Assignment 4: Clustering

Nitya Kari

2023-11-10

Load Dataset

```
# Load packages
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(cluster)
library(NbClust)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.2.1     v dplyr   1.1.1
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(aricode)
```

```
sales <- read.csv("/Users/nitwit/Desktop/Fashion_Retail_Sales.csv")
```

This dataset showcases the sales transactions of a clothing store. It helps with analyzing the shopping behavior of customers through five main variables (along with the customer's ID). These variables include the type of clothing item purchased, the purchase amount in USD, he date of the purchase, the customer's satisfaction levels, and the payment method of the customer. This dataset aims to provide insight into the day-to-day transactions an dinteractions within a fashion clothing store. There are 3400 seperate values included in this dataset.

This dataset was obtained from Kaggle.com (https://www.kaggle.com/datasets/fekihmea/fashion-retail-sales/)

Research Question

From this dataset, I hope to utilize clustering in order to find a pattern of shopping behavior. As such my research question is: "Does the customer satisfaction rating based on the purchase amount cluster into meaningful groups?

Essentially, I would like to explore if there is a pattern of behavior among shopper where higher purchasing amounts would mean that shoppers also have a higher satisfaction rating. The clustering these ratings into groups, I hope to understand the nature of this relationship.

Variables of Interest

Purchasing Amount (USD) -> 'Purchase.Amount..USD.' (numeric, continuous): The amount of money that each customer spent for the purchased item Satisfaction Rating -> 'Review.Rating' (numeric, continuous): The customer's satisfaction level for the purchased item

Data Wrangling

```
# Selecting the variables of interest
salesData <- sales %>%
  select('Purchase.Amount..USD.', 'Review.Rating')

# Renaming the variables of interest
salesData <- salesData %>%
  rename(
    purchaseAmount = Purchase.Amount..USD.,
    reviewRating = Review.Rating)

# Removing the NA values
salesData <- na.omit(salesData)

#Subsetting the Satisfaction Rating variable into different groups
salesData <- salesData %>%
  mutate(satisfaction = case_when(
    reviewRating == 1 | reviewRating < 2 ~ 1,
    reviewRating == 2 | reviewRating < 3 ~ 2,
    reviewRating == 3 | reviewRating < 4 ~ 3,
    reviewRating == 4 | reviewRating < 5 ~ 4,
    reviewRating == 5 ~ 5
  ))

#Frequencies based on the Satisfaction Rating grouping
category_freq <- table(salesData$satisfaction)
print(category_freq)
```

```
##
##   1   2   3   4   5
## 611 602 615 633  26
```

Analysis

```
sales_actual <- salesData$satisfaction
sales_numeric <- salesData %>%
  select(-satisfaction)

# Set seed
set.seed(123)

# Perform k-means with 5 clusters
theoretical_run <- kmeans(
  x = sales_numeric, # numeric data
  centers = 5, # number of clusters
  iter.max = 10, # number of maximum iterations
  nstart = 25 # number of random starting values
)

# Within-cluster sum of squares
theoretical_run$withinss
```

```
## [1]   132880.82 29187419.65   108960.73    99250.79   117691.08
```

```
# Variance
theoretical_run$betweenss / # between sum of squares
theoretical_run$totss # total sum of squares
```

```
## [1] 0.9212997
```

Seeing as the variance is equal to around 0.92 which is very close to 1, it is likely that the clustering of the customer satisfaction is related to the varaiability in the data.

```
# Check out cluster frequencies
table(theoretical_run$cluster)
```

```
##
##   1   2   3   4   5
## 662  34 629 581 581
```

```
# Actual frequencies
table(sales_actual)
```

```
## sales_actual
##   1   2   3   4   5
## 611 602 615 633  26
```

