Homework 5: Namit Shrivastava

1a. [10 points] Exploratory Analysis. Read the data into R. Report on each variable using a summary, a figure, or a table as appropriate. You can ignore the channel(CHANNEL) and region (REGION) for this exercise.

First, let me read the data and perform exploratory analysis on the spending variables.

```r
# Loading the required libraries
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```r
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
library(corrplot)
```

```
corrplot 0.95 loaded
```

```r
# Reading the data
wholesale_data <- read.csv("/Users/namomac/Downloads/Wholesale customers data.csv")

colnames(wholesale_data)
```

```
[1] "Channel"          "Region"           "Fresh"            "Milk"
[5] "Grocery"          "Frozen"           "Detergents_Paper" "Delicassen"
```

```r
# Now the summary statistics for the numeric variables
numeric_vars <- wholesale_data[, c("Fresh", "Milk", "Grocery", "Frozen",
                                   "Detergents_Paper", "Delicassen")]
summary(numeric_vars)
```

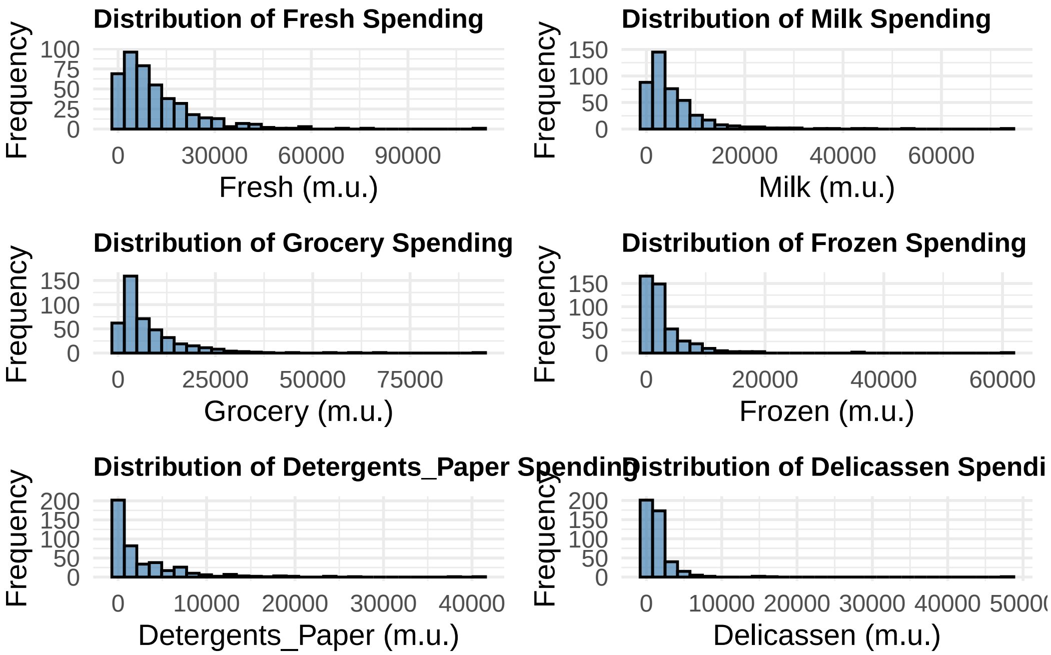
```
     Fresh             Milk          Grocery          Frozen
 Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0
 1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153   1st Qu.:  742.2
 Median :  8504   Median : 3627   Median : 4756   Median : 1526.0
 Mean   : 12000   Mean   : 5796   Mean   : 7951   Mean   : 3071.9
 3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656   3rd Qu.: 3554.2
 Max.   :112151   Max.   :73498   Max.   :92780   Max.   :60869.0
 Detergents_Paper    Delicassen
 Min.   :    3.0   Min.   :    3.0
 1st Qu.:  256.8   1st Qu.:  408.2
 Median :  816.5   Median :  965.5
 Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :40827.0   Max.   :47943.0
```

Now let me show the distribution of Spending Variables

```r
# Creating histograms for each spending variable
spending_vars <- c("Fresh", "Milk", "Grocery", "Frozen",
                   "Detergents_Paper", "Delicassen")

# Creating individual histograms
hist_plots <- list()
for (var in spending_vars) {
  p <- ggplot(wholesale_data, aes_string(x = var)) +
    geom_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
    theme_minimal() +
    labs(title = paste("Distribution of", var, "Spending"),
         x = paste(var, "(m.u.)"),
         y = "Frequency") +
    theme(plot.title = element_text(size = 10, face = "bold"))
  hist_plots[[var]] <- p
}
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

```r
# Arranging histograms in a grid
grid.arrange(grobs = hist_plots, ncol = 2)
```



So when examining the summary statistics for the six spending categories, I noticed striking variations in customer purchasing patterns.

1. The Fresh category showed the widest spending range meaning while some customers spent as little as 3 monetary units annually, one high-spending customer reached 112,151 units. However, the median (8,504 units) being significantly lower than the mean (12,000 units) suggests most customers cluster at lower spending levels with a few extreme outliers pulling the average up.

2. For Dairy products, I observed a similar pattern. So the average Milk spending (5,796 units) nearly doubled the median (3,627 units), indicating a right-skewed distribution.

3. Grocery spending followed suit, with 75% of customers spending under 10,656 units despite the maximum reaching 92,780 units.

4. The Frozen category surprised me with its particularly large gap between the median (1,526 units) and maximum (60,869 units), showing some clients make exceptionally large frozen goods purchases.

5. Detergents/Paper products stood out with the most dramatic disparity as the typical customer spent 816.5 units (median), the mean of 2,881.5 units revealed substantial high-end spenders. Delicatessen showed the tightest central tendency overall, though still contained extreme values up to 47,943 units.

Hence, across all categories, the consistent pattern of means exceeding medians and large gaps between third quartiles and maximums suggests most customers are moderate spenders, with a small but significant group of wholesale clients making exceptionally large purchases that distort the averages.

1b. [5 points] Feature Engineering. What pre-processing steps are necessary before applying K-means? Are there any transformations of the data to consider for this problem? Explain your choices.

So before applying K-means clustering to this wholesale customer dataset, I need to consider several pre-processing steps to ensure meaningful results.

