Assignment 3

Xhesina Hita, Paolo Totaro

2023-10-18

Problem 1: Principle Component Analysis

Analysis

The data set we will be working with in this report contains data from a web-survey conducted on the customers of paper product manufacturer industries. The data set contains 100 observations in 10 variables whose values range from 0 to 10 with 10 being "Excellent" and 0 being "Poor" .

```
              x6          x7           x8          x9          x10          x12
x6   1.000000000 -0.10329258  0.147813122  0.10124768 -0.025694054 -0.123869779
x7  -0.103292581  1.00000000 -0.027747324  0.11081668  0.444693992  0.757603714
x8   0.147813122 -0.02774732  1.000000000  0.13455891  0.003875404  0.002405603
x9   0.101247681  0.11081668  0.134558909  1.00000000  0.149695724  0.257631983
x10 -0.025694054  0.44469399  0.003875404  0.14969572  1.000000000  0.571399283
x12 -0.123869779  0.75760371  0.002405603  0.25763198  0.571399283  1.000000000
x13 -0.446535909  0.24347997 -0.318881554 -0.07408847  0.122810080  0.265074596
x14  0.127911571  0.01509883  0.788384706  0.19491036  0.063988922  0.080360652
x16  0.083927180  0.09344305  0.062924610  0.76828328  0.130652777  0.168732954
x18 -0.002511311  0.15557718  0.062492710  0.84506718  0.228402208  0.317472842
            x13         x14         x16          x18
x6  -0.44653591  0.12791157  0.08392718 -0.002511311
x7   0.24347997  0.01509883  0.09344305  0.155577177
x8  -0.31888155  0.78838471  0.06292461  0.062492710
x9  -0.07408847  0.19491036  0.76828328  0.845067182
x10  0.12281008  0.06398892  0.13065278  0.228402208
x12  0.26507460  0.08036065  0.16873295  0.317472842
x13  1.00000000 -0.30925577 -0.08437362  0.026296616
x14 -0.30925577  1.00000000  0.19077979  0.181460278
x16 -0.08437362  0.19077979  1.00000000  0.744182409
x18  0.02629662  0.18146028  0.74418241  1.000000000
```

Table 1: Correlation matrix of the data set variables

From the correlation matrix we see that some variables are highly correlated among each other:

- *Complain Resolution* (x9) and *Delivery Speed* (x18) with a correlation of 0.84
- *Order & Billing* (x16) and *Delivery Speed* (x18) with a correlation of 0.744
- *Technical Support* (x8) and *Warranty & Claims* (x14) with a correlation of 0.788
- *Complain Resolution* (x9) and *Order & Billing* (x16) with a correlation of of 0.768

- *Salesforce Image* (x12) and *E-Commerce Activities* (x7) with a correlation of 0.75

If we were to apply a multiple linear regression model we would encounter a multicollinearity problem, PCA instead allows us to summarize these variables with a smaller number of representative variables that collectively explain most of the variability in the original set.

Since the pairs of variables above are positively correlated we expect that in the biplot the variables in question will be located close to each other (e.g. customers who give a high rating to *Complain Resolution* tend to give high ratings to *Delivery Speed* as well). This means that in the first PC those variables have similar positive loadings.

We now perform Principal Component Analysis (PCA) on the 10 variables, we standardize the variables in order to get standard deviation equal to 1, note however that standardization is not mandatory since the variables have the same scale (0-10).

In order to determine the number of principal components without using the percentage of explained variance we use the loadings of the principal components which indicate how much each variable contributes to the correspondent principal component. Therefore we want to summarize the set of correlated variables with a smaller number of representative variables (components) that collectively explain most of the variability in the original set.

