**Problem**: The previous email campaign have not been very effective in converting sales. A general message are sent to all the customer in the email list while only few customer are responding to the email campaign.

**Solution**

The Analysts believes for some reason that some customer in the email list have different characteristics while others are not that so different from each other. She believes that if the customer with same characteristics are group together and those with different characteristics are in different group, then suitable marketing message could be sent to the different group that are tailored to their needs. This proposed solution is known as customer segmentation

What is Customer Segmentation?

Customer Segmentation is the processing of separating customers into different segment based on their common characteristics and features. This is important because companies can then markets effectively to these different groups.

Customer can be segmented based on Age, Gender, Marital Status, Location , Life stage for a business to Customer Marketing while for a business to business marketing customer can be segmented based on Industry, Numbers of Employee, previously purchased products and Location.

KMEANS

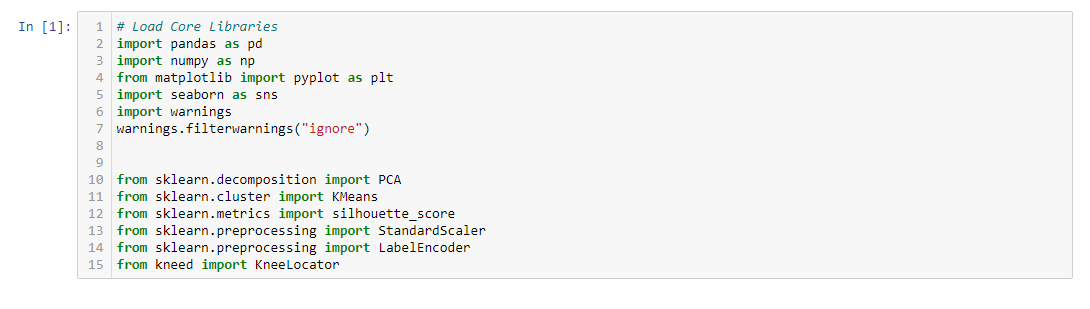
Kmeans is a partitioning method non-hierarchical clustering algorithm , it is less complex and easy to implement when compared to others clustering algorithm. It is an iterative process where you are trying to minimize the distance of the data point to the average data point in the cluster.

K means Algorithm Consist of the following steps

* Partition the items into numbers of clusters
* Determine the center
* Assign point to the cluster which are outside of the cluster according to the distance between the center and the point.
* Calculate the new center
* Repeat the steps until you obtained the desired cluster.

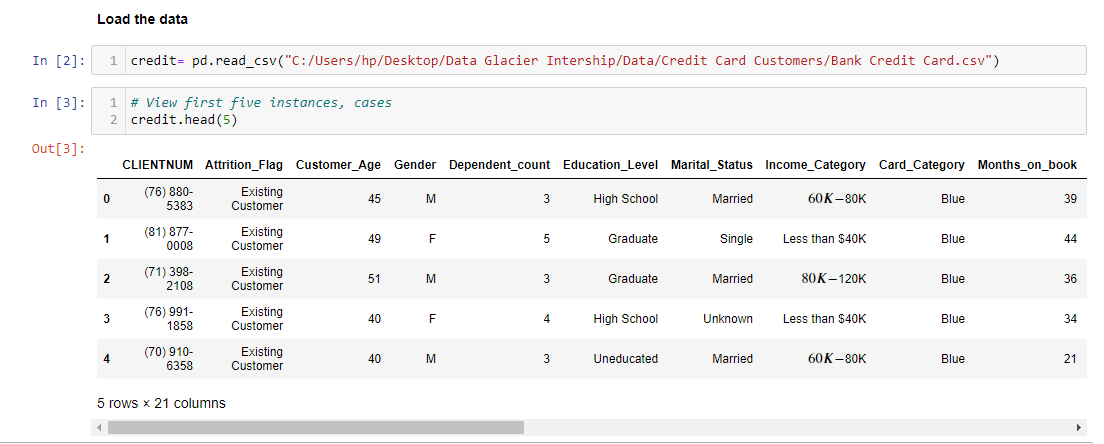
Python Codes

The following packages were used for the KMeans Clustering for Customer Segmentation.



**Data**

The dataset used was sourced from Kaggle.

