university of groningen

Statistical Learning in Marketing

Individual Assignment 2

Faculty of Economics and Business

Methodological Approach

Managerial Question: How can Ixmor optimize their marketing strategy based on information gained from segmentation analysis for their sandwich spread, ispecifically n regards to market size, product attributes, and consumer demographics?

Methodology Approach

1. Defining the Research Purpose

Before starting with cluster analyses, it is important to **define the issue** that is being looked into. This includes the **identification of active variables** and **passive variables.** (Appendix 2.1)

2. Data Preparation

Some extreme values were removed from the dataset, as the active variables were only supposed to be on a 1-7 scale. Before commencing with cluster analysis, **multicollinearity** is checked for. Variables are then **standardized**, and the **euclidean distance matrix** is calculated.

3. Hierarchical Clustering

A number of linkage methods are considered. Finally, the **ward method** hierarchical clustering method is chosen as it minimizes within cluster variance. The Ward method is robust to outliers and creates consistent and interpretable results. The resulting possible cluster solutions are tested for significance across variables. Results were visualized in two dimensions. Finally demographic trends are analyzed.

4. K-means clustering

BIC, Screeplot, and Sillhouette score was used to confirm correct number of clusters. Seed was used to ensure reproducibility. The optimal model's cluster distinctions were validated through **ANOVA**, visualized with **boxplots**, and demographic characteristics were analyzed to interpret each segment. Results were visualized in two dimensions using principal component analysis (**PCA**) for clarity, ensuring distinct and interpretable clusters aligned with the data structure.

3. K-means clustering using hierarchical solution

The 2-cluster solution from **Ward's method** provided stable initial centers for the k-means clustering algorithm. Using hierarchical clustering as a starting point enhances **k-means** stability and avoids random initializations that could lead to inconsistent results. This approach allows k-means to refine clusters further, capturing subtle distinctions while maintaining compactness, resulting in clear, interpretable clusters that leverage the strengths of both methods. Results are visualized using **PCA**.

5. Model-based clustering

Model-based clustering was conducted using the mclust package to explore the data's underlying cluster structure through Gaussian mixture models. This approach identifies clusters by optimizing the **Bayesian Information Criterion (BIC)** across various model structures and cluster numbers. However, mclust yielded a single cluster, indicating a lack of distinct, separable groups under the Gaussian assumption. Consequently, the analysis focuses on the results from hierarchical clustering (Ward's method), k-means clustering, and a combined approach, which provided more interpretable and meaningful segmentation.

Evaluation of the Managerial Question (½)

Hierarchical Clustering	First, multiple linkage methods were considered, but the **Ward method** seems most appropriate as it clearly differentiates segments in an interpretable way (Appendix 2.4 - 2.8) The results of the hierarchical clustering using the ward method point to **2 clear clusters**. During the analysis, the presence of 3 clusters was proven to be insignificant for a number of variables such as in the Taste, Recycling, and Bio.content variables (Appendix 2.13). Additionally, the scree plot (Appendix 2.9, Figure 10) and merge table (Appendix 2.8, Table 14) point to a 2 cluster solution. Cluster 1 shows positive attributes in taste, fattiness, and salt content, while Cluster 2 has negative scores in these areas but scores slightly better in recyclability and bio-content, suggesting Cluster 1 represents a preference for taste and Cluster 2 leans toward eco-friendly or healthier choices (Appendix 2.12, Table 17). Furthermore, cluster 1 has negative values for spreadability and appearance, while cluster 2 has negative values for these physical characteristics of the product. The demographic analysis shows that cluster 1 is mostly male, with a mean age of 39.2, a mode education category of highschool theory oriented degree, and the area not being significant. Cluster 2 on the other hand is both male and female, a bit older with a mean age of 43.3, university degree education, and insignificant area (Appendix 2.14, Table 35). This is somewhat in line line with the two segments the manager proposed, but there are inconsistencies in regards to the first segment being health conscious,, while they instead just care about taste while disregarding health, and are not predominantly female. The first segment constitutes about 42% of the market, while the second makes up 58% (Appendix 2.11, Table 15).
K-Means Clustering	Multiple methods were used to confirm the number of clusters that will be used for the **k-means clustering**. The silhouette test (Appendix 2.17), BIC comparison (Appendix 2.18), and scree plot (appendix 2.19) all point to **2 clusters**. Additionally, the number of clusters can be derived from the hierarchical solution. The results (Appendix 2.20) are somwhat consistent with the hierarchical solution. Group 1 prioritizes bio.content, with a near 0 recycling mean and negative values for taste, fattiness, and salt, indicating health conscious consumers. Group 2 prioritizes taste, with high salt and fat contents. All the active variables are significant across the two clusters (Appendix 2.21). Group 1 has positive values for the physical attributes of the product (Appearance and Spreadability), while Group 2 is negative.The first group is mostly female, while the second group is almost entirely male. The first group mostly has university degree education, while the second group mostly has highschool theory oriented education. (Appendix 2.25, Table 51). Area is insignificant again (Appendix 2.22, Table 49), as well as age (Appendix 2.22, Table 50). The first segment accounts for 56% of the market, while the second takes up 44% (Appendix 2.20, table 39). What is especially interesting in these results is the big difference in gender distribution in the second group, which would allow the manager to easily employ segmentation targeting marketing strategies.
Combined Approach	To develop a robust segmentation strategy, a **combined approach** was undertaken where the hierarchical clustering solution was used as a foundation for applying K-means clustering. Cluster 1 exhibited higher preferences for bio-content and recycling, suggesting a segment inclined toward eco-friendly choices. Conversely, Cluster 2 showed a preference for taste, salt, and fattiness, indicating a group that prioritizes sensory factors over environmental considerations (Appendix 2.26, Table 52). Furthermore, physical attributes of the product (Appearance and Spreadability) were negative for cluster 1 and positive for cluster 2. Cluster 1 is predominantly male (83.2%) with a mean age of 40.2 years and a high-school theory oriented education mode (Appendix 2.31). Cluster 2 has a more balanced gender distribution, with a higher proportion of females (52.4%), an older mean age of 42.9, and a mode education level of university degree. The first segment accounts for 48% of the market, and the second for 52% (Appendix 2.26, Table 53).
Limitations	Given that all three segmentation approaches consistently identified **only two segments**, the manager's estimation that the two target segments make up approximately 40% of the total market cannot be substantiated. Additionally, the **multi-cluster model** yielded only one statistically significant segment (Appendix 2.27, Table 66), suggesting that the data may lack sufficient inherent structure for effective segmentation. This outcome indicates potential limitations in distinguishing distinct consumer groups within the dataset.

Evaluation of the Managerial Question (2/2)

Overall Results	All 3 methods identified 2 major segments: ● **Taste Priority**: Wards cluster 1, K-means cluster 2, and combined approach cluster 2. ● **Environmentally Conscious**: Wards cluster 2, k-means cluster 1, and combined approach cluster 1. Physical Attributes: Across all 3 segmentation approaches, the physical attributes (Appearance and Spreadability) of the product were positive when environmental attributes were positive and taste priority related variables were negative, and vice versa. This is in line with the managers predictions. However, the first customer segment is not health conscious, but actually prefers high salt and fat levels. This goes against the managers predictions. Additionally, the managers predictions about the demographic description of these segments was not entirely accurate. There are inconsistencies between the gender, age, and education distributions of the clusters across the 3 solutions. Also, it is a relevant additional insight for the manager to take into consideration the physical attributes of the product, such as appearance and spreadability.
Recommendation	**Which segmentation solution is most relevant?** The combined approach, using hierarchical clustering as a foundation for K-means, offers a statistically robust solution by leveraging the strengths of both methods. Hierarchical clustering with the Ward method identifies natural groupings and determines an optimal number of clusters through visual tools like dendrograms, but it can be limited by a static cluster structure. Applying K-means afterward refines cluster centroids, enhancing separation and improving within-cluster homogeneity. This approach reduces the risk of overfitting or misclassification seen in single-method clustering, resulting in a more stable and interpretable segmentation model suitable for actionable insights. Therefore, it is recommend for the manager to simply focus on the combined approach solution for the managers solution. **Implications:** Each segment shows clear preferences, allowing for differentiated marketing. Cluster 1, representing eco-conscious consumers who prioritize bio-content and recycling, should be targeted with sustainability-focused messaging and eco-friendly product lines. Additionally, these customers value spreadability and appearance. Conversely, Cluster 2, which values taste and sensory attributes, would respond well to campaigns emphasizing flavor, richness, and indulgence. However, since Ixmor's brand image is primarily health-focused, the second cluster may not align closely with Ixmor's target audience. Given that Ixmor's products are low in salt and fat, the brand could benefit by concentrating on the eco-conscious segment, which aligns better with its health-oriented offerings. Additionally, this eco-conscious segment of the market also values physical attributes of the product (appearance and spreadability), so Ixmor should consider including these product attributes in the marketing strategies. Cluster 2, characterized by a predominantly male gender distribution (83.2%) relatively younger age, and predominantly high-school theory oriented consumers. **Ixmor should therefore direct its marketing efforts toward appealing to this demographic, emphasizing the brand's commitment to health and sustainability to resonate with environmentally conscious, lower education, male customers.** Additionally, it is important to note that the area did not have any statistical effect in this model, as well as others, so Ixmor can take a standardized marketing approach in regards to geographic segmentation techniques.