However, since the loadings in the first principal component are all negative we decide to change the sign for both the loadings and the scores for a better interpretation.

```
           PC1         PC2         PC3         PC4           PC5         PC6
x6  0.02165981 -0.32049650  0.04860871 -0.72939820  0.2398272806 -0.54731012
x7  0.27700321  0.36538591  0.33555404 -0.15960316  0.4869520084  0.23674190
x8  0.13541239 -0.40245687  0.46445831  0.30906779  0.0435811821 -0.15215803
x9  0.47396591 -0.14508598 -0.30520421  0.02695378  0.0572019607 -0.02217456
x10 0.28406630  0.25715964  0.28800178 -0.25880735 -0.8016204219 -0.13504817
x12 0.36045162  0.35061641  0.30470839 -0.11139514  0.1905807332  0.09046645
x13 0.01081700  0.46513401 -0.09999625  0.40264209  0.1368252328 -0.76145685
x14 0.20550318 -0.38653824  0.43386484  0.31114789  0.0003407761 -0.09799111
x16 0.43769822 -0.14728819 -0.33605322  0.01971376  0.0170076781  0.06221520
x18 0.48504901 -0.06508023 -0.29867643  0.08685204 -0.0475646414 -0.03239473
             PC7          PC8          PC9        PC10
x6  -0.005880163 -0.052463100  0.021153757  0.04613444
x7  -0.256111849  0.324337914  0.430093484 -0.05248734
x8   0.125388000  0.560010299 -0.271633608  0.28333309
x9   0.343802973  0.252329193 -0.055811050 -0.68679044
x10 -0.105612238  0.142355036  0.081472454 -0.06602286
x12  0.311386364 -0.411341934 -0.566991915  0.09868132
x13 -0.104049275  0.021306299  0.006703732 -0.03937778
x14 -0.191301250 -0.557018423  0.298684962 -0.27861111
x16 -0.725585471 -0.008303118 -0.348915354  0.14477888
x18  0.344147354 -0.120463345  0.445595399  0.57352005
```

We now want to pick the number of principal components which explain most of the variability of our data set, in order to do this we take into consideration the loading values in each PC above 0.35 (a subjective threshold we consider significantly high).

The first loading vector places weight on x9, x12, x16 and x18. On the other hand, the second loading vector places weight on x7, x8 x12, x13 and x14. The third places weight on x8 and x14, the fourth on x6 and x13 and the fifth on x7 and x10. In the first five principal components, our 10 variables all contribute with large weight (>0.35) to the overall variability. Therefore after checking the loading values on the other PC we conclude that the last five principal components do not add significant insight to our analysis since they explain information that is already explained.

In the first principal component we have all positive loading values, therefore this indicates a positive relationship between variables and PC. A high positive loading value suggests that as the PC score increases the variables tend to increase as well and vice versa for high negative loading values.

In the first PC we notice that x9, x16 and x18 have similar loading values, this suggests that they will contribute with the same weight to the variability captured by the first principal component only. This is a result that we expected since these variables are correlated.

From the significantly high loading values on a same principal component we can capture hidden information.

For instance, the first principal component has high loadings on Complaint Resolution (x9), Salesforce Image (x12), Order & Billing (x16) and Delivery Speed (x18) and can be therefore interpreted roughly as a measure of the overall satisfaction and of the positive experience of a customer toward the customer related services of a company.

In the second principal component we have high loading values in the following variables: Competitive Pricing (x13), Warranty & Claims (x14), Technical Support (x8), Product Quality (x6) and E-Commerce Activity (x7). In general we find that the variables mentioned above can be interpreted as a measure of the product quality and service offered by the company.

In the third principal component the high loading values are found in x8 and x14, both are related to customer support and after-sales service.

In the fourth principal component the high loading values are found in x6 and x13, variables related to the product quality and price.

In the fifth principal component the variables with high loading values are E-Commerce Activity (x7) and Advertising (x10) which are related mostly to digital marketing (in contexts where the advertisement process is carried out online).

Back to the correlation values we observe that the variables that are correlated to one another have some patterns in the loadings. The variables x9 and x18 are correlated and in the first PC they have similar loading values, same goes for x16 and x18. If we consider the variables x8 and x14, they have similar loading values in the second and third PC. The correlated variables x9 and x16 have similar loadings in PC1, PC2 and PC3.

