Machine Learning for clustering (unsupervised)

[EX1] **Let's identify the type of the variables (integer, float, chart...) and the size of the dataset and the file. Which is the variable with more nulls? And with no nulls?**

Variable	Type	Number of nulls	Number of non-nulls
Company_id	integer	0	1175
Reference_date_month	object	0	1175
Product	object	0	1175
Sub_product	object	0	1175
Type_ID_Coverage_GSM	object	47	1128
Type_ID_Coverage_UMTS	object	47	1128
Type_ID_Coverage_LTE	object	47	1128
CNT_EMPLOYEE	float	51	1124
Sector	integer	0	1175
Sub_sector	integer	0	1175
Turnover	float	403	772
Revenue_Scale	object	403	772
ZIP	object	0	1175
Data_usage	integer	0	1175
Voice_usage	float	0	1175
Monthly_expense	float	0	1175
N_lines	integer	0	1175

Dataset size: 17 columns and 1175 rows.
File size: 156.2 KB

The variables with more nulls are *Turnover* and *Revenue_Scale*, which have 403 nulls, and variables *Company_id, Reference_date_month, Product, Sub_product, Sector, Sub_sector, ZIP, Data_usage, Voice_usage, Montly_expense,* and *N_lines* are all filled, they do not have nulls.

[EX6] **Calculate the main statistics (max, min, mean, median and standard deviation) of Data_usage, Voice_usage, Monthly_expense and N_lines variables. Plot a histogram for each of these variables.**

	Data_usage	Voice_usage	Montly_expense	N_lines
max	2568.38	260638.69	9091.51	973
max (NO outliers)	27.28	7657.23	1762.07	58.33
min	0	0	0	0
min (NO outliers)	0	0	0	0
mean	40.95	3412.36	271.70	17.73
mean (NO outliers)	4.91	777.29	102.01	151.22
median	3.02	369.48	65.29	2.0
median (NO outliers)	1.66	274.84	53.78	2.00
std dev	196.46	17114.64	778.99	72.48
st dev (NO outliers)	6.70	1254.31	151.22	6.9

As you can see, we also created another data set called No outliers, in which we get rid of the 15% top values in Data Usage. This is because in the original histogram we are not able to infer many things but in the no outliers we are able to observe in which range lie most part of the points. For example, according to data usage, most of the customers will be between 0 and 10.

[See histograms below]

[EX7] **Calculate and plot the correlation matrix between traffic attributes (i.e., Voice_usage and Data_usage), Monthly_expense, N_lines, Turnover and CNT_EMPLOYEE.**

- **Which are the variables with more and less correlation with respect to the Monthly_expense variable?**
  N_lines is the variable that is more correlated with *Montly_expense* and *CNT_EMPLOYEE* the less.
- **Which are the top 2 variables with higher correlation with Voice_usage?**
  N_lines and *Montly_expense* are the variables with higher correlation with *Voice_usage*.
- **Is Data_usage correlated with Turnover? Does it mean that a company spends more in Data usage when its Turnover increases?**
  It is positively correlated but it is not very strong. There is no strong correlation to ensure that if turnover increases, more money will be spent in data usage.
- **Which is the highest correlated variable with CNT_EMPLOYEE?**
  Turnover is the variable that is higher correlated with *CNT_EMPLOYEE*.
- **A company with high Voice traffic consumption but low Data traffic uses to spend more than other company with high Data traffic and low Voice traffic? Justify your answer.**
  Yes, because the correlation between Voice usage and the monthly expense (0.327104) is higher than the correlation between Data usage and the expense (0.767026).

[See correlation matrix below]

[EX8] **Visualize a scatter plot with Voice_usage vs Monthly_expense variables. Could you visually identify any cluster? How many? Repeat the plot with registers which Voice_usageis between 0 and 10000 minutes. Could you identify any cluster?**
In the first plot we can observe three clusters, and in the second one, in the way data is spread, we cannot clearly identify any cluster, you could say that maybe the three isolated points on top left and top right form other 2 clusters and that the big one belongs to a Gaussian.
[See plots below]

[EX9] **Visualize a scatter plot with Data_usage vs Monthly_expense variables. Could you visually identify any cluster? How many?Repeat the plot with registers which Data_usageis between 0 and 500 GB. Could you identify any cluster?**

- Data_usage vs Monthly_expense: There is clearly one main cluster which has most of the points and then three isolated points which we would not know what to say about them. Depending on the interpretation you give, these could be just outliers of the first cluster, or be clusters of one point, having 3/4 clusters in total.
- Data_usage vs Voice_usage: In this second picture we are zooming in the cluster identified before, and now it is difficult to identify any cluster at all, we can see the many points on the bottom left of the plot but also there are quite points distributed along the whole plot.

