

Obligatorio Machine Learning - CLUSTERING

Descripción del problema:

Un banco desea generar agrupaciones de sus clientes en base a los datos del dataset mencionado en el apartado siguiente de manera de generar políticas de atención al cliente diferenciales para cada tipo de cliente.

Crear el mejor modelo de clustering posible para cumplir con los objetivos definidos anteriormente, utilizando el dataset que se encuentra en el archivo "BankChurners.csv".

Referencias de variables:

- **CLIENTNUM:** Número de cliente → variable cualitativa.
- **Attrition_Flag:** Status de la cuenta al mes siguiente → variable cualitativa.
- **Customer_Age:** Edad del cliente → variable cuantitativa.
- **Gender:** Género del cliente → variable cualitativa.
- **Dependent_count:** Número de personas a cargo → variable cuantitativa.
- **Education_Level:** Nivel educativo → variable cualitativa.
- **Marital_Status:** Estado civil → variable cualitativa.
- **Income_Category:** Categoría de ingresos del cliente → variable cualitativa.
- **Card_Category:** Tipo de tarjeta → variable cualitativa.
- **Months_on_book:** Antigüedad de la cuenta → variable cuantitativa.
- **Total_Relationship_Count:** Cantidad de productos del cliente (cuentas y tarjetas) → variable cuantitativa.
- **Months_Inactive_12_mon:** N° de meses inactivo en los últimos 12 meses → variable cuantitativa.
- **Contacts_Count_12_mon:** N° de contactos en los últimos 12 meses (consultas/reclamos al banco) → variable cuantitativa.
- **Credit_Limit:** Límite de crédito → variable cuantitativa.
- **Total_Revolving_Bal:** Saldo no cubierto de la tarjeta (sería lo que el cliente lleva usado del monto en su tarjeta, es la diferencia entre Credit_Limit y Avg_Open_To_Buy) → variable cuantitativa.
- **Avg_Open_To_Buy:** Disponible de la tarjeta → variable cuantitativa.
- **Total_Amt_Chng_Q4_Q1:** Cambio porcentual de monto de consumos → variable cuantitativa.
- **Total_Trans_Amt:** Monto de consumos en los últimos 12 meses → variable cuantitativa.
- **Total_Trans_Ct:** Cantidad de transacciones en los últimos 12 meses → variable cuantitativa.
- **Total_Ct_Chng_Q4_Q1:** Cambio porcentual de cantidad de consumos → variable cuantitativa.
- **Avg_Utilization_Ratio:** Ratio de utilización de la tarjeta (es el resultado de hacer Total_Revolving_Bal dividido Credit_Limit) → variable cuantitativa.

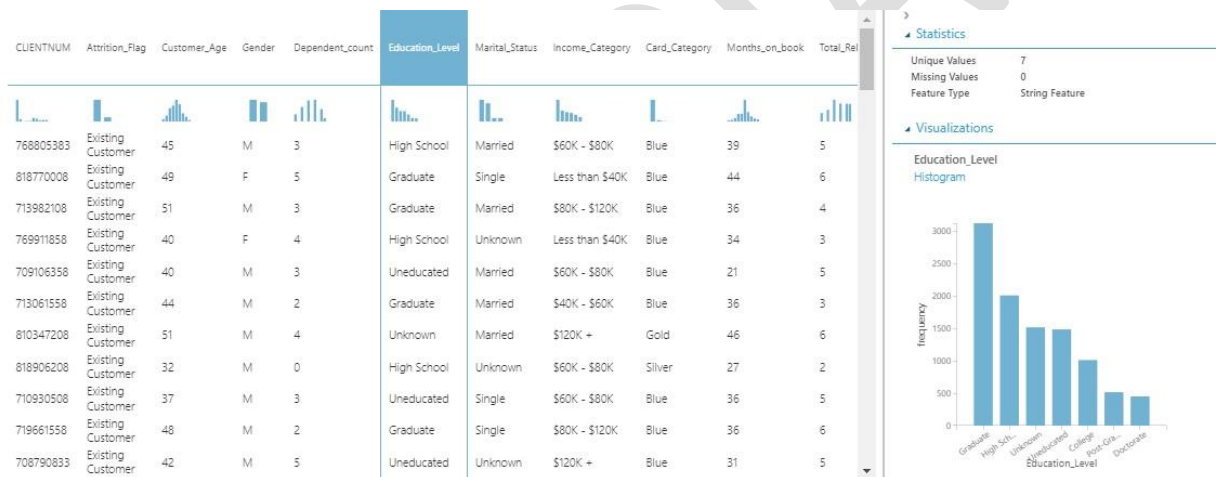
Experimento en Azure:

<https://gallery.cortanaintelligence.com/Experiment/Obligatorio-Clustering-Correa-Lopez-Mosco>

Variables categóricas:

