: (1) based on patterns about library use and preferences, (2) based on demographical and sociological information about library users, (3) based on lifestyle characteristics of library users, (4) based on the information from 1 to 3 combined, by creating computer-generated segments, (5) based on a combination of qualitative and quantitative data in order to create user images (Johannsen, 2014). This research uses the fourth approach. Johannsen (2014) found this approach to have a high reliability, high public library relevance, high measurability, and high predictability, making it the highest scoring approach of the five.

One large scale data science project considering user segments was carried out in the USA. Customer and checkout data from ten public libraries was combined in order to find user segments. The goal of this project was to find core customers, in order to retain those customers and attract more customers of that segment. The data analysis was done by Fuzzy Logix, a big data analysis firm (Futterman & Milam, 2016).

Tempelman-Kluit and Pearce (2014) clustered library user data with k-means clustering, to use as a basis for personas ("hypothetical users that represent the behaviors, goals, and values of actual users", Tempelman-Kluit & Pearce, 2014, p. 616). The data used consisted of 168 transcripts of the "Ask a Librarian" chat service of the New York University library. The segments were based on the information asked by the user and the user's motivation to ask that information. The k-means algorithm was executed two times with an initial value of three and four clusters. The distribution with four clusters was chosen as the basis for the personas, because of the more equal variance within the clusters.

Lai and Zeng (2014) created customer segments of a Chinese digital library, based on behavioral data. The input features they used were; number of downloaded materials, the time between the first and last registered download, and downloading frequency per user. Based on a two-stage segmentation method, three clusters were found, which were described as "key customers", "common customers", and "loyal customers" (Lai & Zeng, 2014, p. 378). The segments differed considerably in size; 2.2, 73.5, and 24.3 percent, respectively.

Hájek and Stejskal (2014) researched customer segments for the largest public library in the Czech Republic. The aim of their study was to find typical library users based on their behavior, which is very similar to this study's aim. However, they used survey data which is very dissimilar from this study's data. The segments were based on users' socio-economic and demographic information, combined with their behavior. The users' behavior was captured in terms of number of books borrowed, occurrence rate of borrowing books, their

actions in the library, and their interaction with library staff. Hájek and Stejskal (2014) used k-means clustering. They chose the fixed number of 11 clusters, taken into account the utility function of their model. 34 features were entered in the model, 30 of which were yes/no questions. The remaining features were gender, age, education and socio-economic status. The largest cluster they discovered covered 20% of all users and the smallest cluster covered 2%. The four biggest clusters were “independent borrowers” (20%), “borrowers requiring assistance in searching” (17%), “users searching for a specific title” (15%), and “users looking for rest and relaxation” (11%). All other segments consisted of less than 10% of the users.

Hsu, Lu, and Lin (2012) proposed a concept hierarchy method for clustering data in order to find customer segments in retail settings. They used library data to test their method. The data used is similar to the data used in this research. The data consisted of three months of library records from a private college in Taiwan. Their features were “record sequence number, reader ID, book ID, book name, category number for the book, and the record date” (Hsu et al., 2012, p. 6225). They found the optimal number of clusters to be eight, based on the combination of three validation indices.

Jin and He (2017) used k-means clustering to create user segments for a Chinese academic library. They focused on data about undergraduates in the period of December 2015 to December 2016, which resulted in a dataset consisting of 8556 users. After 10 iterations, four clusters were found based on the number of books borrowed per user. The first cluster, described as “average readers”, had a size of 51% and the average books borrowed per user were 7.8. Cluster 2 “negative readers” (12%), on average borrowed 1.8 books per user. Cluster 3 “stable readers” (29%), borrowed 19.2 books on average. The smallest cluster (8%), defined as “active readers” had the highest average of books borrowed; namely 35.

Boter and Wedel (2005) used transactional loan data of five Dutch public libraries to find user segments, which has many similarities with the focus of this research. However, the four segments they discovered were based on the books borrowed. This differs from this research, which does not make this limitation, i.e. the segment outcomes can be based on more than book checkouts, for example the time and date of the user actions. The practical goal of Boter and Wedel’s (2005) study was to adjust the library’s book presentation according to customer needs. The data they used is very similar to the data used in this study. Main fields in their dataset were “customer number; a title number and full title details; date and time of borrowing; and age and gender of the borrower” (Boter & Wedel, 2005, p. 191). The main fields in this study’s dataset overlap with Boter and Wedel’s, except for age and gender. They found four segments. The first segment consisted of titles from most categories, but mainly American romantic fiction translated into Dutch. The second segment consisted of mainly literary fiction. The third segment consisted of mainly detective and thriller novels. The fourth segment mostly consisted of Dutch romantic fiction.

Garoufallou et al. (2013) reviewed literature on marketing for LIS, and affirm that segmentation belongs to the most important marketing concepts. Polger (2019) also emphasizes the importance of dividing library users into segments. These segments are important because they allow to communicate with users in an effective way. According to Garoufallou et al. (2013), libraries carry out marketing on research level, strategic level (e.g. mission and visions), promotional level, and at the level of products and services. The research level focusses on investigating the market and customers in play, including customer needs and opinions. They also notice that libraries operate their marketing both offline, in the library, and online, through their website and social media channels.

On a strategic marketing level, the mission of Dutch public libraries is “the elevation of the people” (de Jonge, 2016). According to the guidelines for Dutch public libraries, a public library should be a center for (1) the collection of knowledge and information, (2) development and education, (3) art and culture, (4) inspiration for reading and literature, and (5) meetings and debates (VOB, 2015). The strategic level is developed, nonetheless Malschaert (2016) argues that Dutch public libraries should focus more efforts on marketing campaigns, i.e. the promotional level.

2.3 User segments in Dutch public libraries

Users of Dutch public libraries have been studied regularly by or on behalf of organizations associated with public libraries. In 2008, the VOB has extensively researched customer segments of Dutch public libraries. One main conclusion of this research was that public library users in the Netherlands are mostly female (VOB, 2008). This is also confirmed by Tiessen-Raaphorst and van den Broek (2016). The VOB (2008) found seven customer segments, based on a sample of member data provided by Dutch public libraries, combined with data from a survey among Dutch library members. An overview of these segments can be found in Table 1.

Table 1: Customer segments discovered by VOB (2008).

Segment name	Segment size	Gender	Age
‘Studenten en starters’ (Students and starters)	4%	78% women	Mostly between 21 and 29
‘Stedelingen’ (Urban citizens)	8%	68% women	Mostly between 25 and 45
‘Gepensioneerde gezelligheidszoekers’ (Retired social seekers)	5%	68% women	Mostly between 60 and 80
‘Cultuurgenieters’ (Culture lovers)*	25%	73% women	Mostly older than 50
‘Dynamische gezinnen’ (Dynamic households)	14%	79% women	Mostly between 30 and 60
‘Traditionele gezinnen’ (Traditional households)**	31%	79% women	Mostly between 40 and 60
‘Welvarende genieters’ (Prosperous enjoyers)	11%	78% women	Mostly between 40 and 60
Other	2%		

* This segment can be divided into two subgroups.

** This segment can be divided into three subgroups.

The VOB (2008) found that the library is visited most frequently in the afternoon by most of the segments. The visits mostly take place during the week, with the exception of urban citizens, who visit the library mostly during weekends. Retired social seekers visit the library most often; one quarter of that segment visits the library once a week or more. Culture lovers, dynamic households, traditional households and prosperous enjoyers all visit the library once per two to three weeks, although traditional households go more often when their children still live at home. Students and starters visit either once per week or once per month, and the visits of urban citizens vary from once a week to strongly varying. Students and starters and urban citizens have a relatively short membership duration. All other segments have a long membership duration (mostly over 10 years).

Students and starters borrow a few items per visit. The genres they are most interested in are thrillers and current Dutch literature (for more detailed information, see Table 2). This segment is the only segment that frequently borrows other materials than books, namely CDs and DVDs. Urban citizens borrow at almost every visit five materials, which are most often thrillers or current Dutch literature. Retired social seekers borrow three to five materials at almost every visit, but a small part of this segment never borrows anything. Their two most favorite genres are thrillers and recent releases. The remaining segments all mostly borrow three to five materials each time they visit the library. Among them, thrillers are the most popular genre as well, followed by current Dutch literature.

Table 2: Borrowing behavior as found by VOB (2008). Percentages of respondents that borrow each genre.

	Thrillers	Books about holidays and travel	Current Dutch literature	Romantic books	Recent releases	Foreign literature	Classic Dutch literature	Books about psychology	Books about art and culture	Region and family novels	CDs and DVDs
Students and starters	51%	41%	51%		24%	42%	28%	27%	25%		34%
Urban citizens	58%	37%	44%	29%	27%	31%					
Retired social seekers	60%	37%	41%	33%	60%					34%	
Culture lovers	51-63%	52-53%	52-65%		28-29%	30-42%	26-35%			14-24%	
Dynamical households	69%	45%	44%	37%	30%					27%	
Traditional households	64-67%	40-48%	40-45%	32-40%	26-31%					29-34%	
Prosperous enjoyers	65%	48%	47%	37%	31%					30%	

In recent years, the KB has commissioned target group research. They carried out research specifically for the target groups of low literates, people older than 75 years, people older than 55 years, parents, and children younger than 18 years (KB, 2018).

Buisman and Allen (2015) found that 29% of low literate people in the Netherlands visit public libraries. Of these visitors, 61.7% is female, and their age is mostly between 26 and 45 or between 56 and 65 (39.2% and 34.1%, respectively). 55% is native Dutch. This target group mostly uses support and practice services (e.g. courses and trainings) of the library and borrows less books.

Meijering and Duimel (2016) investigated the target group of people over 75 years old. No data about members of this age group was available, therefore they analyzed library members of the age group 65+. In this age group 33% of all Dutch women have a library membership and 32% of all Dutch men have a library membership. Almost all respondents of this research indicated to have a library membership to borrow books. In addition, they use the library as a place to read magazines and newspapers and to be together with other people.

Scholtz and Witte (2016) found that 20% of all Dutch people older than 55 years visit the library once a month. This is more than the Dutch average. Women of this group more often borrow books whereas men more often read something in the library. 28% of this target group is a library member. Stalpers (as cited in Scholtz & Witte, 2016) found that highly educated people in this group read literature, people with mid-level education read thrillers and romantic books are mostly read by lowly educated women.

Witte and van Nood (2014) found that children mostly between the age of 13 and 17 visit the library, often with their parents. 16% of the children borrow more than three books most of their visits, 39% borrow two to three books, 36% borrow one book, and 10% mostly do not borrow any books each time they visit. The most popular genres for people between 13 and 25 years old are thrillers and adventure novels.

2.4 Cluster tasks with similar data in retail settings

Much retail marketing research can be found on creating customer segments by using clustering algorithms. The link between retail and libraries has become really strong, which makes it relevant to investigate these researches. A trend in Dutch public libraries is to use retail strategies for the library, creating so-called ‘retail libraries’ (van Otterlo, 2016). In addition, transactional retail data has many similarities with transactional data of libraries.

In retail customer segmentation, common clustering methods are k-means (Chiu et al., 2009; Khajvand et al., 2011; Peker, Kocyigit, & Eren, 2017; Chen, Agrawal, and Tirupatikumara, 2015), k-medoids (Brito, Soares, Almeida, Monte, & Byvoet, 2015), hierarchical based clustering methods (Hsu et al., 2012; Chen et al., 2015), and neural networks; mainly Self-Organizing maps (Cuadros & Dominguez, 2014; Chiu et al., 2009). Common ways to choose the number of clusters are based on technical measures (such as the Silhouette Validation Method, C index, and isolation index, Hsu et al., 2012), and intuition or practical relevance (Chen et al., 2015).

In terms of feature engineering, one frequently used approach in retail customer segmentation is to categorize users based on RFM (or slightly different) values (Chiu et al., 2009; Cuadros & Domínguez, 2014; Khajvand et al., 2011; Peker et al., 2017). RFM stands for Recency, Frequency and Monetary value and these values are a way to indicate the value of a customer for a company. Hence, this approach categorizes customers into segments based on how valuable they are for the company. This has the practical implementation to focus marketing efforts on customers with the highest value. However, the present research aims to segment users based on their behavior in order to better serve their needs, which is important for marketing efforts of libraries (Garoufallou et al., 2013). Based on RFM values it would not be possible to take user needs into account. This shortcoming is also noticed by Hsu et al. (2012). They also mention that it turns out to be important to not only look at current customer value, but to potential customer value as well. Nevertheless, general attributes, such as RFM values, still cluster customers together that not necessarily have similar purchase behavior. In addition, it reveals no information about customer preferences and information about products.

One study that focuses on clustering customers based on behavioral aspects, is the study by Chen, et al. (2015). Their data consisted of both transactional data over a two-year period (with records on item type, vendor id, customer id, units sold, and time stamp of day and week) and demographic data. The data was transformed by registering the customers’ purchases for each day within the two-year period. After this step of feature engineering, they clustered the data with k-means and with Ward’s hierarchical clustering. K-means did not work well because the data transformation created a quite sparse representation of the data. Ward’s hierarchical clustering did work, and based on sum of squares distance they found 23 clusters. However, for better marketing implementation (it is more effective to focus marketing efforts on larger segments) they decided on five clusters. These consist of four clusters found on a higher hierarchical level with the addition of one cluster for all remaining customers.

2.5 Conclusion

The aim of this study is to find library user segments based on transactional data revealing details about user behavior. As can be concluded from section 2.2, bibliomining has been used to cluster library data and find user segments. However, to my knowledge, no research has used Dutch transactional library data to find customer segments that depend on the customers' behavior, marked by a variety of indicators. The novelty of this research can be found in the type of data and methods used, in combination with the type of segments aimed for.

The results of this research can be used for marketing implementations. It is more efficient to focus marketing efforts on a larger group of users that have similar characteristics in terms of behaviors and needs. For example, if a cluster is found that visits the library on a particular day and borrows a lot of books from one explicit genre, the library can use visuals, especially on the relevant days, to inspire these users to borrow some new releases of that genre or make them aware of relevant upcoming events.

Section 2.3 revealed that users of Dutch public libraries have been studied and segmented in the past. However, these insights can be complemented with insights found with data mining methods, which are known to in some cases provide unexpected results. In addition, in most cases these insights rely on data from questionnaires. Transactional data has the advantage to be more fact-based. These data mining methods can be inspired by previous bibliomining research in the area of clustering (section 2.2) and similar retail research (section 2.4).

3. Methods

3.1 Data

This section describes the data used and how the data was processed in order to make it compatible with the used algorithms.

3.1.1 Data description

The two datasets used for this research were provided by a Dutch public library which is formed by a collaboration between multiple municipalities. The datasets contain data of all the locations within these municipalities. The first dataset consists of transactional log data about the actions of library users, gathered under a unique user ID. The second dataset consists of data about the libraries' collections. Both datasets are provided in multiple CSV (comma-separated values) files.

The log dataset contains data about library users over a 5-year period (01-01-2014 until 01-01-2019). In total, it consists of 12,049,409 rows, corresponding to user actions of 65,200 unique user IDs. This dataset consists of 29 columns, with information about the action date and time, membership type, title number of the borrowed material, and the location where the action took place (for all columns, see Appendix A). The collection dataset consists of data about the libraries' collections. It contains information such as title, author, publishing year, and genre (all columns can be found in Appendix B). This dataset consists of 108 columns and 827,494 rows and was exported in February 2019. Both datasets do not contain personal information, in terms that users cannot be identified based on these two datasets.

3.1.2 Preprocessing and data cleaning

In order to make the provided data compatible with the clustering algorithm, first some preprocessing and data cleaning have been done. First the 259 separate CSV files of the log dataset were combined, which resulted in one CSV file with all of the log data. Then the 62 CSV files of the collection dataset were combined which resulted in another CSV file with all of the collection data. It turned out that, in most of the collection files, the last row of one file was equal to the first row of the following file. Therefore, all duplicate rows were deleted from the combined dataset. Thereafter, all columns which were not likely to contribute to the final features or to the data analysis (section 3.1.3) were deleted for both sets. Those columns were selected with the features of related studies in mind. This resulted in a reduction of 18 rows for the log dataset and 95 rows for the collection dataset. Furthermore, five additional columns were deleted from the collection dataset, based on their missing value percentages which were above 50%. Ultimately, the log and collection datasets were merged based on title ID. It turned out that for 21,755 rows there was a title ID in the log dataset that did not appear in the collection dataset. Therefore, these rows have been deleted, which resulted in a reduction of 19 users. The columns of which the final dataset consists, together with their final missing value percentages are shown in Table 3.

3.1.2.1 Missing values

For two columns, the library involved had indicated which value belonged to cells with missing values. This applied to the columns containing material type and original language, in case the material type is book. The column containing material type had a missing value percentage of 47.24%. All of the missing values were filled with 'Book'. The column containing the original language of books had a missing value percentage of 89.42%. The missing values in this column that belonged to a book were filled with 'Dutch'. After this missing values treatment, and combining the two datasets, this column had a missing value percentage of 7.32%.

Table 3: Features in the final dataset together with their missing value percentages.