Appendices

Appendix 1: R Code

```r
# Libraries ----
library(cluster)
library(RColorBrewer)
library(plot3D)
library(mclust)
library(dplyr)



# Data Setup ----
## Work Directory and libraries ----
setwd("C:/Users/nilsd/OneDrive/Documents/Statistical Learning in Marketing/Individual Assignments")

## Importing dataset ----
segmentation <- read.csv("4766946_segment.csv")

## Data overview ----
str(segmentation)
summary(segmentation)

segmentation_clean <- segmentation[!(segmentation$Taste == 0 | segmentation$Taste == 8 |
                                     segmentation$Fattiness == 0 | segmentation$Fattiness == 8 |
                                     segmentation$Salt == 0 | segmentation$Salt == 8 |
                                     segmentation$Spreadability == 0 | segmentation$Spreadability == 8 |
                                     segmentation$Appearance == 0 | segmentation$Appearance == 8 |
                                     segmentation$Recycling == 0 | segmentation$Recycling == 8 |
                                     segmentation$Bio.content == 0 | segmentation$Bio.content == 8), ]

summary(segmentation_clean)

cor_matrix <- cor(segmentation_clean[,c(6:12)])
print(cor_matrix)

## Neither active or passive variables ----

# Column 1: Student ID

# Column 13: ID

## Passive Variables ----
# Column 2: Gender: 1 = female, 2 = male, 3 = other

# Column 3: Age: Number of years, continuous variable

# Column 4: Education: 1 = high school profession-oriented degree, 2 = high school theory-oriented degree,
# 3 = higher education non-university degree, 4 = university degree, 5 = other

# Column 5: Area: 1 = metropolitan, 2 = urban, 3 = suburban, 4 = countryside

## Active Variables ----
```

```r
# Column 6: Taste: How strong should the flavor of the sandwich spread be? 1-7, 1 = mild flavor, 7 = strong flavor

# Column 7: Fattiness: How much fat should the sandwich spread contain? 1-7, 1 = low fat level, 7 = high fat level

# Column 8: Salt: How much salt should the sandwich spread contain? 1-7, 1 = low salt level, 7 = high salt level

# Column 9: Spreadability: How spreadable should the sandwich spread be? 1-7, 1 = dense structure, 7 = loose structure

# Column 10: Appearance: How attractive should the packaging be? 1-7, 1 = low attractiveness, 7 = high attractiveness

# Column 11:Recycling: How environmentally-friendly should the packaging be? 1-7,
# 1 = low environmental-friendliness, 7 = high environmental-friendliness

# Column 12: Bio.content: How much of the ingredients of the sandwich spread should be organic/bio?
# 1-7, 1 = low organic/bio level, 7 = high organic/bio level



sum(is.na(segmentation_clean))

# Data Standardization ----

segmentation2 <- segmentation_clean
segmentation2[,6:12] <- data.frame(scale(segmentation_clean[,6:12]))
summary(segmentation2)
segmentation_active <- as.data.frame(c(segmentation2$Taste, segmentation2$Fattiness, segmentation2$Recycling, segmentation2$Bio.content))

# Euclidian distance ----
d <- dist(segmentation2[,c(6:12)])
as.matrix(d)[1:10, 1:10]

# Hierarchicial Clustering ----

segmentation2a.hc <- hclust(d, method = "single")
plot(segmentation2a.hc)
segmentationaggloa <- cbind(as.data.frame(segmentation2a.hc[1]), as.data.frame(segmentation2a.hc[2]))
segmentationaggloa[1:15,]
segmentationaggloa[475:491,]

segmentation2b.hc <- hclust(d, method = "complete")
plot(segmentation2b.hc)
segmentationagglob <- cbind(as.data.frame(segmentation2b.hc[1]), as.data.frame(segmentation2b.hc[2]))
segmentationagglob[1:15,]
segmentationagglob[475:491,]

segmentation2c.hc <- hclust(d, method = "average")
plot(segmentation2c.hc)
segmentationaggloc <- cbind(as.data.frame(segmentation2c.hc[1]), as.data.frame(segmentation2c.hc[2]))
segmentationaggloc[1:15,]
segmentationaggloc[475:491,]
```

```r

segmentation2d.hc <- hclust(d, method = "centroid")
plot(segmentation2d.hc)
segmentationagglod <- cbind(as.data.frame(segmentation2d.hc[1]), as.data.frame(segmentation2d.hc[2]))
segmentationagglod[1:15,]
segmentationagglod[475:491,]

segmentation2e.hc <- hclust(d, method = "ward.D2")
plot(segmentation2e.hc)
segmentationaggloe <- cbind(as.data.frame(segmentation2e.hc[1]), as.data.frame(segmentation2e.hc[2]))
segmentationaggloe[1:15,]
segmentationaggloe[475:491,]

## Screeplot ----
segmentationscreea <- sort(segmentationaggloa[475:491, c(3)], decreasing = TRUE)
plot(segmentationscreea, type="l", col="red", lwd=5, xlab="clusters", ylab="Cluster distance", main="Single Linkage Scree Plot", xaxt="n")
axis(1, at=seq(1,20,by = 1))

segmentationscreeb <- sort(segmentationagglob[475:491, c(3)], decreasing = TRUE)
plot(segmentationscreeb, type="l", col="red", lwd=5, xlab="clusters", ylab="Cluster distance", main="Complete Linkage Scree Plot", xaxt="n")
axis(1, at=seq(1,20,by = 1))

segmentationscreec <- sort(segmentationaggloc[475:491, c(3)], decreasing = TRUE)
plot(segmentationscreec, type="l", col="red", lwd=5, xlab="clusters", ylab="Cluster distance", main="Average Linkage Scree Plot", xaxt="n")
axis(1, at=seq(1,20,by = 1))

segmentationscreed <- sort(segmentationagglod[475:491, c(3)], decreasing = TRUE)
plot(segmentationscreed, type="l", col="red", lwd=5, xlab="clusters", ylab="Cluster distance", main="Centroid Scree Plot", xaxt="n")
axis(1, at=seq(1,20,by = 1))

segmentationscreee <- sort(segmentationaggloe[475:491, c(3)], decreasing = TRUE)
plot(segmentationscreee, type="l", col="red", lwd=5, xlab="clusters", ylab="Cluster distance", main="Ward Scree Plot", xaxt="n")
axis(1, at=seq(1,20,by = 1))

# WARD Method continuation ----

# Dendograms for possible amount of clusters

# see hclust's proposal for 6 groups
plot(segmentation2e.hc)
rect.hclust(segmentation2e.hc, k=6, border="red")

# see hclust's proposal for 5 groups
plot(segmentation2e.hc)
rect.hclust(segmentation2e.hc, k=5, border="red")

# see hclust's proposal for 4 groups
plot(segmentation2e.hc)
rect.hclust(segmentation2e.hc, k=4, border="red")

