Purchase of Office Equipment

Akua Kyeraa Sam

2023-09-24

# Project 1: The Choice of Outlet to Buy Office Equipment

## Background

In this project, I use data from a survey of 200 respondents regarding the importance of several store attributes when choosing where to buy office equipment. The data has the following variables.

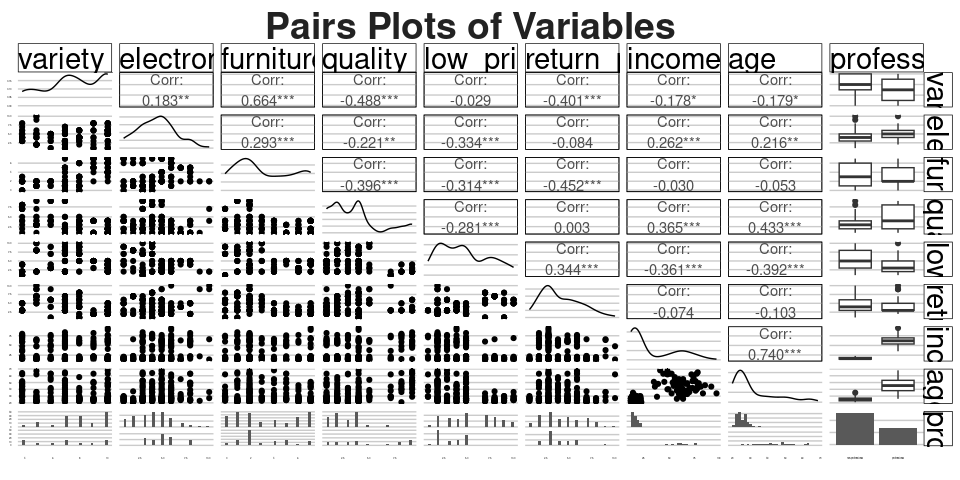
Variables Description

| **Variable** | **Description** |
| --- | --- |
| respondent\_id | An identifier for our observations |
| variety\_of\_choice | Importance of this attribute on a 0-10 scale |
| electronics | Importance of this attribute on a 0-10 scale |
| furniture | Importance of this attribute on a 0-10 scale |
| quality\_of\_services | Importance of this attribute on a 0-10 scale |
| low\_prices | Importance of this attribute on a 0-10 scale |
| return\_policy | Importance of this attribute on a 0-10 scale |
| professional | Whether the respondent is a professional or not (e.g., student) |
| income | Gross annual income expressed in thousands of pound sterling |
| age | Respondents’ age in years |

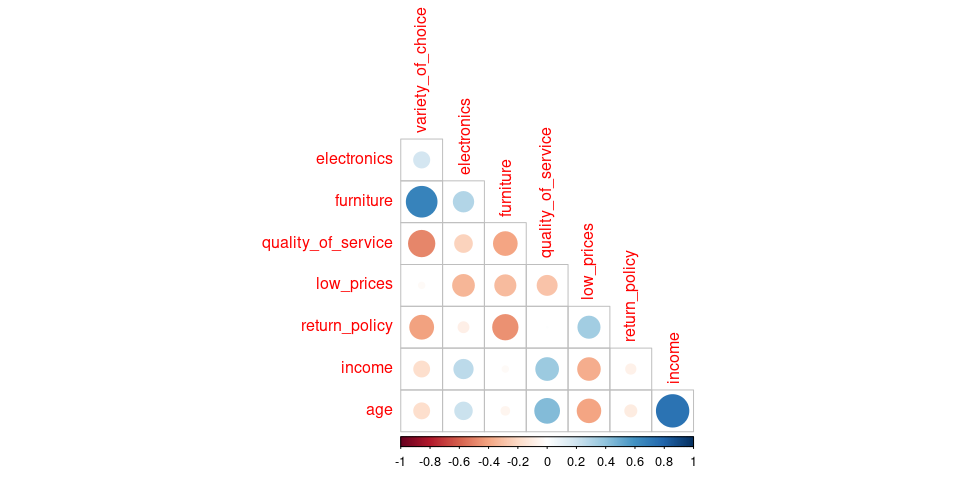
## QUESTION 1: Descriptive Statistics of Each Variable in the dataset

