Assignment 3

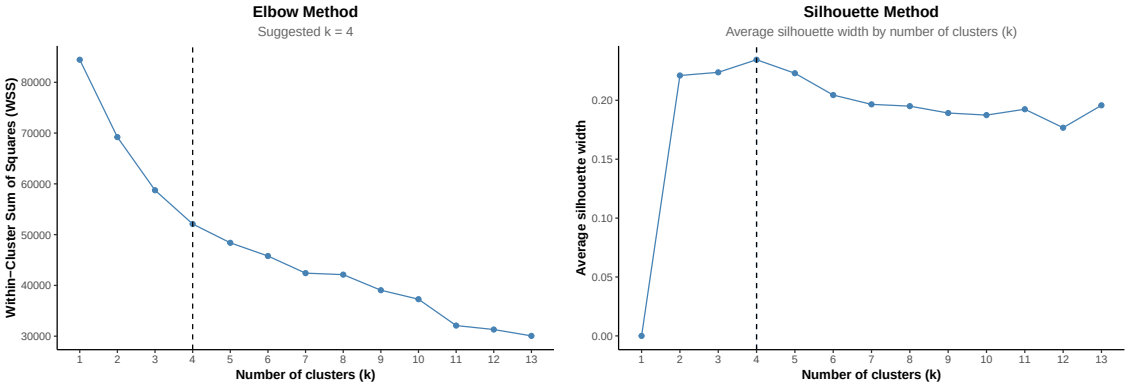
Group Assignment 3

Rishi Ashok Kumar (560527) Nicolas Gonzalez (780037)
Aleksandra Tatko (648925) André van der Meij (589994)

October 06, 2025

Question 1.

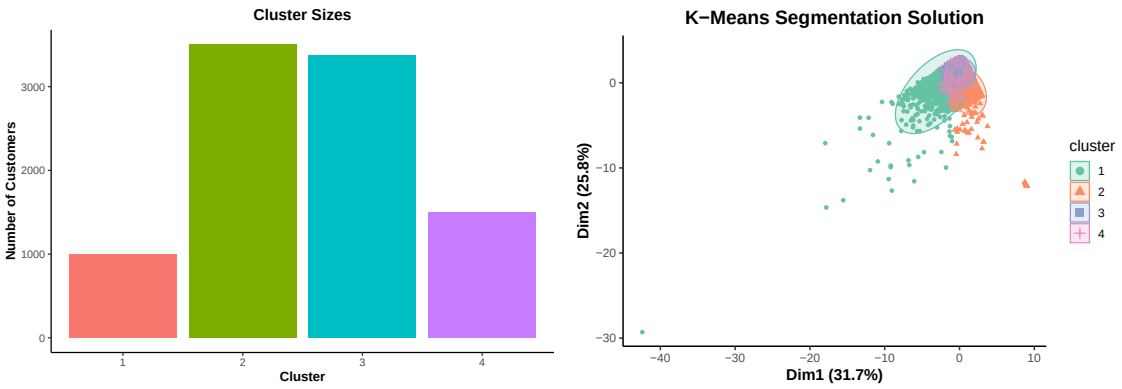
What is the optimal number of clusters? Use the following variables: Age, Average Lead Time, Days Since Creation, Lodging Revenue, Other Revenue, Persons Nights, Room Nights, Days Since Last Stay, Days Since First Stay. Provide relevant figures to support your answer.