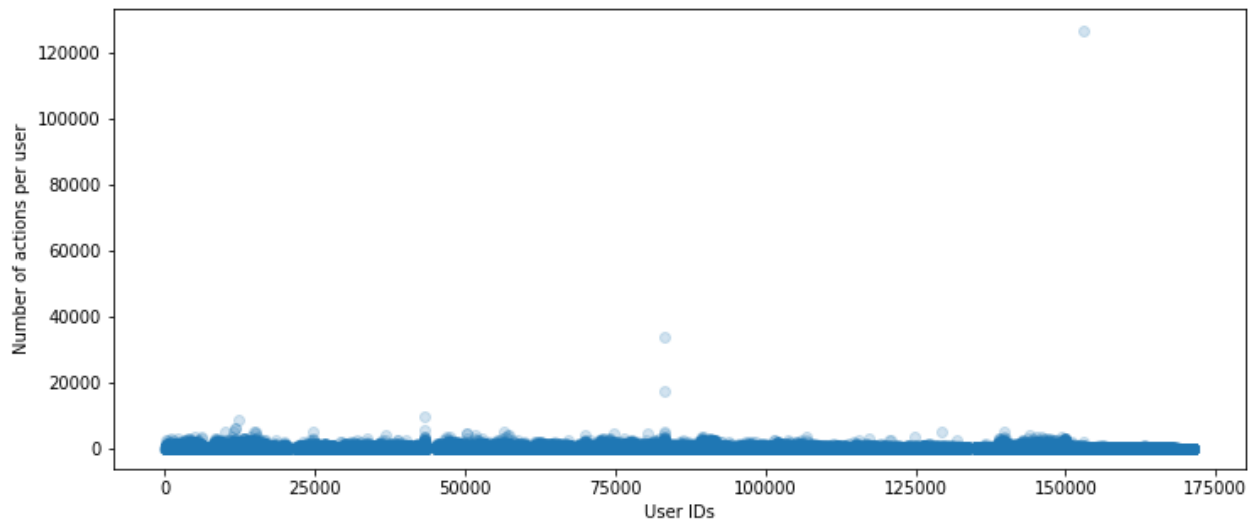
Column name	Description	Missing value percentage
aktie_vest	Location where the action took place.	0%
aktie_datum	Date of the action.	0%
aktie_tijd	Time of the action.	0%
aktie	Type of action (e.g. borrowing, handing in, and buying).	0%
bron	Source of the action's registration (e.g. via the website or at a libraries' counter).	0.0017%
titelnr	Title ID for each collection item (there can be two book copy's, both copies will have the same title ID). Originally both datasets contained this column, which was used to merge both datasets.	0%
scatnr	Codes representing material types and their genres.	0.0040%
leen_dat	Date on which the material was borrowed.	0.0265%
inl_dat	Date on which the material was returned.	0%
actor_id	Unique user ID.	0%
abnr	Membership type.	0%
titel	Title of the material.	0%

ama	Material type.	0%
taal	Language of the material.	3.7267%
leeftijd	Recommended age of the target group in case the material is a book.	45.0632%
pubjaar	Publication date.	0.5561%
oors_taal	Original language of the material.	7.3201%
auteur	Author.	0.0803%

3.1.2.2 Outliers

When looking at the number of actions for each unique user, a couple of users stand out, as can be seen in Figure 1. After consulting with the library it was clarified that some of these users are not natural persons, but user IDs used by the system. It became clear that all users with the membership types 2, 8, 97, and 99 should not be concerned natural library users. Therefore, all users with those membership types were left out of consideration. Over the total dataset, this resulted a reduction of 418 users and 410,735 user actions.

Figure 1: Total number of actions per unique library user from 2014 to 2018.



3.1.3 Data analysis

The log dataset contains data about library members and library users without a membership. Non-members can be users of library services such as printing or internet access. Non-members do have an account linked to a unique user ID in order to access these services. The dataset consists of 75.74% members and 24.26% non-members. Members are considered to be unique users who have a subscription type that is described on the library's website. Considering user actions, 72.97% of the actions are done by members and 27.03% are done by non-members. Subscription type also offers insights about the distribution between people up to 17 years old and people older than 17. In the dataset, 33,045 users have a subscription type for people until 17 years, which is 51.02%.

When looking at each of the years separately, one thing that stands out is that the unique user IDs decreased each year, and the actions decreased since 2015 (as visualized in Figures 2 and 3).

Figure 2: Reduction of library users from 2014 to 2018.

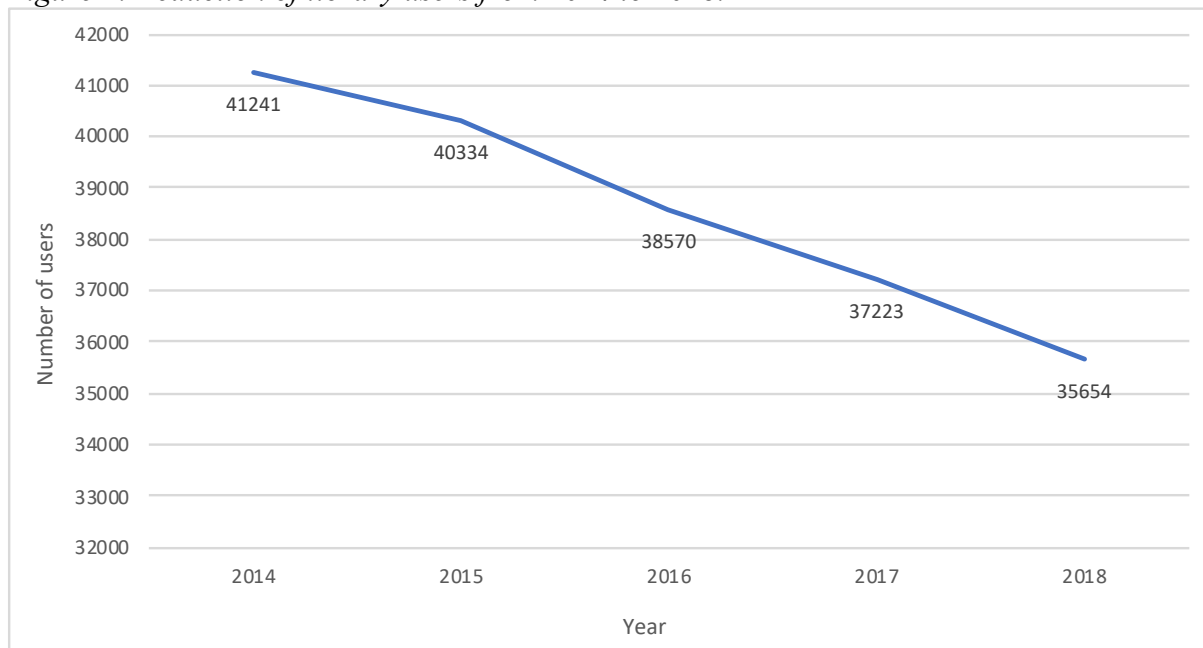
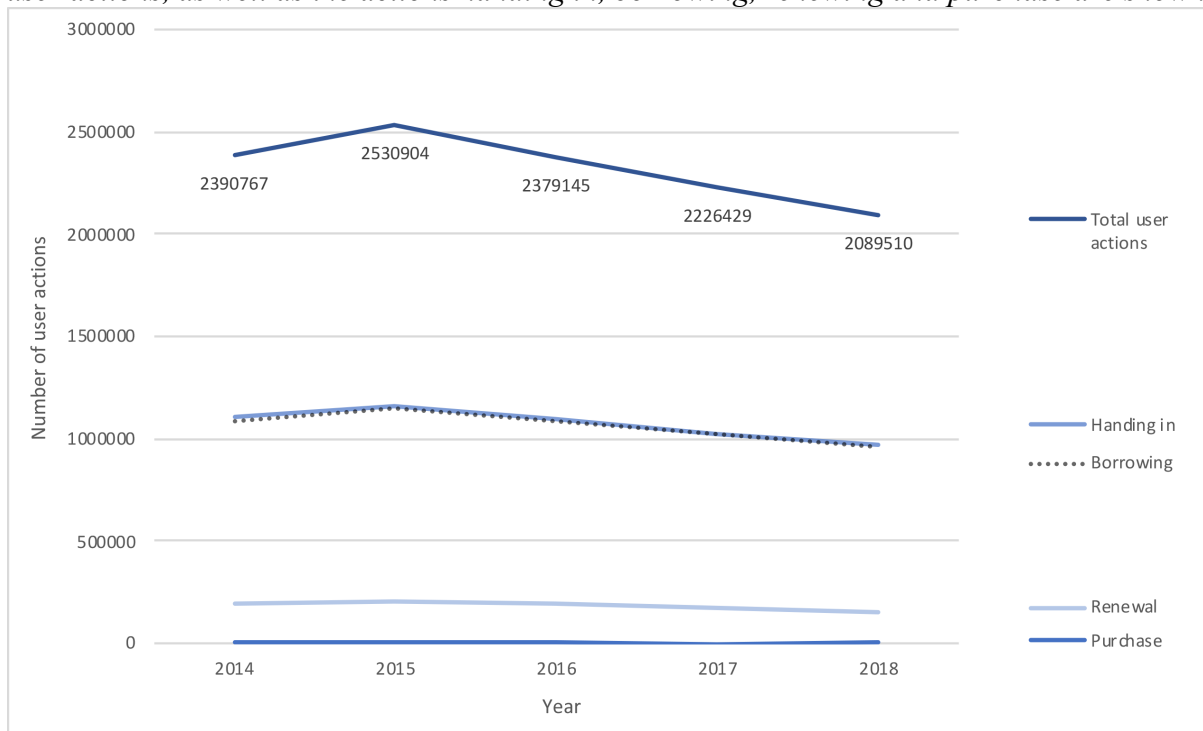
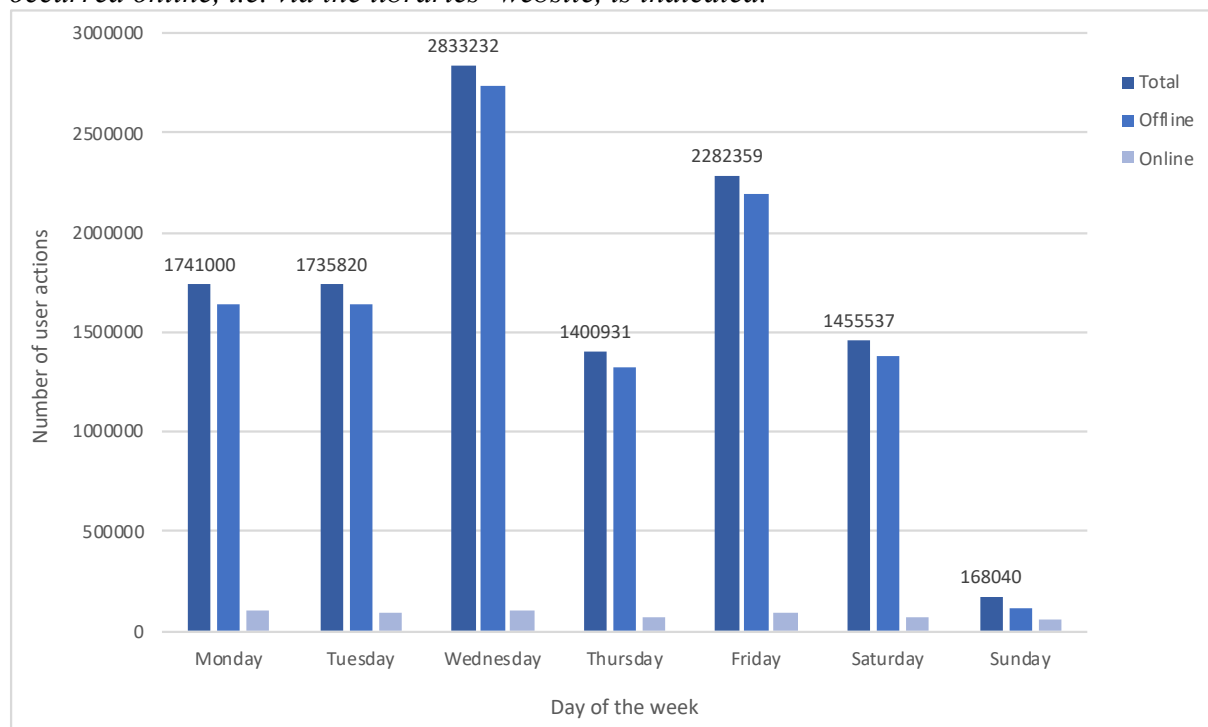


Figure 3: Change in number of library users' actions from 2014 to 2018. The total number of user actions, as well as the actions handing in, borrowing, renewing and purchase are shown.



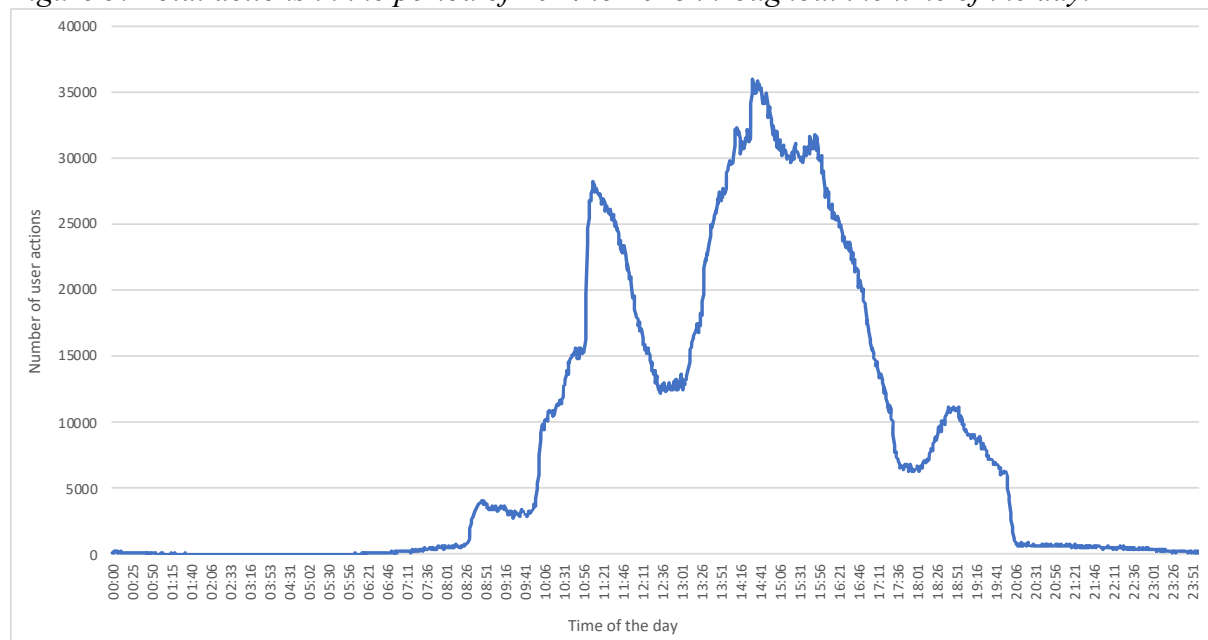
When analyzing the day of the week on which actions take place, it can be concluded that most user actions occur on Wednesdays and least actions occur on Sundays, as can be seen in Figure 4. This can probably be explained by the fact that few library locations open on Sundays. When looking at each year separately, there is only one instance where the distribution differs from the distribution as provided in Figure 4, namely in 2015, when there occurred more actions on Thursdays than on Saturdays.

Figure 4: Total user actions for each day of the week over the total period of 2014 to 2018. Additionally, the division of actions occurred offline, i.e. in a physical library, and actions occurred online, i.e. via the libraries' website, is indicated.



In addition to the days of the week, the time of the day on which actions occur is analyzed. As can be seen in Figure 5, clear moments can be distinguished on which many actions occur.

Figure 5: Total actions in the period of 2014 to 2018 throughout the time of the day.



When analyzing the behavior of users in terms of the locations they visit, it turns out that there is a subset of users who visit more than one location. Possible reasons could be locations being closed for a period, or the trade-off between a location nearby home and a location with a larger collection. On average, users visit 1.31 locations. 76.44% of the users visit only one location, 18.52% of the users visit two locations, and 3.70% visit 3 locations or more.

3.1.4 Feature engineering

For the creation of the features which will be used to enter into the final model, the choice has been made to only consider actions which occurred in the year 2018. This subset has been chosen because a user's behavior might change severely during a five-year period. When these actions are summarized in order to create features, it might be that these summaries do not reflect the user's behavior very well. All of the years in the dataset did not show specific differences. Therefore the most recent year was chosen.

In 2018, 2,089,510 user actions of 35,654 users took place. With these data the following features were created for each individual user; total number of visits for each day of the week, total number of actions, number of different locations visited, total number of actions for each moment of the day, average number of actions per visit, total number of actions per material type/genre, and total number of actions online and offline. This resulted in a feature matrix of 44 columns and 35,654 rows, corresponding to the number of users. The creation of some of the features is discussed more detailed below.

For the creation of the **number of visits for each day of the week**, datetime from the Python Standard Library, was used. The days of the week on which an action occurred were extracted from the action date. For each user we counted how many times they visited the library per day of the week, where all actions occurring on the same date were considered as one visit.

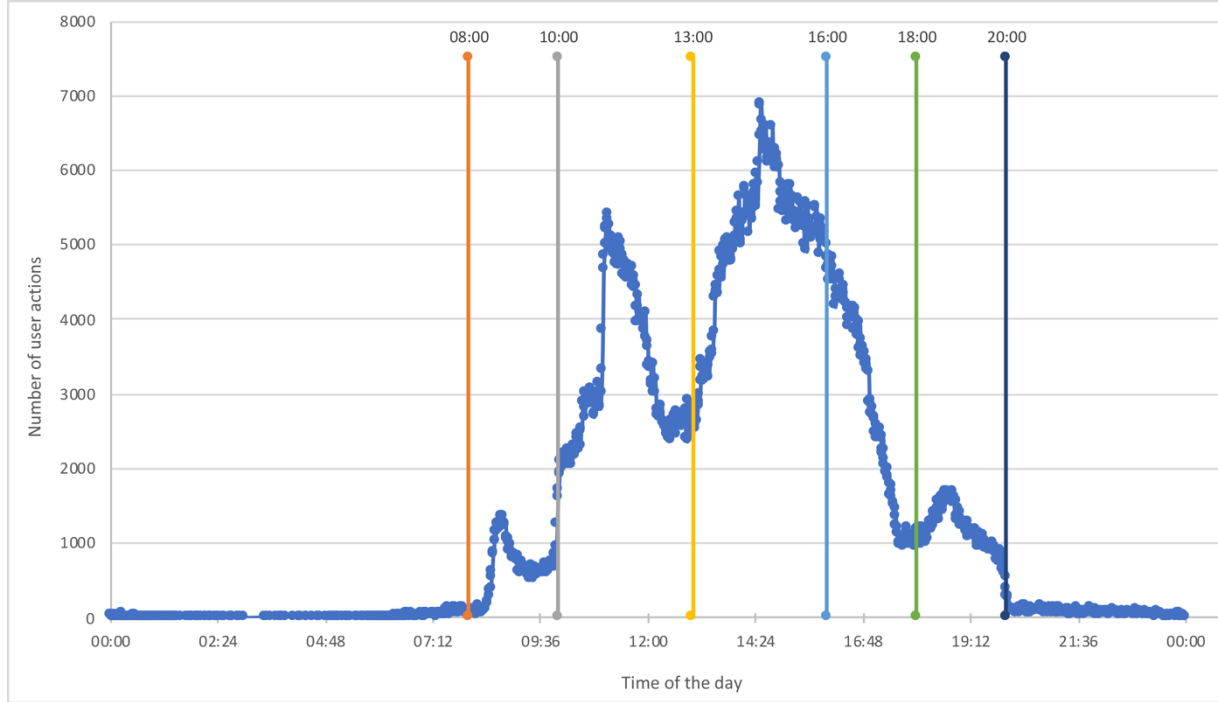
As described in section 3.1.3, a part of the users makes use of more than one library location. In 2018 this is also true; 1.17 locations were visited on average per user. Therefore the **number of different locations visited** per user was included as a feature. No difference has been made between locations that were visited more or less often. For example, if a user visits one library location 100 times, and another library once, the count for that user will be two.

