**Market Segmentation using Cluster Analysis**

We are using market data to understand the clustering profiles that can be done using clustering analysis.

- Firstly, we have loaded the data given in an excel file

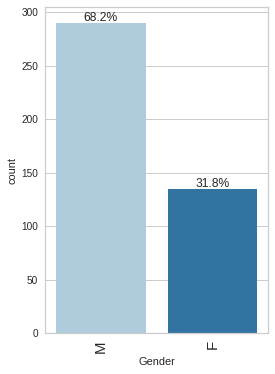
- There are about 425 rows and 10 columns in the data

- There are no null values present in the data so missing value treatment would not be required

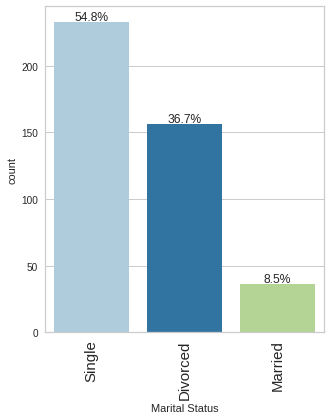
- We can observe that the mean current account values rise about 1000 and mean savings account stay around 1800 which shows money in savings accounts are usually more than current account and most of the people using banks are middle aged people with a mean age of 34.

**EDA**

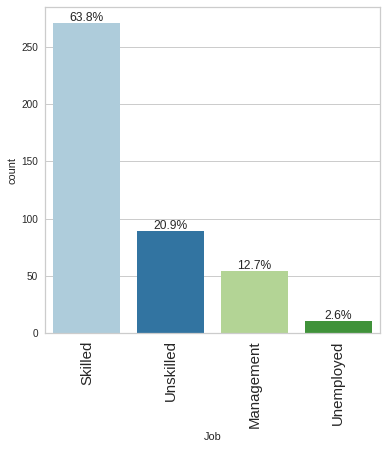
- We can observe from barplots that the number of male customers are twice as more as female customers



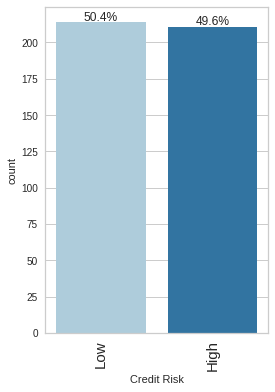
- And most of the customers are single which also shows that most males can be bachelors



- Most of the people are also skilled which means they are all employees maybe software or government officials



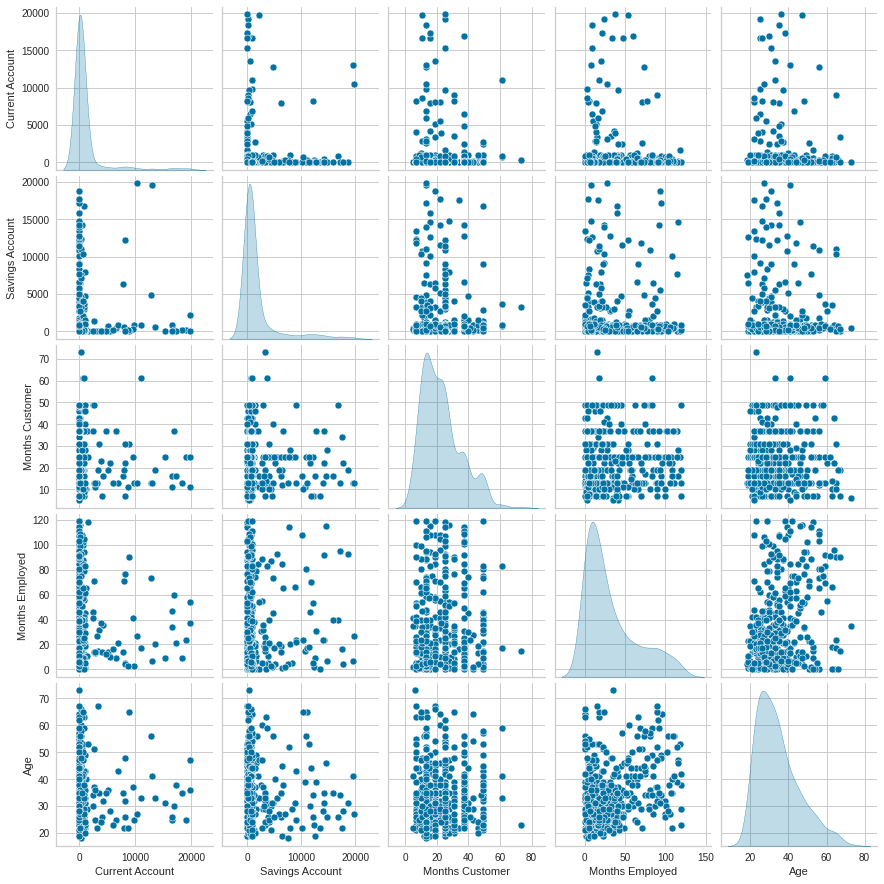
- There is a equal distribution between high and low credit risk



- From the correlation heatmap, we can observe that most of the columns are not correlated but there is a slight correlation between months employed and age which needs to be taken into consideration.