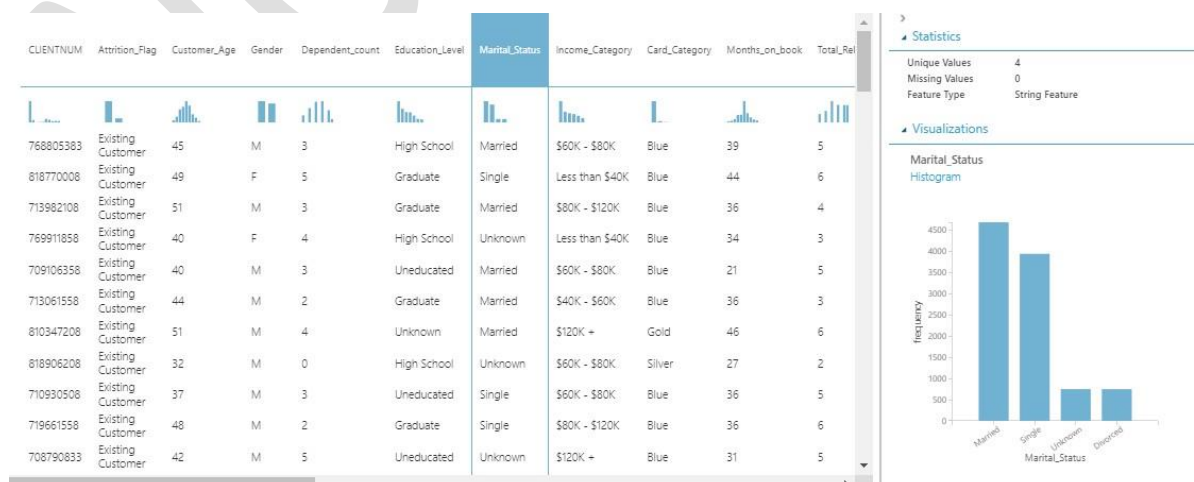
Las variables categóricas son:

- **Attrition_Flag:** Es la variable objetivo.
- **Gender:** Indica las categorías de género de una persona.
- **Education_Level:** Indica las categorías del nivel de educación de una persona.
- **Marital_Status:** Indica las categorías maritales de una persona.
- **Income_Category:** Indica las categorías de ingresos de una persona.
- **Card_Category:** Indica las categorías de tipo de tarjetas que puede tener una persona.



Tenemos 7 niveles educativos.

Observamos que los clientes con mayor ocurrencia son graduados.



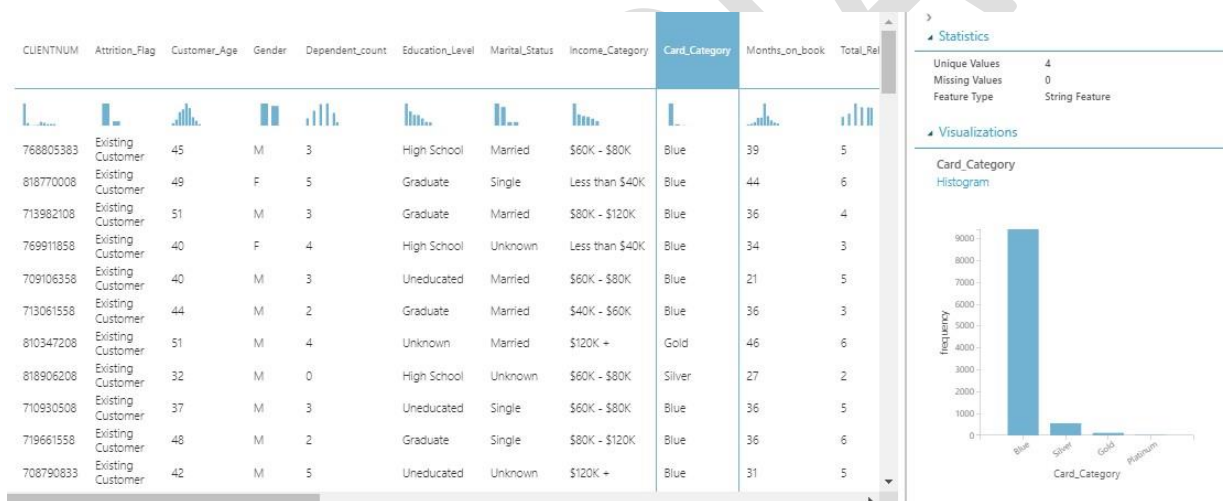
Tenemos 4 estados civiles.

Observamos que la mayoría de los clientes están casados.



Tenemos 6 categorías de ingresos.

Observamos que la mayoría de los clientes ganan menos de 40.000 dólares al año.



Tenemos 4 categorías de tarjetas.

Observamos que la mayoría de los clientes tienen una tarjeta blue.