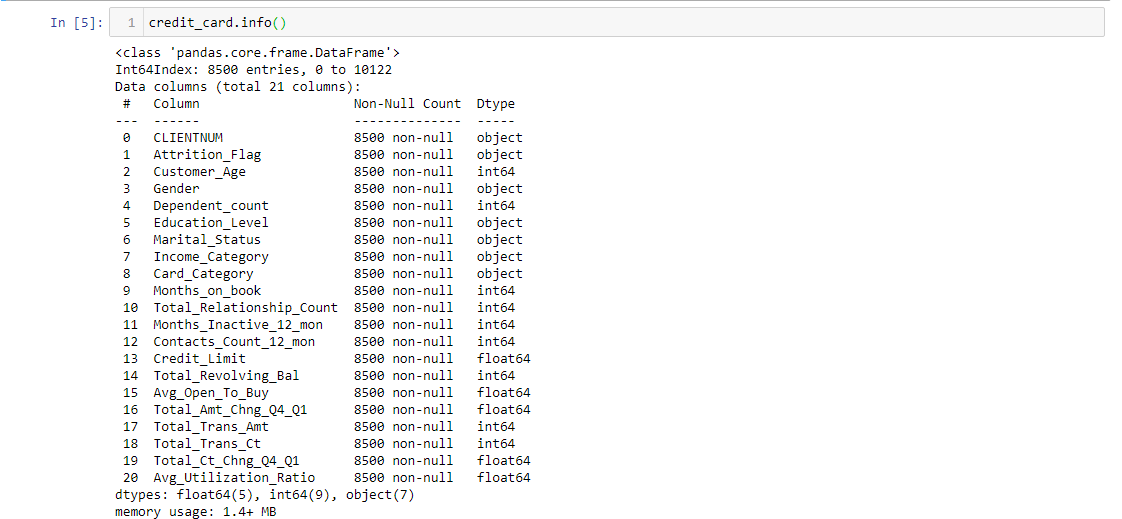


The data contain information about both existing customer and customer who have left the organization. We selects only existing customer for this customer segmentation analysis.

**Exploratory Data Analysis and Data preprocessing**

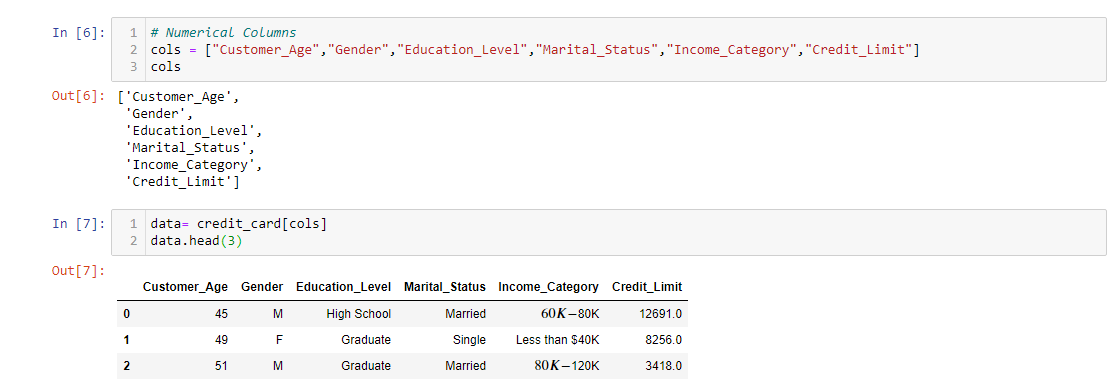
Next we check for missing values, we also check if the variables are in the proper format, we perform EDA on the variables to draw some useful insights. On Numerical variables we calculative the summary statistics, we plot the histogram to give us the distribution of the variables and the boxplot to visualize some summary statistics.

**Data Preprocessing**



There is no missing values in the dataset, the variables are also in their correct format.

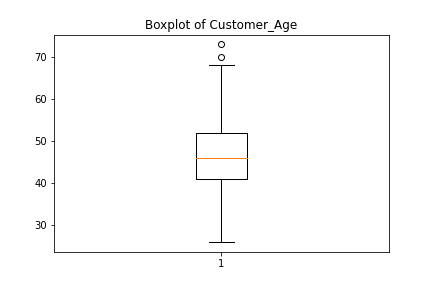
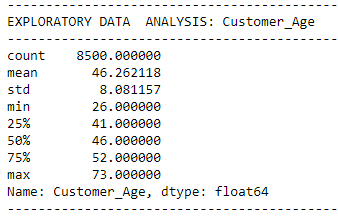
We will be selecting the Customer\_Age, Gender, Education\_Level, Marital\_Status, Income\_Category, Credit\_Limit variables for this analysis, this is because we are doing a business to customer marketing.