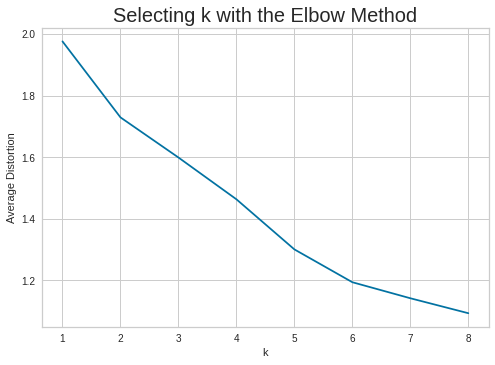


- Also from the pairplot it can be observed that the values of most columns are right skewed so standardization or normalization can be applied

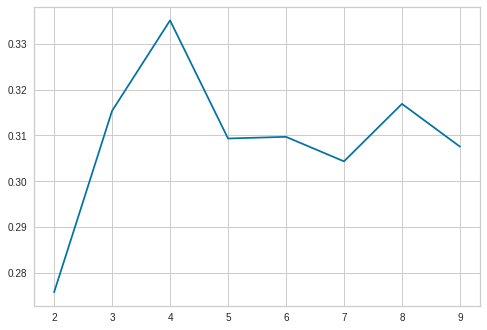


**K-means**

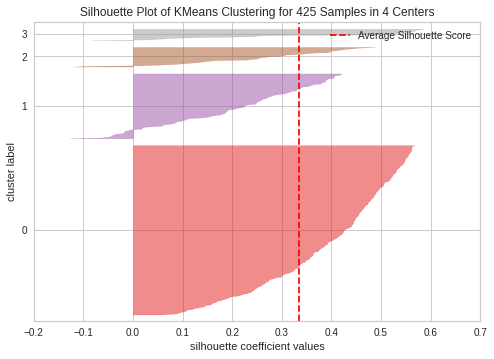
- while fitting the data on the K-Means algorithm, we can observe that the appropriate values of K seem to be 4 or 5.

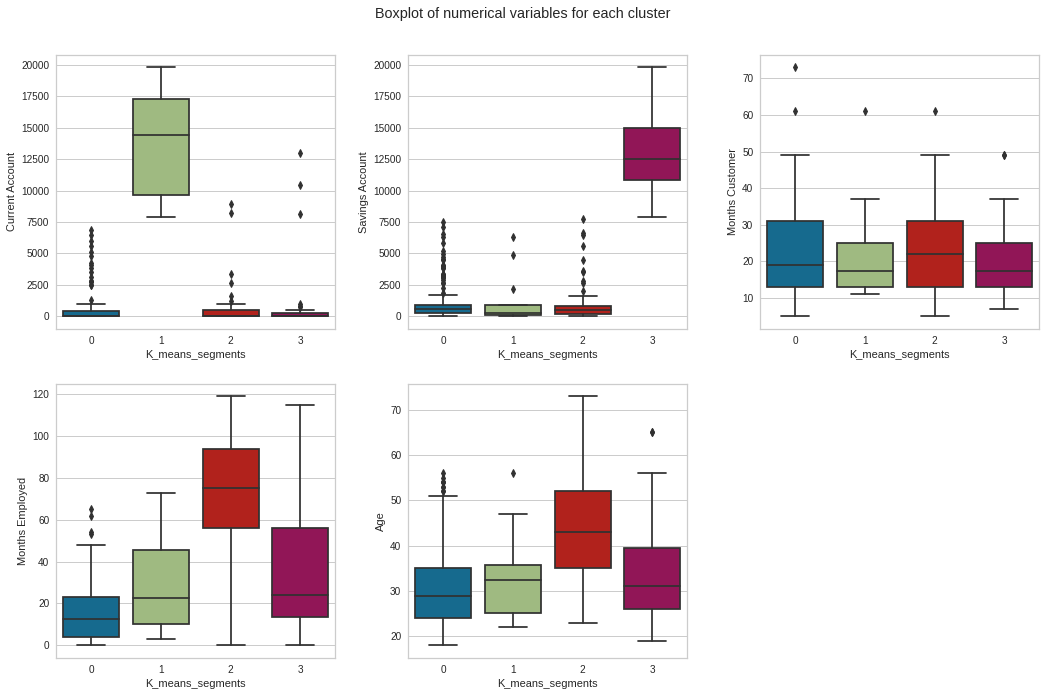


- We will try to confirm the k-value using the silhouette score and we observe that k=4 is the best value