A continuación detallamos el script en donde convertiremos las variables categóricas a variables dummies:

```
select *,
  case when Attrition_Flag = 'Existing Customer' then 1 else 0 end as AF_Existing,
  case when Attrition_Flag = 'Attrited Customer' then 1 else 0 end as AF_Attrited,
  case when Education_Level = 'Graduate' or 'High School' or 'Unknown' or 'Uneducated' then 1 else 0 end as EL_Graduate_HighSchool_Unk_Uned,
  case when Gender = 'M' then 1 else 0 end as G_Male,
  case when Gender = 'F' then 1 else 0 end as G_Female,
  case when Marital_Status = 'Married' or 'Single' then 1 else 0 end as MS_Married_Single,
  case when Income_Category = 'Less than $40K' then 1 else 0 end as IC_less40,
  case when Card_Category = 'Blue' then 1 else 0 end as CC_Blue
from t1;
```

Agrupamos las variables “Income_Category” y “Card_Category” según la categoría de mayor presencia en cada una de ellas, hacemos esto porque las mismas tienen más de 2 categorías. En el caso de “Education_Level” y “Marital_Status” agrupamos según la cantidad de observaciones.

Selección de variables:

Los atributos que vamos a excluir son:





- **CLIENTNUM:** no nos van a dar ningún tipo de información a la hora de correr nuestro modelo porque son identificadores únicos del cliente y de la compra y no los vamos a necesitar para entrenar a nuestro modelo
- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1:** No tenemos información sobre esta variable.
- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2:** No tenemos información sobre esta variable.

Incluimos las variables numéricas.

Seleccionamos n-1 variables dummies de cada categoría y dejamos fuera las variables categóricas.

Creación de modelos:

Creamos modelos con diferentes cantidades de clusters e inicializaciones para poder hacer una comparación entre ellos.

	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
					
Modelo 1 →	Combined Evaluation	3.964208	4.776239	10127	17.082005
	Evaluation For Cluster No.0	3.9167	4.577475	3963	9.062044
	Evaluation For Cluster No.1	4.451164	5.68277	1785	12.766739
	Evaluation For Cluster No.2	3.808706	4.586593	4379	17.082005
	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
					
Modelo 2 →	Combined Evaluation	3.823517	4.683694	10127	16.857249
	Evaluation For Cluster No.0	3.652997	4.455474	3541	13.631124
	Evaluation For Cluster No.1	3.777042	4.429969	3868	16.857249
	Evaluation For Cluster No.2	3.781661	4.928076	1409	6.763225
	Evaluation For Cluster No.3	4.467177	5.787748	1309	8.864923

Modelo 3

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.755299	4.552568	10127	16.923625
Evaluation For Cluster No.0	3.82054	4.566546	2464	16.923625
Evaluation For Cluster No.1	3.73955	4.882459	1376	9.264381
Evaluation For Cluster No.2	3.720525	4.278655	2379	11.520104
Evaluation For Cluster No.3	3.645633	4.294681	3225	13.684122
Evaluation For Cluster No.4	4.190605	6.00931	683	8.971392

Modelo 4

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.664318	4.487967	10127	16.58289
Evaluation For Cluster No.0	4.331757	5.639923	987	8.616508
Evaluation For Cluster No.1	3.531336	4.135216	2621	9.138488
Evaluation For Cluster No.2	3.807243	4.404777	1524	16.58289
Evaluation For Cluster No.3	3.433325	4.165296	2811	8.142655
Evaluation For Cluster No.4	3.730986	4.924397	855	6.122369
Evaluation For Cluster No.5	3.712688	4.825242	1329	6.505187

Modelo 5

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.594814	4.389162	10127	16.683463
Evaluation For Cluster No.0	4.309915	5.586348	969	8.663584
Evaluation For Cluster No.1	3.623942	4.277714	1344	9.012923
Evaluation For Cluster No.2	3.585866	4.17148	2111	16.683463
Evaluation For Cluster No.3	3.283491	3.897258	1339	6.07147
Evaluation For Cluster No.4	3.672448	4.697677	1281	6.505907
Evaluation For Cluster No.5	3.363932	4.063562	2221	10.97827
Evaluation For Cluster No.6	3.730554	4.894791	862	6.101905

Modelo 6

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.520916	4.336695	10127	16.615414
Evaluation For Cluster No.0	3.289297	3.967522	2080	10.982242
Evaluation For Cluster No.1	3.612587	4.818386	791	6.242033
Evaluation For Cluster No.2	3.520041	4.466297	1057	8.13017
Evaluation For Cluster No.3	3.283972	3.889126	1430	6.143182
Evaluation For Cluster No.4	3.625765	4.657121	1227	6.187032
Evaluation For Cluster No.5	3.577249	4.239381	1317	8.988261
Evaluation For Cluster No.6	3.594182	4.180553	1554	16.615414
Evaluation For Cluster No.7	4.165199	5.629598	671	8.976103

