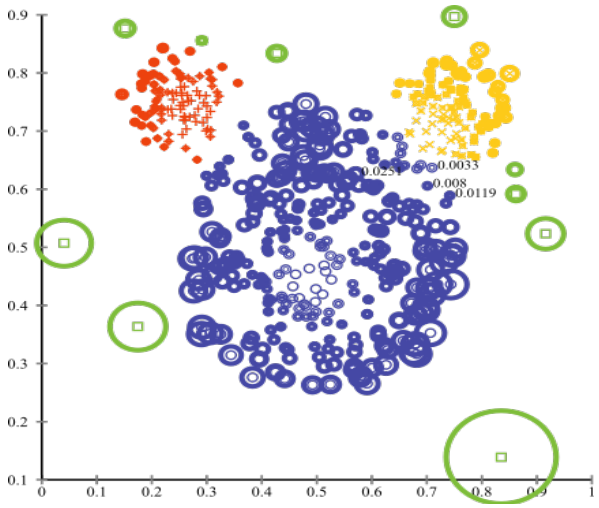
Universidad Rafael Landívar
Facultad de Ingeniera
Análisis de Datos
Ing. Jairo Salazar

# Clustering y Feature Engineering

---






# Clustering

Una empresa ha ido recopilando una base de datos de los clientes que llegan a sus instalaciones. Se desea efectuar un estudio para determinar y clasificar los segmentos de clientes que visitan el centro comercial. Se le ha contratado para efectuar este estudio:
De la base de datos llamada “segmentation data.csv”
i. Genere la estadística General de los datos. (se recomienda usar la función summary de R).
ii. Efectué la limpieza de los datos según las siguientes instrucciones:

- El dataset contiene una columna llamada ID que identifica de manera única a cada registro, esta columna no es útil para poder efectuar clustering por lo que debe de remover esta columna.
- Verifique que no existan datos nulos.
- Vamos a poner en practica el feature scaling, como se menciono en la teoría, al tener características (features) en diferentes escalas, puede ser mas complicado de lograr que los algoritmos logren converger. Efectuar las siguientes actividades:
    1. Con el dataset modificado de acuerdo a lo anteriormente expuesto, efectue un análisis de clustering utilizando la funcion kmeans, evaluando agrupaciones de 2 a 25 clusters. Guardando en un dataset de los errores y la cantidad de clusters (como se efectuo en el kmeans demo en clase).

2. Basandose en la ley del codo determine el numero optimo de clusters, que se deben de utilizar. Grafique los clusters finales con el numero optimo con la funcion plot de la librería USEFUL.
3. Transforme la columna Income de acuerdo a la funcion de standarization, siguiendo la siguiente formula:

$$Z = \frac{x_i - \bar{\mu}}{\sigma}$$

Con el dataset modificado, remueva la columna original y efectue el mismo ejercicio del punto 1, con la columna de income modificada (o reducida en su dimensionalidad). Guarde un dataset de los errores y la cantidad de clusters (como se efecuo en el kmeans demo en clase).

4. Determine en base a la ley del codo el numero optimo de clusters que se deben de utilizar. Grafique los clusters finales con el numero optimo con la funcion plot de la librería USEFUL.

Preguntas:

1. Existe diferencia en el nivel de errores entre el dataset con el Income original o el dataset con el Income estandarizado? Explique lo encontrado y exponga brevemente la razón principal por la cual existen o no existen diferencias.

2. Existe diferencia en el numero de clusters optimos entre el dataset con el Income Original o el dataset con el income estandarizado? Explique lo encontrado y exponga brevemente la razón principal por la cual existen o no existen diferencias.

Entregables:

1. Codigo en R.

2. Documento en PDF, con cada una de las graficas (ley del codo y clustering optimo) y dataset de los errores resultantes del proceso de kmeans con el dataset con el Income original y el dataset con el Income estandarizado. Tambien responder las preguntas planteadas en dicho documento.