As described in section 3.1.3, there is a clear distribution of actions throughout times of a day. For the feature describing the **total number of actions for each moment of the day**, we can divide a day into six intuitive periods, as can be seen in Figure 6. The following periods were selected; 20:00 until 08:00 is considered as night, 08:00 until 10:00 as early morning, 10:00 until 13:00 as morning rush, 13:00 until 16:00 as afternoon rush, 16:00 until 18:00 as afternoon drop, and 18:00 until 20:00 as evening. It is worth mentioning that different library locations have different opening times, this has not been accounted for in making this division.

For the feature **total number of actions online and offline**, the values of the column 'bron' were assigned to either the category offline, when the action occurred in a physical library, or online when the action occurred via the internet. In Appendix C the allocation of the categories is shown. The code L was hard to assign, as this stands for actions occurring in school libraries, which are connected to the library's system. Hence the actions are not executed via the internet, but neither in one of the library locations. It was chosen to assign these instances to the category offline, because the setting shares more similarities with actions occurring in a library than with actions occurring via the library's website.

The feature considering **total number of actions per material type/genre** was created by using the columns 'ama', which contains material type, and 'scatnr' which contains the genre of the material. In the original dataset ama contained 96 different values. These have been allocated, as far as the following material types remained; book, large lettered book, magazine, CD/DVD, playing materials, and other (Appendix D shows how the values were allocated). In this process the number of occurrences per material type was taken into account, such that material types that occurred very often were more likely not to be merged.

Figure 6: Total actions in 2018 throughout the time of the day and the breakdown into 6 moments.



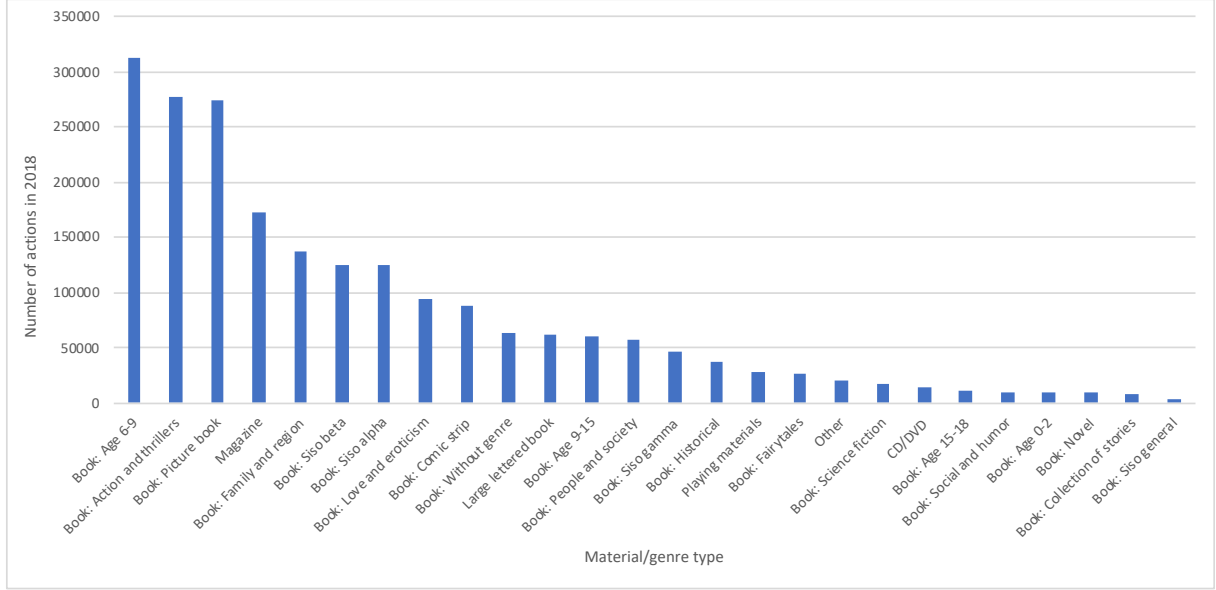
Thereafter, the different genres were designed into features. First, for all rows where the ama value was book, the genre code as provided in the column scatnr was concatenated. The dataset included 2,163 different genre codes. It turned out that 909 of these were unknown. These have been changed to 'without genre'. Within the genre codes, there are also codes for material types outside of books, for example Jazz for CDs. In 34 cases the genre code indicated that the material type was not a book. This might be due to the missing value treatment described in section 3.1.2.1. For these cases the value was assigned to the material type indicated in the genre code. For each reading level, a different genre code exists, leading there to be a lot of codes which indicate the same genre but on a different reading level. All overlapping genre categories were combined, letting go of the reading level. At this point, 420 codes were dedicated to SISO-categories. SISO is a classification system used by Dutch public libraries for non-fiction books with 19 main categories (Wikipedia, 2014). The 420 codes were reduced to those 19 categories. Thereafter, the SISO-categories were merged further to 'siso general', 'siso alpha', 'siso beta', and 'siso gamma', based on their scientific topic. The remaining genre codes at this point were further combined as shown in Appendix E. The process of narrowing down categories has been done with consultation of expert knowledge. After all processing steps, 26 material type/genre categories remained, which are shown in Figure 7.

After all features were created, this resulted in a feature matrix $X = (x_{ij})$, with $i = 1, \dots, 35,654$ rows representing individual users and $j = 1, \dots, 44$ columns representing the features described above. All feature values were normalized in order to make them weigh equal in the model. The normalization has been done using MinMaxScaler from the preprocessing library from Scikit-learn. The used formula is:

$$x_{ij \text{ new}} = \frac{(x_{ij} - \min(x_j))}{\max(x_j) - \min(x_j)}$$

where x_j is a column of the feature matrix and x_{ij} is a column value.

Figure 7: Number of actions associated with different material types/genres.



3.2 Algorithm

For the creation of clusters, the k-means algorithm (MacQueen, 1967) has been used. This partitioning algorithm divides all samples (i.e. users in this situation) into k different clusters, where k is a predetermined number. It does so in the following manner:

- (1) k arbitrary initial cluster centers (centroids) are set;
- (2) all samples ($x_i = 1$ to 35,654) are assigned to the cluster whose centroid they are closest to, based on a distance metric;
- (3) for each cluster, the mean μ_p is calculated and set as the new centroid;
- (4) steps 2 and 3 are repeated until no data points change.

In this process, the k-means algorithm minimizes the within cluster variance, expressed in distance d , for each cluster. This minimization problem can be represented as:

$$\arg \min_c \sum_{p=1}^k \sum_{x_i \in c_p} d(x_i, \mu_p) = \arg \min_c \sum_{p=1}^k \sum_{x_i \in c_p} \|x_i - \mu_p\|^2$$

where x_i is one sample (i.e. user) in cluster p , μ_p is the mean of the cluster, and c are all clusters.

The distance metric used in this research is the Euclidean distance, which calculates the distance d between points u and v . The formula used is as follows:

$$d(u, v) = \sqrt{\sum_{i=1}^n (u_i - v_i)^2}$$

where i is one dimension and n are the total number of dimensions. This distance metric is influenced to a higher degree by large values as compared to smaller values. Therefore the choice was made to normalize the feature values.

Because the initial cluster centroids are arbitrarily chosen, the algorithm might not find the global minimum. In order to prevent this, the initialization parameter of the Scikit-learn k-means algorithm has been set to kmeans++. Instead of choosing completely random initial

centroids, this chooses initial centroids that are far away from each other making it less likely for the algorithm to converge on a local minimum (Igual & Seguí, 2017).

In order to determine the number of clusters, the elbow method is taken into consideration. This method looks at the within cluster sum of squares for different values of k , which expresses the variability of a cluster. The number of clusters to choose, according to this method, is the one from where the sum of squares decreases to a lesser extent than for the number of clusters before. However, the elbow method was not followed blindly, as human interpretability and practical implementation of the clusters are an important goal of this research.

First the algorithm was run with the feature matrix containing actual values, for $k=3$ up to and including $k=9$. Then the same was done for the feature matrix containing normalized values.

The clusters were evaluated in a technical manner using the Silhouette Validation Method. This was implemented using the metrics library from Scikit-learn. The Silhouette score for a cluster distribution is calculated by taking the average of the Silhouette coefficients for all samples. A sample's silhouette coefficient s is calculated as follows:

$$s = \frac{b-a}{\max(a,b)}$$

where a is the distance between a sample and the centroid of its cluster and b is the distance between a sample and the centroid of its nearest cluster. The Silhouette score returns a score between -1 and 1. Scores towards -1 indicate that a different cluster is more similar for a sample, values towards 1 indicate that a sample is similar to the cluster where it is assigned to.

Additionally, the sizes of the clusters and the mean and standard deviation within each cluster were examined. Besides that, the clusters were evaluated on a qualitative level as well. This was done by performing a data analysis of the clusters and examine their properties as compared to similar customer segments described in the literature and by evaluating the clusters with experts from the field.

3.3 Software

The software used for the analysis of this research is Python version 3. In Table 4 an overview is given of the libraries used.

Table 4: Used libraries in Python.

Library	Specifics	Version
Pandas		0.23.4
Numpy		1.15.1
Datetime		3
Scikit-learn	Preprocessing K-means Metrics	0.19.2
Matplotlib	Pyplot	2.2.3
SciPy	Spatial distance	1.1.0

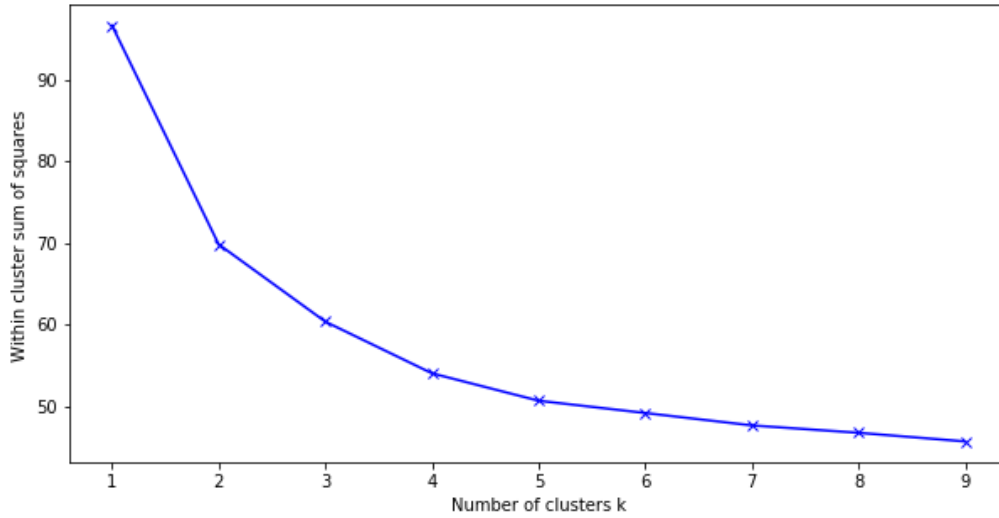
4. Results

In this section, the results of the k-means analysis are presented. As described in section 3.2, first k-means was applied to the feature matrix with unnormalized values, and then to the matrix with normalized values. This was done for $k=3$ up to $k=9$.

4.1 Feature matrix with unnormalized values

For the feature matrix with unnormalized values, it was chosen to start with $k=3$ because the elbow method indicated that to be the optimal number of clusters (Figure 8).

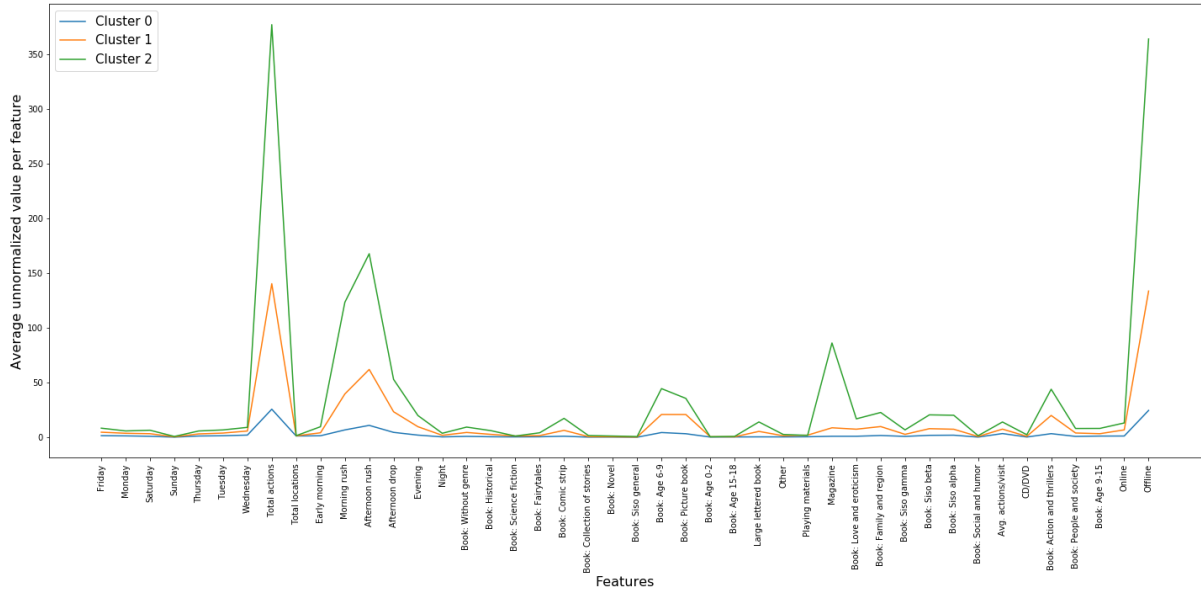
Figure 8: Elbow method showing the within cluster sum of squares for each number of k for the feature matrix with unnormalized values.



After applying k-means with seven different numbers for k , the following results were found. With $k=3$, the clusters were divided in a way that for most features cluster 2 had the highest average values, cluster 0 had the lowest average values and cluster 1 scored in between, as can be seen in Figure 9. For example, the averages for total actions are $M_{cluster0} = 25.64$, $M_{cluster1} = 140.31$, and $M_{cluster2} = 377.04$. This division is only different for the feature considering ‘playing material’ ($M_{cluster0} = 0.44$, $M_{cluster1} = 1.91$, and $M_{cluster2} = 1.71$) and ‘Book: Age 0-2’ ($M_{cluster0} = 0.19$, $M_{cluster1} = 0.47$, and $M_{cluster2} = 0.46$).

For each higher number of k , the division changes on an increasing number of features (for $k=4$ it changes for 3 features, $k=5$ for 8 features, $k=6$ for 17 features). The features where these shifts mostly take place are features considering material type/genre. With $k=6$ such a shift occurs for the first time for another type of feature, namely average number of actions per visit ($M_{cluster0} = 2.89$, $M_{cluster1} = 5.00$, $M_{cluster2} = 7.24$, $M_{cluster3} = 10.14$, $M_{cluster4} = 16.71$, $M_{cluster5} = 12.18$). Up until $k=6$, cluster 0 always has the lowest average for all features, only from $k=7$ that starts to shift. Also, from $k=7$ more than half of the features are distributed in different ways, meaning that the distribution is not dominated anymore by highest to lowest averages. All cluster distributions and their sizes can be found in Appendix F. Considering the sizes of the clusters, especially for a higher number of k some clusters are added that consist of a very small part of the users. For example, at $k=7$ two of the clusters consist of less than 0.60% of the users.

Figure 9: Results for $k=3$ with unnormalized feature matrix.



4.2 Feature matrix with normalized values

For the feature matrix with normalized values, the elbow method also indicated $k=3$ to be the optimal number of clusters (Figure 10). Therefore, k-means was applied as well seven times, with values for k from 3 to 9.

For $k=3$ the results were comparable to those for the unnormalized feature matrix; all clusters had either the highest, lowest, or middle average for all features except for ‘playing material’ ($M_{\text{cluster}0} = 0.45$, $M_{\text{cluster}1} = 1.51$, and $M_{\text{cluster}2} = 1.19$) and ‘Book: age 0-2’ ($M_{\text{cluster}0} = 0.21$, $M_{\text{cluster}1} = 0.36$, and $M_{\text{cluster}2} = 0.34$), as shown in Figure 11. However, from $k=4$ on this division is completely gone, as can be seen in Figures G2 to G7 in Appendix G. From these figures it can be concluded that influential features are days of the weeks, total actions and average actions per visit, and the material type/genre borrowed.

Figure 10: Elbow method showing the within cluster sum of squares for each number of k for the feature matrix with normalized values.

