Entendimiento de los datos:

Se tienen 1068 filas con 20 columas.

Tipos de datos:

Nombre columna	Tipo	Distribución
No.	Entero	Discreta
NIT	Entero	Discreta
RAZON SOCIAL	String	Discreta
SUPERVISOR	String	Discreta
REGIÓN	String	Discreta
DEPARTAMENTO DOMICILIO	String	Discreta
CIUDAD DOMICILIO	String	Discreta
CIIU	String	Discreta
MACROSECTOR	String	Discreta
INGRESOS OPERACIONALES 2018	Numero positivo	Discreta
GANANCIA (PERDIDA) 2018	Numero real	Discreta
TOTAL ACTIVOS 2018	Numero real	Discreta
TOTAL PASIVOS 2018	Numero real	Discreta
TOTAL PATRIMONIO 2018	Numero real	Discreta
INGRESOS OPERACIONALES 2017	Numero real	Discreta
GANANCIA (PERDIDA) 2017	Numero real	Discreta
TOTAL ACTIVOS 2017	Numero real	Discreta
TOTAL PASIVOS 2017	Numero real	Discreta
TOTAL PATRIMONIO 2017	Numero real	Discreta
GRUPO EN NIIF	String	Discreta

	No.	NIT	GANANCIA (PERDIDA) 2017
count	1068.000000	1.068000e+03	1.067000e+03
mean	534.500000	8.590699e+08	1.147441e+08
std	308.449348	3.782354e+07	5.226753e+09
min	1.000000	8.000003e+08	-7.977997e+10
25%	267.750000	8.300113e+08	2.899945e+06
50%	534.500000	8.600590e+08	5.023917e+07
75%	801.250000	8.910851e+08	1.915863e+08
max	1068.000000	9.010975e+08	6.620412e+10

Estos son los datos que podemos sacar al momento respecto al análisis estadistico, debido a que pandas no convierte por defecto los números exponenciales y los toma como string. Después de realizar la preparación de los datos mostraremos el análisis estadístico para esos datos. Las columnas que necesitan esta preparación son:

INGRESOS OPERACIONALES\n2018*

GANANCIA (PERDIDA) 2018

TOTAL ACTIVOS 2018

TOTAL PASIVOS 2018

TOTAL PATRIMONIO 2018

INGRESOS OPERACIONALES\n2017*

TOTAL ACTIVOS 2017

TOTAL PASIVOS 2017

TOTAL PATRIMONIO 2017

El 2% de nuestras casillas están vacías y este porcentaje pertenece a las siguientes columnas:

REGIÓN	0.000936
DEPARTAMENTO DOMICILIO	0.004682
CIUDAD DOMICILIO	0.003745
CIIU	0.000936
MACROSECTOR	0.002809
INGRESOS OPERACIONALES\n2018*	0.000936
GANANCIA (PERDIDA) 2018	0.000936
TOTAL ACTIVOS 2018	0.001873
TOTAL PASIVOS 2018	0.001873
TOTAL PATRIMONIO 2018	0.003745
INGRESOS OPERACIONALES\n2017*	0.004682
GANANCIA (PERDIDA) 2017	0.000936
TOTAL ACTIVOS 2017	0.001873

Existen 68 empresas repetidas.

Categoría para SUPERVISORES: 'SUPERSOCIEDADES', 'SUPERSALUD', 'SUPERVIGILANCIA','SUPERFINANCIERA', 'SUPERSERVICIOS', 'SUPERSUCIEDADES'. Existen errores de escritura que se tratarán en la preparación de datos.

Categoría para REGIÓN: 'Bogotá - Cundinamarca', 'Costa Atlántica', 'Costa Pacífica', 'Centro - Oriente', 'Antioquia', 'Costa Atlantica', 'Otros', 'Eje Cafetero'. Existen errores de escritura que hace que haya información duplicada, esto se tratará en la preparación de los datos.

Categoría para DEPARTAMENTO DOMICILIO: 'BOGOTA D.C.', 'MAGDALENA', 'VALLE', 'CORDOBA','NORTE DE SANTANDER', 'CUNDINAMARCA', 'ANTIOQUIA', 'ATLANTICO', 'GUAJIRA', 'CAUCA', 'SANTANDER', 'BOLIVAR', 'CASANARE', 'RISARALDA', 'BOGOTÁ D.C.', 'CALDAS', 'META', 'BOYACA', 'HUILA','NARIÑO', 'CHOCO', 'SAN ANDRES Y PROVIDENCIA', 'TOLIMA', 'CESAR','SUCRE', 'QUINDIO'. Existen errores de escritura que hace que haya información duplicada, esto se tratará en la preparación de los datos.

CIUDAD DOMICILIO tiene el departamento junto con la ciudad.

Categoría para MACROSECTOR: 'MANUFACTURA', 'COMERCIO', 'CONSTRUCCIÓN', 'SERVICIOS','MINERO-HIDROCARBUROS', nan, 'AGROPECUARIO', 'CONSTRUCCION'. Existen errores de escritura que hace que haya información duplicada, esto se tratará en la preparación de los datos.

Categoría para GRUPO EN NIIF: 'NIIF PLENAS-GRUPO 1', 'NIIF PYMES-GRUPO 2', 'REGIMEN R 414 de 2014 - CGN'.

Preparación de los datos:

Primero nos hicimos cargo de las categorías repetidas por errores en la escritura, estos cambios se aplicaron en las columnas: SUPERVISORES, REGIÓN, DEPARTAMENTO DOMICILIO, MACROSECTOR.

Cambios en SUPERVISOR: Cambiamos el typo de SUPERSUCIEDADES a SUPERSOCIEDADES.