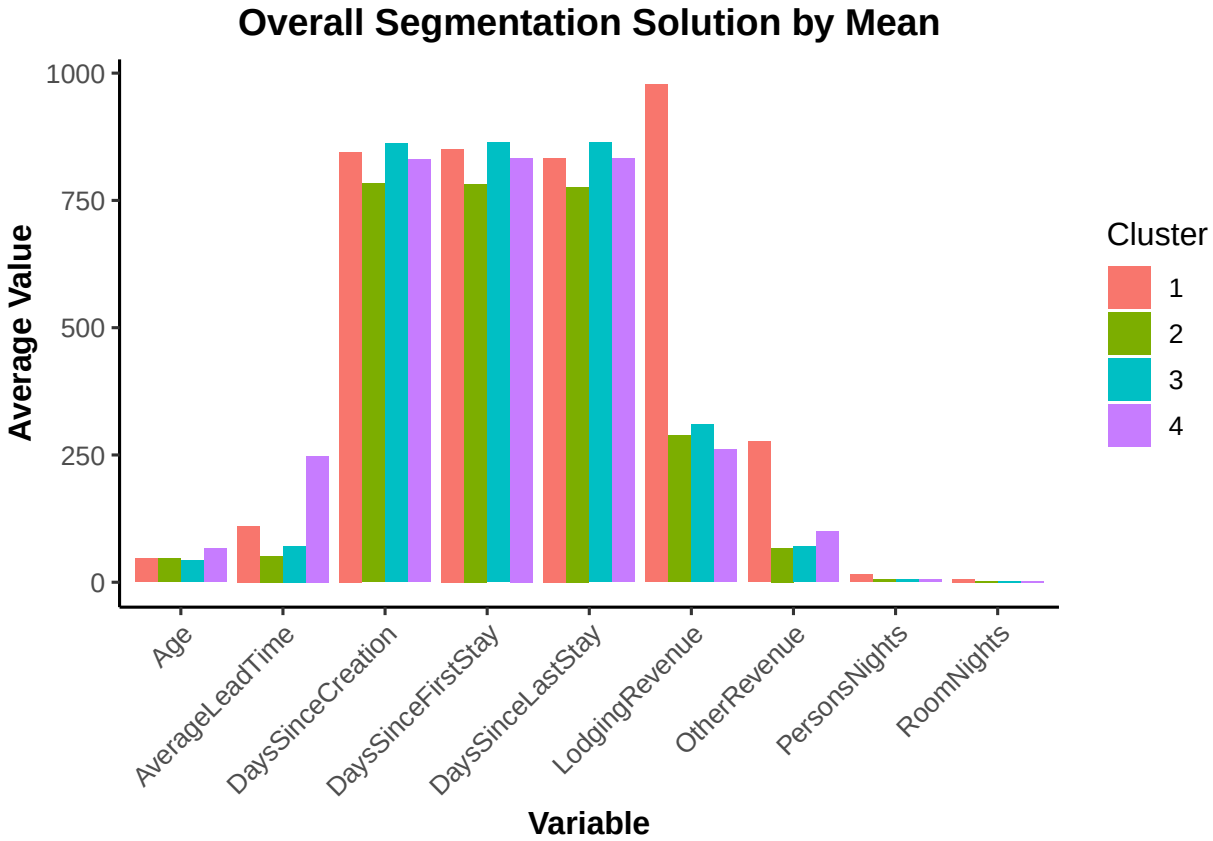
To find the optimal number of clusters we made use of the Elbow and Silhouette method. One one hand, the elbow method shows evaluates how the within-cluster variance decreases as the number of cluster decreases. In the other hands, the silhouette method measures how well each observation fits within its assigned cluster. Both figures are presented above. As we can see, the suggested number of clusters is 4.

Question 2.

Run a k-means cluster analysis using the number of clusters identified in Question 1. Create bar plots to show the cluster sizes and the overall segmentation solution



The figure above summarizes the results of the four-cluster K-Means segmentation. The top-left panel displays the cluster sizes, showing that Cluster 2 and Cluster 3 are the largest segments, while Cluster 1 and Cluster 4 represent smaller portions of the customer base. This indicates that the hotel's clientele is unevenly distributed across segments, with two groups accounting for the majority of customers. The top-right panel visualizes the segmentation solution in two dimensions using principal component analysis (PCA). Each point corresponds to an individual customer, and the color represents their assigned cluster. The visualization shows that the four clusters are partially separated, suggesting that the segmentation captures meaningful differences between groups, although some overlap exists between Clusters 2 and 3. The first two principal components explain about 57% of the total variance, providing a reasonable representation of the overall data structure.

