Evan Phillippi

Data Science Practicum: Project Report

Introduction

Customer segmentation is a common analysis performed in marketing and business. The purpose of customer segmentation is to understand customer behavior, build a profile of customers, and group customers together in a logical and qualitative manner. Customers often have different needs and profiles. The ultimate goal of customer segmentation includes maintaining or increasing customer retention and increasing the likelihood they will consume a specific product.

The Charles Book Club was created in 1986 and was created for the purpose that it could better understand its customers by offering tailored book selections based on individual customer behavior. The club is a only a distributor of books meaning it is not a publisher and advertised to its members through a variety of media and maintained a database of its customers' purchases.

A book called “The Art History of Florence”, is about to be released. The Charles Book club sent a mailing to a random sample of 4,000 customers. The responses of customers were then collated with past purchases data. The result is a dataset that contains demographic and purchase data on the customers and whether or not they purchased “The Art History of Florence”. The purpose of this project was to apply a customer segmentation technique to the customer dataset. Specific emphasis is placed on whether the resulting segments or clusters are able to draw a distinction on whether or not “The Art History of Florence” was purchased. As well as what other variables in the data are associated with purchasing “The Art History of Florence” title.

Materials and Methods

The charles book club dataset contains 4,000 records of unique customer purchase behavior and is summarized in table 1 below. All variables are treated as numeric variables due to the nature of the analysis which will be subsequently explained. The variables “Seq#” and “ID#” were not used in the analysis as they are merely unique identifiers for customers.

Table 1: Overview of Variables on Charles Book Club Data

Variable Name	**Description**
Seq#	Sequence number in the partition
ID#	Identification number in the full (unpartitioned) market test data set
Gender	0=Male 1=Female
M	Monetary- Total money spent on books
R	Recency- Months since last purchase
F	Frequency - Total number of purchases
FirstPurch	Months since first purchase
ChildBks	Number of purchases from the category: Child books
YouthBks	Number of purchases from the category: Youth books
CookBks	Number of purchases from the category: Cookbooks
DoItYBks	Number of purchases from the category: Do It Yourself books I
RefBks	Number of purchases from the category: Reference books (Atlases, Encyclopedias, Dictionaries)
ArtBks	Number of purchases from the category: Art books
GeoBks	Number of purchases from the category: Geography books
ItalCook	Number of purchases of book title: "Secrets of Italian Cooking"
ItalAtlas	Number of purchases of book title: "Historical Atlas of Italy"
ItalArt	Number of purchases of book title: "Italian Art"
Florence	=1 'The Art History of Florence" was bought, = 0 if not
Related purchase	Number of related books purchased

The analysis of the data for customer segmentation primarily employed principal component analysis and k-means clustering. Both were applied following initial exploratory analysis. Principal component analysis is a form of dimensionality reduction. It is often employed with large datasets that contain collinear variables. The result of principal component analysis is a new synthetic set of variables called principal components which are uncorrelated and orthogonal to the original data. Typically, the number of principal components chosen is related to the proportion of the variance explained in the original dataset. After principal component analysis was done, the resulting dataset contained pca scores for every observation in the dataset for every principal component which accounted for approximately 80% of the variance in the data. Next, k-means clustering was applied to the synthetic dataset of 4000 observations. K-menas clustering is a commonly used algorithm to identify natural groups or clusters in a dataset. K-means clustering accomplishes this by iteratively assigning centroids to the data. The goal is to minimize the within cluster sums of squares and maximize the euclidean distance between each cluster centroid. The result should be heterogeneous groups within the dataset.

After clustering was applied, box and whisker plots were generated on each variable in the dataset according to their cluster membership. This allowed for qualification of cluster membership based on the data and so the customer segmentation was accomplished. All analysis was performed in R primarily using the base “stats” package as well a suite of packages from the “tidyverse”.

Results and Discussion

Exploratory Analysis

To determine the correlation structure across the entire dataset, a correlation table along with distributions was generated as can be seen below in figure 1. The strongest positive correlations are associated with “First Purchases”, “F” or frequency, “Youth Books”, “Cookbooks”, and “Do It Yourself Books”. There are no strong negative correlations in the dataset. Interestingly, the purchase of “Art History of Florence” is not strongly correlated with any variable. Furthermore, none of the books that feature specific content about Italy appear to have strong correlations with other variables.