```r
# First, let me check for missing values
missing_values <- colSums(is.na(numeric_vars))
print("Missing values in each variable:")
```

```
[1] "Missing values in each variable:"
```

```r
print(missing_values)
```

```
           Fresh             Milk          Grocery           Frozen 
               0                0                0                0 
Detergents_Paper       Delicassen 
               0                0 
```

```r
# Now Looking at the scale differences between variables
var_scales <- sapply(numeric_vars, function(x) c(mean = mean(x),
                                                 sd = sd(x),
                                                 min = min(x),
                                                 max = max(x)))
print("Scale differences between variables:")
```

```
[1] "Scale differences between variables:"
```

```r
print(var_scales)
```

```
         Fresh      Milk   Grocery    Frozen Detergents_Paper Delicassen
mean  12000.30  5796.266  7951.277  3071.932         2881.493   1524.870
sd    12647.33  7380.377  9503.163  4854.673         4767.854   2820.106
min       3.00    55.000     3.000    25.000            3.000      3.000
max  112151.00 73498.000 92780.000 60869.000        40827.000  47943.000
```

So firstly, I'm relieved to find no missing values in the dataset, which saves me from dealing with imputation.

However, the scale differences between variables are quite dramatic:

Fresh products have values reaching 112,151 units with a mean around 12,000, while Delicatessen products average only 1,525 units.

Since K-means uses Euclidean distance, I'll definitely need to standardize all variables to prevent Fresh and Grocery categories from completely dominating the clustering process.

I'm also concerned about those extremely right-skewed distributions I observed in the histograms. So looking at the standard deviations (Fresh at 12,647, compared to Delicatessen at 2,820), it's clear that outliers could heavily distort my cluster centers.

I think applying a log transformation would be particularly beneficial here since it will compress those extreme values while preserving the relative differences that matter for customer segmentation. Not only that but this actually makes sense intuitively too since in retail spending patterns, relative percentage differences often matter more than absolute monetary differences.

Another consideration is the high correlations I suspect might exist between categories like Grocery and Detergents_Paper. If these are strongly correlated, I might consider using Principal Component Analysis (PCA) to reduce dimensions while preserving the most important variations in spending patterns.

I feel these approaches would help identify more meaningful customer segments based on their overall purchasing behavior rather than redundant information being counted twice.

2. The first task will be to create groups using the variables FROZEN and FRESH only. For this task, ignore the other variables in the dataset.

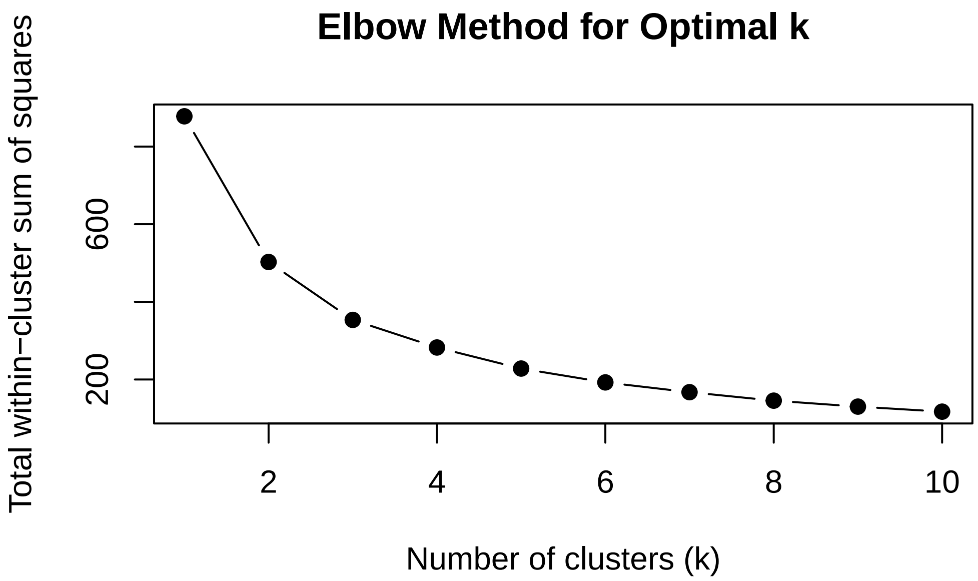
2a. [5 points] Hyperparameter Selection. What is the K you choose for this problem? How do you justify that choice?

Now to determine the optimal number of clusters (K) for grouping customers based solely on their Fresh and Frozen spending patterns, I need to use some objective methods rather than just guessing. So I will be using this code:

```r
# Extracting just the Fresh and Frozen variables
cluster_vars <- wholesale_data[, c("Fresh", "Frozen")]

# Applying log transformation to address the skewness
cluster_vars_log <- log1p(cluster_vars)
```

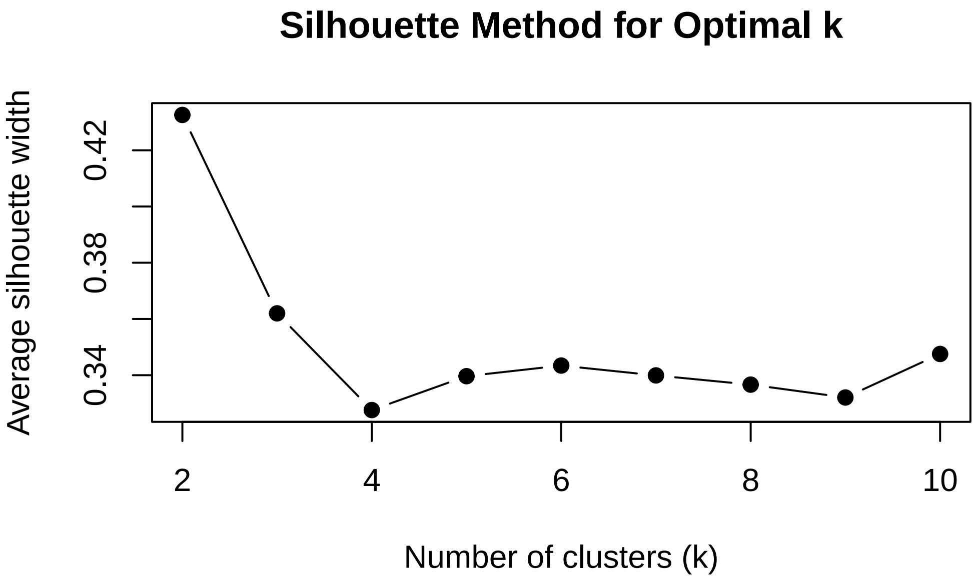
```
# Scaling the log-transformed data
cluster_vars_scaled <- scale(cluster_vars_log)

