Machine Learning Project - CLUSTERING

Description:

A bank wishes to generate groups of its customers based on the data from the dataset mentioned in the following section in order to generate differential customer service policies for each type of customer.

Create the best possible clustering model to meet the objectives defined above, using the dataset found in the "BankChurners.csv" file.

References to variables:

- **CLIENTNUM:** Number of clients → quantitative variable.
- **Attrition_Flag:** Account status the following month → qualitative variable.
- **Customer_Age:** quantitative variable.
- **Gender:** qualitative variable.
- **Dependent_count:** Number of people in charge → quantitative variable.
- **Education_Level:** qualitative variable.
- **Marital_Status:** qualitative variable.
- **Income_Category:** qualitative variable.
- **Card_Category:** qualitative variable.
- **Months_on_book:** Account's age → qualitative variable.
- **Total_Relationship_Count:** Number of customer's products (accounts and cards)→ quantitative variable.
- **Months_Inactive_12_mon:** Number of months inactive in the last 12 months → quantitative variable.
- **Contacts_Count_12_mon:** Number of contacts in the last 12 months (inquiries/claims to the bank)→ quantitative variable.
- **Credit_Limit**: quantitative variable.
- **Total_Revolving_Bal:** Uncovered balance of the card (it would be what the client has used of the amount on his card, it is the difference between Credit_Limit y Avg_Open_To_Buy)→ quantitative variable.
- **Avg_Open_To_Buy:** Available amount in the card → quantitative variable.
- **Total_Amt_Chng_Q4_Q1:** Percentage change in the amount of consumption→ quantitative variable.
- **Total_Trans_Amt:** Amount of consumption in the last 12 months → quantitative variable.
- **Total_Trans_Ct:** Number of transactions in the last 12 months → quantitative variable.
- **Total_Ct_Chng_Q4_Q1:** Percentage change in amount of consumption → quantitative variable.
- **Avg_Utilization_Ratio:** Card utilization ratio (it is the result of doing Total_Revolving_Bal divided by Credit_Limit) → quantitative variable.

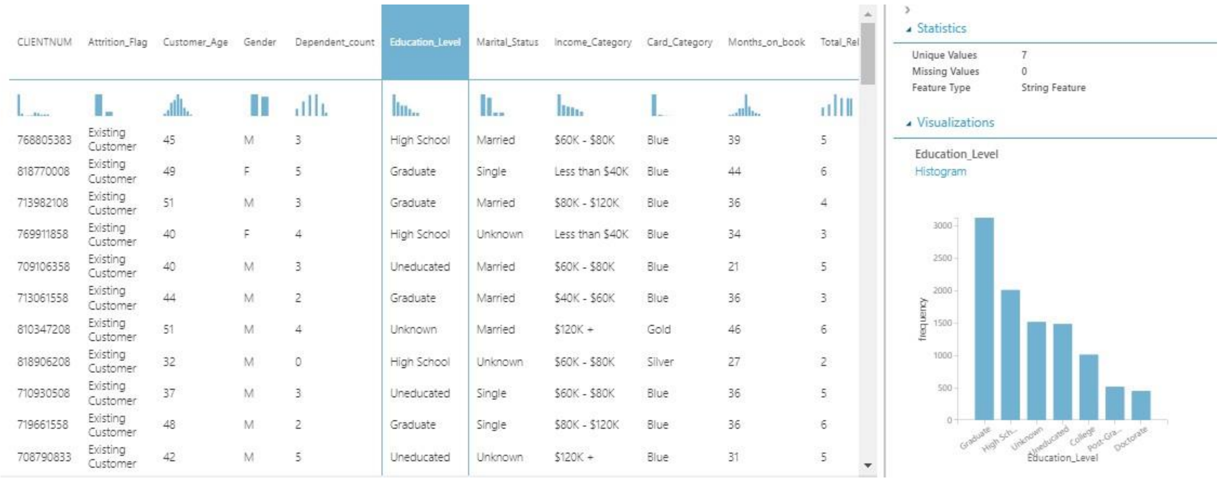
Experiment in Azure:

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

Categorical variables:

The categorical variables are:

- **Attrition_Flag:** It is the target variable.
- **Gender:** Indicates the gender categories of a person.
- **Education_Level:** Indicates the categories of a person's level of education.
- **Marital_Status:** Indicates the marital categories of a person.
- **Income_Category:** Indicates a person's income categories.
- **Card_Category:** Indicates the categories of type of cards that a person can have.