Figure 1: Correlations and Distributions

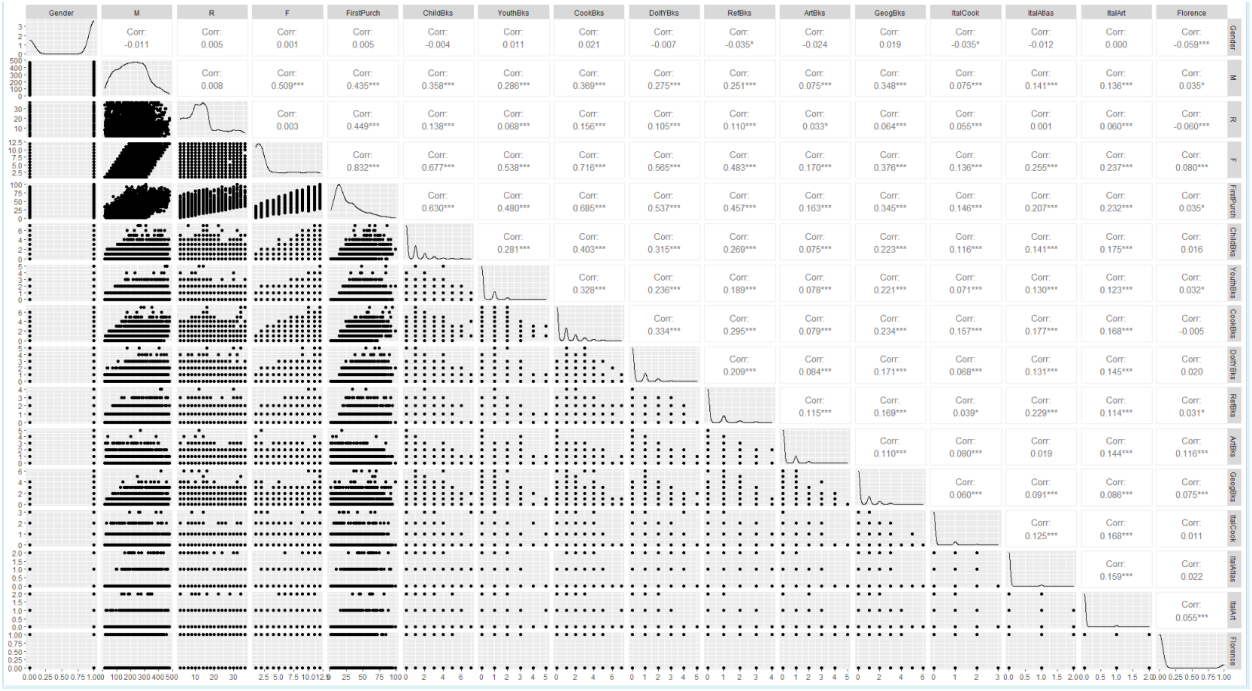
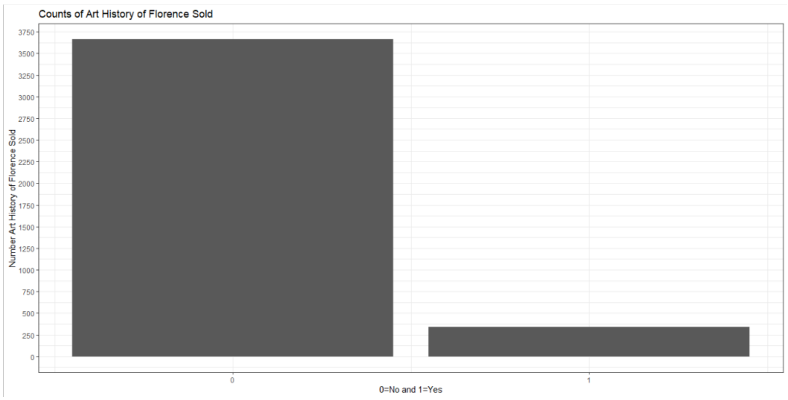


