Assignment 04: Unsupervised Learning

WS24 Data Analytics

Lucas Kook

Goal

The goal of the analysis is to perform customer segmentation using cluster analysis for an online retailer based on the RFM ("Recency, Frequency, Monetary Value") strategy. RFM is used for analyzing and estimating the value of a customer, based on three data points: Recency (How recently did the customer make a purchase?), Frequency (How often do they purchase), and Monetary Value (How much do they spend?).

Dataset

Source: http://archive.ics.uci.edu/ml/datasets/online+retail

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

You can download the data from Canvas and load the data in R by using the following command:

```r
# Note: the path may need to be adapted
rmf <- readRDS("data/rmf.rds")
```

Tasks

- Visualize the data using a pairs plot and a biplot using the first two principle components of a PCA.
- Perform clustering using k-means (choose the number of clusters optimally based on the elbow method), hierarchical clustering and DBSCAN. Produce another biplot (see first task) for each clustering method and color the points based on the clustering. Comment on the results.

- Suggest a final grouping of the data and explain your choice of clustering algorithm and hyperparameters. Interpret the results by identifying the "representative" customer for the segments (what is the average Recency, Frequency and Monetary value in each cluster).

Submission

Note: Only *one* submission per team.

- The assignment should be completed as a notebook containing code and text explanations.
- If you submit an R notebook, please submit the `.html` output together with the `.Rmd` (or `.qmd`) file.