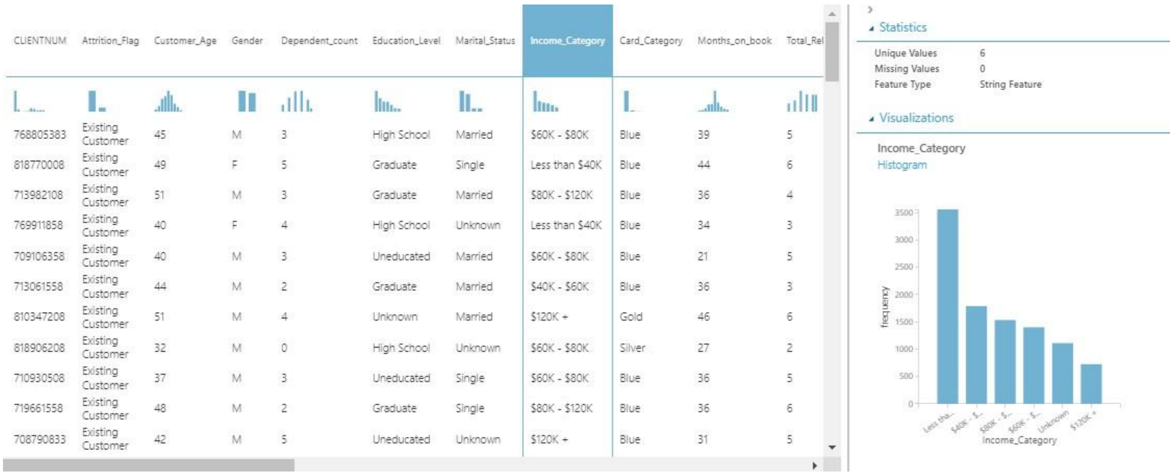


CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Rel
768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5
818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6
713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4
769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3
709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5
713061558	Existing Customer	44	M	2	Graduate	Married	$40K - $60K	Blue	36	3
810347208	Existing Customer	51	M	4	Unknown	Married	$120K +	Gold	46	6
818906208	Existing Customer	32	M	0	High School	Unknown	$60K - $80K	Silver	27	2
710930508	Existing Customer	37	M	3	Uneducated	Single	$60K - $80K	Blue	36	5
719661558	Existing Customer	48	M	2	Graduate	Single	$80K - $120K	Blue	36	6
708790833	Existing Customer	42	M	5	Uneducated	Unknown	$120K +	Blue	31	5

