Advanced Business Analytics, Data Imputation Techniques
Course ID: 229091-D, SMMB, course type: lab

Project 2, deadline: 31th May 2025
You are employed as a data scientist at a retail bank. Your manager has requested you to analyze the data sample on customers who have created an account over the last 3 months in order to perform customer segmentation and determine the key customer segments. The main purpose of the analysis is to allow the bank to better distinguish specific subgroups of clients. This as a result could help to improve targeting and risk assessment.

Information on customers can be found in the dataset ***data_retail.csv***. Key information on the abovementioned dataset is available below.

Feature name	**Description**
TNX_ID	ID of the transaction
CUST_ID	ID of the customer
CUST_BIRTH_DATE	Birth date of the customer
GENDER	Gender of the customer
CITY	Customer location
BALANCE_AMT	Account balance
TNX_DATE_DDMMYY	Transaction date
TNX_TIME_HHMMSS	Transaction time (in hhmmss or hmmss format)
TNX_AMT	Transaction amount in local currency

Task description:

Based on abovementioned dataset prepare the following tasks:

1. ***Exploratory data analysis and feature engineering***

Conduct exploratory data analysis (i.a. missing values (if any), descriptive statistics of characteristics and their distributions, etc.). Create new features that can be used to obtain additional information about the analyzed customer portfolio. Analyze the relationships between the features and generate appropriate visualizations. Based on the analyzes performed, select the variables that you will use to build the segmentation model and briefly justify your choice.

2. ***Building the segmentation model***

Using the selected variables (make relevant transformations if necessary), build a segmentation model using the K-means method. Briefly justify the choice of the optimal number of clusters (using at least 2 different methods), as well as the choice of optimal cluster initialization points.

3. ***Business analysis***

Describe the groups of customers selected on the basis of the model and interpret the statistics for individual segments and their business characteristics. Visualize the segments.

4. ***Alternative clustering methods****

Perform additional analysis using alternative clustering methods (i.e. different than PCA and K-means algorithm). Compare the results with the previously obtained ones.

Score:
10 points maximum, out of which:
Correctness of results and interpretation: 3 (weight: x2)
Programming (possibility of reproducing results, code readability, comments): 2
Aesthetics of work and completeness of materials: 1
Innovation of the proposed solution: 1
Possibility of obtaining additional 2 points for an extra task (*), i.e. max 12/10.

Task submission:
Please upload your solution to the MS Teams group. The solution should include:
a. Text file with the solution to the task, analysis results, model parameter estimates and visualizations along with description and conclusions (.pdf). **All authors should be listed at the very beginning of the report.**
b. Program code (SAS, R or Python) containing the definition of the SAS library (if applicable)/working directory and the libraries used at the beginning of the program code and their versions (if applicable; e.g. for Python based solutions it is best to attach a separate requirements.txt file with versions of packages used) to enable reproduction of results after changing the working directory/input data path.
c. A printout containing the default set of results obtained using the prepared program code (for open source languages could be e.g. jupyter notebook or R markdown report).