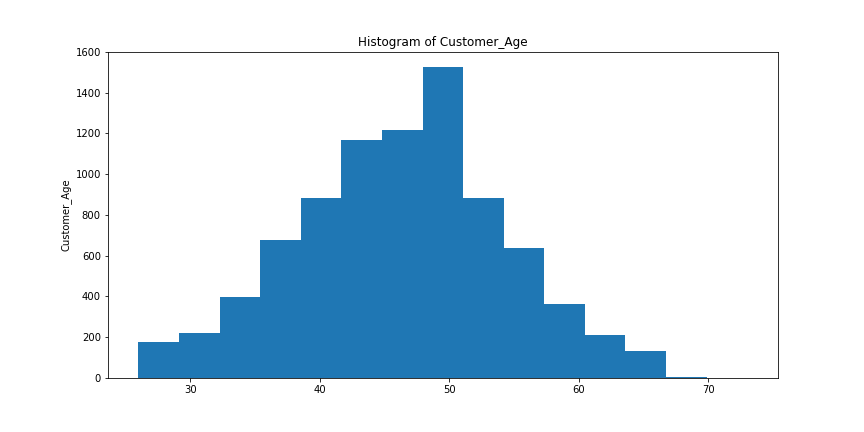


**EDA**

The descriptive summary statistics and the boxplot was performed for the numerical variable while the frequency count and the barchart was used to visualized categorical variables.

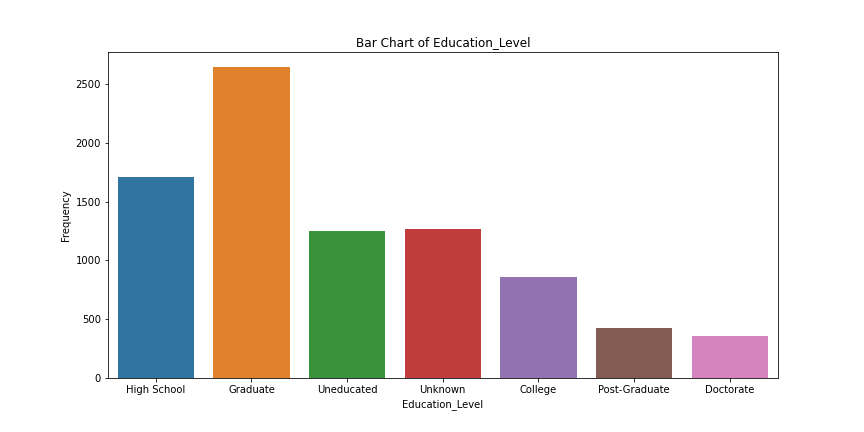
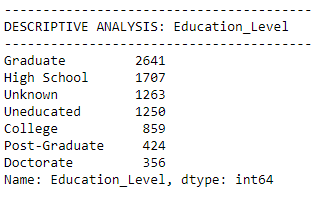
**Customer age** is a numerical variable





The customer age follows approximately the normal distribution, although it is a little bit tailed to the right. The average age of customer in the organization is approximately 46 years.

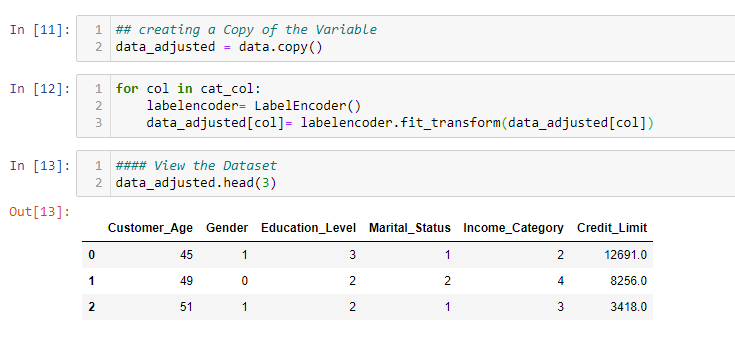
The **Education Level** of Customer is a categorical variable, therefore the frequency count and the barchart was used.



The vast majority of the customer are graduate and high school student, while few of the customer have Post graduate and a doctorate degree, a good chunk of the customer are uneducated.