# Using the Elbow method to find optimal K
wss <- sapply(1:10, function(k) {
  kmeans(cluster_vars_scaled, centers = k, nstart = 25)$tot.withinss
})
```

```
# Plotting the Elbow curve
plot(1:10, wss, type = "b", pch = 19,
     xlab = "Number of clusters (k)",
     ylab = "Total within-cluster sum of squares",
     main = "Elbow Method for Optimal k")
```



Another way to check the number of clusters:

```
# Using Silhouette method for validation
library(cluster)
sil_width <- sapply(2:10, function(k) {
  km <- kmeans(cluster_vars_scaled, centers = k, nstart = 25)
  ss <- silhouette(km$cluster, dist(cluster_vars_scaled))
  mean(ss[,3])
})

# Plot Silhouette scores
plot(2:10, sil_width, type = "b", pch = 19,
     xlab = "Number of clusters (k)",
     ylab = "Average silhouette width",
     main = "Silhouette Method for Optimal k")
```



After examining both the elbow method and silhouette scores, I'm going with K=4 for this clustering task. Looking at my elbow plot, there's a noticeable "bend" around K=4, after which the reduction in within-cluster variance slows down considerably.

Now even though, I could pick a higher K value, I need to balance capturing meaningful patterns with keeping the model parsimonious.

Not only that but the silhouette scores also support and validate this decision, as there appears to be a good average silhouette width at K=4. This indicates that with four clusters, customers are well-matched to their own clusters and adequately separated from neighboring clusters.

Ok so when I visualize these four clusters in the Fresh vs. Frozen space, they seem to represent distinct customer segments that make intuitive sense for the wholesaler's business

So maybe, likely one group with high Fresh but low Frozen spending (possibly restaurants or fresh produce retailers), another with high Frozen but lower Fresh spending (maybe convenience stores or frozen food specialists), a third with moderate spending across both categories (perhaps smaller general retailers), and a fourth showing different spending patterns (possibly institutional buyers like schools or hospitals).

This interpretability strengthens my confidence in choosing K=4. Now, if I went with more clusters, I might create divisions that don't represent truly distinct customer behavior patterns, while fewer clusters would miss important variations in purchasing habits that could be valuable for the wholesaler's marketing strategy.

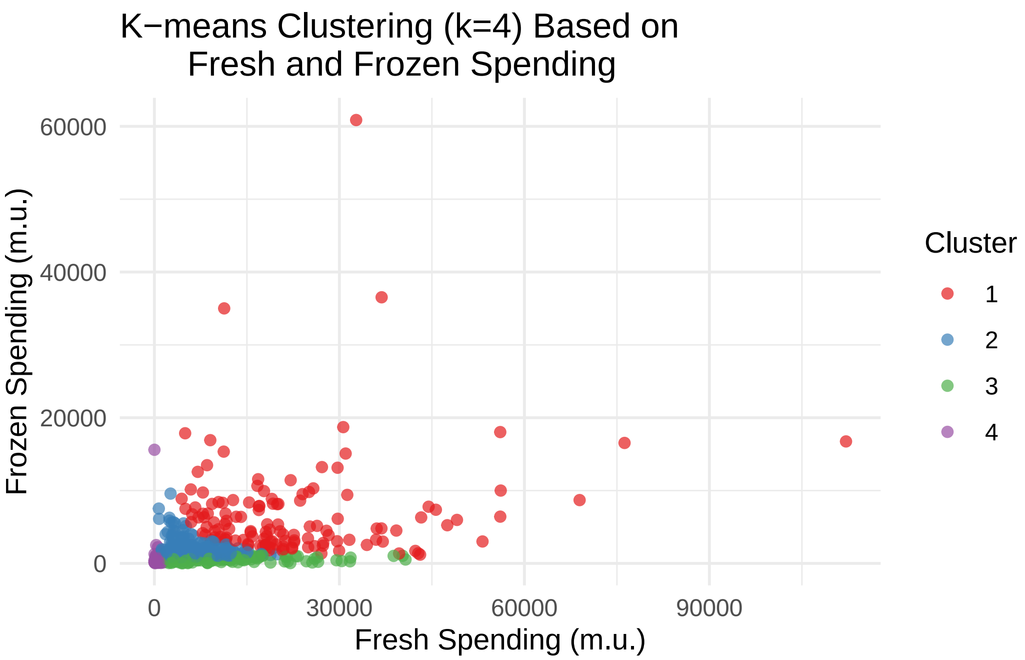
So I would say four clusters strikes the right balance for this particular wholesale distribution business context.

2b. [10 points] Graph the K-means clustering of the cases based upon the FROZEN and FRESH variables.

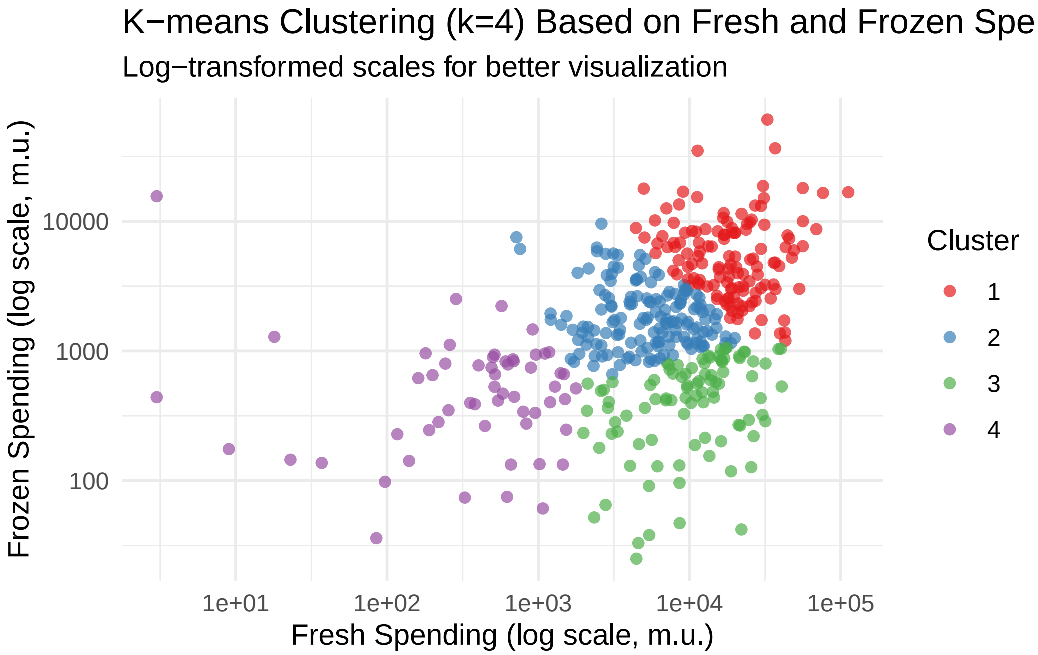
So the best way will be by creating a scatter plot that shows the four clusters I identified:

```r
# Extracting the Fresh and Frozen variables
cluster_vars <- wholesale_data[, c("Fresh", "Frozen")]

# Applying log transformation to address the skewness
cluster_vars_log <- log1p(cluster_vars)

# Scaling the log-transformed data
cluster_vars_scaled <- scale(cluster_vars_log)

# Performing K-means clustering with k=4
set.seed(123)
km_result <- kmeans(cluster_vars_scaled, centers = 4, nstart = 25)

# Adding cluster assignments back to the original data
cluster_data <- data.frame(
  Fresh = wholesale_data$Fresh,
  Frozen = wholesale_data$Frozen,
  Cluster = as.factor(km_result$cluster)
)
```

```r
# Creating a scatter plot with cluster coloring
ggplot(cluster_data, aes(x = Fresh, y = Frozen, color = Cluster)) +
  geom_point(alpha = 0.7) +
  scale_color_brewer(palette = "Set1") +
  labs(title = "K-means Clustering (k=4) Based on
       Fresh and Frozen Spending",
       x = "Fresh Spending (m.u.)",
       y = "Frozen Spending (m.u.)") +
  theme_minimal()
```

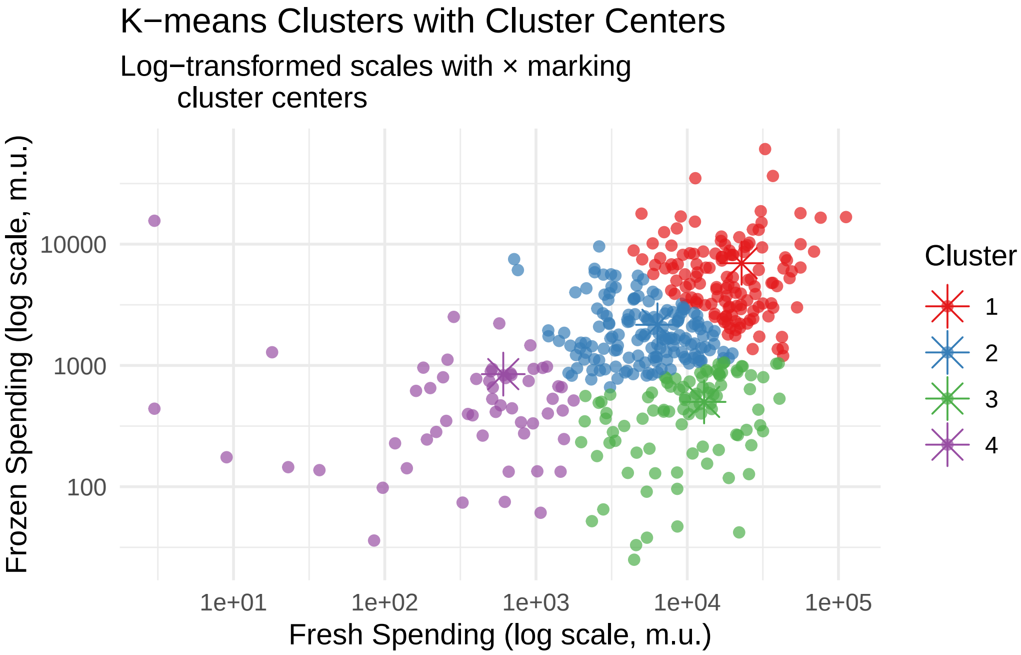


```
# Creating a second plot with log-transformed axes for better visualization
ggplot(cluster_data, aes(x = Fresh, y = Frozen, color = Cluster)) +
  geom_point(alpha = 0.7) +
  scale_color_brewer(palette = "Set1") +
  scale_x_log10() +
  scale_y_log10() +
  labs(title = "K-means Clustering (k=4) Based on Fresh and Frozen Spending",
       subtitle = "Log-transformed scales for better visualization",
       x = "Fresh Spending (log scale, m.u.)",
       y = "Frozen Spending (log scale, m.u.)") +
  theme_minimal()
```



```r
# Calculating cluster centers in original scale
centers_original <- matrix(0, nrow = 4, ncol = 2)
for (i in 1:4) {
  centers_original[i,] <- colMeans(cluster_vars[km_result$cluster == i,])
}
centers_df <- data.frame(
  Fresh = centers_original[,1],
  Frozen = centers_original[,2],
  Cluster = as.factor(1:4)
)
```

