

**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: | P02 |
| Student Name: | Luismika Lim Chieng Hee |
| Student Number: | S10223527J |

**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

Contents

[Deadline for Submission: 1](#_Toc105367145)

[1. Overview 3](#_Toc105367146)

[2. Building the models 4](#_Toc105367147)

[2.1. Model 1 (K-Means) 7](#_Toc105367148)

[2.2. Model 2 (Hierarchical Clustering) 10](#_Toc105367149)

[2.3. Model 3 (K-Means) 14](#_Toc105367150)

[2.4. Model 4 (Heirarchical Clustering) 17](#_Toc105367151)

[3. Summary 21](#_Toc105367152)

# Overview

The goal of this report is to help deliver insights from the findings made through the given dataset using clustering methods like K-Means and Hierarchical Clustering. With the aid of elbow method to find the optimal number of clusters for K-Means and the dendrogram to find the optimal number of clusters for Hierarchical clustering, we will create models to cluster data and at the same time utilise sum of squared errors (SSE) and silhouette score to assess the quality of the data clusters and finally, we will make findings on the dataset and then generate insights.

The dataset used for this report involves customer related data that is used by banks and from this dataset assumptions will be made about the meaning of the data provided and any ambiguity or assumption will be clarified here. So as for the account status attribute, we assume it is indicative of how long has the bank account been receiving salary payouts. For duration in month, we assume that the duration refers to the loan duration made available to the customer.

As for the structure of the report, cluster models using K-Means and hierarchical clustering will be built. First, clustering data for K-Means will be loaded using the elbow method to get the optimal number of clusters based on SSE scores and then calculate the silhouette score of the cluster produced. Next, using the same set of attributes, agglomerative clustering will be performed on the data using different linkage methods and the most suitable one will be chosen by referring to the dendrogram produced. Afterwards, the data will be built based on the number of clusters and displayed on a scatter to help with visualisation of the scatter. We will then check if the model can be improved by plotting silhouette scores on a bar chart to compare the silhouette scores for the range of clusters (which in this report we will use a range of up to 11). If the optimal silhouette score requires a different number of clusters to be built then we will improve the model accordingly. If we have already achieved the optimum silhouette score with our current hierarchical model, we will then proceed to compare the silhouette score with that of the K-Means model and the model with a higher silhouette score would be chosen. However, if there is little or no useful findings derived from the model, the next best model would be chosen for analysis.

# Building the models

First we need to import the different libraries

Text, letter

Description automatically generated

**Fig 1.1**

Then, we explore the data using head() to have a brief look at how the dataset looks like.

Table

Description automatically generated

**Fig 1.2**

We can use dat.describe() to find out information like mean and standard deviation of the numeric data.

Table

Description automatically generated

**Fig 1.3**

dat.info() shows the data types of all the attributes and allows us to know which attributes have numeric data and which attributes have object data.

Text, table

Description automatically generated with medium confidence

**Fig 1.4**

If we want to use the attributes of object datatype, we can convert them to numeric so that we can perform modelling using the data.

Graphical user interface, text

Description automatically generated

**Fig 1.5**

## Model 1 (K-Means)

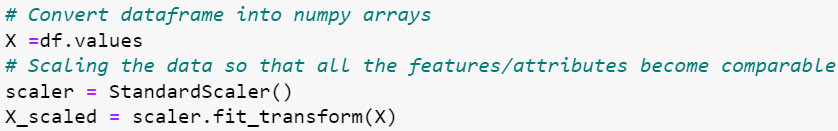
Firstly we will compare the Duration and the Credit Amount attribute.

Table

Description automatically generated

**Fig 2.1**

Afterwards, we convert the dataframe into numpy arrays and then scale the data so that the features are comparable.



**Fig 2.2**

To find the optimal amount of clusters, we can plot a graph compare the number of clusters to be modelled and the SSE score of the model. Through the graph, we can find the inflexion point of the graph to determine the optimal number of clusters to implement inside the model.