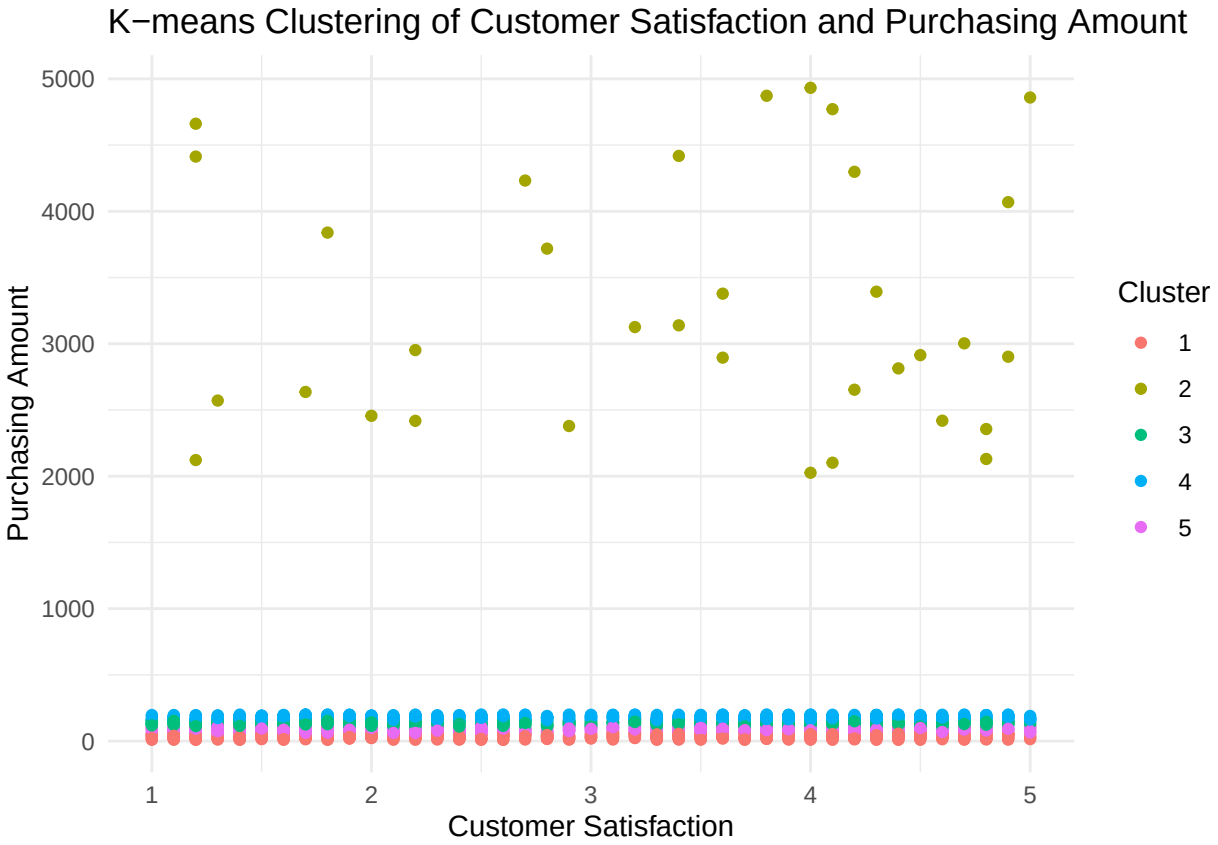
Comparing these frequencies is very interesting because the distributions showcased in the actual and theoretical models are not uniform at all. The clusters do not align with the actual categories.

```r
# Centroids
round(theoretical_run$centers, 1)
```

```
##   purchaseAmount reviewRating
## 1           35.2          3.0
## 2         3290.2          3.4
## 3          133.2          3.0
## 4          178.2          3.1
## 5           85.4          2.9
```

