

**School of InfoComm Technology**

**Applied Analytics Assignment**

Diploma in Cybersecurity & Digital Forensics

Diploma in Financial Informatics

Diploma in Information Technology

Year 2/3 (20212/2023), Semester 3/5

**INDIVIDUAL ASSIGNMENT 1**

(30% of Applied Analytics Module)

**Deadline for Submission:**

**17 June 2022 (Friday), 23:59 HRS**

|  |  |
| --- | --- |
| **Tutorial Group:** | **T03** |
| **Student Name:** | **Seo Shin Youn** |
| **Student Number:** | **S10205100K** |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 24th Jun 2022, 23:59.

**Table of Contents**

[Summary Overview 4](#_Toc106402586)

[Bank Dataset to be analyzed: 5](#_Toc106402587)

[Build Clustering Models using Numerical Data 6](#_Toc106402588)

[Data exploration and Manipulation on numerical data 6](#_Toc106402589)

[Exploratory Data Analysis 6](#_Toc106402590)

[Identifying Null Values in the Bank Dataset (I) 6](#_Toc106402591)

[Identifying Null Values in the Bank Dataset (II) 7](#_Toc106402592)

[Identifying Count and Unique Values of Each Feature 8](#_Toc106402593)

[Potential Irrelevant Features to Drop 8](#_Toc106402594)

[Potential Categorical Features to Drop 9](#_Toc106402595)

[Analyzing Distribution of Numerical Features 11](#_Toc106402596)

[Data Manipulation - Data Cleaning 12](#_Toc106402597)

[Dropping Irrelevant Features 12](#_Toc106402598)

[Dropping Nominal Categorical Features 12](#_Toc106402599)

[Data Manipulation - Data Transformation 13](#_Toc106402600)

[Replacing Column Name Spaces with Underscores 13](#_Toc106402601)

[Transforming Column to the same metric of comparison 13](#_Toc106402602)

[Distribution of the ‘Duration in year’ column before transformation 14](#_Toc106402603)

[Distribution of the ‘Credit amount’ column before log transformation 16](#_Toc106402604)

[Distribution of the ‘Credit amount’ column after log transformation 17](#_Toc106402605)

[Distribution of the ‘Age in years’ column before log transformation 18](#_Toc106402606)

[Distribution of the ‘Credit amount’ column after log transformation 19](#_Toc106402607)

[Label Encoding Ordinal Categorical Variables 20](#_Toc106402608)

[Label Encoding ‘Account Status’ Column 20](#_Toc106402609)

[Label Encoding ‘Credit History’ Column 20](#_Toc106402610)

[Label Encoding ‘Present Employment Since’ Column 21](#_Toc106402611)

[Label Encoding ‘Other Debtors or Guarantors’ Column 21](#_Toc106402612)

[Label Encoding ‘Housing’ Column 21](#_Toc106402613)

[Label Encoding ‘Telephone’ Column 21](#_Toc106402614)

[Label Encoding ‘Foreign Worker’ Column 21](#_Toc106402615)

[Building, Optimizing, and Performing Clustering Analysis on Clustering Models 21](#_Toc106402616)

[K-Means Clustering Model 22](#_Toc106402617)

[Basic K-Means Model 1A (Age and Credit Amount) 22](#_Toc106402618)

[Optimized K-Means Model 1A (Age and Credit Amount) with Silhouette Analysis 23](#_Toc106402619)

[Basic K-Means Model 2A (Age, Credit Amount, and Duration in Years) 25](#_Toc106402620)

[Optimized K-Means Model 2A (Age, Credit Amount, Duration in Years) with Silhouette Analysis 26](#_Toc106402621)

[Basic K-Means Model 3A (Age, Present Employment Since, Credit History) 28](#_Toc106402622)

[Optimized K-Means Model 3A (Age, Present Employment, Credit History) with Silhouette Analysis 29](#_Toc106402623)

[Basic K-Means Model 4A (Credit Amount, Number of Existing Credits, Credit History) 31](#_Toc106402624)

[Optimized K-Means Model 4A (Credit Amount, Number of Existing Credits, Credit History) with Silhouette Analysis 32](#_Toc106402625)

[Basic K-Means Model 5A (Account Status, Housing, Credit\_amount) 34](#_Toc106402626)

[Optimized K-Means Model 5A ( (Account Status, Housing, Credit\_amount) with Silhouette Analysis 35](#_Toc106402627)

[Hierarchical Clustering Models 37](#_Toc106402628)

[Basic Hierarchical Clustering Model 1B (Age and Credit Amount) 37](#_Toc106402629)

[Optimized Hierarchical Clustering Model 1B (Age and Credit Amount) 38](#_Toc106402630)

[Basic Hierarchical Clustering Model 2B (Age, Credit Amount, and Duration in Years) 39](#_Toc106402631)

[Optimized Hierarchical Clustering Model 2B (Age, Credit Amount, Duration in Years) 40](#_Toc106402632)

[Basic Hierarchical Clustering Model 3B (Age, Present Employment Since, Credit History) 41](#_Toc106402633)

[Optimized Hierarchical Clustering Model 3B (Age, Present Employment, Credit History) 42](#_Toc106402634)

[Basic Hierarchical Clustering Model 4B (Credit Amount, Number of Existing Credits, Credit History) 43](#_Toc106402635)

[Optimized Hierarchical Clustering Model 4B (Credit Amount, Number of Existing Credits, Credit History) 44](#_Toc106402636)

[Basic Hierarchical Clustering Model 5B (Account Status, Housing, Credit Amount) 45](#_Toc106402637)

[Optimized Hierarchical Clustering Model 5B (Account Status, Housing, Credit Amount) 46](#_Toc106402638)

[Evaluation and Comparison of the Models 47](#_Toc106402639)

[Summary and Interpretation 48](#_Toc106402640)

[Table Summary 48](#_Toc106402641)

[K-Means Clustering Model 1A 48](#_Toc106402642)

[K-Means Clustering Model 2A 49](#_Toc106402643)

[K-Means Clustering Model 3A 50](#_Toc106402644)

[K-Means Clustering Model 4A 51](#_Toc106402645)

[K-Means Clustering Model 5A 52](#_Toc106402646)

[Interpretation of Each Cluster 53](#_Toc106402647)

[K-Means Clustering Model 1A 53](#_Toc106402648)

[K-Means Clustering Model 2A 54](#_Toc106402649)

[K-Means Clustering Model 3A 56](#_Toc106402650)

[K-Means Clustering Model 4A 58](#_Toc106402651)

[K-Means Clustering Model 5A 60](#_Toc106402652)

# Summary Overview

The purpose of this report seeks to provide readers a comprehensive understanding of how various clustering models could be utilized effectively to generate business insights and drive informed decision making in the context of credit-related data of banking customers. Hence, by the end of this report, readers would be able to better understand the stages of pre-processing data in preparation of building the clustering models, building and optimizing the various clustering models, before finally evaluating and interpreting them. The report would thus be covering the following stages:

**Exploratory Data Analysis**

In this stage, we will be analyzing the bank customer dataset through basic data visualizations to better understand potential issues that may arise when building the various K-Means and Hierarchical Clustering models. Thus, the process would allow us to identify actionable procedures to take when we perform data cleaning and data transformation in the subsequent stages. The procedures include identifying missing values, potential features to drop, and having an overview of the distribution of all the numerical features from the Bank\_Data.csv dataset to understand if the utilization of a normalization technique would be required.

**Data Manipulation**

In this stage, we will be performing data cleaning and data transformation on our dataset to prepare our dataset for effective clustering. Hence, this includes dropping irrelevant and nominal categorical data, replacing feature name spaces with underscores, converting features to the same metrics of comparison with other features, applying normalization techniques on our dataset, alongside label encoding ordinal categorical data to utilize them for clustering purposes.

**Building Clustering Models**

In this stage, we will be focusing on utilizing both K-Means Clustering and Hierarchical Clustering models to identify insights from the various features available in Bank\_Data.csv. Hence, by utilizing the common performance measure of Silhouette Score which takes into account the average intra-cluster distance (i.e. distance between each point in a cluster) and average inter-cluster distance (i.e. average distance between all clusters), we will be able to uncover which of the respective models would be a better choice for further analysis and cluster interpretation. Additionally, as Silhouette Score would be utilized as our common performance measure, the optimization technique would thus include Silhouette Analysis for the K-Means Clustering model in place of the Elbow technique typically utilized as well as the line chart for the Hierarchical Clustering model which allows us to identify the optimal number of clusters that would provide the highest Silhouette Score.

**Cluster Interpretation**

Finally, by identifying the best models with the highest Silhouette Score, we could then derive business insights by performing cluster interpretation techniques such as observing the boxplot distribution of various features and their respective clusters before subsequently identifying customer segments. Thus, this would allow us to identify actionable plans and recommendations that the bank could take advantage of to drive their revenue alongside minimizing their credit risk.

## Bank Dataset to be analyzed:

|  |  |
| --- | --- |
| **Recap on Bank\_Data.csv** | |
| **Variable** | **Definition from Metadata and Assumptions made** |
| Account Status | Duration since salaries were first deposited into the checking account on a regular basis if any in years |
| Duration in month | Loan Tenure in months |
| Credit History | No loans take / Regular Repayment / Late Repayment with Current Bank /Late Repayment with Other Banks in the past |
| Purpose | The reason the individual borrowed credit from the bank |
| Credit amount | The total sum of loan taken from all lines of credit |
| Savings account or bonds | Whether an individual has a savings account or other types of account with the bank if any |
| Present employment since | Duration since an individual was first employed in their current organization in years |
| Installment rate in percentage of disposable income | Percentage of Disposable income that needs to be paid back to the bank from the credit amount borrowed |
| Personal status and sex | Relationship status of an individual based on their gender |
| Other debtors or guarantors | Whether the individual has others who are liable in repaying the loan alongside himself or herself |
| Present residence since | Duration since an individual lived in his or her current residence |
| Property | Available collateral that has been pledged and to be forfeited in the event of default in years |
| Age in years | Age of the individual in years |
| Other installment plans | Other repayments that the individual is contracted to |
| Housing | Whether the individual bought, rented, or was gifted his or her current residence |
| Number of existing credits at this bank | The number of loans/lines of credit taken at the bank |
| Job | Required Skill level of the occupation |
| Number of people being liable to provide maintenance for | How many relatives the individual is legally responsible to provide maintenance for |
| Telephone | Whether the individual can be contactable by the bank through telephone |
| Foreign worker | Whether the individual is a foreign worker or not |
| Target | Whether the individual would be represent a bad or good credit risk to the bank |

# Build Clustering Models using Numerical Data

## Data exploration and Manipulation on numerical data

### Exploratory Data Analysis

#### Identifying Null Values in the Bank Dataset (I)

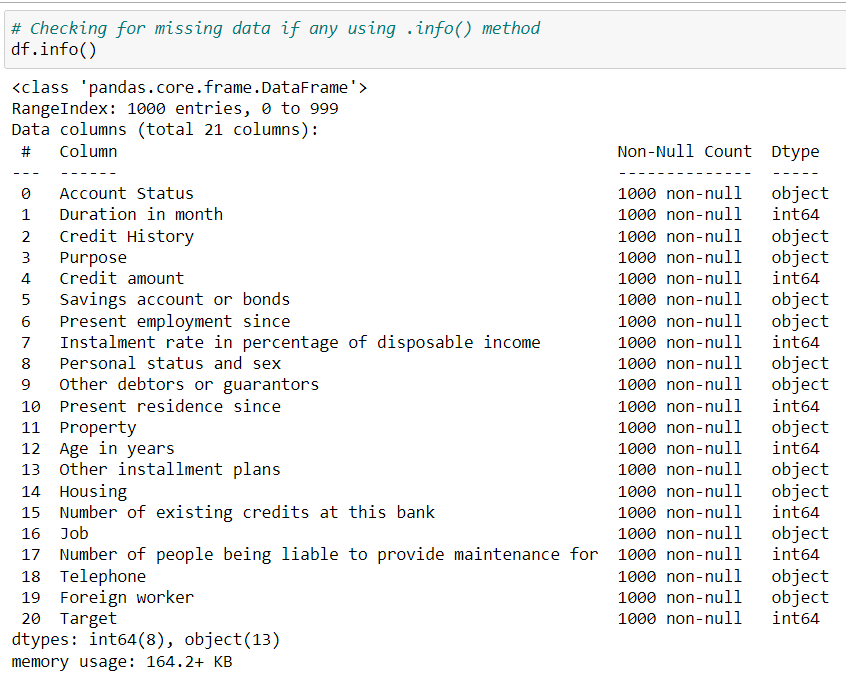


Figure 1

As part of our data exploration process, the first procedure we had taken would be to identify if there are any null values in our dataset that may hinder the subsequent stages of clustering. Hence, by utilizing the .info() method as observed from Figure 1, we could observe that of the 1000 entries in the dataframe, all of the features have a Non-Null Count of 1000 as well. This implies that there are no missing values in our data frame such that no missing value treatment would have to be performed.

#### Identifying Null Values in the Bank Dataset (II)

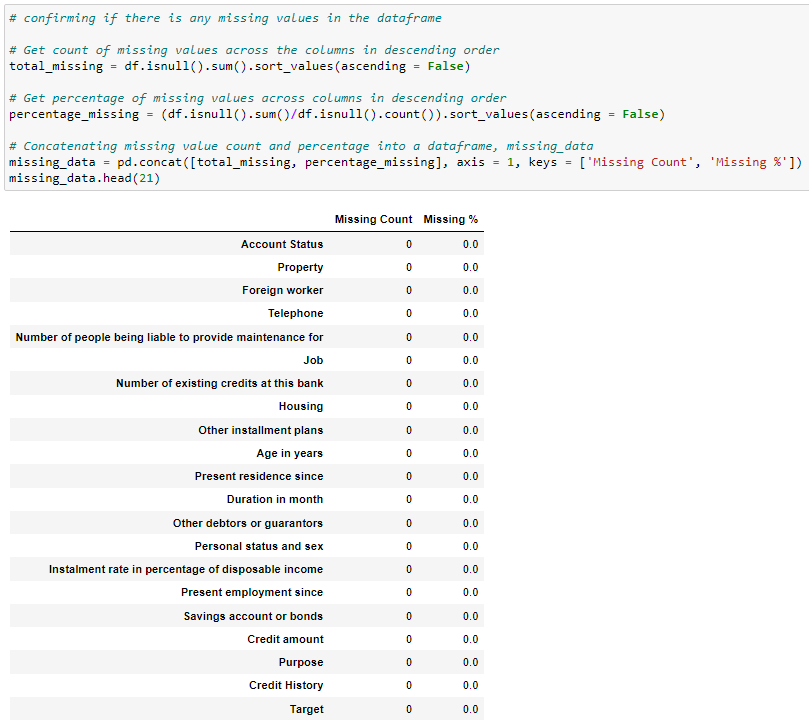


Figure 2

To verify that there are no missing values in our dataset, we could utilize other methods such as df.isnull().sum() in order to count the number of missing values in each column as well as df.isnull().sum()/df.isnull().sum() in order to return the percentage of missing values in each feature. Thus, as observed from figure 2, we could likewise confirm that there are zero counts and percentages of missing values, which is in alignment with our observations from Figure 1.

#### Identifying Count and Unique Values of Each Feature

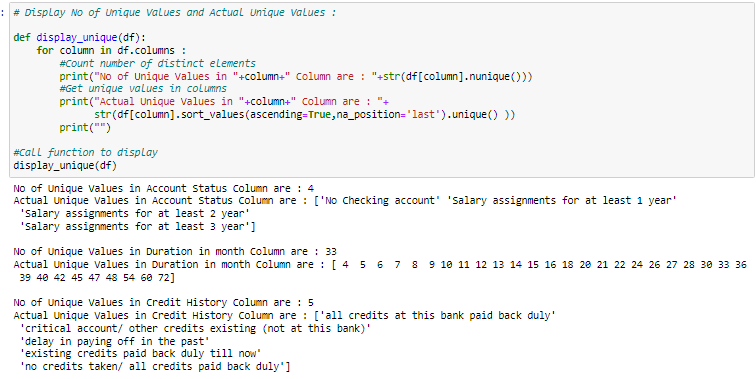


Figure 3

The next part of the exploratory data analysis process includes understanding what are the unique values of each feature that are available in our dataset. By doing the above as observed from Figure 3, readers would be able to view the count and unique values of each feature. Hence, this in turn would allow us to identify irrelevant and meaningless in our dataset that could potentially be removed in the subsequent stages.

#### Potential Irrelevant Features to Drop



Figure 3.1

As per Figure 3.1, we could observe that the ‘Savings account or bonds’ column is irrelevant to our problem of clustering given that it only has one unique value of ‘savings account’. Thus, given that the value of ‘Savings account’ is equally distributed across ‘Savings account or bonds’, it is not a meaningful variable to utilize in order to identify customer segments. Hence, it is likely that we will drop the following feature.

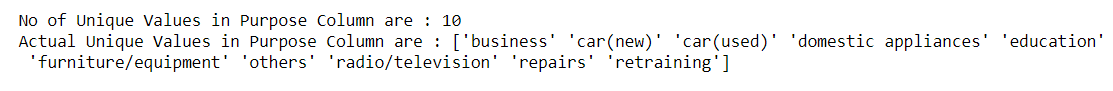


Figure 3.2

As per Figure 3.2, we could observe that the ‘Target’ feature represents the predicted output of whether a customer represents a good or bad credit risk to the bank. Hence, given that the feature would be more suited towards predictive modeling, which is not the focus of this project, it would not be a meaningful variable to utilize to identify customer segments. Hence, it is likely that we will drop the following feature.

#### Potential Categorical Features to Drop

Besides the irrelevant features we have identified in this report that we intend to drop, nominal categorical variables represent another set of features that are to be dropped. This is because nominal categorical variables do not have an intrinsic ranking or natural order, as such, label encoding data is arbitrary and calculations including mean, median, or standard deviation would be meaningless. Thus, it would not be meaningful to build K-Means and Hierarchical Clustering models with nominal categorical data. The nominal categorical data identified include:

  
Figure 3.3

As per Figure 3.3, which illustrates the unique count and values of the ‘Purpose’ column, it is a potential feature to be dropped given that it is a nominal categorical data. Owing to the fact that it is hard to quantify which of the following purposes would require a higher or lower credit amount given that it is subjective to the exact context, there is no intrinsic order to the ‘Purpose’ column.