Graphical user interface, text, application

Description automatically generated

**Fig 2.3**

Chart, line chart

Description automatically generated

**Fig 2.4**

From the results of the elbow method we can deduce that the optimal number of clusters is 2. As such we create a K-Means scatter with 2 clusters as shown below

Graphical user interface, text, application

Description automatically generated

**Fig 2.5**

Chart, scatter chart

Description automatically generated

**Fig 2.6**

From the scatter plot we can see that the graph is split into two and the two red dots represents the centroids which are also the cluster centers. From this scatter we can deduce that the there is a positive relationship between credit amount and duration, the longer the loan duration, the higher the credit amount.

We can check the SSE score using kmeans.inertia\_ and through the score we can evaluate whether the error is too big and from its current score of 885 it is still acceptable.

Graphical user interface, text

Description automatically generated with medium confidence

**Fig 2.7**

Finally, we calculate the silhouette score for the first K-Means model.

**Graphical user interface, text, application

Description automatically generated**

**Fig 2.8**

## Model 2 (Hierarchical Clustering)

Firstly, we insert the scaled data into the algorithm which calculates the distances between data objects and merge the clusters using the specified linkage method, which in this case is “ward”. As for ward method, the clusters are grouped by their sum of squared error (SSE) values

Table

Description automatically generated

**Fig 3.1**

We then use a dendrogram to visualise the hierarchical clustering and to see how many clusters there are. It is indicated in the diagram by different colours. In the code below we can edit the size of the visualisation diagram using plt.figure() and its title using plt.title(). We derive the number of clusters by counting the number of vertical lines in the dendrogram cut by a horizontal line that can traverse the maximum distance verticaly without intersecting a cluster.

Text

Description automatically generated

**Fig 3.2**

Chart, histogram

Description automatically generated

**Fig 3.3**

Next, we build the clustering model using AgglomerativeClustering function from the sk.cluster library and indicate the number of clusters we want and then fit the data inside the model and then use a scatter plot to visualise the clusters

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

**Fig 3.4**

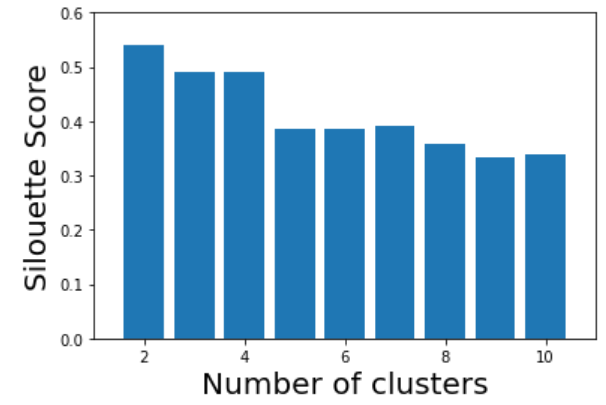
From the clustering we can deduce that credit amount and duration have a positive relationship, when duration increases in length, credit amount also increases.

We then evaluate the sillhoette scores for different number of clusters using a bar graph to find the number of clusters that produces the optimal score.

Graphical user interface, text, application

Description automatically generated

**Fig 3.5**



**Fig 3.6**

Based in the bar chart, using 2 clusters produces the highest silhouette score for this model and as such, no improvement needs to be made to the model in terms of its number on clusters.

This is the final silhouette score of the 1st Hierarchical model

Text

Description automatically generated

**Fig 3.7**

## Model 3 (K-Means)

For the second K-Means model, we will compare the credit amount and age attributes.

Table

Description automatically generated

**Fig 4.1**

Afterwards, we will scale the data to ensure that the attributes have equal weightage such that a change in one of the attributes will have an equivalent change in the other attribute. Credit amount has values in the thousands and age has values in the tens, as such scaling will be useful for comparing these two attributes.

A picture containing text

Description automatically generated

**Fig 4.2**

We will use the elbow method to find the optimal number of clusters by comparing the SSE scores of the model that uses different number of clusters. We want to use the number of clusters that adding another cluster does not give a much better modelling of the data as having too many clusters would defeat the purpose of clustering in the first place so we use the elbow method to find the point of inflexion to discover the optimal number of clusters.