The data has 200 rows and 10 columns. I first read in the data into R. I run a correlation analysis and summary statistics of the variables. The data has no missing observations and no duplicate rows. Figure 1 shows the pairs plot for all the variables in the data set. Some of the variables have very high correlation, for instance, age and income at 0.74 and variety of choices and furniture at 0.66. The main diagonal of the pairs plot shows that the variables are highly skewed. Figure 2 shows a visualisation of the correlations.

| Variable | missing |
| --- | --- |
| respondent\_id | 0 |
| variety\_of\_choice | 0 |
| electronics | 0 |
| furniture | 0 |
| quality\_of\_service | 0 |
| low\_prices | 0 |
| return\_policy | 0 |
| professional | 0 |
| income | 0 |
| age | 0 |



**Figure** **:** Pairs Plots of Variables



**Figure** **:** Correlation Matrix

Summary Statistics for Numeric Variables

| **Variable** | **complete\_rate** | **mean** | **SD** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| respondent\_id | 1 | 100.5 | 57.9 | 1 | 51 | 100.5 | 150 | 200 |
| variety\_of\_choice | 1 | 7.6 | 2.0 | 4 | 6 | 8.0 | 10 | 10 |
| electronics | 1 | 4.5 | 1.9 | 1 | 3 | 4.5 | 6 | 10 |
| furniture | 1 | 3.3 | 2.4 | 0 | 1 | 2.0 | 6 | 7 |
| quality\_of\_service | 1 | 3.5 | 2.3 | 1 | 2 | 3.0 | 4 | 9 |
| low\_prices | 1 | 4.8 | 2.5 | 1 | 2 | 5.0 | 7 | 10 |
| return\_policy | 1 | 4.2 | 2.0 | 1 | 3 | 4.0 | 6 | 10 |
| income | 1 | 32.2 | 22.7 | 13 | 15 | 19.5 | 54 | 95 |
| age | 1 | 32.5 | 11.9 | 21 | 24 | 27.0 | 38 | 68 |

Summary of Character Variables

| **skim\_type** | **skim\_variable** | **n\_missing** | **complete\_rate** | **character.min** | **character.max** | **character.empty** | **character.n\_unique** | **character.whitespace** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| character | professional | 0 | 1 | 12 | 16 | 0 | 2 | 0 |

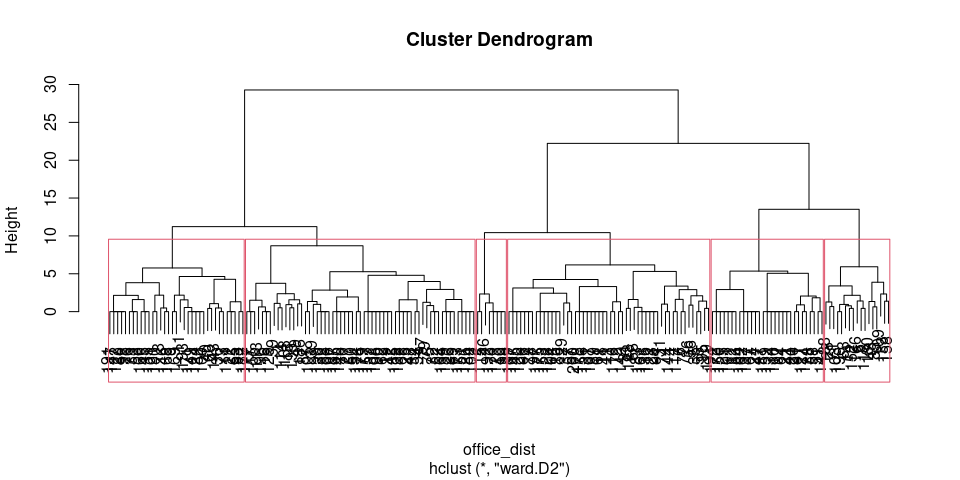
## QUESTION 2: Hierarchical Clustering: Preparing the Data

In this section, I carry out cluster analysis on the dataset. I create a new dataset containing attitudinal variables by eliminating age, income, Professional and respondent\_id. I then standardize the variables.

## QUESTION 3: Hierarchical Clustering: 6 Cluster Solution

I create a distance matrix using the default euclidean method and and then use the hclust function for hierarchical clustering.

Figure () below shows the dendogram for the clustering algorithm.



**Figure** **:** Dendogram for 6 Clusters

## QUESTION 4: Number of Observations in Each of the 6 Clusters

The number of points in each of the clusters in shown in Table () below.

