Laboratorio 2 - BancAlpes

Integrantes:
Sebastian Beltran – K-means
Angelo Valero - Clustering Jerárquico
David Dominguez - GMMs

Perfilamiento de los datos

Al terminar de cargar y explorar los datos verificamos los valores nulos. Luego, se necesitaba asignar a cada dato su tipo correspondiente, se limpiaron los datos de caracteres (‘-’,’?’,’ABC’) que no correspondían a la columna. En determinados casos decidimos eliminar la fila o consérvala, justificando su razón en el noteBook de JupyterLab. Corregimos algunas entradas en la columna de sexo (‘M’,’F’,’Mael’,’f’,’Femael’), de igual forma en la columna de casado (‘1’,’2’,’?’,’0’). Finalmente, los valores de las columnas de sexo y casado fueron convertidas en números para darle un mejor rendimiento al algoritmo de clustering.

El archivo contaba con 660 filas y 11 columnas.

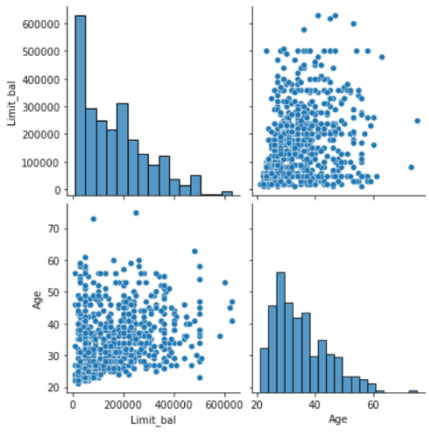
Estadística descriptiva luego de limpiar y asignar debidamente el tipo de cada columna

	Id	Customer	Limit_bal	Sex	Education	Marriage	Age	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	656.000000	656.000000	656.000000	656.000000	656.000000	656.000000	656.000000	6.560000e+02	656.000000	656.000000	656.000000
mean	332.315549	55230.647866	170182.926829	0.425305	1.783537	0.451220	95.493902	1.882010e+05	2.413110	2.597561	3.559451
std	189.813037	25657.650223	129855.927499	0.494766	0.778786	0.550418	1435.931970	4.820178e+06	1.630364	2.941090	2.864812
min	1.000000	11265.000000	10000.000000	0.000000	1.000000	0.000000	21.000000	1.000000e+00	0.000000	0.000000	0.000000
25%	168.750000	33952.500000	50000.000000	0.000000	1.000000	0.000000	28.000000	3.000000e+00	1.000000	1.000000	1.000000
50%	332.500000	53907.000000	140000.000000	0.000000	2.000000	0.000000	33.500000	5.000000e+00	2.000000	2.000000	3.000000
75%	496.250000	77439.000000	240000.000000	1.000000	2.000000	1.000000	41.000000	6.000000e+00	4.000000	4.000000	5.000000
max	660.000000	99843.000000	630000.000000	1.000000	6.000000	2.000000	36745.000000	1.234568e+08	5.000000	15.000000	10.000000

Preparación de datos

Descartamos las columnas de ID y Customer, pues no eran determinantes para el análisis.
Encontramos datos atípicos, como por ejemplos edades mayores 130 años y un número de tarjetas de créditos mayores a 10. Tambien para tener una mejor segmentación en la columna de educacion los valores de 5(unknow) y 6(unknow) se manejaron como 4(others).

Finalmente se graficaron las variables y se busco al grupo natural entre variables, encontrando que la edad y el monto de crédito era un grupo natural de datos.



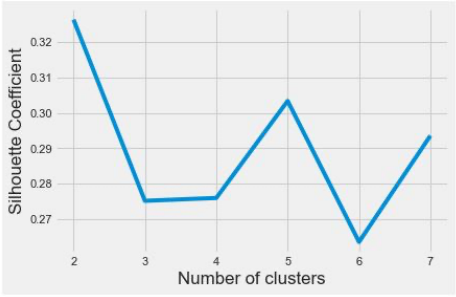
Modelamiento

K-means

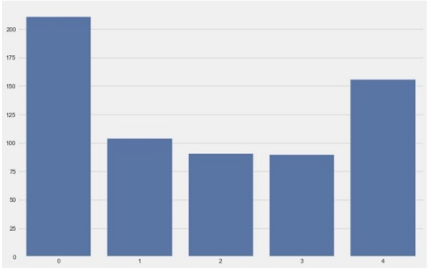
Lo primero que se hace es normalizar los datos.