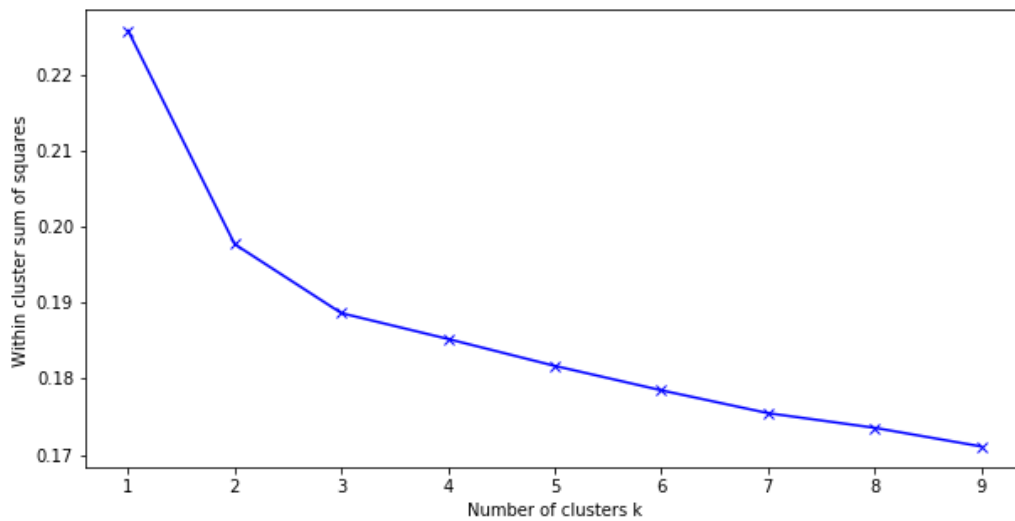
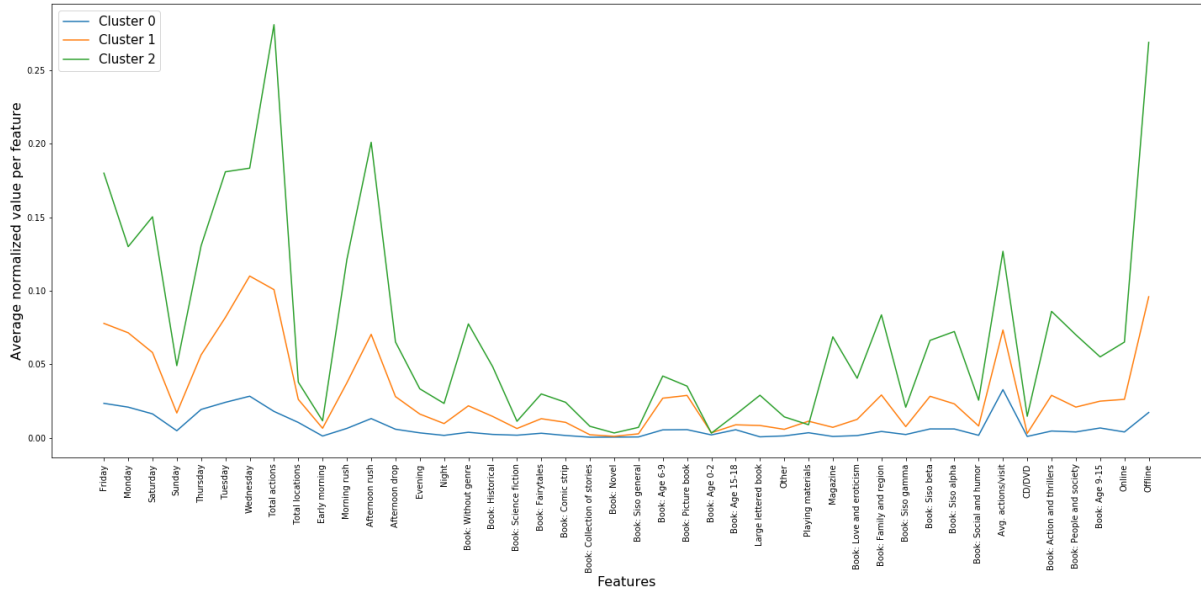
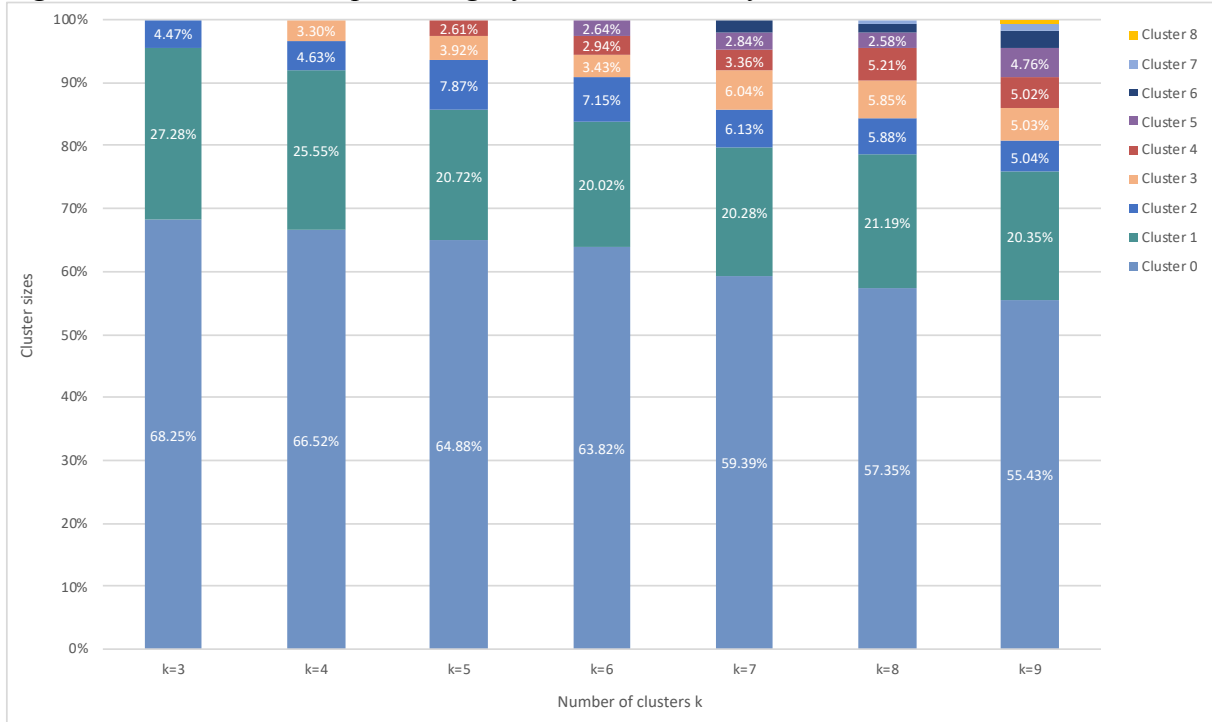


Figure 11: Results for $k=3$ with normalized feature matrix.



Considering the cluster sizes, for all numbers of k there is one cluster consisting of more than half of all users, one cluster which is little below the half of the first cluster's size, and all other clusters only representing less than 8% the users, as can be seen in Figure 12.

Figure 12: Cluster sizes in percentages for each number of k .



To evaluate the technical ‘correctness’ of the clusters, the Silhouette score was calculated for each cluster distribution. Figure 13 shows that the Silhouette score gets lower for a higher number of k , indicating that the technical ‘correctness’ gets worse when creating more clusters. This could have been expected, as the elbow method already indicated $k=3$ to be the optimal number of clusters. Based on the Silhouette scores, and an analysis of the cluster distributions as provided in Appendix G, Figures G1 to G7, the choice was made to consider six clusters as the optimal number of clusters. This was chosen because the cluster

distribution in terms of their averages already show interesting divisions, giving room for human interpretability. This can also be seen by comparing Figures 11 and 13. Moreover, the clusters still have to some extent a ‘good’ Silhouette score. As can be seen in Figure 13, the Silhouette score drops most severely from $k=6$ to $k=7$. In addition to technical measures and human interpretability, cluster size is taken into account when choosing for six clusters. From $k=7$ on, clusters are added that consist of less than 2% of the users, which is considered to be too small to be of value.

Figure 13: Silhouette score for each number of k .

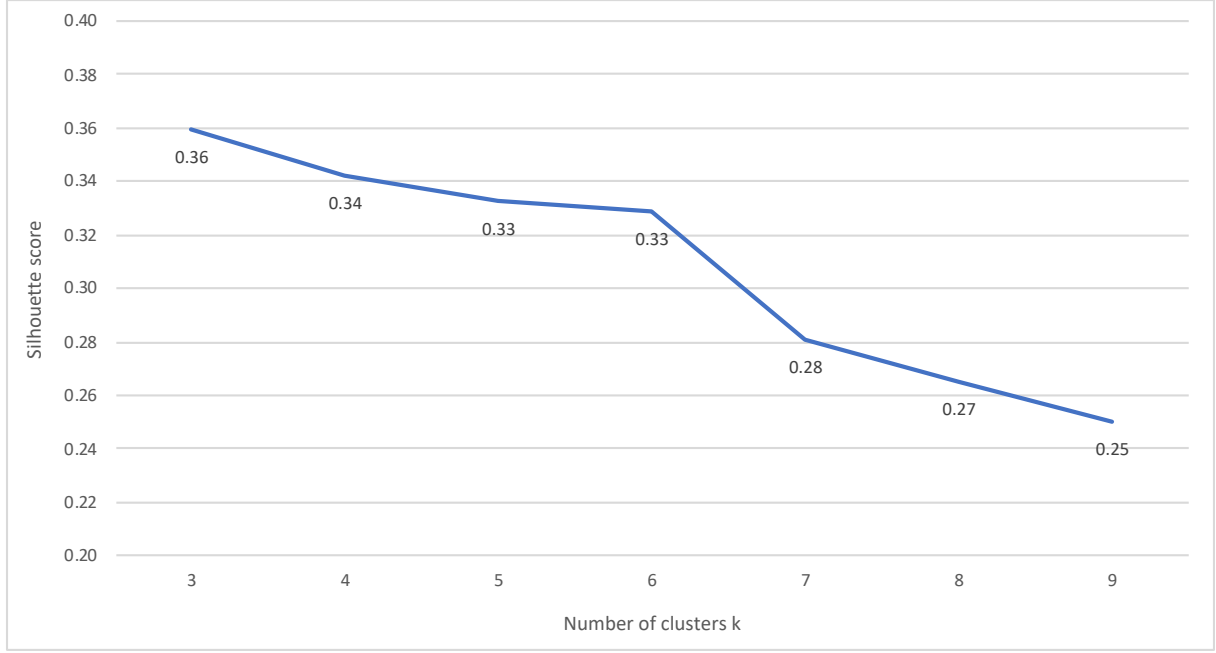
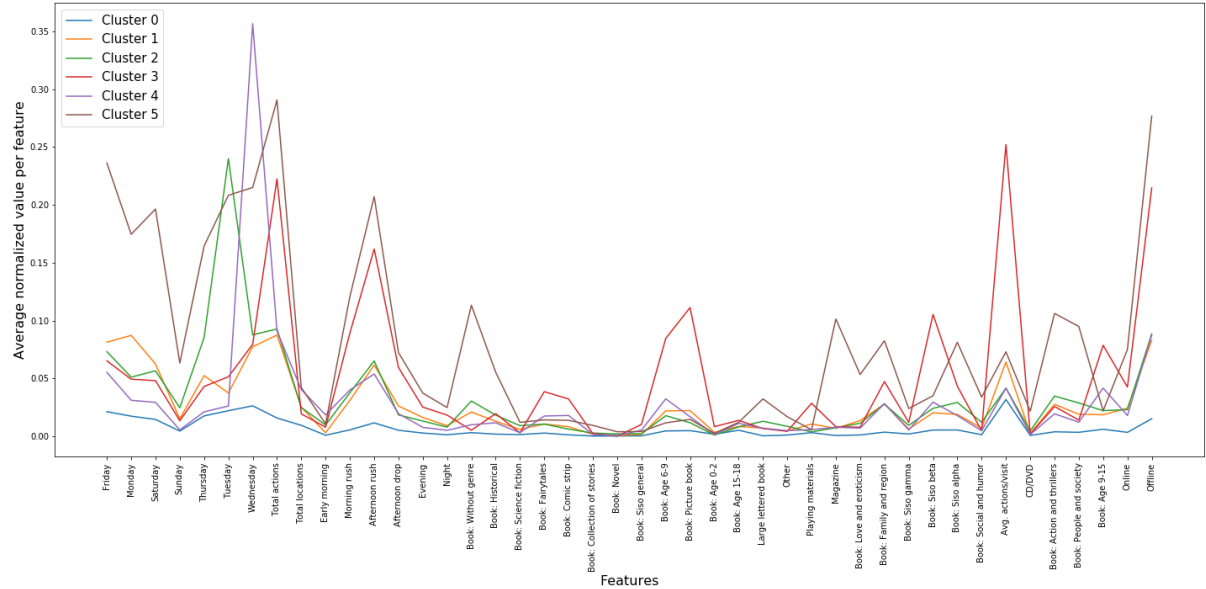


Figure 14: Results for $k=6$ with normalized feature matrix.



In Figure 14 the cluster distribution is shown, based on the averages of the normalized values. All actual averages per feature per cluster and the corresponding standard deviations are provided in Appendix H. When analyzing the cluster distribution based on Figure 14 and the actual cluster means, the following results are found.

All clusters visit most often in the afternoon rush. There is no strong division based on moment of the day, meaning that no clusters behave very different for certain aspects of this feature. For all clusters most actions occur offline, also for this feature there is no strong division. Lastly, there does not appear to be a strong division for different locations visited. The averages of the feature are between 1.10 ($SD = 0.34$) for cluster 0, which generally has low averages, and 1.46 ($SD = 0.88$) for cluster 5, which generally has high averages.

Cluster 0 consists of 63.82% of the users which makes it the biggest cluster. On almost all features this cluster has the lowest number of actions. The average total number of actions for this cluster is 18.50 ($SD = 15.82$). Only for the feature 'Book: Novels' it has the second to lowest average ($M = 0.14$, $SD = 1.29$), and for 'Books: Age 0-2' it has the third to lowest average ($M = 0.20$, $SD = 1.70$).

Cluster 1 is the second largest cluster consisting of 20.02% of the users. This cluster has the second highest average for visits on Mondays ($M = 4.10$, $SD = 4.46$), Fridays ($M = 4.07$, $SD = 3.63$), and Saturdays ($M = 2.81$, $SD = 3.30$) as compared to other clusters. This cluster has the second least total actions ($M = 97.49$, $SD = 40.48$). On average, users in this cluster borrow as third most books for the age 6 to 9 ($M = 14.99$, $SD = 27.28$) and second most picture books ($M = 13.00$, $SD = 27.95$). They have the third most average actions per visit ($M = 5.66$, $SD = 2.97$).

Cluster 2 consists of 7.15% of the users. This cluster has the highest average of visits on Tuesdays ($M = 10.31$, $SD = 4.49$). It has the second highest average for books without genre ($M = 4.51$, $SD = 8.69$), books about action and thrillers ($M = 18.28$, $SD = 31.63$), and large lettered books ($M = 5.76$, $SD = 28.76$). It has the second to lowest average actions per visit ($M = 4.02$, $SD = 2.21$) and second to highest average for the material types other ($M = 1.69$, $SD = 7.34$) and CD/DVD ($M = 0.92$, $SD = 4.81$).

Cluster 3 consists of 3.43% of the users. It has the second highest average for total user actions ($M = 246.35$, $SD = 98.82$). Considering material type/genre, it has the second highest average for historical books ($M = 2.69$, $SD = 6.60$), and the highest average for books about fairytales ($M = 4.25$, $SD = 9.54$), picture books ($M = 64.97$, $SD = 82.20$), comic strip books ($M = 16.21$, $SD = 37.26$), non-fiction beta books ($M = 25.07$, $SD = 34.08$), and books for the age 6 to 9 ($M = 57.70$, $SD = 80.59$) and 0 to 2 ($M = 0.91$, $SD = 5.78$). It also has the highest average actions per visit ($M = 19.30$, $SD = 11.78$).

Cluster 4 and 5 are comparable in size; they consist of 2.94% and 2.64% of the users respectively. Cluster 4 has the highest average for visits on Wednesdays ($M = 18.18$, $SD = 6.66$). For all other days it has the second lowest average. It has the second highest average for books about fairytales ($M = 1.93$, $SD = 5.25$), comic strip books ($M = 9.06$, $SD = 21.74$), and books for the age 6 to 9 ($M = 22.09$, $SD = 33.65$) and 9 to 15 ($M = 5.17$, $SD = 8.85$). This clusters' average actions per visit is 4.03 ($SD = 2.30$).

Cluster 5 is the smallest cluster. It has the highest average of total actions ($M = 321.47$, $SD = 144.13$). This makes it logical that it scores high on all of the days of the week as well. It has the highest average for all days of the week except for Tuesday and Wednesday. Within the cluster, Friday ($M = 11.80$, $SD = 8.48$) has the highest average followed by Wednesday ($M = 10.96$, $SD = 7.49$). In terms of material type/genre this cluster has the highest average value for books without genre ($M = 16.77$, $SD = 21.15$), books about history ($M = 7.59$, $SD = 13.95$), science fiction ($M = 1.77$, $SD = 8.97$), love and eroticism ($M = 22.86$, $SD = 44.61$), family and region ($M = 21.80$, $SD = 38.78$), people and society ($M = 13.29$, $SD = 19.40$), sociality and humor ($M = 2.24$, $SD = 4.40$), action and thriller ($M = 55.75$, $SD = 75.17$),

collections of stories ($M = 2.01$, $SD = 5.14$), novels ($M = 1.53$, $SD = 14.57$), non-fiction alpha books ($M = 20.88$, $SD = 33.49$), non-fiction gamma books ($M = 7.11$, $SD = 18.10$), large lettered books ($M = 14.29$, $SD = 56.68$), magazines ($M = 88.02$, $SD = 146.99$), CDs and DVDs ($M = 4.25$, $SD = 17.63$), and other material types ($M = 3.31$, $SD = 10.52$). It has low values for children's books as compared to other clusters, except for the age 15 to 18 ($M = 0.55$, $SD = 2.40$).

5. Discussion

This section interprets the results presented in section 4, links them to literature described in section 2, and answers the questions raised in the introduction.

5.1 Research questions and results

The goal of this research was to create user segments of Dutch public libraries, based on behavioral log data. These segments can mainly be used for marketing implementations. In the introduction, the following research questions were stated:

1. To what extent can meaningful user segments be found in behavioral log data in terms of moment, place and nature of the users' actions?

1.1 Which features should be entered into the model?

1.2 Which typical characteristic features can be distinguished within the segments?

In order to find the user segments, 44 features describing the users' behavior were created. The features include; visits per day of the week, total number of actions, number of different locations visited, actions per moment of the day, average number of actions per visit, actions per material type/genre, and actions occurring online or offline. In order to find user segments, the data was clustered using k-means.

