Project Report 2 - Explorative Analysis

Seda Ismail

1 Introduction

1.1 Research Question:

"Are there distinct patterns of employment-related characteristics in microcensus data, and how can hierarchical cluster analysis (HCA) be used to uncover these patterns by using a combination of metric and categorical variables, including normal weekly work hours, year of the highest education degree, desired total working hours, current work duration, and interruptions of work longer than three months?"

1.2 Description of the background, motivation and aim of the study

Background: The modern work environment are evolved to have diverse work patterns and changing expectations. Microcensus data, which includes variables normal weekly work hours, current work duration, highest education year, desired total working hours, and interruptions in work, provides a rich source of information to explore these patterns. Hierarchical Cluster Analysis (HCA) is a powerful statistical technique that can help reveal hidden structures and group similar observations within this dataset.

Motivation: The motivation behind this study lies in the need to gain insights into the diverse employment-related characteristics that can be found within the microcensus data for the chosen variables. By employing HCA, motivation for the study is to identify distinct clusters or groups of individuals with similar employment profiles. This analysis examines typical patterns of employment characteristics in the Austrian population. The question whether the chosen variables allow to detect a cluster is answered.

Aim of the study: The aim of this study is to apply hierarchical clustering analysis (HCA) to the variables of normal weekly work hours, current work duration, desired total working hours, and interruptions in work longer than 3 months and unveil meaningful patterns and groupings within the microcensus dataset.

2 Data Collection and Description of the Data

2.1 Type of data collection; circumstances of implementation (time period, etc.)

Type of data collection used for collecting microcensus data are summarized in the "atatmeth-Befragungsmethode" variable as follows:

- 1: Cati: includes those households that were completed in the in-house telephone studio
- 2: F2F: those households that were completed by means of personal interviews conducted by field interviewers.
- 3: Income Call:(s) are those interviews that would originally have been planned as F2F, but are nevertheless conducted in the telephone studio because the household calls in.
- 4: Neutraler Ausfall: Only applies if the person is a single-person household and is neither physically nor mentally able to conduct an interview and no one can actually provide information about this household, it will be treated as a neutral failure.
- 5: Selbstausfülle

Data is collected in the year 2012, which is described in the variable in the dataset as "ajahr-Referenzjahr"

2.2 Description of the data set (sample type, size, characteristics, level of measurement (Skalenniveaus) missing values, etc.).

The microcensus data for Austria refers to a specific survey conducted by Statistik Austria to gather detailed information about various aspects of the population and households.

It provides valuable insights into demographic, social, economic, and housing characteristics of the Austrian population for the reference year 2012.

Microcensus dataset has 9287 rows and 216 columns, corresponding to 9287 respondent's answers.

Level of Measurement (Skalenniveaus): The microcensus data set incorporates only metric variables. Where nominal and ordinal variables with categorical information such as gender, marital status, and occupation is coded as integers. Ordinal variables capture ordered categories, for example, educational levels. The rest of the variables, metric variable, are used for quantitative measures like working hours, employment since in months, and household size.

2.3 Specification and description of the variables used

In total five variables from the microcensus data are selected for the study

bsex: Gender (SPSS: Geschlecht) and is a binary categorical variable (nominal) after its transformation of the gender coding into categories. (1:Men, 2:Woman)

xkartab Highest completed Education (nat.Representation)

dstd: Normal weekly work hours (SPSS: Arbeitsstunden in Referenzwoche d19) and is measured on an ratio scale.

dseitz: Current work since (in months) (SPSS: jetzige Arbeit seit (in Monaten)) and is a metric variable measured on a ratio scale.

cdau: Interruption of work longer than 3 months (SPSS: Unterbrechung länger als 3 Monate c6) is a nominal variable.

The rows with -3 value for the variable, mean that the corresponding question was not asked. They are therefore removed from the data set.

Load microcensus dataset

```
microcensus <- read.csv("mz_2012_testdatensatz_070605.csv")
```

Select the variables of study, convert -3's to NA values and remove NA values from the dataset. In the microcensus dataset, all the non-answered values are defined with -3, indicating that the question was not answered.

```
data <- microcensus %>%
  dplyr::select(bsex, xkartab, cdau, dstd, dseitz) %>%
  mutate_all(~ ifelse(. == -3, NA, .))  %>%
  na.omit()
```

The dimension of the dataset of this study is as follows.

```
dim(data)
```

```
## [1] 292   5
```

2.3.1 Handle Missing Values

Calculate the total missing values for each variable and print out the results

```
missing_counts <- colSums(is.na(data))
print(missing_counts)
```

```
##    bsex xkartab    cdau    dstd  dseitz
##       0       0       0       0       0
```

There are no missing values in the data set.

2.3.2 Transform data

View the structure of the dataset

```
str(data)
```

```
## 'data.frame':    292 obs. of  5 variables:
##  $ bsex   : int  1 2 2 1 1 2 1 1 1 2 ...
##  $ xkartab: int  2 1 4 8 1 8 1 4 1 2 ...
##  $ cdau   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ dstd   : num  40 20 38.5 20 40 50 38.5 4 40 20 ...
##  $ dseitz : int  308 35 39 62 31 159 288 4 2 3 ...
##  - attr(*, "na.action")= 'omit' Named int [1:8995] 1 2 3 6 7 8 9 10 11 12 ...
##   ..- attr(*, "names")= chr [1:8995] "1" "2" "3" "6" ...
```

- bsex (int) needs to be transformed to factor, where 1 represents "Male" and 2 represents "Female".
- xkartab (int) needs to be transformed to factor.
- cdau (int) needs to be transformed to factor, where 1 represents "yes" and 2 represents "no".

Transform variable "bsex"

```
data <- data %>%
  mutate(bsex = case_when(
    bsex == 1 ~ "male",
    bsex == 2 ~ "female"
  )) %>%
  mutate(bsex = as.factor(bsex))
```

Transform variable "cdau"

```
data <- data %>%
  mutate(cdau = case_when(
    cdau == 1 ~ "yes",
    cdau == 2 ~ "no"
  )) %>%
  mutate(cdau = as.factor(cdau))
```

Transform variable "xkartab"

```
data <- data %>%
  mutate(xkartab = case_when(
    xkartab <= 3 ~ 1,  # Pflichtschule/keine Pflichtschule, Lehrabschluss (Berufsschule), Berufsbild. m
    xkartab > 3 & xkartab < 7 ~ 2,
    xkartab >= 7 ~ 3
  )) %>%
  mutate(xkartab = case_when(
    xkartab == 1 ~ "Basic",
    xkartab == 2 ~ "Middle",
    xkartab == 3 ~ "High"
  )) %>%
  mutate(xkartab = as.factor(xkartab))
```

Transform variable "dstd"

```
dstd_999_values <- data %>%
  filter(dstd == 999)
dstd_999_values
```

```
##     bsex xkartab cdau dstd dseitz
## 1 female   Basic   no  999    123
## 2   male    High   no  999    179
```

There is only 2 rows where the dstd value is 999. In the following step, remove these rows.

```
data <- data %>%
  filter(dstd != 999)
```

View the summary of dataset

```
summary(data)
```

```
##      bsex       xkartab     cdau          dstd            dseitz
##  female:139   Basic :190   no :287   Min.   :  3.00   Min.   :  0.00
##  male  :151   High  : 59   yes:  3   1st Qu.: 30.00   1st Qu.: 27.25
##               Middle: 41             Median : 40.00   Median : 89.50
##                                      Mean   : 36.72   Mean   :133.61
##                                      3rd Qu.: 40.00   3rd Qu.:207.25
##                                      Max.   :100.00   Max.   :570.00
```

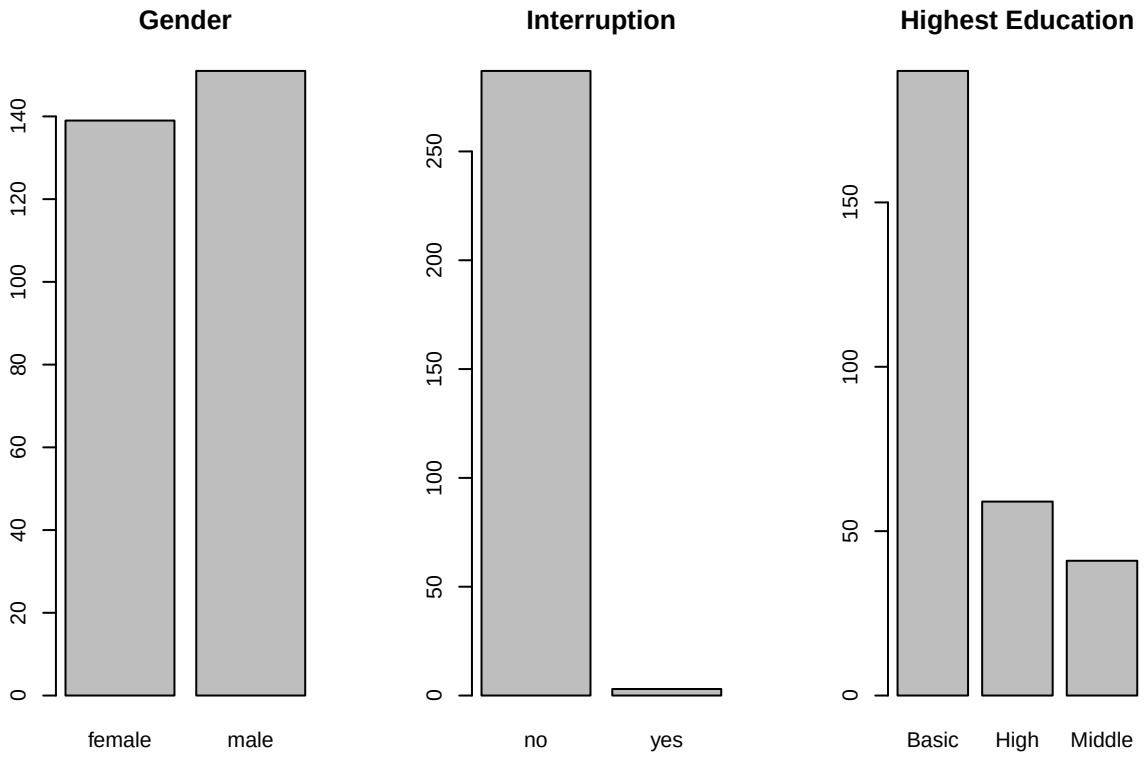
Summary of the data is now looking plausibel and suitable for the later analysis steps.

