Li Tao

850001413

GDDA708 Machine Learning and AI

Assessment 1

Table of Contents

[Part A - Business Decision-Making using Supervised Machine Learning Methods (Classification) 2](#_Toc169037213)

[Task 1 Data Preparation 2](#_Toc169037214)

[Load a selected dataset. Display loading and display the first and last few rows of the dataset. 2](#_Toc169037215)

[Apply two techniques to handle missing values 3](#_Toc169037216)

[Use one effective method to manage different data types 5](#_Toc169037217)

[Apply three specific data visualisation techniques to analyse data 6](#_Toc169037218)

[Task 2 Feature Engineering 9](#_Toc169037219)

[Implement the two most relevant feature selection techniques 9](#_Toc169037220)

[Implement one feature scaling method 11](#_Toc169037221)

[Task 3 Model Building and Prediction 12](#_Toc169037222)

[Explain the steps in constructing and training the model 12](#_Toc169037223)

[Summarize the three key aspects of building the supervised machine learning model. 13](#_Toc169037224)

[Part B - Time Series Trend Analysis and Forecasting 14](#_Toc169037225)

[Task 1 Data Exploration 14](#_Toc169037226)

[Explore the dataset thoroughly 14](#_Toc169037227)

[Task 2 Trend Analysis 17](#_Toc169037228)

[Apply appropriate time series analysis techniques to identify and explain the overall trends 17](#_Toc169037229)

[Task 3 Seasonality Assessment 19](#_Toc169037230)

[Conduct a detailed analysis to identify and describe any seasonality patterns in the data 19](#_Toc169037231)

[Task 4 Anomaly Detection 21](#_Toc169037232)

[Identify and explain any significant anomalies or outliers present in the data 21](#_Toc169037233)

[Task 5 Predictions and Recommendation 22](#_Toc169037234)

[Utilise time series analysis to make informed predictions with one recommendation 22](#_Toc169037235)

[Part C - Business Decision-Making using Un-Supervised Machine Learning Methods (Clustering) 26](#_Toc169037236)

[Task 1 Data Preparation 26](#_Toc169037237)

[Data Cleaning and Pre-processing: Addressing Missing Values and Outliers 26](#_Toc169037238)

[Exploring Data with Techniques and Insights 31](#_Toc169037239)

[Task 2 Unsupervised Algorithm Implementation 34](#_Toc169037240)

[Select two unsupervised segmentation algorithms, justify, explain, and demonstrate 34](#_Toc169037241)

[Apply Principal Component Analysis (PCA) to reduce the dimensionality with justification 39](#_Toc169037242)

[Provide two examples of visualisation techniques 43](#_Toc169037243)

[Task 3 Conclusion 47](#_Toc169037244)

[Summarise the two key aspects of the segmentation analysis 47](#_Toc169037245)

# Part A - Business Decision-Making using Supervised Machine Learning Methods (Classification)

The bank customer churn dataset was downloaded from Kaggle. The process started with data exploration, followed by the implementation of proper algorithms, and closed with evaluating the features and machine learning models’ capability of identifying customer churn. The objective was to identify and predict the customers at risk of leaving the bank, facilitating a well-crafted retention strategy to reduce churn rates.

## Task 1 Data Preparation

### Load a selected dataset. Display loading and display the first and last few rows of the dataset.

I imported the necessary libraries. Using Pandas, I read the dataset into a DataFrame and then explored it by displaying both the first and last few rows.

The code used and the results output are as follows:

A screenshot of a computer screen

Description automatically generated

### Apply two techniques to handle missing values

Two strategies were applied to manage the missing values in this task. One involved dropping the null values, and the other involved imputing the null values with using the mean method.

1. Drop the Null values

Firstly I checked how many data cells are ‘null’. In this case, if the ‘null’ value was more than 80% in one row, the row was considered to lack substantial predictive value. I decided to drop the rows where 80% of the cells’ value is null.