Visualization

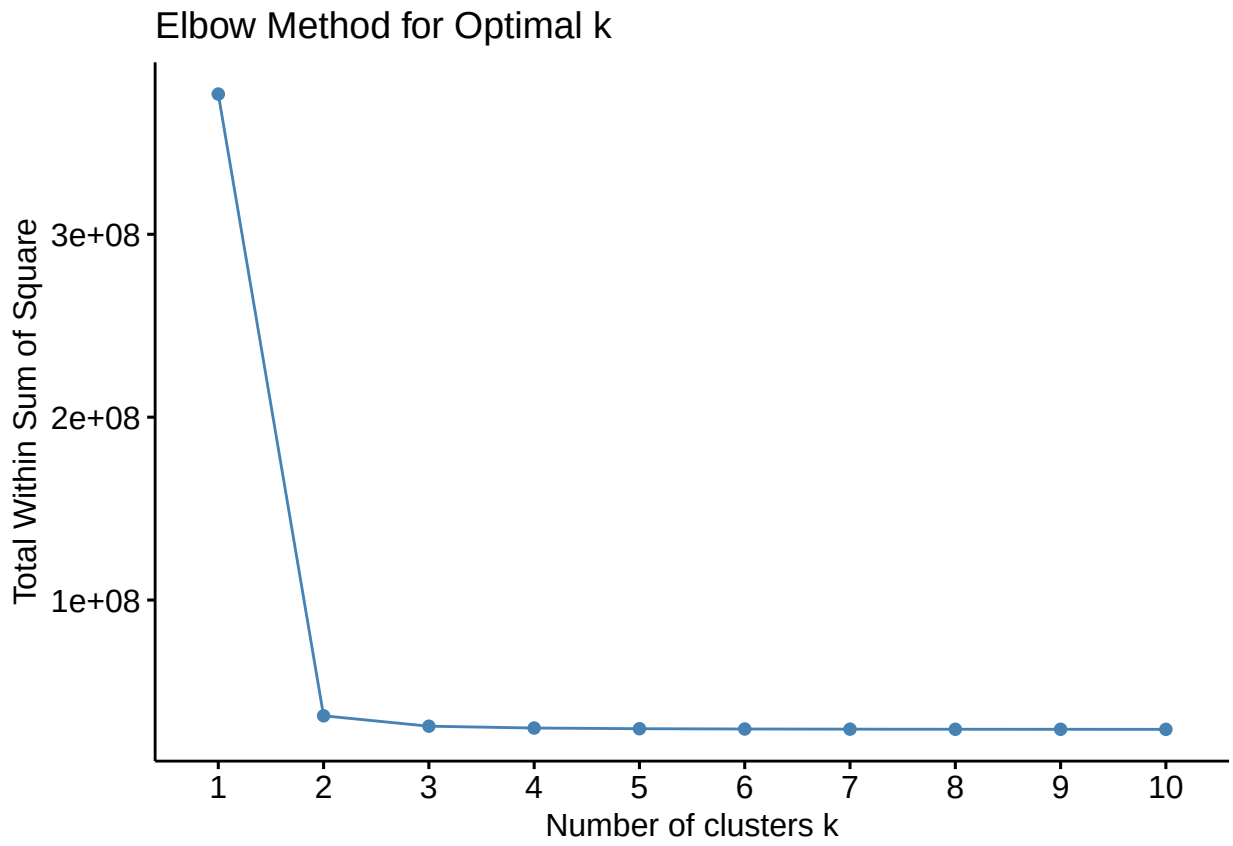
```r
# Visualize the K-means clusters
numeric_with_cluster <- cbind(sales_numeric, Cluster = theoretical_run$cluster)
ggplot(numeric_with_cluster, aes(x = reviewRating, y = purchaseAmount,
                                 color = as.factor(Cluster))) +
  geom_point() +
  labs(x = "Customer Satisfaction", y = "Purchasing Amount", color = "Cluster") +
  ggtitle("K-means Clustering of Customer Satisfaction and Purchasing Amount") +
  theme_minimal()
```



> After visualizing the clusters, it is interesting that many of the clusters, except for cluster 2, have a low purchasing amount.Additionally, the plot points do not showcase a uniform pattern among the different

clusters. This may showcase that there actually is not a relationship between purchasing amount and the customer satisfaction rating.

```
#Elbow Method
wss <- numeric(10)
for (i in 1:10) {
  wss[i] <- kmeans(sales_numeric, centers = i, nstart = 25)$tot.withinss
}

fviz_nbclust(sales_numeric, kmeans, method = "wss", k.max = 10) +
  ggtitle("Elbow Method for Optimal k")
```



