Student Activity 5: Mapping mixed customer profiles

Distance design with mixed variables · Class activity and homework

Livio Finos, University of Padova

Contents

Scenario **1**

Learning goals 1

Part A — In class **2**

1. Inspect and classify the data 2
2. Define a defensible mixed-variable distance 2
3. Map the profiles and inspect the approximation 3
4. Compare with a bad shortcut 4

Part B — Homework extension **4**

Assessment checklist 4

Compare and revise **5**

Required report sections 5

Scenario

A retailer wants to make a map of 24 **synthetic, anonymous** customer profiles. The aim is descriptive: identify profiles that are similar across demographic, behavioural and preference variables. It is not a prediction task and it is not permission to infer sensitive personal characteristics. The data were designed for this activity; they do not describe real people.

The difficulty is fundamental: the columns do not all have the same kind of meaning. Age and income are numeric, loyalty has an order, preferred channel and region are labels, and discount use is binary. Before calculating a distance, we must decide how each type of difference contributes.

Learning goals

By the end, you should be able to:

- classify variables as numeric, ordinal, nominal, binary or identifiers;
- explain why arbitrary numerical codes for categories can create a false geometry;
- calculate and inspect Gower dissimilarities for mixed data;
- make and qualify a two-dimensional MDS map; and
- report the choices of variables, types and weights that define a distance.

Part A — In class

Work in groups of three or four. Submit one short, reproducible R Markdown or Quarto report per group. Every graph must have labelled axes and an interpretation in words. Do not begin by asking an AI assistant or by searching for a ready-made analysis: first record the choices below in your report.

1. Inspect and classify the data

Read `data/customer_profiles.csv`. Make a data dictionary with one row per column. For each variable, state its type and whether it should enter a customer distance. Explain why `customer_id` must be retained for labels but excluded from the calculation.

Also state what one row represents, what one pairwise distance represents, and one use of the resulting map that would be defensible versus one that would not.

> **Plan before prompting.** Save this initial specification before requesting help from an AI assistant. If you use AI later, give it your specification, verify every suggestion, and keep a decision log explaining what you accepted, changed or rejected.

```
## Rows: 24
## Columns: 9
## $ customer_id       <chr> "C01", "C02", "C03", "C04", "C05", "C06", "C07", "C0~
## $ age               <dbl> 24, 29, 35, 42, 51, 47, 32, 27, 39, 56, 61, 54, 45, ~
## $ income_k_eur      <dbl> 24, 31, 43, 58, 76, 69, 38, 28, 52, 88, 105, 92, 64,~
## $ visits_last_month <dbl> 7, 9, 6, 4, 3, 5, 12, 10, 8, 2, 1, 3, 6, 11, 13, 1, ~
## $ loyalty_tier      <ord> Bronze, Bronze, Silver, Silver, Gold, Gold, Bronze, ~
## $ preferred_channel <fct> Mobile, Mobile, Web, Store, Store, Web, Mobile, Mobi~
## $ region            <fct> North, North, North, Central, Central, Central, Sout~
## $ uses_discount     <fct> Yes, Yes, No, No, No, No, Yes, Yes, Yes, No, No, No,~
## $ returns_last_year <dbl> 1, 0, 1, 2, 0, 1, 3, 2, 1, 0, 0, 1, 1, 2, 1, 0, 1, 2~
```

```
##     customer_id          age         income_k_eur    visits_last_month
##  Length   :24   Min.   :23.00   Min.   : 22.00   Min.   : 1.000
##  N.unique :24   1st Qu.:31.50   1st Qu.: 37.00   1st Qu.: 3.000
##  N.blank  : 0   Median :41.50   Median : 56.50   Median : 5.500
##  Min.nchar: 3   Mean   :42.75   Mean   : 61.67   Mean   : 6.042
##  Max.nchar: 3   3rd Qu.:53.25   3rd Qu.: 82.75   3rd Qu.: 8.250
##                 Max.   :67.00   Max.   :118.00   Max.   :14.000
##    loyalty_tier preferred_channel     region  uses_discount returns_last_year
##  Bronze  :7     Mobile:7          Central:7   No :14        Min.   :0.0000
##  Silver  :6     Store :9          North  :9   Yes:10        1st Qu.:0.0000
##  Gold    :6     Web   :8          South  :8                 Median :1.0000
##  Platinum:5                                                Mean   :0.9583
##                                                            3rd Qu.:1.2500
##                                                            Max.   :3.0000
```