[See plots below]

[EX10] **To improve our understanding of the data, plot a 3D visualization between Voice_usage, Data_usageand Monthly_expense for a new subset of the dataset where Voice_usageis below 10000 minutes and Data_usageis <=100 GB.**

- **Could you visually identify any cluster? How many?**
  We can observe one main cluster and then some spread points.
- **Could you identify any outlier?**
  The yellow and green points could be considered outliers.
- **Could you identify a cluster bigger than the others? Describe approximately it in terms of the values of these 3 variables.**
  The main cluster could be considered (0,2000) *Voice_usage*, (0, 20) *Data_usage* and (0, 200) *Montly_expense*.

[See plots below]

[EX14] **Plot the following scatter plots representing the centroids. Describe the minority cluster in terms of Data_usage, Voice_usage and Monthly_expense. How many registers are formed by?**
Black points represent the centroids and the other colors represent the clusters.
The minority cluster is formed by just one register.
Data_usage= 739.68 Voice_usage= 260638.68 Montlhy_expense= 9091.51
As you can see the centroid of this cluster coincides with the position of the register, since there is only one.

[See plots below]

[EX15] **Execute the Sklearn library's KMeans function and compare both Data_usage vs Voice_usage scatter plots. Are they similar?**
After executing the KMeans of the library we obtained the same results as in the Kmeans constructed by ourselves. The label assignment might differ, but the clusters end up being the same. and therefore, we expect to have the Data_usage vs Voice_usage plot exactly as before.
Finally, we obtained the same results but with different colors because the initialized centers were different, but even so they converged in the same final centroids.
[See plot below] (Remember that black is for the centroids)

[EX16] **Repeat the 3D plot visualization between Voice_usage, Data_usageand Monthly_expense after the clustering process. Apply a rotation of (0, -60).**
[See plot below]

[EX17] **Repeat the clustering (using Sklearn's or your own Kmeans function) with K=5 and plot Data_usagevs Voice_usage scatter visualization. Can the new 5 clusters be visually distinguished? From a visual perspective, is this new cluster better than with K=3?**
No, they cannot be easily identified since with 5 clusters we have many centroids together and from a visual perspective we would not think in organizing the data in that way, therefore we can conclude that using K=3 is a better option than 5.
[See plot below]

[EX18] **Repeat the Elbow method for a training dataset formed by Voice_usage and Data_usage only. Which is the optimal K value?**
In both cases, the optimal value of K is 3.
[See plots below]

[EX19] **Calculate the silhouette_scorevalue for a range of KMeans clusters from 2 to 7. The dataset to use is training_dtwith the following variables: Voice_usage, Data_usage and Monthly_expense. Which is the value of K with better Silhouette?**
Silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters, near 0 indicate that the sample is near the decision boundary, and negative values mean that they are assigned in the wrong cluster. Hence, we have to choose a graph with a red line near to 1, the graph with two clusters is nearer to 1 but the second cluster is formed by one point, hence almost all the data is inside the first cluster. In consequence, we will choose the second graph with the highest score that is the one with 3 clusters, now the 3rd cluster has more points, so the division makes more sense.
[See plots below]

[EX20] **Repeat the K-Means clustering with K=3 for the training_dtformed by CNT_EMPLOYEE, Turnover, Voice_usage, Data_usage, Monthly_expense. For each cluster, calculate the mean, standard deviation, min, max for each variable.**

- **Which is the cluster with the highest voice usage?**
  Cluster 1, which has 263764.53 voice usage.
- **Which is the cluster with the highest data usage?**
  Cluster 0l, is the one with highest data usage (4651.69).

- **Customers in the cluster with bigger companies (i.e., bigger number or employees and turnover) use to spend more than the other customers?**
  This corresponds to cluster 1, and yes. Observe that the mean and the minimum values are higher than in the other clusters.
- **As a part of the data scientist team, which is your recommended cluster of customers to sell a new mobile tariff with unlimited data traffic? And for a new mobile tariff with unlimited voice traffic?**

  If the objective is to maximize the customer's benefit, we would sell it to cluster 0, since is the one with higher mean. The maximum value is also the highest but also the standard deviation, which means that there are some customers that will benefit a lot from this.
  In this case, to cluster 1. The reasons are the same as before but in this case applied to cluster 0 instead of 1.

[EX21] **Execute the Mixture of Gaussians function (with number of components=3) to training_dt dataset with Voice_usage, Data_usageand Monthly_expense variables.**

- **Which is the size of each cluster?**
  Cluster 0: 538 registers
  Cluster 1: 3 registers
  Cluster 2: 227 registers
- **Visualize the scatter plot between Data_usage vs Voice_usage. Is it similar to the resulting from K-Means and K=3?**
  Yes, they are equal.
- **Visualize the scatter plot between Data_usage vs Monthly_expense. Is it similar to the resulting from K-Means and K=3?**
  They are similar but some points of the purple cluster now are from the yellow one.
- **Visualize the scatter plot between Voice_usage vs Monthly_expense. Is it similar to the resulting from K-Means and K=3?**
  As before, some points of the purple cluster are now from the yellow.