3 Descriptive Analysis

After data preprocessing, in total there are 290 observations and 5 columns left. In the next section, the descriptive analysis is performed, which allows to discover more hidden details about the variables of study and further may lead to data changes.

Distribution of categorical variables bsex, cdau and highest education

```
par(mfrow=c(1,3))
barplot(table(data$bsex), main="Gender")
barplot(table(data$cdau), main="Interruption")
barplot(table(data$xkartab), main="Highest Education")
```



The data of the study consists of 290 observations, with frequencies of 139 females and 151 males. The distribution of the sample in terms of gender can be considered as balanced, with slightly more percentage of males than females. age category -balt5 has significantly fewer data points in category 60-74 and also in 15-29. The distribution is not uniform across all age groups. The highest proportion is in the 45-59 age group followed by the 30-44. Overall, the sample appears to have a diverse age distribution, with presence of respondents in their pre-retirement years, as well as a smaller representation of younger respondents who are newly entering the labor force. The distribution of the interruption -cdau variable with 287 'no' and 3 'yes' indicates a significant class imbalance. This imbalance can lead in the regression analysis that the model may have a tendency to predict 'no' more frequently, regardless of its true predictive power. The minority class being the 'yes' class and majority class being the 'no' class.

Probability table of gender -bsex

```
table(data$bsex)
```

```
##
## female   male
##    139    151
```

```
round(prop.table(x = table(data$bsex)), 2)
```

```
##
## female   male
##   0.48   0.52
```

The variable gender -bsex, consists of two gender, either a person is a female or male. The frequency tables show that 139 respondents or 47% are female and 161 respondents or 52% are male.

Probability table of cdau

```
table(data$cdau)
```

```
##
##  no yes
```

```
## 287   3
```

```
round(prop.table(x = table(data$cdau)), 2)
```

```
##
##   no  yes
## 0.99 0.01
```

The variable interruption of work longer than three months - cdau, consists of two groups, either a person had an interruption of work longer than three months or not. The frequency tables show that 287 respondents or 98% had no interruption longer than three months and 3 respondents or 1% had. The majority class in this case is the "no" category, while a very small number fall into the "yes" category. This suggests a strong class imbalance, with "no" being the predominant category.

Probability table of xkartab

```
table(data$xkartab)
```

```
##
##  Basic   High Middle
##    190     59     41
```

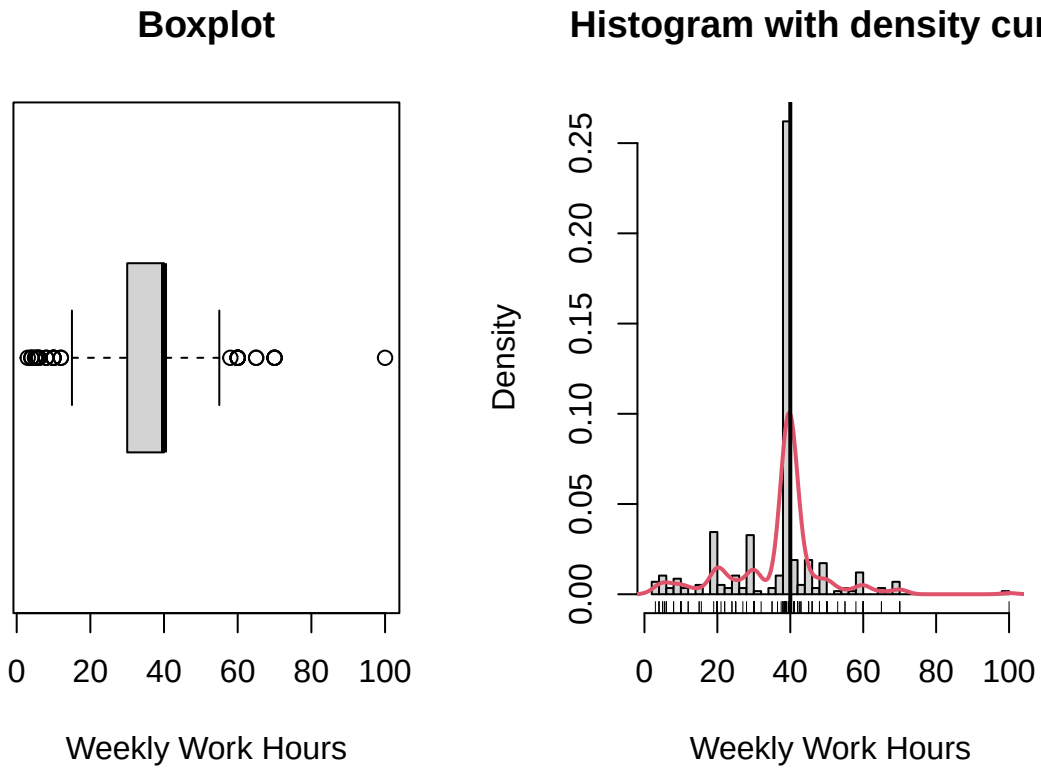
```
round(prop.table(x = table(data$xkartab)), 2)
```

```
##
##  Basic   High Middle
##   0.66   0.20   0.14
```

The variable highest education -xkartab, is divided into three categories. The frequency tables show that 190 respondents or 66% have completed basic level of education, 41 respondents or 14% are have mid-level education, whereas 59 respondents or 20% have completed higher education.

A boxplot and a histogram with density curve for the variable weekly work hours -dstd are created.

```
par(mfrow=c(1,2))
boxplot <- boxplot(data$dstd, horizontal=TRUE, main = "Boxplot", xlab = "Weekly Work Hours")
hist(data$dstd,
     freq = FALSE,
     xlab="Weekly Work Hours",
     main = "Histogram with density curve",breaks=40)
lines(density(data$dstd), col = 2, lwd = 2)
abline(v = median(data$dstd), col='black',lwd=2)
rug(data$dstd)
```



Calculate the skewness

```
skew <- e1071::skewness(data$dstd)
kurt <- e1071::kurtosis(data$dstd)
cat("Skewness:", skew, "\n")
```

```
## Skewness: -0.09035905
```

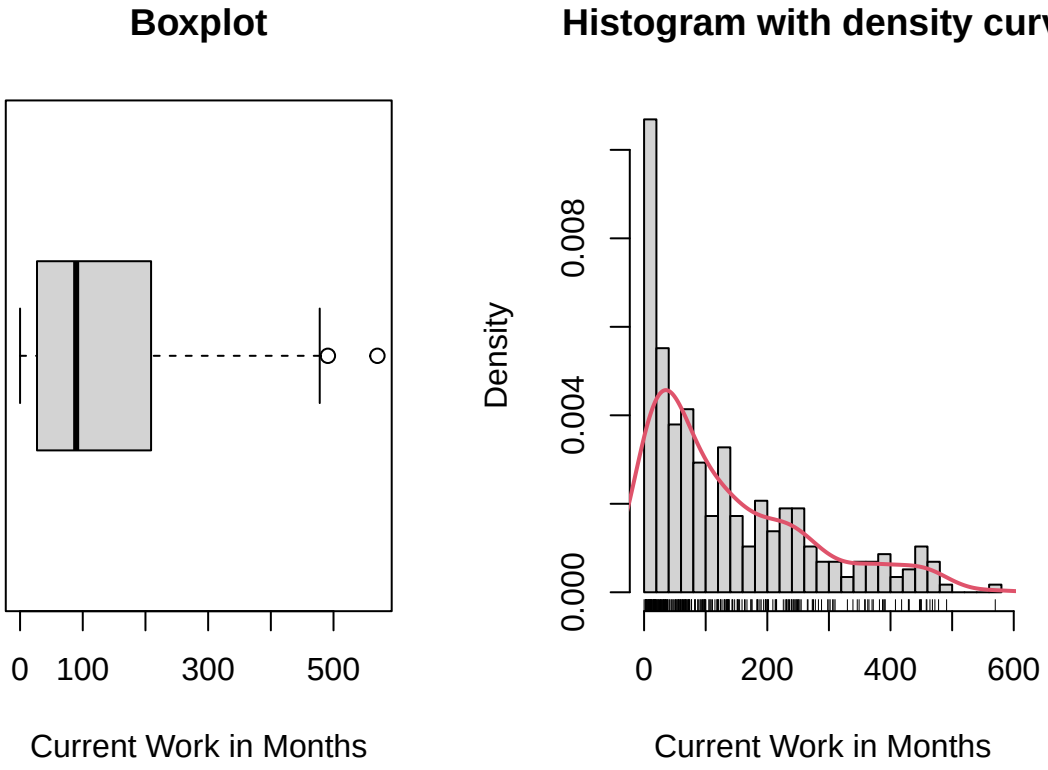
```
cat("Kurtosis:", kurt, "\n")
```

```
## Kurtosis: 2.769042
```

The skewness value of approximately -0.0904 (rounded to four decimal places) indicates that the distribution of the dstd is **slightly** left-skewed (negatively skewed). The kurtosis value of approximately 2.769 indicates that dstd has positive excess kurtosis, which means it has heavier tails and a sharper peak compared to a normal distribution (which has an excess kurtosis of 0). Positive kurtosis suggests that dstd has heavier tails and a sharper peak compared to a normal distribution, indicating a leptokurtic distribution.