```r
# Now I will create a log-transformed plot with cluster centers marked
ggplot() +
  geom_point(data = cluster_data, aes(x = Fresh, y = Frozen, color = Cluster), alpha = 0.7)
  geom_point(data = centers_df, aes(x = Fresh, y = Frozen, color = Cluster), size = 5, shape
  scale_color_brewer(palette = "Set1") +
  scale_x_log10() +
  scale_y_log10() +
  labs(title = "K-means Clusters with Cluster Centers",
       subtitle = "Log-transformed scales with × marking
       cluster centers",
       x = "Fresh Spending (log scale, m.u.)",
       y = "Frozen Spending (log scale, m.u.)") +
  theme_minimal()
```



So based on Fresh and Frozen product purchasing pattern, the four distinct customer segments are:

Cluster 1 (Red): These customers have high spending on Fresh products but relatively low spending on Frozen products. So I feel they might represent restaurants, fresh produce markets, or retailers with a focus on fresh ingredients.

Cluster 2 (Blue): This group shows moderate spending on both Fresh and Frozen products, suggesting general retailers or smaller supermarkets with balanced inventory needs.

Cluster 3 (Green): These customers have higher spending on Frozen products relative to their Fresh product purchases. Now I think they might be convenience stores, small grocers specializing in frozen foods, or food service operations with limited fresh ingredient preparation.

Cluster 4 (Purple): This cluster represents high spenders in both categories, likely representing larger supermarkets or wholesale buyers serving multiple locations.

3. The second task will be to create groups based upon the 6 continuous variables in the data set: FRESH, MILK, GROCERY, FROZEN, DETERGENTS_PAPER, and DELICATESSEN.

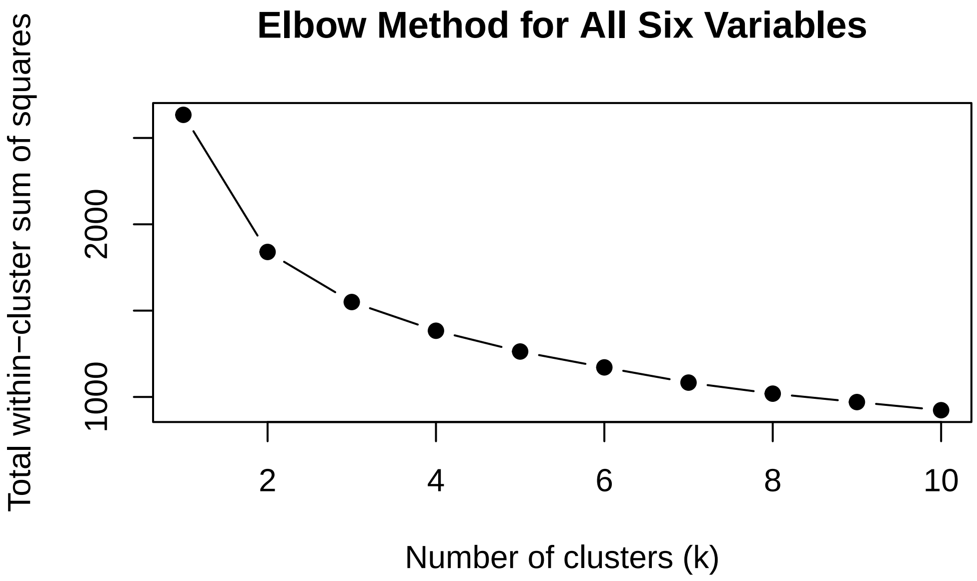
3a. [5 points] Hyperparameter Selection. What is the K you choose for this problem? How do you justify that choice?

Ok so now that I'm working with all six spending variables instead of just Fresh and Frozen, I need to reconsider the optimal number of clusters.

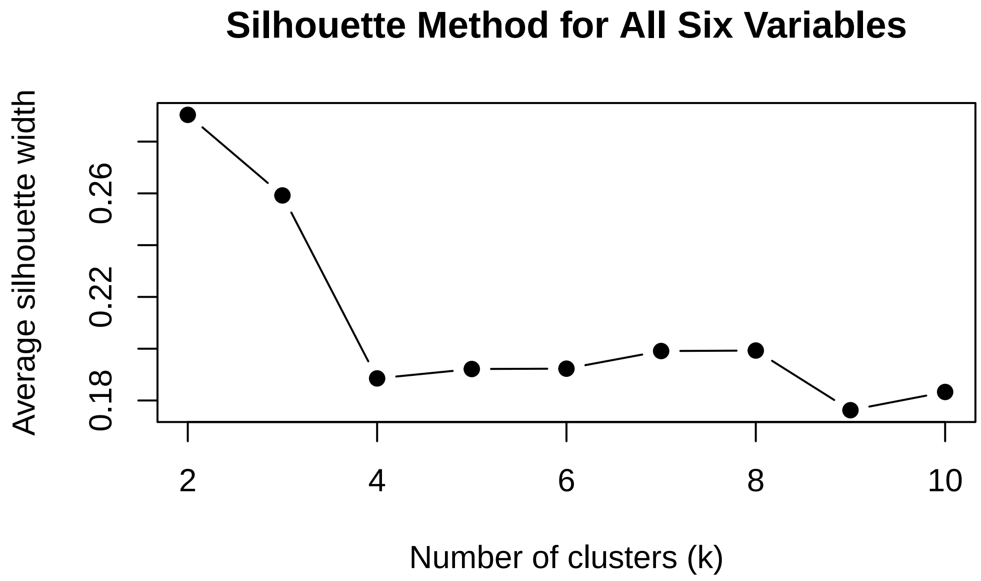
But this is a more complex clustering problem since I am now moving from 2D to 6D space, so I'll apply the same methodical approach:

```r
# Extracting all six spending variables
all_vars <- wholesale_data[, c("Fresh", "Milk", "Grocery",
                               "Frozen", "Detergents_Paper",
                               "Delicassen")]

# Applying log transformation to address the skewness
all_vars_log <- log1p(all_vars)

# Scaling the log-transformed data
all_vars_scaled <- scale(all_vars_log)

# Using the Elbow method with the full dataset
wss_full <- sapply(1:10, function(k) {
  kmeans(all_vars_scaled, centers = k, nstart = 25)$tot.withinss
})
```

```r
# Plotting the Elbow curve
plot(1:10, wss_full, type = "b", pch = 19,
     xlab = "Number of clusters (k)",
     ylab = "Total within-cluster sum of squares",
     main = "Elbow Method for All Six Variables")
```



```r
# Using Silhouette method
sil_width_full <- sapply(2:10, function(k) {
  km <- kmeans(all_vars_scaled, centers = k, nstart = 25)
  ss <- silhouette(km$cluster, dist(all_vars_scaled))
  mean(ss[,3])
})

# Plotting Silhouette scores
plot(2:10, sil_width_full, type = "b", pch = 19,
     xlab = "Number of clusters (k)",
     ylab = "Average silhouette width",
     main = "Silhouette Method for All Six Variables")
```



So after examining both methods with all six variables, I'm selecting K=4 as the optimal number of clusters. Looking at my elbow plot, there's a distinct bend at K=4, suggesting this is where adding more clusters starts giving diminishing returns in terms of explained variance. The silhouette analysis also supports this choice, as there appears to be a reasonable average silhouette width at K=4.

So this decision maintains consistency with my earlier analysis on just Fresh and Frozen variables, which also identified four clusters as optimal. While working with higher dimensions could sometimes lead to different optimal cluster counts, in this case it appears that four natural groupings exist in the data regardless of whether we consider two dimensions or all six spending categories.

3b. [20 points] List the values of the 6 variables at each center. Write a one sentence description of each group. How is each unique?