```

```r
# see hclust's proposal for 3 groups
plot(segmentation2e.hc)
rect.hclust(segmentation2e.hc, k=3, border="red")

# see hclust's proposal for 2 groups
plot(segmentation2e.hc)
rect.hclust(segmentation2e.hc, k=2, border="red")

# Screeplot suggests 2 cluster
# Dendrogram also suggests 2 clusters, however it could be argued up to 3

segmentation2e.hc.segment2 <- cutree(segmentation2e.hc, k=2)
table(segmentation2e.hc.segment2)

segmentation2e.hc.segment3 <- cutree(segmentation2e.hc, k=3)
table(segmentation2e.hc.segment3)

# seg sum function
seg.summ <- function(data , groups) {aggregate(data , list(groups), function(x) mean(as.numeric(x)))}
seg.summ(segmentation2[,c(6:12)], segmentation2e.hc.segment2)
segmentation2e.hc.mean <- seg.summ(segmentation2[,c(6:12)], segmentation2e.hc.segment2)

clusmember2 <- as.factor(segmentation2e.hc.segment2)
segmentation2eaovbase2 <- cbind(clusmember2,segmentation2[,c(6:12)])

clusmember3 <- as.factor(segmentation2e.hc.segment3)
segmentation2eaovbase3 <- cbind(clusmember3,segmentation2[,c(6:12)])

segmentationwithage <- cbind(clusmember2, segmentation2)

## ANOVA Tests ----

# ANOVA test for Taste

#2 cluster
segmentation2.aov_Taste2 <- aov(Taste ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Taste2)
TukeyHSD(segmentation2.aov_Taste2)

#3 cluster
segmentation2.aov_Taste3 <- aov(Taste ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Taste3)
TukeyHSD(segmentation2.aov_Taste3)

# ANOVA Test Fattiness

#2 cluster
segmentation2.aov_Fattiness2 <- aov(Fattiness ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Fattiness2)
TukeyHSD(segmentation2.aov_Fattiness2)
```

```r

# 3 cluster
segmentation2.aov_Fattiness3 <- aov(Fattiness ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Fattiness3)
TukeyHSD(segmentation2.aov_Fattiness3)

# ANOVA Test for Salt

#2 cluster
segmentation2.aov_Salt2 <- aov(Salt ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Salt2)
TukeyHSD(segmentation2.aov_Salt2)

# 3 cluster
segmentation2.aov_Salt3 <- aov(Salt ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Salt3)
TukeyHSD(segmentation2.aov_Salt3)

# ANOVA Test for Spreadability

# 2 cluster
segmentation2.aov_Spreadability2 <- aov(Spreadability ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Spreadability2)
TukeyHSD(segmentation2.aov_Spreadability2)

# 3 cluster
segmentation2.aov_Spreadability3 <- aov(Spreadability ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Spreadability3)
TukeyHSD(segmentation2.aov_Spreadability3)

# ANOVA Test for Appearance

# 2 cluster
segmentation2.aov_Appearance2 <- aov(Appearance ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Appearance2)
TukeyHSD(segmentation2.aov_Appearance2)

# 3 cluster
segmentation2.aov_Appearance3 <- aov(Appearance ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Appearance3)
TukeyHSD(segmentation2.aov_Appearance3)

# ANOVA Test for Reycling

#2 cluster
segmentation2.aov_Recycling2 <- aov(Recycling ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Recycling2)
TukeyHSD(segmentation2.aov_Recycling2)

#3 cluster
```

```r
segmentation2.aov_Recycling3 <- aov(Recycling ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Recycling3)
TukeyHSD(segmentation2.aov_Recycling3)

# ANOVA Test for Bio.content

# 2 cluster
segmentation2.aov_Bio.content2 <- aov(Bio.content ~ clusmember2, data = segmentation2eaovbase2)
summary(segmentation2.aov_Bio.content2)
TukeyHSD(segmentation2.aov_Bio.content2)

# 3 cluster
segmentation2.aov_Bio.content3 <- aov(Bio.content ~ clusmember3, data = segmentation2eaovbase3)
summary(segmentation2.aov_Bio.content3)
TukeyHSD(segmentation2.aov_Bio.content3)

# ANOVA test for age
segmentation2.aov_age <- aov(Age ~ clusmember2, data = segmentationwithage)
summary(segmentation2.aov_age)
TukeyHSD(segmentation2.aov_age)

# Categorical variables significance tests

#Gender
segmentation2e.freq.gender <-table(clusmember2, segmentation2[,c(2)])
segmentation2e.freq.gender
segmentation2e.chisq.gender <- chisq.test(segmentation2e.freq.gender)
segmentation2e.chisq.gender

#Education
segmentation2e.freq.education <-table(clusmember2, segmentation2[,c(4)])
segmentation2e.freq.education
segmentation2e.chisq.education <- chisq.test(segmentation2e.freq.education)
segmentation2e.chisq.education

#Area
segmentation2e.freq.area <-table(clusmember2, segmentation2[,c(5)])
segmentation2e.freq.area
segmentation2e.chisq.area <- chisq.test(segmentation2e.freq.area)
segmentation2e.chisq.area

segmentation2e.hc.segment <- cutree(segmentation2e.hc, k=2)
pca <- prcomp(segmentation2[,c(6:12)])
pca_scores <- predict(pca, segmentation2[,c(6:12)])
plot(pca_scores[, 1], pca_scores[, 2], col = segmentation2e.hc.segment, pch = 16, cex = 2,
     xlab = "PC1", ylab = "PC2", main = "Cluster Plot with PCA Dimensions")

# demographic comparison

segmentation_clean$segment_ward <- segmentation2e.hc.segment2
```

```r
segmentation_clean$segment_ward <- segmentationward$segment
demographic_summary <- segmentation_clean %>%
  group_by(segment_ward) %>%
  summarise(
    gender_proportion_female = mean(Gender == 1, na.rm = TRUE),
    gender_proportion_male = mean(Gender ==2, na.rum= TRUE),
    age_mean = mean(Age, na.rm = TRUE),
    education_mode = names(sort(table(Education), decreasing = TRUE))[1], # Mode for Education
    area_mode = names(sort(table(Area), decreasing = TRUE))[1] # Mode for Area
  )
demographic_summary

# K-means clustering ----

# determining optimal n of clusters

## K means will use 2 clusters, like results from WARD.

# COnfirmation whether 2 clustesr are ideal:

# silhouette scores
set.seed(123)
silhouette_scores <- c()
k_values <- 2:6
for (k in k_values) {
  kmeans_model <- kmeans(segmentation2[, 6:12], centers = k)
  sil <- silhouette(kmeans_model$cluster, dist(segmentation2[, 6:12]))
  avg_sil_width <- mean(sil[, 3])
  silhouette_scores <- c(silhouette_scores, avg_sil_width)
}
plot(k_values, silhouette_scores, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of clusters (k)",
     ylab = "Average silhouette score",
     main = "Silhouette Method for Optimal k")
silhouette_scores

# BIC values
bic_kmeans <- function(kmeans_model, data) {
  # Extract the number of clusters and data points
  k <- length(kmeans_model$centers)  # number of clusters
  n <- nrow(data)                    # number of data points
  d <- ncol(data)
  WSS <- kmeans_model$tot.withinss
  variance <- WSS / (n - k)
  log_likelihood <- -WSS / (2 * variance)
  BIC <- -2 * log_likelihood + log(n) * k * d
  return(BIC)
}
data <- segmentation2[, c(6:12)]
BIC_k2 <- bic_kmeans(segmentatione2.k2, data)
BIC_k3 <- bic_kmeans(segmentatione2.k3, data)
```

```r
BIC_k4 <- bic_kmeans(segmentatione2.k4, data)
BIC_k5 <- bic_kmeans(segmentatione2.k5, data)

BIC_values <- data.frame(
  Clusters = 2:5,
  BIC = c(BIC_k2, BIC_k3, BIC_k4, BIC_k5)
)

print(BIC_values)

## scree plot
data <- segmentation2[, c(6:12)]  # Your data subset for clustering
max_k <- 10
wss <- numeric(max_k)
for (k in 1:max_k) {
  set.seed(123)  # Set seed for reproducibility
  kmeans_model <- kmeans(data, centers = k)
  wss[k] <- kmeans_model$tot.withinss
}
plot(1:max_k, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of Clusters (k)", ylab = "Total Within-Cluster Sum of Squares (WSS)",
     main = "Elbow Method for Optimal k")

## 2 clusters will be used


set.seed(123)
segmentatione2.k2 <- kmeans(segmentation2[,c(6:12)], centers=2)




seg.summ(segmentation2[,c(6:12)], segmentatione2.k2$cluster)
table(segmentatione2.k2$cluster)

# testing significance of 2 cluster solutions


# significance tests
clusmemberk <- as.factor(segmentatione2.k2$cluster)
aovbasek <- cbind(clusmemberk, segmentation2[,c(6:12)])

k_taste <- aov(Taste ~ clusmemberk, data = aovbasek)
summary(k_taste)
TukeyHSD(k_taste)

k_fattiness <- aov(Fattiness ~ clusmemberk, data = aovbasek)
summary(k_fattiness)
TukeyHSD(k_fattiness)
```

```r
k_salt <- aov(Salt ~ clusmemberk, data = aovbasek)
summary(k_salt)
TukeyHSD(k_salt)

k_spreadability <- aov(Spreadability ~ clusmemberk, data = aovbasek)
summary(k_spreadability)
TukeyHSD(k_spreadability)

k_appearance <- aov(Appearance ~ clusmemberk, data = aovbasek)
summary(k_appearance)
TukeyHSD(k_appearance)

k_recycling <- aov(Recycling ~ clusmemberk, data = aovbasek)
summary(k_recycling)
TukeyHSD(k_recycling)

k_bio.content <- aov(Bio.content ~ clusmemberk, data = aovbasek)
summary(k_bio.content)
TukeyHSD(k_bio.content)

## significance check of demographic variables

kclus2 <- as.factor(segmentatione2.k2.segment)


#Gender
segmentation2e2.freq.gender <-table(kclus2, segmentation2[,c(2)])
segmentation2e2.freq.gender
segmentation2e2.chisq.gender <- chisq.test(segmentation2e2.freq.gender)
segmentation2e2.chisq.gender

#Education
segmentation2e2.freq.education <-table(kclus2, segmentation2[,c(4)])
segmentation2e2.freq.education
segmentation2e2.chisq.education <- chisq.test(segmentation2e2.freq.education)
segmentation2e2.chisq.education

#Area
segmentation2e2.freq.area <-table(kclus2, segmentation2[,c(5)])
segmentation2e2.freq.area
segmentation2e2.chisq.area <- chisq.test(segmentation2e2.freq.area)
segmentation2e2.chisq.area

ksegwithage <- cbind(kclus2, segmentation2)

segmentation2e2.aov_age <- aov(Age ~ kclus2, data = ksegwithage)
summary(segmentation2e2.aov_age)
TukeyHSD(segmentation2e2.aov_age)

# plotting the variables
```

```r
# Taste
boxplot(segmentation2$Taste ~ segmentatione2.k2$cluster, ylab="Taste", xlab="Cluster")

# Fattiness
boxplot(segmentation2$Fattiness ~ segmentatione2.k2$cluster, ylab="Fattiness", xlab="Cluster")

# Salt
boxplot(segmentation2$Salt ~ segmentatione2.k2$cluster, ylab="Salt", xlab="Cluster")

# Spreadability
boxplot(segmentation2$Spreadability ~ segmentatione2.k2$cluster, ylab="Spreadability", xlab="Cluster")

# Appearance
boxplot(segmentation2$Appearance ~ segmentatione2.k2$cluster, ylab="Appearance", xlab="Cluster")

# Recycling
boxplot(segmentation2$Recycling ~ segmentatione2.k2$cluster, ylab="Recycling", xlab="Cluster")

# Bio.content
boxplot(segmentation2$Bio.content ~ segmentatione2.k2$cluster, ylab="Bio.content", xlab="Cluster")

#2 clusters pca visualizaiton
segmentatione2.k2.segment <- segmentatione2.k2$cluster
pca2k <- prcomp(segmentation2[,c(6:12)])
pca_scores2k <- predict(pca2k, segmentation2[,c(6:12)])
plot(pca_scores2k[, 1], pca_scores2k[, 2], col = segmentatione2.k2.segment, pch = 16, cex = 2,
     xlab = "PC1", ylab = "PC2", main = "Cluster Plot with PCA Dimensions")

segmentation_clean$segmentk2 <- segmentatione2.k2.segment
demographic_summary_k2 <- segmentation_clean %>%
  group_by(segmentk2) %>%
  summarise(
    gender_proportion_female = mean(Gender == 1, na.rm = TRUE),
    gender_proportion_male = mean(Gender ==2, na.rum= TRUE),
    age_mean = mean(Age, na.rm = TRUE),
    education_mode = names(sort(table(Education), decreasing = TRUE))[1], # Mode for Education
    area_mode = names(sort(table(Area), decreasing = TRUE))[1] # Mode for Area
  )
demographic_summary_k2




#K means with outcomes of hierarchical cluster as starting values ----
hc_clusters <- cutree(segmentation2e.hc, k = 2)
centroids <- aggregate(segmentation2[, 6:12], by = list(cluster = hc_clusters), FUN = mean)[, -1]
kmeans_hierarchical <- kmeans(segmentation2[, 6:12], centers = centroids)
print(kmeans_hierarchical$centers)
```

```r
print(table(kmeans_hierarchical$cluster))
seg.summ(segmentation2[, c(6:12)], kmeans_hierarchical$cluster)


# significance tests
clusmemberkh <- as.factor(kmeans_hierarchical$cluster)
aovbase <- cbind(clusmemberkh, segmentation2[,c(6:12)])

kh_taste <- aov(Taste ~ clusmemberkh, data = aovbase)
summary(kh_taste)
TukeyHSD(kh_taste)

kh_fattiness <- aov(Fattiness ~ clusmemberkh, data = aovbase)
summary(kh_fattiness)
TukeyHSD(kh_fattiness)

kh_salt <- aov(Salt ~ clusmemberkh, data = aovbase)
summary(kh_salt)
TukeyHSD(kh_salt)

kh_spreadability <- aov(Spreadability ~ clusmemberkh, data = aovbase)
summary(kh_spreadability)
TukeyHSD(kh_spreadability)

kh_appearance <- aov(Appearance ~ clusmemberkh, data = aovbase)
summary(kh_appearance)
TukeyHSD(kh_appearance)

kh_recycling <- aov(Recycling ~ clusmemberkh, data = aovbase)
summary(kh_recycling)
TukeyHSD(kh_recycling)

kh_bio.content <- aov(Bio.content ~ clusmemberkh, data = aovbase)
summary(kh_bio.content)
TukeyHSD(kh_bio.content)

## passive variables significance test

#Gender
kh.freq.gender <-table(clusmemberkh, segmentation2[,c(2)])
kh.freq.gender
kh.chisq.gender <- chisq.test(kh.freq.gender)
kh.chisq.gender

#Education
kh.freq.education <-table(clusmemberkh, segmentation2[,c(4)])
kh.freq.education
kh.chisq.education <- chisq.test(kh.freq.education)
kh.chisq.education
```

```r
#Area
kh.freq.area <-table(clusmemberkh, segmentation2[,c(5)])
kh.freq.area
kh.chisq.area <- chisq.test(kh.freq.area)
kh.chisq.area

kh_age <- cbind(clusmemberkh, segmentation2)

kh.aov_age <- aov(Age ~ clusmemberkh, data = kh_age)
summary(kh.aov_age)
TukeyHSD(kh.aov_age)

# plotting

boxplot(segmentation2$Taste ~ kmeans_hierarchical$cluster, ylab="Taste", xlab="Cluster")

# Fattiness
boxplot(segmentation2$Fattiness ~ kmeans_hierarchical$cluster, ylab="Fattiness", xlab="Cluster")

# Salt
boxplot(segmentation2$Salt ~ kmeans_hierarchical$cluster, ylab="Salt", xlab="Cluster")

# Spreadability
boxplot(segmentation2$Spreadability ~ kmeans_hierarchical$cluster, ylab="Spreadability", xlab="Cluster")

# Appearance
boxplot(segmentation2$Appearance ~ kmeans_hierarchical$cluster, ylab="Appearance", xlab="Cluster")

# Recycling
boxplot(segmentation2$Recycling ~ kmeans_hierarchical$cluster, ylab="Recycling", xlab="Cluster")

# Bio.content
boxplot(segmentation2$Bio.content ~ kmeans_hierarchical$cluster, ylab="Bio.content", xlab="Cluster")

#PCA visualization
pca_kmeans <- prcomp(segmentation2[, 6:12])
pca_scores_kmeans <- predict(pca_kmeans, segmentation2[, 6:12])
plot(pca_scores_kmeans[, 1], pca_scores_kmeans[, 2], col = kmeans_hierarchical$cluster, pch = 16, cex = 2,
     xlab = "PC1", ylab = "PC2", main = "K-means Cluster Plot (Hierarchical Initialization)")

# demographic summary
segmentation_clean$khsegment <- kmeans_hierarchical$cluster
demographic_summary_kh <- segmentation_clean %>%
  group_by(khsegment) %>%
  summarise(
    gender_proportion_female = mean(Gender == 1, na.rm = TRUE),
    gender_proportion_male = mean(Gender ==2, na.rum= TRUE),
    age_mean = mean(Age, na.rm = TRUE),
    education_mode = names(sort(table(Education), decreasing = TRUE))[1], # Mode for Education
    area_mode = names(sort(table(Area), decreasing = TRUE))[1] # Mode for Area
```

```r
  )
demographic_summary_kh

# Mclust ----
mod5 <- densityMclust(segmentation2[,c(6:12)])
summary(mod5)
```

Appendix 2: Tables & Figures

Appendix **2.1** - Identification of Active and Passive Variables

Table 1: Classification of Variables

Column Number	Column name	Variable Type	Description
1	Student ID	Neither active nor passive	Unique student identifier
2	ID	Neither active nor passive	Unique response identifier
3	Gender	Passive	1 = female, 2 = male, 3 = other
4	Age	Passive	Number of years, continuous variable
5	Education	Passive	1 = high school profession-oriented degree, 2 = high school theory-oriented degree, 3 = higher education non-university degree, 4 = university degree, 5 = other
6	Area	Passive	1 = metropolitan, 2 = urban, 3 = suburban, 4 = countryside
7	Taste	Active	1-7, 1 = mild flavor, 7 = strong flavor
8	Fattiness	Active	1-7, 1 = low fat level, 7 = high fat level
9	Salt	Active	1-7, 1 = low salt level, 7 = high salt level
10	Spreadability	Active	1-7, 1 = dense structure, 7 = loose structure
11	Appearance	Active	1-7, 1 = low attractiveness, 7 = high attractiveness
12	Recycling	Active	1-7, 1 = low environmental-friendliness, 7 = high environmental-friendliness
13	Bio-content	Active	1-7, 1 = low organic/bio level, 7 = high organic/bio level

Explanation:

The passive variables are demographic. They describe characteristics that are inherent to the respondents which can later be used in the analysis to characterize or interpret the clusters or groups that emerge from the cluster analysis. For example, once clusters are formed based on active variables, demographic information can help describe the type of individuals in each cluster, such as gender distribution or average age in a particular cluster.

The active variables will be standardized to ensure that they are on the same scale before performing the cluster analysis. These will serve as the input for the cluster analysis to help group observations based on similarity in traits.