The codes were used below:

A screenshot of a computer

Description automatically generated

I found 4 rows got dropped.

A screenshot of a computer code

Description automatically generated

1. Imputation with the ‘mean’ value

Then I checked the dataset again, I found ‘Estimated Salary’ and ‘Point Earned’ columns still have ‘null’ values. I decided to handle these missing values by using the ‘mean’ value to impute.

Codes shown below:

A screenshot of a computer

Description automatically generated

I performed the imputation function. We can see after the data cleaning, there is no ‘null’ value in the dataset.

A screenshot of a computer

Description automatically generated

### Use one effective method to manage different data types

Check the data type.

A screenshot of a computer

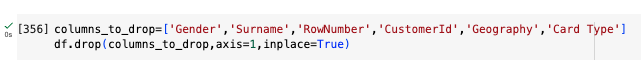
Description automatically generated

There are four columns in the ‘object’ data type. In this case, I wanted to change the ‘Gender’ and ‘Card Type’ data type from object to integer so that later I could use these variables to be compatible with other numeric ones. I replaced these variables with the integer using the ‘one hot encoding’ method.

A screenshot of a computer screen

Description automatically generated

I dropped some useless and duplicated columns.



### Apply three specific data visualisation techniques to analyse data

1. Check the skewness of ‘CreditScore’, ‘Age’, and ‘Balance’.

The skewness of the Credit Score is close to 0, which indicates that the distribution is approximately symmetric.

The skewness of age is 1.02, positive meaning the distribution is right-skewed, with a longer right tail.

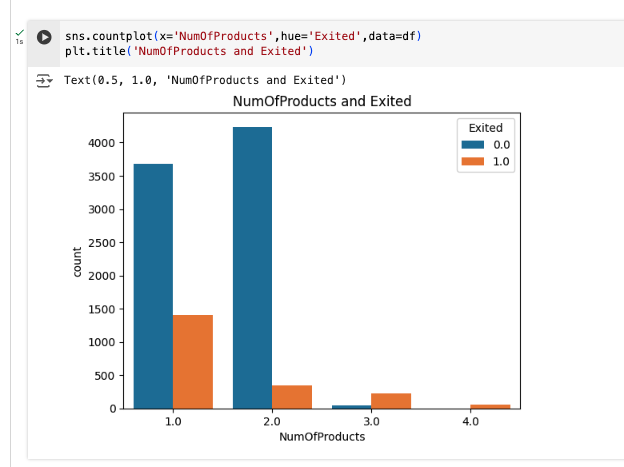
The skewness of balance is -0.14, negative indicating the distribution is left-skewed. However, since the number is close to 0, we can infer that aside from the balance of 0, the distribution is approximately symmetric.

A screenshot of a computer screen

Description automatically generated

1. Check the relationship between the ‘NumOfProducts’ and ‘Exited’.

I found most customers are holding 1-2 products. In addition, there is a possibility that the more products the customer has, the higher the chance he will churn. We can see that customers have 3 products, and the number of churn is high.



1. Check and drop the outliers

I used the z-score method to locate the outliers. And I dropped 205 rows.

A screenshot of a computer program

Description automatically generated

1. Check the correlation

Age exhibits a correlation with both Exited and Complain.

Exited and Complain show a correlation of 100%.

Balance and Number of products also play a significant correlation.

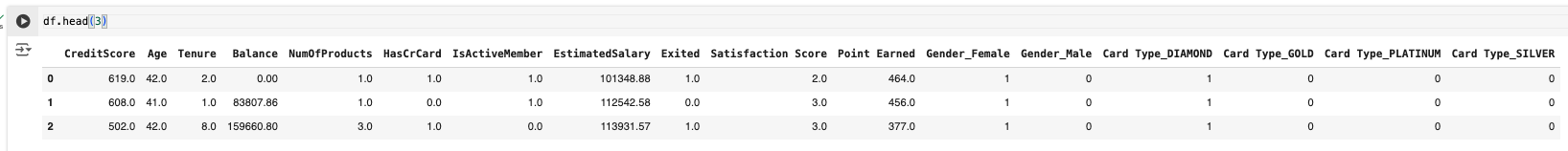
A screenshot of a computer

Description automatically generated

1. Drop column ‘Complain’ due to it being 100% correlated with ‘Exited’, making it unnecessary for the subsequent regression model due to the high correlation score.

