PROPUESTA PROYECTO CURSO APRENDIZAJE NO SUPERVISADO
MAESTRIA EN INTELIGENCIA ANALITICA DE DATOS
ENTREGA 2

Alvaro Pabón
Jorge Lopez
Maria Ana Ortiz

1. **Elegir una de las preguntas utilizando la retroalimentación recibida en el avance anterior y expandir sobre su motivación a responderla.**

La pregunta a realizar es la siguiente:

¿Quiénes son mis clientes ?

Se requiere segmentar a los clientes de tal forma que se pueda entender cuales son los tipos de clientes a los que se les está vendiendo en mayor y también en menor medida. Se considera un aprendizaje no supervisado ya que no existe variable de salida a diferencia del supervisado.

Motivación:

Con esta segmentación podemos responder estas preguntas :

¿Qué tipo de productos se les puede ofrecer ?

Al saber quienes son mis clientes y tenerlos segmentados por ciertas características se podría crear productos financieros nuevos o estrategias comerciales para cada uno de estos segmentos.

Mejorar las capacitaciones de mis asesores de ventas nuevos

La empresa desea incorporar este nuevo conocimiento del cliente en las capacitaciones para aquellos asesores que entren a trabajar para así muy rápidamente perfilar y ejecutar una venta mucho más rápida.

2. **Describir si el problema pertenece a una tarea de reducción de dimensión, clustering, o una combinación de los dos, explicando el por qué.**

El problema pertenece a una combinación de reducción de dimensión y de clustering.

En reducción de dimensión debido a que esto nos puede ayudar a realizar una exploración de datos y empezar a divisar qué variables son más importantes para la segmentación para posteriormente desarrollar una metodología de clustering que nos permita segmentar adecuadamente los clientes.

3. **Proponer al menos un algoritmo/técnica de aprendizaje no supervisado que ustedes crean que es la adecuada para responder la pregunta.**

   Analisis de componentes principales para la exploración de datos y analizar qué variables son más importantes que otras.

   Y para crear grupos puede ser kmeans o kmedois y poder establecer cuales serían el número de segmentos óptimo.

4. **Presentar estadísticas descriptivas utilizando tablas y/o visualizaciones de los datos crudos que tengan a su disposición. Describir el plan que van a seguir para tener los datos listos. Por ejemplo, cómo van a limpiarlos, que van a hacer con los datos faltantes, etc**