Cambios en REGIÓN: Eliminamos el nulo de REGIÓN poniendo la región que le corresponde a su departamento registrado y luego se soluciona el typo que hay entre Costa Atlantica a Costa Atlántica.

Cambios en DEPARTAMENTO DOMICILIO: Llenamos los nulos con la información que hay en CIUDAD DOMICILO, cambiamos el typo que hay entre BOGOTA D.C. a BOGOTÁ D.C.

Cambios en CIUDAD DOMICILIO: Remplazamos todos los nulos con la ciudad que le corresponde al departamento y luego quitamos el departamento adicionado al nombre de la ciudad.

Cambios en CIIU: Eliminamos los nulos debido a que el análisis del modelo será sobre esos sectores y crearía sesgo ponerle alguna otra cosa.

Cambios en ACTIVOS/PASIVOS/PATRIMONIO: Primero convertimos la columna TOTAL PASIVOS 2017 a float quitando el caracter $. Se aplica la formula del patrimonio para hallar los valores faltantes.

Patrimonio = Activos – Pasivos

Cambios para ingresos operacionales y ganancia: Debido a que estos datos hacen una gran contribución al modelo y no hay forma de predecirlos, vamos a eliminar las filas que no tienen este valor.

Cambios para empresas repetidas: Dropeamos las empresas repetidas.

Cambiamos todas las categóricas por valores numéricos para poder hacer un análisis sobre todas las variables.

Modelamiento:

1. **K-Means (María Camila Gómez Hernández – 202011050):** Utilizamos la función log para hacer que los datos no estuvieran tan alejados y escogimos INGRESOS OPERACIONALES 2018_log', 'TOTAL ACTIVOS 2017_log', 'SUPERVISOR', 'REGIÓN', 'TOTAL PASIVOS 2018_log', 'DEPARTAMENTO DOMICILIO', 'CIUDAD DOMICILIO', 'CIIU', 'TOTAL ACTIVOS 2018_log', 'TOTAL PASIVOS 2017_log', 'TOTAL PATRIMONIO 2018_log','MACROSECTOR', 'TOTAL PATRIMONIO 2017_log' para hacer el cluster, debido a la correlación que notamos que tenían. El mejor hiperparametro K que encontramos fue k=2, analizamos el coeficiente de silueta y obtuvimos los siguientes resultados de k:

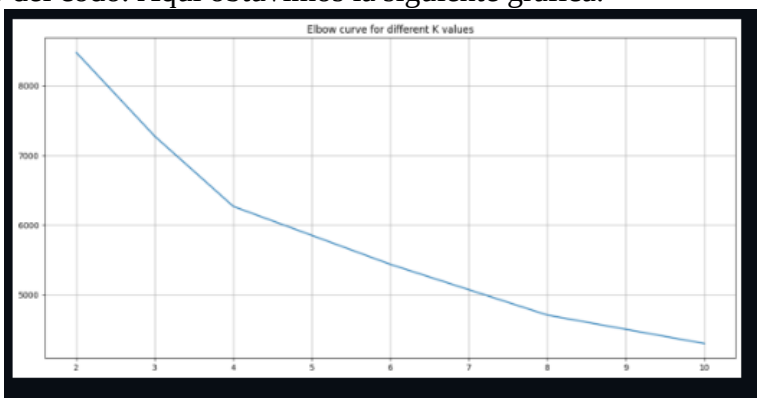
```
For n_clusters = 2 , the average silhouette score is :
0.30515077318353057
For n_clusters = 3 , the average silhouette score is :
0.26373894428529315
For n_clusters = 4 , the average silhouette score is :
0.22413929729567947
For n_clusters = 6 , the average silhouette score is :
0.1981639040309547
For n_clusters = 8 , the average silhouette score is :
0.21868041002790464
For n_clusters = 10 , the average silhouette score is :
0.18086327772588423
```

Lo cual valida cuantitativamente la hipótesis de escoger 2 clústeres. Puede encontrar la ejecución del algoritmo más el análisis con las graficas en k-means.ipnyb.

Validación:

2. **K-means:**

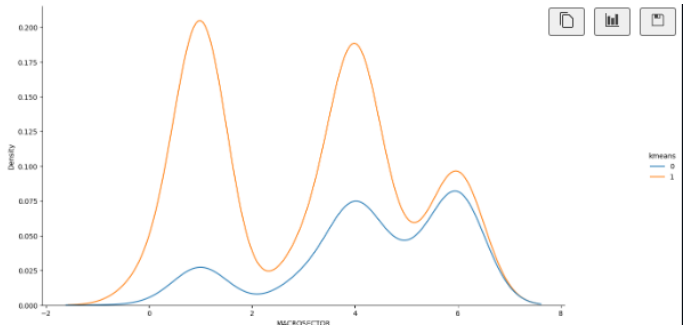
- **Validación cuantitativa:** Para la evaluación cuantitativa utilizamos como guía el método del codo. Aquí obtuvimos la siguiente grafica:



Esta grafica nos dice que se podría utilizar k=4, o k=8 o k>10 pues en esos puntos es donde la grafica se está poniendo más plana. A pesar de que el método del codo apunta que utilicemos k=4, la evaluación con el puntaje de la silueta apunta más hacia k=2, a continuación, está el average silhouette score:

```
For n_clusters = 2 , the average silhouette score is :
0.30515077318353057
For n_clusters = 3 , the average silhouette score is :
0.26373894428529315
For n_clusters = 4 , the average silhouette score is :
0.22413929729567947
For n_clusters = 6 , the average silhouette score is :
0.1981639040309547
For n_clusters = 8 , the average silhouette score is :
0.21868041002790464
For n_clusters = 10 , the average silhouette score is :
0.18086327772588423
```