First k-means was applied to a feature matrix with the actual values for each user (more information about the feature engineering can be found in section 3.1.4). However, this mainly led to clusters divided based on highest to lowest averages for almost all features, which is not very useful for marketing purposes. This was caused by the different scales of the features. Within the model, the distance metric used was the Euclidean distance, which was heavily influenced by large feature values. Therefore k-means was applied again to a feature matrix with normalized values.

From the results of the feature matrix with normalized values, user segments can be distinguished based on feature properties, when four or more clusters are created. As described in section 4.2, three clusters yield the best results based on technical measures. However, in terms of human interpretability and as input for marketing decisions, the results of three clusters are undesirable because they show no feature-based division. The higher the number of clusters, the more characteristics in terms of features appear. However, with a high number of clusters, some really small clusters are added, which again is undesirable for marketing purposes. It is more effective to focus marketing efforts on larger segments (Chen et al., 2015). When taking into consideration the technical measures for the clusters, a trade-off between those three things appears to exist at six clusters. Which is why in this study was chosen to create six clusters.

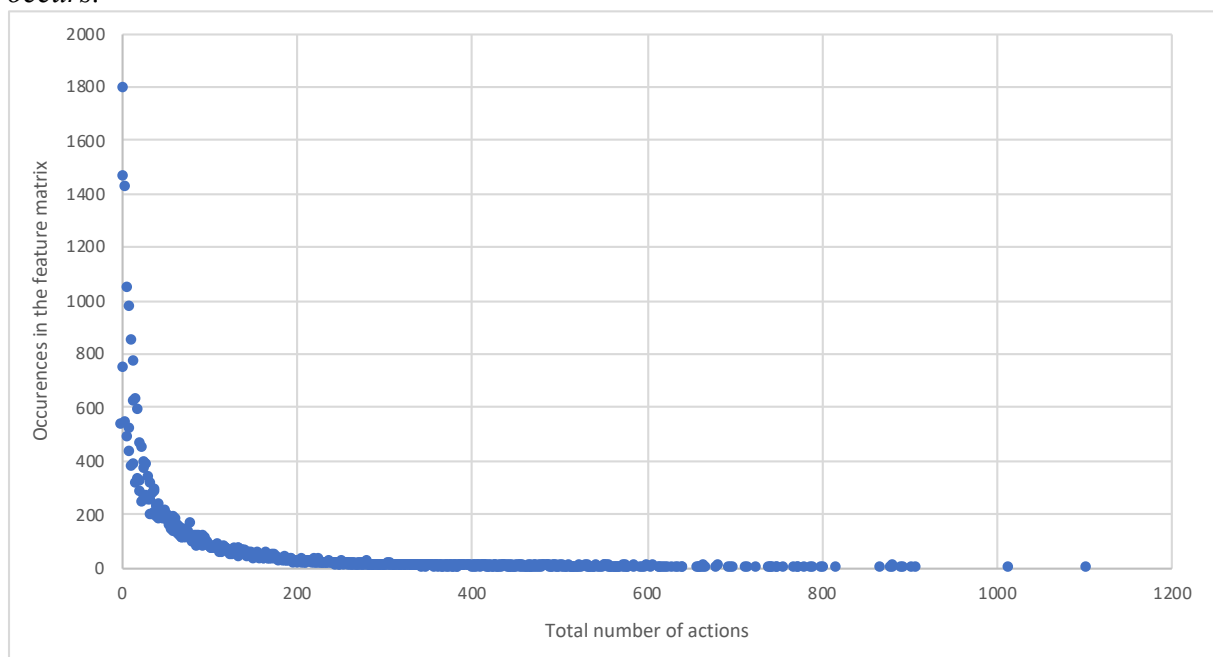
As can be seen in the results presented in section 5, much feature averages have a high standard deviation as compared to their means. This indicates high variances on many features within the clusters. This was to be expected as the silhouette scores did not show very ‘good’ results.

When analyzing the six segments, the following things stand out. Cluster 0 is the largest segment, covering over half of the users. For all but two features, this segment has the lowest average, indicating that users of this segment perform few actions. We call this segment “passive users”.

Cluster 1, the second largest segment, has a low average for most features as well. When looking at total number of actions, it scores second lowest. However, as compared to other segments, users in this segment visit the library relatively much on Mondays, Fridays, and Saturdays. For none of the features this segment stands out in having a highest or lowest average. Therefore we call this segment “average users”.

The two largest segments consist of users with few user actions. When looking at all users in the feature matrix, there are a lot of users with a low number of actions and little users with a high number of actions, as visualized in Figure 15.

Figure 15: On the x-axis all different values that appear in the feature matrix for the feature ‘total actions’ is shown, on the y-axis is shown for how many users in the matrix this value occurs.



Considering cluster 2, it stands out that users in this segment visit the library in particular often on Tuesdays. For that reason, we call this segment “Tuesday visitors”. As compared to other segments, Tuesday visitors have a relatively low average of actions per visit. They have a relatively high average of actions considering books about action and thrillers, and large lettered books. This segment has the second highest average for the material types other and CDs/DVDs. The combination of relatively low average actions per visit and preference in actions and thrillers and non-book materials is similar to the segment students and starters, found by the VOB (2008). However, the VOB (2008) found that students and starters are the only segment interested in these material types, while in these results cluster 5 has a much

higher average for those features. Moreover, the VOB (2008) did not report a clear day preference for students and starters.

Cluster 3 has by far the highest average actions per visit. This segment represents users who on average perform almost 20 actions per visit. In terms of types of materials borrowed, this segment seems to borrow both children's books, a variety of genres and some non-fiction books. It might be that some users in this segment borrow both materials for themselves and for their children, or that this cluster consists of both children and adults. However, only 22% of the users in this segment have a membership type for children. Therefore we carefully call this segment "users with children". Within this cluster, the most popular day to visit the library is Wednesday.

The most typical property of users in cluster 4 is that they visit the library mainly on Wednesday. This characteristic stands out both compared to this segment's averages for other days of the week, as to the other segments' averages for Wednesday. We call this segment "Wednesday visitors". Popular books for this segment are books about fairytales, comic strips, and books for the age 6 to 15. When looking at the averages within this segment, genres that in addition stand out are magazines, books about family and region, and action and thrillers. In the Netherlands, primary schools have the afternoon off on Wednesdays. This segment's preference in books and day of the week might indicate that these users are children. This segment consists for 42% of users with a membership for children from 0 to 12 years and for 30% of users with the membership type adult basic. Hence, it appears that for a large part this segment does consist of children. It might be that the adults in this dataset are parents that visit the library with their children. This could be related to the finding of Witte and van Nood (2014), stating that children of age 13 to 17 visit the library often with their parents. Moreover, Witte and van Nood (2014) found that children between the age of 13 and 25 have a preference for books about adventures and thrillers. In this study, those genres are covered in the category 'action and thrillers', which also is represented in this cluster's results. Finally, Witte and van Nood (2014) found that 39% of the children borrow two to three books each time they visit. The average actions per visit for this segment is 4.03.

It stands out for cluster 5 that it has the most total actions. In particular this segment visits the library often on Fridays and Saturdays. This differs from the VOB's (2008) results. There the segment retired social seekers visits the library most often, however, they visit especially during the week. Cluster 5 has the highest average for a lot of the features, which can logically be explained by the fact that this segment consists of users responsible for many actions. We call this cluster "active users". Jin and He (2017) clustered user data of a Chinese academic library into three clusters, based on the average amount of books borrowed. In their study, the smallest cluster also was responsible for the highest number of books borrowed. Within the active user cluster, high averages in terms of materials borrowed can be distinguished for magazines, books about family and region, action and thrillers, people and society, and non-fiction books in the area of alpha. Material types in particular associated with children score respectively low for this cluster. However, when looking at the membership types within this segment, it consists for a large part of users with a membership type for the age 0 to 12. The segment dynamical households, as found by the VOB (2008) also show a high average for a lot of genres. The two genres they borrowed most are thrillers and romantic books. Dynamical households also indicate that they like Saturday and Sunday openings, which comes close to the days of the week on which this segment goes to the library as well.

As described in section 2.2, Boter and Wedel (2005) used similar data as this study uses to create user segments. However, they only used features considering the genres borrowed. They found four clusters, consisting of users mainly borrowing literary fiction, detective and

thriller novels, American romantic fiction, or Dutch romantic fiction. The results of the clusters in this study do not show a similar distinction; “active users” have the highest average for both the categories love and eroticism, and action and thrillers. This could be caused by the variety of other features which were entered into the model. Another difference exists in the way the categories of the genres were created. Boter and Wedel did not consider nonfiction and children’s books, while this study did.

When looking at the segment sizes, it stands out that the users are unevenly divided over the segments. This is different for the segments found by the VOB (2008). However, when compared to other library user segmentation researches which used data science methods, uneven segment sizes are not unusual. For example, Lai and Zeng (2014) found three user segments with very dissimilar sizes, namely 73.5%, 24.3%, and 2.2%. Moreover, Hájek and Stejskal, found 11 segments, seven of which consisted of less than 10% of the users.

As can be seen in the results presented in section 5 and in Appendix H, much feature averages have a high standard deviation as compared to their means. This indicates high variances on a lot of features within a cluster. This was to be expected as the silhouette scores were not very high either.

With respect to the raised research questions, the following things can be concluded. Considering all of the above, it seems that in this research’s setup, the typical characteristic features per segment are mainly defined by user behavior in terms of the amount of actions performed, the days of the week on which actions occur, and the material type/genre associated with the actions. Whether the action occurred online or offline, on which moment of the day the action took place, and the number of different locations visited do not seem to provide characteristics for the segments, as all segments behave comparable considering these features.

The fact that all segments visit the library mostly in the afternoon corresponds to the results as presented in the customer is queen (VOB, 2008). In addition, the most popular genre found by the VOB (2008) was thriller. In this study, the genre category containing thrillers also has high averages for a number of segments. Some similarities between the results of the VOB (2008) and this study are found, but there are differences as well. The VOB (2008) did not consider users younger than 18 years, while this study did. This causes differences in results. Another dissimilarity is that the VOB (2008) considered a wider variety of features. For example, they also took demographical information into account.

With respect to the main research question, it can be concluded that to a certain extent meaningful user segments are found. As described above, the segments can be distinguished based on behavioral properties, although for some segments these are clearer than for others. Moreover, it seems that the cluster distribution is still influenced by highest to lowest average, although this does not apply to all segments or to all features. It appears that the more meaningful properties a cluster has, the smaller part of the users it covers. This might be due to fact that all features are based on the amount of user actions, and there are severe differences between the number of actions per user.

5.2 Limitations and future work

It can be concluded that with the feature matrix with normalized variables, some meaningful user segments are found. However, because of the way in which the features were created, the feature matrix was quite sparse. As pointed out by Chen et al. (2015), k-means is not very good in clustering sparse data. According to them, hierarchical clustering is more appropriate

for this type of feature matrix. However, due to time limitations, this method was not explored in this study, and thus it becomes a main recommendation for future research.

As described above, there are some segments which have strong feature-based characteristics and some which behavior is captured in terms being average for all features. This might be due to the differences in number of user actions. There are some users with many actions, and many users with only a couple of actions. This might be the cause of the larger clusters not showing strong feature-based characteristics. Therefore it might be useful to treat users with few actions different than users with many actions. It might be that when separately clustered, the smaller differences of the users with few actions get picked up more easily by the clustering algorithm.

In the process of creating the features considering material type and genre, there were a lot of choices made in order to reduce the categories. However, after consulting with people from the field not all choices appeared equally satisfactory. An improvement can be made in the reclassification of all material types/genres. For example, in this study it was chosen to let go of most of the reading levels, future research could investigate whether clusters are divided differently when reading levels become a more prominent part of this feature.

6. Conclusion

The goal of this thesis was to find user segments for Dutch public libraries based on behavioral log data. These can mainly be used for marketing implications, but might be useful to prove libraries' value as well. This research's novelty can be found in the data and methods used, in combination with the segments aimed for. The research questions stated are:

1. To what extent can meaningful user segments be found in behavioral log data in terms of moment, place and nature of the users' actions?

1.1 Which features should be entered into the model?

1.2 Which typical characteristic features can be distinguished within the segments?

The dataset used in this study was from a Dutch public library. After cleaning the dataset, the data from the year 2018 was used to investigate further. From this data 44 features capturing users' behavior were created, in terms of; visits per day of the week, total number of actions, number of different locations visited, actions per moment of the day, average number of actions per visit, actions per material type/genre, and actions occurring online or offline. The feature values were normalized before applying the k-means algorithm. Considering various indicators, six was chosen as the optimal number of clusters. The segments distinguish themselves mainly based on characteristics considering the amount of user actions, the day of the week on which the library is visited and the borrowed material type/genre. With respect to the main research question it can be concluded that to a certain extent there can be found meaningful user segments based on their behavior. These segments are described as "passive users", "average users", "Tuesday visitors", "users with children", "Wednesday visitors", and "active users".

Although the segments show feature-based characteristics, these characteristics are clearer for some clusters than for others. For some clusters it seems that they are characterized by average values throughout the features. This might be due to the sparse feature matrix in combination with the k-means algorithm. Therefore, a recommendation for future research is to investigate hierarchical cluster algorithms, which seem to better deal with sparse data. Other recommendations are to differently treat users with few and many actions, and to further investigate different ways in which to categorize material types and genres.

References

- Al-Daihani, S., & Abrahams, A. (2016). A text mining analysis of academic libraries' tweets. *The Journal of Academic Librarianship*, 42(2), 135-143. doi:10.1016/j.acalib.2015.12.014
- Al-Daihani, S., & Abrahams, A. (2018). Analysis of academic libraries' facebook posts: Text and data analytics. *The Journal of Academic Librarianship*, 44(2), 216-225. doi:10.1016/j.acalib.2018.02.004
- Ball, R. (2019). Big Data and Their Impact on Libraries. *American Journal of Information Science and Technology*, 3(1), 1-9. doi: 10.11648/j.ajist.20190301.11
- Boter, J., & Wedel, M. (2005). User categorization of public library collections. *Library and Information Science Research*, 27(2), 190-202. doi:10.1016/j.lisr.2005.01.004
- Brito, P., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer Integrated Manufacturing*, 36, 93-100. doi:10.1016/j.rcim.2014.12.014
- Buisman, M., & Allen, S. (2015). *Laaggeletterden en de bibliotheek*. Gorinchem/Den Haag: Kohnstamm Instituut en Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA), commissioned by KB | Nationale Bibliotheek.
- Central Bureau for Statistics (2018). *Statistics Netherlands: Public libraries* [Data file]. Last updated on 2018, November 23. Retrieved on 26-03-2019 from <https://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=70763NED&D1=0-1%2c19%2c37-39%2c46-49%2c60%2c74&D2=8-1&VW=T>
- Chen, C. B., Agrawal, D., & Tirupatikumara, S. R. (2015). Retail analytics: Market segmentation through transaction data. In *IIE Annual Conference and Expo 2015* (pp. 1034-1042). Institute of Industrial Engineers.
- Chiu, C., Chen, Y., Kuo, I., & Ku, H. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications: Part 1*, 36(3), 4558-4565. doi:10.1016/j.eswa.2008.05.029
- Cuadros, A., & Domínguez, V. (2014). Customer segmentation model based on value generation for marketing strategies formulation. *Estudios Gerenciales*, 30(130), 25-30. doi:10.1016/j.estger.2014.02.005
- de Jonge, E. (2018). *The mission of the Dutch public library* (Master's thesis, Leiden University). Retrieved from <https://openaccess.leidenuniv.nl/handle/1887/64622>
- Eden, B. L. (2015). Leading the 21st-century academic library: Successful strategies for envisioning and realizing preferred futures (Vol. 1). Rowman & Littlefield.
- Field, N., & Tran, R. (2018). Reinventing the public value of libraries. *Public Library Quarterly*, 37(2), 113-126. doi:10.1080/01616846.2017.1422174
- Futterman, M., Milam, D. P. (March 2016). Core Customer Intelligence: Public Library Reach, Relevance and Resilience. CivicTechnologies. Retrieved on 24-04-2019 from [civictechnologies.com: http://civictechnologies.com/core-customer-intelligence](http://civictechnologies.com/core-customer-intelligence)
- Garoufallou, E., Siatiri, R., Zafeiriou, G., & Balampanidou, E. (2013). The use of marketing concepts in library services: A literature review. *Library Review*, 62(4), 312-334.
- Hájek, P., & Stejskal, J. (2014). Library user behavior analysis—use in economics and management. *WSEAS Transactions on Business and Economics*, 11(1), 107-116.
- Hsu, F., Lu, L., & Lin, C. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, 39(6), 6221-6228. doi:10.1016/j.eswa.2011.12.005
- Igual, L., & Seguí, S. (2017). *Introduction to data science: A python approach to concepts, techniques and applications* (Undergraduate topics in computer science). Cham, Switzerland: Springer. doi:10.1007/978-3-319-50017-1

- Jin, W., & He, S., (2017). Research on Management of Libraries in Universities and Colleges based on K-means Clustering Algorithm under Big Data Environment. *Revista de la Facultad de Ingeniería*, 32(8), 177-181.
- Johannsen, C. C. (2014). Understanding users: from man-made typologies to computer-generated clusters. *New Library World*, 115(9-10), 412-425.
- Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57-63. doi:10.1016/j.procs.2010.12.011
- Knott, M. L., Bradley, D. R., DeGeorge, D. S., & Ottaviani, J. (2007). Catalog information and user expectations in an amazoogole world: Too much? too little? *Against the Grain*, 19(5), 40-44. doi:10.7771/2380-176X.5277
- Koninklijke Bibliotheek (KB). (2018, November 29). *Inzicht in bibliotheekgebruikers*. Retrieved from www.bibliotheekinzicht.nl: <https://www.bibliotheekinzicht.nl/organisatie-en-bedrijfsvoering/ondernemerschap-en-maatschappelijke-waarde/inzicht>
- Lai, Y., & Zeng, J. (2014). Analysis of customer churn behavior in digital libraries. *Program*, 48(4), 370-382. doi:10.1108/PROG-08-2011-0035
- Litsey, R., & Mauldin, W. (2018). Knowing what the patron wants: using predictive analytics to transform library decision making. *The Journal of Academic Librarianship*, 44(1), 140-144.
- MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.
- Malschaert, J. (2016). *Making knowledge / Connecting people*. Paper presented at: IFLA WLIC 2016 – Columbus, OH – Connections. Collaboration. Community in Session 213 - Metropolitan Libraries. Retrieved from <http://library.ifla.org/1546/>
- Massis, B. E. (2012). Using predictive analytics in the library. *New Library World*, 113(9/10), 491-494. doi:10.1108/03074801211273957
- Massis, B. E. (2016). Data-driven decision-making in the library. *New Library World*, 117(1/2), 131-134. doi:10.1108/NLW-10-2015-0081
- Massis, B. E. (2018). Artificial intelligence arrives in the library. *Information and Learning Science*, 119(7/8), 456-459. doi:10.1108/ILS-02-2018-0011
- Meijering, I., & M. Duimel, (2016). *Ouderen en de bibliotheek. Een kwalitatief onderzoek onder 75-plussers*. Den Haag: Koninklijke Bibliotheek.
- Nicholson, S., & Stanton, J. M. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. *Organizational data mining: Leveraging enterprise data resources for optimal performance*. Hershey, PA: Idea Group Publishing Retrieved on 24-04-2019 from <http://hdl.handle.net/10150/106383>
- Niu, X., & Hemminger, B. (2015). Analyzing the interaction patterns in a faceted search interface. *Journal of the Association for Information Science and Technology*, 66(5), 1030-1047. doi:10.1002/asi.23227
- Palmer, S. (2014). Characterizing university library use of social media: A case study of twitter and facebook from australia. *The Journal of Academic Librarianship*, 40(6), 611-619. doi:10.1016/j.acalib.2014.08.007
- Polger, M. A. (2019). *Library Marketing Basics*. Rowman & Littlefield Publishers.
- Peker, S., Kocyigit, A., & Eren, P. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence & Planning*, 35(4), 544-559. doi:10.1108/MIP-11-2016-0210
- Scholtz, A. & Witte, E. (2016). *Boekenbranche meting 36. 1e themameting van 2016 naar het leesgedrag van 55-plussers*. Amsterdam: SMB.
- Showers, B. (Ed.). (2015). *Library analytics and metrics: Using data to drive decisions and services*. London, England: Facet Publishing.