A boxplot and a histogram with density curve for the variable current work in months -dseitz are created.

```
par(mfrow=c(1,2))
boxplot <- boxplot(data$dseitz, horizontal=TRUE, main = "Boxplot", xlab = "Current Work in Months")
hist(data$dseitz,
     freq = FALSE,
     xlab="Current Work in Months",
     main = "Histogram with density curve",breaks=40)
lines(density(data$dseitz), col = 2, lwd = 2)
rug(data$dseitz)
```



Distribution for the variable dseitz is right-skewed, the first quartile is 30, median 40 and the third quartil is 40. The minimum is 0, these can be an indication to the observations where the respondent might have not finished the first months on the new job. The variable is still right skewed, but not sharply. The first quartile is 27.25 and the third is 207.25 . The minimum is 0, the median is 89.50, the mean is 133.61, and the maximum is 570. The majority of the data (75%) is between 0 and 207.25. The values greater than 478 are considered as outliers out of the upper-bound, there are no outliers below the lower-bound.

In summary, the data set consists of two metric variables, three categorical variables and have 290 observations. The metric variables have different scales, where weekly work hours -dstd have between 3 and 100, whereas current work in months -dseitz have values between 0 and 570. This can have an impact on the distances, which is why the variables need to be normalized in order to avoid dominance of the variables whose variation is significantly higher. It is therefore advisable to normalized the individual variables in order to avoid this undesirable effect.

In the next step, the explorative procedure is carried out to cluster the observations using hierarchical cluster analysis.

4 Hierarchical Cluster Analysis

As already mentioned in the the descriptive analysis, the metric variables are normalized in the first step. After normalization the metric variables are in the same scale and have values between 0 and 1.

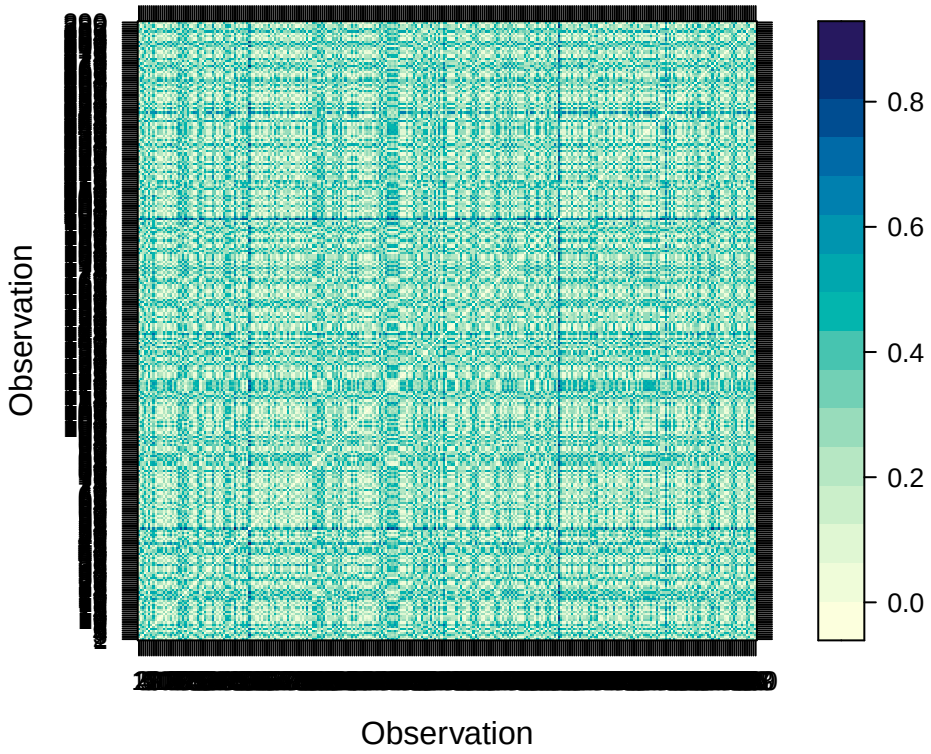
```
data.n <- data
min_max_norm <- function(x) {
    (x - min(x)) / (max(x) - min(x))
  }
data.n$dstd <- min_max_norm(data.n$dstd)
data.n$dseitz <- min_max_norm(data.n$dseitz)
summary(data.n)
```

```
##      bsex       xkartab     cdau          dstd           dseitz
```

```
##  female:139   Basic :190   no :287   Min.   :0.0000   Min.   :0.00000
##  male  :151   High  : 59   yes:  3   1st Qu.:0.2784   1st Qu.:0.04781
##               Middle: 41             Median :0.3814   Median :0.15702
##                                      Mean   :0.3476   Mean   :0.23440
##                                      3rd Qu.:0.3814   3rd Qu.:0.36360
##                                      Max.   :1.0000   Max.   :1.00000
```

In the next step, a distance matrix is created, as the data consists of both metric and categorical variables. Therefore, the Gower distance is used.

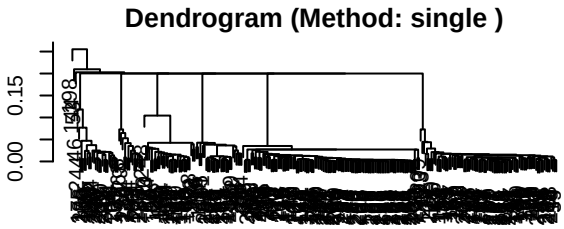
```
distance_matrix <- daisy(data.n, metric = "gower")
```

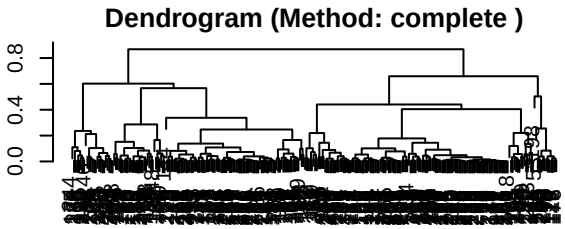


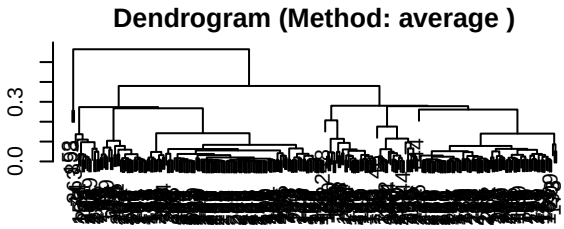
Additionally, a level plot is generated, but even with a total of 290 observations, it appears to be still large. This size making the result challenging to extract meaningful patterns or information.

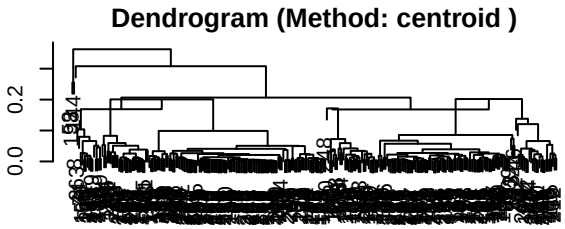
Next, hierarchical clustering will be carried out using the methods 'Single Linkage', 'Complete Linkage', 'Average Linkage', 'Centroid' and the 'Ward Method' and the corresponding dendrograms and silhouette plots are created. Furthermore, the silhouette values for clusters 2-10 are calculated and displayed.

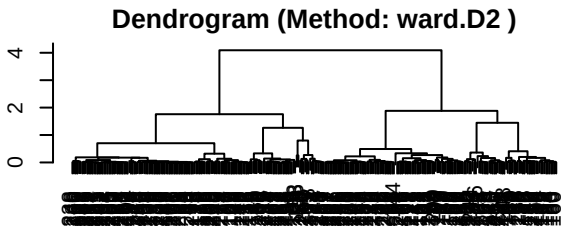
```
par(mfrow = c(3, 2), mar = c(2, 2, 2, 2))
methods <- c("single", "complete", "average", "centroid", "ward.D2")

for (method in methods) {
  cluster <- hclust(distance_matrix, method=method)
  # print(clustering_coefs <- coef(cluster))
  plot(cluster, main=paste("Dendrogram (Method:", method, ")"))
}
```











The many straight lines in the single method on the left side means that the hierarchical clustering algorithm has identified many data point as an individual cluster. Also the low height implies low dissimilarity between the clusters. The complete method and the average method shows larger clusters with higher distances between the clusters compared to the single linkage. The greater the height, the greater is the difference. The ward.D2 method shows the best results, as it shows higher height on the dendrogram (y-axis) indicating that the clusters being merged are less similar to each other.

The clustering coefficients also show that the ward.D2 method has the highest correlation coefficient with 0.995. As higher coefficient indicates a better fit of the method to the data, the choice for the further analyses will be ward.D2 and it will be used to test different k values for the number of clusters.

First, the silhouette coefficients for "Single Linkage" method are calculated. The best coefficients are at two (0.5347) and at nine (0.7276) clusters. Based on the results, the silhouette score suggest that the quality of the clustering solution is relatively better when fewer clusters are considered, with the 2-cluster solution having one of the highest silhouette score. The rest of the cluster values are relatively low.