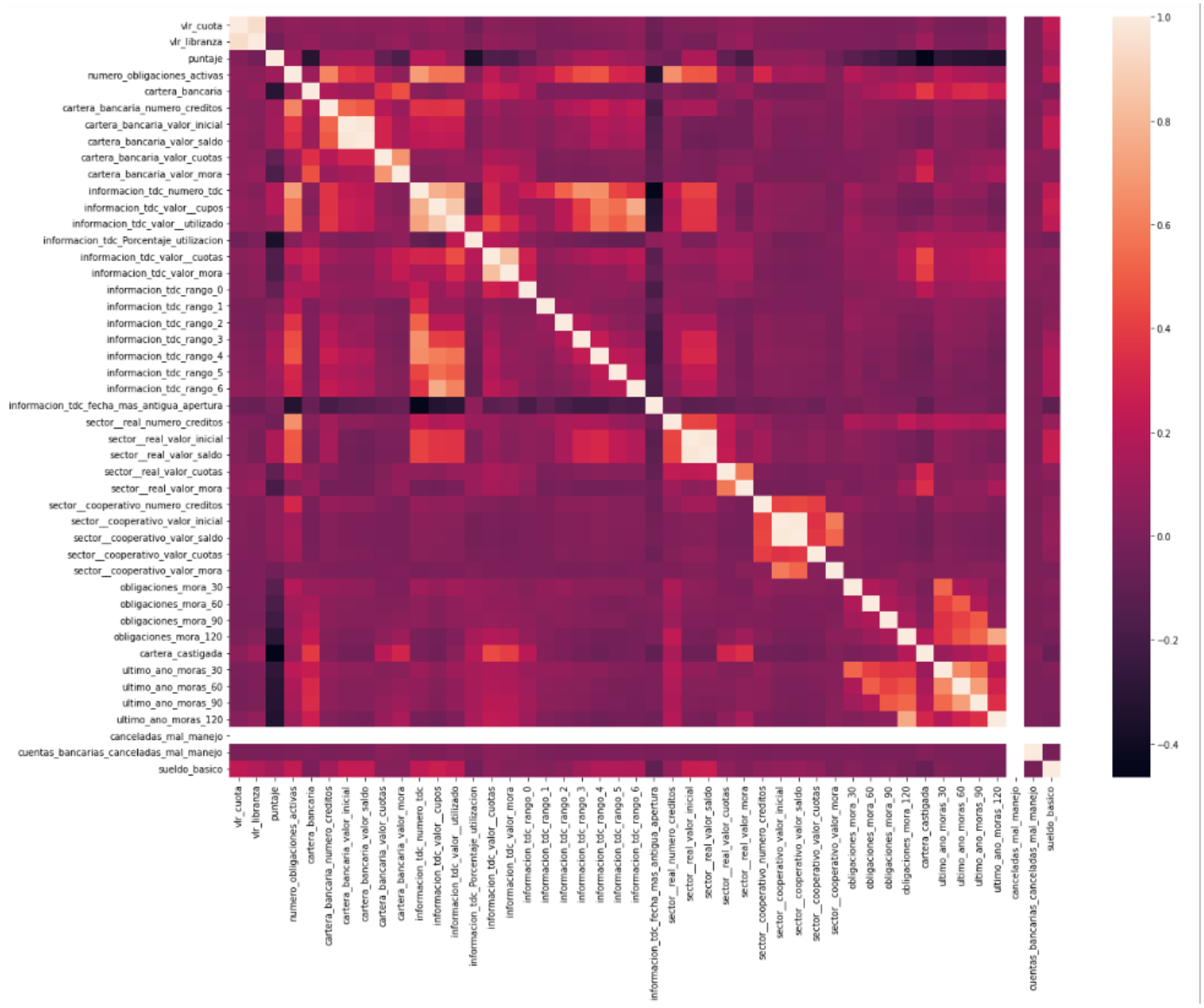
   Nuestro datos tienen 30848 filas y 54 columnas, a continuación se muestran la información de los datos:

```
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   cedula                                  30848 non-null  int64
 1   vlr_cuota                               30848 non-null  int64
 2   vlr_libranza                            30848 non-null  int64
 3   fecha_desem                             30848 non-null  object
 4   pagaduria                               30848 non-null  object
 5   puntaje                                 30766 non-null  float64
 6   fecha_ultima_actualizacion              30766 non-null  object
 7   numero_obligaciones_activas             30766 non-null  float64
 8   genero                                  26626 non-null  object
 9   rango_de_edad                           30482 non-null  object
 10  cartera_bancaria                        30766 non-null  float64
 11  cartera_bancaria_numero_creditos        30766 non-null  float64
 12  cartera_bancaria_valor_inicial          30766 non-null  float64
 13  cartera_bancaria_valor_saldo            30766 non-null  float64
 14  cartera_bancaria_valor_cuotas           30766 non-null  float64
 15  cartera_bancaria_valor_mora             30766 non-null  float64
 16  informacion_tdc_numero_tdc              30766 non-null  float64
 17  informacion_tdc_valor__cupos            30766 non-null  float64
 18  informacion_tdc_valor__utilizado        30766 non-null  float64
 19  informacion_tdc_Porcentaje_utilizacion  16975 non-null  float64
 20  informacion_tdc_valor__cuotas           30766 non-null  float64
 21  informacion_tdc_valor_mora              30766 non-null  float64
 22  informacion_tdc_rango_0                 30766 non-null  float64
 23  informacion_tdc_rango_1                 30766 non-null  float64
 24  informacion_tdc_rango_2                 30766 non-null  float64
 25  informacion_tdc_rango_3                 30766 non-null  float64
 26  informacion_tdc_rango_4                 30766 non-null  float64
 27  informacion_tdc_rango_5                 30766 non-null  float64
 28  informacion_tdc_rango_6                 30766 non-null  float64
 29  informacion_tdc_fecha_mas_antigua_apertura  18820 non-null  float64
 30  sector__real_numero_creditos            30766 non-null  float64
 31  sector__real_valor_inicial              30766 non-null  float64
 32  sector__real_valor_saldo                30766 non-null  float64
 33  sector__real_valor_cuotas               30766 non-null  float64
 34  sector__real_valor_mora                 30766 non-null  float64
 35  sector__cooperativo_numero_creditos     30766 non-null  float64
 36  sector__cooperativo_valor_inicial       30766 non-null  float64
 37  sector__cooperativo_valor_saldo         30766 non-null  float64
 38  sector__cooperativo_valor_cuotas        30766 non-null  float64
 39  sector__cooperativo_valor_mora          30766 non-null  float64
 40  obligaciones_mora_30                    30766 non-null  float64
 41  obligaciones_mora_60                    30766 non-null  float64
 42  obligaciones_mora_90                    30766 non-null  float64
 43  obligaciones_mora_120                   30766 non-null  float64
 44  cartera_castigada                       30766 non-null  float64
 45  ultimo_ano_moras_30                     30766 non-null  float64
 46  ultimo_ano_moras_60                     30766 non-null  float64
 47  ultimo_ano_moras_90                     30766 non-null  float64
 48  ultimo_ano_moras_120                    30766 non-null  float64
 49  canceladas_mal_manejo                   30766 non-null  float64
 50  peor_calificacion__trimestre_1          24107 non-null  object
 51  peor_calificacion__trimestre_2          22868 non-null  object
 52  cuentas_bancarias_canceladas_mal_manejo  30766 non-null  float64
 53  sueldo_basico                           21890 non-null  float64
dtypes: float64(44), int64(3), object(7)
```