[See plots below]

[EX22] **Visualize the 3D plot between Voice_usage, Data_usageand Monthly_expense after the clustering process. Apply a rotation of (0, -60).**
Now you can observe that the clusters are similar but some of there are more points belonging to the yellow cluster than before.
[See plot below]

[EX23] **Evaluate the Silhouette metric for MoG with number of components from 2 to 7. Which is the number of the cluster with the highest score?**
The highest average score is for 2 clusters, since the formula will tend to "favor" lower components but see that with 7 components we observe an increase with respect to the previous iteration, so it is usually an indicator that is also a good choice.
[See plots below]

[EX24] **For n_components=3 and a dataset formed by CNT_EMPLOYEE, Turnover, Voice_usage, Data_usage and Monthly_expense variables, calculate the MoG clustering. For each cluster, calculate the mean, standard deviation, min, max for each variable.**

- **Are these clusters similar to the resulting from KMeans=3?**
  No, they differ in almost all the values.
- **Which is the cluster with the highest voice usage?**
  Cluster 0 is the one with highest voice usage (260638.68).
- **Which is the cluster with the highest data usage?**
  Cluster 2 is the one with highest data usage (2568.38).
- **Customers in the cluster with bigger companies (i.e., bigger number or employes and turnover) use to spend more than the other customers?**

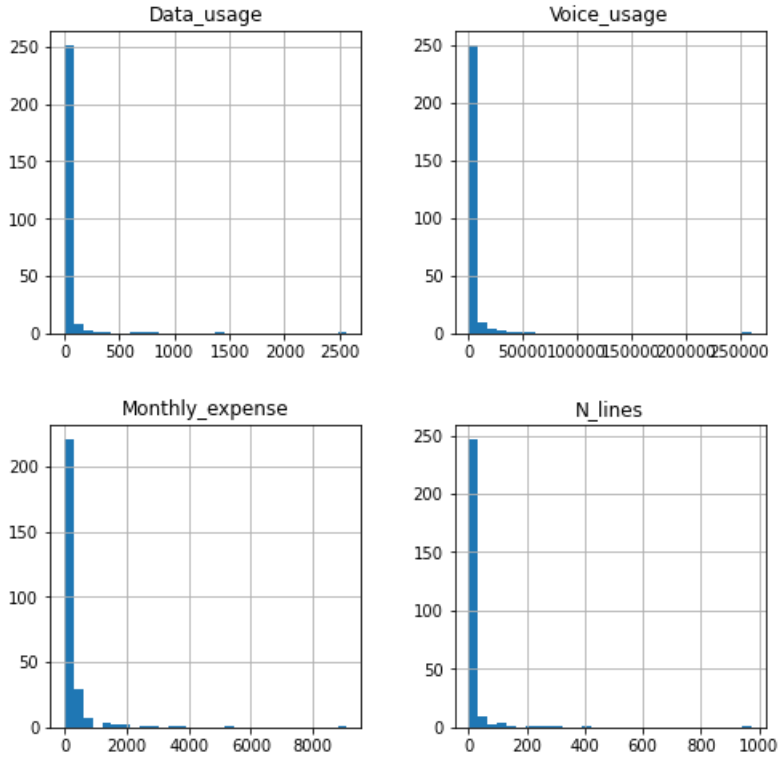
This corresponds to cluster 1, and yes. Observe that the mean and the minimum values are higher than in the other clusters. Maybe cluster 0 has a higher maximum value of monthly expense, but the standard deviation of this cluster is much higher, so in general customers in the cluster with big companies use to spend more.

- **As a part of the data scientist team, which is your recommended cluster of customers to sell a new mobile tariff with unlimited data traffic? And for a new mobile tariff with unlimited voice traffic?**
  If the objective is to maximize the customer's benefit, we would sell it to cluster 2, it is the one with higher mean and maximum values. It's true that it has a very high standard deviation, but we think that in some part, it could be explained by the higher values rather than the lower ones.
  Regarding the unlimited voice traffic, we would sell it to cluster 0 again. In this case the mean is much higher than in the other clusters and the standard deviation is not that big (for example double than cluster 2 but the mean is four times higher.