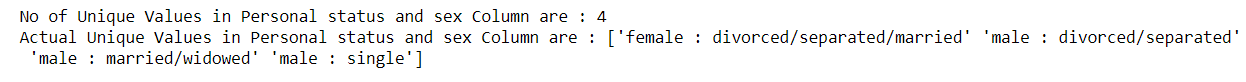


Figure 3.4

As per Figure 3.4, which illustrates the unique count and values of the ‘Personal status and sex’ column, it is a potential feature to be dropped given that it has characteristics of a nominal categorical data. For instance, while gender alone could allow for intrinsic ordering, the introduction of terms used differently such as , ‘divorced,’ ‘separated’, ‘married’, ‘widowed’, ‘single’ makes it complex to order the unique values. (e.g., The ‘married’ term for ‘male’ is grouped differently from how the ‘married’ term is grouped for ‘female’)

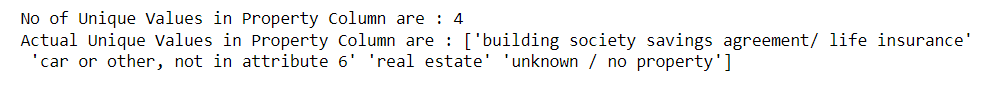


Figure 3.5

Like Figure 3.3, Figure 3.5 which illustrates the unique count and values of the ‘Property’ column is a potential feature to be dropped given that it is a nominal categorical data with ambiguity. Owing to the fact it is hard to quantify whether a building society savings agreement would represent a higher or lower collateral value compared to ‘car or other, not in attribute six’ which is already ambiguous on its own, there is no intrinsic order to the ‘Property’ column.

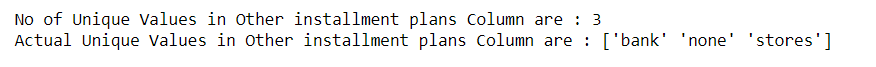


Figure 3.6

Likewise for Figure 3.6, which illustrates the count and unique values of the ‘Installment plans’ column, it is hard to quantify if ‘bank’ or ‘stores’ would represent a more expensive or cheaper installment plan given that it is subjective to the context.

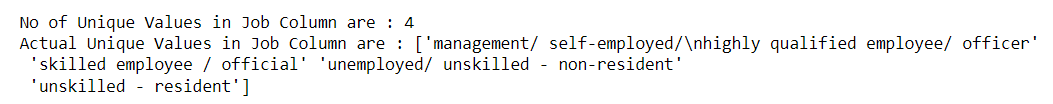


Figure 3.7

Like the problem in Figure 3.4, while it is simple to rank between a skilled employee and unskilled employee, the introduction of other terms such as ‘ officer’, ‘official’, ‘resident’, ‘non-resident’ used sparingly makes it unnecessary complicated task to quantify which of the following values would be higher or lower in nature.

#### Analyzing Distribution of Numerical Features

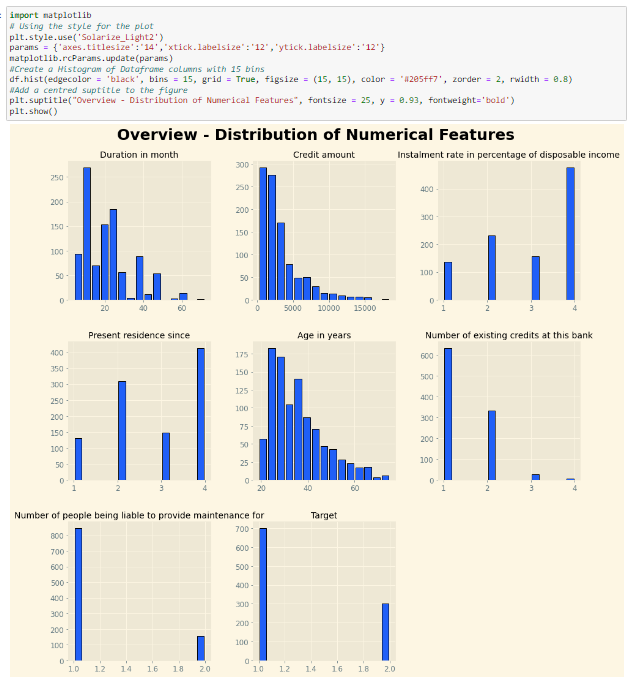


Figure 4

By observing the histogram distribution of all the numerical features, we could observe that of the eight numerical features available in our dataset, only three of the following features including ‘Duration in month’, ‘Credit amount’, and ‘Age in years’ represent non-discrete values with a large difference in scale, while the rest represent discrete values with similar scale. Hence, as observed from Figure 4 which illustrates that the distribution of ‘Duration in month’, ‘Credit amount’, and ‘Age in years’ is positively skewed; to ensure that the data has the same scale, we could potentially perform sklearn.preprocessing.StandardScaler which utilizes Z-Score transformation to scale our data. However, a potential issue with just applying Standard Scaler alone would be that it assumes that the distribution is normal. Hence, to resolve this issue, we could first perform log transformation to transform the positively skewed distribution to a normal distribution before applying Standard Scaler.

### Data Manipulation - Data Cleaning

#### Dropping Irrelevant Features

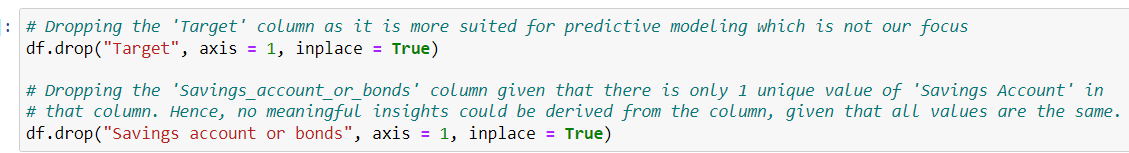
****

Figure 5

Hence, as per our discussion in the exploratory data analysis section, the ‘Target’ feature is more suited towards predictive modeling rather than clustering model which would be our focus for this assignment. As such, in view of the fact that it would not be a relevant feature in identifying customers segments, the following feature would be dropped. Likewise, for the ‘Savings account or bonds’ feature, given that we have previously discussed that the feature only contains one unique value of 'Savings’ which is equally distributed across all entries, it would not be a meaningful feature to create clustering models with. Thus, the following feature would be dropped as well.

#### Dropping Nominal Categorical Features

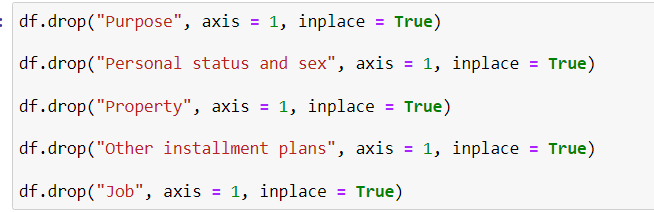


Figure 6

As per our previous discussion that nominal categorical variables would not be effective for clustering given that there is no intrinsic ordering of the various categories, such features would thus be dropped to focus only on meaningful features that could be used effectively for clustering. This includes the ‘Purpose’, ‘Personal status and sex,’ ‘Property’, ‘other installment plans’, and ‘job’ variables.

### Data Manipulation - Data Transformation

#### Replacing Column Name Spaces with Underscores

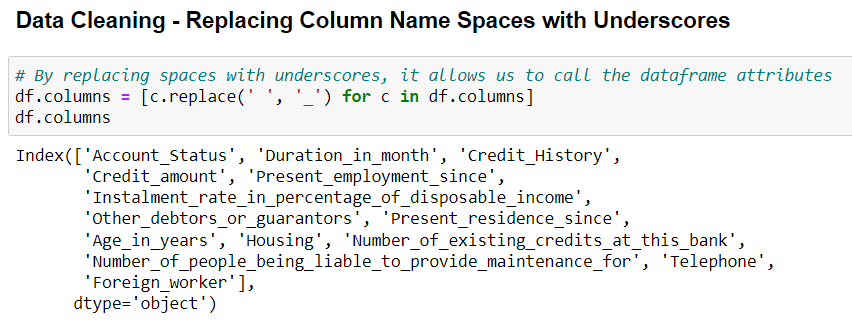


Figure 7

Moving on to the data transformation processes we have performed, the first process we had undertaken would involve replacing the column name spaces with underscores as observed from Figure 7. Hence, by doing so, it would allow us to call upon the various features when accessing their respective attributes and methods (e.g., df.Credit\_amount.mean()), which would not be possible with column name spaces.

#### Transforming Column to the same metric of comparison

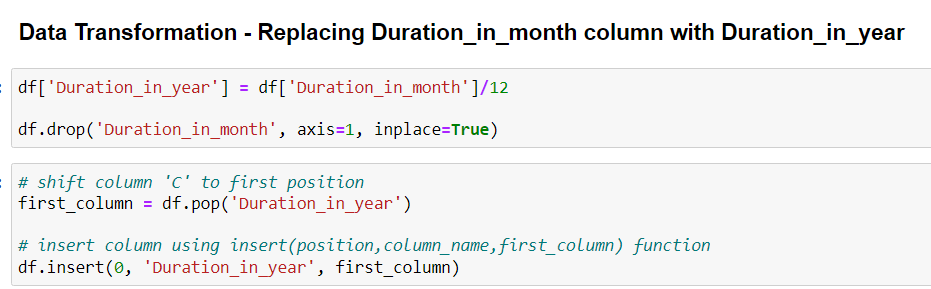


Figure 8

The next part of the data transformation process involves converting the various features to the same metrics of comparison. For instance given that the columns including ‘Account Status’, ‘Present employment since’, and ‘Age in years’ are using years as their default metrics, we would want to convert the ‘Duration in month’ column to ‘Duration in year’ given that it is the only feature using the months metric which may affect the outcome of our clusters. Hence, to convert the ‘Duration in month’ column to ‘Duration in year,’ we could simply divide the existing column by twelve while also dropping the ‘Duration in month’ feature given that it would be further utilized. Additionally, to place the ‘Duration in year’ column to its original position in the dataframe (i.e., First column), we could use the pop and insert method to shift the new column accordingly.

#### Distribution of the ‘Duration in year’ column before transformation

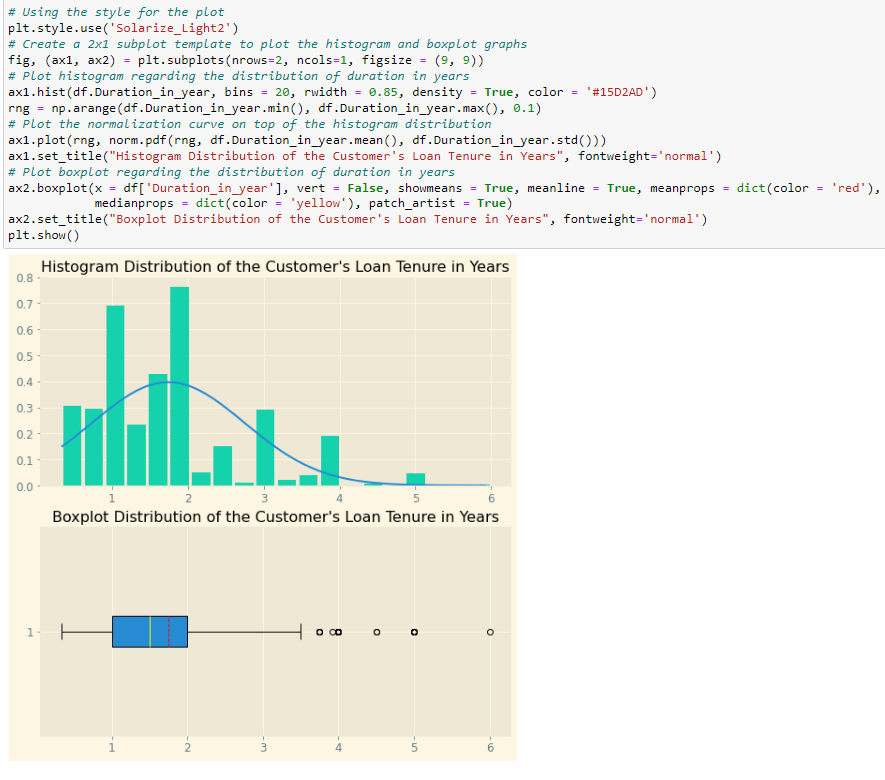


Figure 9.1

By observing the histogram from Figure 9.1 which illustrates the distribution of the ‘Duration in year’ column, we could observe that the feature is positively skewed in nature as seen from the distribution being shifted mostly to the left while its ‘tail’ is on the right. Additionally, by observing the boxplot distribution below it, we could see that there are approximately six outliers on the right of the boxplot whisker. Hence, to correct the positively skewed distribution into a normal distribution, we could apply log transformation which is effective in correcting the positively skewed alongside outliers found before we apply Z-score transformation in subsequent stages to ensure that variables in comparison are on the similar scale.



Figure 9.2

Hence, to perform log transformation on our ‘Duration in year’ column, we could utilize NumPy’s library log function as observed from Figure 9.2 which would convert the positively skewed distribution of ‘Duration in years’ to a normal distribution.

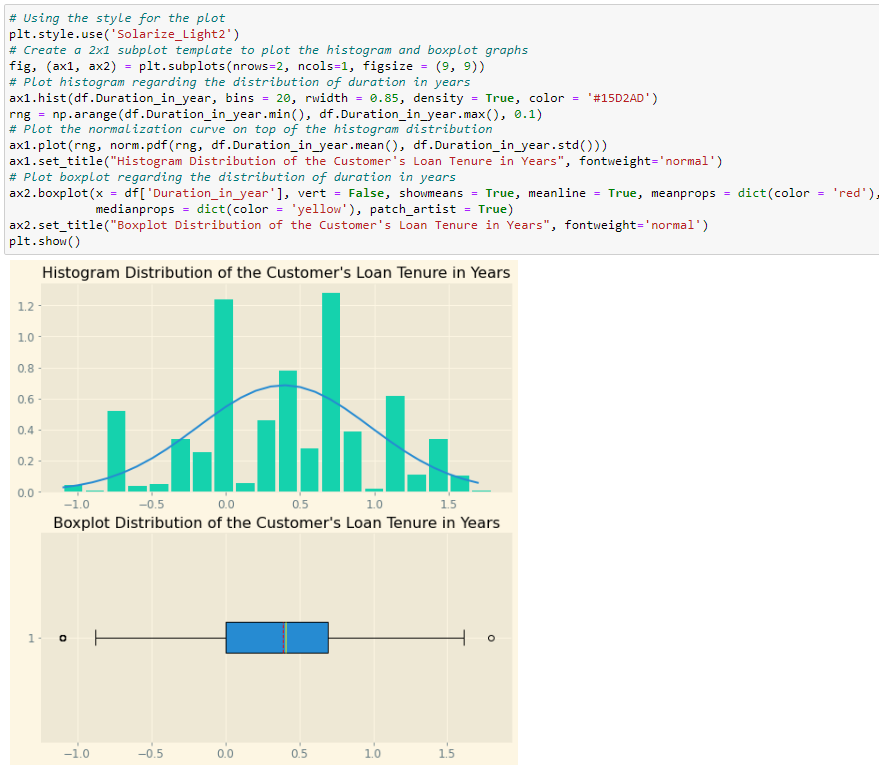


Figure 9.3

By applying log transformation on the ‘Duration in year’ feature and revisiting its distribution, we could observe that we have successfully converted the positively skewed distribution into a normal distribution as evident from the bell-shaped symmetrical pattern observed from Figure 9.3. Furthermore, by reviewing the boxplot distribution we could see that we have successfully removed most of the outliers such that there is only one outlier on both sides of the boxplot whisker.

#### Distribution of the ‘Credit amount’ column before log transformation

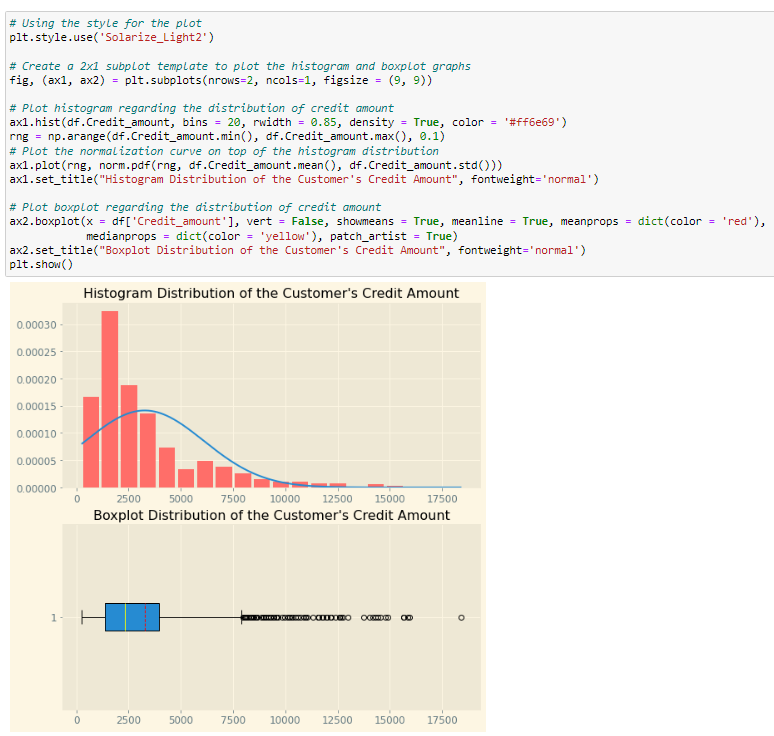


Figure 10.1

By observing the histogram from Figure 10.1 which illustrates the distribution of the ‘Credit amount’ column, we could observe that the feature is positively skewed in nature as seen from the distribution being shifted mostly to the left while its ‘tail’ is on the right. Additionally, by observing the boxplot distribution below it, we could see that there are numerous outliers on the right of the boxplot whisker. Hence, to correct the positively skewed distribution into a normal distribution, we could apply log transformation which is effective in correcting the positively skewed alongside outliers found before we apply Z-score transformation in subsequent stages to ensure that variables in comparison are on the similar scale.



Figure 10.2

Hence, to perform log transformation on the ‘Credit amount’ column, we could utilize NumPy’s library log method as observed from Figure 10.2 which would convert the positively skewed distribution of ‘Credit amounts’ to a normal distribution.