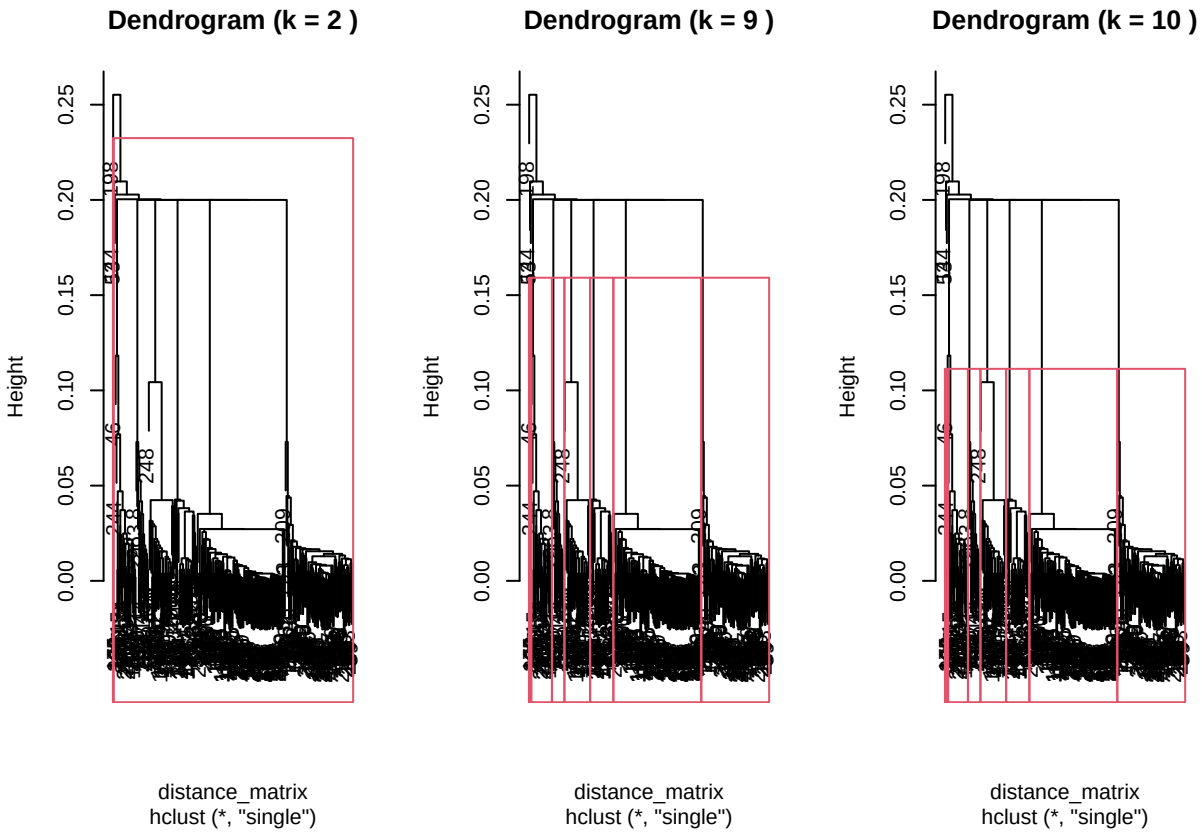
```
cluster.single.sil <- NbClust(distance_matrix, distance=NULL, diss= distance_matrix, min.nc=2, max.nc=1
cluster.single.sil$All.index
```

```
##      2      3      4      5      6      7      8      9     10
## 0.5347 0.2945 0.2037 0.2082 0.1780 0.3286 0.4700 0.7276 0.7167
```

After identifying potential optimal cluster numbers, dendrograms illustrating the corresponding clusters are generated. As the cluster number increased, the height of the rectangles is decrease. This suggests data points within those clusters are relatively similar or have a lower dissimilarity, where as nine and clusters look very similar.

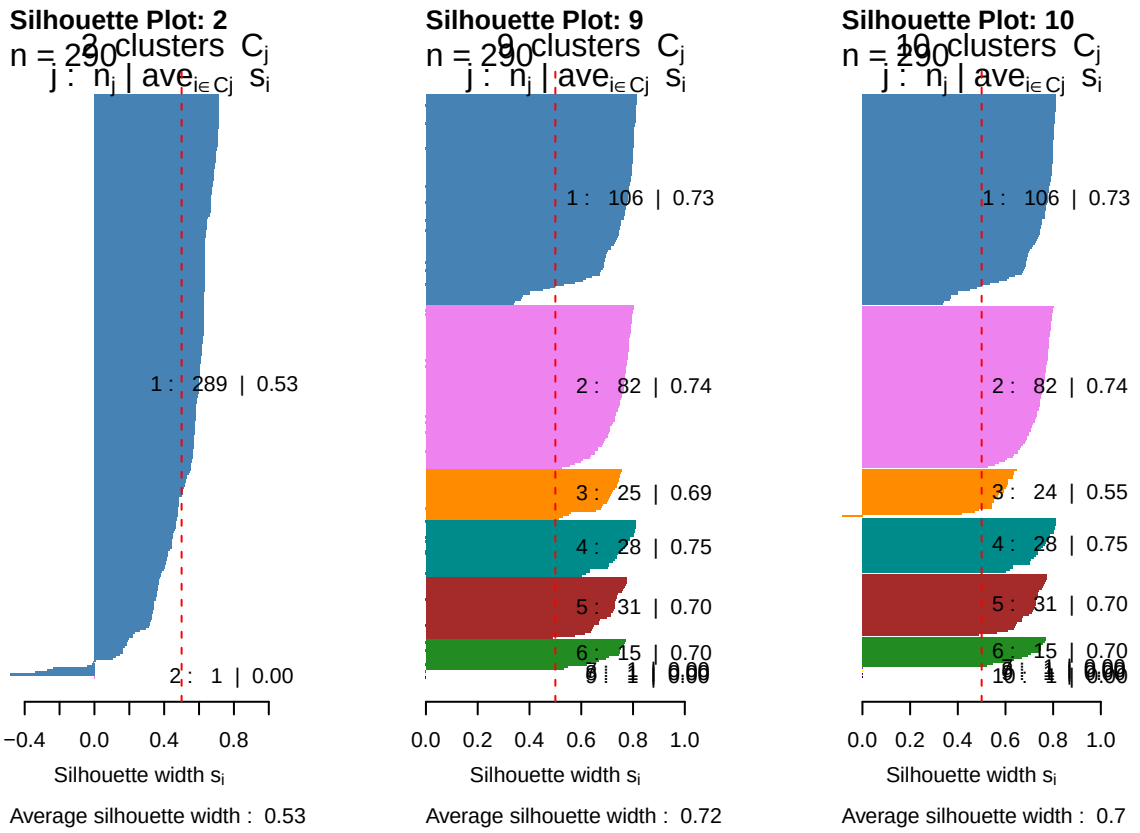
```
hc.single <- hclust(distance_matrix, method = "single")

par(mfrow = c(1, 3))
k <- c(2, 9, 10)
```

```
for (i in k) {
  plot(hc.single, main = paste("Dendrogram (k =", i, ")"))
  rect.hclust(hc.single, k= i)
}
```



Next, the silhouette plots for two, nine and ten clusters are created, as well as a vertical line at 0.5 is drawn. With two cluster almost all the data points are in the first cluster, which shows not a good clustering. With nine and ten clusters average silhouette width is relatively large at 0.7 and most of the cluster sillhouette scores fall after the 0.5 line. There are not many poorly classified data points but overall, single linkage in this case does not efficiently cluster.

```
par(mfrow=c(1,3))
k <- c(2, 9, 10)
for (i in k) {
  plot(silhouette(cutree(hc.single, k = i), distance_matrix), main=paste("Silhouette Plot:", i), col = c
  abline(v = 0.5, lty = 2, col = "red")
}
```