Appendix **2.1** - Correlation Matrix

Table 2: Correlation matrix to confirm absence of multicollinearity

	Taste	Fattiness	Salt	Spreadability	Appearance	Recycling	Bio.content
Taste	1.00000000	0.27391478	0.33317170	0.11554411	0.21789574	-0.02520841	-0.09939103
Fattiness	0.27391478	1.00000000	0.65640104	-0.19858250	-0.06143657	0.14307828	-0.16960208
Salt	0.33317170	0.65640104	1.00000000	-0.28401678	-0.02778909	0.01817045	-0.16304052
Spreadability	0.11554411	-0.19858250	-0.28401678	1.00000000	0.24694597	0.04632358	-0.04380417
Appearance	0.21789574	-0.06143657	-0.02778909	0.24694597	1.00000000	0.38065262	0.18340091
Recycling	-0.02520841	0.14307828	0.01817045	0.04632358	0.38065262	1.00000000	0.35353473
Bio.content	-0.09939103	-0.16960208	-0.16304052	-0.04380417	0.18340091	0.35353473	1.00000000

Appendix 2.2: Data Standardization Check

Table 3: Summary of standardized active variables and normal passive variables

```
    Student            Gender            Age          Education          Area
 Min.   :4766946   Min.   :1.000   Min.   :18.0   Min.   :1.000   Min.   :1.000
 1st Qu.:4766946   1st Qu.:1.000   1st Qu.:23.0   1st Qu.:3.000   1st Qu.:2.000
 Median :4766946   Median :2.000   Median :42.0   Median :4.000   Median :3.000
 Mean   :4766946   Mean   :1.681   Mean   :41.6   Mean   :3.311   Mean   :2.512
 3rd Qu.:4766946   3rd Qu.:2.000   3rd Qu.:58.0   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :4766946   Max.   :3.000   Max.   :65.0   Max.   :5.000   Max.   :4.000
     Taste           Fattiness            Salt           Spreadability
 Min.   :-2.6210   Min.   :-2.3060   Min.   :-2.66881   Min.   :-2.75975
 1st Qu.:-0.7801   1st Qu.:-0.6289   1st Qu.:-0.94338   1st Qu.:-0.88434
 Median : 0.1403   Median : 0.2096   Median :-0.08066   Median : 0.05337
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 1.0607   3rd Qu.: 1.0482   3rd Qu.: 0.78206   3rd Qu.: 0.99107
 Max.   : 2.9016   Max.   : 2.7253   Max.   : 2.50749   Max.   : 2.86648
   Appearance         Recycling        Bio.content            ID
 Min.   :-2.88283   Min.   :-2.4641   Min.   :-3.09939   Min.   :  52
 1st Qu.:-1.01412   1st Qu.:-0.7096   1st Qu.:-1.06754   1st Qu.:1798
 Median :-0.07976   Median : 0.1676   Median :-0.05162   Median :3545
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   :3544
 3rd Qu.: 0.85459   3rd Qu.: 1.0448   3rd Qu.: 0.96430   3rd Qu.:5278
 Max.   : 2.72331   Max.   : 2.7993   Max.   : 2.99615   Max.   :7038
```

Appendix **2.3** - Euclidean Distance Matrix

Table 4: Euclidean Distance Matrix [10:10]

	1	2	3	4	5	6	7	8	9	10
1	0.000000	2.855299	2.427667	3.932945	2.723498	1.284068	4.191965	1.786879	2.880405	1.878327
2	2.855299	0.000000	3.964606	3.702287	1.558736	2.228187	3.339662	1.835494	1.245139	3.605613
3	2.427667	3.964606	0.000000	5.240905	4.447237	3.013550	3.592705	2.414638	3.725551	2.560870
4	3.932945	3.702287	5.240905	0.000000	3.442856	3.535481	4.983073	3.662795	3.056765	4.455880
5	2.723498	1.558736	4.447237	3.442856	0.000000	2.447060	4.519367	2.394679	2.065920	3.964606
6	1.284068	2.228187	3.013550	3.535481	2.447060	0.000000	3.553319	1.817328	2.260268	2.226535
7	4.191965	3.339662	3.592705	4.983073	4.519367	3.553319	0.000000	3.074379	2.760545	4.200333
8	1.786879	1.835494	2.414638	3.662795	2.394679	1.817328	3.074379	0.000000	1.795839	2.562375
9	2.880405	1.245139	3.725551	3.056765	2.065920	2.260268	2.760545	1.795839	0.000000	3.383796
10	1.878327	3.605613	2.560870	4.455880	3.964606	2.226535	4.200333	2.562375	3.383796	0.000000

Appendix **2.4** - Hierarchical Clustering Types (1/5)

Figure 1: Single Linkage Dendrogram

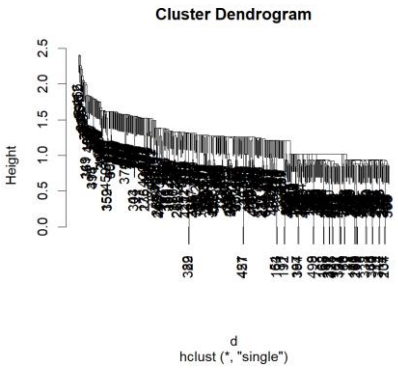


Table 5: Single Linkage Merge Table [1:15]

	merge.1	merge.2	height
1	-8	-327	0
2	-9	-490	0
3	-10	-66	0
4	-11	-190	0
5	-14	-337	0
6	-52	-161	0
7	-62	-335	0
8	-63	-110	0
9	-265	8	0
10	-80	-324	0
11	-89	-172	0
12	-329	11	0
13	-103	-171	0
14	-135	-201	0
15	-160	-166	0

Table 6: Single Linkage Merge Table [475:491]

	merge.1	merge.2	height
475	-442	474	1.828598
476	-36	-77	1.828598
477	462	475	1.830850
478	-94	477	1.837893
479	383	478	1.839597
480	476	479	1.842038
481	-153	480	1.985820
482	-342	481	2.023930
483	-195	482	2.023930
484	-458	483	2.049334
485	-309	484	2.057013
486	-83	485	2.110761
487	-152	486	2.207413
488	-293	487	2.230601
489	-244	488	2.261781
490	-12	489	2.395987
491	-459	490	2.406213

Appendix **2.5** - Hierarchical Clustering Types (2/5)

Figure 2: Complete Linkage Dendrogram

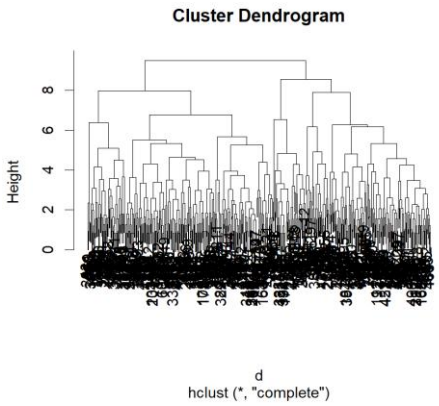


Table 7: Complete Linkage Merge Table [1:15]

	merge.1	merge.2	height
1	-8	-327	0
2	-9	-490	0
3	-10	-66	0
4	-11	-190	0
5	-14	-337	0
6	-52	-161	0
7	-62	-335	0
8	-63	-110	0
9	-265	8	0
10	-80	-324	0
11	-89	-172	0
12	-329	11	0
13	-103	-171	0
14	-135	-201	0
15	-160	-166	0

Table 8: Complete Linkage Merge Table [475:49

	merge.1	merge.2	height
475	428	467	4.911411
476	461	469	5.091449
477	458	468	5.254623
478	465	471	5.335215
479	466	472	5.345458
480	449	479	5.531074
481	442	477	5.668077
482	451	475	6.167862
483	473	478	6.172152
484	456	483	6.290939
485	368	476	6.389976
486	374	474	6.545578
487	480	481	6.770915
488	482	484	7.897316
489	485	487	7.979768
490	486	488	8.549427
491	489	490	9.503158

Appendix **2.6** - Hierarchical Clustering Types (3/5)

Figure 3: Average Linkage Dendrogram

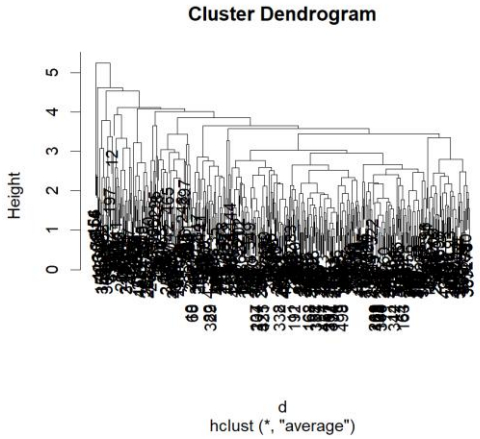


Table 9: Average Linkage Merge Table [1:15]

	merge.1	merge.2	height
1	-8	-327	0
2	-9	-490	0
3	-10	-66	0
4	-11	-190	0
5	-14	-337	0
6	-52	-161	0
7	-62	-335	0
8	-63	-110	0
9	-265	8	0
10	-80	-324	0
11	-89	-172	0
12	-329	11	0
13	-103	-171	0
14	-135	-201	0
15	-160	-166	0

Table 10: Average Linkage Merge Table [475:49

	merge.1	merge.2	height
475	448	471	3.334704
476	390	462	3.347831
477	423	466	3.378548
478	469	472	3.432479
479	468	476	3.442473
480	474	479	3.583242
481	463	480	3.648901
482	459	464	3.651131
483	470	477	3.777225
484	478	481	3.851825
485	-12	475	3.930843
486	482	484	4.012829
487	473	486	4.091000
488	485	487	4.106827
489	483	488	4.538980
490	424	489	4.622918
491	431	490	5.241496

Appendix **2.7** - Hierarchical Clustering Types (4/5)

Figure 4: Centroid Linkage Dendrogram

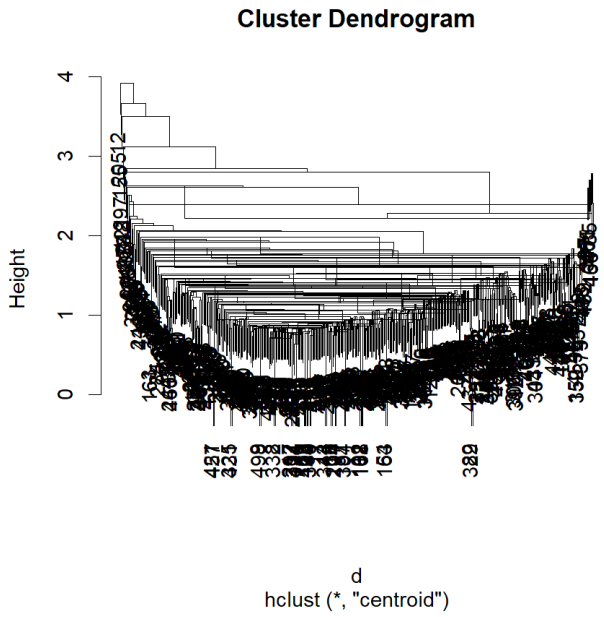


Table 11: Centroid Linkage Merge Table [1:15]

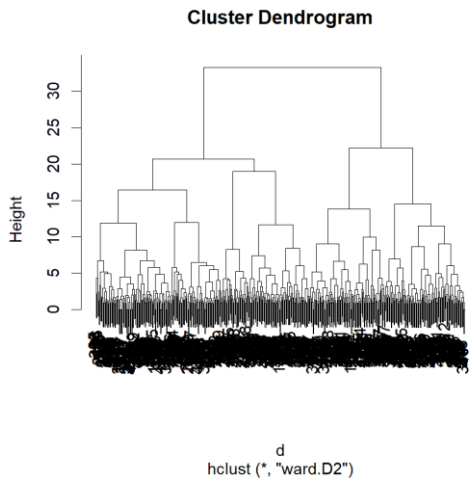
	merge.1	merge.2	height
1	-8	-327	0
2	-9	-490	0
3	-10	-66	0
4	-11	-190	0
5	-14	-337	0
6	-52	-161	0
7	-62	-335	0
8	-63	-110	0
9	-265	8	0
10	-80	-324	0
11	-89	-172	0
12	-329	11	0
13	-103	-171	0
14	-135	-201	0
15	-160	-166	0

Table 12: Centroid Linkage Merge Table [475:4

	merge.1	merge.2	height
475	-152	-459	2.406213
476	-442	474	2.428472
477	-398	476	2.511898
478	-195	-342	2.596381
479	477	478	2.391026
480	-263	479	2.611170
481	-470	480	2.628911
482	-183	-395	2.702283
483	481	482	2.214762
484	-83	475	2.779482
485	483	484	2.277499
486	-244	485	2.801428
487	-119	486	2.843873
488	-293	487	3.114696
489	-153	488	3.498761
490	-202	489	3.662341
491	-12	490	3.920087

Appendix **2.8**- Hierarchical Clustering Types (5/5)

Figure 5: Ward Linkage Dendrogram



Explanation: Dendogram of ward method is much clearear than other linkage methods. Biggest jump of merging cost in merge table can be seen between the 2nd and 3rd merge, indicating 2 clusters is ideal.