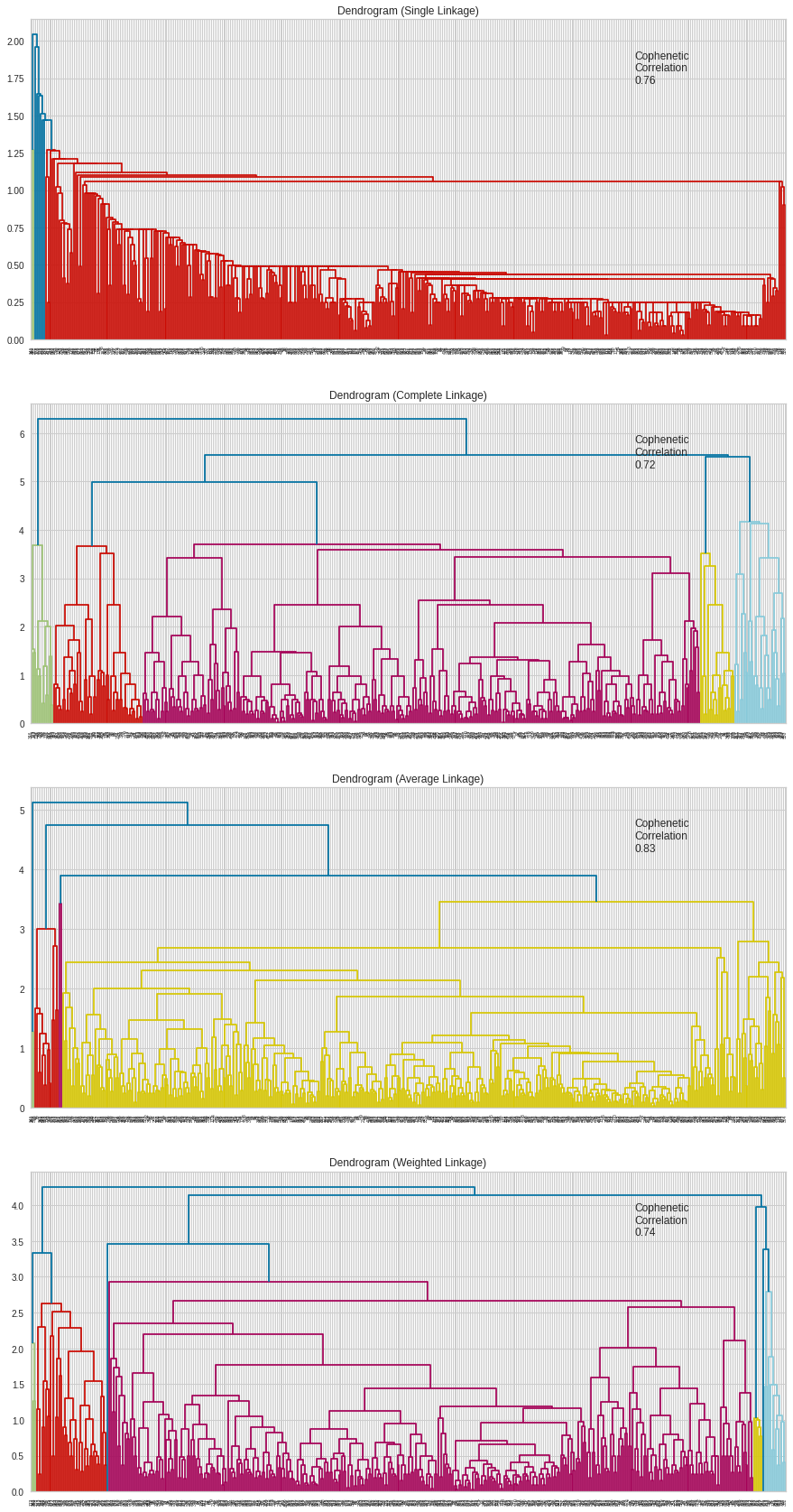
- We have also visualized the values using silhouette visualizer and we have also performed cluster profiling for each cluster and observed the numerical variables for each cluster and observed that the segment corresponding to months cluster seemed the best.



