FAQ - AllLife Bank Customer Segmentation

1. How should one approach the AllLife Bank Customer Segmentation project?

Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric.

Once you understand the task, download the dataset and import it into a Jupyter notebook to get started with the project.

To work on the project, you should start with data preprocessing and EDA using descriptive statistics and visualizations.

Once the EDA is completed and data is preprocessed, you can use clustering algorithms to do customer segmentation and analyze these segments to gain insights.

It is important to close the analysis with key findings and recommendations to the business.

2. We only have a handful of columns. Do we need to do feature extraction for this project?

Since there are very few columns in the dataset, feature extraction using PCA is not mandatory.

3. I don't understand how to do cluster profiling. Please help.

Cluster profiles inform you about the characteristics of different variables with respect to each cluster.

This can be done by first grouping the data based on the clusters formed, and then checking the summary statistics of the variables or visualizing the variable distributions for each group to gain insights.

(**Hint:** You can check the mean, median, and boxplot of each variable w.r.t clusters)

4. I don't know what to do with outliers for clustering. Please help.

There is no one answer with regards to outliers and it depends upon several factors.

You can perform your analysis with and without outliers and see if there is any difference in the results. Sometimes there is no significant difference and in that case, there is no need to treat outliers. However, there may be a scenario where outliers can form separate segments which could be useful for the business.

You can also explore different distance measures, like Manhattan distance, which minimizes the effect of outliers.

5. I am unable to use silhouettevisualizer to visualize agglomerative clustering because there is no "predict" method for AgglomerativeClustering. I get the following error.

```
"AttributeError: 'AgglomerativeClustering' object has no attribute 'predict'"
```

Any suggestions on how to plot the silhouette scores?

The *silhouettevisualizer* of the yellowbrick library is only designed for k-means clustering. It does not support HC algorithms. Also, there is no such open-source library to make a similar plot of silhouette score for HC algorithms.

We would suggest you calculate the silhouette score for different numbers of clusters obtained from hierarchical clustering and make a simple line chart to show where is the highest silhouette score. This would help you to decide how many clusters should be made.

6. The column 'Customer Key' contains duplicate values. What does the column represent and how to deal with these duplicate values??

The 'Customer Key' is a unique ID given to each customer in the database. The duplicate values might correspond to customer profile changes, and as such, there is no need to delete these records as these are actual occurrences at some point in the time. The column can be dropped during the analysis.

7. How do we compare K-means clusters with Hierarchical clusters?

You compare several things, like:

Which clustering technique took less time for execution?
Which clustering technique gave you more distinct clusters, or are they the same?
How do the silhouette scores vary?
How many observations are there in the similar clusters of both algorithms?
How many clusters are obtained as the appropriate number of clusters from both algorithms?

You can also mention any difference you obtained in the cluster profiles from both the clustering techniques.

8. I am not able to load the yellowbrick library and getting the following error:

```
ModuleNotFoundError: No module named 'yellowbrick'
```

Can you help?

If the yellowbrick library is already installed, then the following command can be run in the Anaconda prompt:

```
pip install -U yellowbrick
```

9. I have installed the yellowbrick package and updated it to the latest version. But I am still having issues importing the KElbow visualizer and getting the following error:

```
ImportError: cannot import name 'safe_indexing' from 'sklearn.utils'
```

How to fix this?

This error generally occurs due to version differences between sklearn and yellowbrick. Please update the sklearn library by running the below code in the Anaconda prompt:

```
pip install -U scikit-learn —user
```