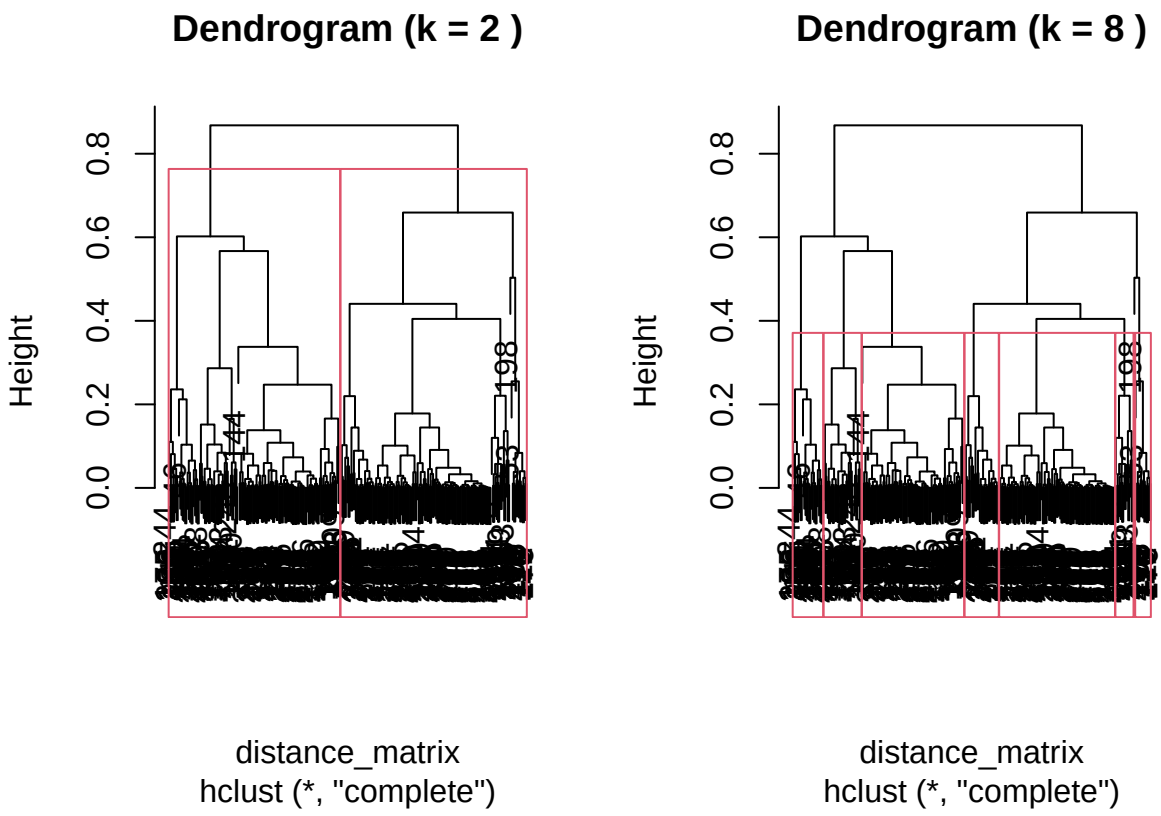
Secondly, the silhouette coefficients for 'Complete Linkage" method are calculated. The best coefficients are at two (0.5460) and at eight (0.6862) clusters, which is similar to the single linkage method.

```
cluster.complete.sil <- NbClust(distance_matrix, distance=NULL, diss= distance_matrix, min.nc=2, max.nc=
cluster.complete.sil$All.index
```

```
##      2      3      4      5      6      7      8      9     10
## 0.5460 0.4199 0.4125 0.5058 0.4707 0.5959 0.6862 0.6862 0.6683
```

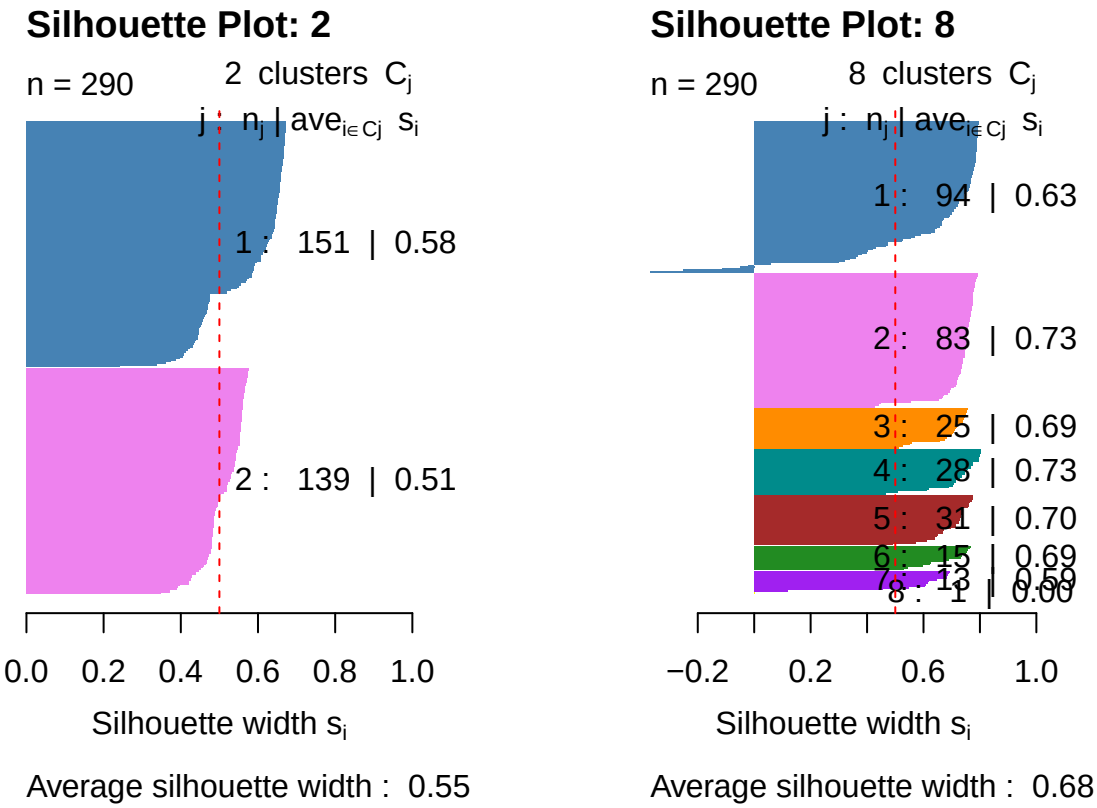
After identifying potential optimal cluster numbers, dendrograms illustrating the corresponding clusters are generated. The rechtangle widths with two clusters has almost the same length suggesting relatively well-defined clusters. As the cluster number increased, the height of the rectangles is decrease. A higher rectangle for two clusters suggests that it comprises data points that are less similar to each other compared to the data points with eight cluster, which were merged at a lower height and are more similar.

```
hc.complete <- hclust(distance_matrix, method = "complete")

par(mfrow = c(1, 2))
k <- c(2, 8)
for (i in k) {
  plot(hc.complete, main = paste("Dendrogram (k =", i, ")"))
  rect.hclust(hc.complete, k= i)
}
```



Next, the silhouette plots for two, and eight clusters are created, as well as a vertical line at 0.5 is drawn. With two cluster data is divided into two groups almost evenly, which indicates a good clustering. With eight clusters average silhouette width is relatively larger at 0.68 and most of the cluster sillhouette scores fall after the 0.5 line, except for the eigth cluster. There are not many poorly matched data points. Overall, complete linkage in this case also does not efficiently cluster for eight clusters. Therefore, in this case two cluster can be chosen for the complete linkage method.

```
par(mfrow=c(1,2))
k <- c(2, 8)
for (i in k) {
  plot(silhouette(cutree(hc.complete, k = i), distance_matrix), main=paste("Silhouette Plot:", i), col =
  abline(v = 0.5, lty = 2, col = "red")
}
```



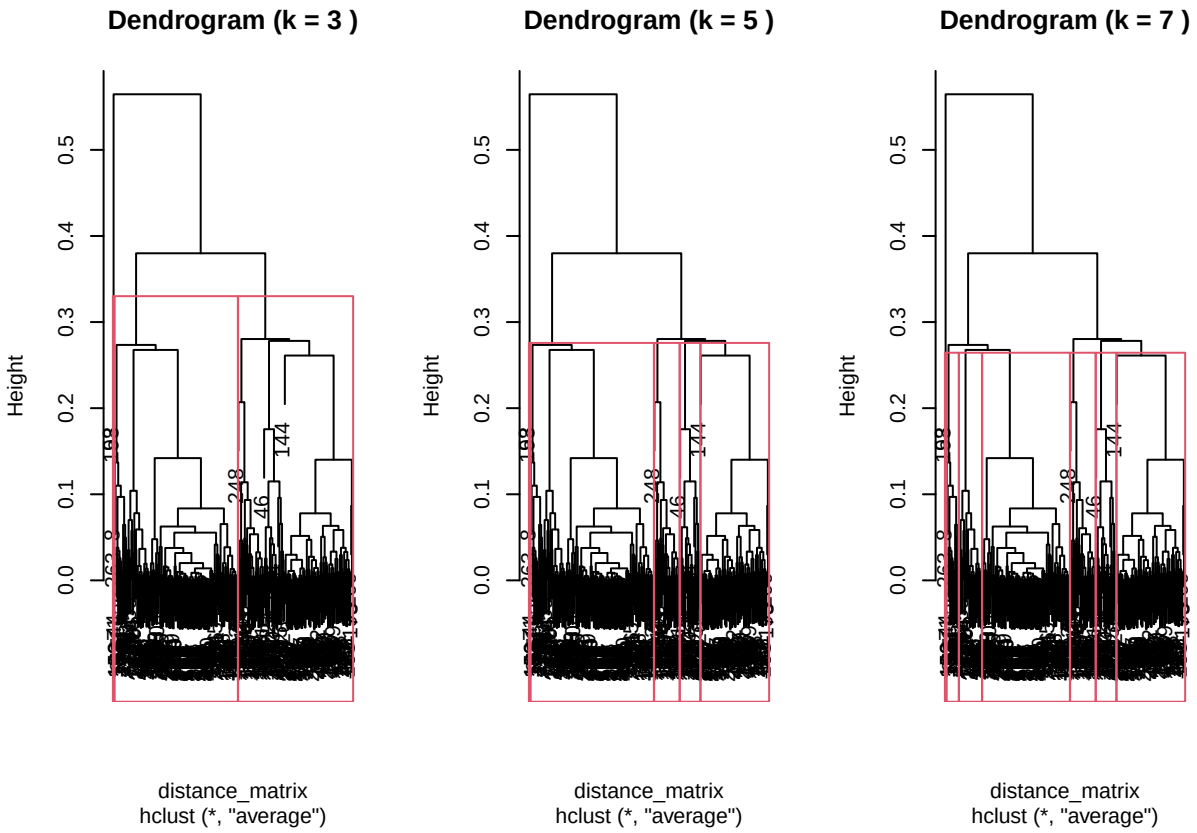
Thirdly, the silhouette coefficients for 'Average Linkage" method are calculated. The average silhouette coefficients show that the best results are achieved at three (0.5424), five (0.5711) clusters and at seven (0.7313) clusters.

```
cluster.average.sil <- NbClust(distance_matrix, distance=NULL, diss= distance_matrix, min.nc=2, max.nc=
cluster.average.sil$All.index
```

```
##      2      3      4      5      6      7      8      9     10
## 0.4901 0.5424 0.5143 0.5711 0.6211 0.7313 0.7307 0.7276 0.7170
```