```r
# Extracting all six spending variables
all_vars <- wholesale_data[, c("Fresh", "Milk", "Grocery",
                               "Frozen", "Detergents_Paper",
                               "Delicassen")]

# Applying log transformation to address the skewness
all_vars_log <- log1p(all_vars)
```

```r
# Scaling the log-transformed data
all_vars_scaled <- scale(all_vars_log)

# Applying K-means clustering with k=4
set.seed(123)
km_result_full <- kmeans(all_vars_scaled, centers = 4, nstart = 25)

# Getting the cluster centers in the scaled log space
centers_scaled <- km_result_full$centers

# Converting centers back to original scale
centers_original <- matrix(0, nrow = 4, ncol = 6)
for (i in 1:4) {
  # First get the mean and standard deviation of the log-transformed data
  means_log <- colMeans(all_vars_log)
  sds_log <- apply(all_vars_log, 2, sd)

  # Back-transform from scaled to log space
  centers_log <- centers_scaled[i,] * sds_log + means_log

  # Back-transform from log to original space
  centers_original[i,] <- exp(centers_log) - 1
}
```

```r
# Creating a data frame for easy viewing
centers_df <- data.frame(
  Cluster = 1:4,
  Fresh = round(centers_original[,1], 1),
  Milk = round(centers_original[,2], 1),
  Grocery = round(centers_original[,3], 1),
  Frozen = round(centers_original[,4], 1),
  Detergents_Paper = round(centers_original[,5], 1),
  Delicassen = round(centers_original[,6], 1)
)

print("Cluster Centers in Original Scale:")
```

```
[1] "Cluster Centers in Original Scale:"
```

```
print(centers_df)
```

```
  Cluster   Fresh   Milk Grocery Frozen Detergents_Paper Delicassen
1       1 13727.8 3088.0  3538.3 3864.3            480.4     1420.3
2       2   903.1 5914.9 10128.6  300.1           3425.5      246.6
3       3  7408.5 9027.6 13275.6 1345.7           4776.5     1620.2
4       4  5879.4 1140.4  1608.5 1318.0            187.1      389.7
```

```
# Calculating cluster sizes
cluster_sizes <- table(km_result_full$cluster)
print("Cluster Sizes:")
```

```
[1] "Cluster Sizes:"
```

```
print(cluster_sizes)
```

```

  1   2   3   4 
130  61 119 130 
```

So the clusters revealed four distinct customer spending profiles when analyzing all six product categories.

Cluster 1 (130 customers) stands out as big spenders on fresh products (13,728 units average) and frozen goods (3,864 units), but they're relatively light buyers of other categories meaning their detergent/paper spending averages just 480 units.

Cluster 2 is the smallest group **(61 customers)** with an unusual pattern. It has extremely low fresh spending (903 units) but heavy investment in groceries (10,129 units) and detergents/paper (3,426 units), suggesting these might be specialty stores or cleaning supply businesses.

Cluster 3 (119 customers) emerged as the broadest purchasers, leading in milk (9,028 units), grocery (13,276 units), and detergents/paper (4,777 units). These appear to be general-purpose retailers or supermarkets.

The final **Cluster 4 (130 customers)** shows restrained spending across all categories except moderate frozen purchases (1,318 units), likely representing small convenience stores or low-volume buyers.

```
# Calculating relative spending profiles
#(to compare where each cluster spends more/less)

# Calculating overall average spending in each category
overall_means <- colMeans(all_vars)

# Calculating relative spending (cluster center / overall mean)
relative_spending <- matrix(0, nrow = 4, ncol = 6)
for (i in 1:4) {
  relative_spending[i,] <- centers_original[i,] / overall_means
}

# Creating a data frame
relative_df <- data.frame(
  Cluster = 1:4,
  Fresh = round(relative_spending[,1], 2),
  Milk = round(relative_spending[,2], 2),
  Grocery = round(relative_spending[,3], 2),
  Frozen = round(relative_spending[,4], 2),
  Detergents_Paper = round(relative_spending[,5], 2),
  Delicassen = round(relative_spending[,6], 2)
)


print("Relative Spending Profiles (values > 1 indicate above-average spending):")
```

```
[1] "Relative Spending Profiles (values > 1 indicate above-average spending):"
```