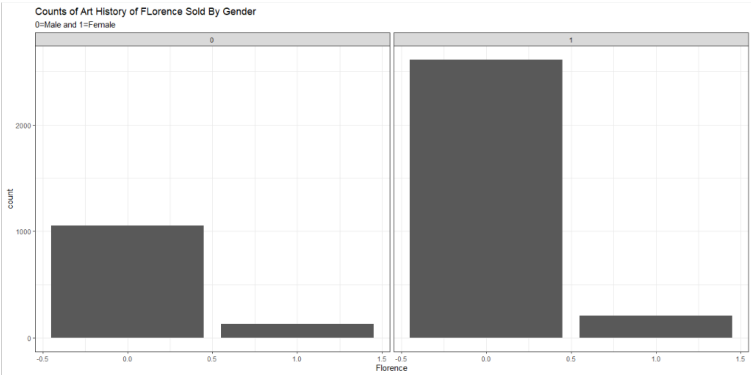
Figure 2 seen below shows the counts of customers who purchased “The Art History of Florence”. As can be seen, customers overwhelmingly did not purchase the book. And those that did represent less than 10% of the customers.

Figure 2: Counts of members who purchases The Art History of Florence



The only demographic type data is the gender of the member. As can be seen below in figure 3, the majority of members are female and while more women than men purchased “The Art of Florence”, the proportions still resemble the overall dataset when segmenting by gender.

Figure 3: Purchases of the Art History of Florence by Gender

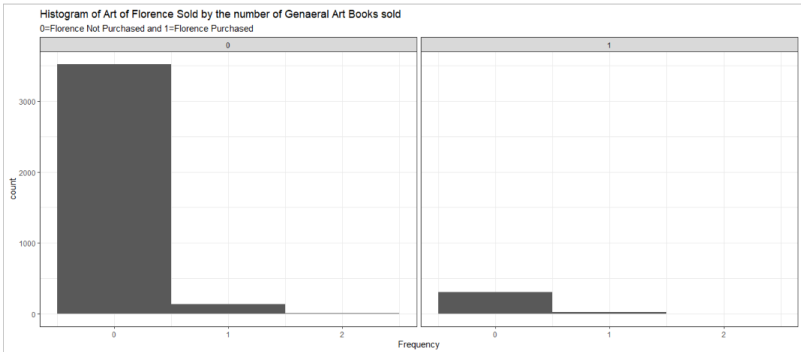


Of particular interest in this dataset are the variables recency (R), frequency (F), and monetary value (M). These variables are commonly used in marketing and business and represent the general purchasing behavior of customers. Recency represents the months that have elapsed since the most recent purchases. Frequency are the total number of purchases over agiven time period, and monetary value is the cumulative sum of those purchases. Typically in customer segmentation, it is the goal to identify those customers with a high score for each. These represent the most loyal customers who spend the most and make the most purchases in shorter periods of time. Figure 4 shows histograms, density functions and means for R, F, and M in the dataset. As can be seen, both R and F are right skewed with means on the left side of their distributions. Monetary value is somewhat less right skewed with a mean falling closer to the center.

Figure 4: Histogram and Density Plots with Means of Recency, Frequency, and Monetary Value

In addition to the variables initially examined above, the number of purchases of Italian art Books was also examined. It could be reasonably hypothesized that the purchase of or interest in Italian art books might be an indicator of whether or not “The Art History of Florence was purchased” . As can be seen below in figure 5, only a small proportion of members purchased an Italian Art Book as well as “The Art History of Florence”.

Figure 5: Counts of Italian art books purchased segmenting on purchase of “The Art History of Florence”



Principal Component Analysis

Principal component analysis was accomplished with the use of the prcomp function of the stats package in R. All data was centered and scaled due to different measures for each variable. As can be seen below in table 2, summary output from the pca object generated from the book data indicate that components 1 through 9 account for approximately 80% of the variance in the original dataset and so were selected for clustering analysis.