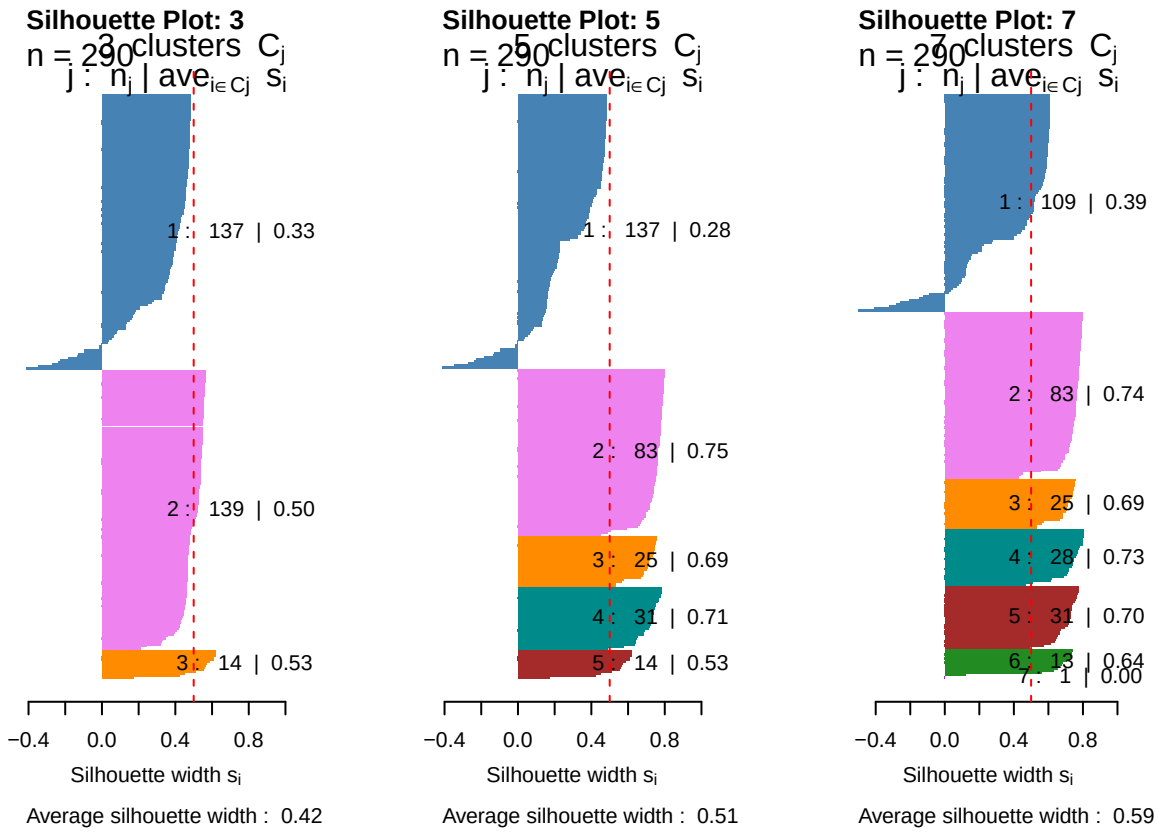
After identifying potential optimal cluster numbers, dendrograms illustrating the corresponding clusters are generated. The rechtangle widths with there clusters has almost the same length in its two clusters and includes little amount of data points in the other cluster. This does not indicate good clustering. As the cluster number increased, the height of the rectangles is decrease, although from five to seven there was not much decrease, indicating data points are as dissimilar as in five and in seven clusters.

```
hc.average <- hclust(distance_matrix, method = "average")

par(mfrow = c(1, 3))
k <- c(3, 5, 7)
for (i in k) {
  plot(hc.average, main = paste("Dendrogram (k =", i, ")"))
  rect.hclust(hc.average, k= i)
}
```



Next, the silhouette plots for three, and seven clusters are created, as well as a vertical line at 0.5 is drawn. With three cluster, data is divided into two groups almost evenly with an additional cluster. With five and seven clusters average silhouette width is largest at 0.59 and most of the cluster sillhouette scores fall after the 0.5 line, except for the seventh cluster. There are some poorly matched data points in each cluster number. Overall, average linkage in this case also does not necessarily cluster well for seven clusters, where the seventh cluster has only one data ppint. Therefore, in this case five cluster can be chosen for the average linkage method.

```
par(mfrow=c(1,3))
k <- c(3, 5, 7)
for (i in k) {
  plot(silhouette(cutree(hc.complete, k = i), distance_matrix), main=paste("Silhouette Plot:", i), col =
  abline(v = 0.5, lty = 2, col = "red")
}
```



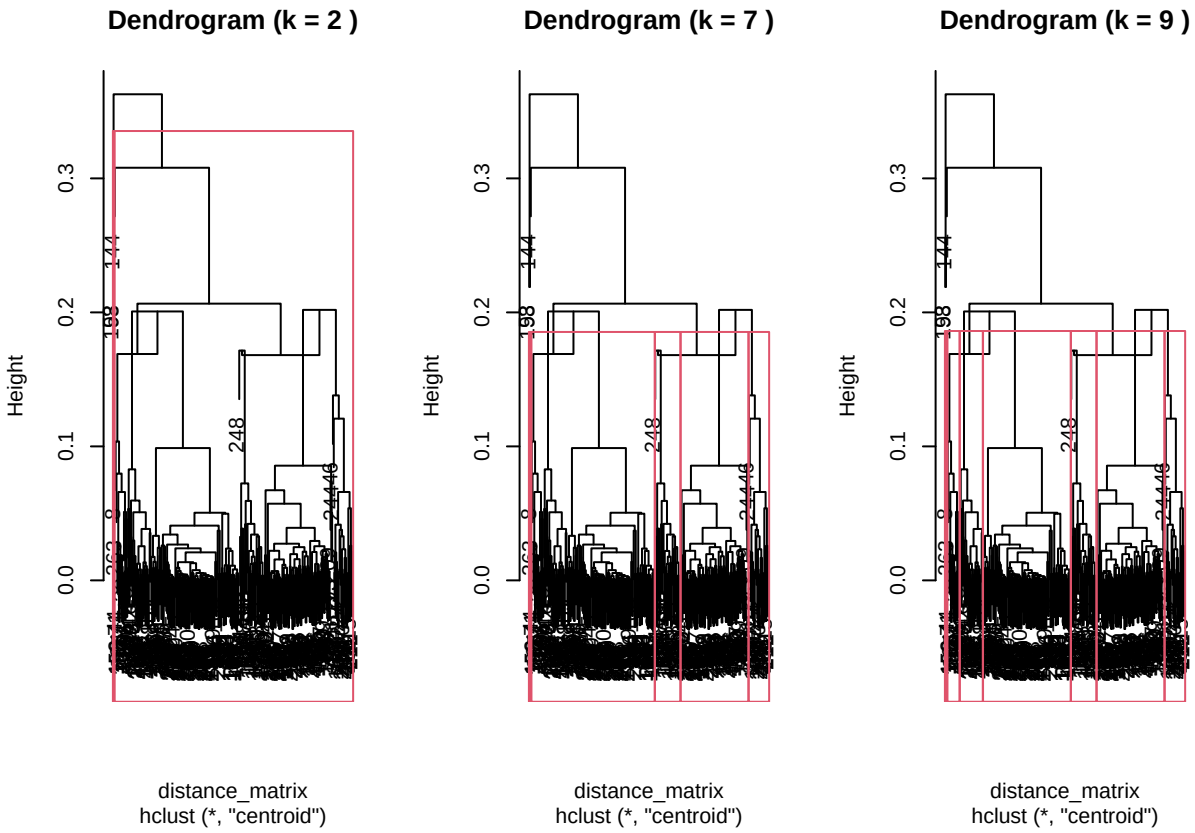
For the fourth method, the silhouette coefficients for 'Centroid Linkage" method are calculated. The best coefficients are from the small numbers at two (0.4901) and at nineth (0.7276) from the higher number of clusters, seven clusters (0.5578) will be plotted as well.

```
cluster.centroid.sil <- NbClust(distance_matrix, distance=NULL, diss= distance_matrix, min.nc=2, max.nc=
cluster.centroid.sil$All.index
```

```
##      2      3      4      5      6      7      8      9     10
## 0.4901 0.2885 0.2037 0.4868 0.4894 0.5578 0.6135 0.7276 0.7170
```