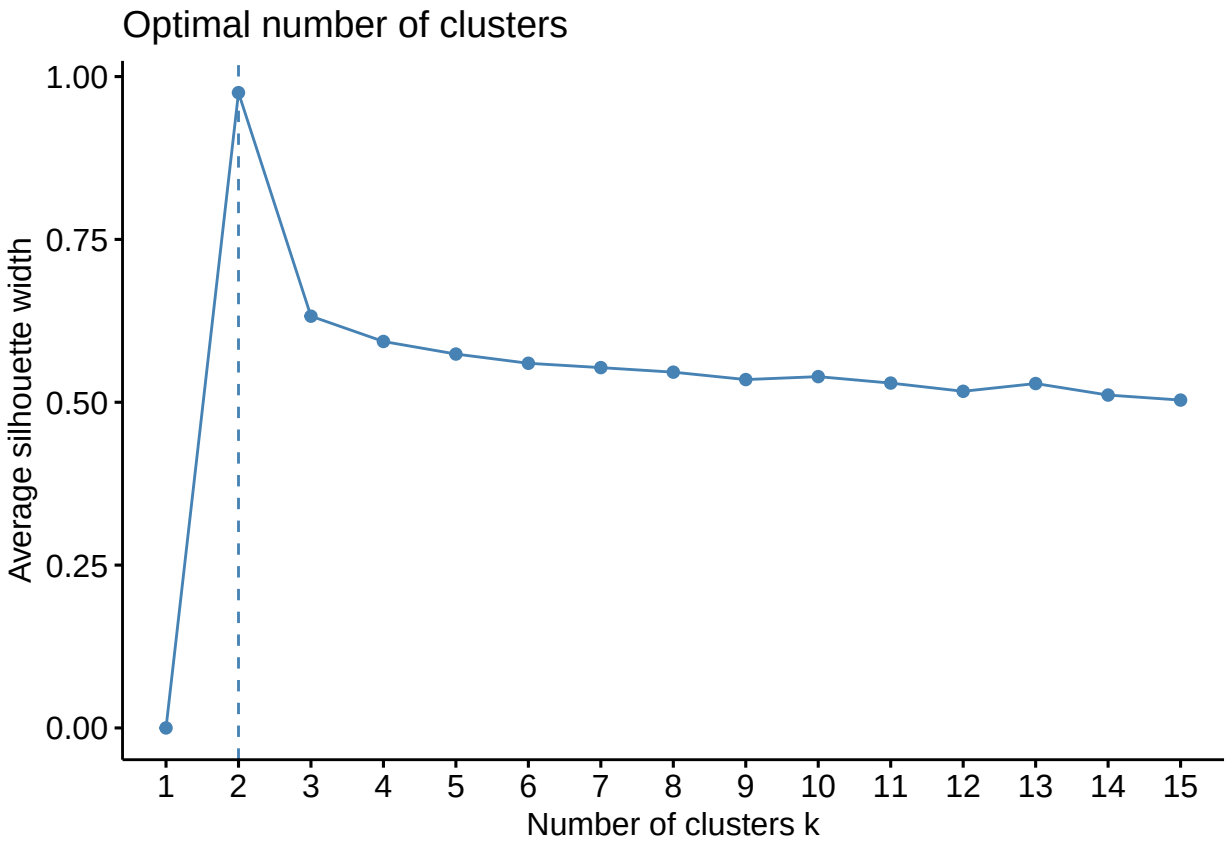
> The elbow method showcases that adding clusters beyond 2 diminishes returns. As such, the optimal amount of clusters would be 2.

```
# Silhouette Method
silhouettes <- sapply(
  2:15, # 2 to number of clusters
  # Needs minimum 2
  FUN = function(centers){

    # Obtain k-means output
    output <- kmeans(
      x = sales_numeric,
      centers = centers,
      iter.max = 10,
      nstart = 25
```

```
    )

    # Compute silhouettes
    silhos <- silhouette(
      output$cluster,
      dist(sales_numeric)
      # computes Euclidean's distance by default
    )

    # Obtain average width
    mean(silhos[,3], na.rm = TRUE)

  }
)

# Print values
silhouettes
```

```
##  [1] 0.9752990 0.6321264 0.5932932 0.5739930 0.5593483 0.5544417 0.5474024
##  [8] 0.5388618 0.5391813 0.5287893 0.5316330 0.5122053 0.5114652 0.5128873
```

```
# Plot silhouette method
fviz_nbclust(
  x = sales_numeric,
  FUNcluster = kmeans, # cluster function
  method = "silhouette", # silhouette
  k.max = 15,  # maximum number of clusters
  iter.max = 10, # same as our k-means setup
  nstart = 25 # same as our k-means setup
)
```



> The silhoutte method also showcases that the optimal amount of clusters would be 2.