#### Distribution of the ‘Credit amount’ column after log transformation

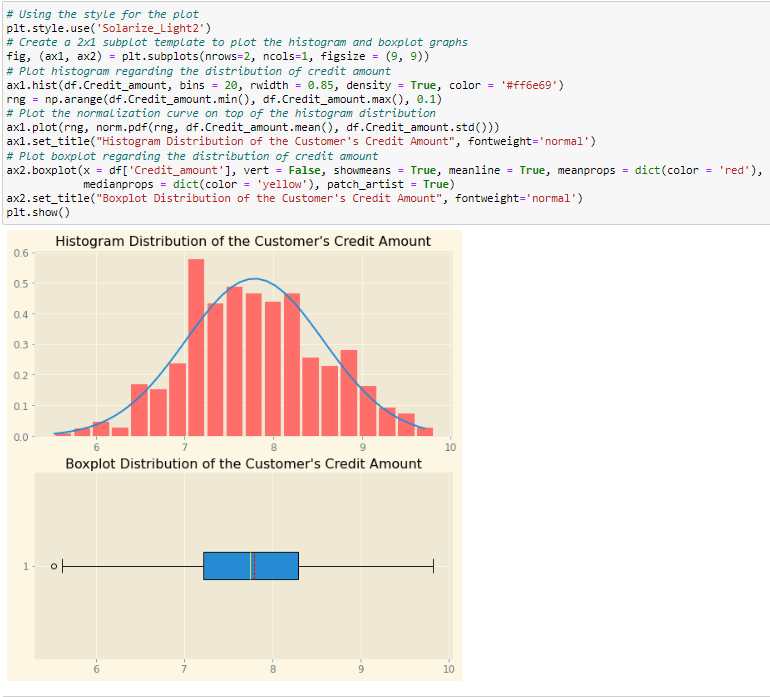


Figure 10.3

By applying log transformation on the ‘Credit amount’ feature and revisiting its distribution, we could observe that we have successfully converted the positively skewed distribution into a normal distribution as evident from the bell-shaped symmetrical pattern observed from Figure 10.3. Furthermore, by reviewing the boxplot distribution we could see that we have successfully removed most of the outliers such that there is only one outlier on the left side of the boxplot whisker.

#### Distribution of the ‘Age in years’ column before log transformation

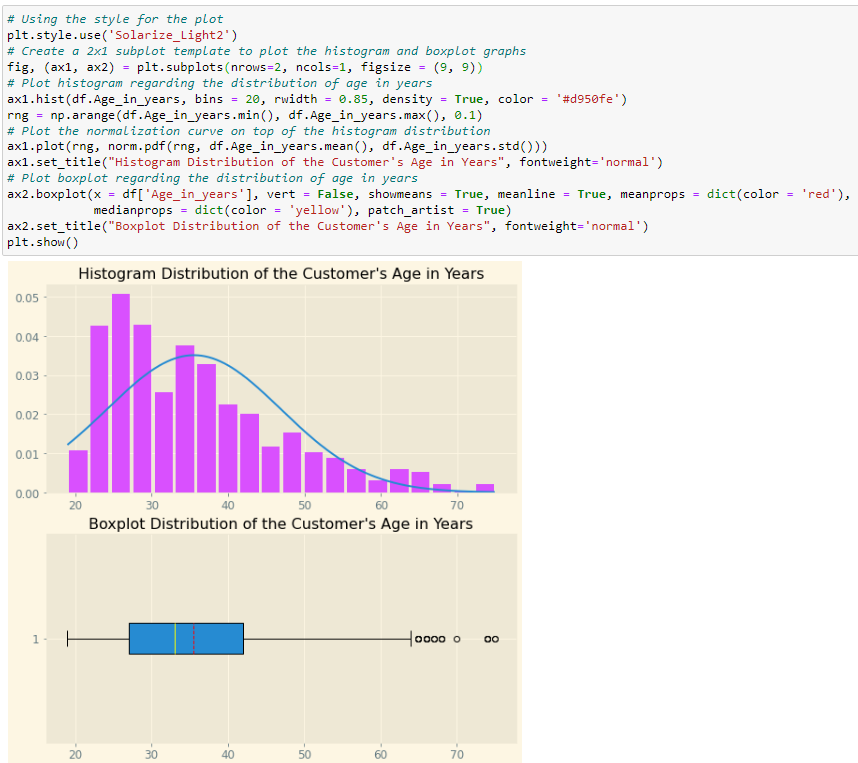


Figure 11.1

By observing the histogram from Figure 11.1 which illustrates the distribution of the ‘Age in Years’ column, we could observe that the feature is positively skewed in nature as seen from the distribution being shifted mostly to the left while its ‘tail’ is on the right. Additionally, by observing the boxplot distribution below it, we could see that there are approximately seven outliers on the right of the boxplot whisker. Hence, to correct the positively skewed distribution into a normal distribution, we could apply log transformation which is effective in correcting the positively skewed alongside outliers found before we apply Z-score transformation in subsequent stages to ensure that variables in comparison are on the similar scale.



Figure 11.2

Hence, to perform log transformation on the ‘Age in years’ column, we could utilize NumPy’s library log method as observed from Figure 11.2 which would convert the positively skewed distribution of the ‘Age in years’ column to a normal distribution.

#### Distribution of the ‘Credit amount’ column after log transformation

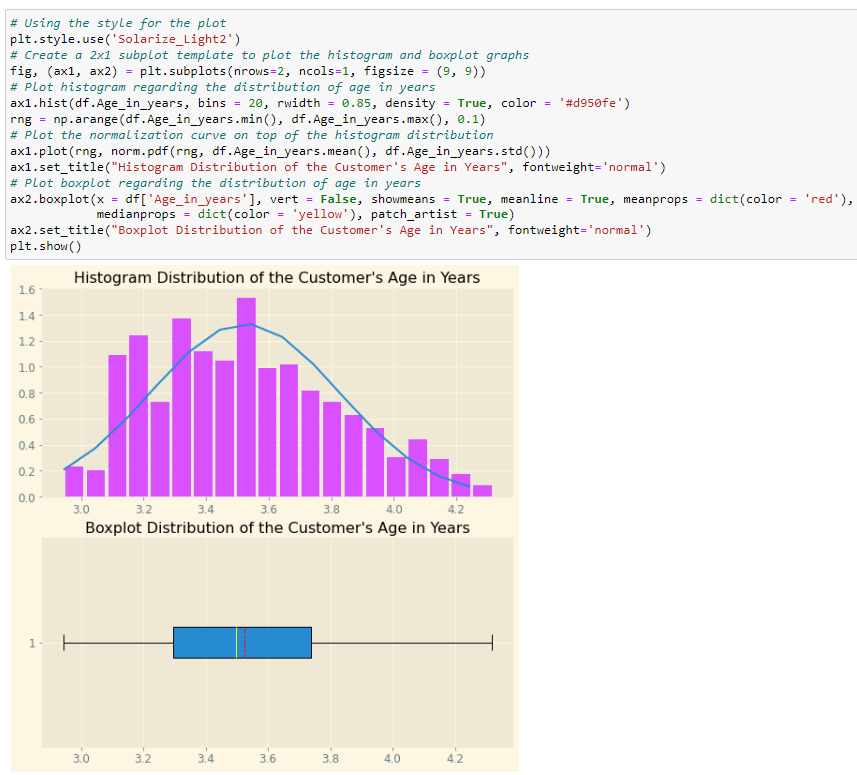


Figure 11.3

By applying log transformation on the ‘Age in years’ feature and revisiting its distribution, we could observe that we have successfully converted the positively skewed distribution into a normal distribution as evident from the bell-shaped symmetrical pattern observed from Figure 11.3. Furthermore, by reviewing the boxplot distribution we could see that we have successfully removed all the outliers previously present.

#### Label Encoding Ordinal Categorical Variables

To increase the variety of features available to build clustering models, we could utilize label encoding which is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on how we intend to map the ordering. Thus, by doing so, it would allow us to derive a larger number of insights from both numerical and ordinal categorical features. The ordinal categorical variables to be encoded would thus include the following as per Figure 12.

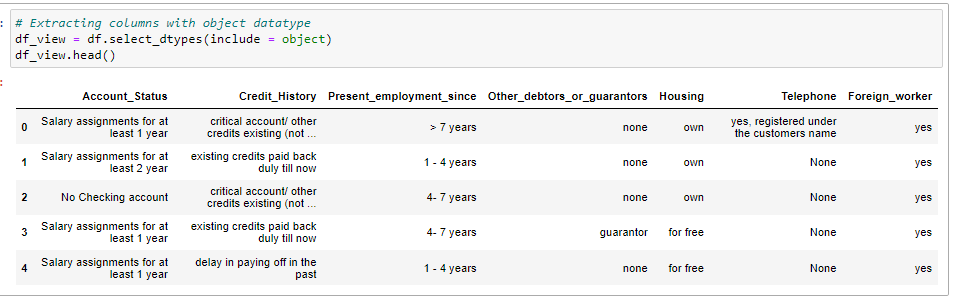


Figure 12

#### Label Encoding ‘Account Status’ Column

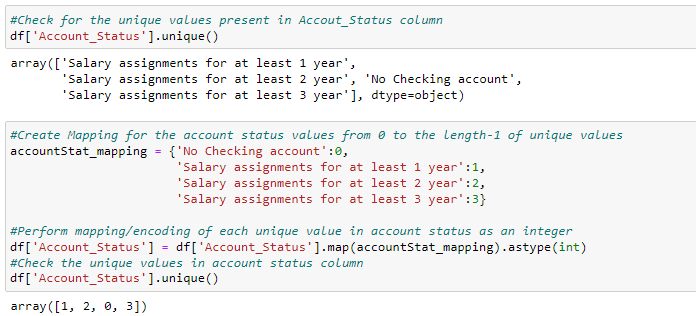


Figure 12.1

#### Label Encoding ‘Credit History’ Column



Figure 12.2

For the label encoding of ‘Credit History’ as observed from Figure 12.2, we will be treating ‘all credit at this bank paid back duly’ and ‘no credits take/ all credits paid back duly’ as the same value by encoding them to the same integer value of 3.

#### Label Encoding ‘Present Employment Since’ Column

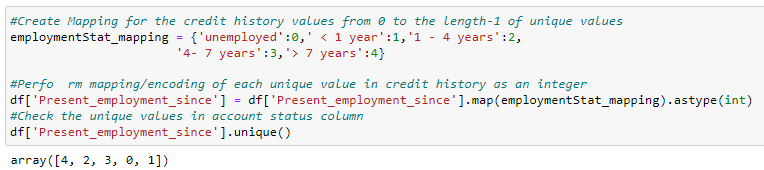


Figure 12.3

#### Label Encoding ‘Other Debtors or Guarantors’ Column

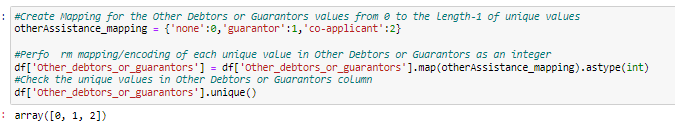


Figure 12.4

#### Label Encoding ‘Housing’ Column

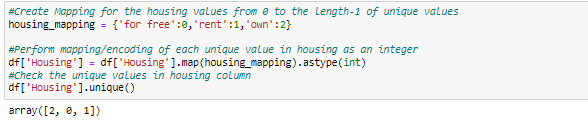


Figure 12.5

#### Label Encoding ‘Telephone’ Column

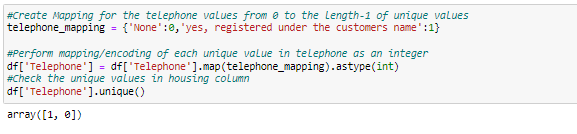


Figure 12.6

#### Label Encoding ‘Foreign Worker’ Column

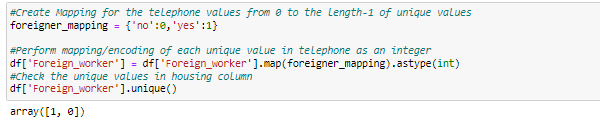
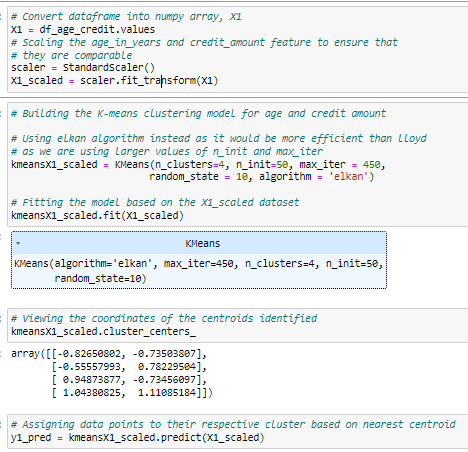


Figure 12.7

## Building, Optimizing, and Performing Clustering Analysis on Clustering Models

### K-Means Clustering Model

#### Basic K-Means Model 1A (Age and Credit Amount)

Figure 13.1

As per Figure 13, to build the K-Means model for Age and Credit Amount, we can firstly scale both variables to the same range and magnitude using Standard Scaler (i.e., Z-Score Transformation). Thereafter, to build the K-Means model, we would start with initially four clusters, while using the ‘Elkan’ algorithm which accelerates the clustering process by avoiding redundant distance calculation. Moving forward, we can fit the model created to the customer dataset of Age and Credit Amount which we have scaled using the fit() method before using the predict() method to assign the entries to their respective clusters.

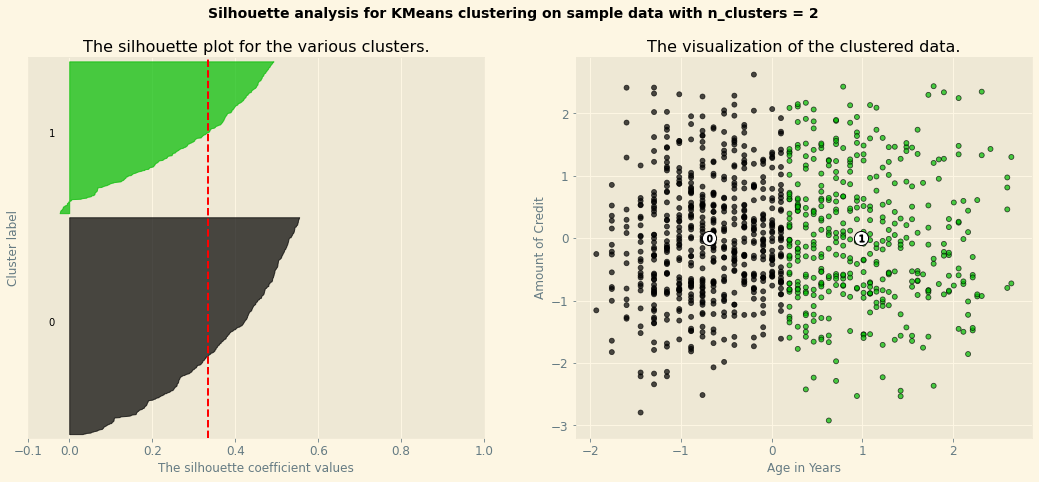


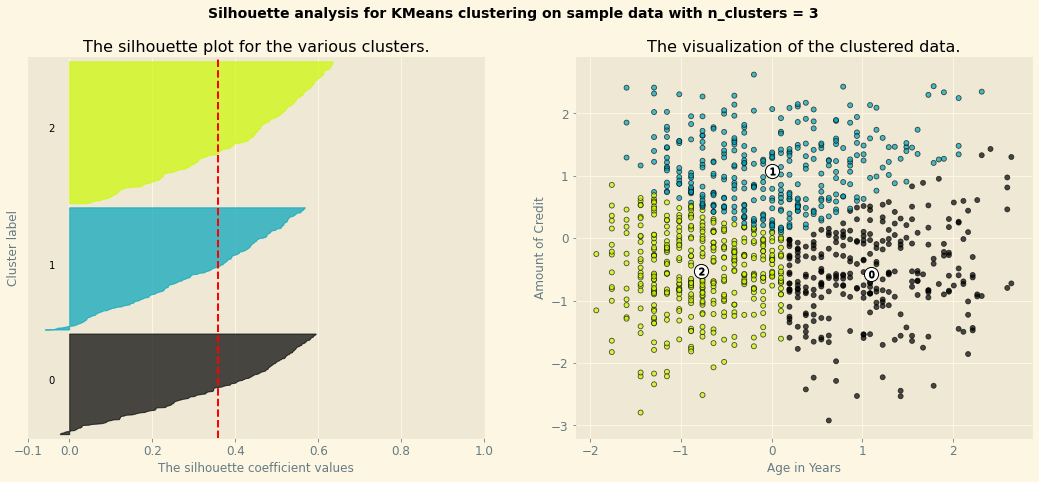
Figure 13.2

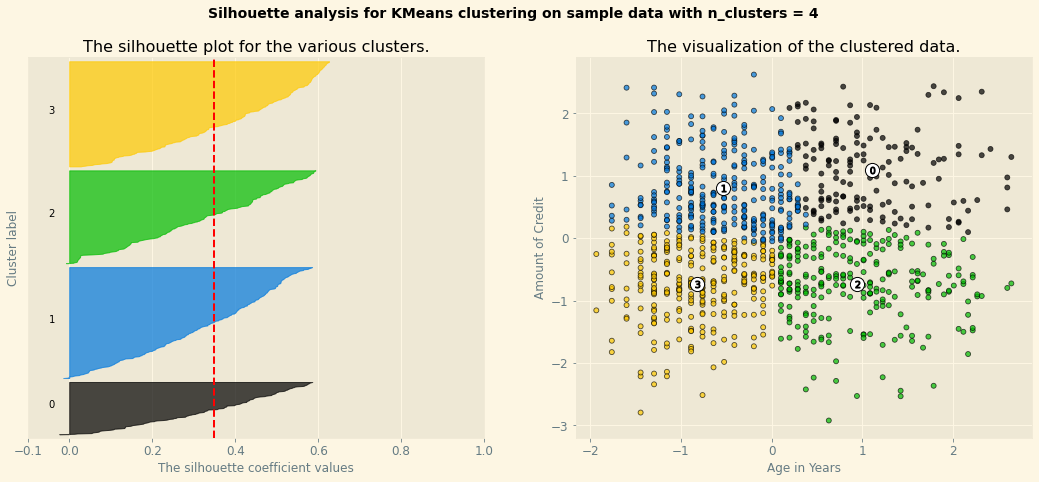
Thus, given that Silhouette Score would be used as our common performance measure against the Hierarchical Clustering model that will be built subsequently, we could observe that the K-Means Cluster with a cluster value of 4 has returned a Silhouette Score of 0.348 (1 - Best Possible Score).