All the information above can be easily observed in the biplot (*Figure 1*):

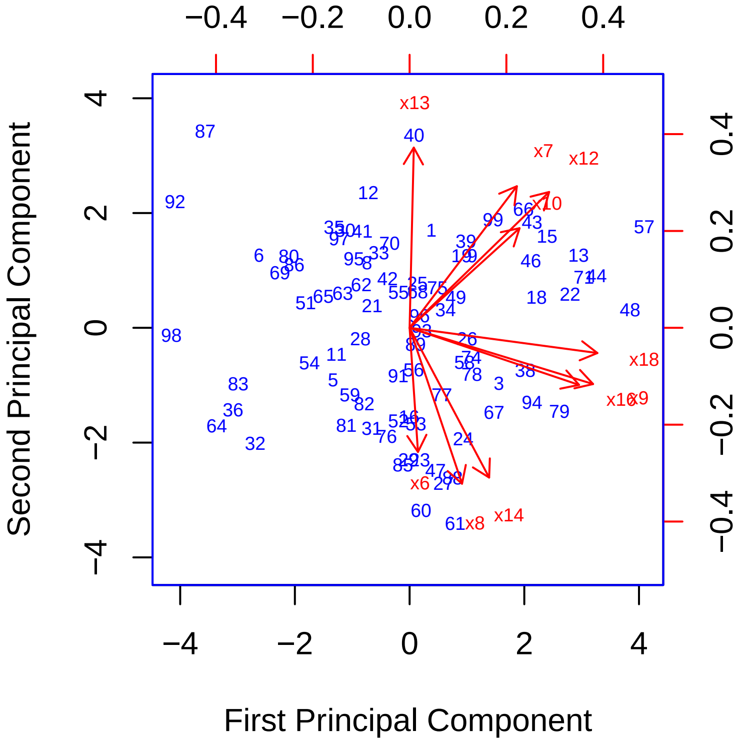

Figure 1: Biplot of the first and second principal components

We can see from *Figure 1* (biplot) that all the arrows lay in the right half of the plot since the loading values of the first PC are all positive, while for the second PC they lay both on the top and on the lower part since the loadings are both positive and negative.

We observe that the arrows corresponding to the variables x9, x16 and x18 are very close to each other, this is not surprising since as we mentioned above they are correlated and have similar loading values both for the first and the second principal component. Same goes for some other variables such as x7, x12, x10 and x8, x14.

The variables x6 and x13 have opposite loading sign in the second PC and indeed the arrows in the biplot point in opposite directions. These two variables also have different arrow lengths, this indicates the magnitude with which a variable has an impact in explaining the variation in the data. Indeed, in the second PC their loading values in absolute values are 0.32 for x6 and 0.46 for x13.

Data points that are closer to the direction of an arrow are positively associated with the corresponding variable. Oppositely, data points pointing away from the arrow are negatively associated with the corresponding variable. If we consider the observations 57 and 48, they have a large positive score on the first principal component, therefore they correspond to customers who assigned high positive feedback to the customer related services of a company.

Instead, observations 98, 92 and 87, which have negative scores on the first principal component, correspond to customers who aren't satisfied with the customer related services offered by the company.

The values 87 and 40 have positive scores for the second principal component which means that these two customers were satisfied in terms of pricing (x13) but not of product quality (x6).

Observations which are close to the origin of the arrows such as 89 and 96 have approximately average satisfaction levels for all the variables in both the first two principal components.

As mentioned above, we usually are not interested in all of the principal components, we would like to use just few of them in order to visualize and interpret the data. When computing the percentages of variance explained by each principal component we obtain that:

- The 1st PC explains 30.51%
- The 2nd PC explains 23.45%
- The 3rd PC explains 16.30%
- The 4th PC explains 10.73%
- The 5th PC explains 5.81%

The remaining five PCs explain even less variability and we don't include them since we decide to take only the first five PCs which explain in total 86.80% of the variability (we think that a percentage of explained variability of over 85% is a good amount).