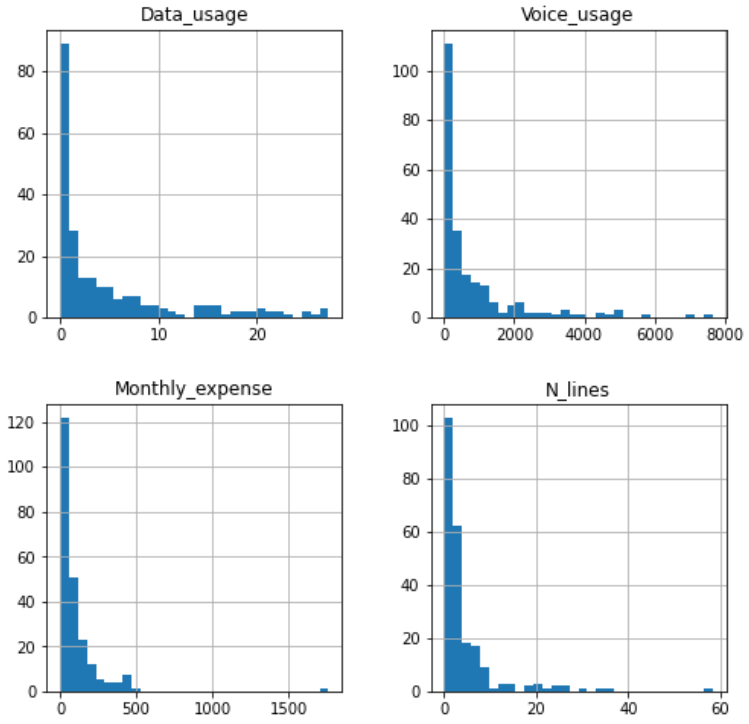
Annex

[EX 6] Histograms

- Original

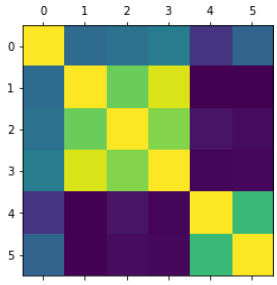


- Without 15%

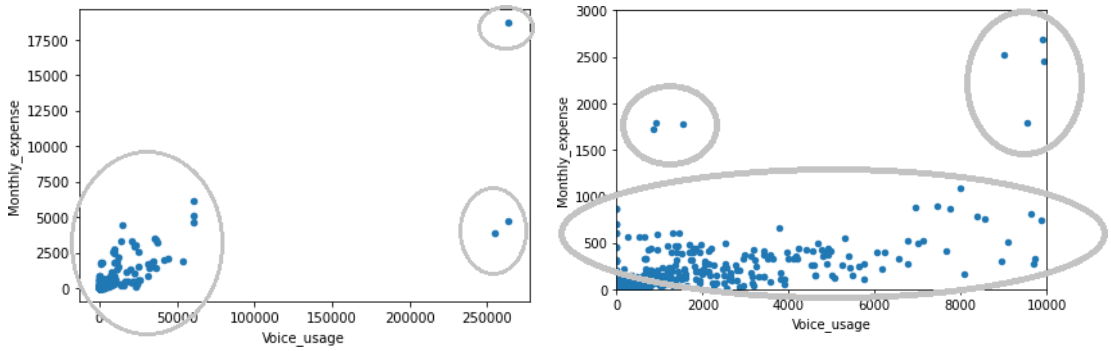


[EX7] Correlation matrices

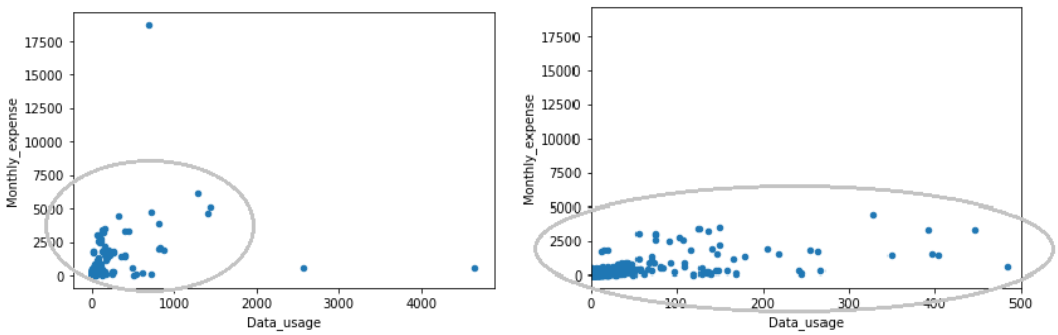
	Data_usage	Voice_usage	Monthly_expense	N_lines	Turnover	CNT_EMPLOYEE
Data_usage	1.000000	0.327104	0.355372	0.402734	0.126128	0.298305
Voice_usage	0.327104	1.000000	0.767026	0.940066	-0.034378	-0.032440
Monthly_expense	0.355372	0.767026	1.000000	0.810076	0.018335	0.001247
N_lines	0.402734	0.940066	0.810076	1.000000	-0.016893	-0.011199
Turnover	0.126128	-0.034378	0.018335	-0.016893	1.000000	0.663925
CNT_EMPLOYEE	0.298305	-0.032440	0.001247	-0.011199	0.663925	1.000000