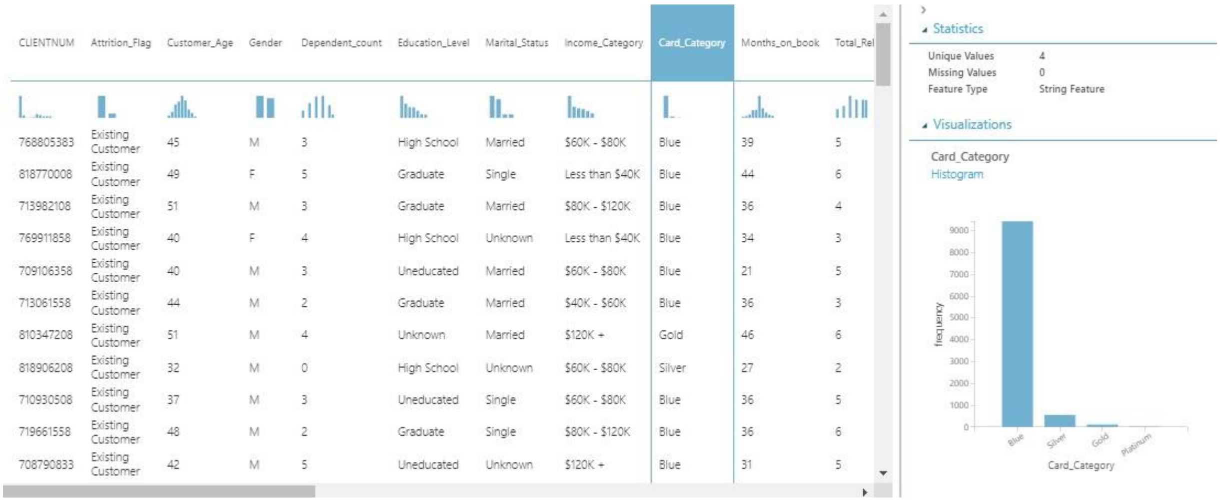
We have 7 educational levels.
We observe that the clients with the highest occurrence are graduates.

CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Rel
768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5
818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6
713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4
769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3
709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5
713061558	Existing Customer	44	M	2	Graduate	Married	$40K - $60K	Blue	36	3
810347208	Existing Customer	51	M	4	Unknown	Married	$120K +	Gold	46	6
818906208	Existing Customer	32	M	0	High School	Unknown	$60K - $80K	Silver	27	2
710930508	Existing Customer	37	M	3	Uneducated	Single	$60K - $80K	Blue	36	5
719661558	Existing Customer	48	M	2	Graduate	Single	$80K - $120K	Blue	36	6
708790833	Existing Customer	42	M	5	Uneducated	Unknown	$120K +	Blue	31	5

We have 4 marital statuses.
We note that most of the clients are married.



We have 6 income categories.
We note that most clients earn less than $40,000 per year.



We have 4 categories of cards.
We note that most customers have a blue card.

We detail the script below, where we will convert the categorical variables to dummy variables:

```
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;
```

We group the variables "Income_Category" and "Card_Category" according to the category with the greatest presence in each of them, we do this because they have more than 2 categories. In the case of "Education_Level" and "Marital_Status", we group according to the number of observations.

Variables selection:

The attributes that we are going to exclude are:

- ***CLIENTNUM:*** it is not going to give us any type of information when running our model because it is a unique identifier of the customer and of the purchase and we are not going to need it to train our model.
- ***Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1:*** We have no information on this variable.
- ***Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2:*** We have no information on this variable.

We include all the numerical variables.

We select n-1 dummy variables from each category and leave out the categorical variables.

Models creation:

We create models with different amounts of clusters and initializations to be able to make a comparison between them.

Model 1

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
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

Model 2

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
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

Model 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

Model 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

Model 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

Model 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

Model 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

Model 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

Model	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Model 1	3	Evenly	Combined Evaluation	3.964208	4.776239	10127	17.082005
Model 2	4	K-Means++	Combined Evaluation	3.823517	4.683694	10127	16.857249
Model 3	5	Random	Combined Evaluation	3.755299	4.552568	10127	16.923625
Model 4	6	First N	Combined Evaluation	3.664318	4.487967	10127	16.58289
Model 5	7	Evenly	Combined Evaluation	3.594814	4.389162	10127	16.683463
Model 6	8	K-Means++	Combined Evaluation	3.520916	4.336695	10127	16.615414
Model 7	9	Random	Combined Evaluation	3.501791	4.262666	10127	16.619757
Model 8	10	First N	Combined Evaluation	3.445449	4.229891	10127	13.714954

We tried from k=3 to k=10 and different initialization methods.

The best model is the one with the smallest intra-cluster distance (Average distance to cluster center) and the largest distance between clusters (Average distance to other center).