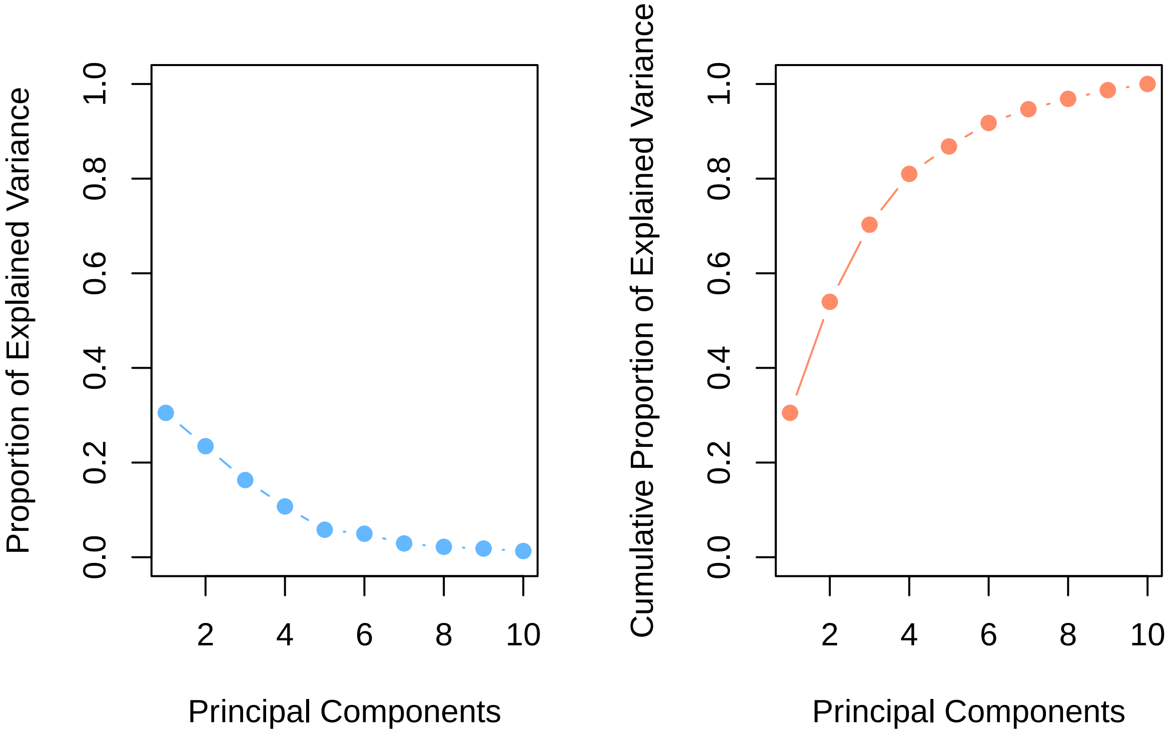

Figure 2: Scree plots of the Proportion of Explained Variance (left) and Cumulative Proportion of Explained Variance (right)

If we look at the **scree plot of cumulative proportion of variance explained** (another method to define the number of PC to keep) in *Figure 2 (right)* we notice that more than 85% of the variability is explained by the first 5 principal components and that after that the cumulative proportion of explained variability stabilizes. By looking at *Figure 2 (left)* we notice that in the Proportion of Explained Variance scree plot there is an elbow point in correspondence of the fifth principal component, therefore we keep the first five PCs.

After this analysis we are left with 5 principal components that explain most of the variability in the data. Those can be used for low-dimensional representation of the data, for data visualization, or can be used in supervised learning problems.

Problem 2: Cluster Analysis

Hierarchical Cluster Analysis

In this second part of the assignment we try to classify customers of a women clothing shop from a new data set using cluster analysis. The data set we will be analyzing is obtained from a survey of purchasing behavior of 275 women asked about their attitudes regarding choice of clothing store. They answered to 18 questions by assigning a value from 1 to 5, 1="Not at all important" 5="Very important".

We were asked to define 3 groups of typical customers of a clothing store for women and we believe that the 3 main groups are:

- Fashion addicted
- On-a-budget shoppers
- Occasional shopper

In order to distinguish as much as possible the 3 groups we defined above we chose 7 variables:

- x6: 'that the store is highly fashion orientated', again for the fashion customers
- x7: 'that the store keeps high prices', we believe high prices influence the choice of customers who don't want to spend a fortune
- x8: 'that the store has a lot to choose from', we think that a shop with a lot to choose from might suit fashion customers and occasional customers
- x11: 'that there are often sales in the store', occasional and on a budget customers can be very interested in sales
- x12: 'that the store often has attractive ads', is an important factor for fashion addicted and occasional customers who go shopping once in a while
- x13: 'that in the store you can always find news', fashion customers are always looking for new clothes
- x15: 'that this is a store of price aware customers' fits mostly cheap customers

Therefore, we reduce the data set according to the 7 chosen variables.