- **Validación cualitativa:** En general todos los grupos que se obtuvieron están muy cercanos a los demás, lo cual podría indicar que hay que mejorar en el modelo, sin embargo, los grupos que generaron en el modelo han sido los mejores de acuerdo a una comparación con un K diferente, tal vez tenga mucho que ver también la cantidad de datos proporcionados. Debido a que el objetivo de la empresa era perfilar mejor las empresas por sectores, podemos ver que no se puede lograr con el enfoque que tienen en este momento, se puede ver en la siguiente grafica de la agrupación de los macrosectores:

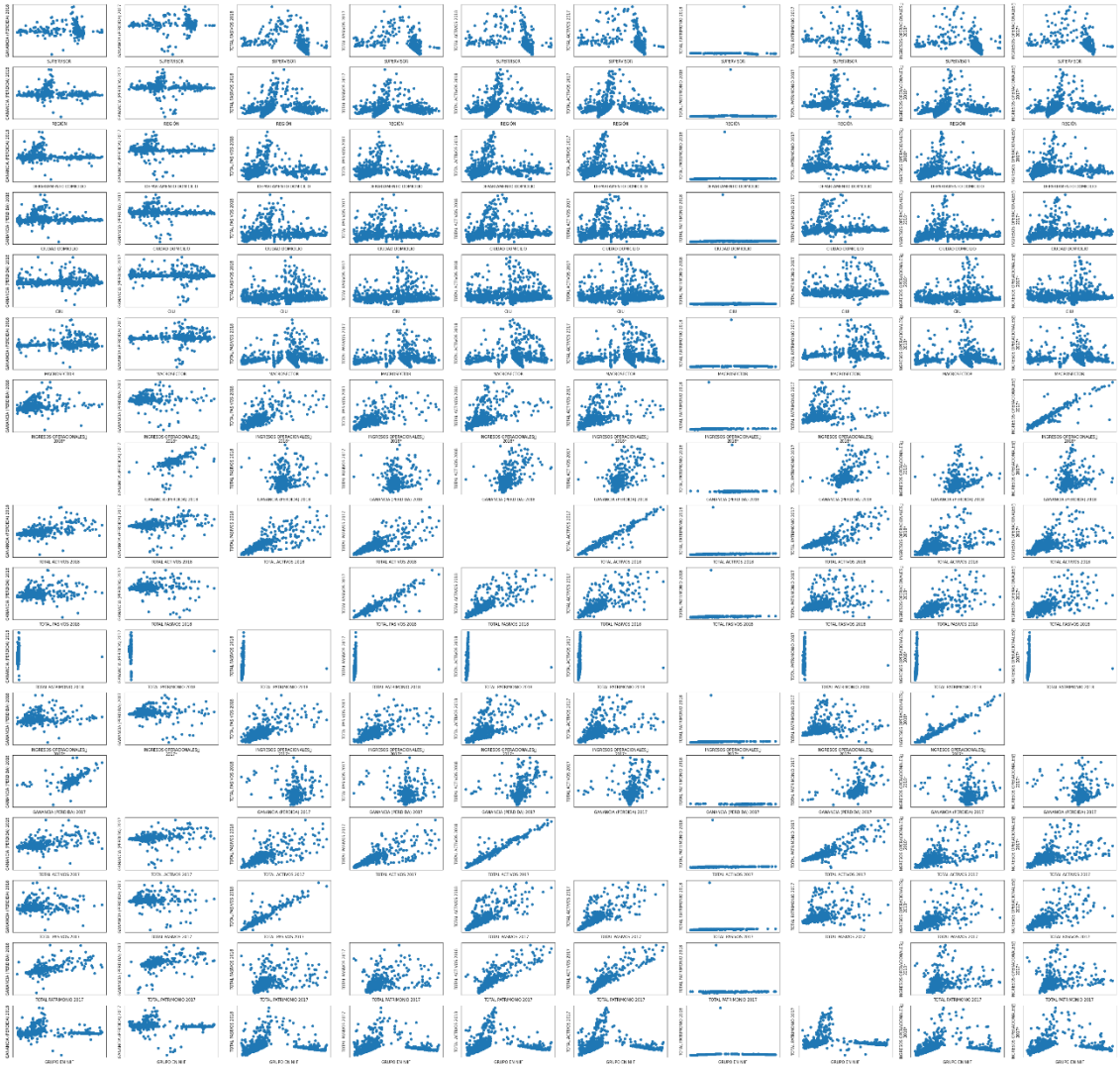


Recomendamos que se adicionen más datos al modelo y que además se agreguen más cosas que ayuden a diferenciar los sectores.

Visualización:

- **DBScan (WIlliam Felipe Mendez Ardila -202012662):** Se hizo un escalamiento y una normalización con uso de las herramientas de sklearn para obtener las

siguientes relaciones en los datos:

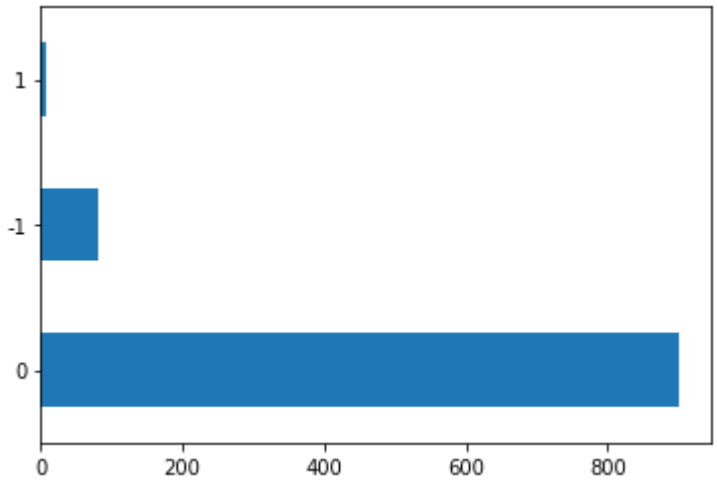


- Por esto se tomaron como columnas importantes para el analisis las siguientes: GANANCIA (PERDIDA) 2018, GANANCIA (PERDIDA) 2017, TOTAL PASIVOS 2018, TOTAL PASIVOS 2017, TOTAL ACTIVOS 2018 , TOTAL ACTIVOS 2017, TOTAL PATRIMONIO 2018, TOTAL PATRIMONIO 2017, INGRESOS OPERACIONALES\r\n2018*, INGRESOS OPERACIONALES\r\n2017*

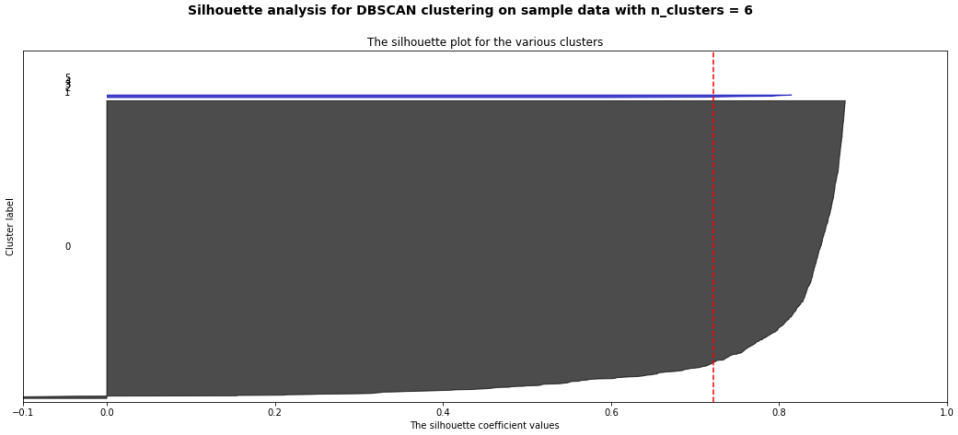
Luego se probó con 11^2 combinaciones de hiperparámetros de entrada para las variables min_samples y epsilon par aobtener los siguientes resultados:

Además, se probó con diferentes algoritmos de cálculo de distancias, pero estos no generaron diferencia en el resultado final.

Con esto se tomaron los hiperparámetros epsilon igual a 0,2 y min_samples igual a 6 por generar dos clusters diferentes al de ruido, esta combinación tiene esta distribución:

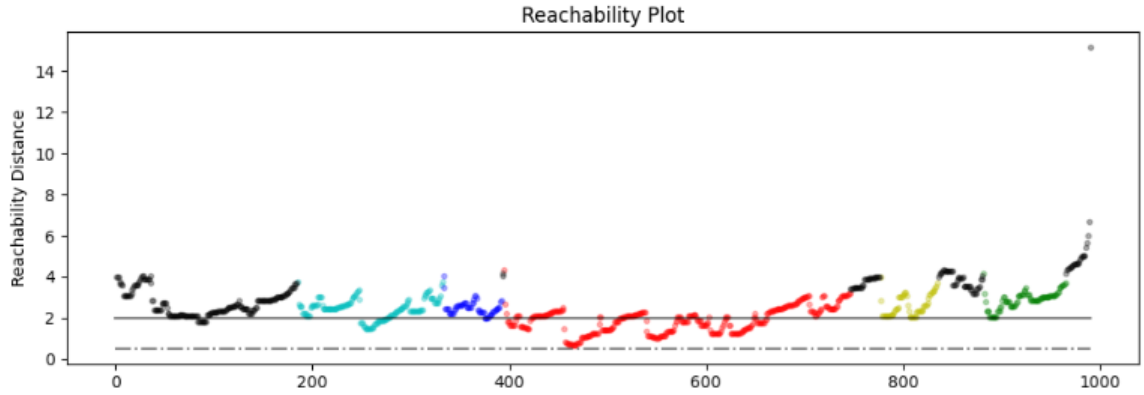


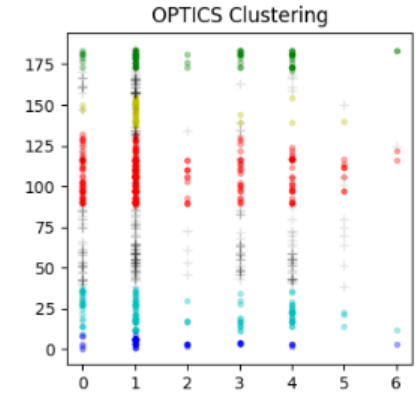
Luego, se realizó el analisis de siluetas con el modelo, pero este solo reforzó que el modelo no es útil para interpretar los datos, porque se obtuvo el siguiente resultado:

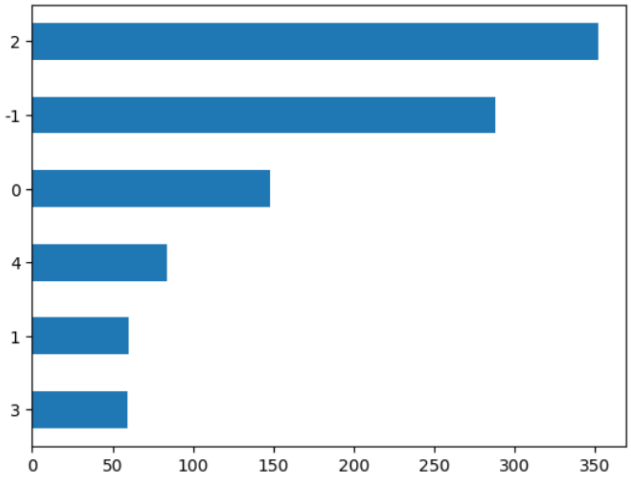


Finalmente se revisaron las curvas de distribución de los clusters con respecto a las columnas de datos, pero no se encontró un patrón significativo con los clusters generados.