2. Define a defensible mixed-variable distance

Gower dissimilarity is a useful default for this setting. For each pair of customers it is the weighted average of variable-specific dissimilarities:

$$d_G(i,j) = \frac{\sum_{k=1}^{p} w_k \, \delta_{ijk} \, d_{ijk}}{\sum_{k=1}^{p} w_k \, \delta_{ijk}}.$$

Here w_k is the chosen weight for variable k, $\delta_{ijk} = 1$ when that variable is observed for both customers (and 0 otherwise), and $0 \leq d_{ijk} \leq 1$ is its variable-specific difference. For a numeric variable,

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{\max(x_{\cdot k}) - \min(x_{\cdot k})};$$

for a nominal variable it is 0 for a match and 1 for a mismatch. An ordered factor is compared through the distance between its ordered ranks, scaled to the range from 0 to 1. Thus a difference of 10 years and a difference of 10 thousand euros are not compared in raw units. The formula also makes two decisions visible: which variables enter the distance, and which weights they receive.

Before using software, complete the following audit in your report:

1. Decide whether discount use is symmetric binary (both values informative) or asymmetric binary (only one value informative). Defend the choice.
2. State whether all selected variables initially have equal weights, and what "equal" means after numeric variables have been range-scaled.
3. Compute by hand the contributions of **age**, **income**, and **preferred channel** for two chosen pairs. Show the ranges used and what the partial calculation does not yet include.
4. Predict one close pair and one distant pair before computing the full matrix.

Use `cluster::daisy()` below. Check the help page and state how the function treats each of your columns. Then identify the three closest and three most distant customer pairs. Are the results plausible from the original rows?

In addition, check for missing values, duplicated profiles and zero distances between distinct customers. A distance value alone is not an interpretation: every claim about a close or distant pair must be checked against the original variables. Compare your prediction with the calculated results and identify one agreement and one surprise.

Table 1: Three closest and three most distant profile pairs under Gower dissimilarity.

Comparison	Customer 1	Customer 2	Gower dissimilarity
Closest	C11	C24	0.012
Closest	C04	C21	0.033
Closest	C10	C20	0.058
Most distant	C16	C22	0.902
Most distant	C07	C24	0.913
Most distant	C07	C16	0.934

3. Map the profiles and inspect the approximation

Fit classical MDS to the Gower matrix. Colour points by loyalty tier and use point shape for preferred channel. Describe two visible patterns and one important limitation of the map. Remember that proximity is the evidence; a horizontal or vertical axis has no intrinsic interpretation.

Your report must also include a distance-reproduction (Shepard-style) plot: original Gower dissimilarities on the horizontal axis and two-dimensional map distances on the vertical axis. Report their correlation and identify a pair with a substantial discrepancy from the diagonal.

```
## [1] 0.9447721
```