Text

Description automatically generated

**Fig 4.3**

Chart, line chart

Description automatically generated  
**Fig 4.4**

From the diagram we can deduce that 3 (the inflexion point) is the optimal number of cluster for this model.

We then build the K-Means model and then fit the scaled data into the model. The centroids are highlighted in red for identification and easier visualisation. Kmeans.predict is used to group data to their clusters by assigning them to their closest centroid. The clusters are then visualised using the scatter.

Text

Description automatically generated

**Fig 4.5**

Chart, scatter chart

Description automatically generated

**Fig 4.6**

From this scatter we can deduce that more customers who have higher credit amounts are of lower age and that people of higher age are at more concentrated at the extreme ends in terms of their credit amount.

We then evaluate the SSE Score of the K-Means model

A picture containing text

Description automatically generated

**Fig 4.7**

Finally, we evaluate the silhouette score of last K-Means model

**Graphical user interface, text, application

Description automatically generated**

**Fig 4.8**

## Model 4 (Heirarchical Clustering)

Firstly, we add scaled data into the algorithm and define the linkage method used which will determine the clustering of the model.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig 5.1**

We then use a dendrogram to visualise the hierarchical clustering and determine the number of clusters to use. The number of clusters is determined by the number of vertical lines cut by a horizontal line that can traverse the maximum distance vertically without intersecting a cluster. It is indicated for us in the visualisation through the use of different colours to represent the different clusters.

Chart

Description automatically generated

**Fig 5.2**

Next we build the hierarchical clustering using the number of clusters we determined from the dendrogram followed by inserting the scaled data into the model and finally we visualise the model using a scatter plot.

Text

Description automatically generated

**Fig 5.3**

Chart, scatter chart

Description automatically generated

**Fig 5.4**

Then, we evaluate the number of clusters by comparing this model with other models that use different number of clusters by their silhouette score and finally we plot it on a bar graph visualisation for easier comparison.

Graphical user interface, text, application, email

Description automatically generated

**Fig 5.5**

Chart, bar chart

Description automatically generated

**Fig 5.6**

From this bar graph we can gather that the number of clusters used in the model, 3 is optimal and no further changes is required to be made to the number of clusters in the current model.

This is the final silhouette score of the 2nd Hierarchical model

Text

Description automatically generated

**Fig 5.7**

# Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Number of Clusters | Attributes | Linkage  Method | Silhouette Score | Remarks |
| Model 1  (K-Means) | 3 | Duration & Credit Amount | NA | 0.58 |  |
| Model 2  (Hierarchical) | 3 | Duration & Credit Amount | Ward | 0.49 |  |
| Model 3  (K-Means) | 2 | Credit Amount & Age | NA | 0.54 |  |
| Model 4 | 2 | Credit Amount & Age | Ward | 0.48 |  |

## Justifying models chosen for evaluation

Intepretations and insights can be derived from all 4 models. However only two models will be chosen to be interpreted and evaluated. Both models picked will be models that are using different attributes and have the highest silhouette score among the models that use the same attributes. As such the models we will be evaluating are models 1 & 3. Notice that both models are K-Means Models.

## Model 1

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Fig 6.1 Fig 6.2**

From duration boxplot, we can see that cluster 0 has a mean of approxmimately 15 and upper and lower quartiles of approximately 24 and 11 respectively. Cluster 1 has a mean of about 35 and upper and lower quartiles of about 48 and 30 respectively. The spread of cluster 0 is smaller than that of cluster 1 most of cluster 0 is made up of lower duration than that of cluster 1. Majority of the data in cluster 1 including the mean is in between the 30 and 50 range.

From Credit Amount boxplot, there are many outliers for cluster 0 which have higher credit amounts than the cluster itself. Cluster 1 has a mean of about 2000 and an upper quartile of approximately 2500 and a lower quartile of approximately 1500. Cluster 2 has a mean of about 6500 and an upper quartile of approximately 8500 and a lower quartile of approximately 4000.There are a few outliers for cluster 2 with higher credit amounts and one outlier with an exceptionately high credit amount.