##   
## 1 2 3 4 5 6   
## 59 8 52 17 35 29

Number of Observations per Cluster

| **Cluster** | **Points** |
| --- | --- |
| Cluster 1 | 59 |
| Cluster 2 | 8 |
| Cluster 3 | 52 |
| Cluster 4 | 17 |
| Cluster 5 | 35 |
| Cluster 6 | 29 |

## QUESTION 5A: Summary Statistics per Cluster and Flexclust package

I compute the average of each of the variables per cluster.

variable Means for the 6 Cluster Hierarchical Clustering Solution

| **cluster** | **variety\_of\_choice** | **electronics** | **furniture** | **quality\_of\_service** | **low\_prices** | **return\_policy** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 9.8 | 4.4 | 6.02 | 2.6 | 3.4 | 2.7 |
| 2 | 7.5 | 3.8 | 0.38 | 4.0 | 8.2 | 1.0 |
| 3 | 6.8 | 2.8 | 1.33 | 2.9 | 8.1 | 6.5 |
| 4 | 4.9 | 6.5 | 2.12 | 3.7 | 2.8 | 6.8 |
| 5 | 8.2 | 6.7 | 4.37 | 1.9 | 4.5 | 3.5 |
| 6 | 5.2 | 3.7 | 1.31 | 8.3 | 2.1 | 3.6 |

The output shows that the clustering does distinguish the clusters. It is notable that each of the clusters appear to have a different mean for each of the variables. Next, I use the Flexclust package to repeat the clustering.

### QUESTION 5B: Clustering Using the FlexClust Package

This section I use the flexclust package for clustering.

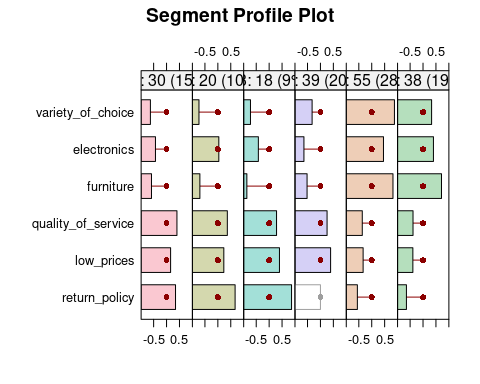
## kcca object of family 'ejaccard'   
##   
## call:  
## kcca(x = office\_cl, k = 6, family = kccaFamily("ejaccard"), control = new("flexclustControl"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5 6   
## 24 24 21 42 48 41

The cluster assignments in this case are as follows;

Number of Observations per Cluster

| **Cluster** | **Points** |
| --- | --- |
| Cluster 1 | 24 |
| Cluster 2 | 24 |
| Cluster 3 | 21 |
| Cluster 4 | 42 |
| Cluster 5 | 48 |
| Cluster 6 | 41 |

I develop a segment separation plot. The first step in making a segment separation plot is to do a Principal Component Analysis (PCA). Plotting the first two principal components allows us to develop the desired plot which depict the position of each data point in the clusters using two dimensions.



**Figure** **:** Segment Profile Plot

It is notable that we have a change in cluster memberships between Hierarchical clustering and using flexclust. Table () below compares the cluster assignments between the two techniques. Except for the first two clusters, there are significant shifts in cluster memberships.

Number of Observations per Cluster

| **Cluster** | **Points\_HC** | **Points\_Flextable** |
| --- | --- | --- |
| Cluster 1 | 59 | 24 |
| Cluster 2 | 8 | 24 |
| Cluster 3 | 52 | 21 |
| Cluster 4 | 17 | 42 |
| Cluster 5 | 35 | 48 |
| Cluster 6 | 29 | 41 |

## QUESTION 6: Concordance of Hierarchichal and KCCA Procedure

In this section, I examine the degree of overlap between the clustering results for the two clustering methods. There is low overlap between the techniques. For instance, in cluster 1, the two place only four (4) points in the same cluster. Only clusters 5 shows a higher degree of agreement between the two methods.

