HOMEWORK 3 – CLUSTERING ANALYSIS

(Total points 10)

Cluster Analysis, Customer Segmentation

Business Context: Telecommunication Companies

Assume you work as a data scientist for a telco company (for example, Verizon, AT&T, etc.) and you are asked to analyze the dataset **TelcoData.csv** containing information about current customers of the company. You are asked to mine the dataset to segment customers into interesting groups, to provide the company with new, initial, ideas on how to update its plan offering, on which services could be bundled together, and so on.

Data Description

The data contains information about a telco company that provided home phone and Internet services to a sample of customers in California. Data Source: IBM.

The data provided to you refers to Q3 of a given year (undisclosed) and contains the following attributes:

- CustomerID: a unique ID that identifies each customer
- Age: The customer's current age, in years
- Senior Citizen: 1 = if the customer is 65 or older; 0 = otherwise
- Tenure.Months: Indicates the total amount of months that the customer has been with the company
- Multiple.Lines: 1 = Indicates if the customer subscribes to multiple telephone lines with the company; 0 = otherwise
- Unlimited.Data: 1 = indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads; 0 = otherwise
- Online.Security: 1 = Indicates if the customer subscribes to an additional online security service provided by the company; 0 = otherwise
- Online Backup: 1 = Indicates if the customer subscribes to an additional online backup service provided by the company; 0 = otherwise
- Device.Protection: 1 = Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company; 0 = otherwise
- Tech Support: 1 = Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times; 0 = otherwise
- Streaming TV: 1 = Indicates if the customer uses their Internet service to stream television programing. Currently, the company does not charge an additional fee for this service.
- Streaming Movies: 1 = Indicates if the customer uses their Internet service to stream movies. Currently, the company does not charge an additional fee for this service.
- Monthly.Charge: Indicates the customer's current total monthly charge for all their services from the company
- Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.
- Partner: 1 = Indicates if the customer lives with a partner; 0 = otherwise
- Parent_cat = Categorical version of the Partner. Yes = if the customer lives with a partner; No = otherwise

Analysis

Perform the following analysis in R.

Pre-Processing/Data Visualization (NOTE: for this part, you need to know how to perform basic descriptives and how to use ggplot. Review the R Tutorial Video if you do not know how to do this):

a) (0.4) Load the data and summarize only the attributes Age, Tenure.Months and Monthly.Charges and comment on their distribution in terms of mean, min and max.

b) (0.5) Use ggplot2 to produce a histogram for the attribute Tenure.Months. Your plot should have: proper title, proper axis, labels. Fill the histogram with a color of your choice. Comment on the distribution of the attribute.

c) (0.6) Use ggplot2 to produce a box-plot of Monthly.Charges, by Partner_cat (you need to use the categorical version of the Partner variable for this). Your plot should have: proper title, proper axis, labels. Comment on the distribution of Monthly.Charges for the two groups and identify whether there are any outliers. (It is ok if on your box plot do not appear specific labels/numbers for median, min, max, etc.)

d) (0.3) Assess which attributes/variables you should include in your cluster analysis and whether you need to exclude some. Justify your choices. (For this question, you do not need to code but you must explain your choices). Also: include only one of the attributes for Partner.

e) (0.2) Assess whether you need to normalize the data. If yes, explain why and create the function in R to implement min-max normalization and perform the normalization on the appropriate attributes (you may want to create a copy of your dataset). Simply report the code used. If you do not think normalization is necessary, explain why.

Hierarchical Clustering

f) (0.3) Generate a distance matrix using Euclidian distance and using only the attributes you chose to include in your cluster analysis, in point d). Report the code and a screenshot of the first 5 columns/rows of the distance matrix.

g) (0.2) Run hierarchical clustering using the distance matrix produced in point f) and using the Ward method. Only report the code for this question.

h) (0.3) Plot the dendrogram, make sure to use the following options: labels = FALSE, hang = 0. Attach the output to the answer. It is ok if the labels do not appear and the bottom looks a little cluttered.

i) (0.3) Use the rect.hlcust() function to draw on the dendrogram the best 4-clusters solution. The borders of each cluster should have a different color. Report the code and the plot.

j) (0.2) Cut the dendrogram into the 4 clusters and report the size of each cluster.

k) (1) Add a column to the original dataset with the cluster's number to which each customer belongs to. Use the function ddply() to get the means, for each cluster, for the attributes Monthly_Charges, Tenure.Months, Age; and the proportions of the binary attributes Partner, Senior.Citizen, Online.Backup, Tech.Support, Streaming.Movies, Streaming.TV, Online.Security, Unlimited.Data. To get the proportions for the binary attributes, first, you need to get the total number of observations, n = length (CustomerID); then, you can sum the binary variable and divide it by the total number. Example:

ddply(*data*, *clusterColumn,* summarize, n = length (CustomerID), Partner = sum(Partner)/n, …… and so on, including all the binary variables and also the mean for the numerical variables).

For this question, simply copy and paste the code and the table obtained.

K-Means Clustering

Next, you decide to perform k- means clustering. First, set your seed to be 123.

l) (0.5) Run four different versions of k-means clustering, with centers 2, 4, 6 and 8 and nstart = 15. Make sure to use the same attributes you used for hierarchical clustering and decide whether to use the normalized data. For this question report the code only.

m) (0.7) Next, use the fviz_cluster function to get a visual representation of the four different version of k-means clustering, and plot the four of them, together, in a grid. Report the graphs and by simply looking at them, pick which one looks more appropriate to you and explain why.

n) (1) Find what you think is the most appropriate number of clusters by computing the WSS (only) and plotting the Elbow plot. Use a loop to run k-means for 10 times, and set nstart = 10. When you plot the results, make sure the y-axis starts from 0. Report the code, the plot obtained and explain what you think it may be an appropriate number of clusters. You need to motivate your choice.

o) (0.5) Finally, run again k-means using the number of clusters you decided in point n) and nstart = 10. Add a column to the original dataset with the cluster's number to which each customer belongs to. Use the function ddply() to get a summary, for each cluster, of the same attributes as in point k). Attach the table obtained to the answer.

Produce Insights

After you complete the analysis, you need to interpret the results obtained for the clusters so to provide your company with an initial direction on how to update its plan offering, which services could be bundled and offered together and so on.

(0.3) First, you need to decide whether to use the results from Hierarchical Clustering or the results from K-Means. For this, compare the tables produced in point k) and o), and decide which clustering results, in your opinion, are more meaningful and are going to lead to newer, interesting ideas. Explain briefly your choice.

(1.5) Once you have chosen which results to focus on (that is, either the table produced in point k) or in o), explain, in details, how the individual clusters contained in your chosen solution differ from each other, in terms of attributes.

(1.2) Finally, use such analysis to propose at least 2 interesting plans / bundles your company could offer. Try to be specific in your answer.

For example, let us assume among your attributes you have Age and Minutes of Calls. You find a cluster that distinguish itself for a very high average Age and high average Minutes of Calls, while the others clusters have similar averages for age and Minutes of Calls. This may indicate that there is a segment of customers that tend to include more senior individuals and that use intensively phone calls, compared to other services. As such, the company could think of introducing a Plan for Seniors, that targets people above a certain age and that only includes unlimited phone-calls (while any other service would need to be paid by consumption, for example).