	Limit_bal	Sex	Education	Marriage	Age	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000
mean	0.258955	0.423313	0.257669	0.225460	0.260850	0.415133	0.484356	0.172699	0.354141
std	0.209257	0.494463	0.245250	0.273952	0.173326	0.240131	0.325941	0.196464	0.285972
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.076613	0.000000	0.000000	0.000000	0.129630	0.222222	0.200000	0.066667	0.100000
50%	0.209677	0.000000	0.333333	0.000000	0.222222	0.444444	0.400000	0.133333	0.300000
75%	0.370968	1.000000	0.333333	0.500000	0.370370	0.555556	0.800000	0.266667	0.500000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Con el siguiente grafico de terminamos los mejores números de clusters y elegimos N = 5



Los clusters quedan de la siguiente forma

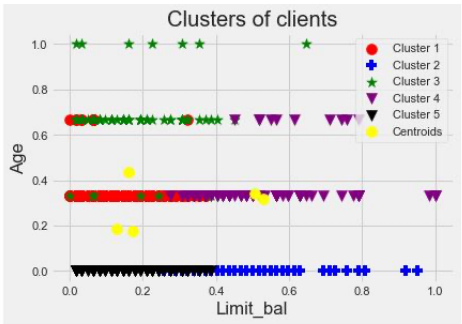


Cluster	Limit_bal	Sex	Education	Marriage	Age	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	211	211	211	211	211	211	211	211	211
1	104	104	104	104	104	104	104	104	104
2	91	91	91	91	91	91	91	91	91
3	90	90	90	90	90	90	90	90	90
4	156	156	156	156	156	156	156	156	156

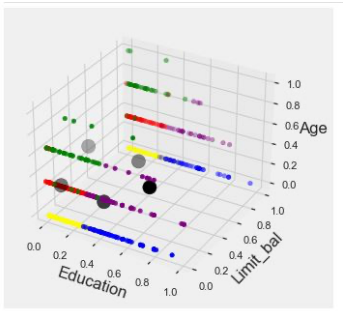
Elegimos las columnas de educacion, limit_bal y edad para graficar los clusters.

```
name_cols_number = ['Education','Limit_bal','Age']
```

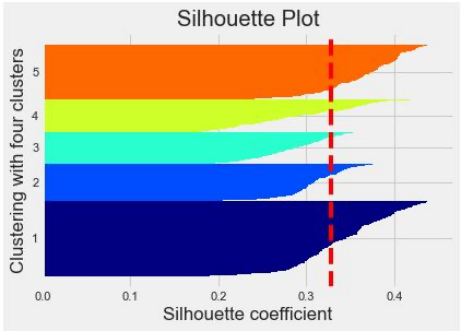
Hacemos un grafico 2d Edad vs Limit_bal



Hacemos un grafico 3d de los clusters según la edad, educacion y limit_bal.
Aquí podemos observar que los clusters si tienden a estar agrupados por limit_bal y edad pero no por educacion.



Realizamos un gráfico de la silueta para saber el coeficiente de cada cluster y encontramos que la media de eficiencia es menor a 0.35, mientras que los cluster 1 y 5 pasan del 0.4



Clustering Jerárquico

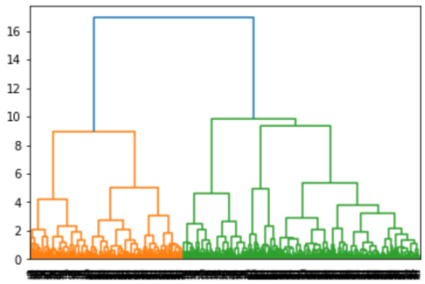
Primero normalizamos los datos.