We now perform clustering using four hierarchical clustering methods (Single, Complete, Average Linkage and Ward's Mehtod) and we plot their dendrograms.

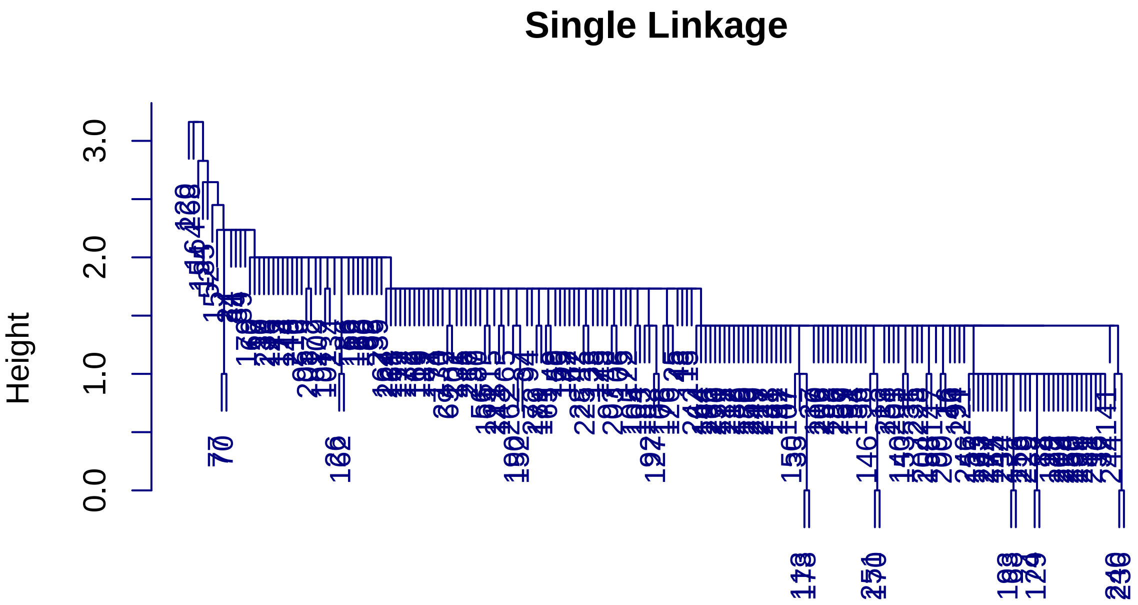

Figure 3: Single Linkage clustering Dendrogram

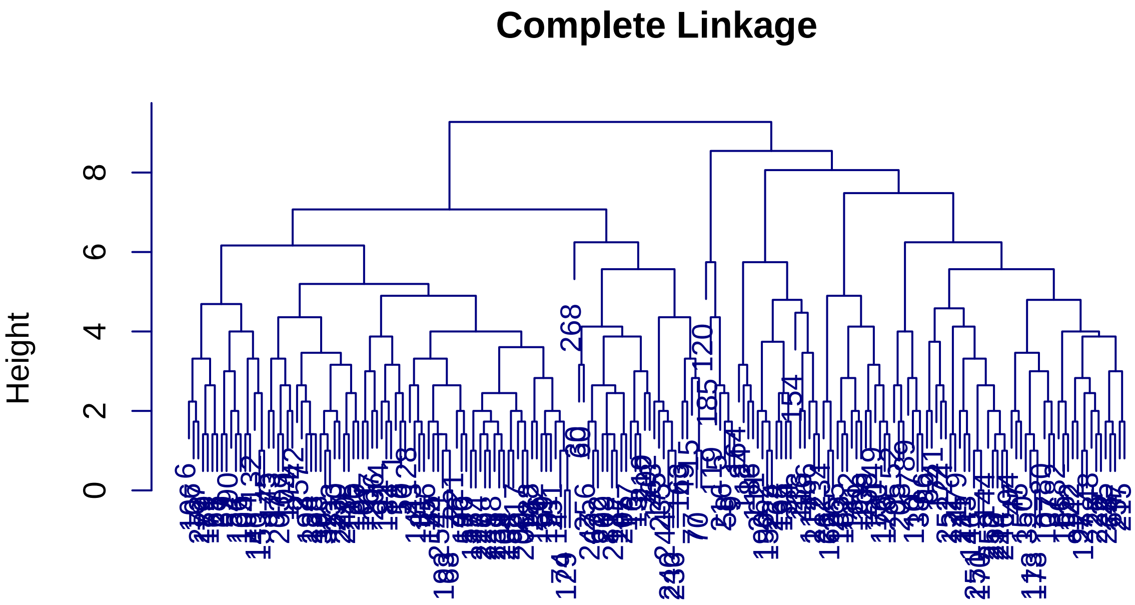

Figure 4: Complete Linkage clustering Dendrogram

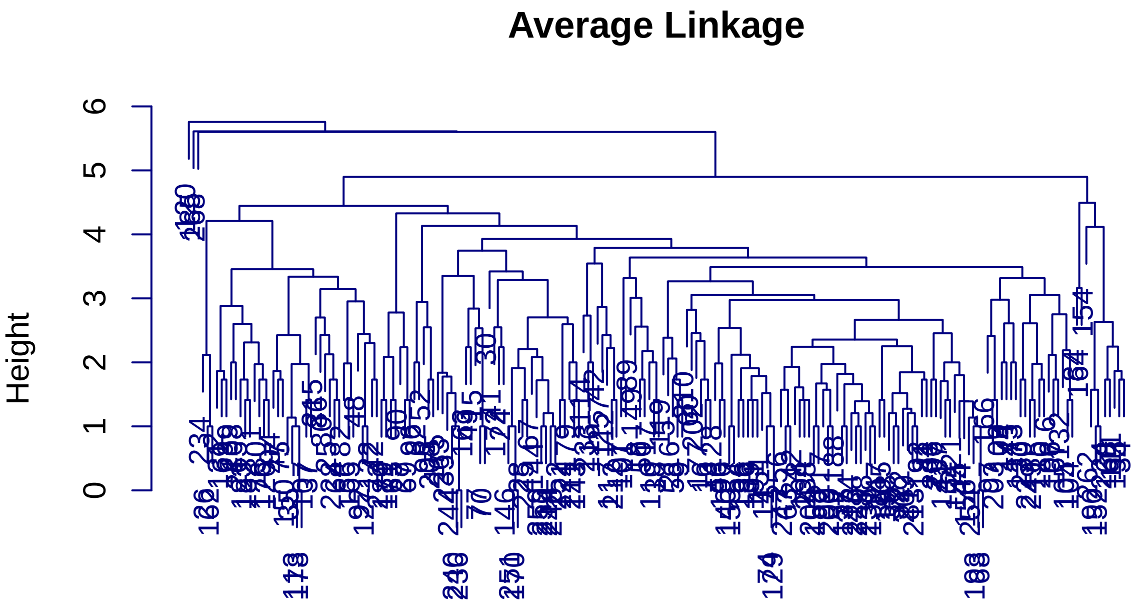

Figure 5: Average Linkage clustering Dendrogram

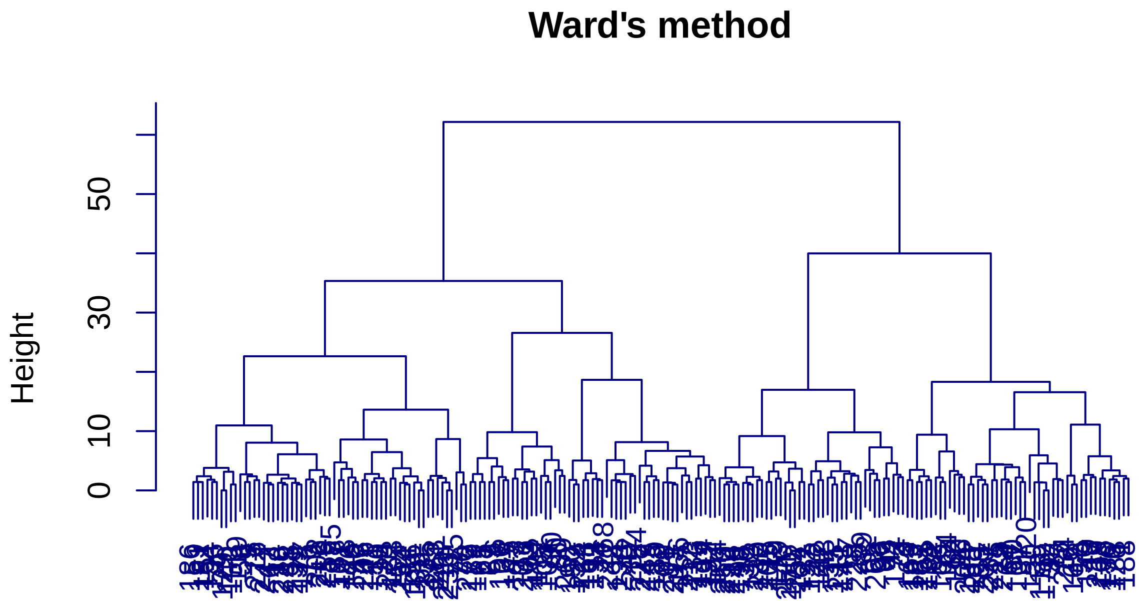

Figure 6: Ward's Linkage clustering Dendrogram

Dendrograms are used for cluster representations, they help in visualizing groups of items according to their similarities.

Each leaf of the dendrogram represents one observation. However, as we move up the tree, some leaves begin to group into branches. These groups correspond to observations that are similar to each other. As we move higher up the tree, branches themselves keep fusing with each other.

The lower we are in the tree the more similar the groups of observations are to each other.

On the other hand, observations that fuse near the top of the tree can be quite different. The height of this fusion indicates how different the two observations are.

The information above come in handy when choosing the number of clusters for our model. Looking at the dendrograms obtained from single and average linkage we notice how hard it is to distinguish groups of