- Silva, D. V., & Miguéis, V. L. (2017). Combining Data Analytics with Layout Improvement Heuristics to Improve Libraries' Service Quality. *Exploring Services Science*, 223–234. doi:10.1007/978-3-319-56925-3_18
- Tempelman-Kluit, N., & Pearce, A. (2014). Invoking the user from data to design. *College & Research Libraries*, 75(5), 616-640. doi:10.5860/crl.75.5.616
- Tiessen-Raaphorst, A., & van den Broek, A. (2016). *Sport en cultuur. Patronen in belangstelling en beoefening*. Den Haag: Sociaal en Cultureel Planbureau. Retrieved from www.scp.nl:
https://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2016/Sport_en_cultuur
- Tsuji, K., Yoshikane, F., Sato, S., & Itsumura, H. (2014, August). Book recommendation using machine learning methods based on library loan records and bibliographic information. In *2014 IIAI 3rd International Conference on Advanced Applied Informatics*(pp. 76-79). IEEE. doi:10.1109/IIAI-AAI.2014.26
- Ullah, A., & Ameen, K. (2018). Account of methodologies and methods applied in LIS research: A systematic review. *Library and Information Science Research*, 40(1), 53-60. doi:10.1016/j.lisr.2018.03.002
- Vakkari, P., Aabø, S., Audunson, R., Huysmans, F., & Oomes, M. (2014). Perceived outcomes of public libraries in Finland, Norway and the Netherlands. *Journal of Documentation*, 70(5), 927-944. doi:10.1108/JD-06-2013-0072
- van Otterlo, M. (2016). Project BLIIPS: Making the Physical Public Library more Intelligent through Artificial Intelligence. *Qualitative and Quantitative Methods in Libraries*, 5(2), 287-300.
- van Otterlo, M., & Warnaar, M. (2017). Towards Optimizing the Public Library: Indoor Localization in Semi-Open Spaces and Beyond. Benelearn
- Vereniging Openbare Bibliotheken (VOB) (2005). Richtlijn voor basisbibliotheken. Retrieved on 29-04-2019 from www.debibliotheken.nl:
http://www.debibliotheken.nl/fileadmin/documenten/branche/2005_richtlijn-voor-basisbibliotheken_vob_vng.pdf
- Vereniging Openbare Bibliotheken (VOB) (2008). De klant is koningin. Retrieved on 11-02-2019 from www.debibliotheken.nl:
http://www.debibliotheken.nl/fileadmin/documenten/vereniging/2008_de-klant-is-koningin_handboek_openbare_bibliotheken.pdf
- Witte, E., & van Nood, B. (2014). *Boekenbranche meting 26. Herhaalmeting van de 2e themameting van 2013, naar het kopen, lezen en lenen van boeken onder jongeren tussen de 13 en 25 jaar*. Amsterdam: SMB.
- Wikipedia. (2014). *SISO*. Retrieved from www.wikipedia.org:
<https://nl.wikipedia.org/wiki/SISO>
- Yi, K., Chen, T., & Cong, G. (2017). Library personalized recommendation service method based on improved association rules. *Library Hi Tech*, 36(3), 443-457. doi:10.1108/LHT-06-2017-0120
- Zickuhr, K., Rainie, H., Purcell, K., & Pew Internet & American Life Project. (2013). *Library services in the digital age*. Washington, D.C.: Pew Research Center's Internet & American Life Project. (2013). Retrieved on 19-03-2019 from <https://files.eric.ed.gov/fulltext/ED539071.pdf>

Appendix A: Original columns of the transactional log dataset

loghfd_id
aktie_vest

aktie_datum
aktie_tijd
aktie
punt
bron
aktie_bits
exem_id
titelnr
eigenaar
oorsprong
scatnr
materiaal
kast
col_id
leen_dat
inl_dat
originldat
actor_id
stat_code
wijk_buurtcode
instantie_id
abnr
thuis_vest
adm_nr
hoofd_vest
plaatsing
poort_id

Appendix B: Original columns of the collection dataset

titelnr	leesnivo	sysbits
sleutel	matvolume	mediumbits
titbits	leverbaar	nivobits
titstat	sanadv1	taalbits
presoort	sanadv2	vestbits0
titel	titl_resv	vestbits1
ama	oors_taal	vestbits2
titel2	land_uitg	vestbits3
deelaand	picastamp	updated
deeltitel	auteur	tt_info2
taal	reeks	vestbits5
genre_rub	siso	vestbits7
materiaal	momkeys	vestbits4
materiaalj	blok_publ	vestbits6
leeftijd	bapop	bacultaut
clib	baoordeel	bacult
ppn	banivo	onderttl
titsoort	nbdmat	bameu
type6	bainit	vestbits9
bron	kijkwijzer	vestbits11
type8	navivan	vestbits10
winstdat	navitm	vestbits8
prijs	banbd	rdacarrier
serienr	kijkwreden	rdacontent
pubjaar	cwisenr	class94
pubjaartm	pim	class93
hwoord	tt_info	class92
mutdat	mak_id	expected
plaatsing	mak_loc	
media_srt	mut_id	
scatnr	ewstatus	
freq	ewcat1	
sortpos	ewcat2	
status	ewlftvan	
pid	ewlfttm	
regio_code	ewavivan	
scatj	ewavitm	
ex_koppel	ewleesbel1	
nbd_cat	ewleesbel2	
bijlage	ewgender	

Appendix C: Allocation of codes in the column ‘bron’

Original value	Explanation	Allocated value
Z	Zelfbediening	Offline
S	Zelfservice (Sip2)	Offline
B	Balie/Wiseclient	Offline
E	Extern/Boekenbus	Offline
I	Internet	Online
J	WiseApp	Online
N	Noodverwerking	Offline
X	Extra	Offline
D	Dubbele uitleentermijn	Offline
P	Pubcat	Online
L	Leeslog (schoolWise)	Offline
C	Leeslog	Offline

Appendix D: Allocation of codes in the column ‘ama’

Original value	Allocated value
+ CD	CD/DVD
+CD	CD/DVD
+CD'	CD/DVD
BLADMUZIEK	Other
Bladmuziek	Other
BLU RAY	CD/DVD
BLU-RAY	CD/DVD
BOEK	Book
BOEK + CD	Book
Boek + CD	Book
BOEK + CD-ROM	Book
BOEK + CD'S	Book
BOEK + DVD	Book
Boek + DVD	Book
BOEK + GELUIDSCASSETTE	Book
Boek + Geluidscassette	Book
BOEK + GELUIDSOPNAME	Book
Bookstart	Book
CD	CD/DVD
CD-ROM	CD/DVD
CD-Rom	CD/DVD
Compact disc	CD/DVD
DAISY-ROM	Other
Daisy-rom	Other
DUIT	Other
DVD	CD/DVD
DVD (INFO)	CD/DVD
DVD DVD	CD/DVD
DVD INFO	CD/DVD
DVD MUZIEK	CD/DVD
DVD SPEELFILM	CD/DVD
DVD VIDEO	CD/DVD
DVD video	CD/DVD
DVD-MUZIEK	CD/DVD
DVD-ROM	CD/DVD
DVD-VIDEO	CD/DVD
DVD-Video	CD/DVD
DVD-Video + CD	CD/DVD
dwarsligger	Book
Dwarsligger	Book
E-reader	Other
eBookstick	Other
ENGE	Other
Grootletter boek	Large lettered book
GROOTLETTERBOEK	Large lettered book
grootletterboek	Large lettered book
GROTE LETTER	Large lettered book

HARDLOPER	Other
Hardloper	Other
HARDLOPERS	Other
Hardlopers	Other
IBL	Other
IFL	Other
LUISTERBOEK	Other
Luisterboek	Other
Luisterboek op CD	Other
Luisterboek op MP3-CD	Other
MP3	CD/DVD
MUZIEK-DVD	CD/DVD
NED	Other
NINTENDO WII	Other
PLAYSTATION 2	Other
PLAYSTATION 3	Other
Playstation 3	Other
Playstation 4	Other
PUZZEL	Playing materials
SPEELGOED	Playing materials
SPEELLEERMATERIAAL	Playing materials
SPEELLEERMATERIAAL	Playing materials
Speelleermateriaal	Playing materials
SPEELLEERMATERIALEN	Playing materials
Speelleermaterialen	Playing materials
SPEELLEERMATERIALEN	Playing materials
SPEL/PUZZEL	Playing materials
Spraakmaker	Other
SPRAAKMAKERS	Other
Spraakmakers	Other
STERBOEK	Book
TAALCURSUS	Other
Tijdschrift	Magazine
TIJDSCHRIFT	Magazine
Tijdschrift J	Magazine
VE-krat	Other
verk.	Other
Vertelplaten	Other
VIDEO	CD/DVD
Wepboek	Other
WEPBOEK	Other
WII	Other
WIIU	Other
XBOX	Other
Xbox	Other
Xbox 360	Other
XBOX 360	Other
Xbox One	Other
Z&B	Other

Appendix E: Allocation of genre codes (originally from the column ‘scatnr’)

Original value	Allocated value
avonturen	Action and thrillers
biografie	People and society
bijbels verhaal	People and society
detective	Action and thrillers
dierenleven	Family and region
doktersverhaal	People and society
erotisch	Love and eroticism
familie	Family and region
feministisch verhaal	People and society
griezelverhaal	Action and thrillers
historisch	Historical
humor	Social and humor
kerstverhaal	People and society
kinderleven	Family and region
oorlog	Action and thrillers
paasverhaal	People and society
politiek	People and society
paarden en pony's	Family and region
protestants milieu	People and society
psychologisch verhaal	People and society
rassenvraagstuk	People and society
romantisch	Love and eroticism
rooms-katholiek milieu	People and society
streekverhaal	Family and region
schoolverhaal	Family and region
science fiction	Science fiction
sinterklaas	People and society
sprookjes	Fairytales
sociaal verhaal	Social and humor
politiek verhaal	People and society
spionage	Action and thrillers
sport	People and society
stripverhaal	Comic strip
thriller	Action and thrillers
verhalenbundel	Collection of stories
verzet	Action and thrillers
western	Action and thrillers
zeeverhaal	Action and thrillers
islamitisch	People and society
pinksterverhaal	People and society
arabische roman	Novel
bosnische roman	Novel

chinese roman	Novel
duitse roman	Novel
engelse roman	Novel
franse roman	Novel
friese roman	Novel
griekse roman	Novel
kroatische roman	Novel
poolse roman	Novel
perzische roman	Novel
portugese roman	Novel
russische roman	Novel
servische roman	Novel
spaanse roman	Novel
turkse roman	Novel
vietnamese roman	Novel
italiaanse roman	Novel
overige talen roman	Novel
siso 000 algemeen	Siso general
siso 100 wijsbegeerte	Siso gamma
siso 200 godsdienst	Siso gamma
siso 300 - 330 sociale wetenschappen	Siso gamma
siso 340 - 380 economie	Siso gamma
siso 390 - 398 rechtswetenschap	Siso gamma
siso 399 krijgswetenschap	Siso gamma
siso 410 - 420 psychologie - occulte wetenschappen	Siso gamma
siso 430 - 490 opvoeding - onderwijs - vorming	Siso gamma
siso 500 wiskunde - informatica - natuurwetenschappen	Siso beta
siso 600 - 610 geneeskunde - gezondheidszorg - lichaams oefening en sport	Siso beta
siso 620 bord-, denk - en kansspellen enz. - handenarbeid enz. - huishoudkunde enz.	Siso alpha
siso 630 land- en tuinbouw - jacht en visserij	Siso beta
siso 640 - 690 techniek	Siso beta
siso 700 - 770 kunst; algemeen - beeldende kunst	Siso alpha
siso 780 - 790 muziek - dans - toneel - film	Siso alpha
siso 800 taal- en letterkunde	Siso alpha
siso 900 - 940 geschiedenis	Siso alpha
siso 950 - 990 land- en volkenkunde	Siso alpha
A-boek	Age 6-9
Prentenboek	Picture book
Boekstart	Age 0-2
B-boek	Age 9-15
C-boek	Age 9-15
Jeugd (D)	Age 15-18

Appendix F: Cluster sizes and distributions for each number of k for the unnormalized feature matrix

Table F1: Cluster sizes in percentages for each number of k .

$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	Cluster
68.25%	66.52%	64.88%	63.82%	59.39%	57.35%	55.43%	0
27.28%	25.55%	20.72%	20.02%	20.28%	21.19%	20.35%	1
4.47%	4.63%	7.87%	7.15%	6.13%	5.88%	5.04%	2
	3.30%	3.92%	3.43%	6.04%	5.85%	5.03%	3
		2.61%	2.94%	3.36%	5.21%	5.02%	4
			2.64%	2.84%	2.58%	4.76%	5
				1.96%	1.21%	2.50%	6
					0.73%	1.15%	7
						0.72%	8

Figure F1: Cluster distribution for $k=3$.

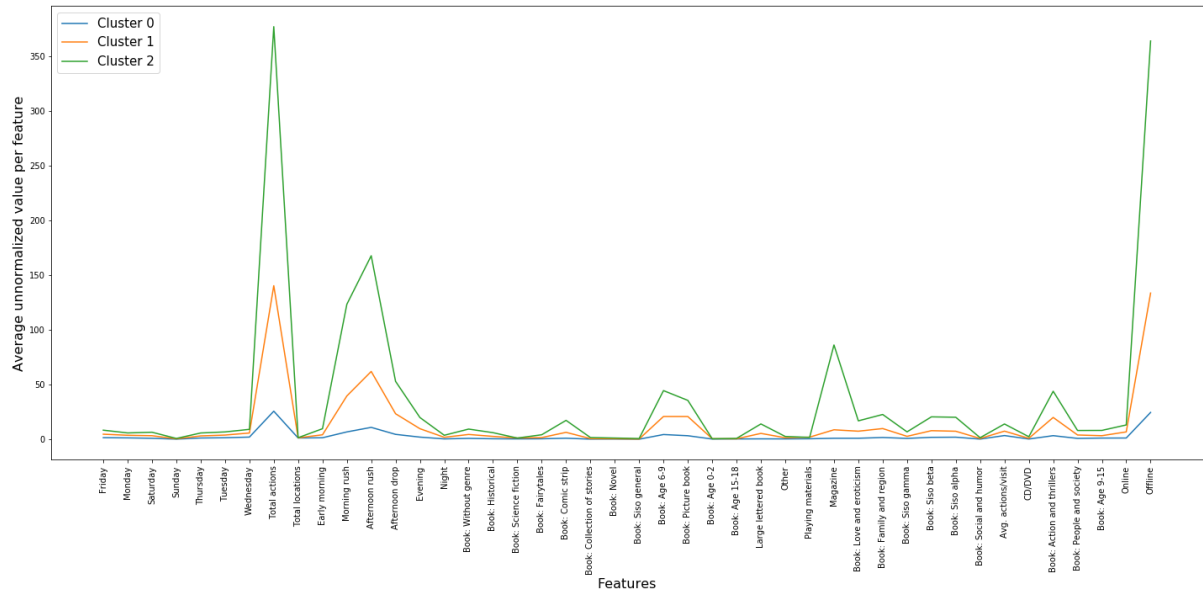


Figure F2: Cluster distribution for $k=4$.