	Limit_bal	Sex	Education	Marriage	Age	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	0.016129	0.0	0.333333	0.5	0.055556	0.111111	0.2	0.066667	0.0
1	0.177419	0.0	0.333333	0.0	0.092593	0.222222	0.0	0.666667	0.9
2	0.129032	0.0	0.333333	0.0	0.240741	0.666667	0.2	0.200000	0.4
3	0.064516	0.0	0.333333	0.5	0.296296	0.444444	0.2	0.066667	0.1
4	0.064516	1.0	0.333333	0.5	0.666667	0.555556	0.0	0.800000	0.3
...	...	...	...	...	...	...	...	...	...
655	0.435484	1.0	0.333333	0.5	0.333333	1.000000	0.2	0.666667	0.0
656	0.451613	0.0	0.666667	0.0	0.111111	1.000000	0.2	0.866667	0.2
657	0.290323	0.0	0.000000	0.0	0.129630	0.777778	0.2	0.600000	0.1
658	0.112903	0.0	0.000000	0.0	0.037037	1.000000	0.2	1.000000	0.0
659	0.435484	0.0	0.333333	0.0	0.129630	0.888889	0.0	0.800000	0.2

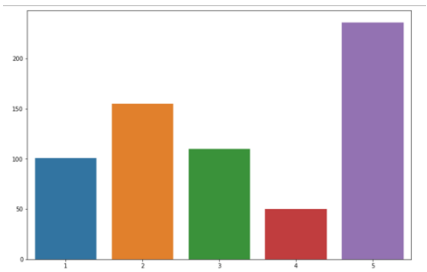
Para segmentar adecuadamente los grupos realizamos una prueba del método silhouette para identificar el número óptimo de clusters.

Podemos observar que el método nos indica que el número de clusters apropiado para segmentar es de 5.

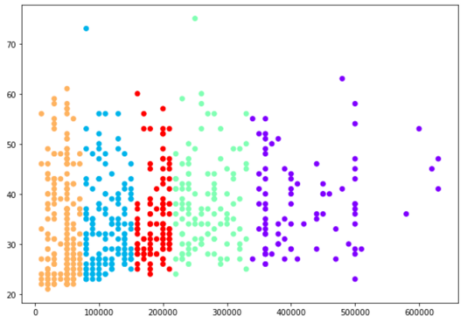
Creamos el Dendrograma, el cual es el árbol mediante el cual se presenta el clustering jerarquico.



Ahora con el dendrograma creamos los clusters y realizamos el corte en 8 para dividir el conjunto de datos en 5 grupos



Ahora utilizaremos agglomerative clustering con nuestros datos ya diferenciados para encontrar una mejor segmentación.

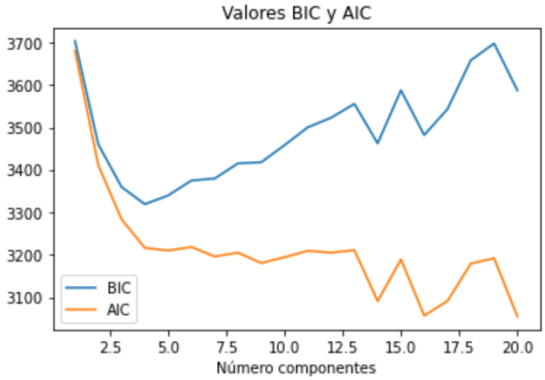


Se puede ver una clara segmentación de los grupos, determinado por el monto del crédito en dólares.

Gaussian mixture models (GMMs)

Dado que es posible ver un grupo natural cuando observamos **Age** contra **Limit_bal**. Consideramos que para lograr una buena segmentación con este algoritmo, se deben usar estas dos variables.

Debido a que este modelo es probabilístico, recurrimos a métricas como la Akaike infromation criterion (AIC) y Bayesian information criterion (BIC), para identificar el numero apropiado de clusters y que tan bien se ajustan los datos al modelo.



Encontramos que el numero óptimo de clusters según la métrica BIC es de 4.