As a business condition we randomly say that the number of observations cannot be less than 800. Then we must discard the models:

- ***Model 3:*** Cluster 5 (Cluster 4 in Azure) has 683 observations.
- ***Model 6:*** Cluster 2 (Cluster 1 in Azure) has 791 observations and Cluster 8 (Cluster 7 in Azure) has 671 observations.
- ***Model 7:*** Cluster 1 (Cluster 0 in Azure) has 459 observations, Cluster 6 (Cluster 5 in Azure) has 669 observations and Cluster 7 (Cluster 6 in Azure) has 790 observations.
- ***Model 8:*** Cluster 2 (Cluster 1 in Azure) has 345 observations, Cluster 7 (Cluster 6 in Azure) has 664 observations and Cluster 8 (Cluster 7 in Azure) has 748 observations.

In this instance, without having made a characterization that is the optimal method to select a model, we would choose model 4 because it is the one with the second smallest intra-cluster distance. We can call this model "hand-made" because we vary "k" values manually.

We perform the characterization of the best model found:

To decide on which model, we must make a characterization, so we write a SQL script and this way be able to characterize the representative individual of each 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
```

Cantidad_casos = Amount_of_cases.
For each cluster we will calculate the average value of each attribute.

Name each of the groups:

rows	columns
10127	2

view as

Attrition_Flag	AF_Existing
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

As Attrition_Flag is a categorical variable (it can take two values: 1 or 0, which after Normalization are 0.437506 and -2.285681 respectively), its interpretation is not the same as with numeric variables.
The same thing happens with all categorical variables.

Gender	G_Male
M	1.059956
F	-0.943436
M	1.059956
F	-0.943436
M	1.059956
M	1.059956
M	1.059956

Education_Level	EL_Graduate_HighSchool_Unk_Uned
High School	-0.668521
Graduate	1.495838
Graduate	1.495838
High School	-0.668521
Uneducated	-0.668521
Graduate	1.495838
Unknown	-0.668521
Post-Graduate	-0.668521
Uneducated	-0.668521
Doctorate	-0.668521
Uneducated	-0.668521
Unknown	-0.668521
College	-0.668521

Marital_Status	MS_Married_Single
Married	1.077338
Single	-0.928214
Married	1.077338
Unknown	-0.928214
Married	1.077338
Married	1.077338
Married	1.077338
Unknown	-0.928214
Single	-0.928214
Single	-0.928214
Divorced	-0.928214

Income_Category	IC_less40
$60K - $80K	-0.736437
Less than $40K	1.35789
$80K - $120K	-0.736437
Less than $40K	1.35789
$60K - $80K	-0.736437
$40K - $60K	-0.736437
$120K +	-0.736437
$60K - $80K	-0.736437
$60K - $80K	-0.736437
Unknown	-0.736437

Card_Category	CC_Blue
Blue	0.270611
Blue	0.270611
Blue	0.270611
Blue	0.270611
Blue	0.270611
Blue	0.270611
Gold	-3.695345
Silver	-3.695345
Platinum	-3.695345

If we analyze each attribute (using color rules), we can see if they help to distinguish between 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	0,044410641	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

Cantidad_casos =Amount_cases

If the attributes have values very close to 0 and therefore very similar to the mean, then they do not "help" to differentiate between clusters; we observe that most are close to zero.

On the other hand, if we use color rules by cluster, we can see within each one which attributes are most striking, the variables that best separate them and how they are characterized. Based on the table we see above for k=6:

- **Cluster 1** (Assigment 0) has most of the values of the attributes close to the mean because they are around zero, so we say that the cluster does not stand out in any of them and they are not useful for the analysis, but the values that do rise above the mean are Credit_Limit (2,28510158) and Avg_Open_To_Buy (2,283833635). We can consider it as a client with a high credit limit and with an available balance on his/her card. It has 987 observations.
- **Cluster 2** (Assigment 1) has most of the values of the attributes close to the mean because they are close to zero, so we say that the cluster does not stand out in any of them, and they are not useful for the analysis. However, CC_Blue has a value of 0.270610758, so we say that all observations are of CC_Blue=1 from what we see in the analysis of this attribute. In short, we can consider it as an average customer who has a blue card. It has 2621 observations.
- **Cluster 3** (Assigment 2) has all the values of the attributes close to the mean because they are close to zero, so we say that the cluster does not stand out in any of them, and they are not useful for the analysis.