observations. Single linkage indeed can often result in extended, trailing clusters in which observations are fused one-at-a-time and our case is no exception. Average linkage as well is not a suitable clustering method for this data set, wherever we draw an horizontal line on the dendrogram we will end up with a very unbalanced dendrogram which doesn't fulfill our needs.

It should be noted that the choice of clusters also depends on the context and the data. In our case we want to distinguish clusters of women by referring to their preferences when it comes to choosing a clothing shop. By looking at the dendrograms of Complete linkage and Ward's method we believe that a number of two or four clusters is good.

Moreover, let's check how many observations are in each cluster for each of the four methods we used.

Table 1: Number of observations in each cluster using Ward's method

Clusters	Observations
1st	53
2nd	40
3rd	59
4th	48

Table 2: Number of observations in each cluster using Complete linkage

Clusters	Observations
1st	110
2nd	90

Table 3: Number of observations in each cluster using Average linkage

Clusters	Observations
1st	1
2nd	197
3rd	1
4th	1

Table 4: Number of observations in each cluster using Single linkage

Clusters	Observations
1st	1
2nd	197
3rd	1
4th	1

From *Table 1* and *Table 2* we notice how the two clustering methods, Ward’s and Complete linkage, produce balanced clusters (similar number of observations in each cluster); whereas in *Table 3* and *Table 4*, Average and Single linkage produce very unbalanced clusters. Note that even if we chose 2 as number of clusters in both Average and Single linkage we would have still obtained unbalanced clusters with 1 observation in one cluster and 199 in the other.

It is therefore reasonable, in our case, to prefer Ward’s method and Complete linkage over Single and Average linkage.

We now perform a profiling of our clusters in order to gain information regarding groups of customers for the clothing shop.

```
Ward’s method
```

```
  Group.1       x6       x7       x8     x11      x12      x13      x15
1       1 2.754717 1.452830 4.132075 4.00000 3.490566 3.207547 3.566038
2       2 3.500000 1.325000 3.800000 2.30000 2.425000 4.025000 2.125000
3       3 3.457627 3.016949 4.084746 3.40678 2.644068 3.711864 3.389831
4       4 2.312500 2.000000 3.395833 1.81250 1.541667 2.500000 3.312500
```

By looking at the mean for each variable in each cluster we can come up with some interpretations.

If we consider Ward's method, we notice that people who belong to the first group are subjects **more interested in the price**. For instance they believe that it is important for a clothing shop to have Sales (x11) and a lot to choose from (x8), in stores with a lot of things you are more likely to find cheaper stuff.