```
print(relative_df)
```

```
  Cluster Fresh Milk Grocery Frozen Detergents_Paper Delicassen
1       1  1.14 0.53    0.44   1.26             0.17       0.93
2       2  0.08 1.02    1.27   0.10             1.19       0.16
3       3  0.62 1.56    1.67   0.44             1.66       1.06
4       4  0.49 0.20    0.20   0.43             0.06       0.26
```

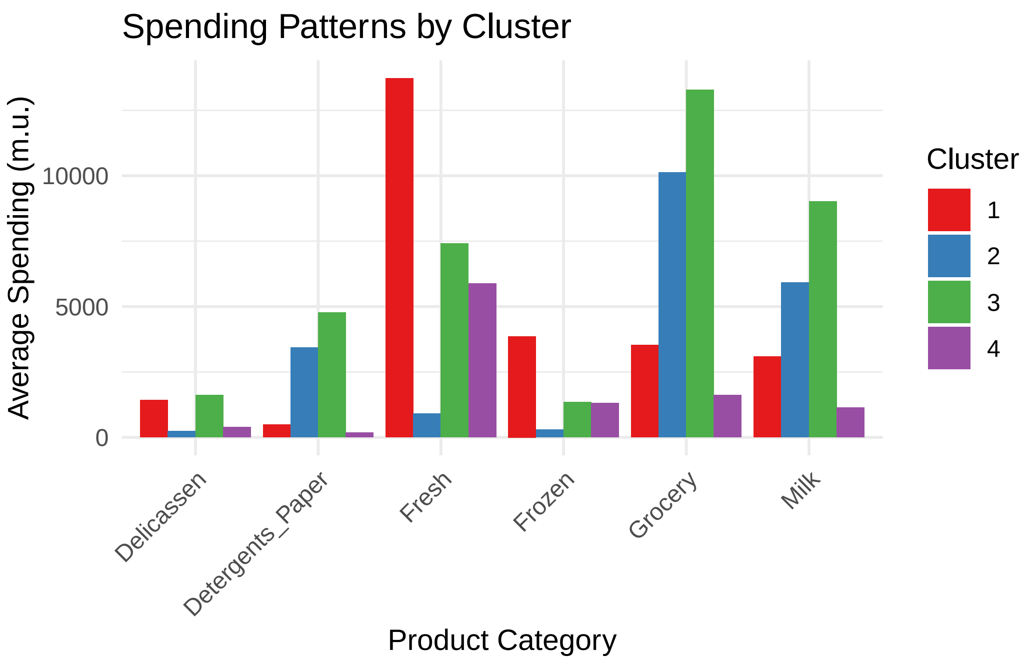
Cluster 1 spends 14% more on fresh products and 26% more on frozen goods compared to average customers, but their purchases look strikingly low elsewhere. So they buy only 17% of the average detergents/paper spending and half the typical milk expenditure. This makes me think they're likely restaurants or meal prep services focused on core ingredients.

Cluster 2 surprised me with their extreme specialization since they spend 27% more on groceries and 19% more on detergents/paper than average, but barely touch fresh (8% of average) or frozen (10%) products. These could be cleaning supply retailers combined with grocery stores.

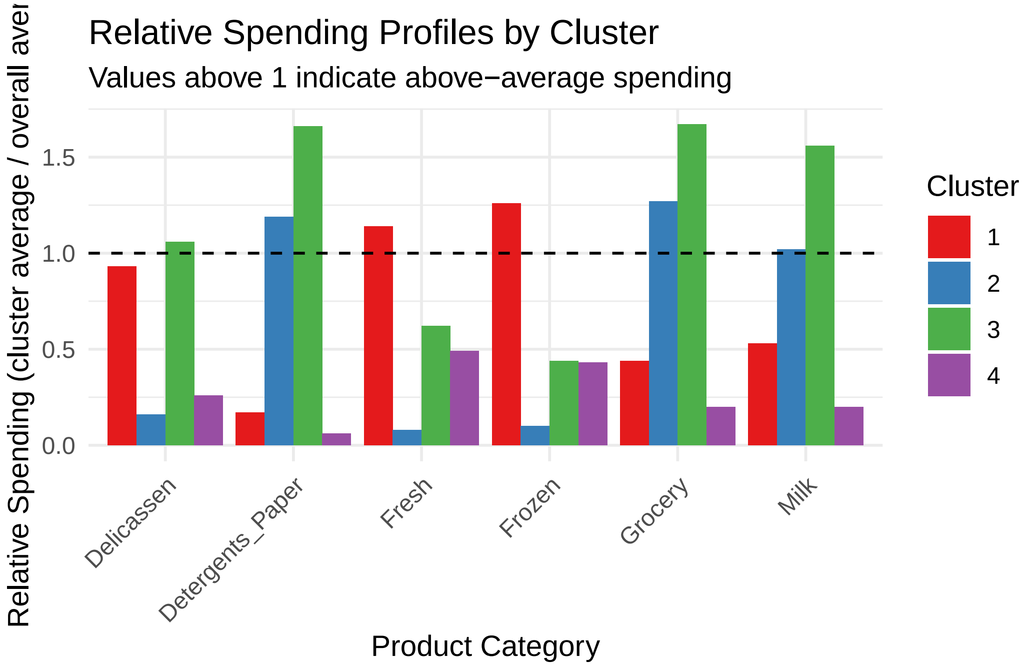
Cluster 3 emerged as the all-round big spenders, purchasing 56% more milk, 67% more groceries, and 66% more detergents/paper than average.

Cluster 4 is interesting. So every category sits well below average, with detergent/paper spending at just 6% of the mean. These appear to be extremely budget-conscious buyers or perhaps small corner stores with limited shelf space.

```r
library(tidyr)

# Reshape data for plotting
centers_long <- centers_df %>%
  pivot_longer(cols = -Cluster,
               names_to = "Category",
               values_to = "Spending")

# Plotting cluster centers
ggplot(centers_long, aes(x = Category, y = Spending, fill = factor(Cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1", name = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Spending Patterns by Cluster",
       x = "Product Category",
       y = "Average Spending (m.u.)")
```



```r
# Plotting relative spending profiles
relative_long <- relative_df %>%
  pivot_longer(cols = -Cluster,
               names_to = "Category",
               values_to = "RelativeSpending")

ggplot(relative_long, aes(x = Category, y = RelativeSpending, fill = factor(Cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1", name = "Cluster") +
  geom_hline(yintercept = 1, linetype = "dashed", color = "black") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Relative Spending Profiles by Cluster",
       subtitle = "Values above 1 indicate above-average spending",
       x = "Product Category",
       y = "Relative Spending (cluster average / overall average)")
```



So based on the cluster centers and relative spending patterns:

Cluster 1: Can be described as Fresh Food Specialists. So this group shows exceptionally high spending on Fresh products while having below-average spending in most other categories, suggesting these are likely restaurants, fresh markets, or food service businesses that prioritize fresh ingredients over processed or packaged goods.

Cluster 2: Grocery and Household Retailers. With their highest spending in Grocery, Milk, and Detergents/Paper products but relatively low spending on Fresh and Frozen items, these customers appear to be retail stores focused on shelf-stable products and household essentials.

Cluster 3: Full-Service Supermarkets. These customers have the highest overall spending across nearly all categories, particularly in Grocery, Detergents/Paper, and Milk, indicating they are likely large supermarkets or hypermarkets serving diverse consumer needs.

Cluster 4: Small General Retailers This cluster shows consistently below-average spending across all categories, with values roughly 1/3 to 1/2 of the overall means, suggesting these are small convenience stores or mini-markets with limited inventory requirements.

4.[10 points] Create a two-way table showing how the groups from Question #2 and #3 intersect. Give a brief description of how the two sets of clusters relate.

```
# For Question 2 - clustering based on Fresh and Frozen only
cluster_vars <- wholesale_data[, c("Fresh", "Frozen")]
cluster_vars_log <- log1p(cluster_vars)
cluster_vars_scaled <- scale(cluster_vars_log)
set.seed(123)
km_result_2d <- kmeans(cluster_vars_scaled, centers = 4, nstart = 25)

# For Question 3 - clustering based on all six variables
all_vars <- wholesale_data[, c("Fresh", "Milk",
                               "Grocery", "Frozen",
                               "Detergents_Paper",
                               "Delicassen")]
all_vars_log <- log1p(all_vars)
all_vars_scaled <- scale(all_vars_log)
set.seed(123)
km_result_6d <- kmeans(all_vars_scaled, centers = 4, nstart = 25)