  We can consider it as an average customer. It has 1524 observations.

At this point we do NOT stop considering this cluster, as we could stop considering an attribute for not "helping" to differentiate between clusters.

- **Cluster 4** (Assigment 3) has all the values of the attributes close to the mean because they are close to zero, so we say that the cluster does not stand out in any of them, and they are not useful for the analysis. We can consider it as an average customer. It has 2811 observations.
- **Cluster 5** (Assigment 4) has most of the values of the attributes close to the mean because they are close to zero, so we say that the cluster does not stand out in any of them, and they are not useful for the analysis. However, the values that do rise above the mean are Total_Trans_Amt (2.5600544) and Total_Trans_Ct (1.74878995).

  We can consider it as a customer who has a high amount consumed in the last 12 months and a high number of transactions in the last 12 months. It has 855 observations.
- **Cluster 6** (Assigment 5) has all the values of the attributes close to the mean because they are close to zero, so we say that the cluster does not stand out in any of them, and they are not useful for the analysis. We can consider him as an average customer. It has 1329 observations.

Variables that contribute discriminating groups:

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	-1,171165053	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	-1,069079181	-0,128116756	-0,275220021	0,284330326
5	0,074414508	0,064823461	0,07548405	-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	

Cantidad_casos = Amount_cases
Varianza = Variance

We calculate the variance of each attribute. **High variances indicate different values, more separated, so the more separated they are, the better because they discriminate the groups better.**

If <u>we see each attribute by its variance</u>, the category that contributes the most, because it is the one with the greatest variance and greater than 1, is Total_Trans_Amt (1.130416657) and it was the one that separated the clusters the most. Those that contribute to a lesser extent and do not completely detach from the average because they are still values close to zero are: Credit_Limit, Avg_Open_To_Buy, Total_Trans_Ct, AF_Existing and CC_Blue_avg (the green ones in the Variance row).

We look for the new "best model" using the Hyper Parameter Tuning technique (Sweep Clustering):

In the "Sweep Clustering" module we are going to define different values of "k" and Azure is going to create as many models as values of "k" we have given it. Once finished, it will throw which of them is the best.

	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Model 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
Model 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

	Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
Model 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
Model 4: k = 6, Evenly	Combined Evaluation	3.655149	4.498165	10127	16.88163
	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

Model	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Model 1	8	First N	Combined Evaluation	3.541218	4.366469	10127	14.166088
Model 2	4	Random	Combined Evaluation	3.823517	4.683694	10127	16.857249
Model 3	5	K-Means++	Combined Evaluation	3.730944	4.626191	10127	16.820254
Model 4	6	Evenly	Combined Evaluation	3.655149	4.498165	10127	16.88163

We choose model 3 since the intra-cluster and inter-cluster distances do not vary as much respect to model 2 and neither with respect to the extremes (models 1 and 4).

Conclusion:

Model	k	Initialization	Result description	Average distance to cluster center	Average distance to other center	Number of points	Maximal distance
Model 4 – "hand-made"	6	First N	Combined Evaluation	3.664318	4.487967	10127	16.582890
Model 3 - Sweep Clustering	5	K-Means++	Combined Evaluation	3.730944	4.626191	10127	16.820254

Model 4 – "hand-made":
k = 6, First N

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

Model 3 – Sweep Clustering:
k = 5, KMeans++

Result Description	Average Distance to Cluster Center	Average Distance to Other Center	Number of Points	Maximal Distance To Cluster Center
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

The distribution of observations is very similar and greater than 800 as defined above, so it is not a decisive factor in our selection.
Therefore, we say that the best model is the one with the smallest intra-cluster distance (Average distance to cluster center) → Model 4 – "hand-made".