Table 13: Ward Linkage Merge Table [1:10]

	merge.1	merge.2	height
1	-8	-327	0
2	-9	-490	0
3	-10	-66	0
4	-11	-190	0
5	-14	-337	0
6	-52	-161	0
7	-62	-335	0
8	-63	-110	0
9	-265	8	0
10	-80	-324	0
11	-89	-172	0
12	-329	11	0
13	-103	-171	0
14	-135	-201	0
15	-160	-166	0

Table 14: WardLinkage Merge Table [475:491]

	merge.1	merge.2	height
475	437	454	8.226649
476	455	464	8.423403
477	415	458	8.606625
478	453	468	9.030566
479	450	463	9.046058
480	466	474	9.955024
481	477	479	11.478969
482	472	476	11.680134
483	469	473	11.872867
484	461	467	11.973118
485	478	480	13.821387
486	471	481	14.498294
487	483	484	16.425787
488	475	482	18.945029
489	487	488	20.684672
490	485	486	22.226570
491	489	490	33.265948

Appendix **2.9** - Scree Plots

Figure 6: Scree Plot of Single Hierarchical Clustering

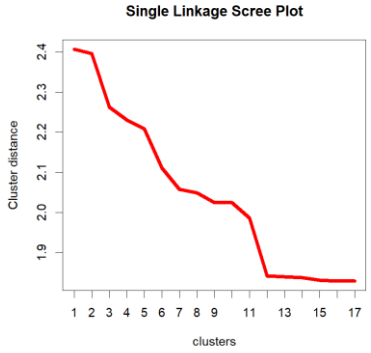


Figure 7: Scree Plot of Complete Hierarchical Clustering

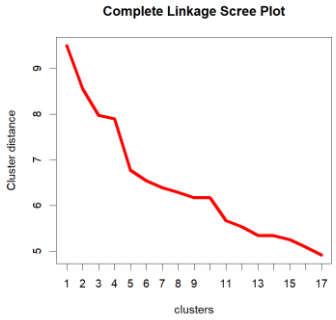


Figure 8: Scree Plot of Average Hierarchical Clustering

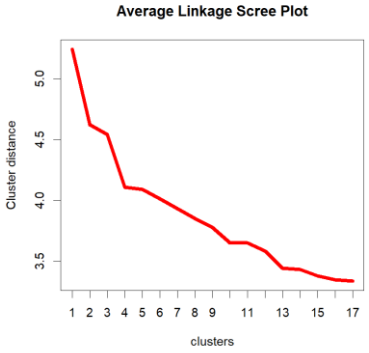


Figure 9: Scree Plot of Centroid Hierarchical Clustering

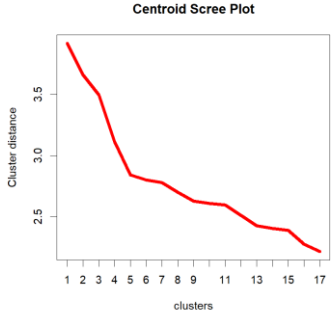
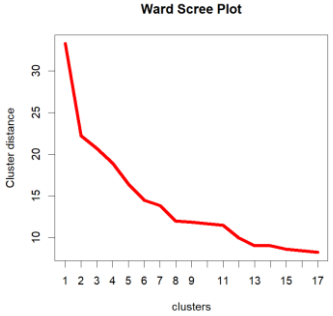


Figure 10: Scree Plot of Ward Hierarchical Clustering



Appendix **2.10** - Dendograms for Ward

Figure 11: Dendrogram Ward 6 clusters

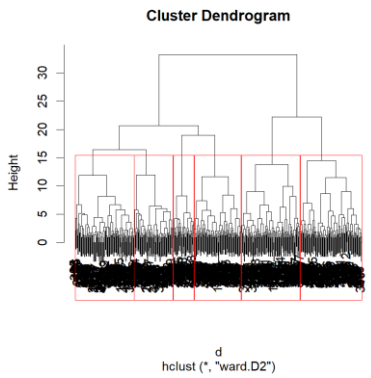


Figure 12: Dendrogram Ward 5 clusters

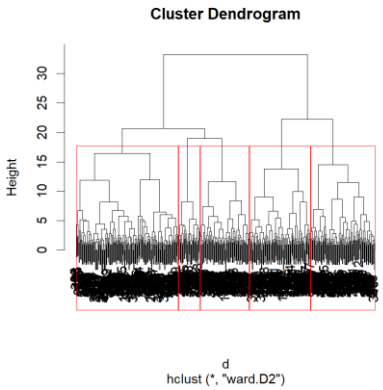


Figure 13: Dendrogram Ward 4 clusters

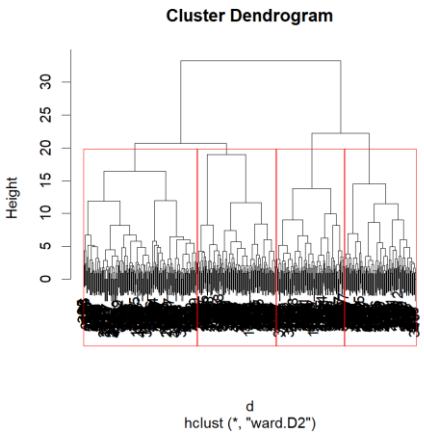


Figure 14: Dendrogram Ward 3 clusters

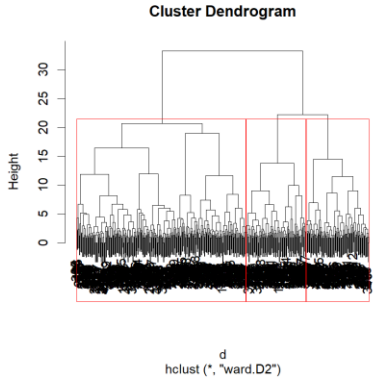
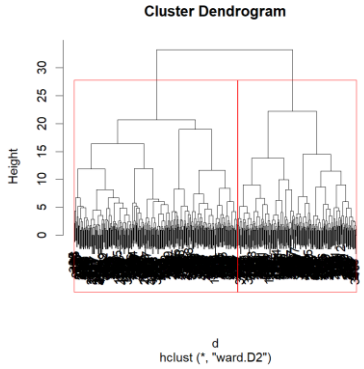


Figure 15: Dendrogram Ward 2 clusters



Appendix **2.11** - Distribution of observations across clusters for Ward

Table 15: Distribution of observations across 2 clusters

```
segmentation2e.hc.segment
  1   2
207 285
```

Table 16: Distribution of observations across 3 clusters

```
segmentation2e.hc.segment
  1   2   3
101 285 106
```

Appendix **2.12** - Segment means for ward 2 and 3 clusters

Table 17: Segment Means Ward 2 cluster

	Group.1	Taste	Fattiness	Salt	Spreadability	Appearance	Recycling	Bio.content
1	1	0.3359550	0.6268961	0.7070366	-0.3588619	-0.2828830	-0.2561836	-0.5178668
2	2	-0.2440094	-0.4553246	-0.5135319	0.2606471	0.2054624	0.1860702	0.3761348

Table 18: Segment Means Ward 3 cluster

	Group.1	Taste	Fattiness	Salt	Spreadability	Appearance	Recycling	Bio.content
1	1	-0.1968769	0.4006003	0.4147606	-0.5036863	-1.0603735	-0.6922632	-0.6450818
2	2	-0.2440094	-0.4553246	-0.5135319	0.2606471	0.2054624	0.1860702	0.3761348
3	3	0.8436533	0.8425177	0.9855261	-0.2208688	0.4579334	0.1593262	-0.3966524

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 19: ANOVA & Tukey Test for Taste, 2 clusters

```
> summary(segmentation2.aov_Taste2)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember2   1   40.3   40.33   43.85 9.33e-11 ***
Residuals   490  450.7    0.92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Taste2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Taste ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
          diff        lwr        upr p adj
2-1 -0.5799644 -0.7520431 -0.4078856     0
```

Table 20: ANOVA & Tukey Test for Taste, 3 clusters

```
> summary(segmentation2.aov_Taste3)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember3   2   96.3   48.16   59.68 <2e-16 ***
Residuals   489  394.7    0.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Taste3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Taste ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
            diff        lwr       upr     p adj
2-1 -0.04713246 -0.2917019 0.1974369 0.8930724
3-1  1.04053020  0.7468578 1.3342026 0.0000000
3-2  1.08766266  0.8473899 1.3279355 0.0000000
```

The difference between clusters 2 and 1 for the Taste variable in the 3 cluster solution is insignificant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 21: ANOVA & Tukey Test for Fattiness, 2 clusters

```
> summary(segmentation2.aov_Fattiness2)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember2   1  140.4  140.44   196.3 <2e-16 ***
Residuals   490  350.6    0.72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Fattiness2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fattiness ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff       lwr       upr p adj
2-1 -1.082221 -1.233989 -0.930452     0
```

Table 22: ANOVA & Tukey Test for Fattiness, 3 clusters

```
> summary(segmentation2.aov_Fattiness3)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember3   2  150.5   75.27   108.1 <2e-16 ***
Residuals   489  340.5    0.70
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Fattiness3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fattiness ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
          diff        lwr        upr     p adj
2-1 -0.8559248 -1.0830783 -0.6287713 0.0000000
3-1  0.4419174  0.1691576  0.7146773 0.0004615
3-2  1.2978422  1.0746794  1.5210051 0.0000000
```

Here all differences between clusters are significant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 23: ANOVA & Tukey Test for Salt, 2 clusters

```
> summary(segmentation2.aov_Salt2)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember2   1  178.6  178.64   280.2 <2e-16 ***
Residuals   490  312.4    0.64
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Salt2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salt ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff      lwr       upr p adj
2-1 -1.220569 -1.36383 -1.077308     0
```

Table 24: ANOVA & Tukey Test for Salt, 3 clusters

```
> summary(segmentation2.aov_Salt3)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember3   2  195.5   97.74   161.7 <2e-16 ***
Residuals   489  295.5    0.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Salt3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salt ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
          diff        lwr        upr p adj
2-1 -0.9282924 -1.1399203 -0.7166645 0e+00
3-1  0.5707656  0.3166484  0.8248827 6e-07
3-2  1.4990580  1.2911480  1.7069680 0e+00
```

Here all differences between clusters are significant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 25: ANOVA & Tukey Test for Spreadability, 2 clusters

```
> summary(segmentation2.aov_Spreadability2)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember2   1     46   46.02   50.68 3.9e-12 ***
Residuals   490    445    0.91
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Spreadability2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Spreadability ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff       lwr       upr p adj
2-1 0.6195089 0.4485195 0.7904984     0
```

Table 26: ANOVA & Tukey Test for Spreadability, 3 clusters

```
> summary(segmentation2.aov_Spreadability3)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember3   2   50.2  25.078   27.82 3.61e-12 ***
Residuals   489  440.8   0.902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Spreadability3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Spreadability ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
          diff         lwr        upr     p adj
2-1  0.7643334  0.50585340  1.0228133 0.0000000
3-1  0.2828175 -0.02755832  0.5931934 0.0825638
3-2 -0.4815158 -0.73545482 -0.2275769 0.0000306
```

Here all differences between clusters are significant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 27: ANOVA & Tukey Test for Appearance, 2 clusters

```
> summary(segmentation2.aov_Appearance2)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember2   1   28.6  28.596    30.3 5.97e-08 ***
Residuals   490  462.4   0.944
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Appearance2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Appearance ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff       lwr       upr p adj
2-1 0.4883454 0.3140404 0.6626504 1e-07
```

Table 28: ANOVA & Tukey Test for Appearance, 3 clusters

```
> summary(segmentation2.aov_Appearance3)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember3   2  147.8   73.91   105.3 <2e-16 ***
Residuals   489  343.2    0.70
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Appearance3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Appearance ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
         diff        lwr       upr    p adj
2-1 1.265836 1.03777876 1.4938930 0.000000
3-1 1.518307 1.24446198 1.7921518 0.000000
3-2 0.252471 0.02842039 0.4765216 0.022644
```

Here all differences between clusters are significant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 29: ANOVA & Tukey Test for Recycling , 2 clusters

```
> summary(segmentation2.aov_Recycling2)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember2   1   23.5  23.453   24.58 9.85e-07 ***
Residuals   490  467.5   0.954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Recycling2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Recycling ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff       lwr       upr p adj
2-1 0.4422538 0.2669821 0.6175255 1e-06
```

Table 30: ANOVA & Tukey Test for Recycling, 3 clusters

```
> summary(segmentation2.aov_Recycling3)
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmember3   2     61  30.480   34.66 8.38e-15 ***
Residuals   489    430   0.879
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Recycling3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Recycling ~ clusmember3, data = segmentation2eaovbase3)

$clusmember3
           diff        lwr       upr     p adj
2-1  0.87833337  0.6230402 1.1336265 0.0000000
3-1  0.85158943  0.5450403 1.1581386 0.0000000
3-2 -0.02674395 -0.2775521 0.2240642 0.9659555
```

The difference between clusters 3 and 2 in the 3 cluster solution are insignificant.