Modelo 7

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.501791	4.262666	10127	16.619757
Evaluation For Cluster No.0	3.697956	4.262214	459	9.194493
Evaluation For Cluster No.1	3.491492	4.138816	854	6.56281
Evaluation For Cluster No.2	3.585117	4.168153	1547	16.619757
Evaluation For Cluster No.3	3.545472	4.176825	1305	8.990393
Evaluation For Cluster No.4	3.494618	4.446631	1046	8.124066
Evaluation For Cluster No.5	4.160079	5.625996	669	8.976396
Evaluation For Cluster No.6	3.612864	4.817784	790	6.240901
Evaluation For Cluster No.7	3.268567	3.950198	2032	10.981204
Evaluation For Cluster No.8	3.281524	3.880988	1425	6.137569

Modelo 8

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Combined Evaluation	3.445449	4.229891	10127	13.714954
Evaluation For Cluster No.0	3.495047	4.407154	908	6.451621
Evaluation For Cluster No.1	4.287735	5.331909	345	13.714954
Evaluation For Cluster No.2	3.539103	4.116333	885	5.789362
Evaluation For Cluster No.3	3.520454	4.163197	1063	6.066266
Evaluation For Cluster No.4	3.149588	3.81526	1043	5.257454
Evaluation For Cluster No.5	3.161535	3.794018	1783	5.556301
Evaluation For Cluster No.6	4.132433	5.567458	664	7.722296
Evaluation For Cluster No.7	3.607802	4.776258	748	6.223827
Evaluation For Cluster No.8	3.151672	3.629424	1491	5.360525
Evaluation For Cluster No.9	3.593312	4.596087	1197	6.038759

Modelo	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Modelo 1	3	Evenly	Combined Evaluation	3,964208	4,776239	10127	17,082005
Modelo 2	4	K-Means++	Combined Evaluation	3,823517	4,683694	10127	16,857249
Modelo 3	5	Random	Combined Evaluation	3,755299	4,552568	10127	16,923625
Modelo 4	6	First N	Combined Evaluation	3,664318	4,487967	10127	16,58289
Modelo 5	7	Evenly	Combined Evaluation	3,594814	4,389162	10127	16,683463
Modelo 6	8	K-Means++	Combined Evaluation	3,520916	4,336695	10127	16,615414
Modelo 7	9	Random	Combined Evaluation	3,501791	4,262666	10127	16,619757
Modelo 8	10	First N	Combined Evaluation	3,445449	4,229891	10127	13,714954

Probamos desde k=3 hasta k=10 y distintos métodos de inicialización.

El mejor modelo es el que tiene menor distancia intra-cluster (Average distance to cluster center) y mayor distancia entre clusters (Average distance to other center).

Como condición del negocio decimos, aleatoriamente, que la cantidad de observaciones no puede ser menor a 800. Entonces debemos descartar los modelos:

- **Modelo 3:** el cluster 5 (cluster 4 en la notación de Azure) tiene 683 observaciones.
- **Modelo 6:** el cluster 2 (cluster 1 en la notación de Azure) tiene 791 observaciones y el cluster 8 (cluster 7 en la notación de Azure) tiene 671 observaciones.
- **Modelo 7:** el cluster 1 (cluster 0 en la notación de Azure) tiene 459 observaciones, el cluster 6 (cluster 5 en la notación de Azure) tiene 669 observaciones y el cluster 7 (cluster 6 en la notación de Azure) tiene 790 observaciones.
- **Modelo 8:** el cluster 2 (cluster 1 en la notación de Azure) tiene 345 observaciones, el cluster 7 (cluster 6 en la notación de Azure) tiene 664 observaciones y el cluster 8 (cluster 7 en la notación de Azure) tiene 748 observaciones.

En esta instancia, sin haber hecho una caracterización que es el método óptimo para seleccionar un modelo, nos quedaríamos con el modelo 4 porque es el que tiene la segunda menor distancia intra-cluster.

Realizamos la caracterización del mejor modelo encontrado:

Para saber con cual modelo quedarnos debemos hacer una caracterización, por lo que agregamos un script de SQL y de ese modo poder caracterizar al individuo representativo de cada cluster.

```
select assignments
, avg(Customer_Age) as Customer_Age_avg
, avg(Dependent_count) as Dependent_count_avg
, avg(Months_on_book) as Months_on_book_avg
, avg(Total_Relationship_Count) as Total_Relationship_Count_avg
, avg(Months_Inactive_12_mon) as Months_Inactive_12_mon_avg
, avg(Contacts_Count_12_mon) as Contacts_Count_12_mon_avg
, avg(Credit_Limit) as Credit_Limit_avg
, avg(Total_Revolving_Bal) as Total_Revolving_Bal_avg
, avg(Avg_Open_To_Buy) as Avg_Open_To_Buy_avg
, avg(Total_Amt_Chng_Q4_Q1) as Total_Amt_Chng_Q4_Q1_avg
, avg(Total_Trans_Amt) as Total_Trans_Amt_avg
, avg(Total_Trans_Ct) as Total_Trans_Ct_avg
, avg(Total_Ct_Chng_Q4_Q1) as Total_Ct_Chng_Q4_Q1_avg
, avg(Avg_Utilization_Ratio) as Utilization_Ratio_avg
, avg(AF_Existing) as AF_Existing_avg
, avg(EL_Graduate_HighSchool_Unk_Uned) as EL_Graduate_HighSchool_Unk_Uned_avg
, avg(G_Male) as G_Male_avg
, avg(MS_Married_Single) as MS_Married_single_avg
, avg(IC_less40) as IC_less40_avg
, avg(CC_Blue) as CC_Blue_avg
, count(*) as Cantidad_casos
from t1
```

Por cada cluster vamos a calcular el valor promedio de cada atributo.

