Assignment 1

Due date: 31 Jan 2022, 23:59:59
Submission format: R file

Here is the data description of the dataset “bank-market.csv”.

Column	Name	Description
1	customerid	Customer ID
2	age	continuous: age of the customer
3	marital	categorical: "married","divorced","single" (divorced include widowed)
4	education	categorical: "unknown","secondary","primary","tertiary"
5	balance	continuous: average yearly balance, in euros
6	housing	has housing loan? (binary: "yes","no")
7	loan	has personal loan? (binary: "yes","no")
8	duration	continuous: last contact duration, in seconds
9	campaign	no of contacts performed during this campaign (numeric)
10	pdays	no of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
11	previous	no of contacts performed before this campaign and for this client (numeric)
12	poutcome	outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
13	current	Is the client a active client? (binary: 1 = “yes”, 0 = “No”)
14	deposit	has the client subscribed a term deposit? (binary: "yes","no")

Question 1 (Data cleaning and validation) (10 Bonus marks)

(a) Clean the provided dataset (do not remove any column), store it as “df”.

Question 2 (Clustering Comparison 1) (10 marks)

(a) Remove the column of “customerid” from the “df”. Use daisy() to calculate the Gower Distance of “df” and store the result as “gd1”.
(b) Use the “gd1” to perform k-means with k = 2 with seed = 123. Store the result as “km1”.
(c) Remove the column “current” and store the new dataset as “df2”. Use daisy () to calculate the Gower Distance of “df2” and store the result as “gd2”.
(d) Use the “gd2” to perform k-means with k = 2 with seed = 123. Store the result as “km2”.
(e) Compare the clustering result of “km1” and “km2” (no need to see the clusters’ characteristics). What do you observe and why? (Word limit: 50)

Question 3 (Clustering Comparison 2) (10 marks)

(a) Select all numeric columns from “df”1 and store them as a new dataset “df_num”. Use daisy() to calculate the Gower Distance of “df_sum” and store the result as “gd_num”.
(b) Use the “gd_num” to perform k-means with k = 2 with seed = 123. Store the result as “km_num”.
(c) Use scale() on “df_num” and store the result as “df_scale”. Use daisy() to calculate the Gower Distance of “df_scale” and store the result as “gd_scale”.
(d) Use the “gd_scale” to perform k-means with k = 2 with seed = 123. Store the result as “km_scale”.
(e) Compare the clustering result of “km_num” and “km_scale” (no need to see the clusters’ characteristics). What do you observe and why? (Word limit: 50)

Question 4 (Clustering Comparison 3) (6 marks)

(a) Use dist() to calculate the Euclidean Distance of “df_num” and “df_scale”, and store the results as “eu” and “eu_scale” respectively.
(b) Use “eu” and “eu_scale” to perform k-means with k = 2 with seed = 123 respectively. Store the results as “km_eu” and “km_eu_scale”.
(c) Compare the clustering result of “km_eu” and “km_eu_scale” (no need to see the clusters’ characteristics). What do you observe and why? (Word limit: 50)