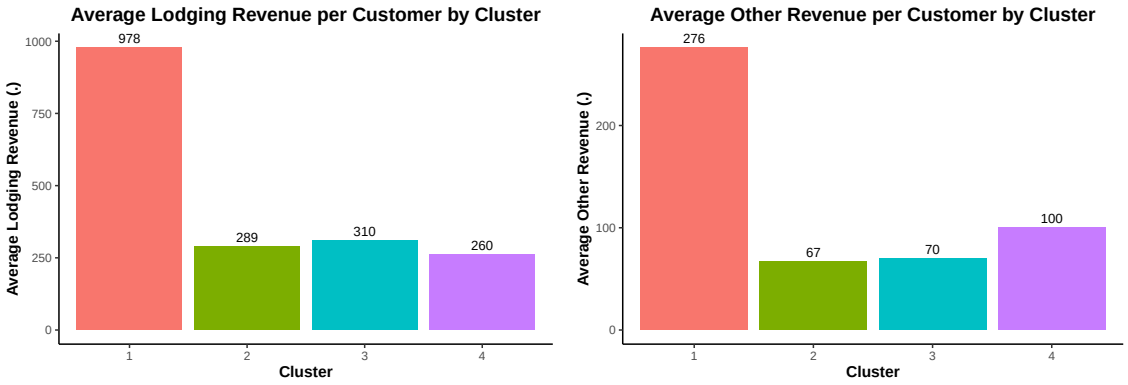


This plot illustrates the average values of key variables by cluster. Cluster 1 stands out with the highest Lodging Revenue and Other Revenue, suggesting that it represents the most valuable and profitable customers. Cluster 2 shows moderate revenues but higher values for Days Since First Stay and Days Since Last

Stay, which may correspond to loyal or repeat customers. Cluster 3 has a similar profile but slightly lower spending levels, possibly representing mid-value regular guests. Finally, Cluster 4 shows the lowest values across most variables, indicating low-spending or infrequent guests.

Question 3.

The hotel management aims to increase revenue and needs help identifying which customer segments to target. Based on your cluster analysis: - Which cluster generates the highest lodging revenue? - Which cluster contributes the most to other revenues (e.g., food, beverages, spa)?



Based on the results of the cluster analysis, Cluster 1 clearly generates the highest revenues in both categories. The first chart shows that customers in Cluster 1 have an average lodging revenue of €978, which is significantly higher than the other clusters (all below €320). This indicates that Cluster 1 represents high-value guests who spend substantially more on accommodation. Similarly, the second chart reveals that Cluster 1 also contributes the most to other revenues such as food, beverages, and spa services, with an average of €276 per customer. In contrast, Clusters 2, 3, and 4 generate noticeably lower ancillary revenues (ranging from €67 to €100 on average). Overall, Cluster 1 stands out as the hotel's most profitable customer segment, both in lodging and in additional services. Therefore, this group should be prioritized for retention and targeted marketing, as it offers the greatest potential to drive total revenue growth.

Question 4.

Management also requests guidance on how to position the hotel's services for the selected target segment(s). Propose two specific, actionable recommendations that are clearly grounded in your cluster analysis.

1. Develop a Premium Loyalty and Experience Program for Cluster 1 (High-Value Guests) Since cluster 1 customers generate the highest lodging and revenues, we could state that they have a stronger willingness to spend on comfort and exclusive experiences. For this reason, the hotel should position itself as a premium destination by offering personalized loyalty rewards and value-added experiences, such as suite upgrades, priority access to spa and dining or personalized travel packages. While targeting this cluster, the hotel should emphasize exclusivity and high service quality.

2. Introduce Upselling and Bundling Strategies for Clusters 2–3 (Moderate-Spending Regular Guests) Since clusters 2 and 3 represent loyal or frequent guests with less expenditure than cluster 1, the hotel could offer service bundles such as including access to spa, activities or dining deals to encourage their expenditure.