Nombrar a cada uno de los grupos:






Obligatorio Clustering - Correa Lopez Mosco > Normalize Data > Transformed dataset

rows	columns
10127	2

	Attrition_Flag	AF_Existing
view as		
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Existing Customer		0.437506
Attrited Customer		-2.285681

Como Attrition_Flag es una variable categórica (puede tomar dos valores: 1 ó 0, que luego de la Normalización son 0.437506 y -2.285681 respectivamente), su interpretación no es la misma que con las variables numéricas.

Esto mismo ocurre con todas las variables categóricas.

Gender	G_Male	Education_Level	EL_Graduate_HighSchool_Unk_Uned	Marital_Status	MS_Married_Single
					
M	1.059956	High School	-0.668521	Married	1.077338
F	-0.943436	Graduate	1.495838	Single	-0.928214
M	1.059956	Graduate	1.495838	Married	1.077338
F	-0.943436	High School	-0.668521	Unknown	-0.928214
M	1.059956	Uneducated	-0.668521	Married	1.077338
M	1.059956	Graduate	1.495838	Married	1.077338
M	1.059956	Unknown	-0.668521	Married	1.077338
M	1.059956	Post-Graduate	-0.668521	Unknown	-0.928214
		Uneducated	-0.668521	Unknown	-0.928214
		Doctorate	-0.668521	Single	-0.928214
		Uneducated	-0.668521	Single	-0.928214
		Unknown	-0.668521	Divorced	-0.928214
		College	-0.668521		
Income_Category	IC_less40	Card_Category	CC_Blue		
					
\$60K - \$80K	-0.736437	Blue	0.270611		
Less than \$40K	1.35789	Blue	0.270611		
\$80K - \$120K	-0.736437	Blue	0.270611		
Less than \$40K	1.35789	Blue	0.270611		
\$60K - \$80K	-0.736437	Blue	0.270611		
\$40K - \$60K	-0.736437	Blue	0.270611		
\$120K +	-0.736437	Blue	0.270611		
\$60K - \$80K	-0.736437	Gold	-3.695345		
\$60K - \$80K	-0.736437	Silver	-3.695345		
Unknown	-0.736437	Platinum	-3.695345		

Si analizamos cada atributo (usando reglas de color), podemos ver si los mismos sirven para distinguir entre clusters.

Assignments	Customer_Age_avg	Dependent_count_avg	Months_on_book_avg	Total_Relationship_Count_avg	Months_Inactive_12_mon_avg	Contacts_Count_12_mon_avg	Credit_Limit_avg
0	-0,049382294	0,221767911	-0,04089464	-0,147326141	-0,020785317	0,060086317	2,28510158
1	0,464358818	0,008695361	0,440040105	0,285286012	-0,070675587	-0,040488456	0,037565113
2	-1,2184445	-0,30746784	-1,17116505	0,363317925	-0,194751214	0,103865094	-0,323686197
3	0,245038095	0,058838604	0,2309372	0,082814913	0,030618677	-0,190354657	-0,600279693
4	-0,115945682	-0,028818921	-0,090771154	-1,06907918	-0,128116756	-0,275220021	0,284330326
5	0,074414508	0,064823461	0,07548405	-0,357223684	0,39580692	0,495804894	-0,313221875

Assignments	Total_Revolving_Bal_avg	Avg_Open_To_Buy_avg	Total_Amt_Chng_Q4_Q1_avg	Total_Trans_Amt_avg	Total_Trans_Ct_avg	Total_Ct_Chng_Q4_Q1_avg	Utilization_Ratio_avg
0	0,008791567	2,283833635	0,022767076	0,265917605	0,170526641	-0,013713906	-0,841425749
1	-0,066108717	0,043483926	-0,135978666	-0,326646719	-0,186019255	0,014502014	-0,297735111
2	0,215338013	-0,342923464	0,648167783	-0,383885704	-0,280501247	0,377914946	0,290706065
3	0,225309064	-0,6203528	-0,063749984	-0,118763259	0,182503022	0,139844787	0,731281395
4	0,281888176	0,258999124	0,055741955	2,56000544	1,74878995	0,067426327	-0,269231044
5	-0,78099295	-0,243139458	-0,393030015	-0,50883292	-0,949210557	-0,790947884	-0,494826609