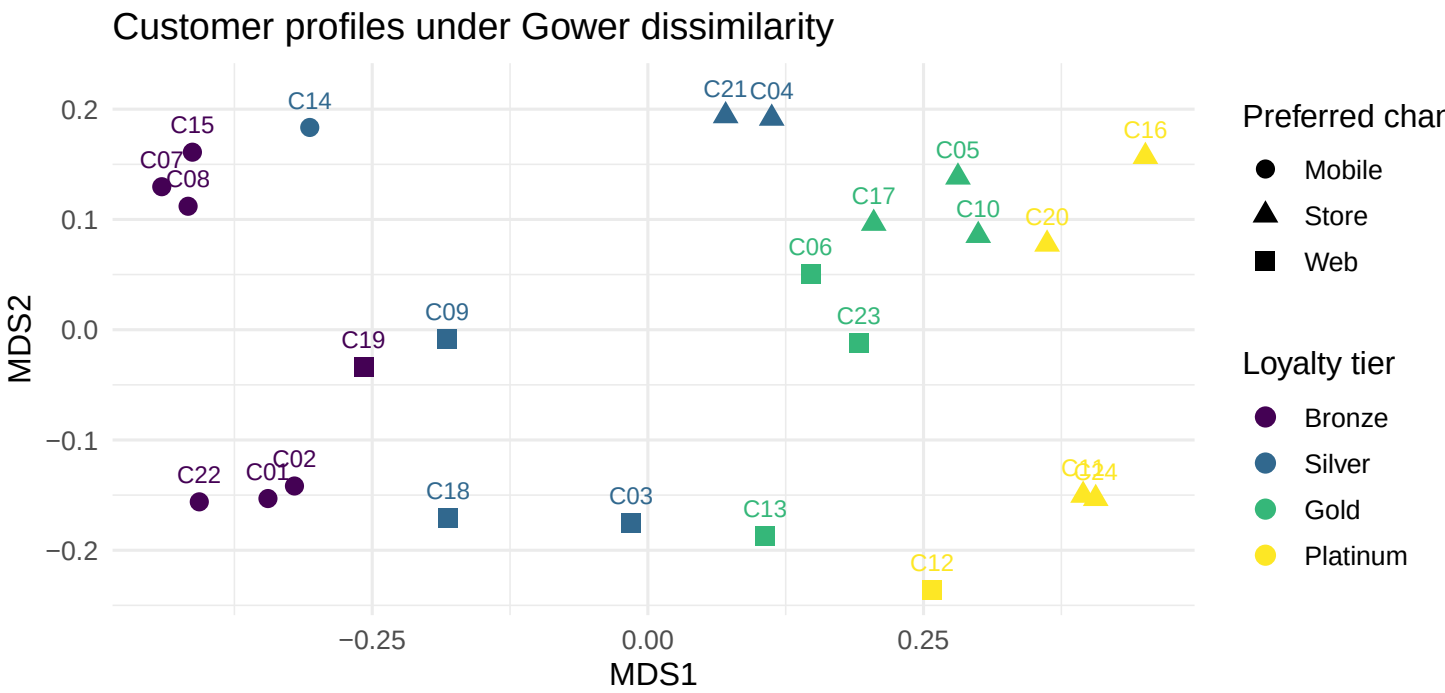

Figure 1: A two-dimensional classical-MDS display of mixed customer profiles. It is an approximation to the Gower dissimilarities.

4. Compare with a bad shortcut

Someone proposes encoding `preferred_channel` as Mobile = 1, Store = 2, Web = 3, and then computing Euclidean distance. Explain why this imposes an invented order and why the size of the gaps is arbitrary. Demonstrate the issue by changing the codes, or compare it with Gower. Do not present the shortcut as an alternative recommended analysis.

Part B — Homework extension

Complete these questions individually. Limit the report to four pages of text, tables and figures (appendix code does not count).

1. **Sensitivity to weights.** The retailer considers recent visits twice as important as region. Recompute Gower dissimilarity using `weights` in `daisy()`. Which nearest-neighbour relationships change? Is the change useful for the stated purpose, or merely convenient?
2. **Sensitivity to variables.** Refit the map without age and income. Explain what substantive question the revised distance now answers.
3. **Map quality.** Plot original Gower dissimilarities against distances on the two-dimensional map. Report their correlation and identify a pair whose distance is noticeably poorly reproduced.
4. **Decision memo.** Write 150–200 words specifying the objects, variables, variable types, weights, distance, map limitation and one ethical limitation of customer profiling.

Assessment checklist

Criterion	Evidence
Data understanding	Accurate data dictionary; identifier excluded; types justified.
Distance design	Clear explanation of Gower and of at least one modelling choice.
Visual reasoning	Labelled MDS map, explicit interpretation, no invented axis meaning.
Reproducibility	Code runs from the course folder and uses the supplied CSV.
Critical reasoning	Sensitivity comparison and a limitation stated clearly.

Compare and revise

Exchange reports or displays with another group. Compare variable types, binary convention, weights, close/distant pairs and map claims. If results differ, trace the difference to a documented choice. Revise your report and state one point of agreement, one difference and one change you made (or explain why no change was needed).

Required report sections

Section	Required content
Initial distance plan	Objects, purpose, data dictionary, types, binary convention and weights
Manual audit	Two partial hand calculations, predictions and explanation of Gower components
Baseline results	Pair table, source-variable comparison, MDS map and distance-reproduction plot
Sensitivity analysis	Weight and scope comparisons, with stable and changed conclusions
Comparison and revision	What another group revealed and how the final report changed
Reproducibility and AI appendix	Code, package versions, decision log, relevant prompts and AI statement

An AI-assistance statement can follow this template, adapted to the work actually performed:

> **AI assistance.** We used [tool/model, if available; date] to assist with [specific tasks]. We checked [specific code, calculations, and interpretations] and revised [main issues]. The group takes responsibility for the submitted analysis and conclusions.

Note on AI assistance

These teaching materials were prepared with assistance from ChatGPT (OpenAI) in drafting explanations, developing examples, and generating R code. The author retains responsibility for the final content.