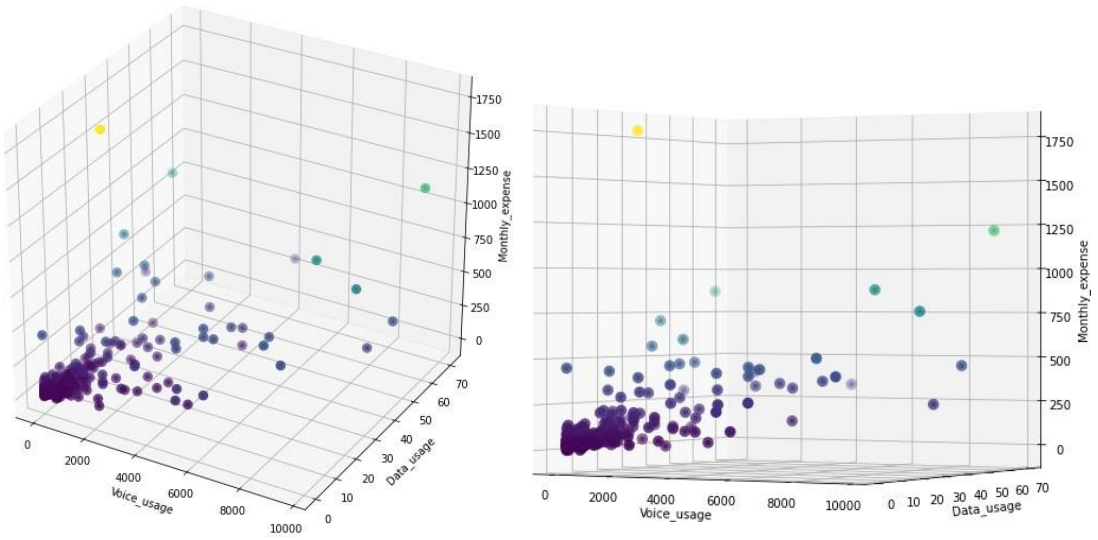
[EX8] Clusters



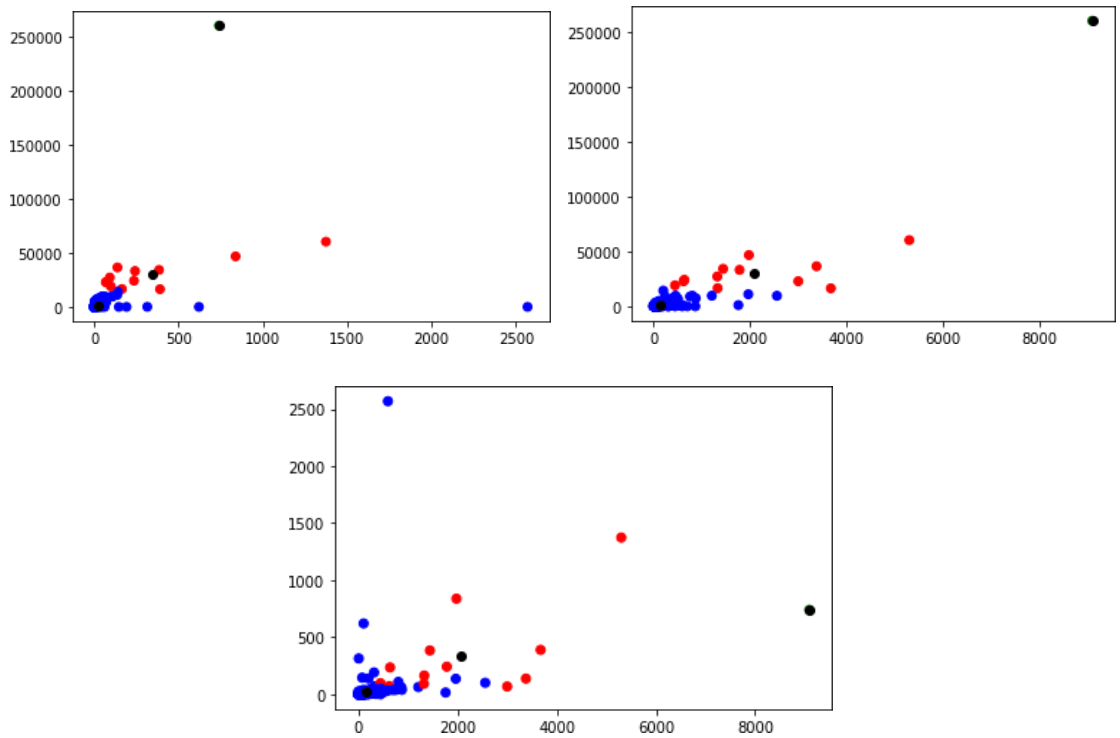
[EX9] Clusters



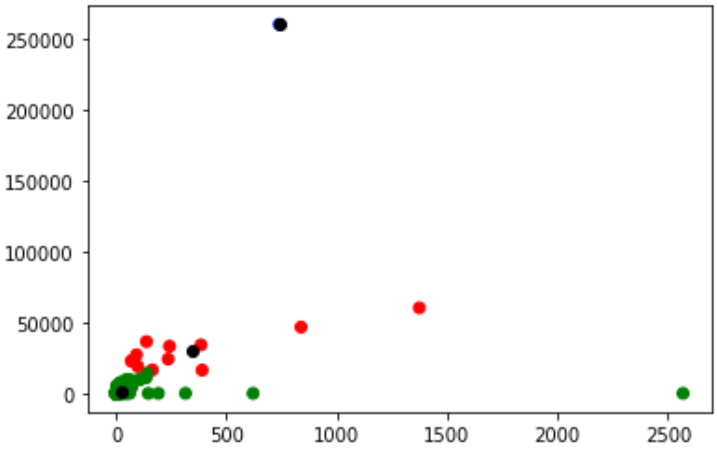
[EX10] 3D plots between Voice_usage, Data_usage and Monthly_expense