Tambien encontramos una metrica correcta con el metodo de SILHOUETTE

```
print("SILHOUETTE: ", silhouette_score(X, clusters))
```

```
SILHOUETTE:  0.2914786074901753
```

Aquí podemos observar las estadísticas de la covarianza y la media del modelo.

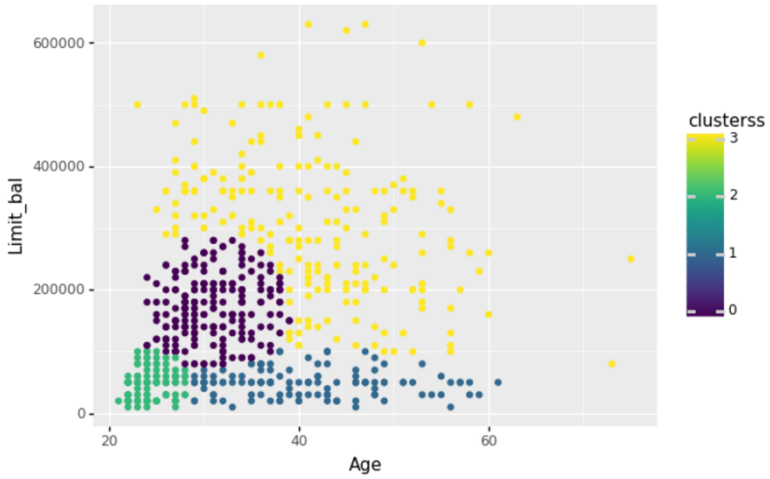
```
# Matriz de covarianza de cada componente
modelo_gmm.covariances_
```

```
array([[[ 0.03798946,  0.01223033],
        [ 0.01223033,  0.04721895]],

       [[ 0.20386799,  0.01185037],
        [ 0.01185037,  0.33240893]],

       [[ 0.91260013, -0.00430241],
        [-0.00430241,  0.02927656]],

       [[ 0.95595807, -0.26156232],
        [-0.26156232,  0.9899343 ]]])
```

```
# Media de cada componente
modelo_gmm.means_
```

```
array([[-1.1338169 , -0.90401434],
       [-0.37416534,  0.02167358],
       [ 0.57176215, -0.93064099],
       [ 0.71925287,  0.96291883]])
```

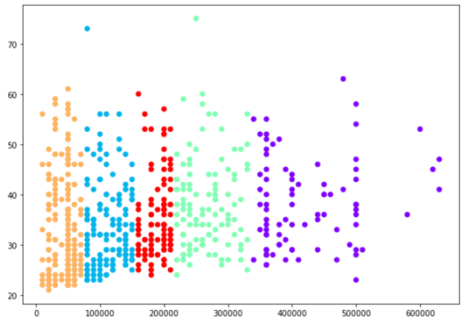
Finalmente podemos ver los grupos segmentados por el modelo, los cuales pueden ser visibles mediante la gráfica del monto de crédito y la edad.



conclusiones

Encontramos y aconsejamos a la organización el uso del modelo de Clustering Jerárquico, debido a que este modelo permite tener una clara segmentación de los clientes. Es importante aclarar que, aunque el algoritmo GMMs también tiene una representación gráfica apropiada de la segmentación, este fue construido solo con las variables de monto de crédito y edad, caso contrario con el modelo jerárquico que incluyo todas las variables para su análisis y fue determinado por todo el conjunto de datos.

Recomendaos al área de marketing segmentar todos sus clientes en 5 grupos, los cuales están determinados por los montos indicados en la gráfica.



Finalmente, no descartamos que en las campañas de marketing de la entidad sea necesario tener en cuenta la edad, pues consideramos que esta variable es importante para la segmentación de los grupos. Esperamos que la organización nos permita incluir muchos más datos de sus clientes, pues es necesario tener una cantidad mayor de datos para una mejor segmentación, por ejemplo, el número de uso de las tarjetas de crédito, el número de su núcleo familiar, el número de transacciones por fechas y su historial crediticio.