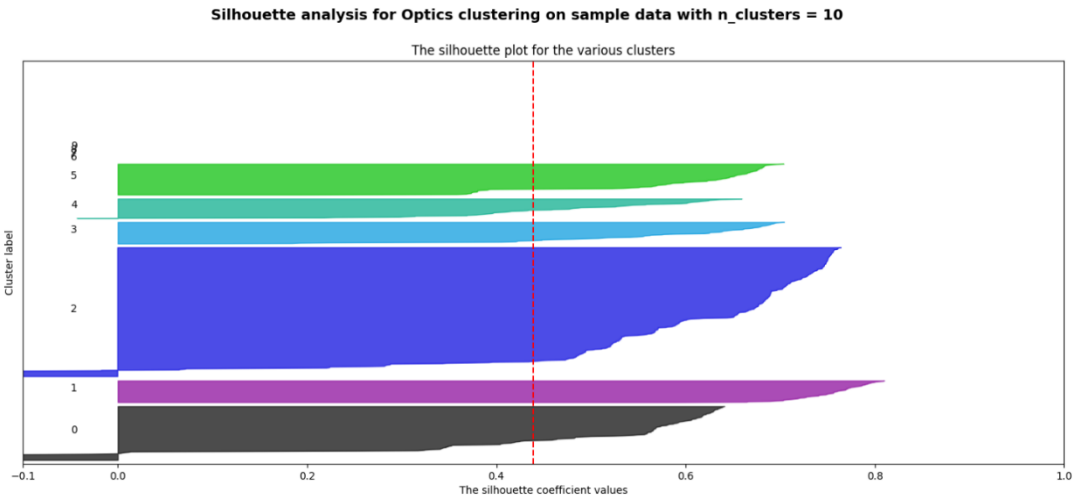
- **Optics (Juliana Andrea Galeano Caicedo – 202012128):** Se utilizó la función de log sobre los datos que no tuvieran negativos, a su vez con los datos que contenían negativos se utilizó una estandarización seguida de una normalización, todo esto con el fin de hacer todos los valores comparables entre sí. Su utilizaron los siguientes features: Total activos 2017, Total activos 2018, Ingresos operacionales 2018, Total pasivos 2018, Total patrimonio 2018, Ganancia 2017, Ganancia 2018, Región y CIIU. Utilizando los hiperparamentos min_samples = 10, xi = 0.05, min_cluster_size = 0.05.
  Se obtuvo la siguiente distribución de clústeres:



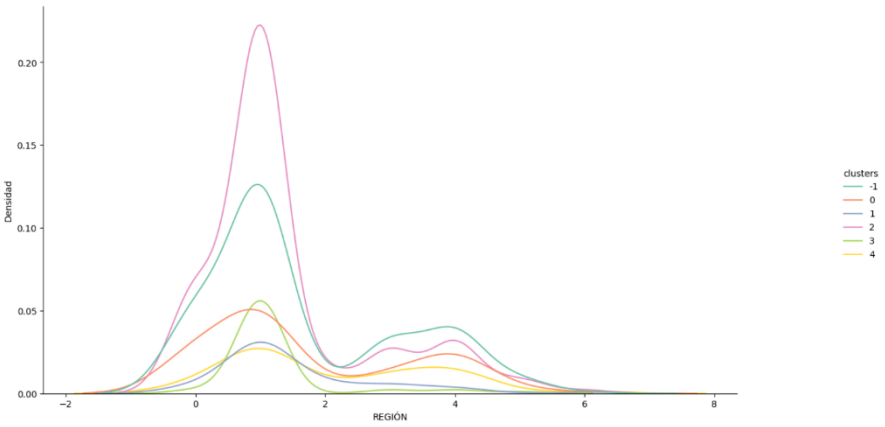


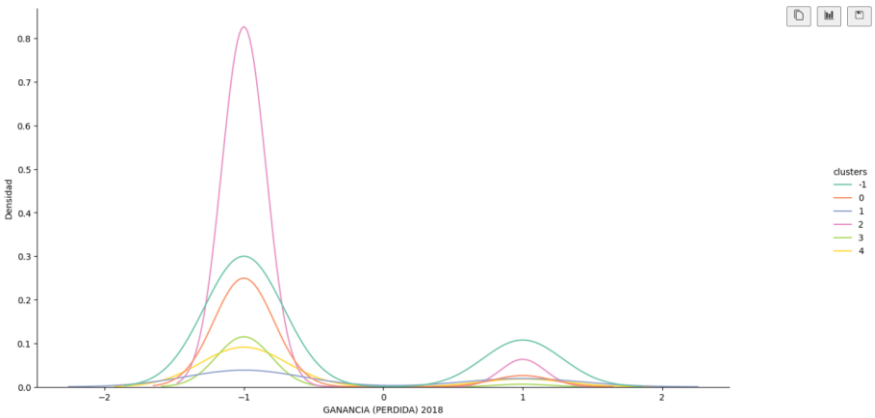


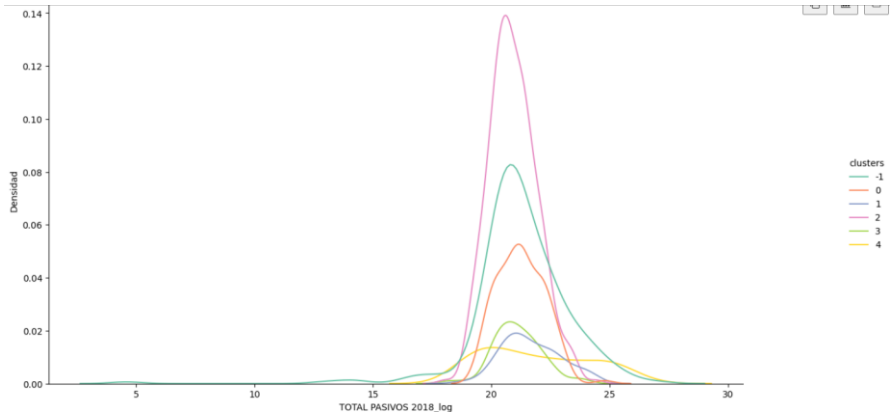
Se realizó el análisis de siluetas con el modelo y se obtuvo el siguiente resultado:



Aunque los resultados son buenos porque se obtiene un promedio de silueta de 0.4391276972364632, con el análisis de curvas de distribución podemos observar que el modelo no brinda datos importantes para el negocio.

