Data Preprocessing

Project Description

2024/2025

NOVA IMS Information Management School

Joana Neves

jneves@novaims.unl.pt

Project Description

In today's healthcare scene, patient care and satisfaction are vital. Hospitals must continuously seek ways to differentiate themselves and understand patient needs. Therefore, City Hospital, which provides services across multiple departments, aims to leverage the data collected by its information systems to enhance patient care and operational efficiency.

The data available represents patient interactions and treatments across various departments, reflecting the hospital's overall performance and patient demographics. To harness this data effectively, City Hospital's management has assembled a team of data scientists to analyze and segment patient information. Within this team, there is a dedicated subgroup focused on data preprocessing (DP Team).

The DP Team's role is to prepare the data for advanced analytical methods and provide initial insights into hospital operations and patient care patterns. This is crucial as the hospital currently lacks comprehensive information on its activities and patient behaviors.

City Hospital requires an exploratory analysis to address fundamental operational questions and an analytic-based table (ABT) for descriptive analysis and patient segmentation. Essentially, the DP Team aims to utilize data from the hospital's information systems to create an ABT, which will then be handed over to the next team for further analysis and implementation.

Description of the transactional table variables:

Variable	Description
Patient ID	Unique identification of the patient
Age	Patient age
Gender	Patient gender (Male, Female, Other)
City of Residence	Patient city of residence
Profession	Patient profession
Insurance Provider	Patience insurance provider
Family History	Patient family history diseases
Education Level	Patient education level
Marital Status	Patient marital status
Visit Date	Date of the consultation
Department	Consultation department
Consultation Duration	Consultation duration in minutes
Satisfaction Level	Patient evaluation of the satisfaction level with the consultation (1-5)
Approximate Annual Income	Patient approximate annual income
Consultation Price	Consultation price (pounds)
Insurance Coverage	Amount of the consultation price covered by the insurance provider (pounds)

Requirements:

1. Preprocess the data in order to do clusters with the patients
2. Build an ABT
3. Withdraw some insights using visualization tools

Notes:

- The transactional table will be given by the professor
- The software that can be used are Excel, SAS (Enterprise Guide and/or Miner), PowerBI or any other you may want to use
- Here are some of the requirements to perform clustering, make sure you follow all the requirements. However: (1) you don't need to standardize the data, (2) you can leave some qualitative variables; (3) check for multicollinearity and if exist, highlight the most important correlations (but keep all variables that you created)

Requirement	Clustering
Quantitative Variable	DESIRABLE
No Missing values	**YES**
No Outliers	**YES**
Homocedascity	NA
Low Multicollinearity	**YES**
Normally distributed variables	NA
Standardized variables	**YES**

Suggestion (guidelines):

1. Perform some initial descriptive statistics (SAS Miner)
2. Treat outliers (SAS Miner)
3. Treat missing values (SAS Miner)
4. Check final statistics (SAS Miner)
5. Check the coherence (SAS Miner/SAS Guide)
6. Transform and create derived variables (SAS Guide/Miner)
7. Create the final ABT (SAS Guide)
8. Check for multicollinearity (SAS Guide/SAS Miner)
9. Create some visualizations (PowerBI)

Deliverables:

- ABT in Excel;
- PDF Report (reporting all steps of your project and interpretation of each step);
- Document with visualizations (in PowerBI)
- Presentation document (in PDF)

All documents must be **submitted to moodle** until **December 10 (23h59m)**.