Appendix **2.13** - ANOVA & Tukey Tests for categorical variables, 2 vs. 3 clusters

Table 31: ANOVA & Tukey Test for Bio.content, 2 clusters

```
> summary(segmentation2.aov_Bio.content2)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember2   1   95.8   95.84   118.8 <2e-16 ***
Residuals   490  395.2    0.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Bio.content2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bio.content ~ clusmember2, data = segmentation2eaovbase2)

$clusmember2
         diff       lwr      upr p adj
2-1 0.8940016 0.7328673 1.055136     0
```

Table 32: ANOVA & Tukey Test for Bio.content, 3 clusters

```
> summary(segmentation2.aov_Bio.content3)
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember3   2     99   49.51   61.77 <2e-16 ***
Residuals   489    392    0.80
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_Bio.content3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bio.content ~ clusmember3, data = segmentation2eaovb

$clusmember3
          diff         lwr        upr     p adj
2-1  1.0212166  0.77748463  1.2649486 0.0000000
3-1  0.2484294 -0.04423745  0.5410963 0.1143642
3-2 -0.7727872 -1.01223731 -0.5333371 0.0000000
```

The difference between clusters 3 and 1 in the 3 cluster solution are insignificant.

Appendix **2.14** - Significance Tests for passive variables

Table 33: Gender

```
> segmentation2e.freq.gender

clusmember2   1   2   3
          1  32 172   3
          2 137 139   9
> segmentation2e.chisq.gender <- chisq.test(segmentation2e.freq.gender)
> segmentation2e.chisq.gender

        Pearson's Chi-squared test

data:  segmentation2e.freq.gender
X-squared = 60.903, df = 2, p-value = 5.957e-14
```

Table 34: Education Segment Frequency

```
> segmentation2e.freq.education

clusmember2   1   2   3   4   5
          1  22  64  63  45  13
          2   0  18  66 194   7
> segmentation2e.chisq.education <- chisq.test(segmentation2e.freq.education)
> segmentation2e.chisq.education

        Pearson's Chi-squared test

data:  segmentation2e.freq.education
X-squared = 133.56, df = 4, p-value < 2.2e-16
```

Table 35: Area

```
> segmentation2e.freq.area

clusmember2   1   2   3   4
          1  34  75  65  33
          2  46  84 109  46
> segmentation2e.chisq.area <- chisq.test(segmentation2e.freq.area)
> segmentation2e.chisq.area

        Pearson's Chi-squared test

data:  segmentation2e.freq.area
X-squared = 3.292, df = 3, p-value = 0.3488
```

Table 36: Age

```
             Df Sum Sq Mean Sq F value Pr(>F)
clusmember2   1   1984  1983.6   6.639 0.0103 *
Residuals   490 146397   298.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(segmentation2.aov_age)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Age ~ clusmember2, data = segmentationwithage)

$clusmember2
        diff       lwr      upr     p adj
2-1 4.067226 0.9657762 7.168676 0.0102678
```

Appendix **2.15** - PCA Results for Ward Segmentation

Figure 16: PCA Cluster Plot ward 2 clusters

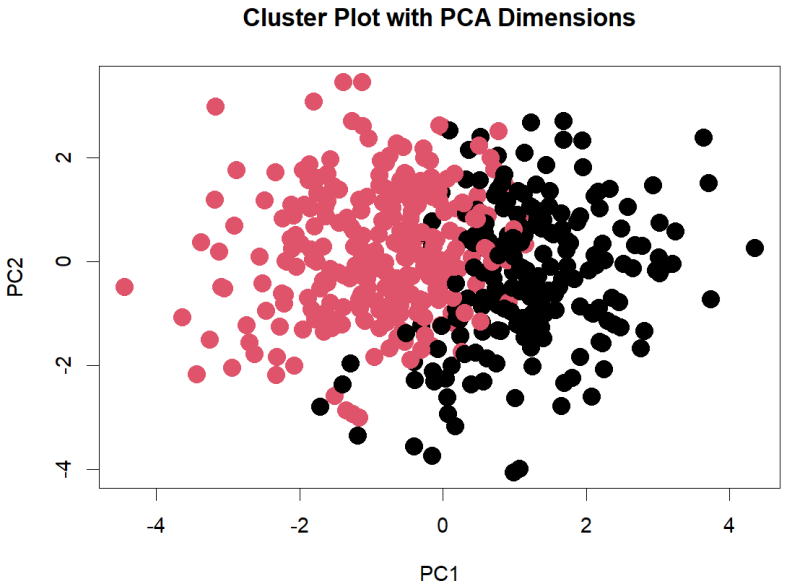
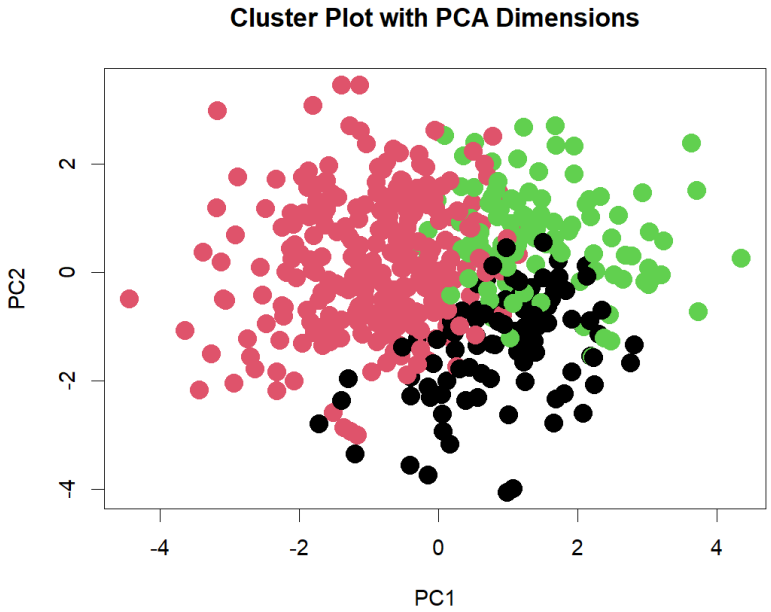


Figure 17: PCA Cluster Plot ward 3 clusters



Appendix **2.16** - Demographic Analysis of Ward analysis

Table 37: Demographic description of 2 cluster ward solution

	segment_ward	gender_proportion_female	gender_proportion_male	age_mean	education_mode	area_mode
	<int>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	1	0.155	0.831	39.2	2	2
2	2	0.481	0.488	43.3	4	3

Appendix **2.17** - Silhouette

Figure 18: Silhouette Graph

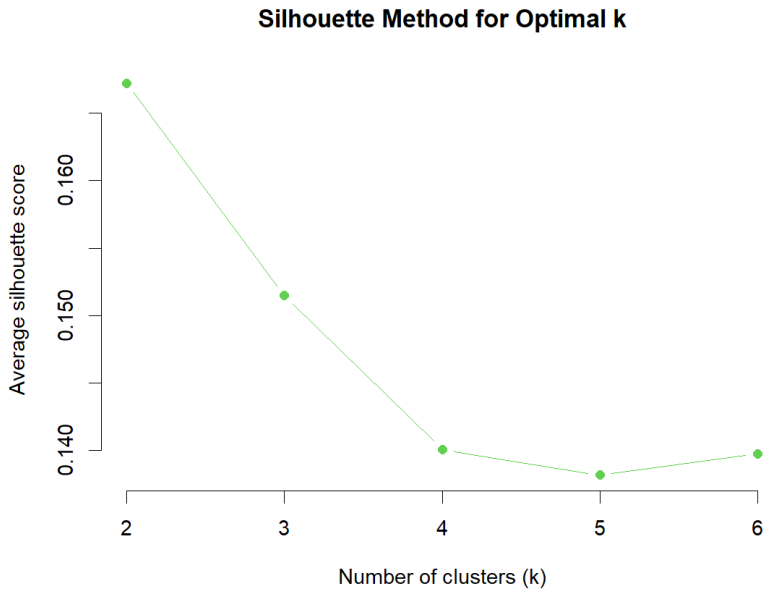


Table 36: Silhouette Scores

```
> silhouette_scores
[1] 0.1672258 0.1514647 0.1400424 0.1381377 0.1397336
```

Explanation: In line with the hierarchical solution, 2 clusters seems to be the best solution for k-means clustering

Appendix **2.18** - BIC

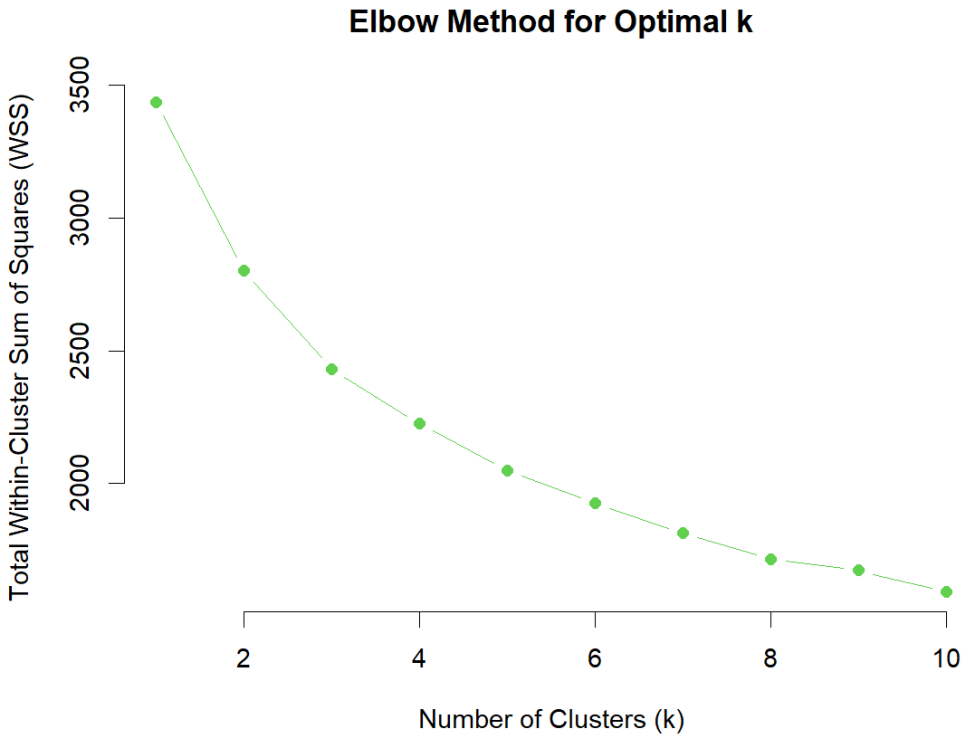
Table 37: BIC Values for different clusters

	clusters	BIC
1	2	1085.451
2	3	1382.176
3	4	1678.902
4	5	1975.627

Explanation: In line with the hierarchical solution and silhouette values, the BIC table shows that the BIC is lowest for a 2 cluster solution, indicating that a 2 cluster solution has the best fit.

Appendix **2.19** - Scree plot for k-means

Figure 19: Scree plot



Explanation: In line with the hierarchical solution and silhouette values, and the BIC table, the scree plot confirms that 2 clusters is an optimal amount. It can be see that the WSS drops significantly from 2 to 3 clusters.