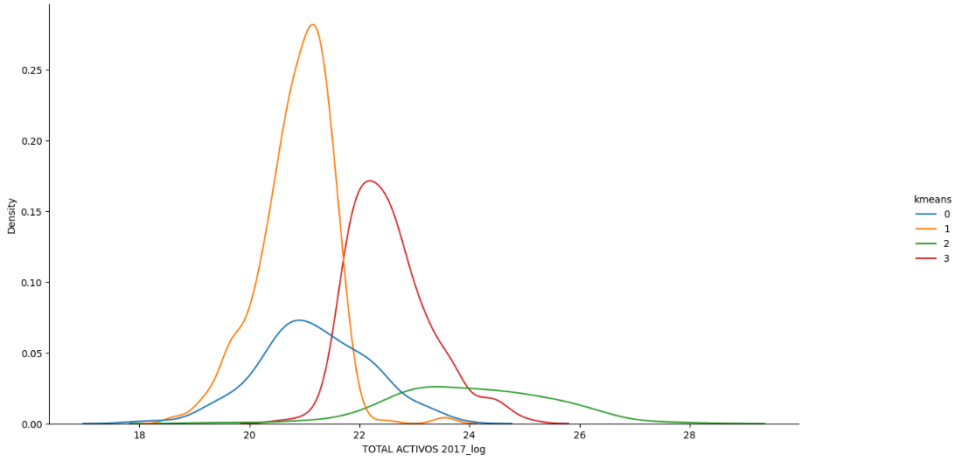


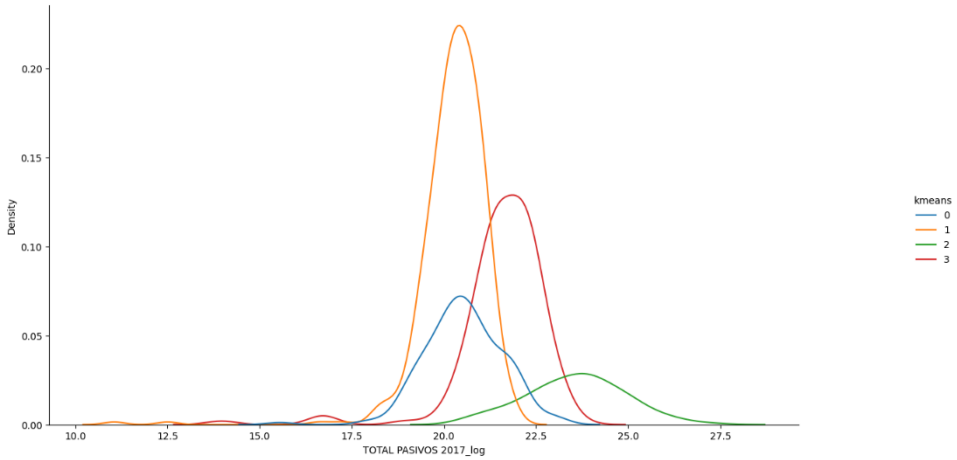


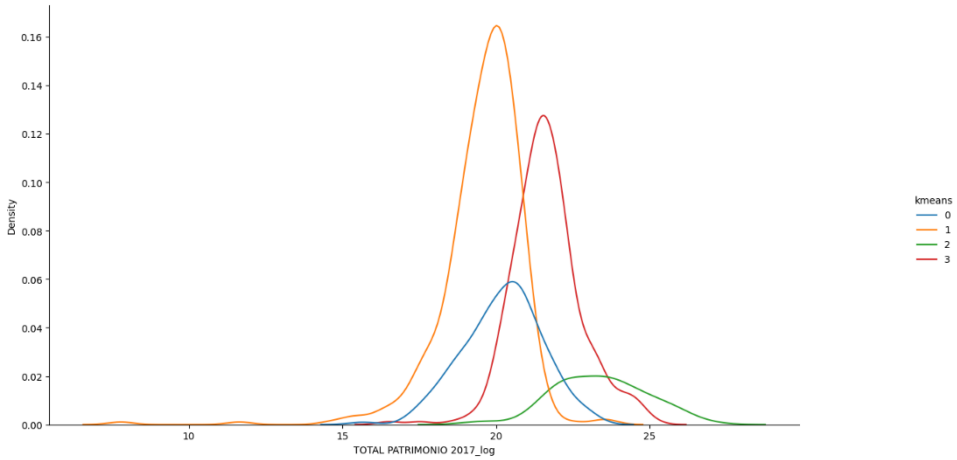


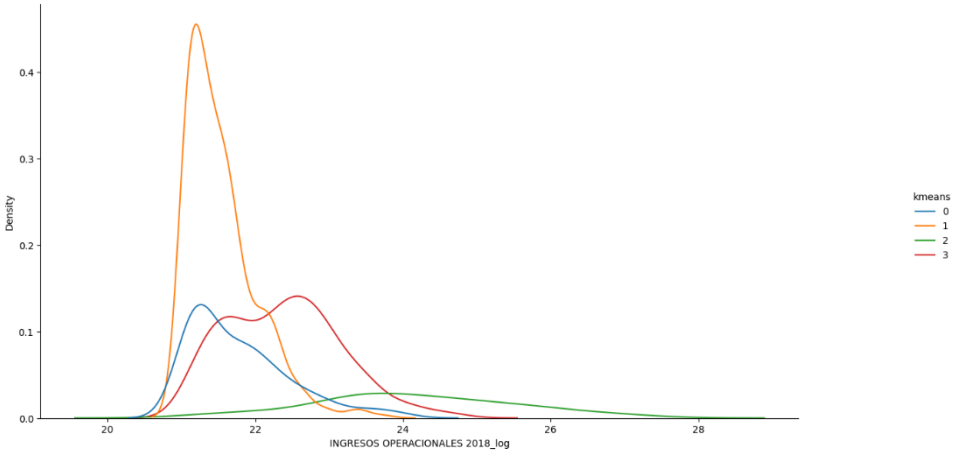
Análisis de resultados:

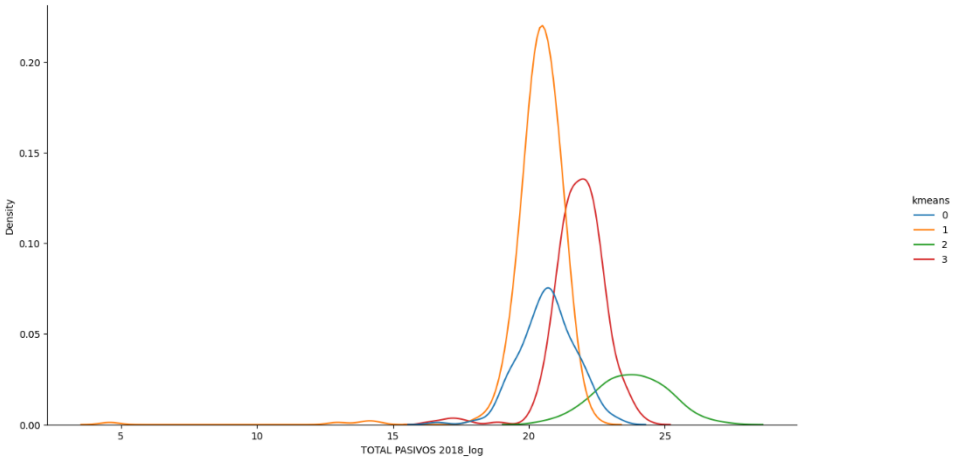
Haciendo uso del modelo de K-Means obtenemos un cluster que muestra que una gran cantidad de las empresas tienen las características: Ser de Bogotá, tener activos, pasivos, patrimonio e ingresos menores a el resto de las empresas y pertenecer al sector comercio. Estas características se unen en el cluster 1 como se muestra en las imágenes a continuación y con estas podemos decir que Consultalpes podría centrarse en ofrecer sus servicios a este perfil de empresas para mejorar su situación. Además, se puede observar que el cluster 3 (que pertenece a unas características de ser menos de Bogotá con respecto al 1 y estar más centrado en los sectores de manufactura y servicios) tiene un mejor desempeño económico que el 1, por lo que este perfil puede tener menos prioridad para Consultalpes a la hora de ofreces sus servicios.

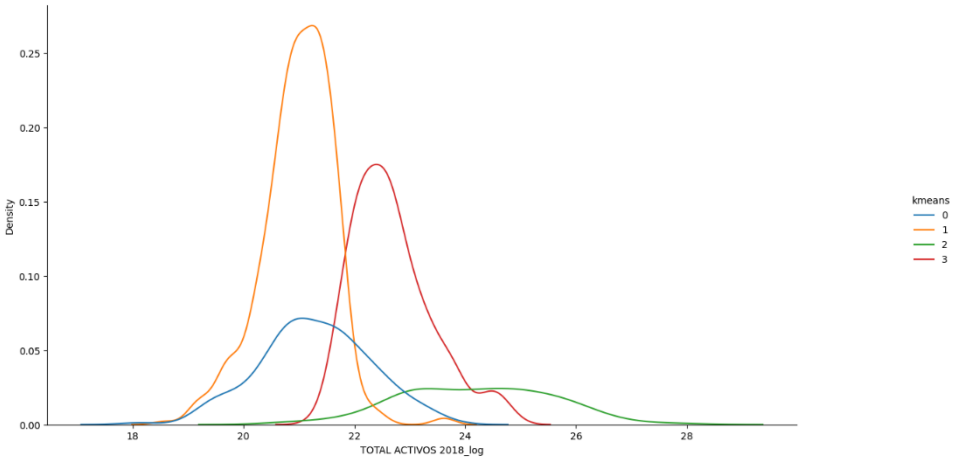

Density
0.25
0.20
0.15
0.10
0.05
0.00
18
20
22
24
26
28
TOTAL ACTIVOS 2017_log
kmeans
0
1
2
3