**What are the findings we can derive from the model?**

Cluster 0 consist of customers with lower credit amount and lower loan duration as compared to cluster 1. Cluster 0 has a smaller, tighter spread as compared to cluster 1. Based on their loan duration and credit amount, customers in cluster 0 are more likely to be of the lower-income bracket as they are provided with less credit amount and a shorter loan duration. The bank can offer to implement financial products like prepaid cards and check-cashing. Customers in cluster 1 have a higher credit amount and higher duration which implies that they are of middle to high income brackets. The banks can try to find out more about the preferences of this group of customers by observing their purchases of big-ticket items or even their daily spendings to up-sell services and credit cards to these customers. The bank can also offer these customers a premium banking option where they can be served and attended to seperately and provide faster and more personalised services.

## Model 3

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

**Fig 6.3 Fig 6.4**

From the credit amount boxplot we can see that cluster 0 and cluster 2 are quite similar their spread and mean are very close and they have nearly the same range. Cluster 1 however, differs drastically from the other two. It has a range that lies almost completely outside of the range of cluster 0 and 2. However, notice that cluster 1 has a mean that is near the lower end of the range of cluster 1. Cluster 1 has an excpetionally high anomaly with a very high credit amount. Cluster 0 has a mean of around 2000 and a upper and lower quartile of approximately 2700 and 1000 respectively. Cluster 1 has a mean of around 7600 and a uppper and lower quartile of approximately 10000 and 6500 respectively. It has an exceptionately high outlier with 17500 credit amount. Cluster 2 has a mean of 2000 and an upper and lower quartile of approximately 2700 and 1000 respectively.

From the age boxplot diagram we can observe that cluster 0 and 2 are at two ends of the age spectrum while cluster 1 has a rather large range that overlaps both cluster 0 and 2. Cluster 0 has a range of about 18 to 40 while it has a mean of around 29. Cluster 1 has a upper quartile of around 42, a lower quartile of about 27 and a mean of approximately 24. Cluster 2 has a mean of 49 and a upper and lower quartile of approximately 56 and 45 respectively. Cluster 0 and cluster 2 are clusters indicative of very different age groups as they do not overlap one another. Cluster 1 on the other hand consists of customers of all ages. Although cluster 1 has a large range its mean and upper and lower quartiles are considerably lower than the highest values within its range.

**What are the findings we can derive from the model?**

Cluster 0 consists of customers who have a low credit amount and are of young to middle age. Similarly, cluster 2 consists of customers who have a low credit amount. However, the customers in cluster 2 are of older age, ranging from 40 years to 70 over years. Customers in cluster 1 have high credit amounts ranging from 5000 to 15000 with the exception of a few anomalies. And has an age range of between 20 and 60. Based on this information we can can infer that cluster 0 consists of young people that have only just started their careers and are mostly people working in junior postions in their respective organisations and as such draw a lower salary which could explain their lower credit amounts. As for customers in this cluster, the bank can consider offering mortgage loans as many people in this cluster’s age group are looking to buy their first home. They can also look to offer financial advice through seminars and talks to young adults looking to kickstart their lifelong financial management journey to make a name for themselves and to win the trust of their customers and to maintain customer loyalty. They can also look to sell these financial services through social media as the younger generation are more tech savvy and advertising through social media can reach a large portion of customers in this cluster. Cluster 1 consists of a range of high income earners and looking at the upper and lower quartiles, it mainly consists of middle aged people who are at the peak of their careers where they are drawing in high salaries and have the highest spending power. There are also some younger individuals in their 20s that are inside the high-income bracket. The bank can offer more ways for these high income earners to spend by offfering lines of credit like tiered credit cards where cardholders have member priviliges such as premium lounges in airports or receive VIP treatment at hotels and clubs. As customers in this cluster probably have more disposable income, the bank can offer higher rebates for higher expenditure. The bank can also cross-sell auto loans to the customers as they are more likely to make vehicle purchases. Cluster 2 consists of customers with lower credit amounts and are of older age. For this group of customers, the bank can offer to help them learn how to navigate the new digital banking landscape through traditional media like television, radio and posters. The banks can also station staff specialy catered for to help elderly with their banking related issues at bank branches.