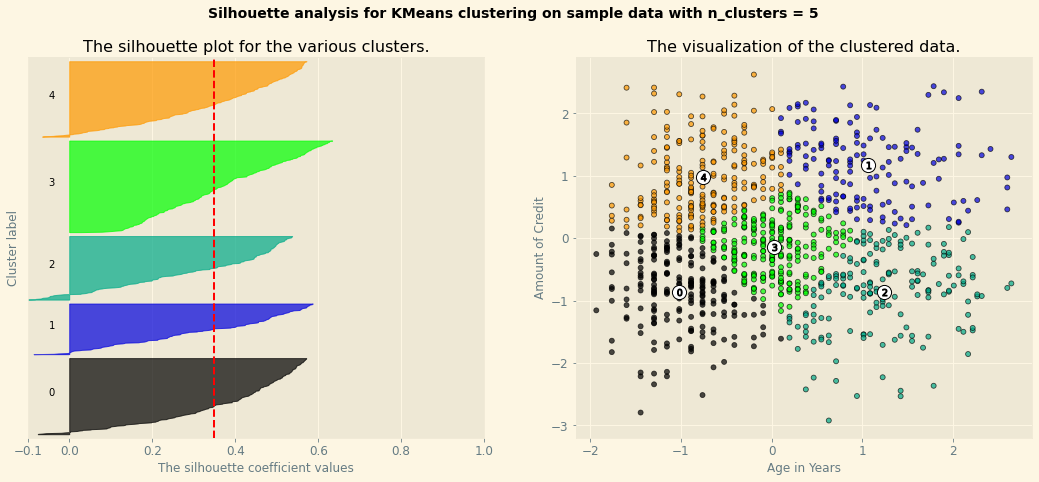
#### Optimized K-Means Model 1A (Age and Credit Amount) with Silhouette Analysis

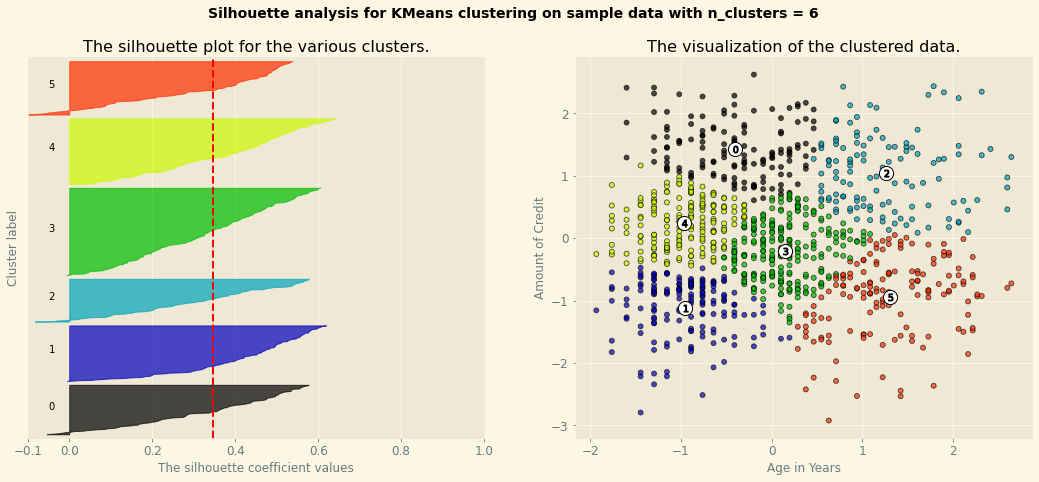
Given that Silhouette Score would be utilized as the common performance measure against Hierarchical Clustering, instead of using the Elbow Technique that we are familiar with which plots the Sum of Squared Errors, we can make use of the Silhouette Analysis technique which plots the average Silhouette Score across the clusters. Additionally, the K-Value that we will evaluate on ranges from 2 to 6 given that anything above 6 will be difficult to derive any meaningful insight from.











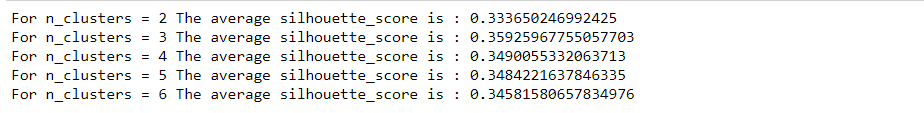


Figure 13.3

By plotting out the Silhouette Graphs, we could observe that among the clusters values of 2 to 6, clusters = 3 has provided the highest average Silhouette Score of 0.359 ( 3 d.p.) as observed from the red dotted line that cuts through the plots. Hence, we could observe that it has performed slightly better than the basic K-means model we had previously built which had an average Silhouette Score of 0.348 (3 d.p.).

#### Basic K-Means Model 2A (Age, Credit Amount, and Duration in Years)

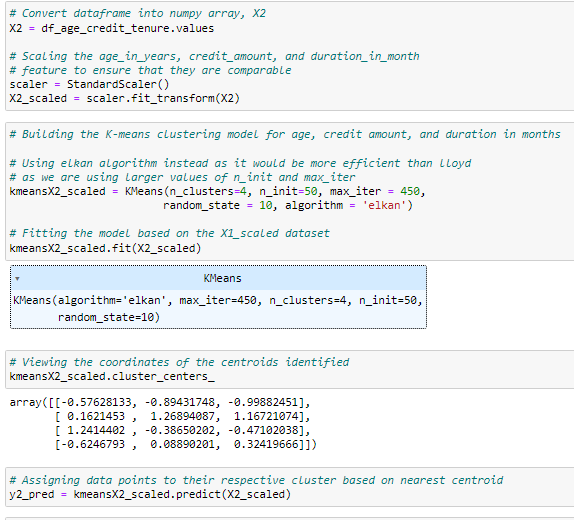
****

Figure 14.1

As per Figure 14.1, to build the K-Means model for Age, Credit Amount, and Duration in years, we can firstly scale the variables to the same range and magnitude using Standard Scaler (i.e., Z-Score Transformation). Thereafter, to build the K-Means model, we would start with initially four clusters, while using the ‘Elkan’ algorithm which accelerates the clustering process by avoiding redundant distance calculation. Moving forward, we can fit the model created to the customer dataset of Age, Credit Amount, and Duration in Years which we have scaled using the fit() method before using the predict() method to assign the entries to their respective clusters.

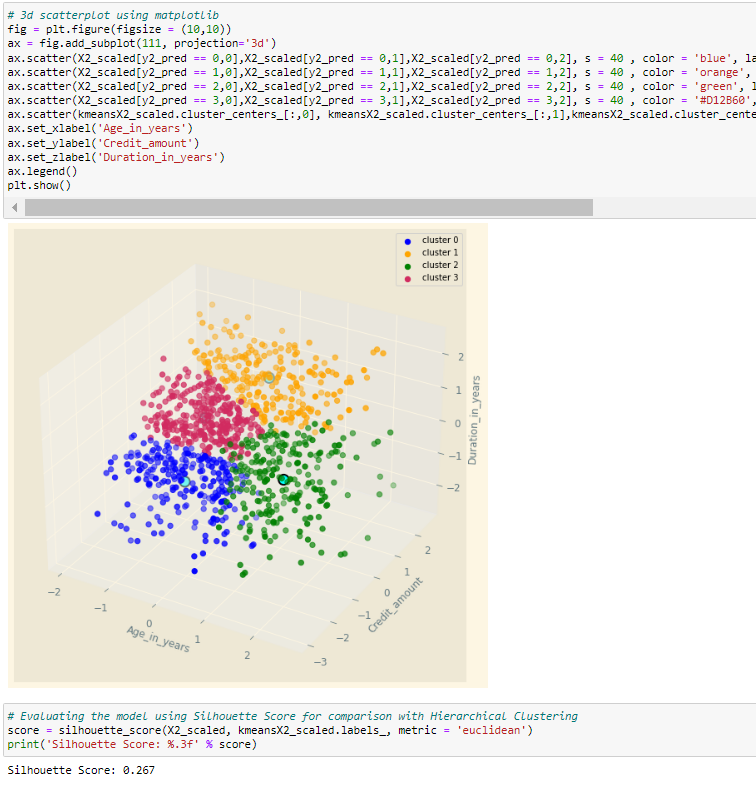
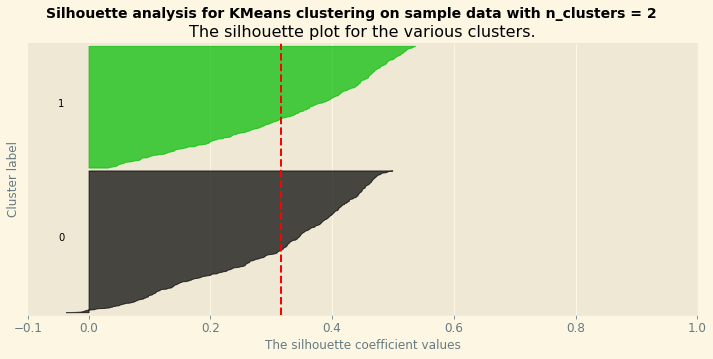


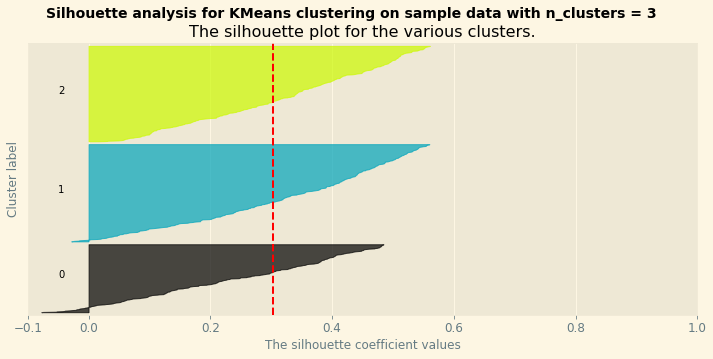
Figure 14.2

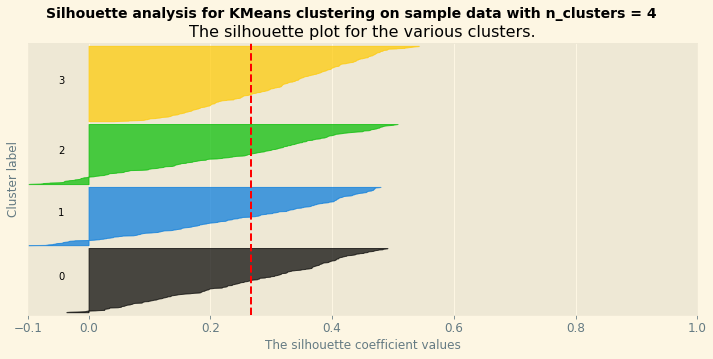
Thus, given that Silhouette Score would be used as our common performance measure against the Hierarchical Clustering model that will be built subsequently, we could observe that the K-Means Cluster with a cluster value of 4 has returned a Silhouette Score of 0.267 (1 - Best Possible Score).

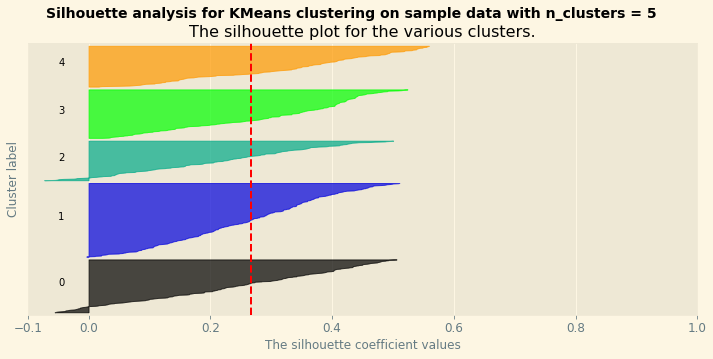
#### Optimized K-Means Model 2A (Age, Credit Amount, Duration in Years) with Silhouette Analysis

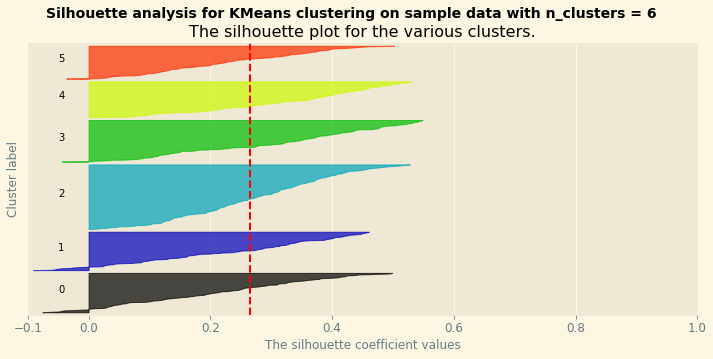
Given that Silhouette Score would be utilized as the common performance measure against Hierarchical Clustering, instead of using the Elbow Technique that we are familiar with which plots the Sum of Squared Errors, we can make use of the Silhouette Analysis technique which plots the average Silhouette Score across the clusters. Additionally, the K-Value that we will evaluate on ranges from 2 to 6 given that anything above 6 will be difficult to derive any meaningful insight from.











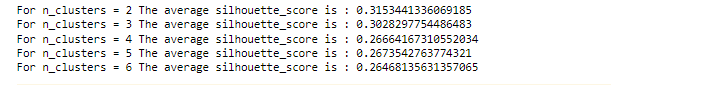


Figure 14.3

By plotting out the Silhouette Graphs, we could observe that among the clusters values of 2 to 6, clusters = 2 has provided the highest average Silhouette Score of 0.315 ( 3 d.p.) as observed from the red dotted line that cuts through the plots. Hence, we could observe that it has performed significantly better than the basic K-means model we had previously built which had an average Silhouette Score of only 0.267 (3 d.p.).

#### Basic K-Means Model 3A (Age, Present Employment Since, Credit History)

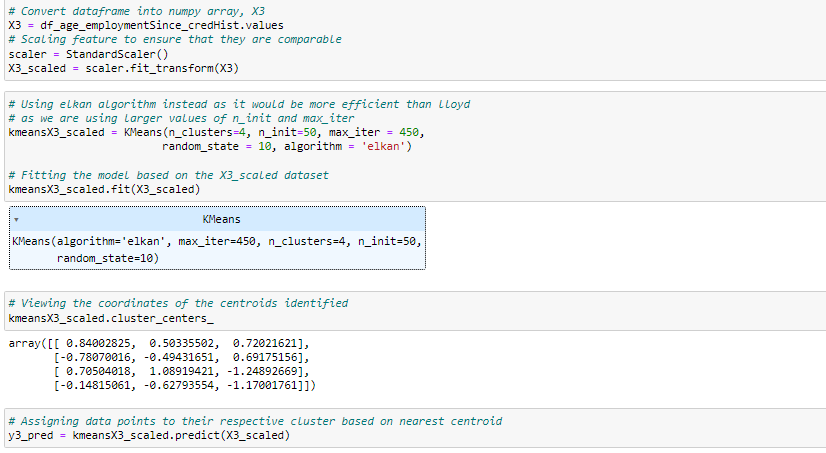
****

Figure 15.1

As per Figure 15.1, to build the K-Means model for Age, Present Employment Since and Credit History, we can firstly scale the variables to the same range and magnitude using Standard Scaler (i.e., Z-Score Transformation). Thereafter, to build the K-Means model, we would start with initially four clusters, while using the ‘Elkan’ algorithm which accelerates the clustering process by avoiding redundant distance calculation. Moving forward, we can fit the model created to the customer dataset of Age, Present Employment Since, and Credit History, which we have scaled using the fit() method before using the predict() method to assign the entries to their respective clusters.

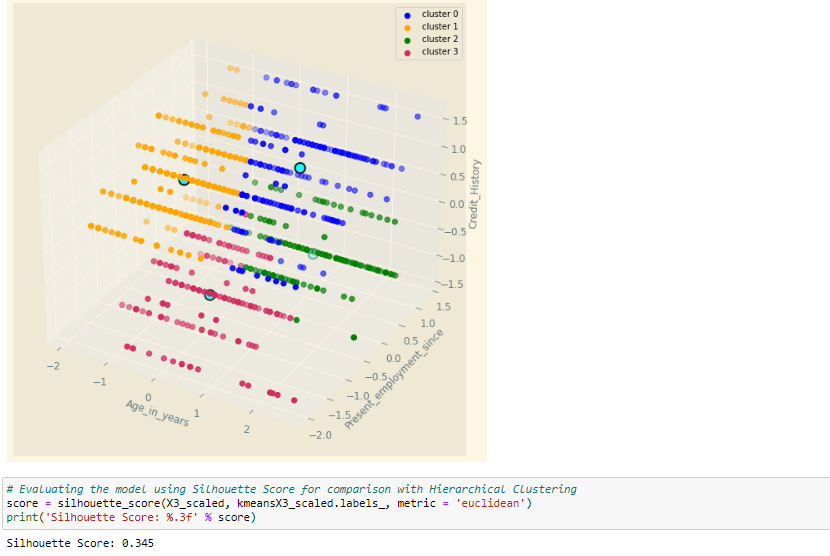
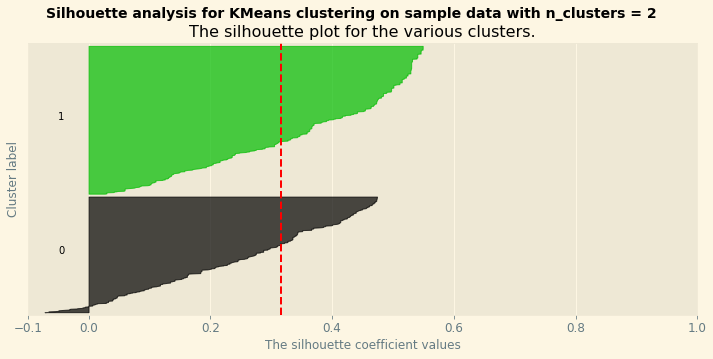


Figure 15.2

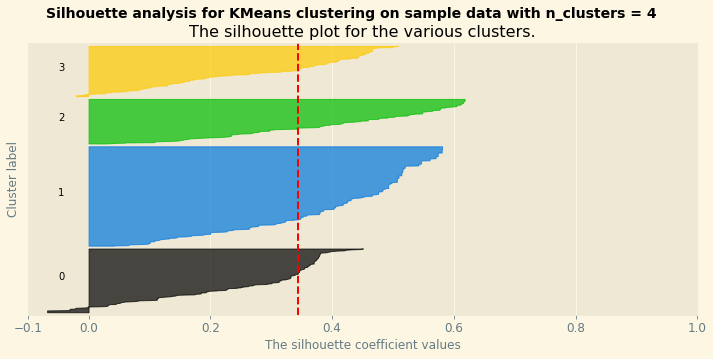
Thus, given that Silhouette Score would be used as our common performance measure against the Hierarchical Clustering model that will be built subsequently, we could observe that the K-Means Cluster with a cluster value of 4 has returned a Silhouette Score of 0.345 (1 - Best Possible Score).