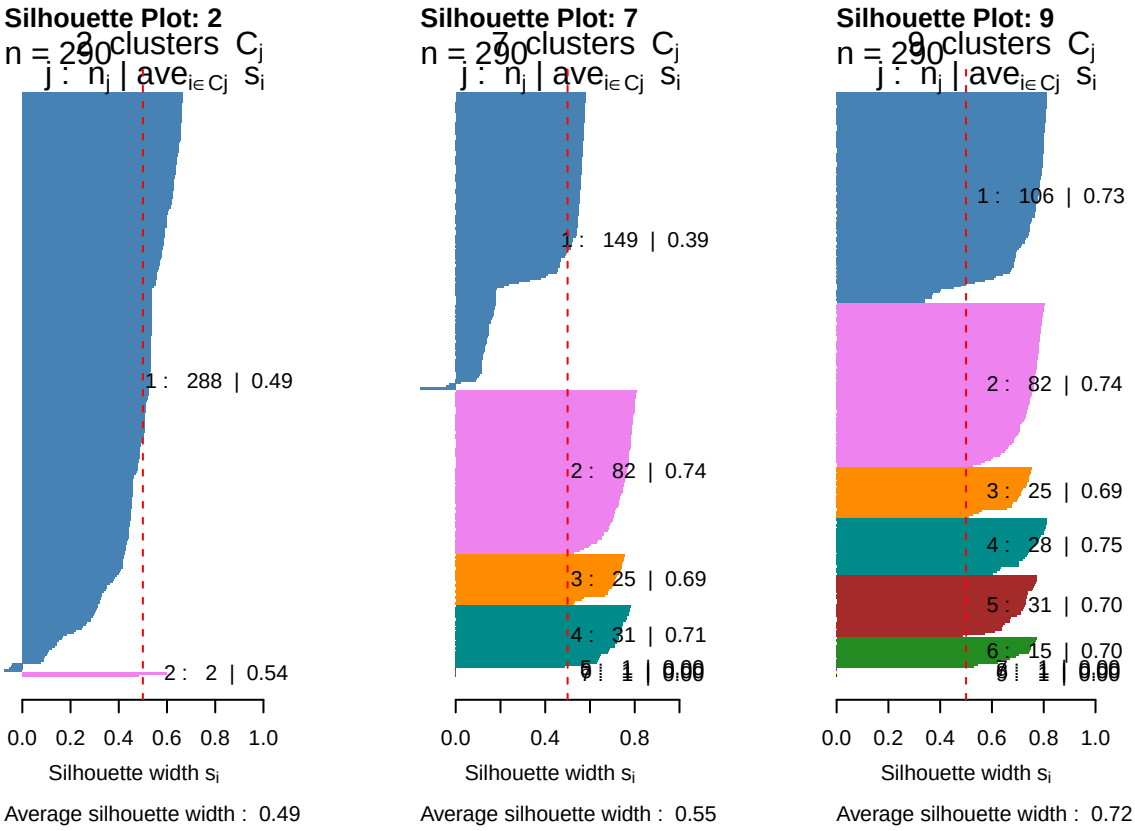
After identifying potential optimal cluster numbers, dendrograms illustrating the corresponding clusters are generated. With two cluster almost all the data points are in the first cluster, this does not suggest well-defined clusters. The height of the rectangles did not differ as much as before, indicating data points are as dissimilar as in seven and in nine clusters.

```
hc.centroid <- hclust(distance_matrix, method = "centroid")

par(mfrow = c(1, 3))
k <- c(2, 7, 9)
for (i in k) {
  plot(hc.centroid, main = paste("Dendrogram (k =", i, ")"))
  rect.hclust(hc.centroid, k= i)
}
```



Next, the silhouette plots for two, seven and nine clusters are created, as well as a vertical line at 0.5 is drawn. With two cluster, data is divided into two very uneven classes. With seven clusters average silhouette width has become larger at 0.55 and most of the cluster sillhouette scores fall after the 0.5 line, except after the fifth cluster, whereas each cluster after the fifth only has one data point. With nine clusters, similarly to seven, after the sixth cluster the rest of the data points builds individual clusters. There are not many poorly matched data points in two and seven clusters and also in nine clusters no datapoint is falsely classified. In this case there are not really a good cluster number.

```
par(mfrow=c(1,3))
k <- c(2, 7, 9)
for (i in k) {
  plot(silhouette(cutree(hc.centroid, k = i), distance_matrix), main=paste("Silhouette Plot:", i), col =
  abline(v = 0.5, lty = 2, col = "red")
}
```



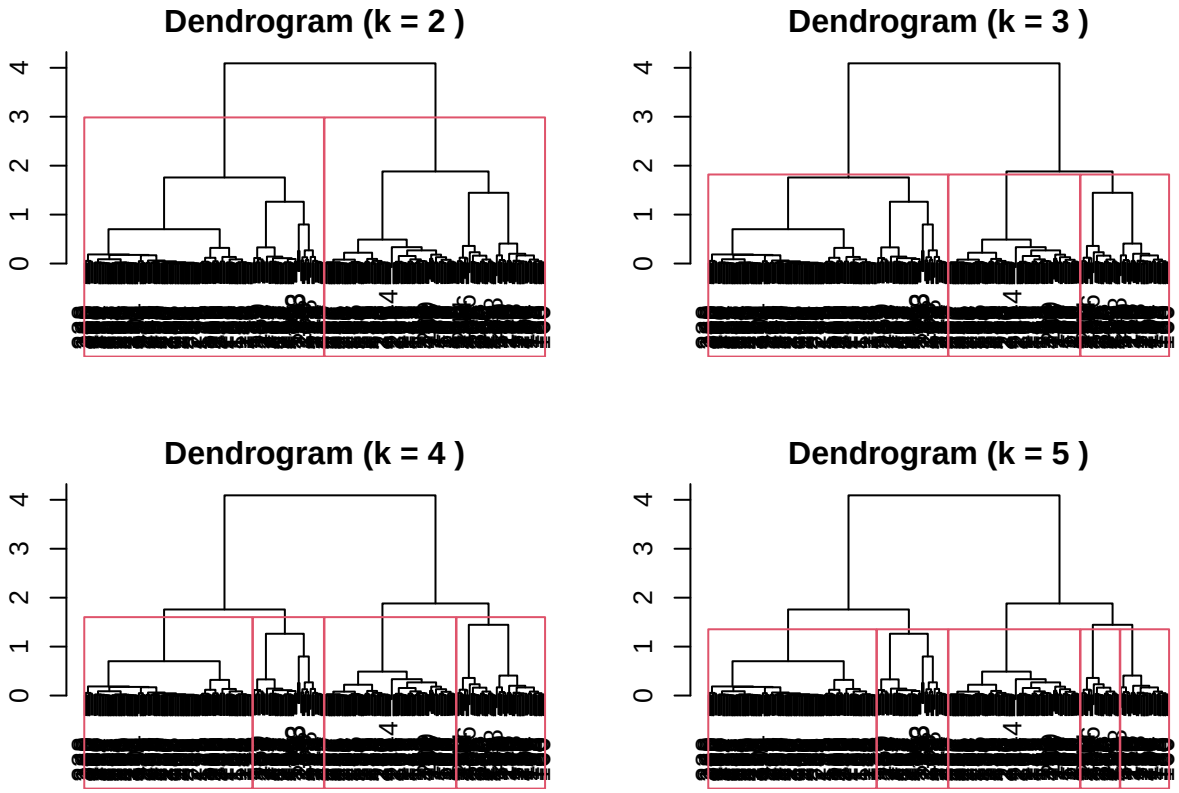
Lastly, the silhouette coefficients for 'Ward.D2" method are calculated. This method attempts to minimize the sum of the squared Euclidean distances in all clusters, meaning that it minimizes the sum of the squared differences between the points within each cluster. Compared to the other methods, the method showed much larger distances between the clusters.

```
cluster.ward.sil <- NbClust(distance_matrix, distance=NULL, diss= distance_matrix, min.nc=2, max.nc=10,
cluster.ward.sil$All.index
```

```
##      2      3      4      5      6      7      8      9     10
## 0.5460 0.5265 0.6023 0.6724 0.7219 0.7313 0.6693 0.5712 0.5558
```

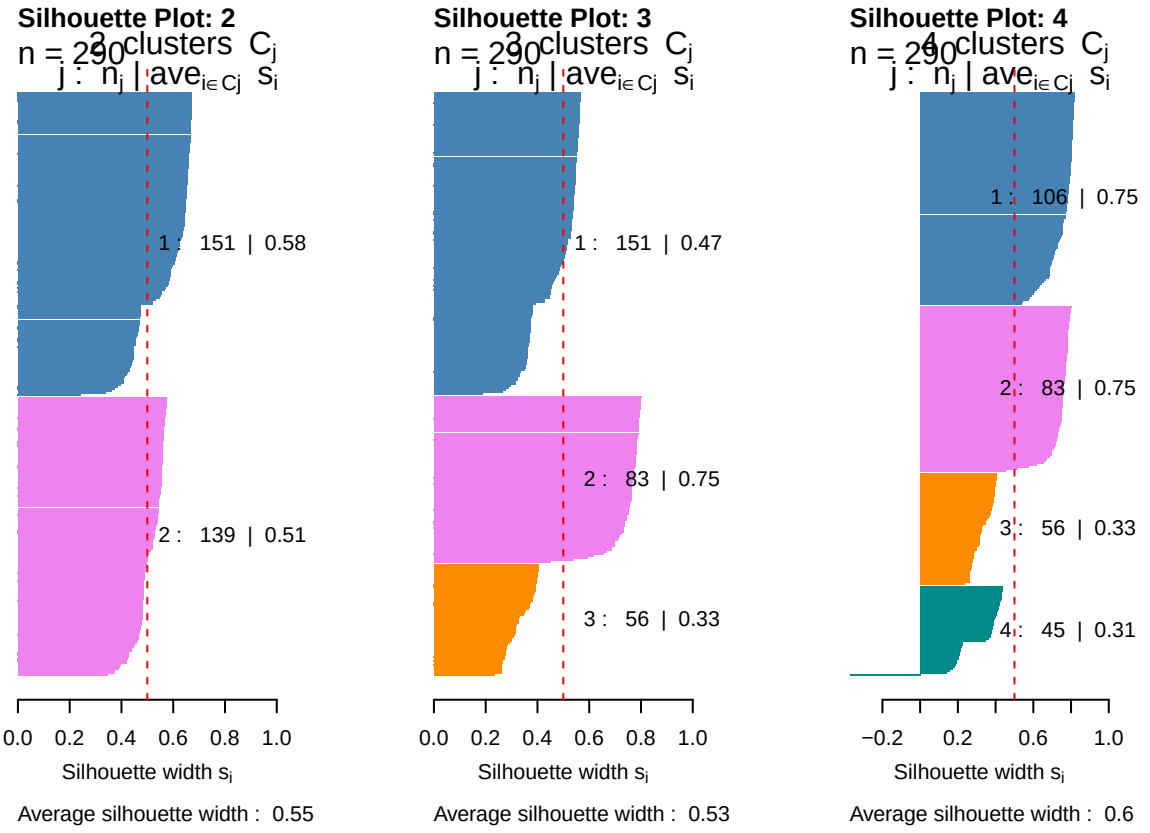
After identifying potential optimal cluster numbers, dendrograms illustrating the corresponding clusters are generated. The five and four cluster has almost at the same height, whereas three is relatively higher. The two clusters, is almost one unit higher than three and therefore not a good choice. The three cluster solution has relatively small cluster, whereas in four cluster solution clusters width closer in length. Therefore, the hierarchical cluster is divided into four to achieve a good clustering.

```
hc.ward <- hclust(distance_matrix, method = "ward.D2")

par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
for (i in 2:5) {
  plot(hc.ward, main = paste("Dendrogram (k =", i, ")"))
  rect.hclust(hc.ward, k= i)
}
```