Appendix **2.20** - Segment Means K-means 2 clusters

Table 38: Segment Means k-means 2 clusters

	Group.1	Taste	Fattiness	Salt	Spreadability	Appearance	Recycling	Bio.content
1	1	-0.2565415	-0.6319599	-0.6370496	0.2164441	0.1030467	0.05000154	0.2833379
2	2	0.3278030	0.8075043	0.8140079	-0.2765675	-0.1316708	-0.06389085	-0.3620429

Table 39: Distribution of observations across segments

1	2
276	216

Appendix **2.21** - ANOVA & TukeyHSD for k-means

Table 40: Taste ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberk   1   41.4   41.37   45.09 5.23e-11 ***
Residuals   490  449.6    0.92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_taste)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Taste ~ clusmemberk, data = aovbasek)

$clusmemberk
         diff       lwr       upr p adj
2-1 0.5843445 0.4133625 0.7553266     0
```

Table 41: Fattiness ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value Pr(>F)
clusmemberk   1  251.1  251.07   512.8 <2e-16 ***
Residuals   490  239.9    0.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_fattiness)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fattiness ~ clusmemberk, data = aovbasek)

$clusmemberk
        diff      lwr      upr p adj
2-1 1.439464 1.314563 1.564365     0
```

Table 42: Salt ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value Pr(>F)
clusmemberk   1  255.1  255.13     530 <2e-16 ***
Residuals   490  235.9    0.48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_salt)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salt ~ clusmemberk, data = aovbasek)

$clusmemberk
        diff      lwr      upr p adj
2-1 1.451057 1.327218 1.574897     0
```

Table 43: Spreadability ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberk   1   29.5  29.452   31.27 3.74e-08 ***
Residuals   490  461.5   0.942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_spreadability)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Spreadability ~ clusmemberk, data = aovbasek)

$clusmemberk
          diff        lwr        upr p adj
2-1 -0.4930116 -0.6662458 -0.3197773     0
```

Appendix **2.21** - ANOVA & TukeyHSD for k-means

Table 44: Appearance ANOVA and TukeyHSD

```
             Df Sum Sq Mean Sq F value  Pr(>F)
clusmemberk   1    6.7   6.676   6.754 0.00964 **
Residuals   490  484.3   0.988
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_appearance)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Appearance ~ clusmemberk, data = aovbasek)

$clusmemberk
          diff        lwr         upr     p adj
2-1 -0.2347175 -0.4121746 -0.05726037 0.0096367
```

Table 45: Recycling ANOVA and TukeyHSD

```
             Df Sum Sq Mean Sq F value Pr(>F)
clusmemberk   1    1.6  1.5718   1.574   0.21
Residuals   490  489.4  0.9988
> TukeyHSD(k_recycling)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Recycling ~ clusmemberk, data = aovbasek)

$clusmemberk
          diff        lwr       upr     p adj
2-1 -0.1138924 -0.2922821 0.0644973 0.2102831
```

Table 46: Bio.content ANOVA and TukeyHSD

```
             Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberk   1   50.5   50.47   56.14 3.18e-13 ***
Residuals   490  440.5    0.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(k_bio.content)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bio.content ~ clusmemberk, data = aovbasek)

$clusmemberk
          diff        lwr        upr p adj
2-1 -0.6453808 -0.8146247 -0.4761368     0
```

Appendix **2.22** - Significance Tests for passive variables

Table 47: Gender ChiSquare

```
kclus2   1   2   3
     1 144 123   9
     2  25 188   3
> segmentation2e2.chisq.gender <- chisq.test(seg
> segmentation2e2.chisq.gender

        Pearson's Chi-squared test

data:  segmentation2e2.freq.gender
X-squared = 94.466, df = 2, p-value < 2.2e-16
```

Table 48: Education ChiSquare

```
kclus2   1   2   3   4   5
     1   0   2  57 210   7
     2  22  80  72  29  13
> segmentation2e2.chisq.education <- chisq.test(seg
> segmentation2e2.chisq.education

        Pearson's Chi-squared test

data:  segmentation2e2.freq.education
X-squared = 232.96, df = 4, p-value < 2.2e-16
```

Explanation: Area and Age are insignificant

Table 49: Area ChiSquare

```
kclus2  1  2  3  4
     1 47 87 98 44
     2 33 72 76 35
> segmentation2e2.chisq.area <- chisq.test(segme
> segmentation2e2.chisq.area

        Pearson's Chi-squared test

data:  segmentation2e2.freq.area
X-squared = 0.36031, df = 3, p-value = 0.9483
```

Table 50: Age ANOVA and TukeyHSD

```
             Df Sum Sq Mean Sq F value Pr(>F)
kclus2        1    476   475.7   1.576   0.21
Residuals   490 147905   301.8
> TukeyHSD(segmentation2e2.aov_age)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Age ~ kclus2, data = ksegwithage)

$kclus2
        diff       lwr      upr     p adj
2-1 -1.98128 -5.082383 1.119823 0.2099636
```

Appendix **2.23** - Boxplot of variables for k-mean clustering

Figure 20: K means box plot of Taste

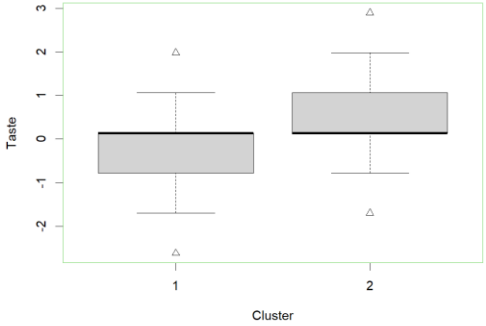


Figure 21: K means box plot of Fattiness

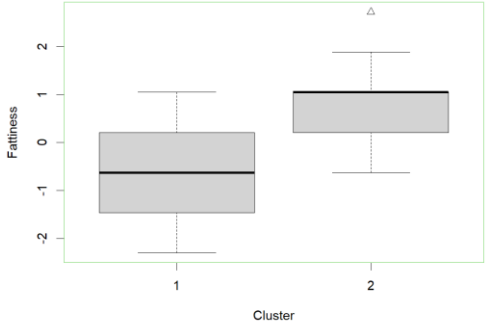


Figure 22: K means box plot of Salt

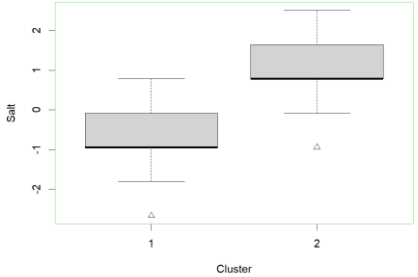
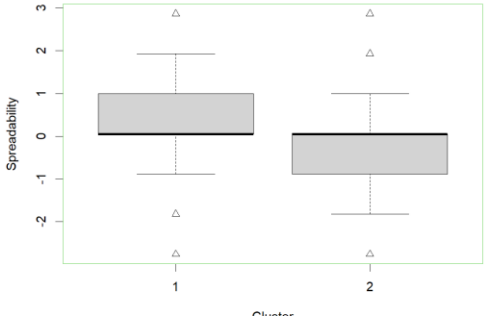


Figure 23: K means box plot of Spreadability



Appendix **2.23** - Boxplot of variables for k-mean clustering

Figure 24: K means box plot of Appearance

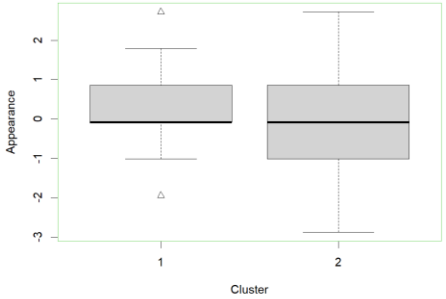


Figure 25: K means box plot of Recycling

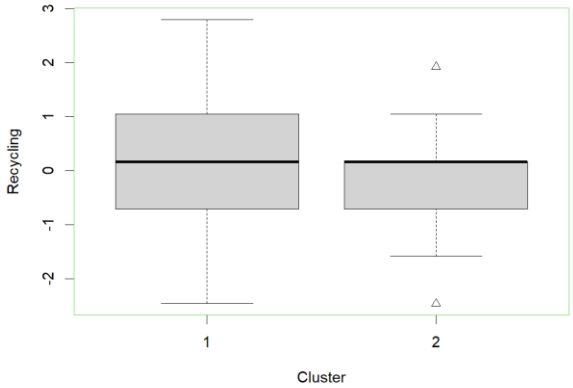
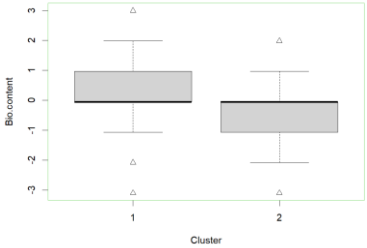
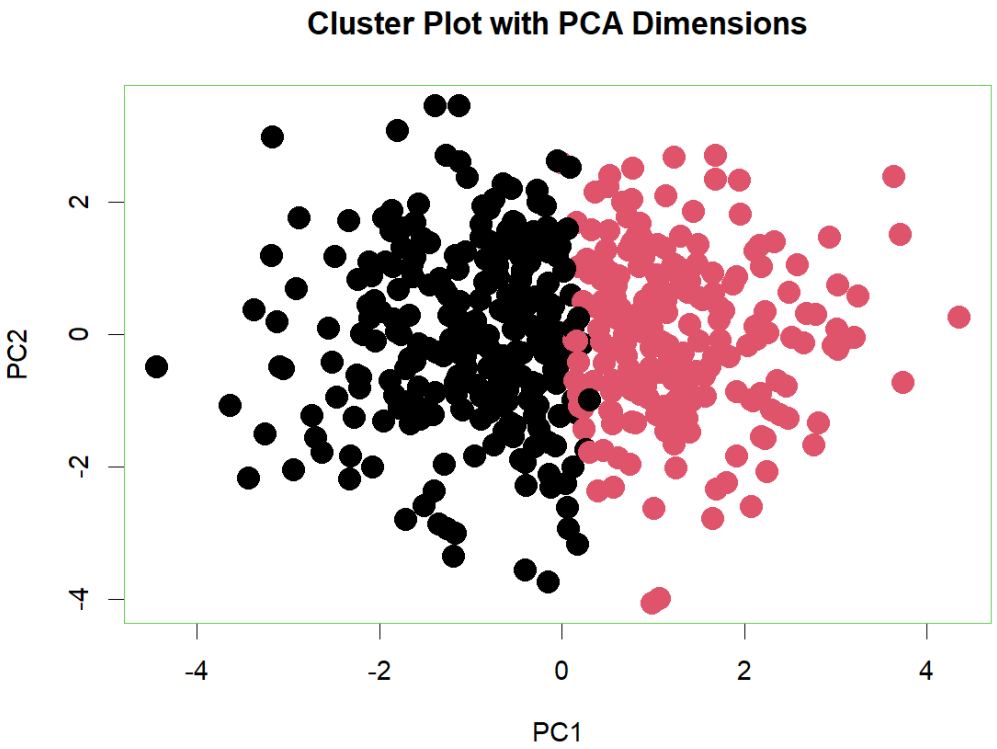


Figure 26: K means box plot of Bio.content



Appendix **2.24** - PCA Visualization for K-means Segmentation

Figure 27: PCA Cluster Plot k-means 2 clusters



Appendix **2.25** - PCA Visualization for K-means Segmentation

Table 51: Demographic summary of 2 cluster K-means solution

	segmentk2	gender_proportion_female	gender_proportion_male	age_mean	education_mode	area_mode
	<int>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	1	0.522	0.446	42.5	4	3
2	2	0.116	0.870	40.5	2	3

Appendix **2.26** - K-means using hierarchical solution as start

Table 52: Segment Means k-means 2 clusters

	Taste	Fattiness	Salt	Spreadability	Appearance	Recycling	Bio.content
1	0.2640636	0.699389	0.7095583	-0.3366879	-0.2407226	-0.1788694	-0.3418857
2	-0.2474297	-0.655333	-0.6648617	0.3154792	0.2255590	0.1676020	0.3203496

Table 53: Distribution of observations across segments

1	2
238	254

Appendix **2.27** - ANOVA & TukeyHSD for k-means

Table 54: Taste ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberkh   1   32.1   32.15   34.33 8.55e-09 ***
Residuals    490  458.9    0.94
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_taste)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Taste ~ clusmemberkh, data = aovbase)

$clusmemberkh
          diff        lwr        upr p adj
2-1 -0.5114933 -0.6830227 -0.3399639     0
```

Table 55: Fattiness ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value Pr(>F)
clusmemberkh   1  225.5  225.50   416.2 <2e-16 ***
Residuals    490  265.5    0.54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_fattiness)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fattiness ~ clusmemberkh, data = aovbase)

$clusmemberkh
         diff       lwr       upr p adj
2-1 -1.354722 -1.485199 -1.224245     0
```

Table 56: Salt ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value Pr(>F)
clusmemberkh   1  232.1  232.10   439.3 <2e-16 ***
Residuals    490  258.9    0.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_salt)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Salt ~ clusmemberkh, data = aovbase)

$clusmemberkh
        diff       lwr       upr p adj
2-1 -1.37442 -1.503264 -1.245576     0
```

Table 57: Spreadability ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberkh   1   52.3   52.26   58.37 1.15e-13 ***
Residuals    490  438.7    0.90
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_spreadability)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Spreadability ~ clusmemberkh, data = aovbase)

$clusmemberkh
         diff       lwr      upr p adj
2-1 0.6521671 0.4844392 0.819895     0
```

Appendix **2.27** - ANOVA & TukeyHSD for k-means

Table 58: Appearance ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberkh   1   26.7  26.714   28.19 1.67e-07 ***
Residuals    490  464.3   0.948
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_appearance)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Appearance ~ clusmemberkh, data = aovbase)

$clusmemberkh
         diff       lwr       upr p adj
2-1 0.4662816 0.2937399 0.6388233 2e-07
```

Table 59: Recycling ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberkh   1   14.7  14.750   15.18 0.000112 ***
Residuals    490  476.3   0.972
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_recycling)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Recycling ~ clusmemberkh, data = aovbase)

$clusmemberkh
         diff       lwr       upr     p adj
2-1 0.3464715 0.1717207 0.5212222 0.0001116
```