Assignments	AF_Existing_avg	EL_Graduate_HighSchool_Unk_Uned_avg	G_Male_avg	MS_Married_single_avg	IC_less40_avg	CC_Blue_avg	Cantidad_casos
0	-0,001183503	0,03319594	0,658059511	-0,196705558	-0,560318271	-2,12423158	987
1	0,418804518	-0,018635538	0,604396227	0,076475051	-0,69168976	0,270610758	2621
2	0,351736612	0,04410641	0,171312241	0,010079122	-0,26645026	0,257599093	1524
3	0,365817976	-0,00943508	-0,88713264	0,004999191	0,989091417	0,252269449	2811
4	0,418396214	-0,0052908	0,270314847	-0,006363688	-0,133858766	-0,026256075	855
5	-2,271338008	-0,015467887	-0,17463992	-0,02277282	0,079861898	0,231816612	1329

Si los atributos tienen valores muy cercanos a 0 y por tanto muy similares a la media, entonces no “ayudan” a diferenciar entre clusters; observamos que la mayoría resultan cercanos a cero.

Por otro lado, si usamos reglas de color por cluster podemos ver dentro de cada uno qué atributos son más llamativos, las variables que mejor separan y cómo están caracterizados. Basándonos en la tabla que vemos arriba para k=6:

- **Cluster 1** (Assignment 0) tiene la mayoría de los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis, pero los valores que sí se disparan de la media son Credit_Limit (2,28510158) y Avg_Open_To_Buy (2,283833635). Lo podemos considerar como un cliente con alto límite crediticio y con saldo disponible en su tarjeta. Tiene 987 observaciones.
- **Cluster 2** (Assignment 1) tiene la mayoría de los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis. Sin embargo, CC_Blue tiene un valor de 0,270610758, entonces decimos que todas las observaciones son de CC_Blue=1 por lo que vimos en el análisis de este atributo.
En resumen, lo podemos considerar como un cliente promedio que tiene tarjeta blue. Tiene 2621 observaciones.
- **Cluster 3** (Assignment 2) tiene todos los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis.
Lo podemos considerar como un cliente promedio. Tiene 1524 observaciones.

En este punto NO dejamos de considerar a este cluster, como sí podríamos dejar de considerar un atributo por no “ayudar” a diferenciar entre clusters.

- **Cluster 4** (Assignment 3) tiene todos los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis.
Lo podemos considerar como un cliente promedio. Tiene 2811 observaciones.
- **Cluster 5** (Assignment 4) tiene la mayoría de los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis. Sin embargo, los valores que sí se disparan de la media son Total_Trans_Amt (2,5600544) y Total_Trans_Ct (1,74878995).
Lo podemos considerar como un cliente que tiene un alto monto consumido en los últimos 12 meses y alta cantidad de transacciones en los últimos 12 meses. Tiene 855 observaciones.
- **Cluster 6** (Assignment 5) tiene todos los valores de los atributos cercanos a la media porque dan en el entorno de cero, entonces decimos que el cluster no se destaca en ninguno y no nos sirven para el análisis. Lo podemos considerar como un cliente promedio. Tiene 1329 observaciones.

Variables que aportan a la hora de discriminar los grupos:

Assignments	Customer_Age_avg	Dependent_count_avg	Months_on_book_avg	Total_Relationship_Count_avg	Months_Inactive_12_mon_avg	Contacts_Count_12_mon_avg	Credit_Limit_avg
0	-0,049382294	0,221767911	-0,04089464	-0,147326141	-0,020785317	0,060086317	2,285101584
1	0,464358818	0,008695361	0,440040105	0,285286012	-0,070675587	-0,040488456	0,037565113
2	-1,218444585	-0,30746784	-1,171165063	0,363317925	-0,194751214	0,103865094	-0,323686197
3	0,245038095	0,058838604	0,2309972	0,082814913	0,030618677	-0,190354657	-0,600279693
4	-0,115945682	-0,028818921	-0,090771154	-1,069079181	-0,128116756	-0,275220021	0,284330326
5	0,074414508	0,064823461	0,075484005	-0,357223684	0,39580692	0,495804894	-0,313221875
Varianza	0,286951215	0,025372441	0,263768885	0,232378454	0,036224207	0,061650693	0,925738494

Assignments	Total_Revolving_Bal_avg	Avg_Open_To_Buy_avg	Total_Amt_Chng_Q4_Q1_avg	Total_Trans_Amt_avg	Total_Trans_Ct_avg	Total_Ct_Chng_Q4_Q1_avg	Utilization_Ratio_avg
0	0,008791567	2,283833635	0,022767076	0,265917605	0,170526641	-0,013713906	-0,841425749
1	-0,066108717	0,043483926	-0,135978666	-0,326646719	-0,186019255	0,014502014	-0,297735111
2	0,215338013	-0,342923464	0,648167783	-0,383885704	-0,280501247	0,377914946	0,290706065
3	0,225309064	-0,6203528	-0,063749984	-0,118763259	0,182503022	0,139844787	0,731281395
4	0,281888176	0,258999124	0,055741955	2,56000545	1,748789951	0,067426327	-0,269231044
5	-0,78099295	-0,243139458	-0,393030015	-0,50883292	-0,949210557	-0,790947884	-0,494826609
Varianza	0,131459875	0,921510653	0,099630803	1,130416664	0,676080809	0,130986191	0,267306114