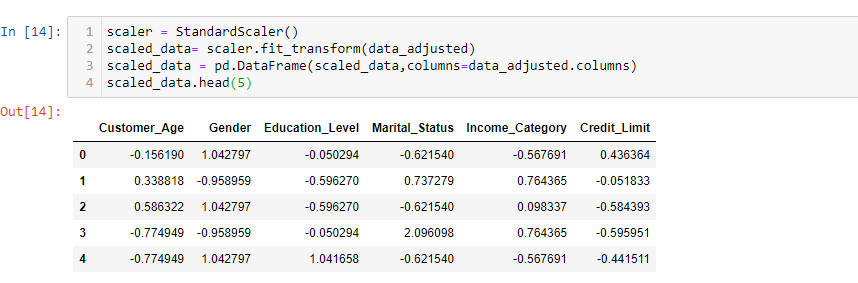
**Label Encoding**

The clustering algorithm will not work optimally with the categorical variables, therefore all the categorical variables in the data was encoded into numerical variables.



**Feature Scaling**:

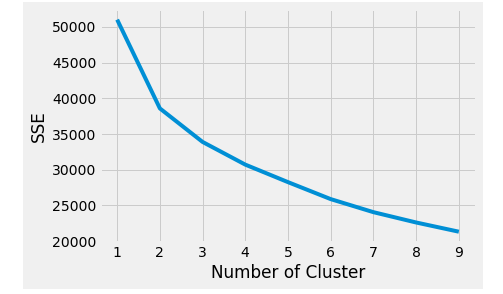
The Kmeans algorithm uses Euclidean for calculating the distance between two points, this is a problem because the algorithm will be influence more by the variables with a high magnitude. Therefore to suppress this effects the dataset set is scaled. Standardization will perform on the dataset to transform it.



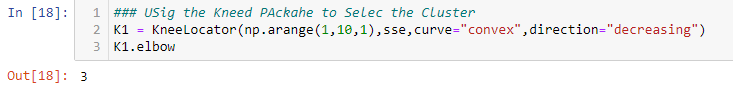
**Selecting K**

The elbow method will be used in selecting the K number of clustering for our customer segmentation.



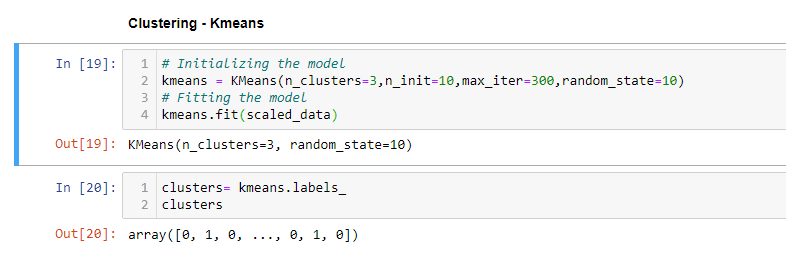


To select the optimal k, we choose the value of k in the elbow, from the data we can conclude that the choice of k is 3.Alternatively, KneeLocator from the kneed package can be used to select the optimum value of k.



The optimal value of k is 3. Therefore, 3 clusters can be generated from the dataset.