Table 60: Bio.content ANOVA and TukeyHSD

```
              Df Sum Sq Mean Sq F value   Pr(>F)
clusmemberkh   1   53.9   53.89   60.41 4.58e-14 ***
Residuals    490  437.1    0.89
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh_bio.content)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bio.content ~ clusmemberkh, data = aovbase)

$clusmemberkh
         diff       lwr       upr p adj
2-1 0.6622353 0.4948185 0.8296521     0
```

Appendix **2.28** - Significance Tests for passive variables

Table 61: Gender ChiSquare

```
clusmemberkh   1   2   3
           1  36 198   4
           2 133 113   8
> kh.chisq.gender <- chisq.test(kh.freq.gender)
> kh.chisq.gender

        Pearson's Chi-squared test

data:  kh.freq.gender
X-squared = 79.803, df = 2, p-value < 2.2e-16
```

Table 62: Education ChiSquare

```
clusmemberkh   1   2   3   4   5
           1  22  79  78  46  13
           2   0   3  51 193   7
> kh.chisq.education <- chisq.test(kh.freq.education)
> kh.chisq.education

        Pearson's Chi-squared test

data:  kh.freq.education
X-squared = 189.99, df = 4, p-value < 2.2e-16
```

Table 63: Area ChiSquare

```
clusmemberkh  1  2  3  4
           1 38 82 82 36
           2 42 77 92 43
> kh.chisq.area <- chisq.test(kh.freq.area)
> kh.chisq.area

        Pearson's Chi-squared test

data:  kh.freq.area
X-squared = 1.033, df = 3, p-value = 0.7933
```

Table 64: Age ANOVA and TukeyHSD

```
             Df Sum Sq Mean Sq F value Pr(>F)
clusmemberkh   1    920   920.0   3.057  0.081 .
Residuals    490 147460   300.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(kh.aov_age)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Age ~ clusmemberkh, data = kh_age)

$clusmemberkh
        diff        lwr      upr     p adj
2-1 2.736386 -0.3385697 5.811342 0.0810072
```

Appendix **2.29** - Boxplot of variables for k-mean clustering

Figure 28: K means box plot of Taste

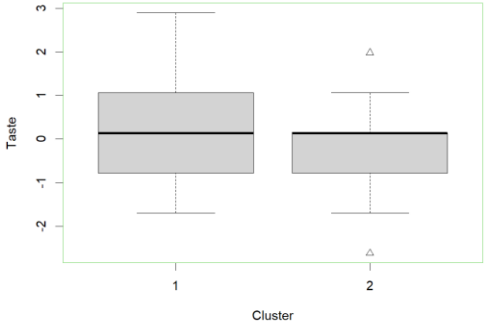


Figure 29: K means box plot of Fattiness

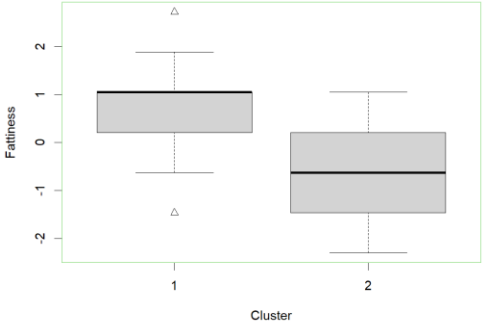


Figure 30: K means box plot of Salt

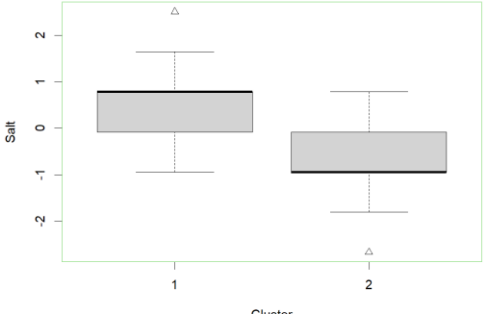
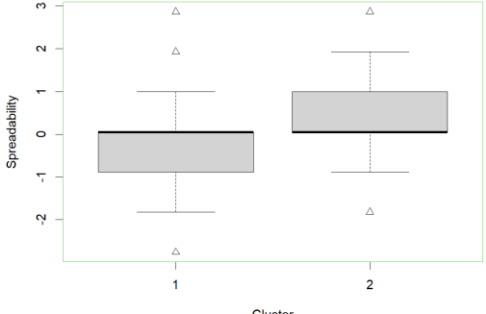


Figure 31: K means box plot of Spreadability



Appendix **2.29** - Boxplot of variables for k-mean clustering

Figure 32: K means box plot of Appearance

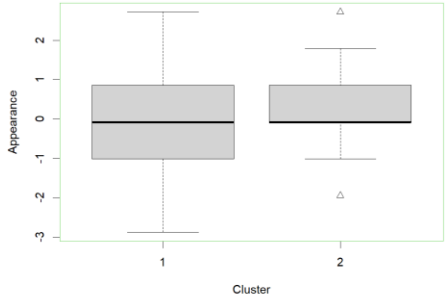


Figure 33: K means box plot of Recycling

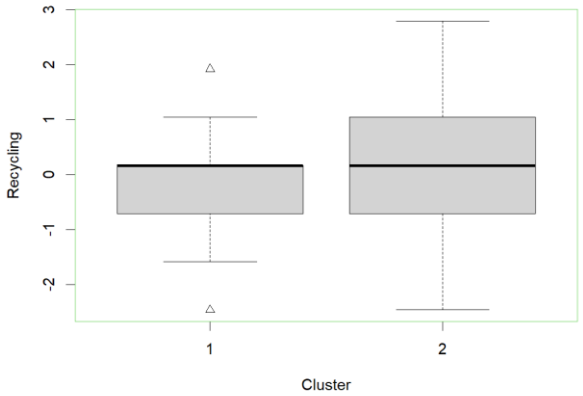
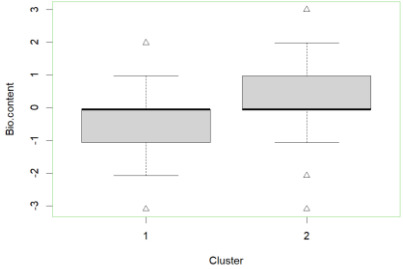
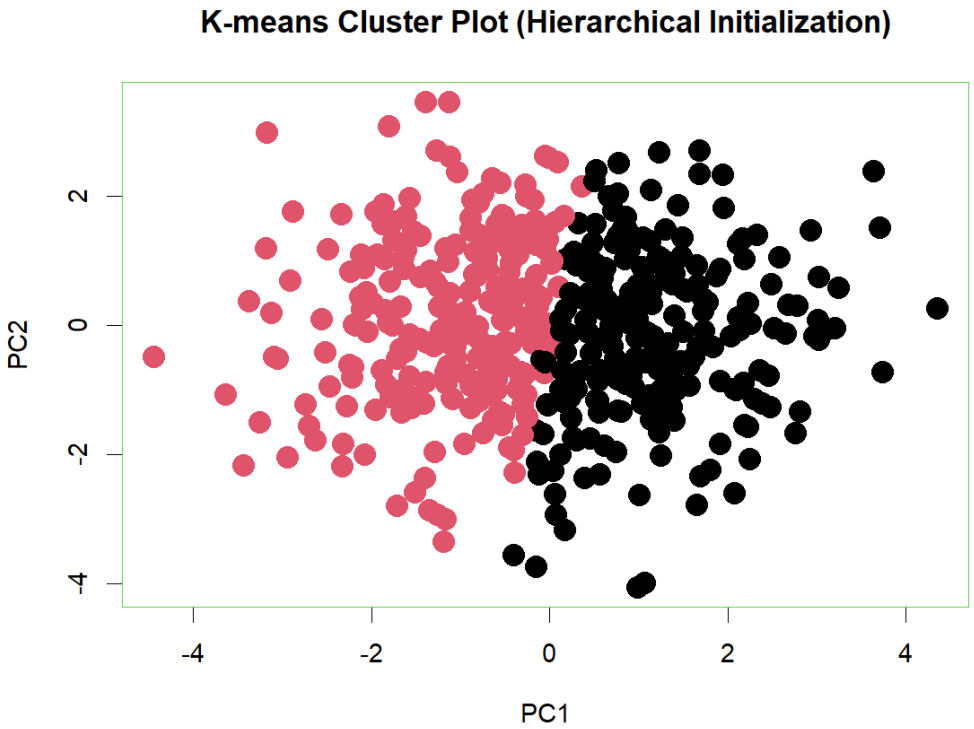


Figure 34: K means box plot of Bio.content



Appendix **2.30** - PCA Visualizaiton of combined approach

Figure 35: PCA Visualization combined approach



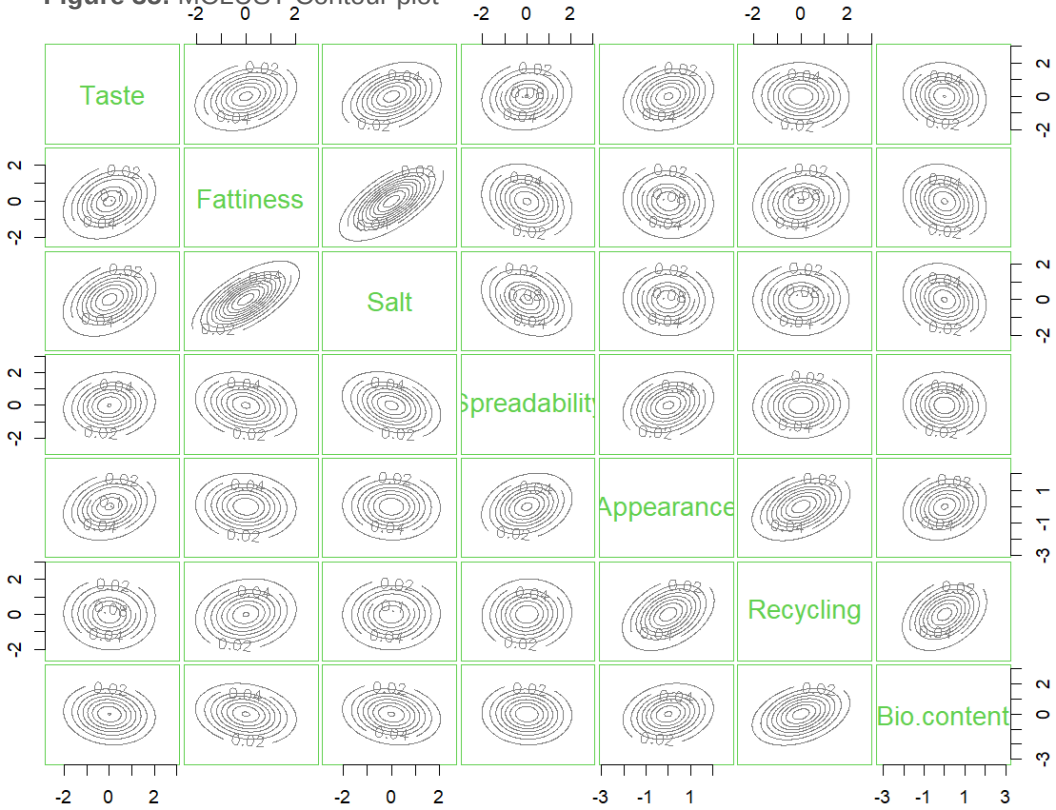
Appendix **2.31** - Demographic summary of combined approach

Table 65: Age ANOVA and TukeyHSD

	khsegment	gender_proportion_female	gender_proportion_male	age_mean	education_mode	area_mode
	<int>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	1	0.151	0.832	40.2	2	2
2	2	0.524	0.445	42.9	4	3

Appendix **2.26** - MCLUST Contour Plot

Figure 35: MCLUST Contour plot



Appendix **2.27** - MCLUST Summary

Table 66: Summary of MCLUST

```
----------------------------------------------------
Density estimation via Gaussian finite mixture modeling
----------------------------------------------------

Mclust XXX (ellipsoidal multivariate normal) model with 1 component:

 log-likelihood   n df       BIC       ICL
      -4543.355 492 35 -9303.657 -9303.657
```

Explanation: The identification of a single cluster suggests that the data points are relatively homogeneous, meaning they share similar characteristics across the dimensions being analyzed. This indicates that there isn't enough variance or difference in the data to warrant the separation into multiple clusters under gaussian assumptions