## office\_6\_clust  
## 1 2 3 4 5 6  
## 1 4 2 9 1 0 8  
## 2 0 1 9 7 0 7  
## 3 0 2 12 3 1 3  
## 4 3 2 21 4 3 9  
## 5 27 1 0 0 20 0  
## 6 25 0 1 2 11 2

## QUESTION 7: Why I Would Not go with the 6 Cluster Solution

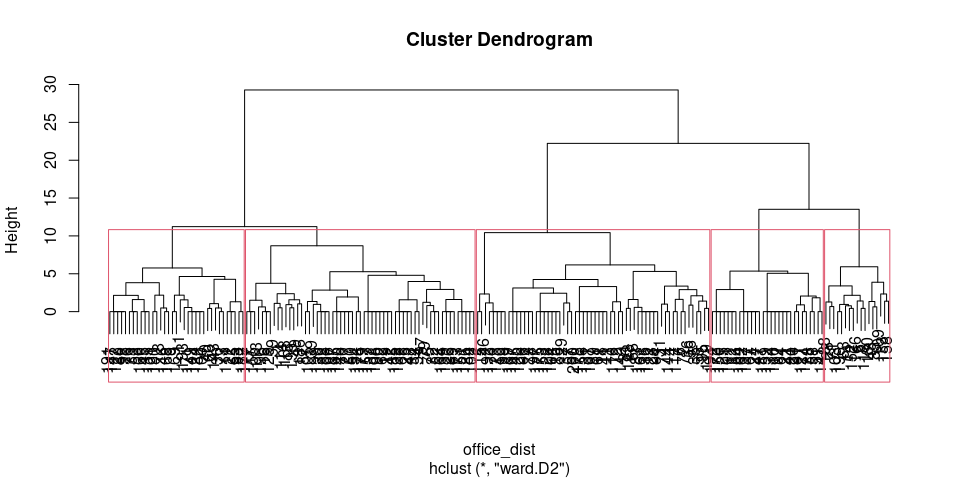
From the segment profile plot, some clusters appear very similar to each other. For instance, cluster 1 and 2 seem to have a set of similar preferences- favourable return policy, good service quality, and low prices. Furthermore, the variables variety of choice, electronics and furniture appear to follow a similar pattern. It would be better to have the two clusters combined.

There are techniques to estimate the optimal number of clusters although none is perfect. In this case, I use the NbClust package to determine the optimal number of clusters.

The technique recommends an optimal cluster size of between 3 and 4 based on the elbow method. The code chunk in line 392 shows this technique of determining optimal clusters.

## QUESTION 8: Implementing a 5 Cluster Solution

In this section, I implement the 5 cluster solution.



**Figure** **:** Dendogram for 5 Clusters

## QUESTION 9: Cluster Summary Statistics

I calculate the cluster means for each of the 5 clusters.

variable Means for the 5 Cluster Hierarchical Clustering Solution

| **cluster5** | **variety\_of\_choice** | **electronics** | **furniture** | **quality\_of\_service** | **low\_prices** | **return\_policy** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 9.8 | 4.4 | 6.0 | 2.6 | 3.4 | 2.7 |
| 2 | 6.9 | 2.9 | 1.2 | 3.1 | 8.2 | 5.7 |
| 3 | 4.9 | 6.5 | 2.1 | 3.7 | 2.8 | 6.8 |
| 4 | 8.2 | 6.7 | 4.4 | 1.9 | 4.5 | 3.5 |
| 5 | 5.2 | 3.7 | 1.3 | 8.3 | 2.1 | 3.6 |

This section shows clustering using the flexclust package using 5 clusters.

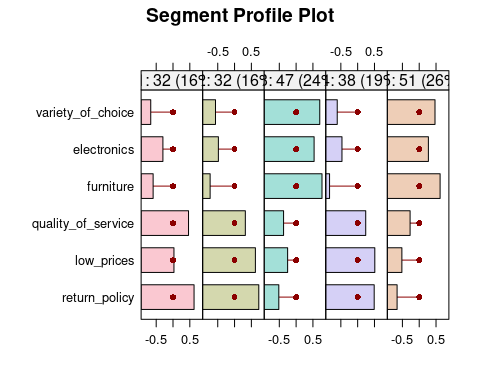
## kcca object of family 'ejaccard'   
##   
## call:  
## kcca(x = office\_cl, k = 5, family = kccaFamily("ejaccard"), control = new("flexclustControl"))  
##   
## cluster sizes:  
##   
## 1 2 3 4 5   
## 31 32 57 34 46

The cluster assignments in this case are as follows;

Number of Observations per Cluster

| **Cluster** | **Points** |
| --- | --- |
| Cluster 1 | 31 |
| Cluster 2 | 32 |
| Cluster 3 | 57 |
| Cluster 4 | 34 |
| Cluster 5 | 46 |

The segment profile plot gives a business friendly way to describe the clustering. In our case, there appears to be distinct preferences for each of the clusters.