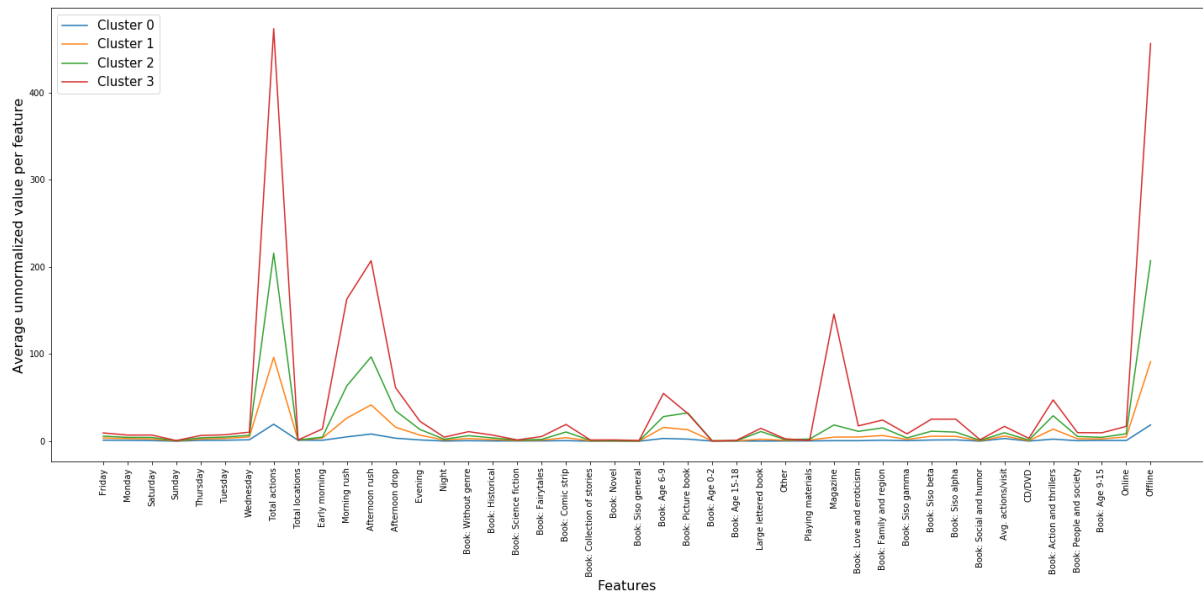


Figure F3: Cluster distribution for $k=5$.

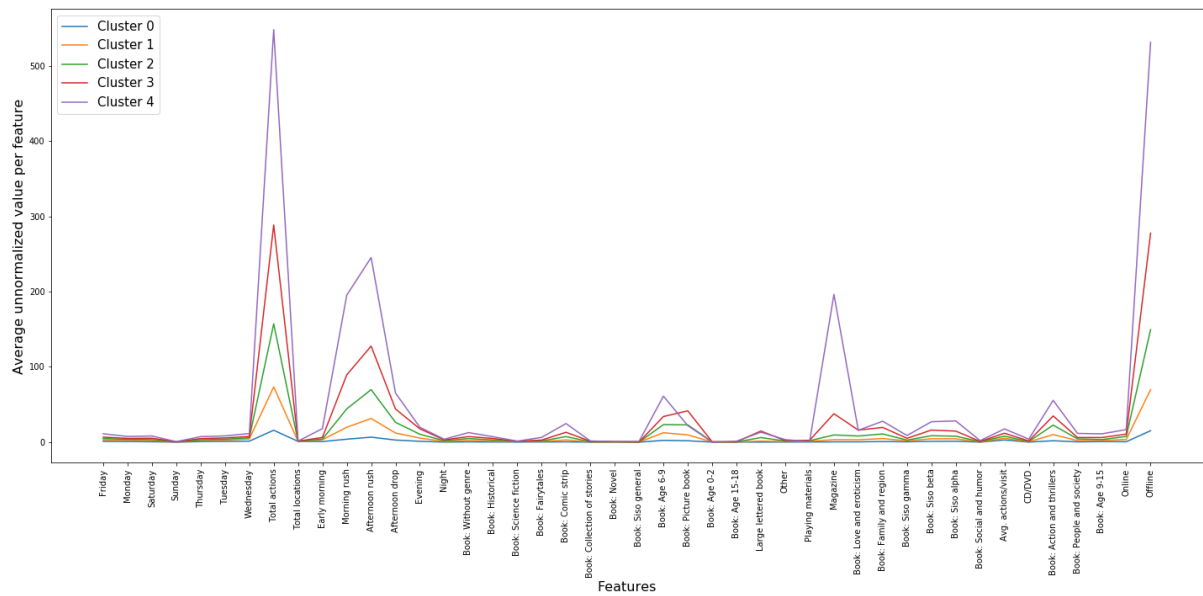


Figure F4: Cluster distribution for $k=6$.

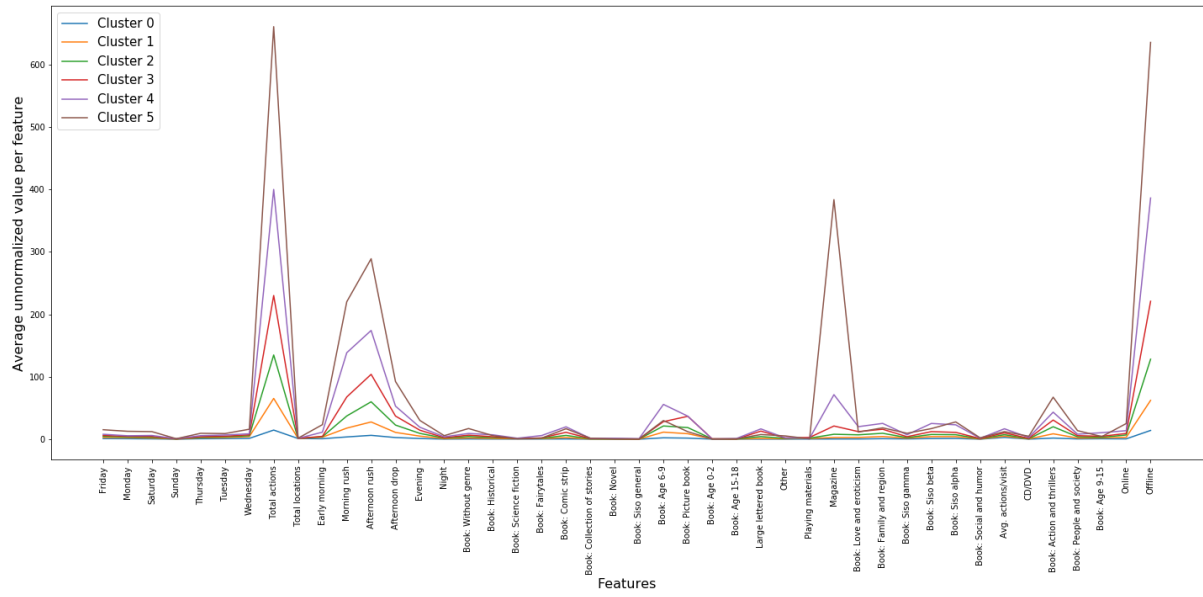


Figure F5: Cluster distribution for $k=7$.

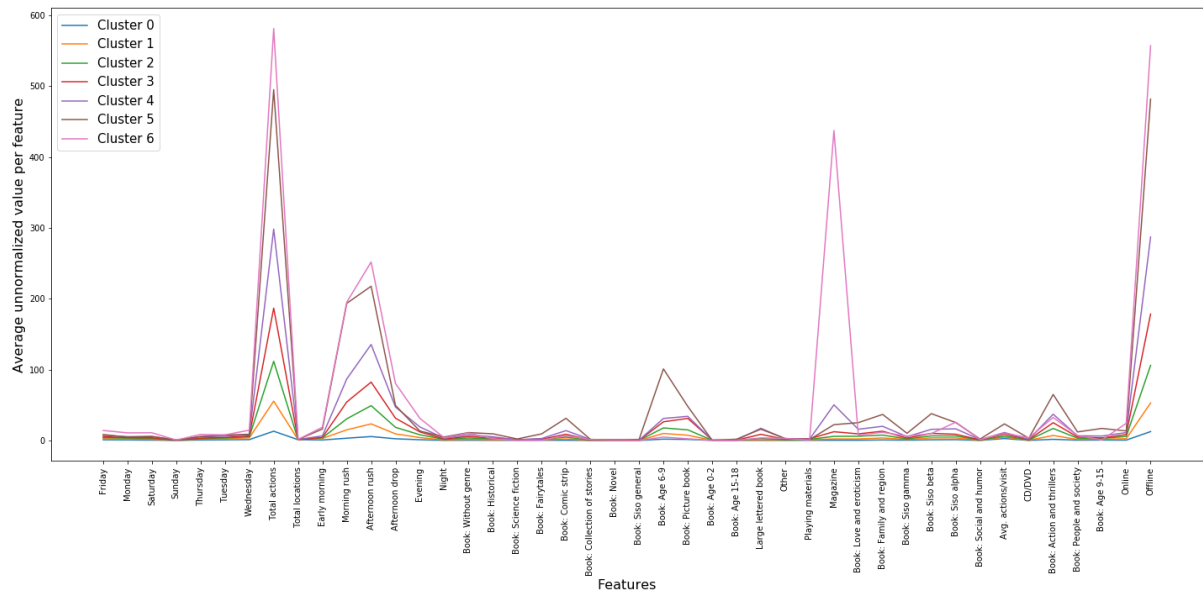


Figure F6: Cluster distribution for $k=8$.

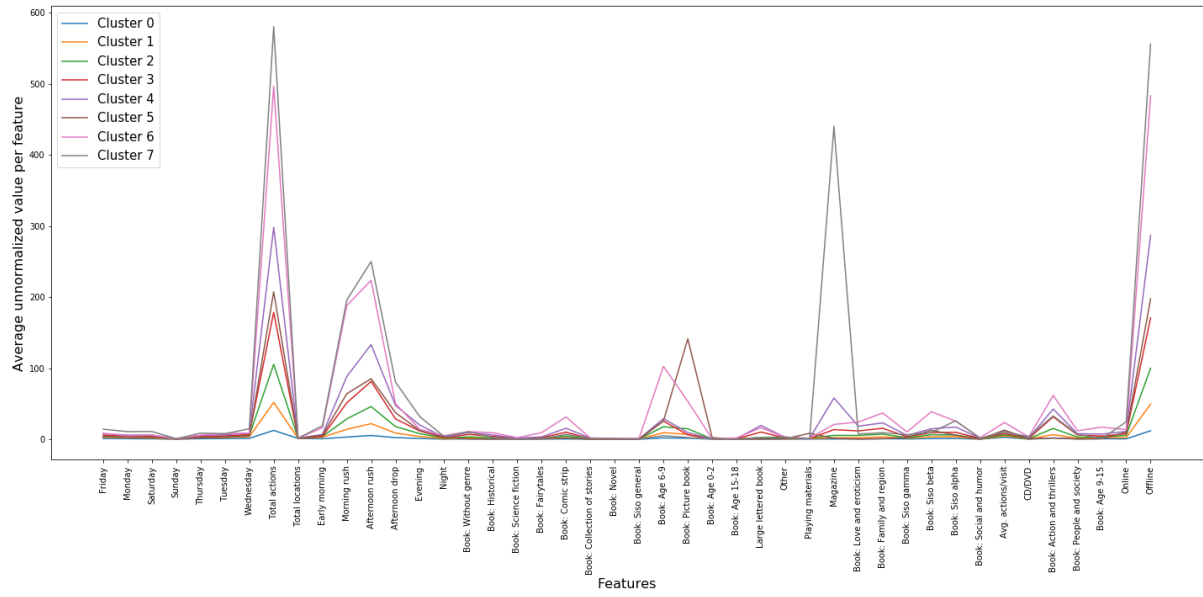
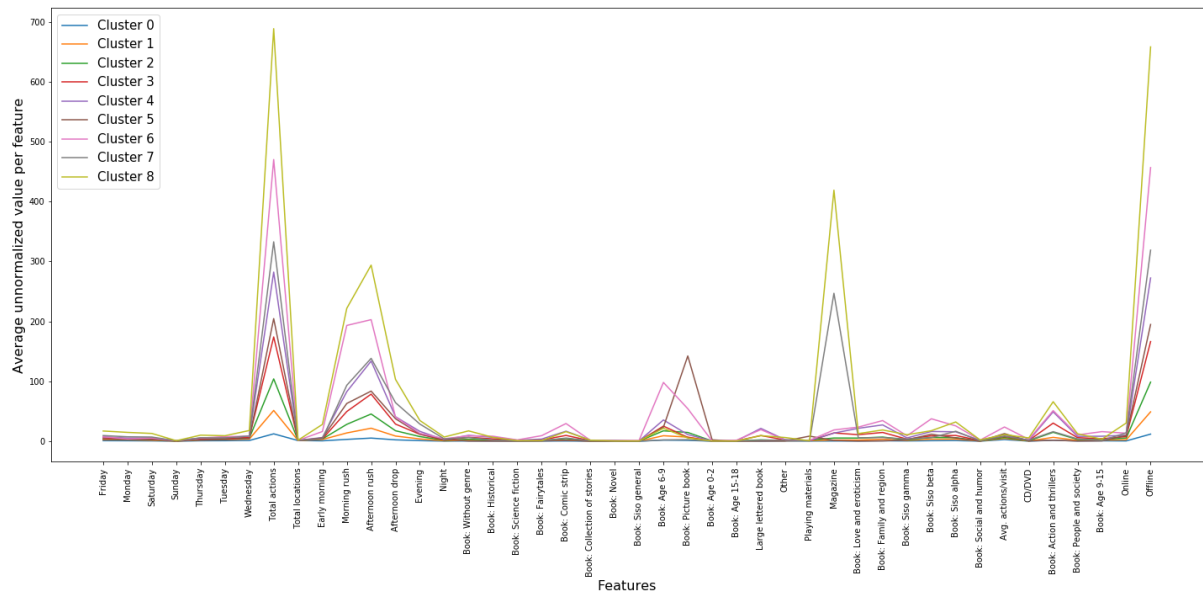


Figure F7: Cluster distribution for $k=9$.



Appendix G: Cluster sizes and distributions for each number of k for the normalized feature matrix

Table G1: Cluster sizes in percentages for each number of k .

$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	Cluster
68.25%	66.52%	64.88%	63.82%	59.39%	57.35%	55.43%	0
27.28%	25.55%	20.72%	20.02%	20.28%	21.19%	20.35%	1
4.47%	4.63%	7.87%	7.15%	6.13%	5.88%	5.04%	2
	3.30%	3.92%	3.43%	6.04%	5.85%	5.03%	3
		2.61%	2.94%	3.36%	5.21%	5.02%	4
			2.64%	2.84%	2.58%	4.76%	5
				1.96%	1.21%	2.50%	6
					0.73%	1.15%	7
						0.72%	8

Figure G1: Cluster distribution for $k=3$.

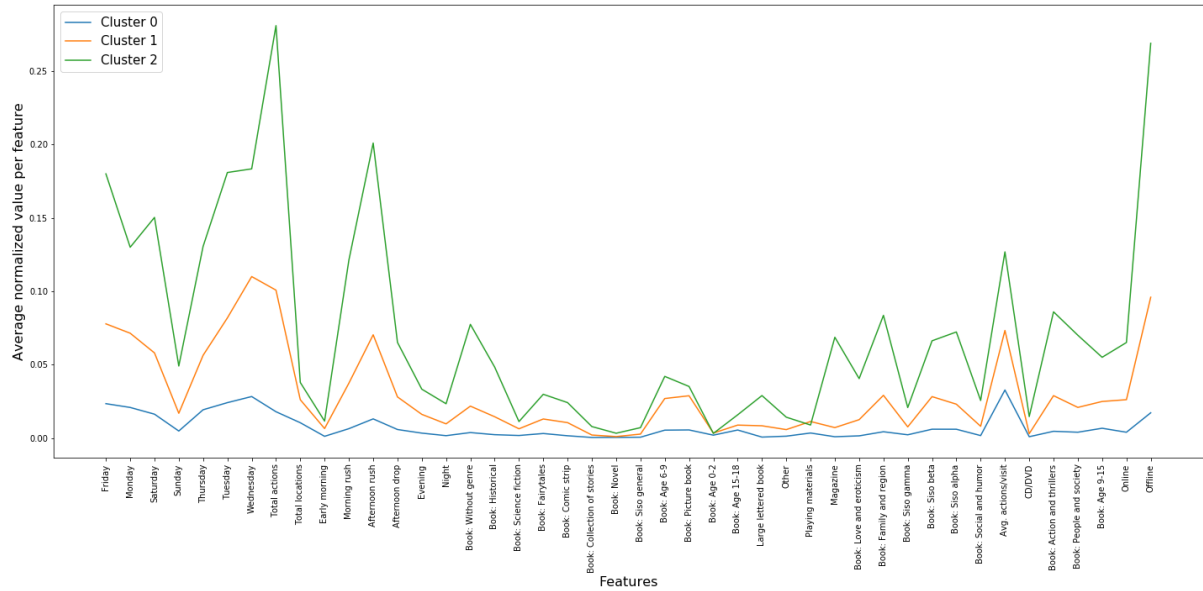


Figure G2: Cluster distribution for $k=4$.

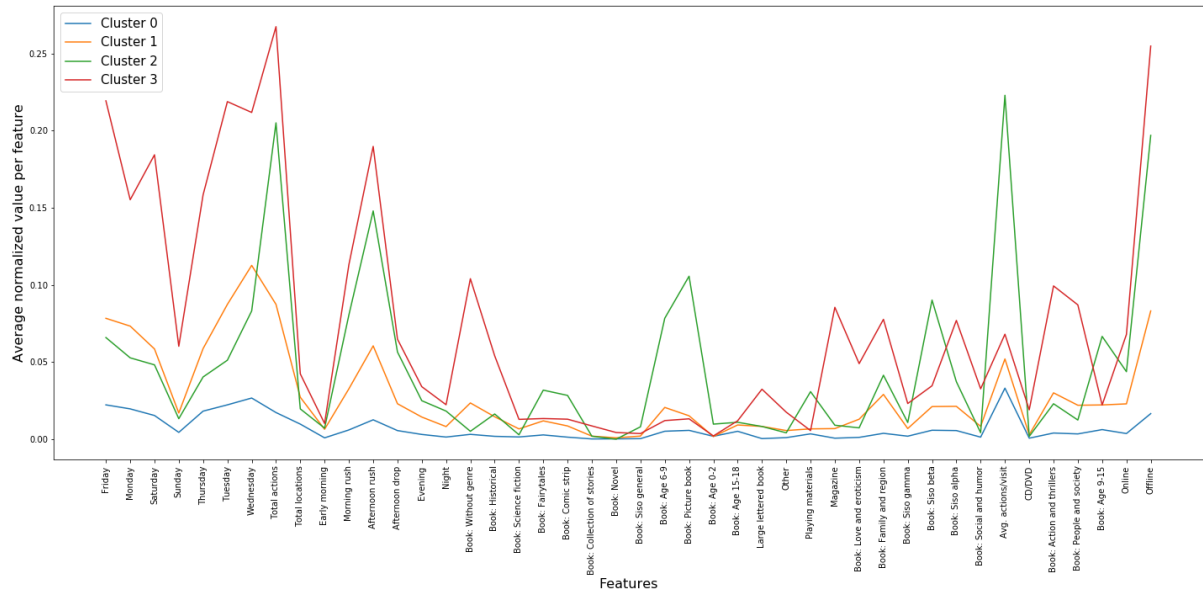


Figure G3: Cluster distribution for $k=5$.