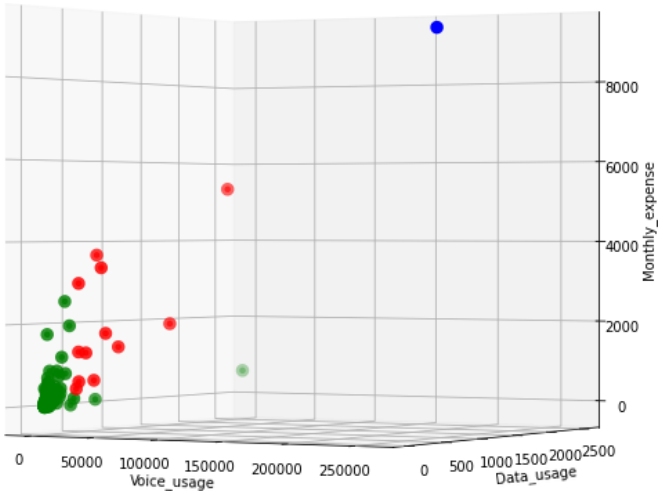
[EX14] Plots Data_usage vs Voice_usage, Monthly_expense vs Voice_usage and Monthly_expense vs Data_usage



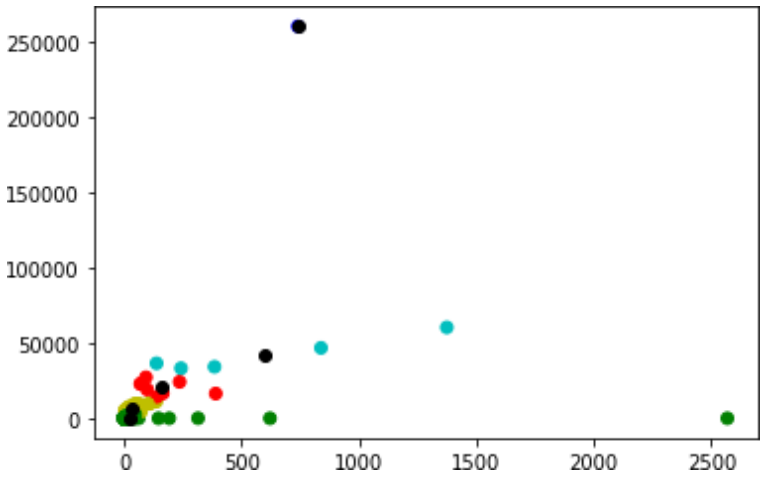
[EX15] Data_usage vs Voice_usage scatter plot



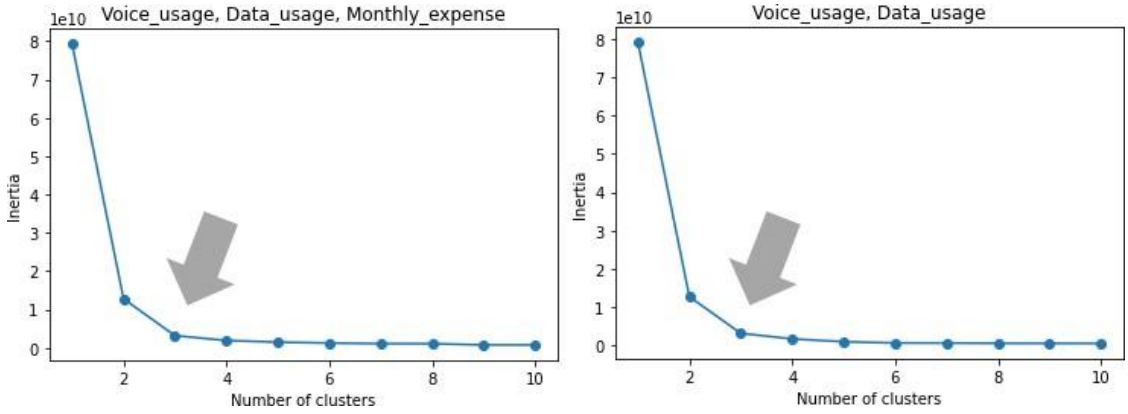
[EX16] 3D plot Voice_usage, Data_usage and Monthly_expense