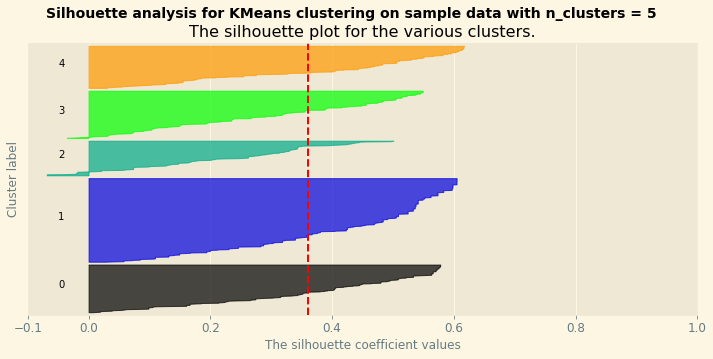
#### Optimized K-Means Model 3A (Age, Present Employment, Credit History) with Silhouette Analysis

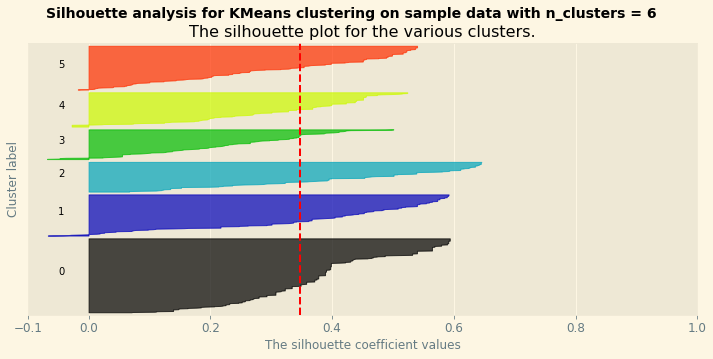
Likewise, given that Silhouette Score would be utilized as the common performance measure against Hierarchical Clustering, instead of using the Elbow Technique that we are familiar with which plots the Sum of Squared Errors, we can make use of the Silhouette Analysis technique which plots the average Silhouette Score across the clusters. Additionally, the K-Value that we will evaluate on ranges from 2 to 6 given that anything above 6 will be difficult to derive any meaningful insight from.











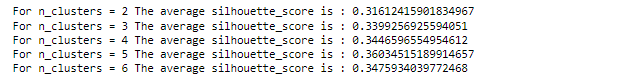


Figure 15.3

By plotting out the Silhouette Graphs, we could observe that among the n\_clusters values of 2 to 6, clusters = 5 has provided the highest average Silhouette Score of 0.360 ( 3 d.p.) as observed from the red dotted line that cuts through the plots. Hence, we could observe that it has performed significantly better than the basic K-means model we had previously built which only had an average Silhouette Score of only 0.345 (3 d.p.).

#### Basic K-Means Model 4A (Credit Amount, Number of Existing Credits, Credit History)

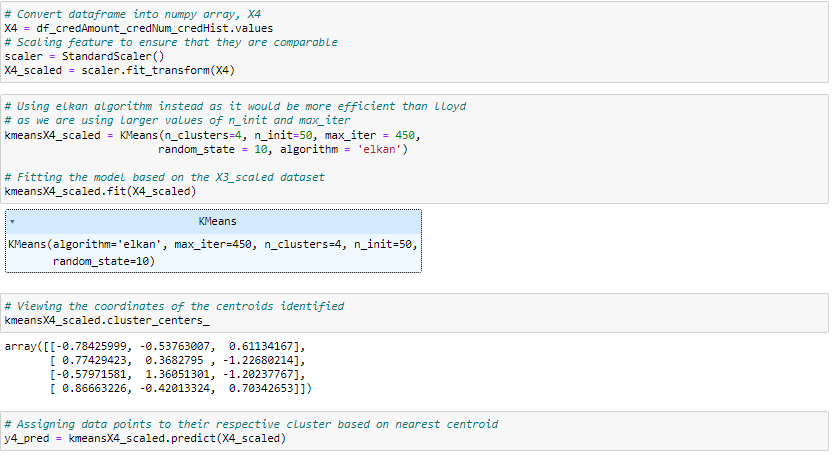
****

Figure 16.1

As per Figure 16.1, to build the K-Means model for Credit Amount, Number of Existing Credits, and Credit History, we can firstly scale the variables to the same range and magnitude using Standard Scaler (i.e., Z-Score Transformation). Thereafter, to build the K-Means model, we would start with initially four clusters, while using the ‘Elkan’ algorithm which accelerates the clustering process by avoiding redundant distance calculation. Moving forward, we can fit the model created to the customer dataset of Credit Amount, Number of Existing Credits, and Credit History, which we have scaled using the fit() method before using the predict() method to assign the entries to their respective clusters.

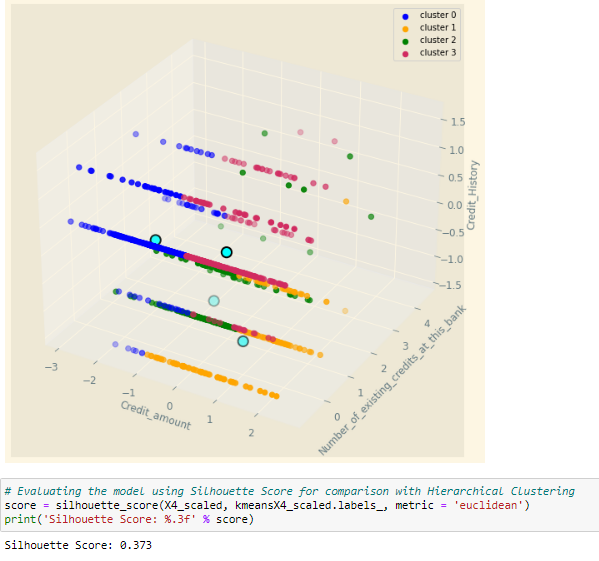
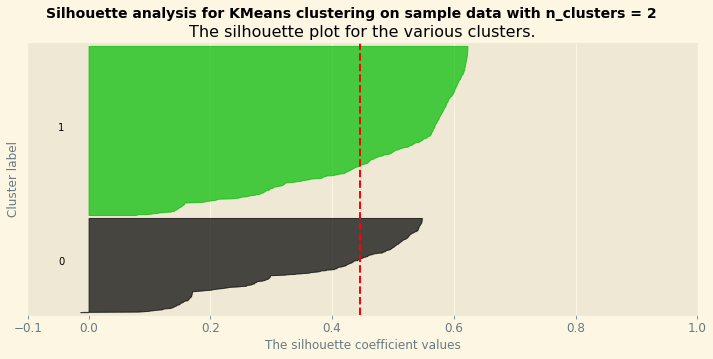


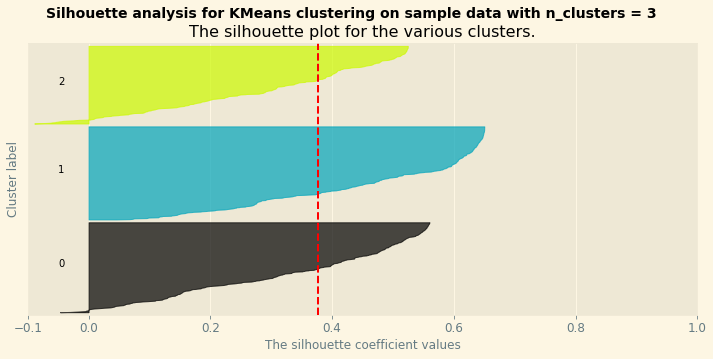
Figure 16.2

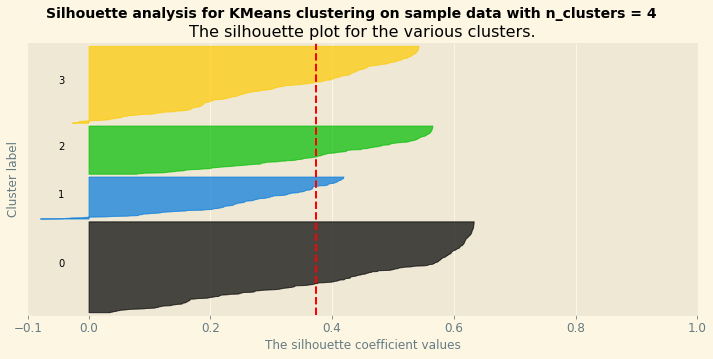
Thus, given that Silhouette Score would be used as our common performance measure against the Hierarchical Clustering model that will be built subsequently, we could observe that the K-Means Cluster with a cluster value of 4 has returned a Silhouette Score of 0.373 (1 - Best Possible Score).

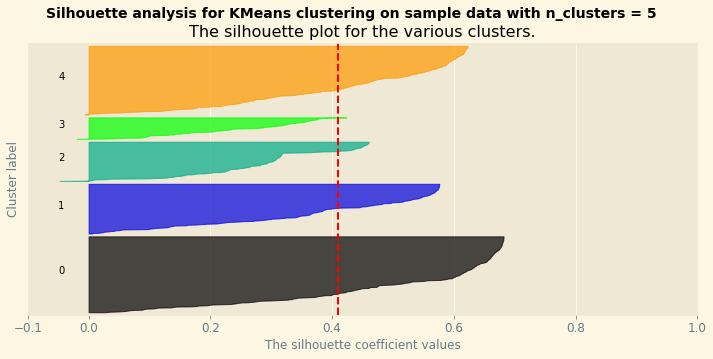
#### Optimized K-Means Model 4A (Credit Amount, Number of Existing Credits, Credit History) with Silhouette Analysis

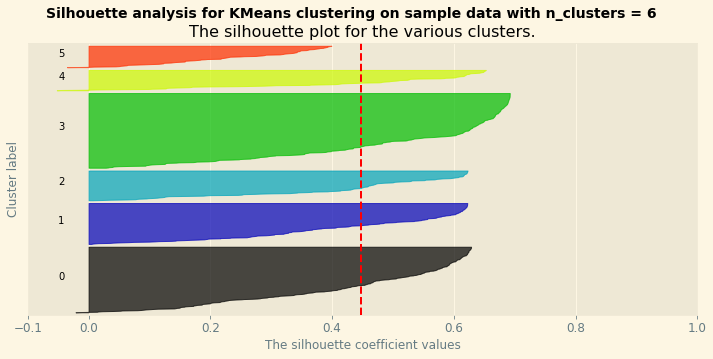
Likewise, given that Silhouette Score would be utilized as the common performance measure against Hierarchical Clustering, instead of using the Elbow Technique that we are familiar with which plots the Sum of Squared Errors, we can make use of the Silhouette Analysis technique which plots the average Silhouette Score across the clusters. Additionally, the K-Value that we will evaluate on ranges from 2 to 6 given that anything above 6 will be difficult to derive any meaningful insight from.











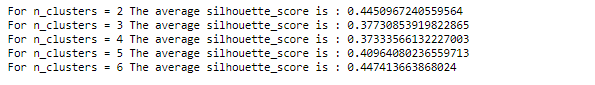


Figure 16.3

By plotting out the Silhouette Graphs, we could observe that among the clusters values of 2 to 6, clusters = 6 has provided the highest average Silhouette Score of 0.447 ( 3 d.p.) as observed from the red dotted line that cuts through the plots. Hence, we could observe that it has performed significantly better than the basic K-means model we had previously built which only had an average Silhouette Score of only 0.373 (3 d.p.).

#### Basic K-Means Model 5A (Account Status, Housing, Credit\_amount)

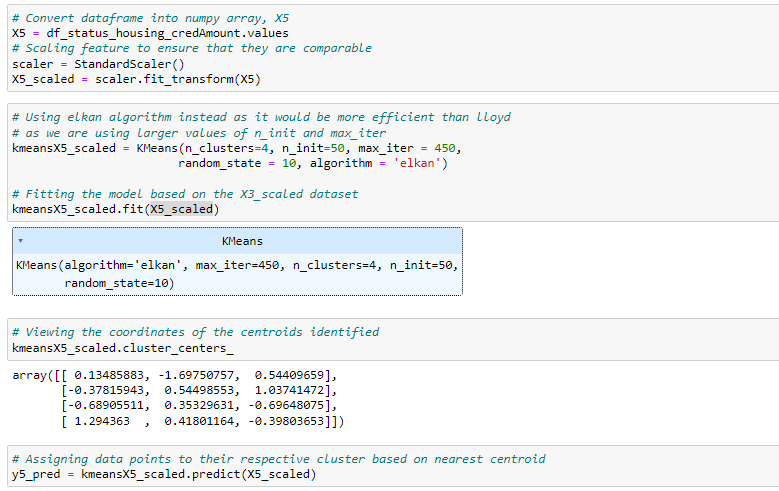
****

Figure 17.1

As per Figure 17.1, to build the K-Means model for Account Status, Housing, and Credit\_amount we can firstly scale the variables to the same range and magnitude using Standard Scaler (i.e., Z-Score Transformation). Thereafter, to build the K-Means model, we would start with initially four clusters, while using the ‘Elkan’ algorithm which accelerates the clustering process by avoiding redundant distance calculation. Moving forward, we can fit the model created to the customer dataset of Account Status, Housing, and Credit\_amount which we have scaled using the fit() method before using the predict() method to assign the entries to their respective clusters.

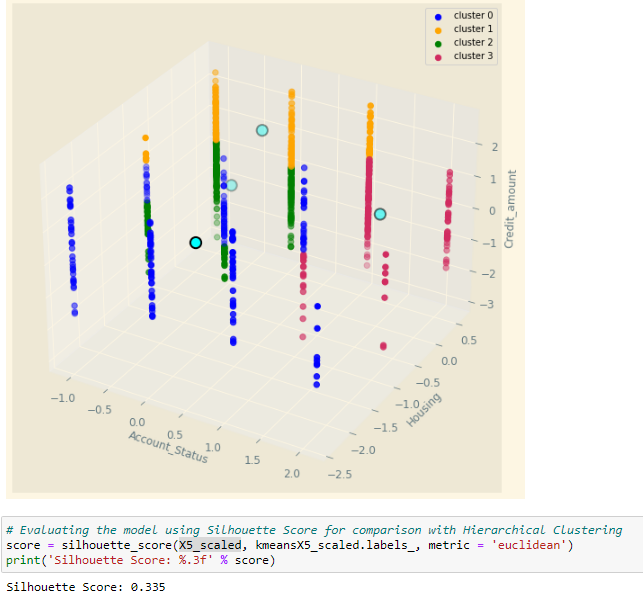
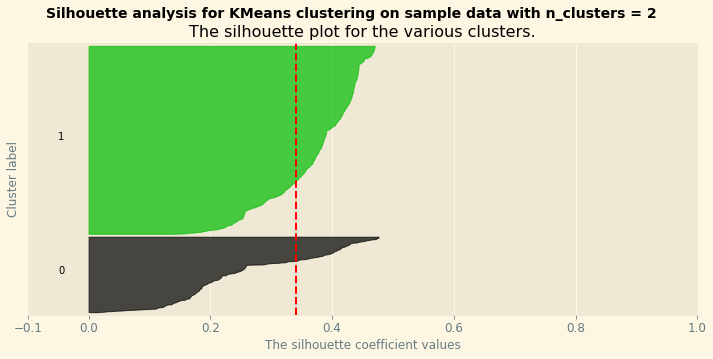
****

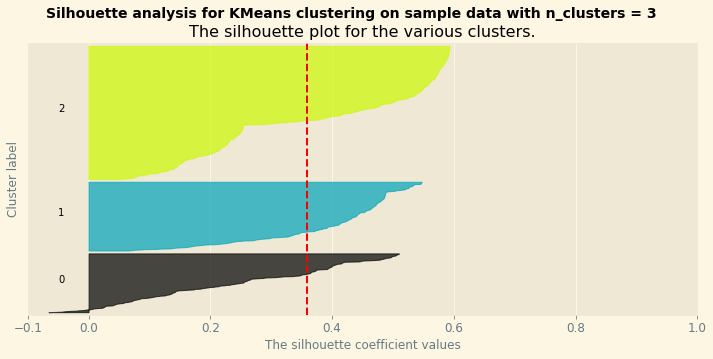
Figure 17.2

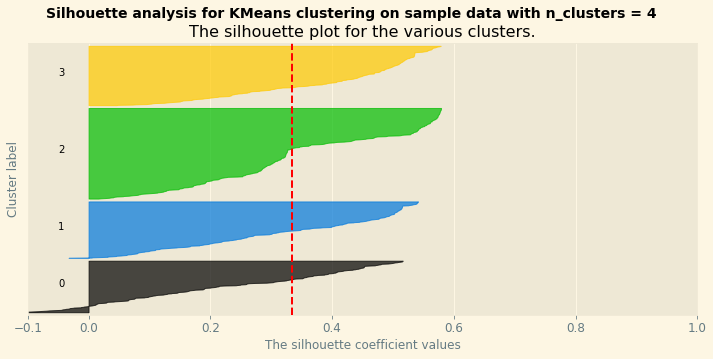
Thus, given that Silhouette Score would be used as our common performance measure against the Hierarchical Clustering model that will be built subsequently, we could observe that the K-Means Cluster with a cluster value of 4 has returned a Silhouette Score of 0.335 (1 - Best Possible Score).

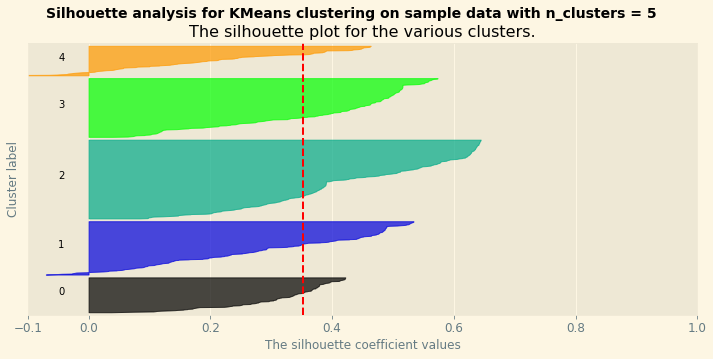
#### Optimized K-Means Model 5A ( (Account Status, Housing, Credit\_amount) with Silhouette Analysis

Similarly, given that Silhouette Score would be utilized as the common performance measure against Hierarchical Clustering, instead of using the Elbow Technique that we are familiar with which plots the Sum of Squared Errors, we can make use of the Silhouette Analysis technique which plots the average Silhouette Score across the clusters. Additionally, the K-Value that we will evaluate on ranges from 2 to 6 given that anything above 6 will be difficult to derive any meaningful insight from.









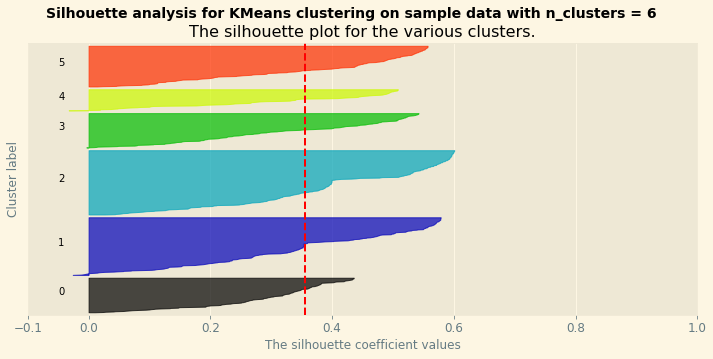




Figure 17.3

By plotting out the Silhouette Graphs, we could observe that among the cluster’s values of 2 to 6, clusters = 3 has provided the highest average Silhouette Score of 0.358 ( 3 d.p.) as observed from the red dotted line that cuts through the plots. Hence, we could observe that it has performed significantly better than the basic K-means model we had previously built which only had an average Silhouette Score of only 0.335 (3 d.p.).