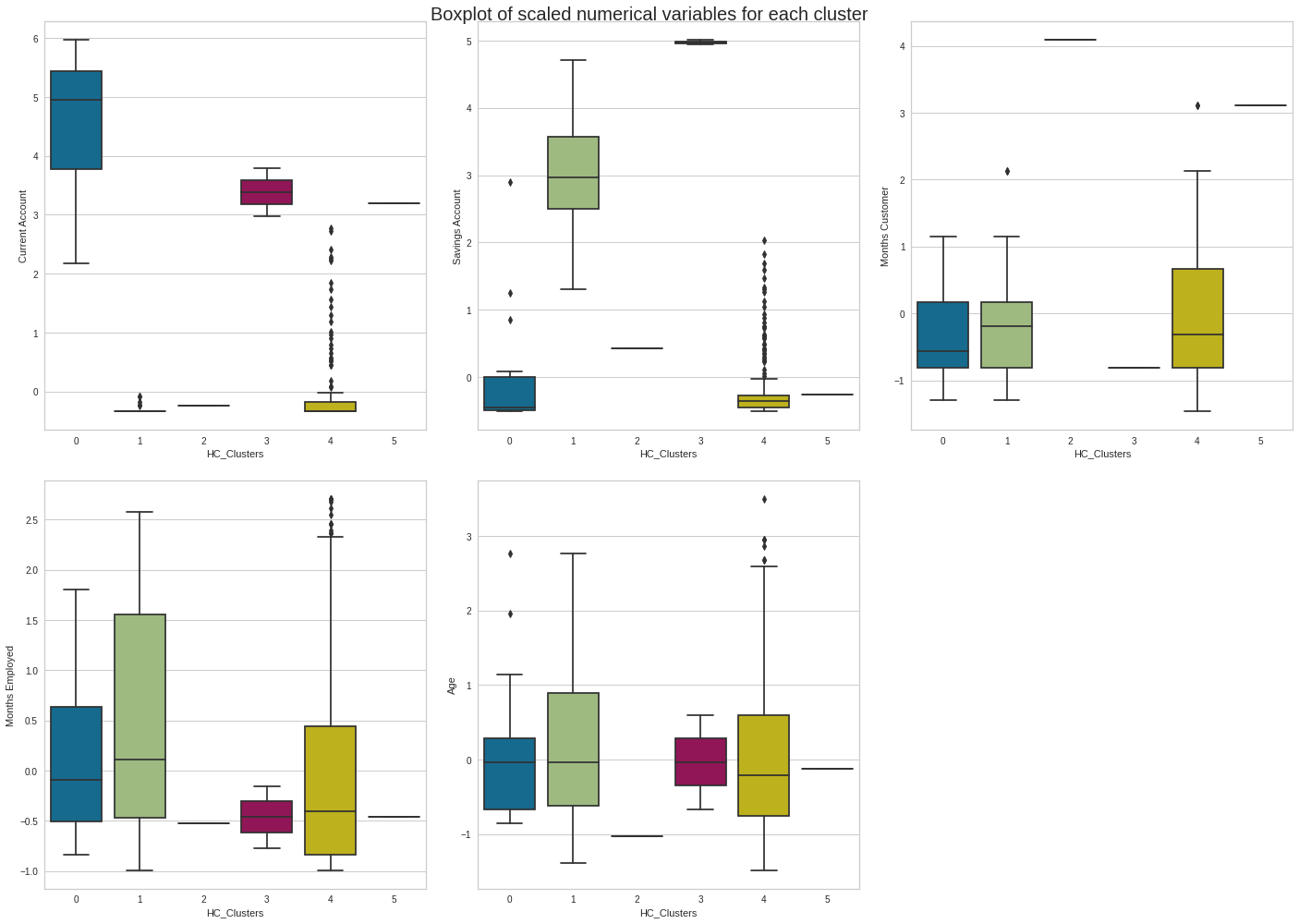


**Hierarchical Clustering**

- We built the hierarchical clustering using the 4 different distance metrics - "euclidean", "chebyshev", "mahalanobis", "cityblock" and 4 linkages - "single", "complete", "average", "weighted" and we found the Highest cophenetic correlation as 0.8319055185745458 which is obtained with Chebyshev distance and average linkage



- We also visualized dendograms and also obtained clustering profiles for the same



- Further we also tried PCA for dimensionality reduction to reduce the dimensions to two components and observe them and we have clearly observed the three components differently. We can observe all the components distinguishably except the third and fifth and second. SO they can even be outliers.