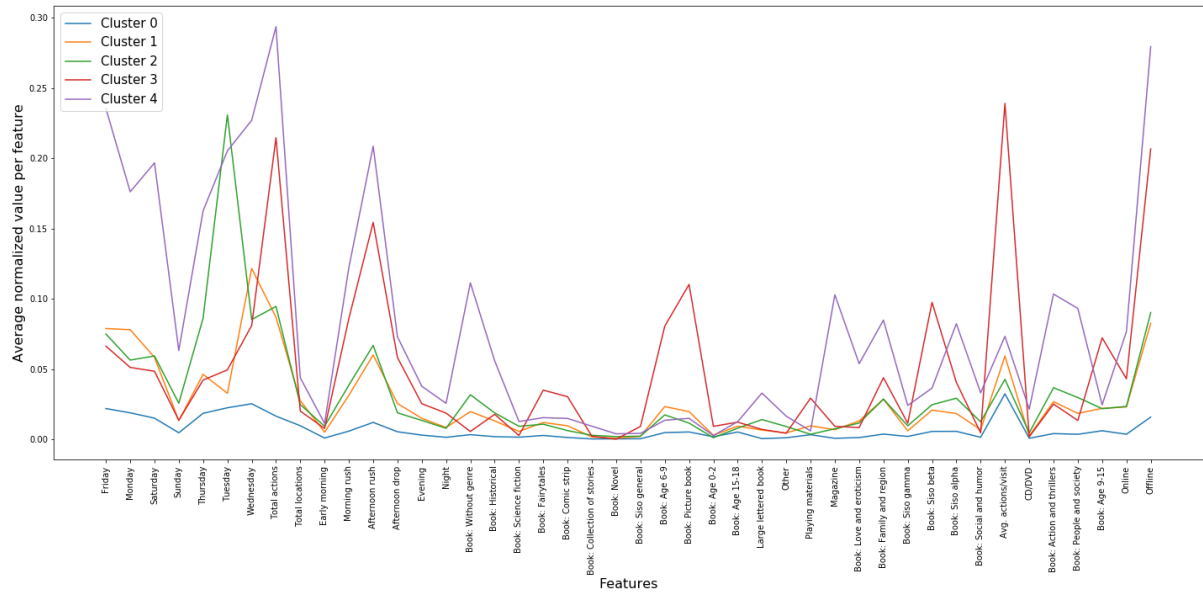


Figure G4: Cluster distribution for $k=6$.

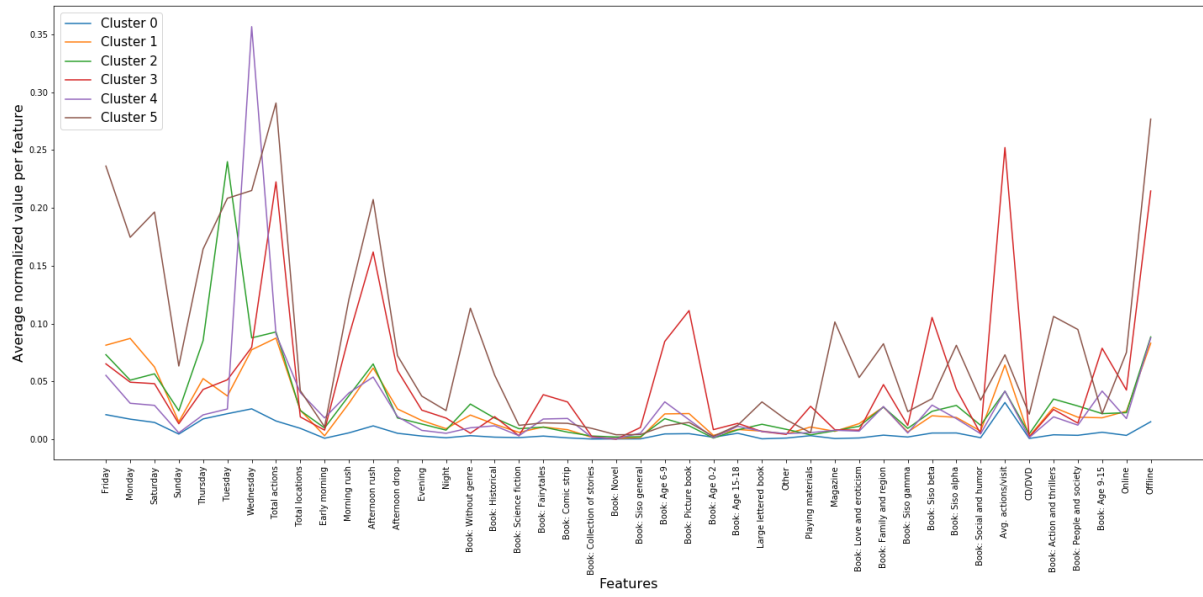


Figure G5: Cluster distribution for $k=7$.

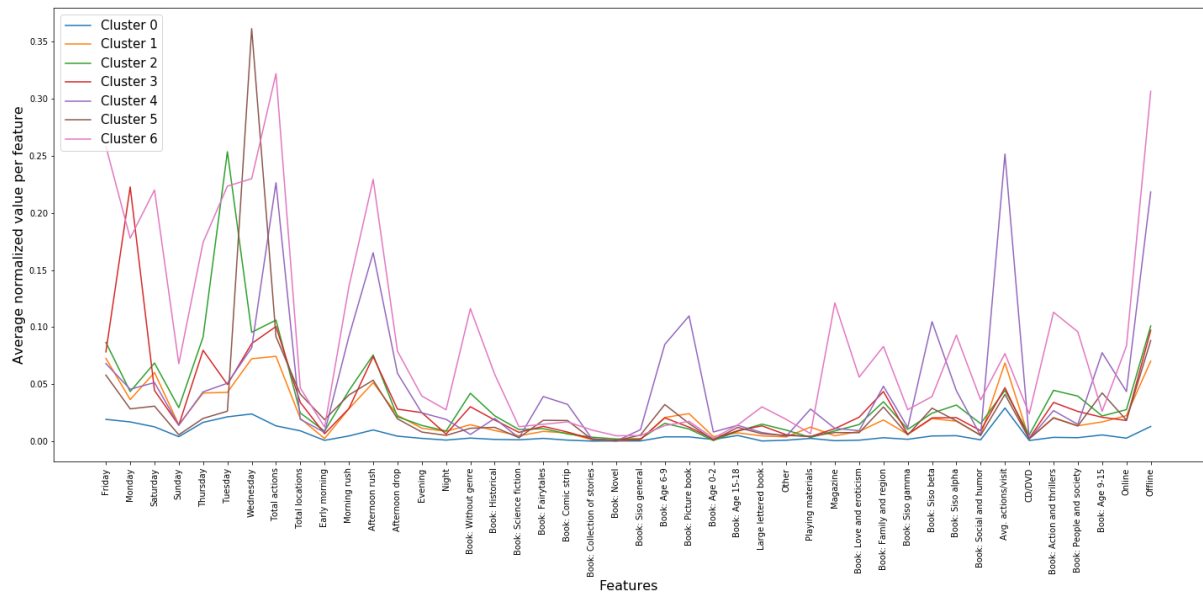


Figure G6: Cluster distribution for $k=8$.

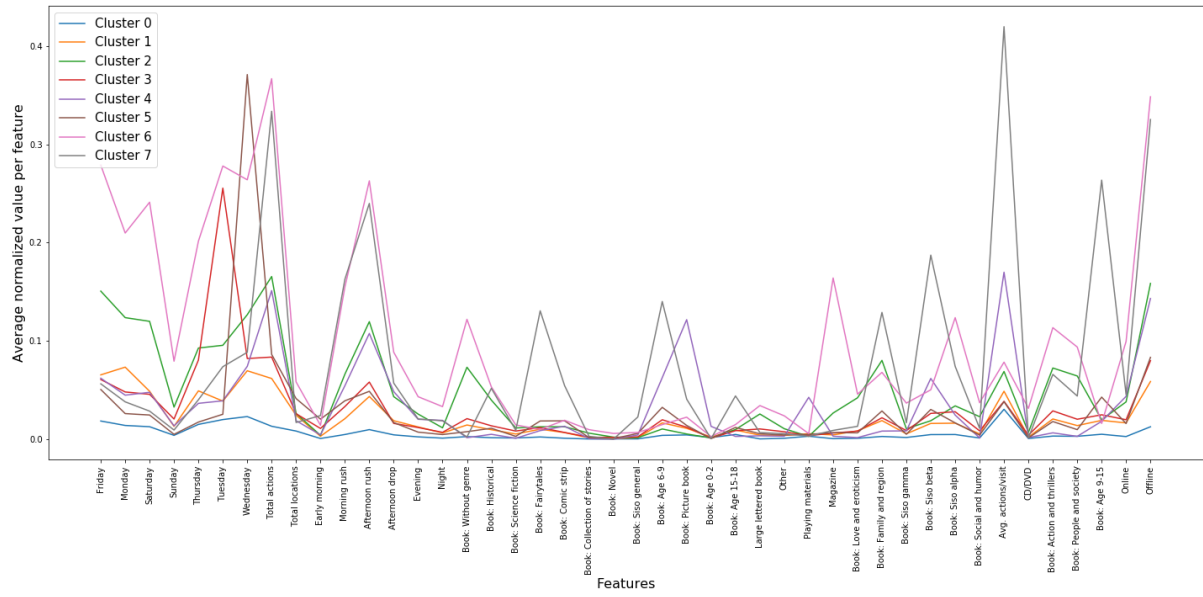
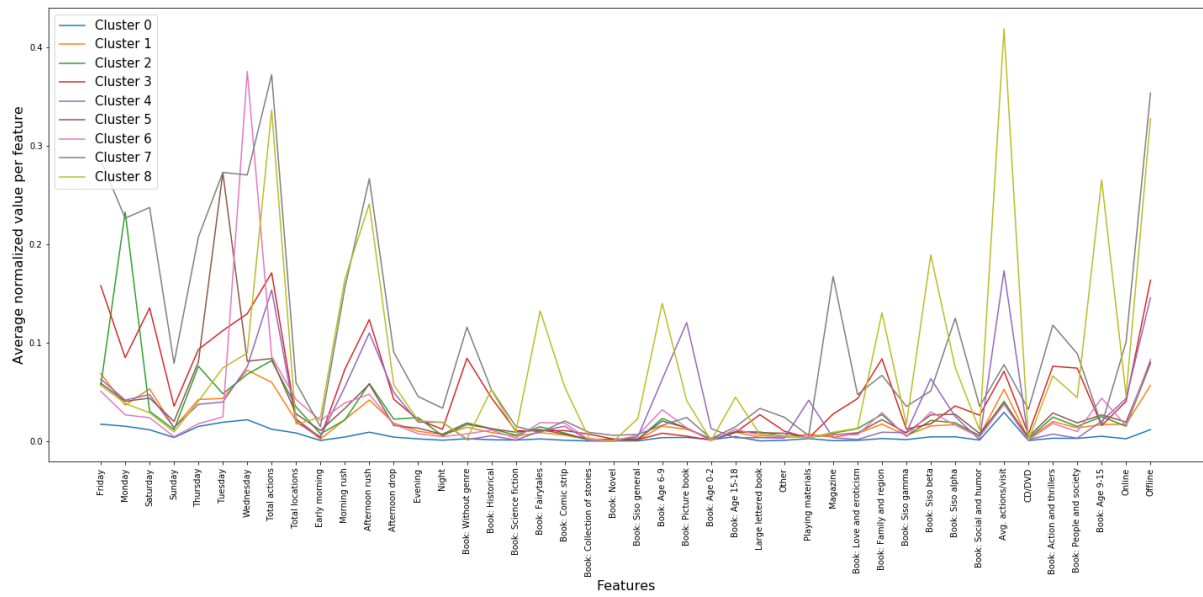


Figure G7: Cluster distribution for $k=9$.



Appendix H: Mean and standard deviation per feature per cluster for $k=6$ with the feature matrix with normalized values

Table H1: Mean and standard deviation per feature per cluster for $k=6$ with the feature matrix with normalized values.

Features	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Friday	1.07	1.45	4.07	3.63	3.66	3.51	3.26	3.27	2.76	3.69	11.80	8.48
Monday	0.82	1.48	4.10	4.46	2.40	3.72	2.32	3.52	1.47	2.74	8.20	8.86
Saturday	0.66	1.20	2.81	3.30	2.55	3.03	2.17	3.21	1.32	2.45	8.84	7.42
Sunday	0.08	0.33	0.27	0.68	0.44	0.96	0.24	0.68	0.10	0.40	1.14	2.02
Thursday	0.85	1.40	2.52	3.04	4.09	3.80	2.07	3.36	1.02	1.88	7.89	6.60
Tuesday	0.95	1.45	1.61	1.83	10.31	4.49	2.21	3.13	1.13	1.93	8.95	8.01
Wednesday	1.34	1.67	3.95	2.93	4.47	3.46	4.06	3.78	18.18	6.66	10.96	7.49
All actions	18.50	15.82	97.49	40.48	103.32	51.37	246.35	98.82	101.30	60.84	321.47	144.13
Locations	1.10	0.34	1.28	0.58	1.27	0.54	1.21	0.51	1.45	0.62	1.46	0.88
Early morning	0.68	2.93	2.16	6.81	6.95	15.55	5.62	21.96	13.08	27.17	7.63	33.93
Morning rush	4.67	7.29	25.90	25.00	31.37	31.32	74.29	83.68	33.38	32.56	100.63	83.51
Afternoon rush	8.04	9.29	42.54	28.37	44.95	33.42	111.67	81.20	37.17	37.07	142.93	91.57
Afternoon drop	3.50	5.87	17.46	18.71	12.32	16.39	39.55	40.74	13.05	22.31	48.12	57.76
Evening	1.41	3.88	8.04	15.20	6.54	17.91	12.45	32.33	3.83	14.29	18.43	36.52
Night	0.20	1.06	1.38	3.78	1.19	3.48	2.76	8.97	0.79	2.70	3.73	9.27
Book: Without genre	0.47	1.74	3.11	6.91	4.51	8.69	0.78	3.85	1.48	5.08	16.77	21.15
Book: Historical	0.26	1.17	1.79	5.15	2.51	6.98	2.69	6.60	1.61	3.88	7.59	13.95
Book: Science fiction	0.22	1.59	0.88	5.59	1.34	7.52	0.45	1.51	0.46	4.22	1.77	8.97
Book: Fairytales	0.31	1.32	1.17	4.04	1.16	3.67	4.25	9.54	1.93	5.25	1.57	7.46
Book: Comic strip	0.65	3.18	4.20	13.35	3.14	10.85	16.21	37.26	9.06	21.74	6.95	31.65
Book: Collection of stories	0.06	0.50	0.37	1.43	0.61	2.02	0.51	6.15	0.37	1.80	2.01	5.14
Book: Novel	0.14	1.29	0.29	3.41	0.79	6.58	0.05	1.69	0.23	3.02	1.53	14.57
Book: Siso general	0.03	0.30	0.11	0.60	0.15	0.76	0.70	2.25	0.37	1.53	0.28	2.36
Book: Age 6-9	3.20	7.72	14.99	27.28	12.20	25.41	57.70	80.59	22.09	33.65	7.98	31.20
Book: Picture book	2.87	8.02	13.00	27.95	6.87	19.92	64.97	82.20	9.90	23.82	8.54	35.32
Book: Age 0-2	0.20	1.70	0.38	3.22	0.12	1.80	0.91	5.78	0.16	1.42	0.29	2.49
Book: Age 15-18	0.24	1.23	0.38	2.12	0.37	1.85	0.62	2.68	0.51	1.62	0.55	2.40
Large lettered book	0.21	2.56	3.16	19.00	5.76	28.76	2.97	29.31	3.06	20.46	14.29	56.68
Other	0.22	1.36	0.91	5.08	1.69	7.34	0.85	4.67	0.96	4.49	3.31	10.52
Playing materials	0.43	2.22	1.44	5.55	0.52	3.20	3.86	11.45	0.78	4.38	0.77	5.45
Magazine	0.61	3.44	6.03	20.48	6.45	21.80	7.23	33.61	6.82	26.70	88.02	146.99
Book: Love and eroticism	0.52	2.50	5.82	15.10	4.84	13.38	3.39	14.41	3.04	11.34	22.86	44.61
Book: Family and region	0.96	2.85	7.38	15.61	7.36	16.01	12.49	25.34	7.49	12.50	21.80	38.78
Book: Siso gamma	0.60	2.32	1.88	5.66	2.81	7.79	3.56	6.26	1.63	5.09	7.11	18.10
Book: Siso beta	1.30	3.52	4.85	9.41	5.77	11.13	25.07	34.08	7.04	10.73	8.36	15.02
Book: Siso alpha	1.42	3.88	4.86	9.96	7.53	13.52	11.14	15.61	4.60	8.86	20.88	33.49
Book: Social and humor	0.09	0.56	0.47	1.56	0.80	2.21	0.32	1.13	0.32	1.24	2.24	4.40
Avg. actions/visit	3.31	2.50	5.66	2.97	4.02	2.21	19.30	11.78	4.03	2.30	6.30	3.45
CD/DVD	0.15	1.08	0.49	3.39	0.92	4.81	0.36	2.28	0.31	1.92	4.25	17.63
Book: Action and thrillers	2.07	4.90	14.53	25.16	18.28	31.63	13.54	27.23	10.22	19.56	55.75	75.17
Book: People and society	0.49	1.62	2.68	6.54	4.04	8.62	1.97	4.90	1.72	4.83	13.29	19.40
Book: Age 9-15	0.76	2.14	2.31	5.77	2.75	6.22	9.76	18.38	5.17	8.85	2.74	10.98
Online	0.70	2.19	4.96	9.52	4.71	9.41	8.69	20.17	3.67	9.10	15.32	25.18
Offline	17.80	15.30	92.53	39.54	98.62	49.42	237.66	98.05	97.63	58.26	306.15	142.42