Appendix — Full R Code

```r
rm(list = ls())
# Load required packages
library(tidyverse)
library(factoextra)
library(cluster)
library(ggplot2)

# Load dataset
setwd("/Users/nico/Documents/EUR/Marketing/Assignment/Strategic-Marketing/Assignments")
hotel_customers <- read.csv("/Users/nico/Documents/EUR/Marketing/Assignment/Strategic-Marketing/Assignments/HotelCustomersSubset

# Quick look at the data
str(hotel_customers)
head(hotel_customers)
summary(hotel_customers)

# Set seed
set.seed(123)

# Convert Age to integer
hotel_customers$Age <- as.numeric(hotel_customers$Age)

# Remove missing and unrealistic age values
hotel_customers <- hotel_customers %>%
  filter(!is.na(Age) & Age >= 18 & Age <= 95)

# Select the variables
selected_variables <- c(
  "Age","AverageLeadTime","DaysSinceCreation",
  "LodgingRevenue","OtherRevenue",
  "PersonsNights","RoomNights",
  "DaysSinceLastStay","DaysSinceFirstStay"
)

# Keep columns of selected variables and check for missing values
analysis_data <- hotel_customers %>%
  select(all_of(selected_variables)) %>%
  drop_na()

# Standardize all variables
analysis_data_scaled <- scale(analysis_data)

# Check with elbow method
k_opt <- 4

p1 <-  fviz_nbclust(analysis_data_scaled, kmeans, method = "wss", k.max = 13) +
  labs(
    title = "Elbow Method",
    subtitle = paste("Suggested k =", k_opt),
    x = "Number of clusters (k)",
    y = "Within-Cluster Sum of Squares (WSS)"
  ) +
  scale_color_manual(values = c("#2E86AB", "#7FB069"))+
  geom_vline(xintercept = k_opt, linetype = "dashed") +
  theme_classic() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )


# Check with silhouette method
p2 <- fviz_nbclust(analysis_data_scaled, kmeans, method = "silhouette", k.max = 13)+
```

```
  labs(
    title = "Silhouette Method",
    subtitle = "Average silhouette width by number of clusters (k)",
    x = "Number of clusters (k)",
    y = "Average silhouette width"
  ) +
  scale_color_manual(values = c("#2E86AB", "#7FB069"))+
  geom_vline(xintercept = k_opt, linetype = "dashed") +
  theme_classic()+
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )

print(p1)
print(p2)

# Perform k
kmeans_result <- kmeans(analysis_data_scaled, centers = 4, nstart = 25)

# Add cluster to original data set
hotel_customers$Cluster <- as.factor(kmeans_result$cluster)

# Create data frame
cluster_sizes <- data.frame(
  Cluster = factor(1:4),
  Size = kmeans_result$size
)

# Plot cluster sizes
p3 <- ggplot(cluster_sizes, aes(x = Cluster, y = Size, fill = Cluster)) +
  geom_col(show.legend = FALSE) +
  labs(
    title = "Cluster Sizes",
    x = "Cluster",
    y = "Number of Customers"
  ) +
  theme_classic()+
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )

# Visualize the 4-cluster segmentation solution
p4 <-  fviz_cluster(kmeans_result, data = analysis_data_scaled,
             ellipse.type = "norm",
             geom = "point",
             palette = "Set2",
             main = "K-Means Segmentation Solution") +
  theme_classic(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )


print(p3)
print(p4)

# Compute the mean of each variable per cluster
cluster_profile <- hotel_customers %>%
  group_by(Cluster) %>%
  summarise(
    Age = mean(Age, na.rm = TRUE),
```

```r
    AverageLeadTime = mean(AverageLeadTime, na.rm = TRUE),
    DaysSinceCreation = mean(DaysSinceCreation, na.rm = TRUE),
    LodgingRevenue = mean(LodgingRevenue, na.rm = TRUE),
    OtherRevenue = mean(OtherRevenue, na.rm = TRUE),
    PersonsNights = mean(PersonsNights, na.rm = TRUE),
    RoomNights = mean(RoomNights, na.rm = TRUE),
    DaysSinceLastStay = mean(DaysSinceLastStay, na.rm = TRUE),
    DaysSinceFirstStay = mean(DaysSinceFirstStay, na.rm = TRUE)
  ) %>%
  pivot_longer(-Cluster, names_to = "Variable", values_to = "Average")

# Bar plot of average variable values per cluster
p5 <- ggplot(cluster_profile, aes(x = Variable, y = Average, fill = Cluster)) +
  geom_col(position = "dodge") +
  labs(
    title = "Overall Segmentation Solution by Mean",
    x = "Variable",
    y = "Average Value"
  ) +
  theme_classic(base_size = 12) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )

print(p5)

# Compute average and total revenues per cluster
p6 <- hotel_customers %>%
  group_by(Cluster) %>%
  summarise(avg_lodgingrevenue = mean(LodgingRevenue, na.rm = TRUE)) %>%
  ggplot(aes(x = Cluster, y = avg_lodgingrevenue, fill = Cluster)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = round(avg_lodgingrevenue, 0)), vjust = -0.5) +
  labs(
    title = "Average Lodging Revenue per Customer by Cluster",
    x = "Cluster",
    y = "Average Lodging Revenue (€)"
  ) +
  theme_classic(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )

p7 <- hotel_customers %>%
  group_by(Cluster) %>%
  summarise(avg_otherrevenue = mean(OtherRevenue, na.rm = TRUE)) %>%
  ggplot(aes(x = Cluster, y = avg_otherrevenue, fill = Cluster)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = round(avg_otherrevenue, 0)), vjust = -0.5) +
  labs(
    title = "Average Other Revenue per Customer by Cluster",
    x = "Cluster",
    y = "Average Other Revenue (€)"
  ) +
  theme_classic(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "gray40", hjust = 0.5),
    axis.title = element_text(face = "bold")
  )

print(p6)
```

```r
print(p7)

code <- readLines("Appendix_code.R")
knitr::asis_output(paste0(
  "\\begingroup\\footnotesize\n```r\n",
  paste(code, collapse = "\n"),
  "\n```\n\\endgroup\n"
))
```