Project 2 Helper Guide: Cluster Analysis Workflow

Overview

This guide will help you plan and run your cluster analysis for Project 2. Use it to:

- Choose a *clear* and *specific* research question
- Select and prepare relevant variables
- Clean and summarize your data
- Use cluster analysis to segment your data
- Interpret your results

Step 1: Choose a Question You Want to Answer

Think about **who** or **what** you want to cluster.

- Are you clustering **customers**, **stores**, or **products**?

- What is the goal of your analysis? (e.g., segmenting high vs. low activity stores, grouping customers by shopping behavior, high-selling vs. low-selling brands)

Write your question here:

My research question is:

What types of [X] exist based on [Y]? ("X" = the thing you are clustering (store, customer, or product segments); "Y" = spending habits, product diversity, visit frequency, etc.)

Step 2: Identify Your Unit of Analysis

What are the rows in your final dataset?

I am clustering:

☐ Stores

☐ Customers

☐ Products

☐ Something else: _____

Step 3: Join and Clean Your Data

Join the datasets you need to build the variables for your clustering question.

- shopper_info
- store_info
- gtin

What is my unit of analysis? (from Step 2)

What tables will I join and why?

Filter or transform your data:

- Remove invalid prices (e.g., `unit_price <= 0`)
- Keep or remove `gtin = NA` and `gtin = 0` (fuel) depending on your question
- Create total spending: `unit_price * unit_quantity`

What cleaning steps did you take and why?

Step 4: Build Your Summary Dataset

Now summarize your data to create one row per unit (store, customer, or product).

Write down 3–5 variables that might be useful for clustering:

Variable Name	What It Measures	Notes / Formula

Are these numeric variables?

Are they all available in the data?

Do you need to aggregate them (e.g., total sales per store, average items per visit)?

Step 5: Explore Your Variables

Use `ggpairs()` to check for skew and correlation.

```
library(GGally)

your_data %>%
  select(var1, var2, var3, ...) %>%
  ggpairs()
```

What do you notice?

- Are any variables very skewed?

- Are any variables highly correlated?

- Do any variables need to be log-transformed?

Make a list of any variables you want to log-transform:

Step 6: Transform and Scale Your Data

Log-transform skewed variables (add +1 to avoid log(0))

```
your_data <- your_data %>%
  mutate(
    log_var1 = log(var1 + 1),
    log_var2 = log(var2 + 1)
  )

## Create the dataset you'll use for clustering {.unnumbered}
cluster_data <- your_data %>%
  select(log_var1, log_var2, var3, ...) %>%
  drop_na()
```

Scale the data

```
cluster_scaled <- scale(cluster_data)
```

Which variables are in your scaled clustering data?

Step 7: Determine the Number of Clusters

Use these plots to decide on an optimal number of clusters:

```
fviz_nbclust(cluster_scaled, kmeans, method = "wss")
fviz_nbclust(cluster_scaled, kmeans, method = "silhouette")
```

What does the elbow plot suggest?

What does the silhouette plot suggest?

How many clusters will you use? Why?

Step 8: Run the Clustering Algorithm

```r
set.seed(123)
kmeans_fit <- kmeans(cluster_scaled, centers = X, nstart = 25)

final_clusters <- your_data %>%
  mutate(cluster = kmeans_fit$cluster)
```

```r
final_clusters %>%
  group_by(cluster) %>%
  summarise(across(everything(), mean))
```

How many observations are in each cluster?

What patterns do you notice when you group by cluster?

Step 9: Interpret and Visualize Clusters

Now use your original variables (not log-transformed) to summarize each cluster.

What makes each cluster different?

Do the results support your original question?

How would you describe each cluster in plain language?

Step 10: Add Context from Other Tables

Join back to gtin or shopper_info to see what's driving patterns.

Example: Most purchased subcategory by cluster.

```
your_data %>%
  left_join(cluster_info, by = "id") %>%
  group_by(cluster, subcategory) %>%
  summarise(count = n()) %>%
  filter(count == max(count))  # most common
```

Final Checklist

- ☐ I have a clear and specific research question
- ☐ I created a clean dataset at the right level of analysis
- ☐ I chose 3–5 good variables for clustering
- ☐ I transformed and scaled the data
- ☐ I chose an appropriate number of clusters
- ☐ I interpreted each cluster in context
- ☐ I documented my decisions throughout