Table 2: Importance of principal components

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     2.1407 1.10815 1.07653 1.03186 1.00374 0.96671 0.93827 0.92899 0.89656 0.87238 0.8514 0.80243
Proportion of Variance 0.2864 0.07675 0.07243 0.06655 0.06297 0.05841 0.05502 0.05394 0.05024 0.04757 0.0453 0.04024
Cumulative Proportion  0.2864 0.36315 0.43559 0.50213 0.56510 0.62351 0.67853 0.73247 0.78271 0.83027 0.8756 0.91582
                          PC13    PC14    PC15    PC16
Standard deviation     0.77076 0.75981 0.35720 0.21880
Proportion of Variance 0.03713 0.03608 0.00797 0.00299
Cumulative Proportion  0.95295 0.98903 0.99701 1.00000
```

Additionally, principal component 1 accounts for the most variability (figure 6), this is common with pca. A loading plot shown in figure 7 displays the relative importance of each variable in each variable. As can be seen, the first principal component is primarily built from “First Purchase”, “F”, and “Cook Books”. Indicating there is some linear relationship amongst these variables. While principal component 2 primarily accounts for “R”, “Art Books”, “The Art History of Florence”. Indicating a linear relationship between these variables.

Figure 6: Scree plot indicating the percentage of variance capture in each principal component.

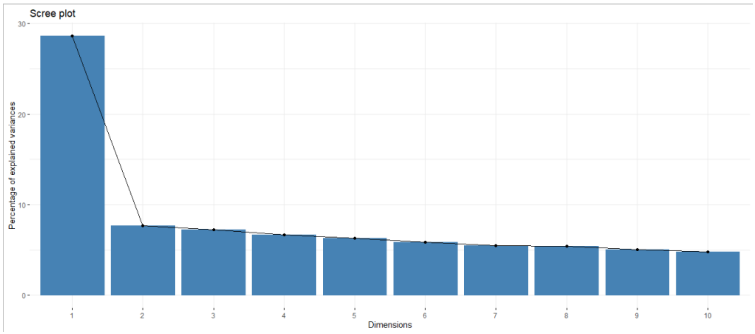