### Hierarchical Clustering Models

#### Basic Hierarchical Clustering Model 1B (Age and Credit Amount)

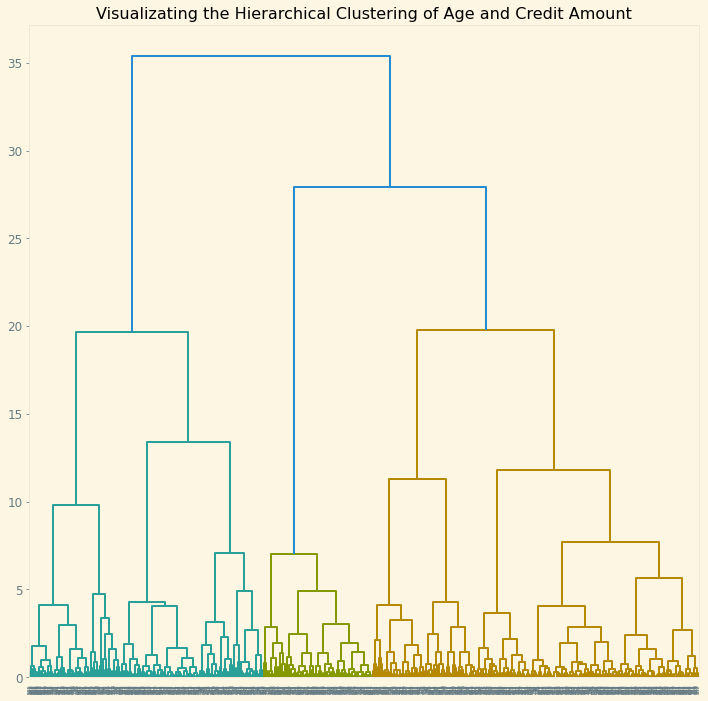
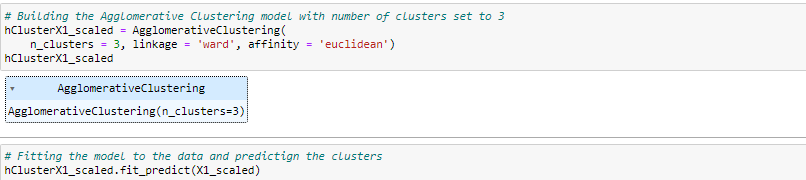
****

Figure 18.1

By observing Figure 18.1, which illustrates the dendrogram created using the X1\_scaled dataset through the Agglomerative Clustering Linkage, we could observe that the clusters value we should start off with would be three given that it cuts the maximum vertical distance observed.

****

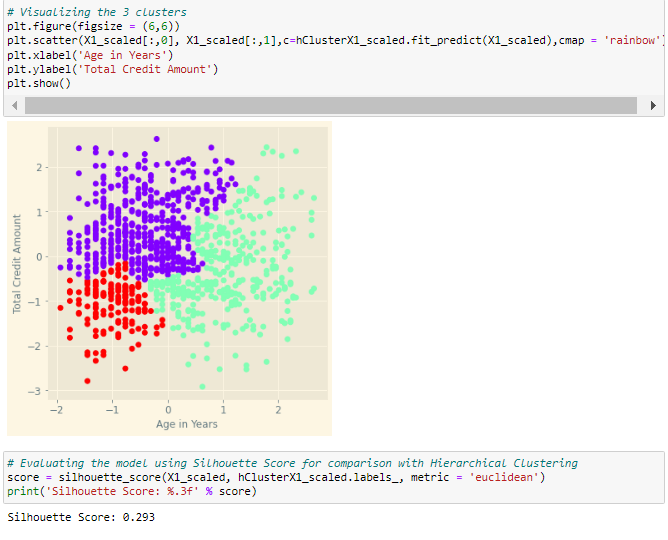
****

Figure 18.2

Hence, by using the three clusters and ward linkage to build the agglomerative clustering model before fitting the entries to their respective clusters, we could observe that the hierarchical clustering model is able to produce a Silhouette Score of 0.293.

#### Optimized Hierarchical Clustering Model 1B (Age and Credit Amount)

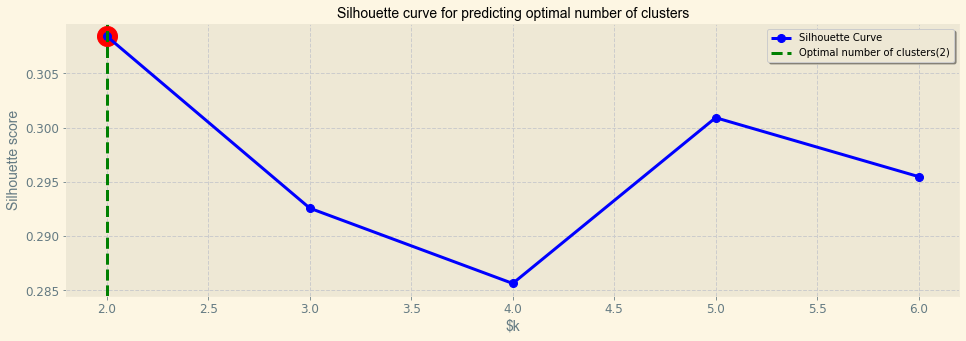
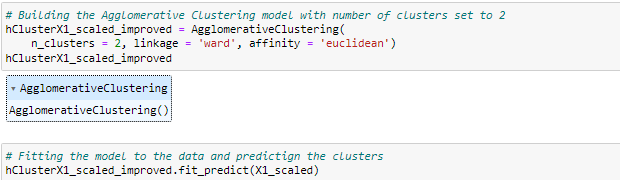


Figure 18.3

To optimize the following hierarchical clustering model, we could evaluate the various number of cluster values from 2 to 6 in order to determine the number of clusters which would produce the highest Silhouette Score by plotting a line chart as observed from Figure 18.3.



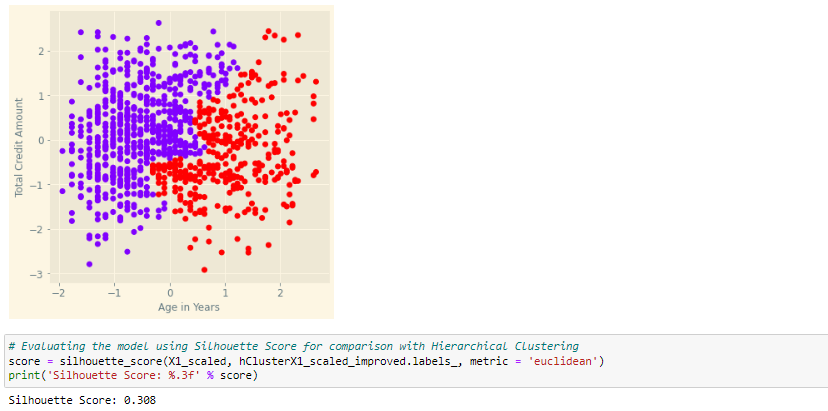


Figure 18.4

Thus, by creating a function to loop through a range of clusters from 2 to 6 and storing and plotting the list of Silhouette Scores, we could observe that cluster = 2 would provide us the highest silhouette score of 0.308 through observing the new hierarchical clustering model created which is slightly better than the initial model which had a Silhouette Score of 0.293.

#### Basic Hierarchical Clustering Model 2B (Age, Credit Amount, and Duration in Years)

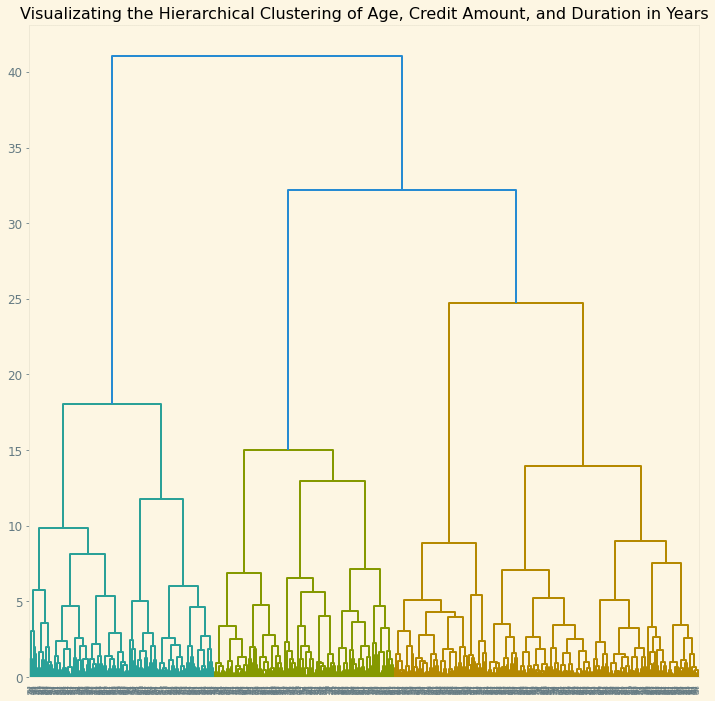
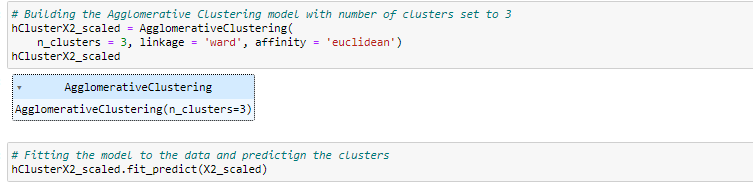
****

Figure 19.1

By observing Figure 19.1, which illustrates the dendrogram created using the X2\_scaled dataset through the Agglomerative Clustering Linkage, we could observe that the clusters value we should start off with would be three given that it cuts the maximum vertical distance observed.

****

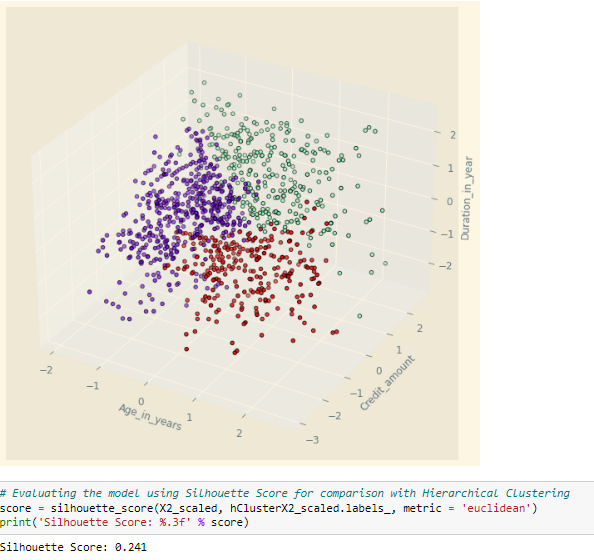
****

Figure 19.2

Hence, by using the three clusters and ward linkage to build the agglomerative clustering model before fitting the entries to their respective clusters, we could observe that the hierarchical clustering model is able to produce a Silhouette Score of 0.241 as observed from Figure 19.2.

#### Optimized Hierarchical Clustering Model 2B (Age, Credit Amount, Duration in Years)

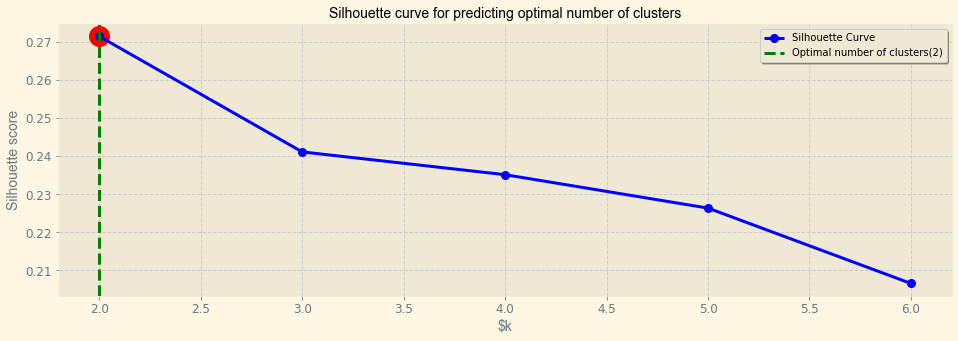
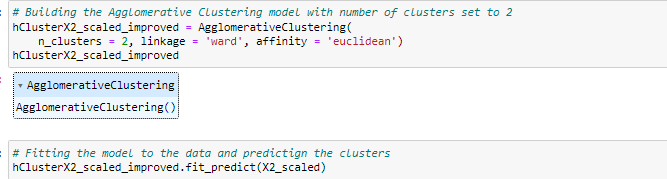


Figure 19.3

To optimize the following hierarchical clustering model, we could evaluate the various number of cluster values from 2 to 6 in order to determine the number of clusters which would produce the highest Silhouette Score by plotting a line chart as observed from Figure 19.3.



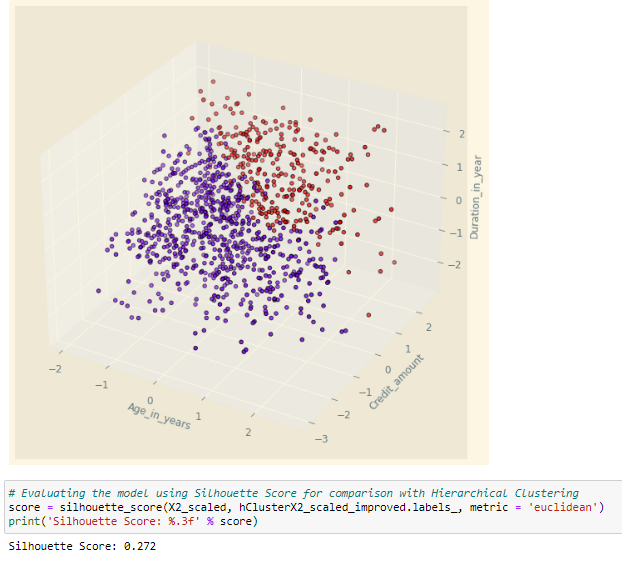


Figure 19.4

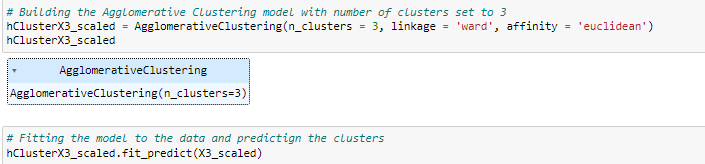
Thus, by creating a function to loop through a range of clusters from 2 to 6 and storing and plotting the list of Silhouette Scores, we could observe that cluster = 2 would provide us the highest silhouette score of 0.272 through observing the new hierarchical clustering model created which is slightly better than the initial model which only had a Silhouette Score of 0.241.

#### Basic Hierarchical Clustering Model 3B (Age, Present Employment Since, Credit History)

****

Figure 20.1

By observing Figure 20.1, which illustrates the dendrogram created using the X3\_scaled dataset through the Agglomerative Clustering Linkage, we could observe that the clusters value we should start off with would be three given that it cuts the maximum vertical distance observed.



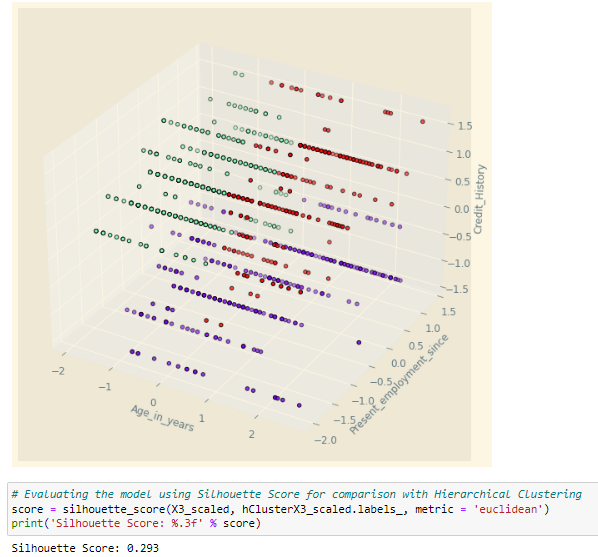


Figure 20.2

Hence, by using the three clusters and ward linkage to build the agglomerative clustering model before fitting the entries to their respective clusters, we could observe that the hierarchical clustering model is able to produce a Silhouette Score of 0.293 as observed from

#### Optimized Hierarchical Clustering Model 3B (Age, Present Employment, Credit History)

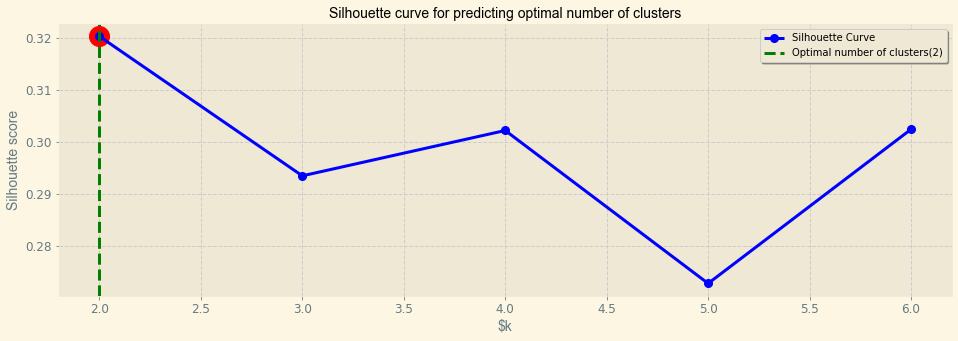
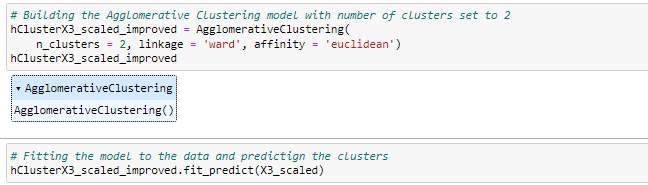
****

Figure 20.3

To optimize the following hierarchical clustering model, we could evaluate the various number of cluster values from 2 to 6 in order to determine the number of clusters which would produce the highest Silhouette Score by plotting a line chart as observed from Figure 20.3.

****

****

Figure 20.4

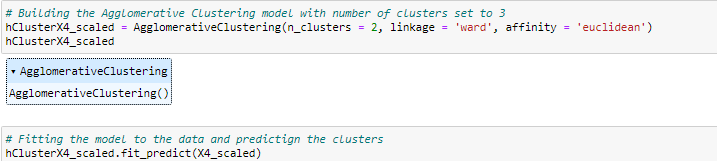
Thus, by creating a function to loop through a range of clusters from 2 to 6 and storing and plotting the list of Silhouette Scores, we could observe that cluster = 2 would provide us the highest silhouette score of 0.320 through observing the new hierarchical clustering model created which is slightly better than the initial model which only had a Silhouette Score of 0.293.