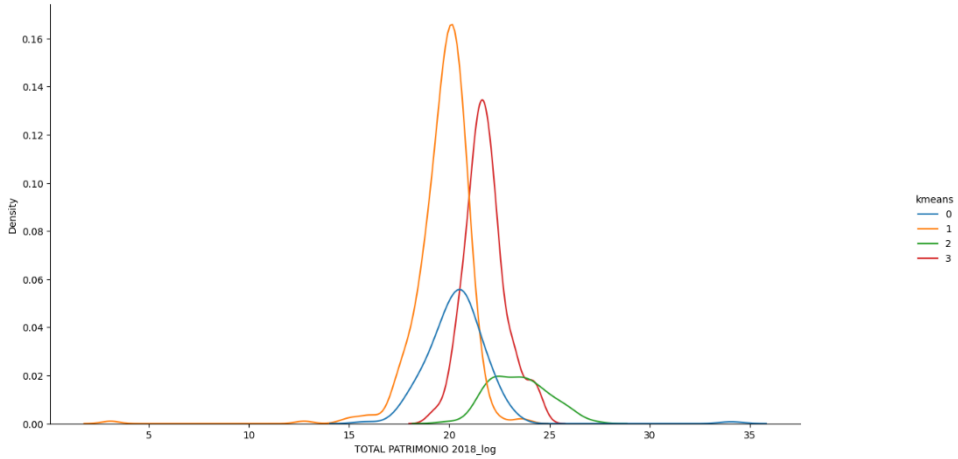

Density
0.20
0.15
0.10
0.05
0.00
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
TOTAL PASIVOS 2017_log
kmeans
0
1
2
3

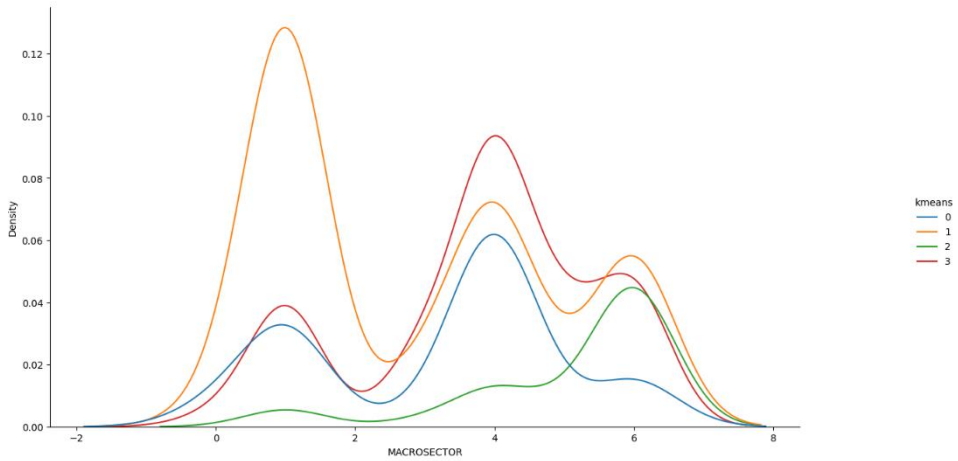

Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10
15
20
25
TOTAL PATRIMONIO 2017_log
kmeans
0
1
2
3

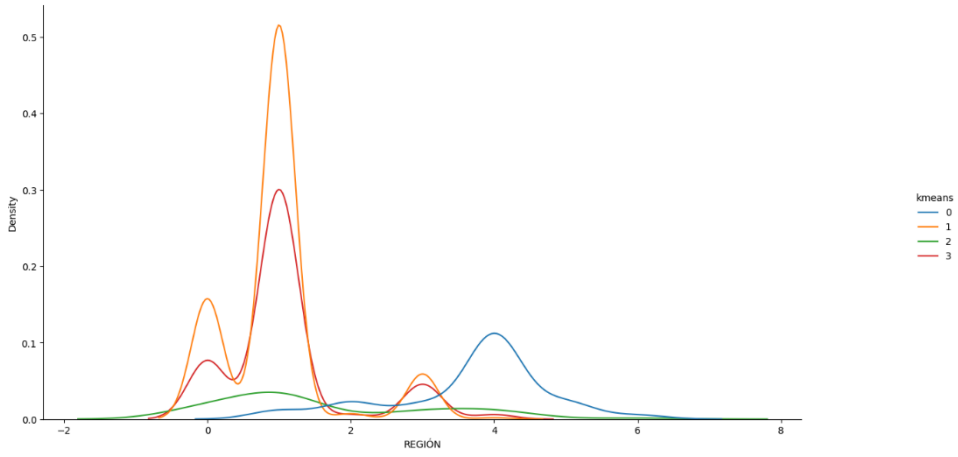

Density
0.0
0.1
0.2
0.3
0.4
20
22
24
26
28
INGRESOS OPERACIONALES 2018_log
kmeans
0
1
2
3


Density
0.00
0.05
0.10
0.15
0.20
5
10
15
20
25
TOTAL PASIVOS 2018_log
kmeans
0
1
2
3


Density
0.00
0.05
0.10
0.15
0.20
0.25
18
20
22
24
26
28
TOTAL ACTIVOS 2018_log
kmeans
0
1
2
3


Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
5
10
15
20
25
30
35
TOTAL PATRIMONIO 2018_log
kmeans
0
1
2
3


Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−2
0
2
4
6
8
MACROSECTOR
kmeans
0
1
2
3


Density
0.5
0.4
0.3
0.2
0.1
0.0
−2
0
2
4
6
8
REGIÓN
kmeans
0
1
2
3