Back to Analysis

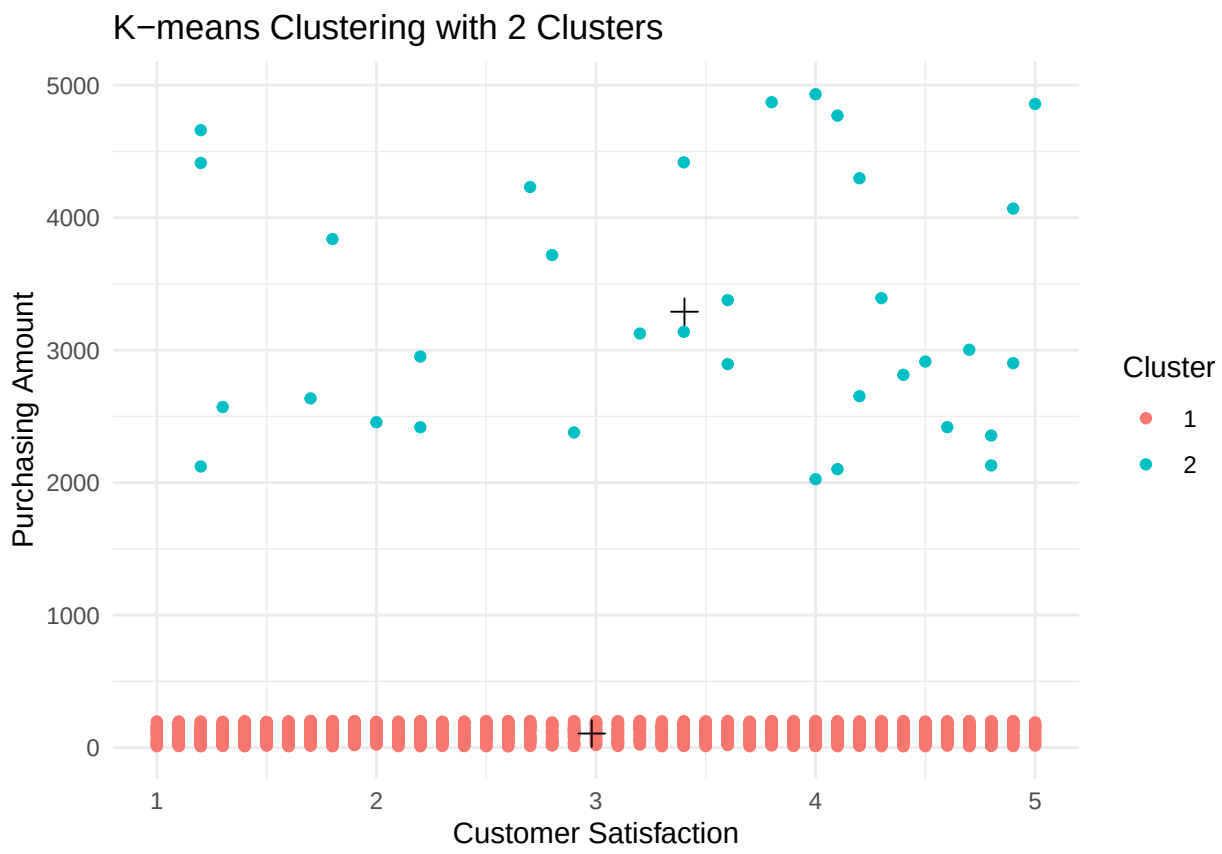
```
#K Means with 2 clusters
set.seed(123)
kmeans_output_two <- kmeans(sales_numeric, centers = 2)
numeric_salaries_with_cluster_two <- cbind(sales_numeric,
                                           Cluster = kmeans_output_two$cluster)

#Centroids
centroidsII <- kmeans_output_two$centers
centroidsII
```

```
##   purchaseAmount reviewRating
## 1       106.0799     2.980473
## 2      3290.1765     3.402941
```

Back to Visualization

```
# Visualize the K-means clusters
ggplot(numeric_salaries_with_cluster_two, aes(x = reviewRating,
                                              y = purchaseAmount,
```

```
                                        color = as.factor(Cluster))) +
geom_point() +
labs(x = "Customer Satisfaction", y = "Purchasing Amount", color = "Cluster") +
ggtitle("K-means Clustering with 2 Clusters") +
theme_minimal() +
# Add centroids to the plot
geom_point(data = data.frame(centroidsII), aes(x = reviewRating, y = purchaseAmount),
           color = "black", size = 3, shape = 3)
```



> This is similar to the original visualization with 5 clusters where there is no apparent pattern.

Discussion

It is quite interesting to look at the visualizations of the clustering. Despite the high varaince of 0.92 which implies that clustering has an impact on the variance in the dataset, the visualizations of both K-means clusters (5 clusters and the optimal 2 clusters) do not showcase s clear discernable pattern in customer satisfaction based on the purchase amount. It could be argued that there are more plot point in both visualizations closer to the higher customer satisfaction rating, but there is a good amount of distribution amount the purchasing amount with just slightly more points around the $2000 to $3000 purchasing amount range.

However, it is important to note that it is is difficult to discern the disrtibution amount clusters with lower purchasing amounts in these plots. As such, we can look at the centroid distribution in order to understand the relationship between the purchase amount and the customer satisfaction rating. For the K-means with 5 clusters, each cluster had a customer satisfaction rating very close

to the median of 3. However, cluster 2, the cluster with the highest average purchase amount of upwards of $3000, had a customer satisfaction rating of 3.4. This is also apparent in the K-means with the optimal amount of clusters (2). The cluster with a higher average purchase amount of $3000+ had a customer satisfaction rating of 3.4 while the other cluster had a customer satisfaction rating of 2.9.

When looking at all of this information together, it is reasonable to conclude that purchase amount does have a slight impact on customer satisfaction but it is not the deciding factor nor is it the main factor. The relationship between the two is not enough to cluster into very meaningful groups. The purchase amount does not have enough of an impact on the customer satisfaction.