In the second group women are **very interested in fashion**, they like having new clothes and don't care about the price. (High value in x6 and x13, but low value in x7 and x15).

In the third group there are **people that like fashion but they are also sensitive to the price**, they prefer big stores where you can find more things and with different prices. (high x6, x7,x8,x11,x13,x15).

In the fourth group we can distinguish people that are not attracted by advertising and sales in the store (low value for x12 and for x11). It seems **they are occasional customers**.

```
Complete Linkage
```

```
  Group.1       x6       x7       x8      x11      x12      x13      x15
1       1 3.300000 2.227273 4.154545 3.418182 3.072727 3.718182 3.327273
2       2 2.644444 1.766667 3.533333 2.400000 1.933333 2.900000 2.966667
```

Ragarding Complete linkage clustering, where we chose a number of two clusters, we observe that most of the variables have the same mean so there isn't a clear difference between the two groups. The only difference we can observe is that the variables x11, x12 and x13 have different values between the two groups. In the first group subjects believe that attractive advertising, sales and new collections in the store are very important (high values), whereas in the second group the opinion about it is quite the opposite (low values). Anyway the clustering performance using Complete linkage is not efficient as the Ward's method one.

After these considerations we believe that the correct number of clusters for this kind of analysis is four (we can distinguish four different groups of customers).

We started the analysis by defining three groups of customers, *fashion addicted*, *on-a-budget shoppers* and *occasional shoppers*. When applying the clustering methods above we discovered that it is hard to distinguish two groups of observations from our data set but a good choice is four clusters, *on-a-budget shoppers*, *fashion addicted on-a-budget*, *fashion addicted NOT on-a-budget* and *occasional shoppers*. We could actually group them in three clusters only by merging together *fashion addicted on-a-budget* and *fashion addicted NOT on-a-budget*.

K-means Cluster Analysis

In this second part we perform a K-means cluster procedure with K=4 clusters as we believe that with four we can optimize the distinction among different clusters.

Since the K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial (random) cluster assignment of each observation, for this reason we run the algorithm 25 times from different random initial configurations.

The presented result is the best among those 25 runs.

```
  Group.1       x6       x7       x8      x11      x12      x13      x15
1       1 3.555556 2.611111 4.555556 3.583333 3.722222 4.083333 4.361111
2       2 3.813559 2.033898 4.016949 2.830508 2.474576 4.118644 2.101695
3       3 2.416667 1.812500 3.291667 1.562500 1.520833 2.604167 3.375000
4       4 2.315789 1.807018 3.789474 3.877193 2.789474 2.719298 3.333333
```

After performing the K-means algorithm to determine the four clusters we look at the mean of each group and see if they are different from the ones obtained using Ward's method.

Just like in Ward's method (group 3), for K-means clustering in group 1 we find *fashion addicted on-a-budget* customers. These subjects are interested both in low prices and in new collections, stores with a lot of stuff to choose from and attractive ads.

Group 2 is made up of people that are very interested in fashion and in having always something new, they don't care about the price, basically *fashion addicted* (Group 2 in Ward's method).

Group 3 is composed by customers who believe that prices are important and who like big stores. They are not attracted by advertising and sales in the store and can be defined as *occasional shoppers.* They are equal to customers in group 4 of Ward's method.

In Group 4 we find *on-a-budget shoppers* who are more interested in prices and sales like the ones in Group 1 of Ward's method.

By analyzing both methods we obtain the same clusters of people but they change places. There are costumers that are interested only in the price and occasional costumers. Then we have also two groups that both like fashion a lot, in this group we can distinguish people that like fashion but also are sensitive to the price and people that like fashion and don't care about the price.

Just like we mentioned above we can also choose to use three number of clusters, merging *fashion addicted on-a-budget* and *fashion addicted NOT on-a-budget.*

One way to use these information is market segmentation, i.e. identifying subgroups of people and understanding what they are most interested in. For example, if we were faced with people who loved fashion then we would increase advertising on the new garments and send it to them. On the other hand, if a store had some old collection clothes, they could be put on sales fitting customers who don't care about fashion and are on-a-budget.