Revisión primeras columnas:

	vlr_cuota	vlr_libranza	puntaje	numero_obligaciones_activas	cartera_bancaria	cartera_bancaria_numero_creditos
count	3.084800e+04	3.084800e+04	30766.000000	30766.000000	30766.000000	30766.000000
mean	3.120107e+05	1.426512e+07	605.719171	6.956250	0.259865	1.164467
std	2.870526e+05	1.659078e+07	213.546296	4.917678	0.941269	1.862438
min	8.690000e+03	4.269460e+05	0.000000	0.000000	0.000000	0.000000
25%	1.393080e+05	4.712679e+06	446.000000	3.000000	0.000000	0.000000
50%	2.400000e+05	1.109969e+07	696.000000	6.000000	0.000000	1.000000
75%	3.896420e+05	1.501459e+07	776.000000	10.000000	0.000000	2.000000
max	7.306652e+06	3.291371e+08	932.000000	74.000000	4.000000	62.000000

Matriz de correlación:



Pasos para datos limpios:

- Realizar la estandarización de las variables manual o vía pca sklearn ya que se ven que las variables están en diferente medidas.
- Revisión del concepto de la variable para ver si en realidad son datos perdidos o cuando es NA es porque es cero.
- En caso que de verdad sean valores perdidos procederemos a intentar buscar alguna correlación con variables para imputar o alguna metodología de imputación como medio, moda o incluso KNN.

5. **Con base en los roles definidos en el documento de la semana anterior, delinear las actividades que se llevaron y llevarán a cabo para la primera entrega calificada del proyecto. Ser preciso, y tomar este punto como un contrato entre los miembros del equipo.**

- **Alvaro Pabón:**
  - Planteamiento del problema a solucionar (Realizado)
  - Consecución de la información (Realizado)
  - Limpieza de datos (En proceso)
  - Apoyo a modelación no supervisada (Por iniciar)
- **Maria Ana Ortiz:**
  - Modelación del modelo no supervisado elegido (Por iniciar)
- **Jorge López:**
  - Creando la documentación y estructura de la entrega semana 3 (Realizado)
  - Modelación del modelo supervisado (Por iniciar)
  - Creación del reporte final para ser entregado a la universidad (Por iniciar)