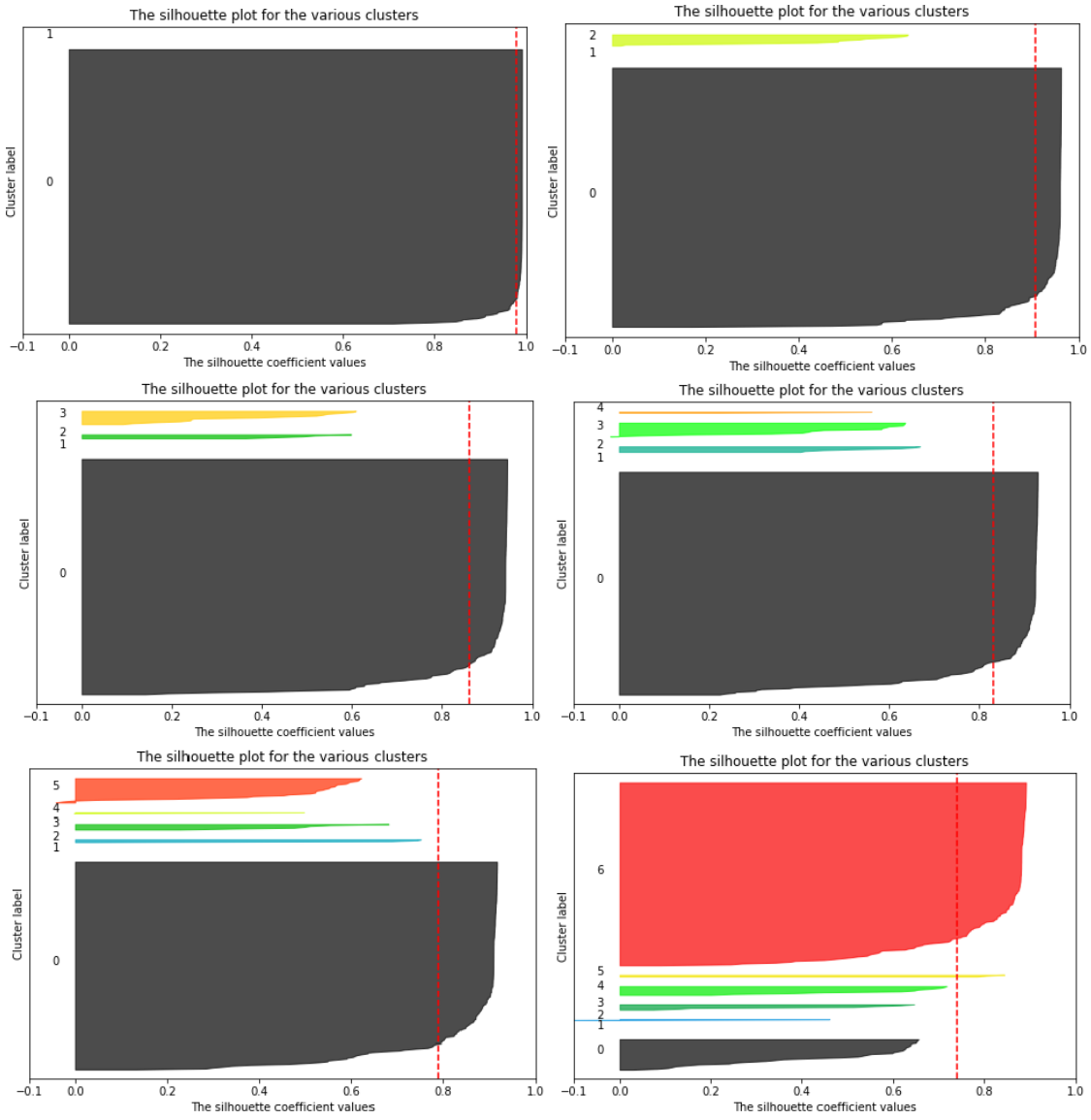
[EX17] K=5 Data_usage vs Voice_usage



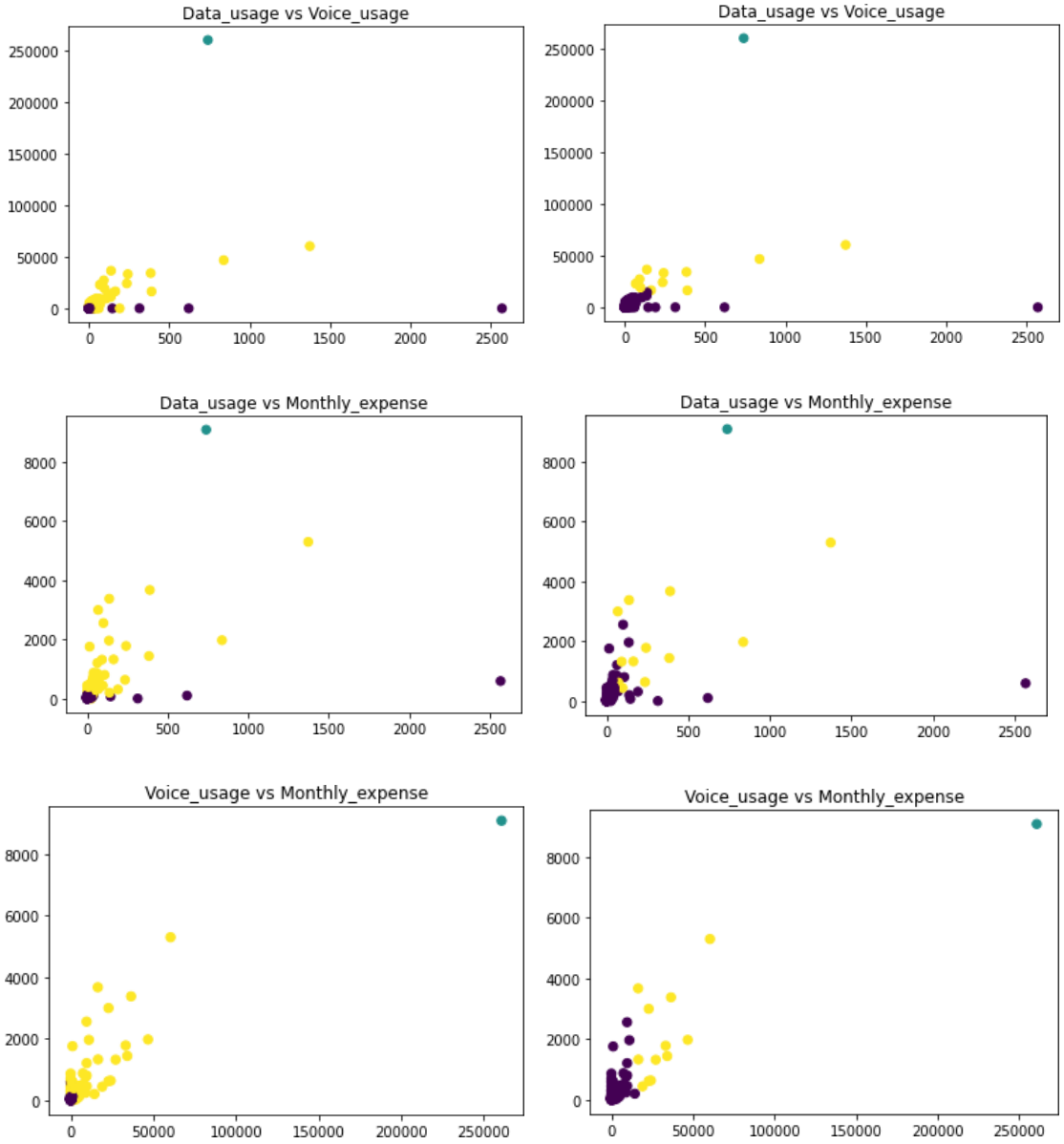
[EX18] Elbow method plot



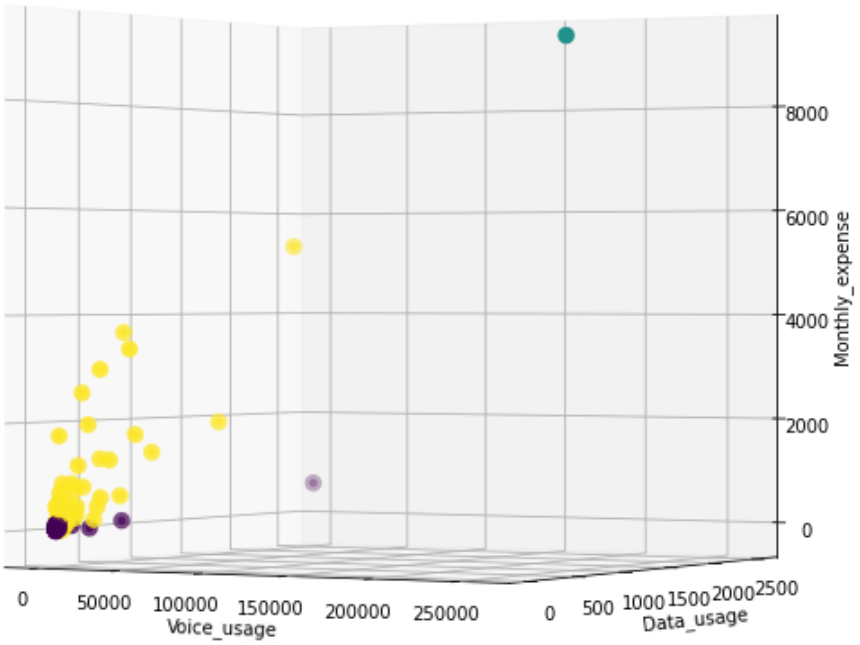
[EX19] Silhouette plots



[EX21] Plots Data_usage vs Voice_usage, Data_usage vs Monthly_expense and Voice_usage vs Monthly_expense (Mixture of Gaussians – K Means)



[EX22] 3D plot Voice_usage, Data_usage and Monthly_expense



[EX23] Silhouette plots for Mixture of Gaussians from 2 to 7 components