Assignment 1

Khadija Ramzi & Mohit Gupta

April 2023

### Question 1

Code related to first answer in the grey area below

setwd("/Users/khadija/Desktop/MSBA excel files/QUARTER 4 Spring/Group Assignment")  
survey <- read.csv('assg1.csv')  
library(psych)  
survey.sc <- survey  
survey.sc[,2:6] <- scale(survey.sc[,2:6])  
describe(survey.sc)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## Resp 1 30 15.5 8.8 15.50 15.50 11.12 1.00 30.00 29.00 0.00 -1.32  
## X1 2 30 0.0 1.0 0.20 0.02 1.01 -1.50 1.56 3.06 -0.18 -1.53  
## X2 3 30 0.0 1.0 0.28 0.05 0.97 -1.68 1.26 2.94 -0.46 -1.18  
## X3 4 30 0.0 1.0 -0.24 -0.06 1.08 -1.34 1.94 3.28 0.49 -0.94  
## X4 5 30 0.0 1.0 0.21 0.06 1.02 -1.86 1.24 3.10 -0.48 -1.17  
## X5 6 30 0.0 1.0 -0.19 -0.01 1.04 -1.59 1.57 3.16 0.19 -1.19  
## se  
## Resp 1.61  
## X1 0.18  
## X2 0.18  
## X3 0.18  
## X4 0.18  
## X5 0.18

To ensure comparability across individuals and samples, it is crucial to scale the data. This involves applying a z-scoring procedure to each ratings variable, which means subtracting the variable’s mean and dividing by its standard deviation for each observation. First, we create a new data frame called ‘survey.sc’ and copy survey into it. Then we modify scale.sc by scaling columns 2 through 6. To validate that the scaling was done correctly, we use the describe command. With this all means become zero and standard deviation is one.

### 

### Question 2

Write the code related to second answer within the grey area below

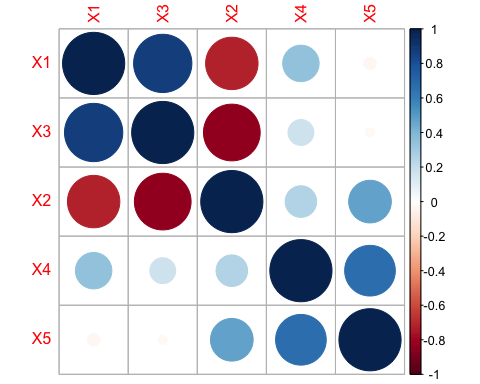
cor(survey.sc[,2:6])

## X1 X2 X3 X4 X5  
## X1 1.00000000 -0.7107386 0.87895393 0.3432966 -0.03865034  
## X2 -0.71073859 1.0000000 -0.83182655 0.2573651 0.47015073  
## X3 0.87895393 -0.8318265 1.00000000 0.1727462 -0.02056604  
## X4 0.34329663 0.2573651 0.17274619 1.0000000 0.66421386  
## X5 -0.03865034 0.4701507 -0.02056604 0.6642139 1.00000000

library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(survey.sc[,2:6]), order = 'hclust')



In the above matrix we see that the variables have a fair amount of correlation. A few common themes can be seen like X1 ‘I never go back to a store that had treated me with disrespect’ can be related with X2 ‘I find salespeople who fawn over me irritating, I just expect them to courteous’ and less likely with X4 ‘I care little for the fancy displays in departmental stores.’ We can conclude by saying that people are more likely to go to the discount store which is welcoming & if the salesperson is polite and understanding and things like display or overly friendly attitude have little impact for the people visiting the store.

### 

### Question 3

survey.pc = prcomp(survey.sc[,2:6])  
summary(survey.pc)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.6345 1.3610 0.58478 0.30601 0.20148  
## Proportion of Variance 0.5343 0.3704 0.06839 0.01873 0.00812  
## Cumulative Proportion 0.5343 0.9048 0.97315 0.99188 1.00000

PCA helps us map our initial set of four variables onto a new four-dimensional space, with each dimension being orthogonal. The second row in the above presented table indicates the amount of variance each component accounts for in the data, ranked in descending order. The first component explains 53.43% of the variance, followed by the second component, which explains 37.04%, and so on. The third row shows the cumulative variance. In our case, the first and second components account for 90.47% of the variance. This implies that we can decrease the dimensionality of the original set of four variables to two and still maintain 90.47% of the explanatory power. However, the table above does not specify how the PCA-generated factors correspond to the initial attributes.