Assignments	AF_Existing_avg	EL_Graduate_HighSchool_Unk_Uned_avg	G_Male_avg	MS_Married_single_avg	IC_less40_avg	CC_Blue_avg	Cantidad_casos
0	-0,001183503	0,03319594	0,658059511	-0,196705558	-0,560318271	-2,12423158	987
1	0,418804518	-0,018635538	0,604396227	0,076475051	-0,69168976	0,270610758	2621
2	0,351736612	0,044410641	0,171312241	0,010079122	-0,26645026	0,257599093	1524
3	0,365817976	-0,00943508	-0,88713264	0,004999191	0,989091417	0,252269449	2811
4	0,418396214	-0,0052908	0,270314847	-0,006363688	-0,133858766	-0,026256075	855
5	-2,271338008	-0,015467887	-0,17463992	-0,02277282	0,079861898	0,231816612	1329
Varianza	0,94685118	0,000606634	0,274916354	0,007036935	0,301544346	0,759017012	

Calculamos la varianza de cada atributo. **Varianzas altas indican valores diferentes, más separados, por lo que cuanto más separados estén mejor porque mejor discriminan los grupos.**

Si vemos cada atributo por su varianza, la categoría que aporta en mayor medida, porque es la de mayor varianza y superior a 1, es Total_Trans_Amt (1,130416657) y fue la que más separo los clusters. Los que aportan en menor medida y no llegan a despegarse completamente de la media porque siguen siendo valores cercanos a cero son: Credit_Limit, Avg_Open_To_Buy, Total_Trans_Ct, AF_Existing y CC_Blue_avg (los verdes en la fila de Varianza).

Buscamos el nuevo “mejor modelo” utilizando la técnica Hyper Parameter Tuning (Sweep Clustering):

En el módulo “Sweep Clustering” vamos a definir distintos valores de “k” y Azure va a crear tantos modelos como valores de “k” le hayamos dado. Una vez terminado va a arrojar cuál de ellos es el mejor.

