Universidad Rafael Landivar
Facultad de ingenieria
Analisis de datos

Laboratorio Clustering en R

En este laboratorio se trabajaran dos secciones, descritas a continuacion:

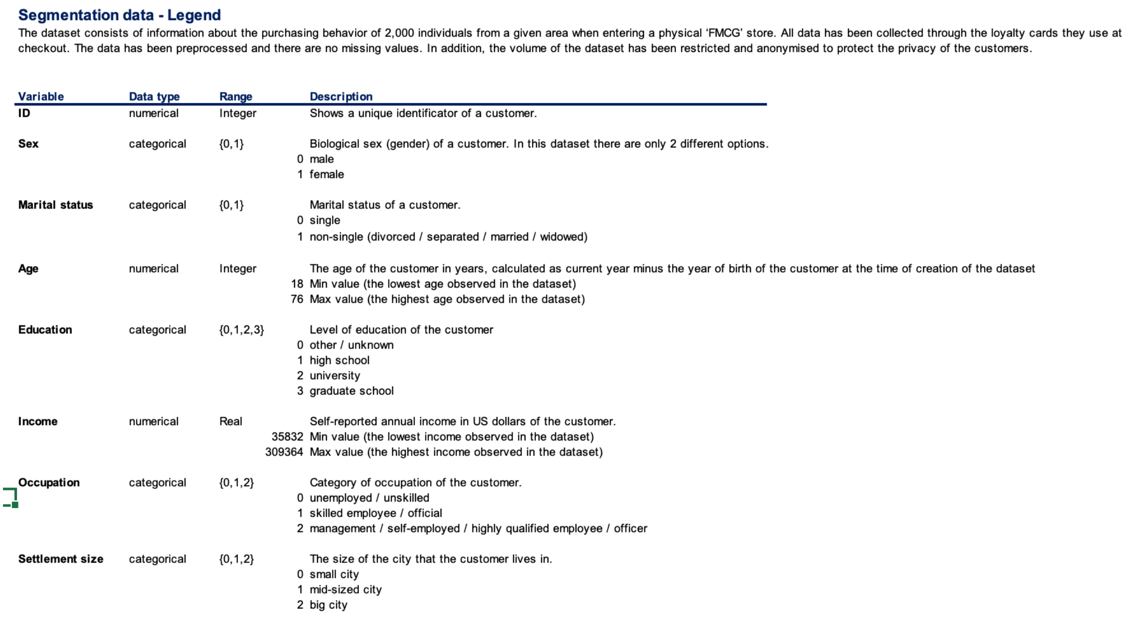
Parte I – clustering y la ley del codo

Para el dataset *“segmentation data.csv”* efectuar un analisis de clustering determinando lo siguiente:

1. Determinar el numero ideal de clusters, de acuerdo a la ley codo, justificando de manera grafica su decision.
2. Efectuar el clustering con el numero ideal de clusters determinado por su persona y graficar (con las herramientas vistas en clase) como quedarian los datos agrupados.

Puede utilizar la funcion kmeans para efectuar su trabajo. Tambien determine bajo una inspeccion de los datos si todas las columnas son necesarias. Si alguna no es necesaria removerla del dataset.

Se adjunto el significado de cada columna a continuacion:



Segmentation data - Legend

The dataset consists of information about the purchasing behavior of 2,000 individuals from a given area when entering a physical ‘FMCG’ store. All data has been collected through the loyalty cards they use at checkout. The data has been preprocessed and there are no missing values. In addition, the volume of the dataset has been restricted and anonymised to protect the privacy of the customers.

Variable	Data type	Range		Description
ID	numerical	Integer		Shows a unique identificator of a customer.
Sex	categorical	{0,1}		Biological sex (gender) of a customer. In this dataset there are only 2 different options.
			0	male
			1	female
Marital status	categorical	{0,1}		Marital status of a customer.
			0	single
			1	non-single (divorced / separated / married / widowed)
Age	numerical	Integer		The age of the customer in years, calculated as current year minus the year of birth of the customer at the time of creation of the dataset
			18	Min value (the lowest age observed in the dataset)
			76	Max value (the highest age observed in the dataset)
Education	categorical	{0,1,2,3}		Level of education of the customer
			0	other / unknown
			1	high school
			2	university
			3	graduate school
Income	numerical	Real		Self-reported annual income in US dollars of the customer.
			35832	Min value (the lowest income observed in the dataset)
			309364	Max value (the highest income observed in the dataset)
Occupation	categorical	{0,1,2}		Category of occupation of the customer.
			0	unemployed / unskilled
			1	skilled employee / official
			2	management / self-employed / highly qualified employee / officer
Settlement size	categorical	{0,1,2}		The size of the city that the customer lives in.
			0	small city
			1	mid-sized city
			2	big city

Entregables:

1. Su codigo en un R notebook.
2. Pdf explicando sus decisiones en relacion al numero de clusters y alguna otra decision de diseño que haya podido tomar.

Parte 2 – Clustering y feature engineering

Para el dataset “marketing_campaign.csv” que contiene la siguiente estructura de columnas:

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Efectuar un analisis de clustering, pero antes de poderlo efectuar, es necesario tratar el dataset mediante las tecnicas de feature engineering vistos en clase (a excepcion de la normalizacion que aun no lo hemos tratado). Encontraran datos faltantes, variables categoricas (alfanumericas), caracteristicas (columnas) que posiblemente no son aceptables para nuestro analisis, por lo que es necesario efectuar este proceso como paso previo a efectuar el clustering.

Una vez efectuado este paso previo:

1. Determinar el numero ideal de clusters, de acuerdo a la ley codo, justificando de manera grafica su decision.
2. Efectuar el clustering con el numero ideal de clusters determinado por su persona y graficar (con las herramientas vistas en clase) como quedarian los datos agrupados.

Entregables:

1. Su codigo en un R notebook.
2. Pdf explicando sus decisiones en relacion al numero de clusters y alguna otra decision de diseño que haya podido tomar.