### Question 4

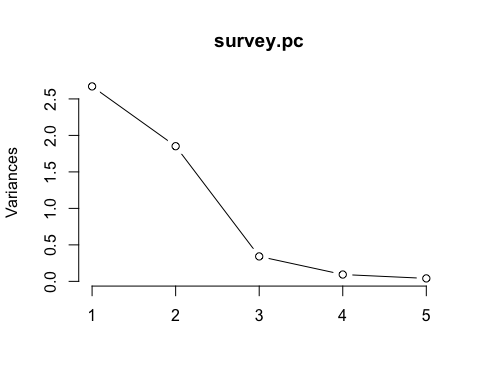
survey.pc$rotation

## PC1 PC2 PC3 PC4 PC5  
## X1 -0.55491712 0.2324895 -0.2795639 -0.7234233 0.19109786  
## X2 0.57063147 0.2117721 -0.1832902 -0.4621638 -0.61833777  
## X3 -0.57964069 0.1567520 0.3106045 0.1866707 -0.71282845  
## X4 -0.01698458 0.6859778 -0.5610983 0.4611287 0.04092556  
## X5 0.17370927 0.6371534 0.6906101 -0.1248104 0.26709683

This tell us the exact relationship between each factor and the original variables. The first two components (that explain of the variance in the data can be expressed as:

### Question 5

plot(survey.pc, type = 'l')



#keep the variables over 1 i.e. PC1 and PC22

The common criterion is to keep factors that are above the “kink” in the scree-plot. The second criteria often used is to keep factors which are over 1 in the scree plot. In this case, this would equate to keeping two factors i.e. PC1 and PC2. Thus, we can reduce the dimensionality of our data from four to two in this case.

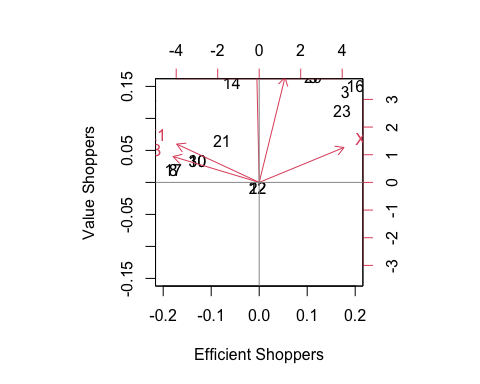
### 

### Question 6

For PC1 we can see a high correlation with X2 at 57.06% It can be seen from the above equations that X1 and X3 have a more significant negative connection with PC1, while X2 has a positive association with PC1. Since PC1 indicates customers who respect courteous service but do not ask numerous inquiries, they can be referred to as “Efficient Shoppers” based on the questions asked. On the other hand, PC2 is positively impacted by all variables, with X4 and X5 having the most impact. They can be called “Value Shoppers” since they favor bargain retailers and are unconcerned with how items are displayed at department stores.

### Question 7

biplot(survey.pc, xlim=c(-.2,.2), xlab = "Efficient Shoppers", ylab = "Value Shoppers", ylim=c(-.15,.15))  
abline(h = 0, v = 0, col = "gray60")



### Question 8

Although the fourth segment of the figure is not well defined and could require additional data to validate, there appear to be four segments in total. X1 and X3 are part of the first segment, X4 and X5 are part of the second segment, and X2 is the only member of the third segment. Samples 24, 1, and 9 appear to be in a different segment at first; however, they most likely belong in the same segments as X1 and X3.

### Question 9

There are four separate client categories based on the plot. There is a significant association between X4 and X5 in the first segment, characterized by low consumption and trimmed spending. The second segment, which is represented by X1 and X3, on the other hand, is highly connected and exhibits traits that are dissimilar from those of PC1. These clients prefer to converse with salespeople and count on them to offer guidance or solutions. Customers in the third sector, represented by X2, are typically less social than those in PC1. Customers in the final phase (11, 28) may not be chatty, but they make significant purchases. They might not be price sensitive and will buy from you again.

### Question 10

The client group (11, 28) that has the potential to be the most profitable for the store is the one that is less clearly defined on the map. These clients are willing to purchase more expensive things because they are not overly price sensitive. Additionally, they don’t require a lot of attention from salesmen, which allows the company to reduce the cost of its sales team. This would allow the shop to direct salesperson resources toward clients needing more help, thereby satisfying a broader clientele.