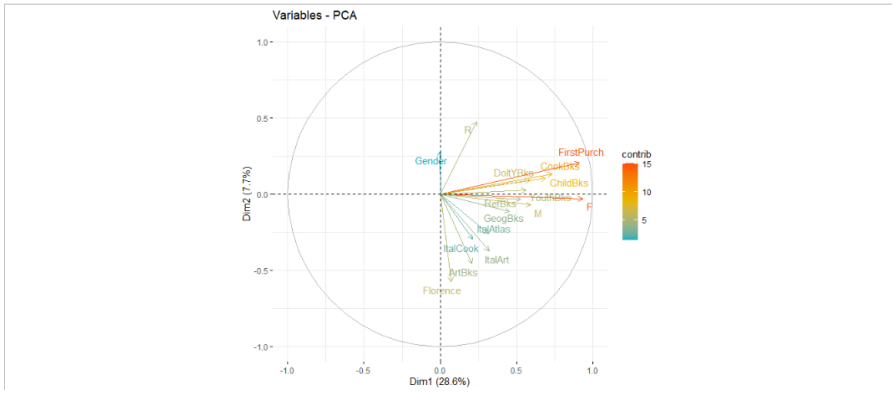


Figure 7: Loading plot for principal components 1 and 2.



K-means Clustering

After principal component analysis was conducted, the pca scores for every observation in the first nine principal components was used to be clustered. To choose the optimal number of clusters the elbow and silhouette methods were used. Elbow and silhouette methods are both gap statistics used to indicate the number of clusters at which the the within group sum of squares decreases the most. Figure 8 and 9 show agreement that the optimal number of clusters for the pca scores is 2.

Figure 8: Elbow method for optimal number of clusters

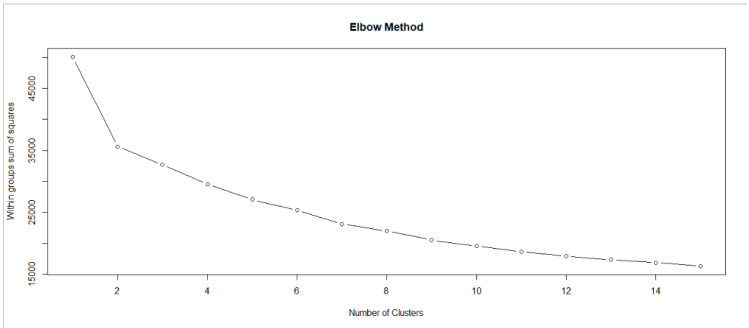
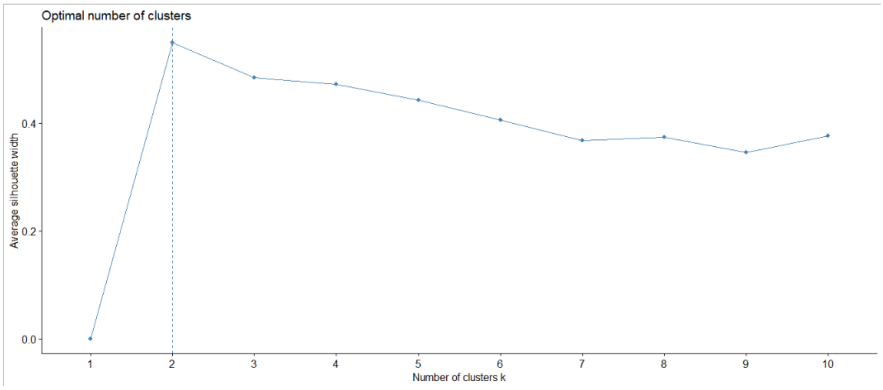


Figure 9: Silhouette method for optimal number of clusters



Clustering Results

Figures 10 and 11 show exploratory graphs that are used to qualify membership in each cluster based on the dataset after pca and k-means were applied. The results indicate that there are two customer segments which are summarized beneath the figures.

Figure 10: Cluster membership based on gender and purchase of The Art History of Florence.

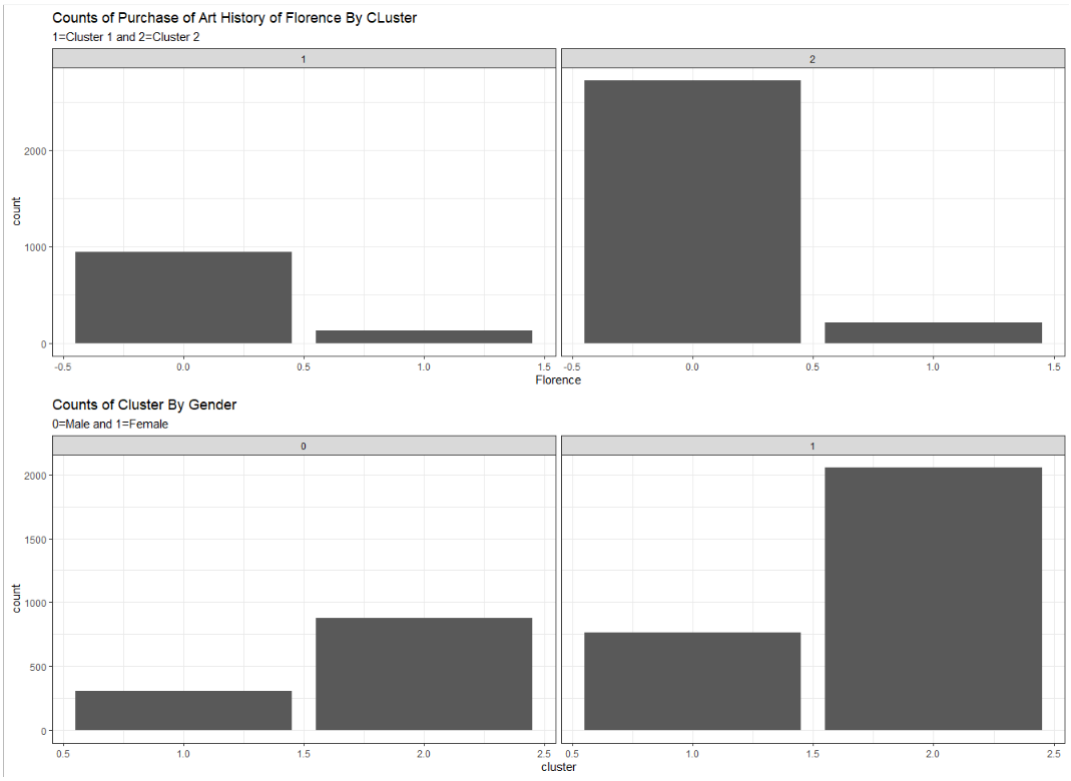
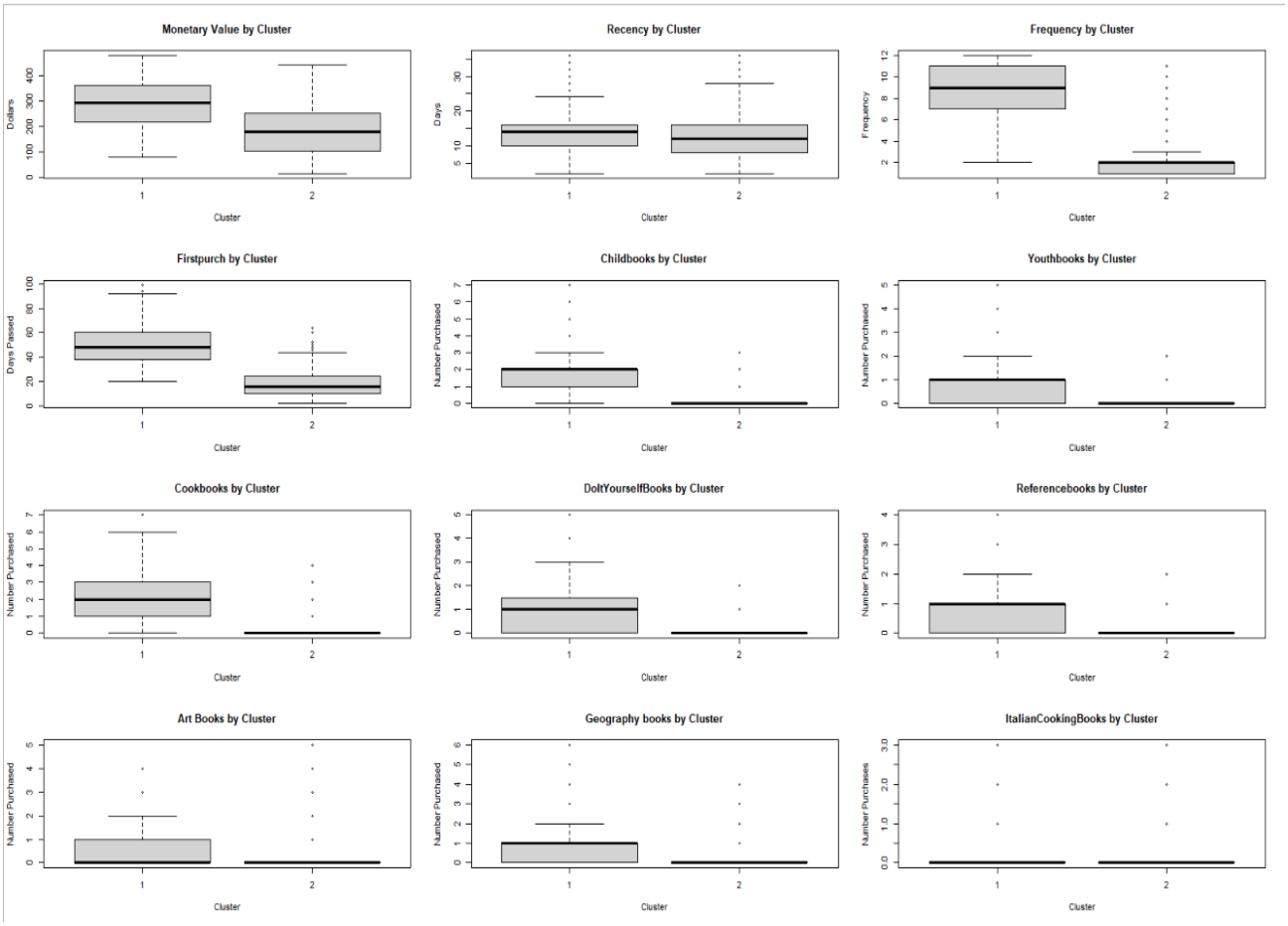


Figure 11: Box and Whisker plot of remaining variables by cluster membership



Customer Segment 1: This segment is a mix of male and females and has fewer people who purchased “The Art History of Florence” than segmen 2. Customers in this segment have spent more money on the Charles Book Club and purchase more books . However, more days pass between purchases in this segment. This segment also purchases significantly more Childrens, Youth, Cookbooks, Do It yourself Books, and Reference books. This segment clearly utilizes the Charles Book club and should be primarily targeted for future purchases.

Customer Segment 2: This segment is overwhelmingly female and contains few people who purchased books. In general these are the least active customers. They spend slightly less than segment 1 and in general they purchase books at similar time intervals. These customers should not be the focus of future targeted advertising as they are unlikely to purchase.

Conclusion

This project conducted customer segmentation of book sales data using pca and k-means clustering. After the analysis customers were segmented into two distinct groups. The segments were not driven by gender and surprisingly, also were not driven by the purchase of “The Art

History of Florence”. In general, the segments were differentiated by those customers who are active and frequent buyers of books across genres.