Silhouette plots are created for two to four clusters. In the three cluster solution, the third cluster contains the remaining 56 observations; whereas the second cluster from the two cluster solution is split. The four cluster solution there are two clusters that have less than 0.5 average silhouette score, and some poorly matched data points to its own cluster. The choice remains with the four cluster solution.

```
par(mfrow=c(1,3))
for (i in 2:4) {
  plot(silhouette(cutree(hc.ward, k = i), distance_matrix), main=paste("Silhouette Plot:", i), col = co
  abline(v = 0.5, lty = 2, col = "red")
}
```



Additionally, the cluster memberships are analyzed and interpreted.

```
clust <- cutree(hc.ward, k = 4)
aggregate(. ~ clust, data = data, mean)
```

```
##   clust bsex  xkartab     cdau     dstd   dseitz
## 1     1    2 1.000000 1.000000 41.09670 134.8208
## 2     2    1 1.000000 1.012048 31.03434 116.7229
## 3     3    1 2.446429 1.000000 34.59821 132.9107
## 4     4    2 2.333333 1.044444 39.51333 162.7778
```

The clusters can be interpreted as follows:

Cluster 1 predominantly consists of male individuals with basic level of education, the average work hours are around 41.09670 and average work experience in months is around 134.82, which is around 11 years.

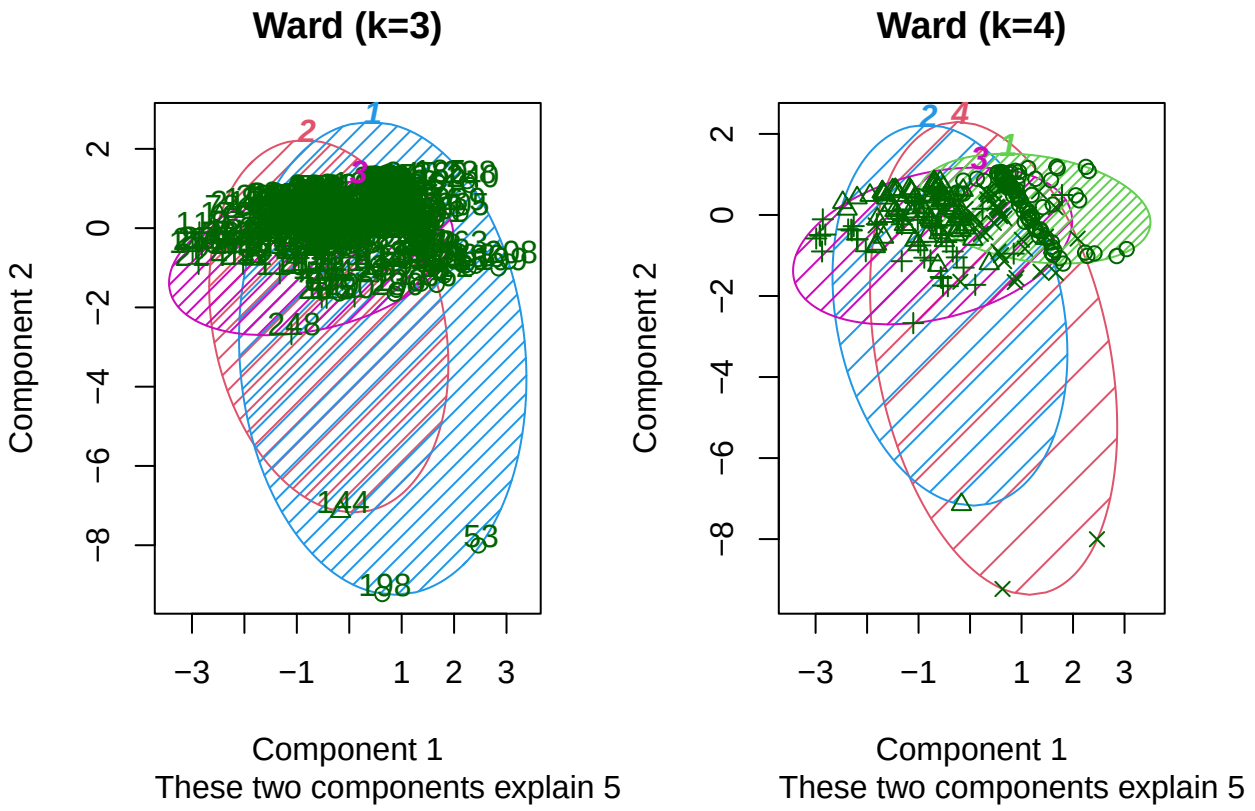
Cluster 2 predominantly consists of female individuals with basic level of education, the average work hours are around 31.03434 which is less than the first cluster, average work experience in months is around 116.7229, which is around 9 years, also less than the first cluster.

Cluster 3 predominantly consists of female individuals with middle and higher education levels, the average work hours are around 34.59821, and average work experience in months is around 132.9107, which is around 11 years. The difference between cluster one and three is that the first cluster consists of only male individuals having basic level of education.

Cluster 4 predominantly consists of male individuals with middle and higher education levels, the average work hours are around 39.51333, and average work experience in months is around 162.7778, which is around 13 years, having the highest averaged years in experience. This shows that males with higher education work slighlty less than males with basic education level (cluster 1) and have relatively longer averaged experience. Compared to cluster 3, males having similar education as woman, work more and have longer averaged experience.

Next, two cluster plots are created. The two optimal cluster numbers three and four, for the best performing model ward.D2 is plotted respectively. In 3-cluster solution the cluster one is almost overlapping with all other clusters, does not show a good clustering, as in the silhouette score (0.47). Cluster one and two in 4-cluster solution almost distinctively clustered, where as there are some points in the second cluster that fall into the third cluster. Cluster one and two have high silhouette coefficients.

```
par(mfrow = c(1, 2))
clusplot(data, cutree(hc.ward, k = 3), color = TRUE, shade = TRUE, labels = 2, lines = 0,  main='Ward (
clusplot(data, cutree(hc.ward, k = 4), color = TRUE, shade = TRUE, labels = 4, lines = 0,  main='Ward (
```



5 Conclusion

In this report, a cluster analysis of the five selected variables was carried out to answer the following research question: *Are there distinct patterns of employment-related characteristics in microcensus data, and how can hierarchical cluster analysis (HCA) be used to uncover these patterns by using a combination of metric and categorical variables, including normal weekly work hours, year of the highest education degree, desired total working hours, current work duration, and interruptions of work longer than three months?*

To answer this question, the data from the 2012 microcensus is examined. In the next step, all variables were transformed, and all implausible observations were removed, this included 290 observations in total. Next, a descriptive analysis was performed in which the characteristics were described univariately. The metric variables have found to be at different scale levels and need to be normalized.

After the descriptive analysis, the hierarchical cluster analysis was carried out with total of five different methods of agglomerative clustering including Single Linkage, Complete Linkage, Average Linkage, Centroid Linkage and Ward method. Considering the data included both metric and categorocal variable to calculate distance the Gower metric is used. Although a level plot was generated, extracting clear patterns was not still possible due to the high dimensionality. Next, for each method average silhouette coefficients were calculated from two to ten clusters and silhouette plots were created for the best results. The worst performing method

were for this analysis was the single linkage. Average results were achieved with Complete Linkage and the centroid method, where the best silhouette coefficients were between 0.30 and 0.36. The best results were achieved with ward method. With the ward method, the highest silhouette coefficient was 0.75 for two clusters in the 4-cluster solution. With three cluster the results were relatively worse than the 4-cluster solution.

Next, two cluster plots were created for the two best cluster numbers. It was shown that the variables used can be grouped into relatively separable clusters with number of four clusters. The cluster interpretations showed some distinct groups females with basic level/higher level as well as with men. Work hours or current work in months showed little variation in the clusters. The interruption variable could be excluded which had a severe imbalance.

Possible problems in the data and the analysis is the choice of variables for clustering, considering they may not capture all relevant information for employment-related characteristics. As for future work, additional attemps can be made to form more meaningful clusters with other features.