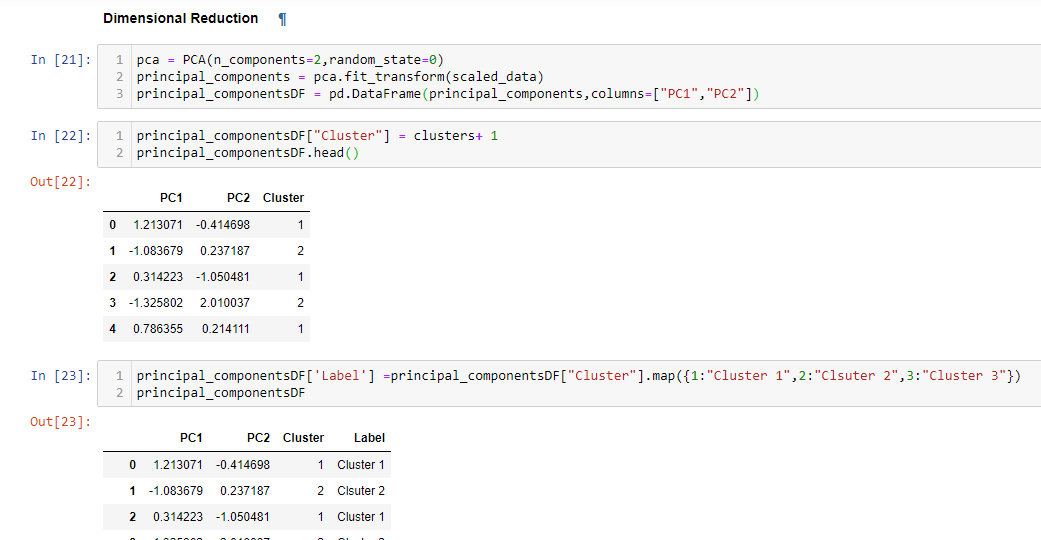
**Clustering**



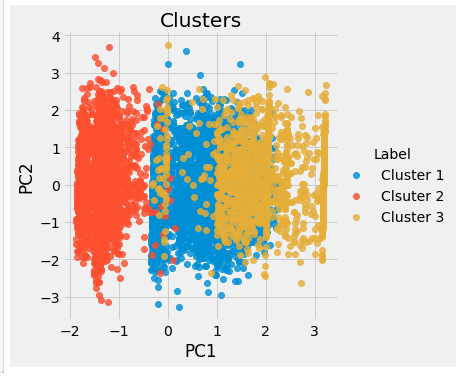
The data point was cluster into three group. The data point in the same clusters have similar characteristics while the data point in the different clusters have a different characteristics.

**Visualizing the Cluster on the X-Y Plane**

The Principal Component analysis will be used for dimensional reduction into two plane (PC1 and PC2). Then we visualized the PC1 and PC2 as a scatter diagram and then identify for the clusters.

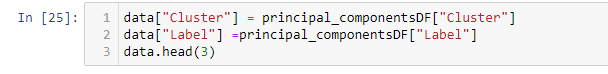


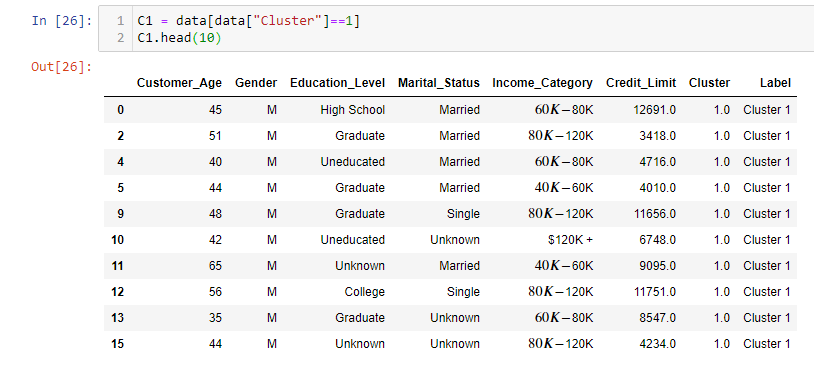
Plotting on the X-Y Plane



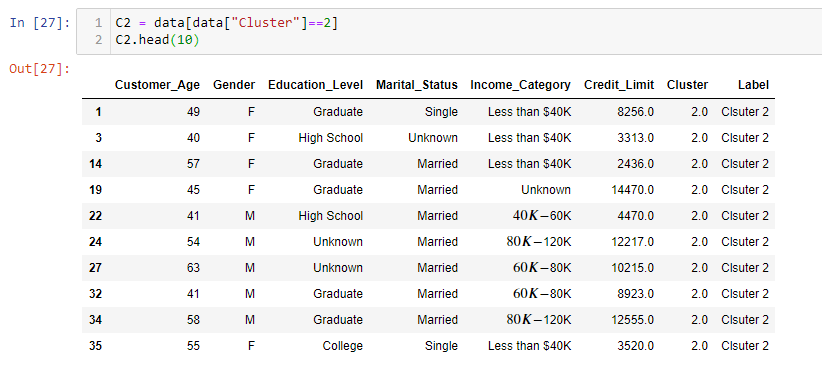
Selecting Data Point in Each Cluster.

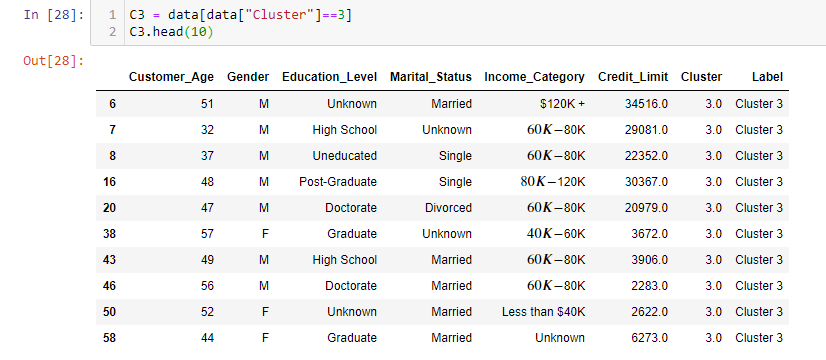
**Cluster One**

****

****

**Cluster two**

**Cluster Three**

****

**Thank you.**