#### Basic Hierarchical Clustering Model 4B (Credit Amount, Number of Existing Credits, Credit History)

****

Figure 21.1

By observing Figure 21.1, which illustrates the dendrogram created using the X4\_scaled dataset through the Agglomerative Clustering Linkage, we could observe that the clusters value we should start off with would be two given that it cuts the maximum vertical distance observed.

****

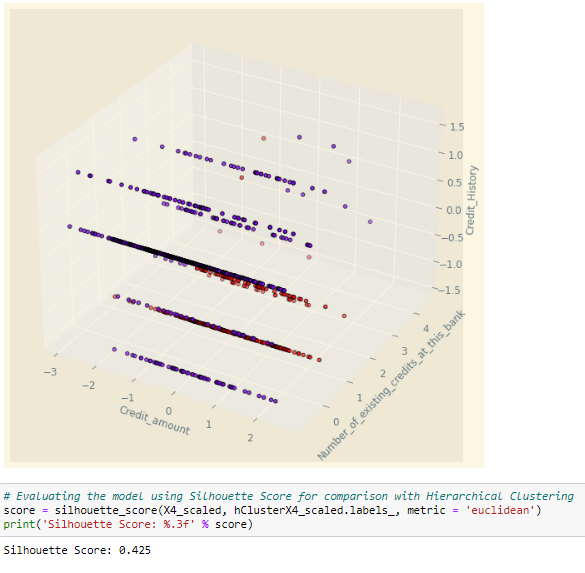
****

Figure 21.2

Hence, by using the two clusters and ward linkage to build the agglomerative clustering model before fitting the entries to their respective clusters, we could observe that the hierarchical clustering model is able to produce a Silhouette Score of 0.425 as observed from Figure 21.2.

#### Optimized Hierarchical Clustering Model 4B (Credit Amount, Number of Existing Credits, Credit History)

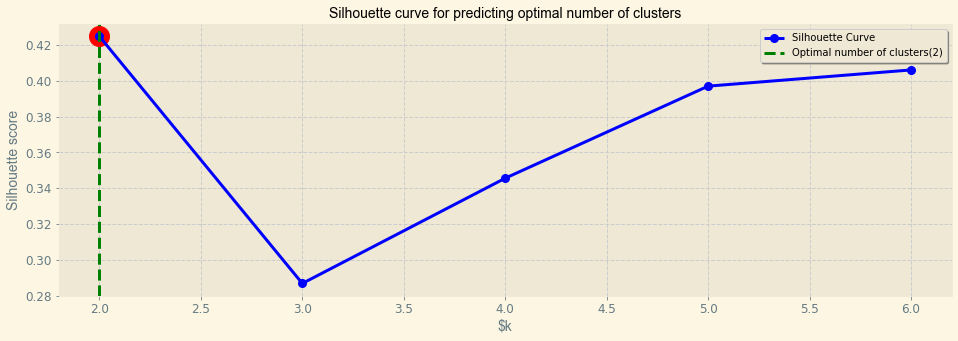
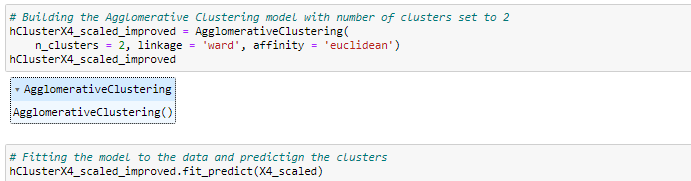
****

Figure 21.3

To optimize the following hierarchical clustering model, we could evaluate the various number of cluster values from 2 to 6 in order to determine the number of clusters which would produce the highest Silhouette Score by plotting a line chart as observed from Figure 21.3.



****

Figure 21.4

Thus, by creating a function to loop through a range of clusters from 2 to 6 and storing and plotting the list of Silhouette Scores, we could observe that cluster = 2 would provide us the highest silhouette score of 0.425 through observing the new hierarchical clustering model created which is exactly the same as the initial model given that both utilizes the same number of cluster.

#### Basic Hierarchical Clustering Model 5B (Account Status, Housing, Credit Amount)

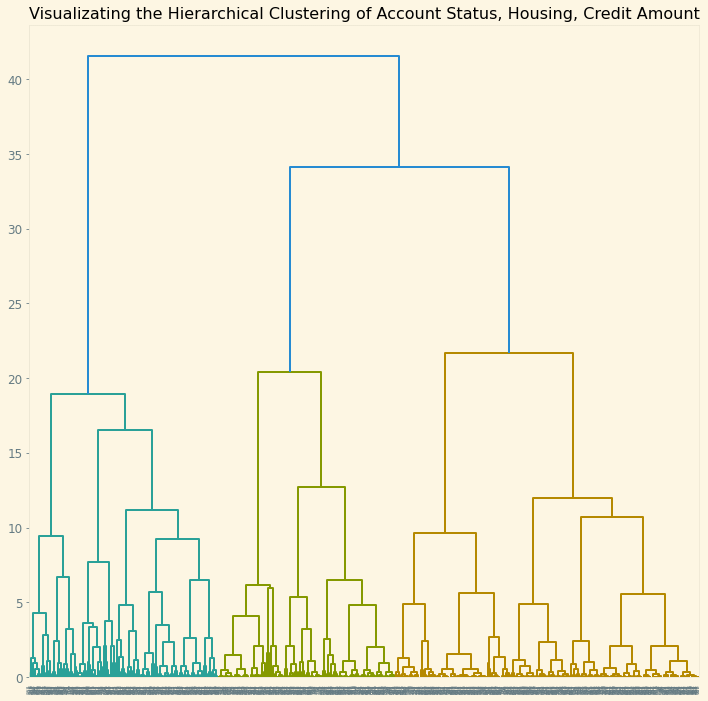
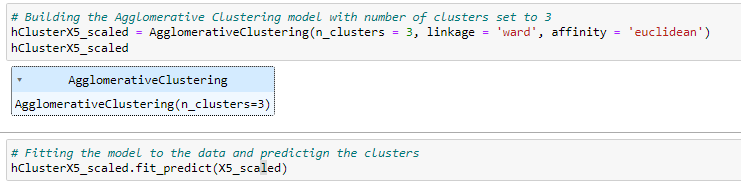
****

Figure 22.1

By observing Figure 22.1, which illustrates the dendrogram created using the X5\_scaled dataset through the Agglomerative Clustering Linkage, we could observe that the clusters value we should start off with would be three given that it cuts the maximum vertical distance observed.

****

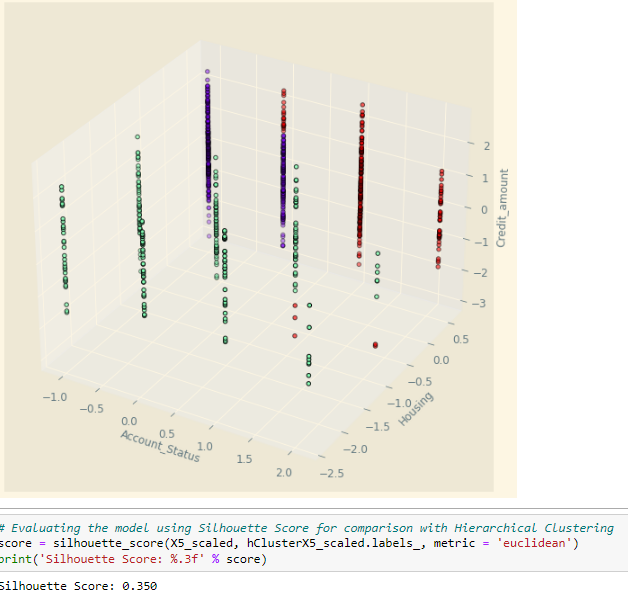
****

Figure 22.2

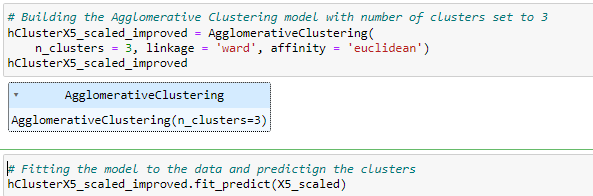
Hence, by using the three clusters and ward linkage to build the agglomerative clustering model before fitting the entries to their respective clusters, we could observe that the hierarchical clustering model is able to produce a Silhouette Score of 0.350 as observed from Figure 22.2.

#### Optimized Hierarchical Clustering Model 5B (Account Status, Housing, Credit Amount)

****

Figure 22.3

To optimize the following hierarchical clustering model, we could evaluate the various number of cluster values from 2 to 6 in order to determine the number of clusters which would produce the highest Silhouette Score by plotting a line chart as observed from Figure 22.3.

****

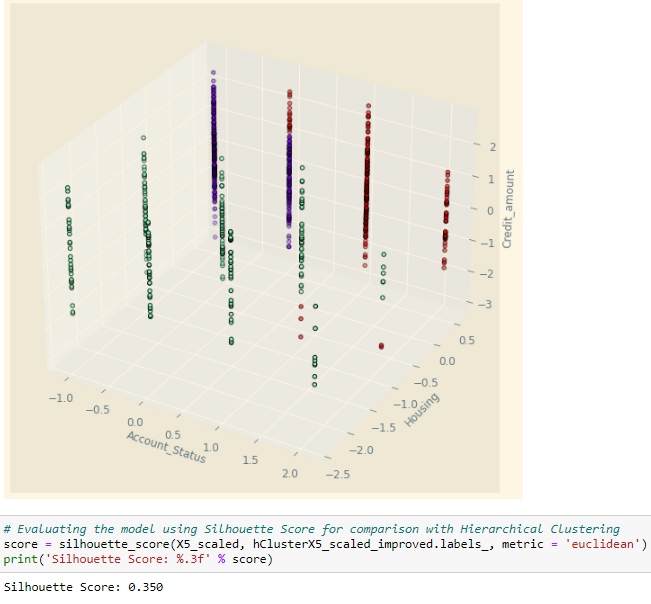
****

Figure 22.4

Thus, by creating a function to loop through a range of clusters from 2 to 6 and storing and plotting the list of Silhouette Scores, we could observe that cluster = 3 would provide us the highest silhouette score of 0.350 through observing the new hierarchical clustering model created which is exactly the same as the initial model given that both utilizes the same number of cluster.

## Evaluation and Comparison of the Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Features** | **Linkage** | **Silhouette Score** | **Remarks** |
| K-Means Clustering Model 1A | Age\_in\_Years + Credit\_Amount | NA | 0.3593 | All K-Means Clustering models have performed much better compared to the alternative Hierarchical Clustering models as observed from the Silhouette Score.  Will be used for further analysis from which to derive business insights. |
| K-Means Clustering Model 2A | Age\_in\_Years, Credit\_amount, Duration\_in\_year | NA | 0.3153 |
| K-Means Clustering Model 3A | Age\_in\_Years, Present\_employment\_since, Credit\_History | NA | 0.3603 |
| K-Means Clustering Model 4A | Credit\_amount, Number\_of\_existing\_credits\_at\_this\_bank, Credit\_History | NA | 0.4474 |
| K-Means Clustering Model 5A | Account\_Status, Housing, Credit\_Amount | NA | 0.358 |
| Hierarchical Clustering Model 1B | Age\_in\_Years + Credit\_Amount | Ward | 0.3084 |  |
| Hierarchical Clustering Model 2B | Age\_in\_Years, Credit\_amount, Duration\_in\_year | Ward | 0.2715 |  |
| Hierarchical Clustering Model 3B | Age\_in\_Years, Present\_employment\_since, Credit\_History | Ward | 0.3203 |  |
| Hierarchical Clustering Model 4B | Credit\_amount, Number\_of\_existing\_credits\_at\_this\_bank, Credit\_History | Ward | 0.4247 |  |
| Hierarchical Clustering Model 5B | Account\_Status, Housing, Credit\_Amount | Ward | 0.350 |  |

By collating the performance of the various K-Means and Hierarchical Clustering models, we could observe that for all the features that we have compared with, all the K-means models have outperformed the hierarchical clustering models in terms of the Silhouette Score. For instance, while K-Means Clustering Model 1A has a Silhouette Score of 0.3593, Hierarchical Clustering Model 1B only has a Silhouette Score of 0.3084. Thus, given that the K-means Clustering models have all outperformed their counterparts, we will be using them for further analysis from which to derive business insights.

# Summary and Interpretation

## Table Summary

|  |  |  |
| --- | --- | --- |
| K-Means Clustering Model 1A | | |
| **Clusters** | **Description** | **Analysis & Interpretation** |
| Cluster 0:  Low Age + Low Credit Amount | Median Age of approximately twenty-seven  Median Loan Amount of approximately $2,000 | Low credit amount leads to lower interest earned and revenue. Young Adults are not as profitable. |
| Cluster 1:  Medium Age + Medium-High Credit Amount | Median Age of approximately thirty-four  Median Loan Amount of approximately $5,200 | High credit amount leads to higher interest earned and revenue. Medium-Aged Adults Bank Customers are more profitable. |
| Cluster 2:  High Age + Low Credit Amount | Median Age of approximately forty-six  Median Loan Amount of approximately $2,000 | Low credit amount leads to lower interest earned and revenue. Senior Adults are not as profitable. |

|  |  |  |
| --- | --- | --- |
| K-Means Clustering Model 2A | | |
| **Clusters** | **Description** | **Analysis & Interpretation** |
| Cluster 0:  Medium Age in Years + Low Credit Amount + Low Duration in Years | Median Age of approximately 33 years  Median Credit Amount of approximately $2,000  Median Loan Tenure of approximately 1 Year | On average, middle-aged adults with smaller amounts of credit take shorter loan tenure.  Less profitable for banks due to lower interest earned from smaller credit and shorter time for incurring interest |
| Cluster 1:  Medium Age in Year + High Credit Amount + High Duration in Years | Median Age of approximately 33 years  Median Credit Amount of approximately $4,000  Median Loan Tenure of approximately 2 Years | On average, middle-aged adults with larger amounts of credit take longer loan tenure.  More profitable for banks due to larger amount of interest earned from larger credit and more interest charges over time |

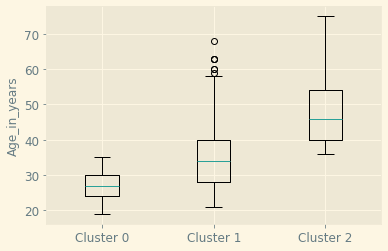
|  |  |  |
| --- | --- | --- |
| K-Means Clustering Model 3A | | |
| **Clusters** | **Description** | **Analysis & Interpretation** |
| Cluster 0: Medium Age, High Present Employment Since, High Credit History | Median Age of approximately 39, Median Years of Present Employment: > 7, Median Credit History: 2 (encoded) | We could observe that middle-aged customers with employment of more than 7 years since is among the groups least likely to default on loans |
| Cluster 1: Low Age, Medium Present Employment Since, High Credit History | Median Age of approximately twenty-seven, Median Years of Present Employment: 1 to 4, Median Credit History: 2 (encoded) | We could observe that young adults with employment of between 1 to 4 years since is among the groups least likely to default on loans |
| Cluster 2: (High Age, Medium Present Employment Since, High Credit History) | Median Age of approximately forty-four, Median Years of Present Employment: 1 to 4, Median Credit History: 2 (encoded) | We could observe that senior-aged adult with employment between 1 to 4 years is among the groups least like to default on loans |
| Cluster 3: (Medium Age, Medium Present Employment Since, Low Credit History) | Median Age of approximately thirty, Median Years of Present Employment: 1 to 4, Median Credit History: 0 (encoded) | We could observe that middle-aged adult with employment between 1 to 4 years is among the groups most likely to default on loans |
| Cluster 4: High Age, High Present Employment Since,  Low Credit History | Median Age of approximately 44, Median Years of Present Employment: >7, Median Credit History: 0 (encoded) | We could observer that senior-aged adult with employment of more than 7 years since is among the groups most likely to default on loans |

|  |  |  |
| --- | --- | --- |
| K-Means Clustering Model 4A | | |
| **Clusters** | **Description** | **Analysis & Interpretation** |
| Cluster 0: Medium Credit Amount, Low Number of Existing Credit, High Credit History | Median Credit Amount: Approximately $3800, Median no. of Credit: one, Median Credit History: 2 (Encoded) | In general, customers with multiple credits have a higher chance of defaulting on loans compared to customers holding onto a single loan regardless of the amount of money borrowed |
| Cluster 1: Low Credit Amount, High Number of Existing Credit,  Low Credit History | Median Credit Amount: Approximately $2000 , Median no. of Credit: two, Median Credit History: 0 (Encoded) |
| Cluster 2: High Credit Amount, High Number of Existing Credit, Low Credit History | Median Credit Amount: Approximately $4900, Median no. of Credit: two, Median Credit History: 0(Encoded) |
| Cluster 3: Low Credit Amount, Low Number of Existing Credit,  High Credit History | Median Credit Amount: Approximately $2000, Median no. of Credit: one, Median Credit History: 2(Encoded) |
| Cluster 4: Medium Credit Amount, Low Number of Existing Credit, Low Credit History | Median Credit Amount: $2500, Median no. of Credit: one, Median Credit History: 0(Encoded) | Cluster four is an exception whereby customers holding on to one loan also have a poor credit history |
| Cluster 5: Medium Credit Amount, High Number of Existing Credit, High Credit History | Median Credit Amount: $2900, Median no. of Credit: two, Median Credit History: 2(Encoded) | Cluster five is another exception whereby customers hold on to more than one line of credit but have a good credit history |

|  |  |  |
| --- | --- | --- |
| K-Means Clustering Model 5A | | |
| **Clusters** | **Description** | **Analysis & Interpretation** |
| Cluster 0: Medium Account Status, Medium Housing, High Credit Amount | Median account Status of approximately 1  Median Housing of approximately 1  Median Credit Amount of $3,000 | Represent the least interest among the bank given that customers in this cluster rent their housing. As such, housing may not be used as collateral. |
| Cluster 1: High Account Status, High Housing, Low Credit Amount | Median account Status of approximately 2  Median Housing of approximately 2  Median Credit Amount of $2,000 | Best group of customers to target due to housing which may be used as collateral alongside consistent salary assignment |
| Cluster 2: Low Account Status, High Housing, Low Credit Amount | Median account Status of approximately 0  Median Housing of approximately 2  Median Credit Amount of $2,300 | Second Best group of customers to target due to housing which may be used as collateral, but no salary assignment with the bank |

## Interpretation of Each Cluster

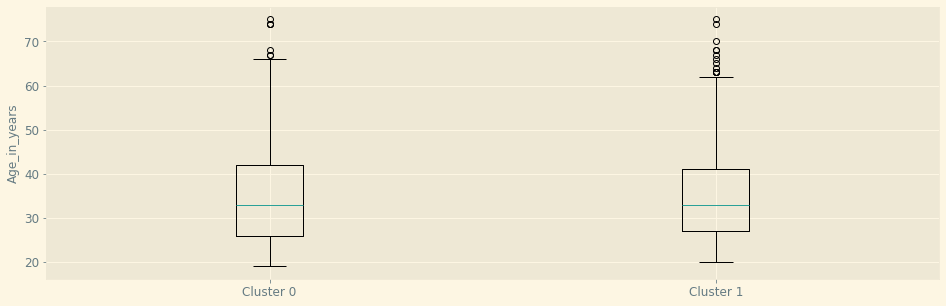
### K-Means Clustering Model 1A

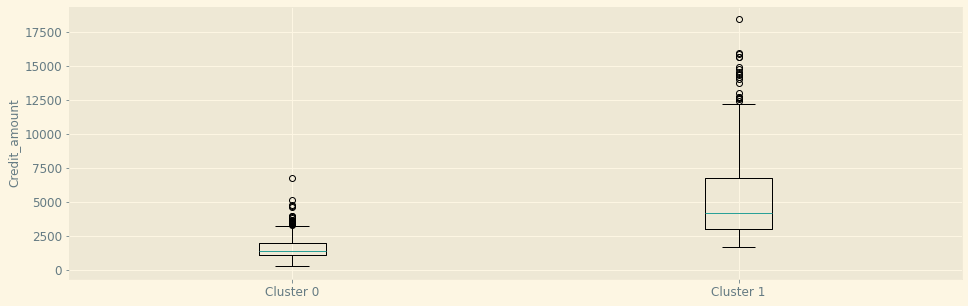


By observing the boxplot distribution of the three clusters as part of K-means clustering model 1A, we could observe that Cluster 0 has a low median age of twenty-seven with a low median credit amount of approximately $2000. Whereas for Cluster 1, customers within it have a average median age of 34 with a Medium to High median credit amount of $5,200. Additionally, for cluster two, we could observe that customers within it have a high median age of forty-six, with a low median credit amount $2,000. Thus, through analyzing the observation, we could see that Medium-aged customers are much more profitable compared to young and senior adult customers given that would take on higher credit amount which implies a larger amount of interest earned for the bank, while less is earned from young adult and seniors given they borrow lower amount of credits which leads to lower interest earned as well.