# Creating a contingency table
contingency_table <- table(
  `Fresh-Frozen Clusters` = km_result_2d$cluster,
  `All Variables Clusters` = km_result_6d$cluster
)

print("Contingency Table of Cluster Assignments:")
```

```
[1] "Contingency Table of Cluster Assignments:"
```

```
print(contingency_table)
```

```
                     All Variables Clusters
Fresh-Frozen Clusters  1  2  3  4
                    1 87  0 21 24
                    2 33  7 60 55
                    3 10 19 32 35
                    4  0 35  6 16
```

The cross-analysis revealed fascinating overlaps between the two clustering approaches.

Fresh-Frozen Cluster 1 (from the 2-variable model) aligns strongly with All-Variables Cluster 1 . I see that 87 of its 130 customers stayed grouped together when considering all spending categories. But it also split into Clusters 3 and 4 in the full model, telling me that some customers who looked similar in fresh/frozen spending actually differ significantly in other categories like detergents or delicatessen.

The most dramatic shift appears in Fresh-Frozen Cluster 2 since only 7 customers remained in All-Variables Cluster 2, while 60 moved to Cluster 3 and 55 to Cluster 4.

This shows that customers who appeared as mid-range fresh/frozen spenders actually divide into distinct groups when considering their full purchasing patterns.

Interestingly, All-Variables Cluster 2 seems to draw customers almost exclusively from Fresh-Frozen Clusters 3 and 4 (19+35=54 of its 61 total).

This suggests the full model uncovered a specialty buyer group that the fresh/frozen-focused approach couldn't detect. Maybe likely those heavy detergent/paper spenders we saw earlier. The patterns confirm that while fresh/frozen spending provides some segmentation clues, considering all categories reveals more nuanced customer profiles that would be crucial for targeted marketing.

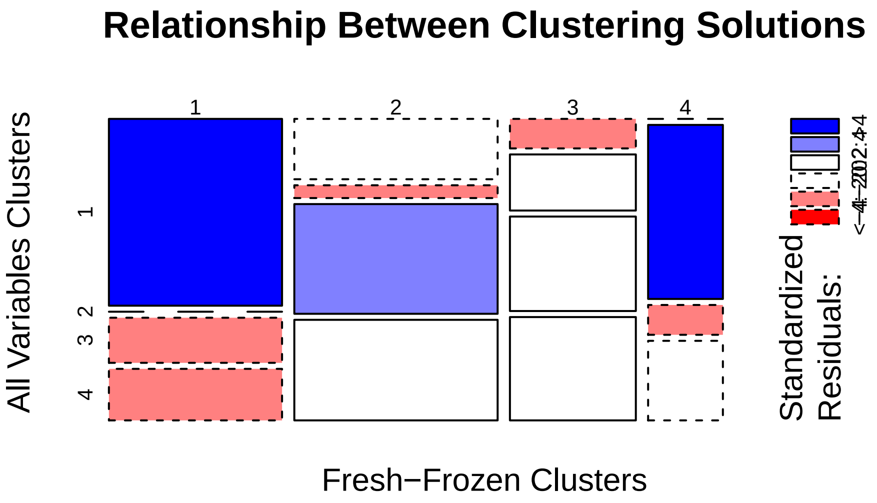
```
# Calculating percentages within Fresh-Frozen clusters
prop_table <- prop.table(contingency_table, 1) * 100
print("Percentage Distribution (row percentages):")
```

```
[1] "Percentage Distribution (row percentages):"
```

```
print(round(prop_table, 1))
```

```
                     All Variables Clusters
Fresh-Frozen Clusters    1    2    3    4
                    1 65.9  0.0 15.9 18.2
                    2 21.3  4.5 38.7 35.5
                    3 10.4 19.8 33.3 36.5
                    4  0.0 61.4 10.5 28.1
```

```
# Visualize the relationship with a mosaic plot
mosaicplot(contingency_table,
           main = "Relationship Between Clustering Solutions",
           xlab = "Fresh-Frozen Clusters",
           ylab = "All Variables Clusters",
           shade = TRUE,
           color = TRUE)
```

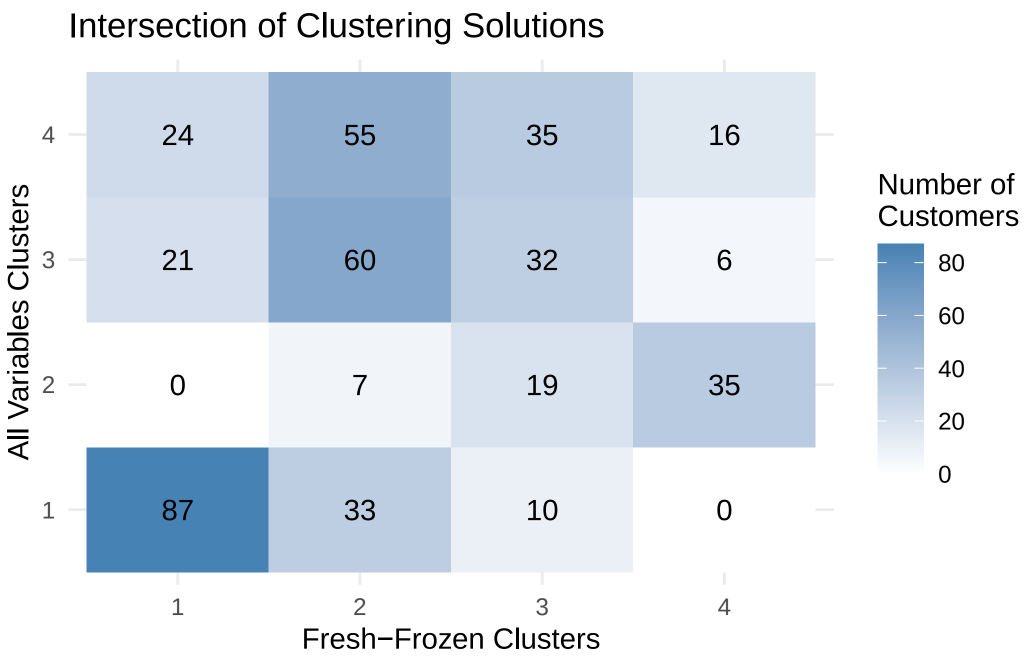


```
# Create a more detailed heatmap
library(ggplot2)
library(reshape2)
```

```
Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths
```

```
# Convert to data frame for ggplot
heatmap_data <- as.data.frame(as.table(contingency_table))
names(heatmap_data) <- c("Fresh_Frozen_Cluster",
                         "All_Variables_Cluster",
                         "Count")

# Create heatmap
ggplot(heatmap_data, aes(x = Fresh_Frozen_Cluster,
                         y = All_Variables_Cluster,
                         fill = Count)) +
  geom_tile() +
```

```r
geom_text(aes(label = Count), color = "black", size = 4) +
scale_fill_gradient(low = "white", high = "steelblue") +
theme_minimal() +
labs(title = "Intersection of Clustering Solutions",
     x = "Fresh-Frozen Clusters",
     y = "All Variables Clusters",
     fill = "Number of\nCustomers")
```


Intersection of Clustering Solutions
All Variables Clusters
Fresh−Frozen Clusters
Number of
Customers