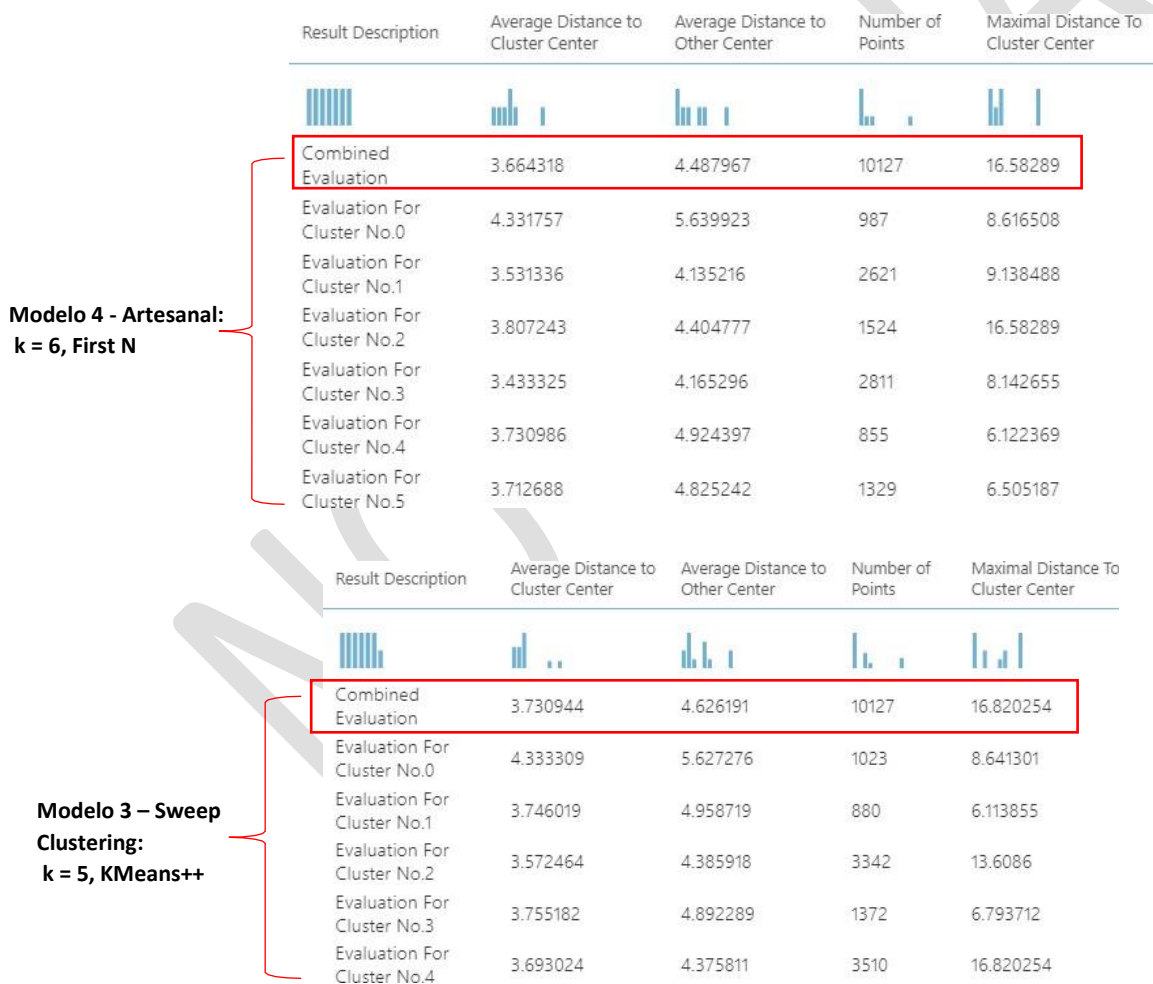
	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Modelo 1: k = 8, First N	Combined Evaluation	3.541218	4.366469	10127	14.166088
	Evaluation For Cluster No.0	3.486849	4.434759	984	6.462139
	Evaluation For Cluster No.1	3.660728	4.743036	1273	6.772846
	Evaluation For Cluster No.2	4.190358	5.120721	454	14.166088
	Evaluation For Cluster No.3	3.553018	4.199001	1473	6.133601
	Evaluation For Cluster No.4	3.3654	3.923137	2122	6.065379
	Evaluation For Cluster No.5	3.33267	3.997909	2379	6.391235
	Evaluation For Cluster No.6	4.145965	5.603634	668	8.149312
	Evaluation For Cluster No.7	3.611664	4.817122	774	6.238283
Modelo 2: k = 4, Random	Combined Evaluation	3.823517	4.683694	10127	16.857249
	Evaluation For Cluster No.0	3.781661	4.928076	1409	6.763225
	Evaluation For Cluster No.1	3.777042	4.429969	3868	16.857249
	Evaluation For Cluster No.2	3.652997	4.455474	3541	13.631124
	Evaluation For Cluster No.3	4.467177	5.787748	1309	8.864923
Modelo 3: k = 5, KMeans++	Combined Evaluation	3.730944	4.626191	10127	16.820254
	Evaluation For Cluster No.0	4.333309	5.627276	1023	8.641301
	Evaluation For Cluster No.1	3.746019	4.958719	880	6.113855
	Evaluation For Cluster No.2	3.572464	4.385918	3342	13.6086
	Evaluation For Cluster No.3	3.755182	4.892289	1372	6.793712
	Evaluation For Cluster No.4	3.693024	4.375811	3510	16.820254
	Combined Evaluation	3.655149	4.498165	10127	16.88163
Modelo 4: k = 6, Evenly	Evaluation For Cluster No.0	4.16935	5.667779	673	8.973435
	Evaluation For Cluster No.1	3.665691	4.221102	3438	16.88163
	Evaluation For Cluster No.2	3.519335	4.268446	2666	13.594982
	Evaluation For Cluster No.3	3.720331	4.849086	1324	6.76979
	Evaluation For Cluster No.4	3.583449	4.506438	1217	12.558593
	Evaluation For Cluster No.5	3.631342	4.87287	809	6.253418

Modelo	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Modelo 1	8	First N	Combined Evaluation	3,541218	4,366469	10127	14,166088
Modelo 2	4	Random	Combined Evaluation	3,823517	4,683694	10127	16,857249
Modelo 3	5	K-Means++	Combined Evaluation	3,730944	4,626191	10127	16,820254
Modelo 4	6	Evenly	Combined Evaluation	3,655149	4,498165	10127	16,88163

Nos quedamos con el modelo 3 ya que las distancias intra-cluster y entre clusters no varían tanto respecto al modelo 2 y tampoco frente a los extremos (modelos 1 y 4).

Conclusión:

Modelo	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Modelo 4 - Artesanal	6	First N	Combined Evaluation	3,664318	4,487967	10127	16,582890
Modelo 3 - Sweep Clustering	5	K-Means++	Combined Evaluation	3,730944	4,626191	10127	16,820254



La distribución de las observaciones es muy similar y mayor a 800 como definimos anteriormente, por lo que no es un factor decisivo en nuestra selección.

Por lo tanto, determinamos que el mejor modelo es el que tiene menor distancia intra-cluster (Average distance to cluster center) → Modelo 4 – Artesanal.