Thus, in terms of the actionable plans, our recommendation for the bank would be to increase the interest rate for loans lent to young and senior adult customers given that the amount of credit earned is low because of the low credit amount that they would take. Hence, this would assist the banks in increasing their revenue from high interest rate small loans. Additionally, for medium aged customers, given that they are likely to borrow large credit amounts, the bank could potentially attract more medium-aged customers to borrow copious amounts through promotional discounted interest rates of loan. Thus, this would allow the bank to earn a larger amount of interest from large loans taken which would outweigh the cost from lowering interest rate slightly. However, as part of our consideration, to remain competitive, the price elasticity of demand for consumer credit should be calculated comprehensively to determine the magnitude of change in interest rate so that customers would not find alternatives.

### K-Means Clustering Model 2A

****

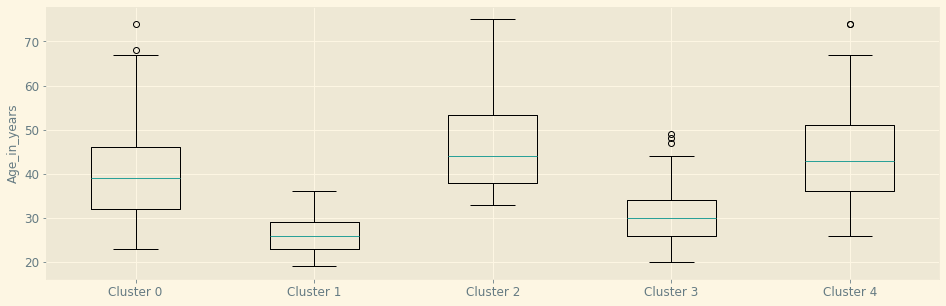
****

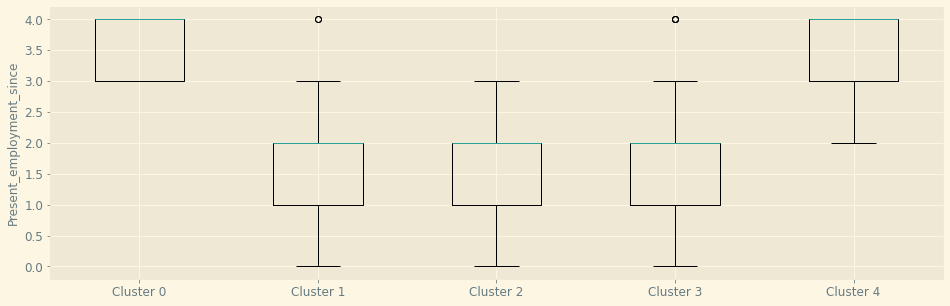
****

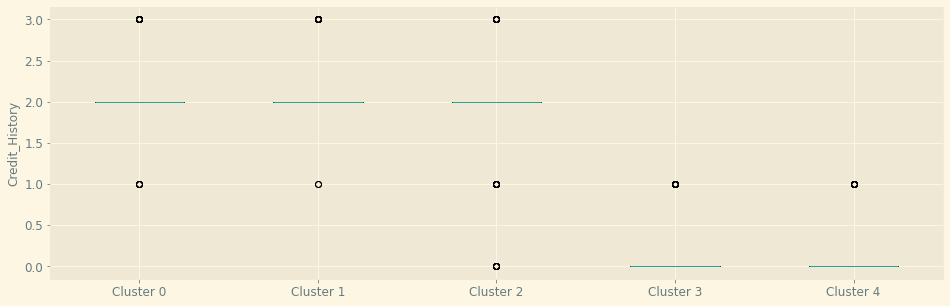
By observing the boxplot distribution of the two clusters as part of K-Means Clustering model 2A, we could observe that Cluster 0 has a average median age of 33 years, a relatively low median credit amount of $2,000, and a relatively low median loan tenure of 1 year. Whereas for Cluster 0, we could observe that the customers within it have an average median age of 33 years as well, a high median credit amount of $4,000, with a high median loan tenure of 2 years. Thus, through analyzing the two clusters, we could see that an average age person who takes a small credit is likely to take short loan tenure while those who take a large credit amount is likely to take a longer loan tenure. Thus, we conclude that those who take a small credit amount with short loan tenure are not as profitable compared to customers who take a larger credit amount and longer loan tenure given that the bank is able to earn more interest from cluster one through larger amounts of interest charged and frequency of interest charged over time.

Hence, our objective for cluster zero would be to attract such customers to take on longer loan tenure. Thus, one of the solutions would be to offer new loan products that have a lower interest rate when taking longer loan tenures while also making such new loan products collateralized to ensure that the bank does not take on additional risk of default. Additionally, to attract customers to take longer loan tenure, marketing campaigns could be held to promote them by explaining the benefits of it (e.g., better cash flow management). As for cluster one, given that they are already profitable customers due to the fact they take large amounts of loan in the span of longer loan tenure, our objective would be to increase customer loyalty for cluster 1. Hence, to strengthen the customer base for cluster one, the bank could potentially offer loan interest rebates for customers who have managed to maintain a good credit history. Thus, this would allow the bank to maintain and increase demand for such loan products while at the same time not taking additional credit risk by filtering customers.

### K-Means Clustering Model 3A

****

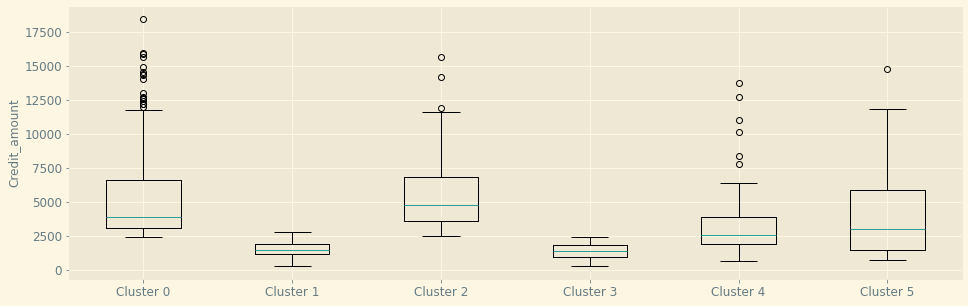
****

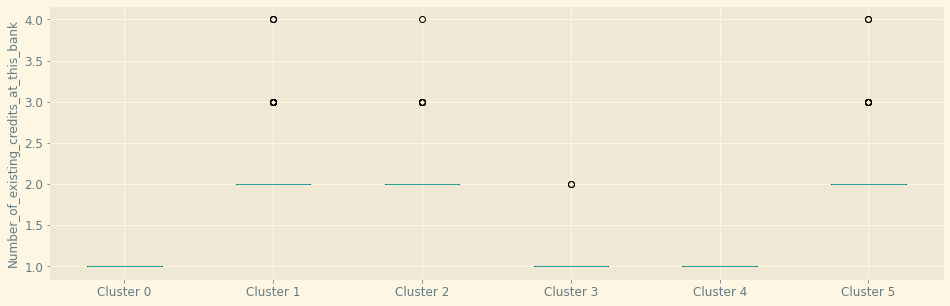
****

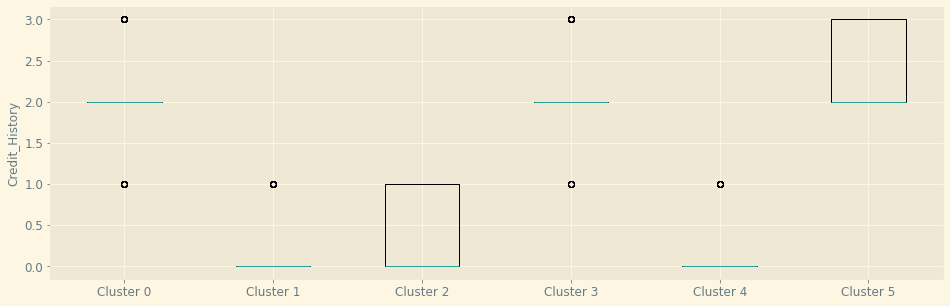
By observing the box plot distribution of the five clusters as part of K-Mean Clustering model 3A, we could observe that Cluster 0 has an average median age of 39 years, a high median present employment since of more than 7 years, and with a high median credit history of two which has been encoded. Whereas, for cluster one, we could observe that the customers within it have a low median age of 27 years, a relatively average median present employment of 1 to 4 years, and a high median credit history of 2 which has been label encoded. As for cluster two, we could observe that the customers within it have a high median age of 44 years, a relatively average median present employment of 1 to 4 years, and a high median credit history of 2 which has been label encoded. Subsequently for cluster three, we could observe that the customers within it have an average median age of 30 years, a relatively average median present employment of 1 to 4 years, and a low median credit history of 0 which has been label encoded. Finally for cluster four, we could observe that the customers within it have a high median age of 44 years, a relatively high median present employment of more than 7 years, and a low median credit history of 0 which has been label encoded. Thus, through analyzing the five clusters we could break them down into their respective age groups such that for young adult customers, they pose relatively low risk of credit default as observed from their high credit history in spite of the fact that they only have 1 to 4 years of employment. Whereas for medium aged customers, those with employment between 1 to 4 years are likely to default on loans as evident of their poor credit history while those with more than 7 years of employment are least likely to default on loans as evident of their high credit history. As for senior adults, the employment since has an opposite effect on their credit history as compared to medium aged customers given that those with more than 7 years of employment are more likely to default compared to those employed for only 1 to 4 years.

Hence, our objective for senior aged customers with more than 7 years of employment would be to reduce the risk of loan default which could potentially be achieved by reducing their debt servicing ratio at the banks to 10% of their gross monthly income. Hence, this would prevent customers from borrowing more than they could afford to effectively. As for medium aged customers with employment between 1 to 4 years, our objective for this group of customers would also be to reduce their risk of loan default which could potentially be achieved by limiting such customers to only collateralized and or small loans. Hence, this would aid the bank in minimizing their credit risk for this group of customers which are likely to default on their loans. Finally, given that young adult customers are less likely to default on their loan, our objective for this group of customers would be to attract more young adults to take on loans at the bank given that they do not pose much risk of credit default. Hence, this could be achieved by offering innovative loan products such as a work-study program loan that encourage working young adults to continue upskilling through financial support provided.

### K-Means Clustering Model 4A

****

****

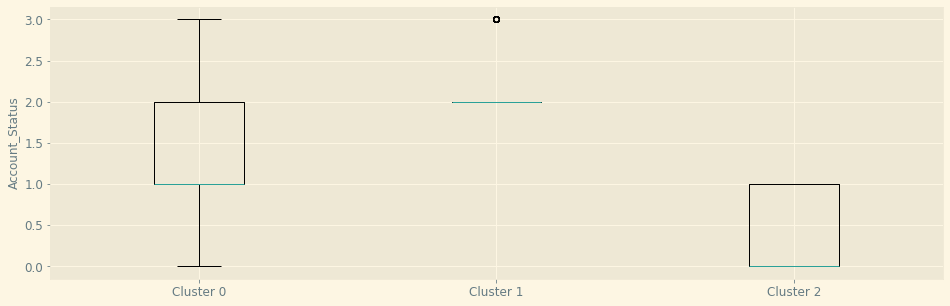
****

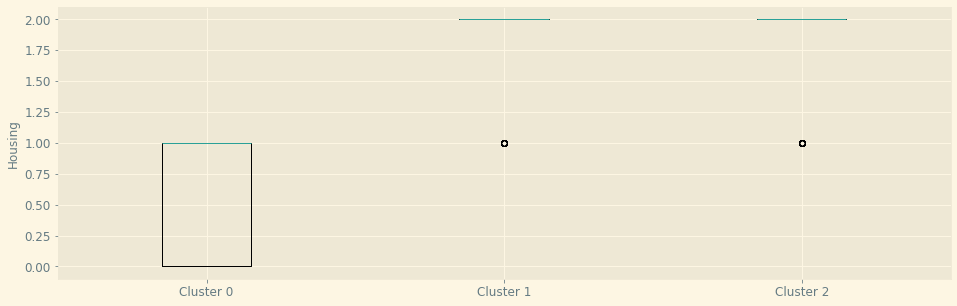
By observing the box plot distribution of the 6 clusters as part of the K-Means Clustering model 4A, we could observe that Cluster 0 has a average median credit amount for $3,000, a low median number of credit of 1, and a high median credit history of 3 which has been label encoded. Whereas, for cluster 1, we could observe that the customers within it have a low median credit amount of $2,000, a high median number of existing credit of 2, and a low median credit history of 0 which has been label encoded. As for cluster 2, we could observe that the customers within it have a high median credit amount of $4,900, a high median number of existing credit of 2, and a low median credit history of 0 which has been label encoded. Subsequently for cluster 3, we could observe that the customers within it have a low median credit amount of $2,000, a low median number of existing credit of 1, and a high median credit history of 2 which has been label encoded. After which, for cluster 4 and 5 which form the exceptions of our interpretation, we could observe that they both have a medium credit amount of $2,500 and $2,900 respectively, a number of existing credits of 1 and 2 respectively, and a median credit history of 0 and 2 respectively.

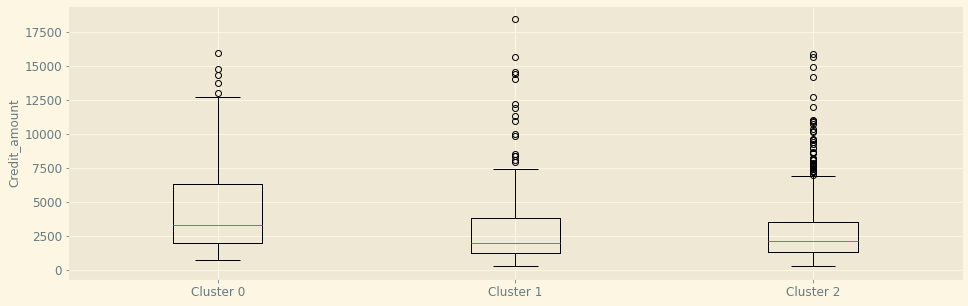
Thus, through analyzing the six clusters we could break them down into two groups which are mainly customers with a single line of credit and customers with multiple lines of credit given that through our analysis, credit amount does not play an effective role in determining credit history of a customer. Thus, in general we could observe that customers with multiple lines of credit are more likely to default on loans compared to those with a single line of credit as observed from their respective credit history.

Hence, as our goal for customers with multiple lines of credit would be to reduce their risk of default on loans as evident of their credit history, and in view of the fact that our objective for customers with a single line of credit would be to encourage them to take up 1 and not more 1 line of credit, our recommendation for the bank would be offer an attractive loan through discounted interest for bank customers who have not take up their first loan while increasing the interest on subsequent lines of credit taken. Hence, this would disincentivize customers from taking more than one line of credit. Rather than disincentive customers from taking more than one loan, another way would be to only limit customers to collateralized loans if they wish to take subsequent loans. Thus, this would increase the revenue of banks while also reducing their credit risk from the customer’s likelihood to default on loans.

### K-Means Clustering Model 5A







By observing the boxplot distribution of the three clusters as part of K-Means Clustering Model 5A, we could observe that Cluster 0 has a relatively average median Account Status of 1 (1 year of salary assignment, a relatively average median Housing of 1 (rental), alongside a relatively high credit amount of approximately $3,000. Whereas for Cluster 1, customers within it have a high account status of 2 (2 year of salary assignment), a high median Housing of 2 (owned), alongside a low median credit amount of approximately $2,000. Additionally, for cluster two, we could observe that customers within it have a relatively low median account status of 0 (No Checking Account), a relatively high median housing of 2 (owed), alongside a relatively low median credit amount of approximately $2,200.

Thus, in terms of the actionable plans the bank could take, given that cluster 1 would represent the least credit risk to the bank owing to the fact that they have a consistent salary assignment of 2 years while they also own a home which may be utilized as a collateral, loan interest rebates could be offered to such customer segment as it would be a generally risk adverse action to take. As for cluster two, it would represent the second least credit risk given that their housing may be used as a collateral. However, they do not have any salary assignment in that bank which may be attributed to the fact that such customers rely on other banks instead. Hence, for such customers, more marketing effort could be done for such customers to make the switch to the bank that they had taken a credit from. Thus, this would encourage such individuals to make use of their bank services rather than the services of their competitors. Finally, Cluster 0 would represent the highest credit risk to the bank given that although they have a consistent salary assignment of at least 1 year, there is limited collateral which may be utilized given that they rent instead of owning a home. Thus, for such customers, we would recommend the bank to prevent them from taking large loans as they are likely to default on loans and may not be able to compensate accordingly.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

**THE END**