**Figure** **:** Segment Profile Plot

I attach expressive labels to each of the clusters in Table () below.

Number of Observations per Cluster

| **Cluster** | **Top\_3\_Preferences** | **Tag** |
| --- | --- | --- |
| Cluster 1 | Return Policy, Quality of Service, & Low Prices | Return and Service |
| Cluster 2 | Return Policy, Low Prices, & Quality of Service | Return and prices |
| Cluster 3 | Furniture, Variety of Choice & Electronics | Furniture & Electronics Nerds Favouring Variety of Choice at Low Prices |
| Cluster 4 | Low Prices, Return Policy & Quality of Service | Price Sensitive Customers favoring Good Return Policy |
| Cluster 5 | Furniture, Variety of Choice & Electronics | Furniture & Eletronics Nerds Favouring Variety of Choice with Good Service |

## QUESTION 10 Why Prefer the 5 Cluster Solution

Cluster 5 distinguishes the clusters better than the six cluster solution going by the segment profile plot. While cluster 3 and 5 appear similar, customers in cluster 3 favour low prices while those in cluster 5 are keen on the quality of services. The clustering algorithm distinctly describes all the other clusters.

Next I check the concordance between hierarchical clustering and the flexclust technique.

## office\_5\_clust  
## 1 2 3 4 5  
## 1 1 14 6 0 10  
## 2 0 19 6 2 5  
## 3 32 1 0 22 2  
## 4 0 25 3 0 6  
## 5 26 1 2 11 6

There are 26 data points in common between the hierarchical clustering and the flexclust technique.

In this section I will incorporate all the other variables, professional, age and income to evaluate the usefulness of the 5 cluster solution. First, I group the data by cluster and compute the mean of age and income. As table () shows, the clustering appears to clearly differentiate customers by age and income. For instance, customers in profile 1 have an average age of 29.5 years and with a mean income of 26. By contrast, cluster 5 consists of customers with a mean age of 46 years and an income of 54.8.

Table () shows the distribution of professionals versus non-professionals by cluster. The bulk of of professionals (48) fall in cluster 1 while most non-professionals fall into cluster 2 (57). Likewise, most high income , relatively old people are in cluster 5. Low income youngsters are in cluster 2.

Summary per Cluster

| **cluster5** | **income** | **age** |
| --- | --- | --- |
| 1 | 26 | 30 |
| 2 | 19 | 26 |
| 3 | 46 | 39 |
| 4 | 40 | 35 |
| 5 | 55 | 46 |

Summary per Cluster

| **cluster5** | **professional** | **n** |
| --- | --- | --- |
| 1 | non-professional | 48 |
| 1 | professional | 11 |
| 2 | non-professional | 57 |
| 2 | professional | 3 |
| 3 | non-professional | 5 |
| 3 | professional | 12 |
| 4 | non-professional | 17 |
| 4 | professional | 18 |
| 5 | non-professional | 5 |
| 5 | professional | 24 |

## QUESTION 11: Targeting Each Cluster

We can target the segments as follows;

Cluster 1: Promote mainly electronics equipment with favourable return policies and quality services. This means that the sales to these customers should have suitable warranties.

Cluster 2: Again, we should target these customers with suitable return policies but, unlike cluster 1, favourable prices.

Cluster 3: Target these customers a wide variety of choices of furniture and electronics.

Cluster 4: For this cluster, we should put equal emphasis on return policy and low prices of electronics.

Cluster 5: These customers prefer a wide variety of furniture and to a lesser extent electronics with good quality service.

## QUESTION 12 K Means Clustering

Finally, I run the kmeans clustering using 5 centres. Here, we notice that there are 60 points in cluster 1, 17 in cluster 2, 33 in cluster 3, 29 in cluster 4, and 61 in cluster 5.

##   
## 1 2 3 4 5   
## 60 17 33 29 61

## office\_5\_clust  
## 1 2 3 4 5   
## 59 60 17 35 29

## QUESTION 13 Corcordance K Means and HC

I then check for the concordance between kmeans and hierarchical clustering and compute the hit rate.

## office\_5\_clust  
## 1 2 3 4 5  
## 1 0 60 0 0 0  
## 2 0 0 17 0 0  
## 3 0 0 0 33 0  
## 4 0 0 0 0 29  
## 5 59 0 0 2 0

There appears to be little concordance between the two techniques.