A close-up of a sign

Description automatically generated



## Task 2 Feature Engineering

### Implement the two most relevant feature selection techniques

I used two techniques to select features for the prediction model. One is Recursive Feature Elimination (RFE) and the other is SelectKBest.

I imported the related libraries and separated the variables X and Y.

A screenshot of a computer program

Description automatically generated

1. SelectKBest

By using the SelectKBeat function find and print the 5 features

A screenshot of a computer program

Description automatically generated

Get the accuracy score if using these features to run a logistic regression model

A screenshot of a computer program

Description automatically generated

1. Recursive Feature Elimination (RFE):

Initialize the RFE function with the logistic regression model, find out the 5 features and print out the accuracy score

A screenshot of a computer program

Description automatically generated

### Implement one feature scaling method

I employed the features selected via the SelectKBest technique, which yielded a higher accuracy score. These features include ‘Age’, ‘Balance’, ‘IsActiveMember’, ‘Gender\_Female’, and ‘NumOfProducts’.

I utilized the Min-Max Scaling method to scale the features to a fixed range. This scaling method will affect the data distribution because after scaling, all the features will have similar scales and distributions, which will help the model to learn without being disturbed by some individual features. Moreover, by scaling the features, I could make sure that each feature contributes proportionally to the machine-learning process, thereby reducing the biases.

A screenshot of a computer

Description automatically generated

## Task 3 Model Building and Prediction

### Explain the steps in constructing and training the model

Logistic regression is popular in predicting binary outcomes, which means it is very suitable for analysing customer churn (whether they exited or were retained). Based on the last step, identifying potential predictor variables by implementing feature selection and scaling, I found that the SelecKBest method returned a higher score than the RFE method on logistic regression. Thus, I utilized the features selected by the SelectKBest function: ‘Age’, ‘Balance’, ‘IsActiveMember’, ‘Gender\_Female’, and ‘NumOfProducts’.

I fitted a logistic model to the data, with binary churn status as the dependent variable and the selected predictor variables as the independent variables. The model estimated the probability of churn based on the values of the predictor variables. I split the dataset into training and testing, and set the test size is 30%.

The performance of the model is evaluated below:

Accuracy: overall correctness of the model’s prediction, indicating 0.82.

Precision: the proportion of true positive predictions among all positive predictions (true and false), indicating 0.83 for retained and 0.61 for churn.

Recall (sensitivity): measure the model’s ability to capture positive instances, indicating 0.97 for retained and 0.2 for churn.

F1-score: a balanced measure is the mean of precision and recall. Especially when there is an imbalance, indicating 0.9 for retained and 0.3 for churn.

A screenshot of a computer program

Description automatically generated

Overall, the logistic model demonstrates strong performance. Although the score for churn customers is slightly low due to some data shortages, the score for retained customers is notably high, indicating that it successfully predicts the majority of instances. The bank can utilize the mode to predict the customers who wish to stay with the bank. Therefore, it can be inferred that customer churn in this bank is strongly related to age, gender, account balance, active member status, and the number of products they hold. Armed with this information, the bank can effectively predict the likelihood of customer churn

### Summarise the three key aspects of building the supervised machine learning model

First, I believe when performing data preparation, especially handling the missing value, it is vital to understand the dataset’s overall status before making the decision, such as dropping the null value or imputation with the mean, median, or mode. There is always a reason behind the action without bias so that the action would not change the dataset direction.

Secondly, I think selecting the right model with proper evaluation is important. Evaluating the model can support an insight into the effectiveness, meaningfulness, and accuracy of the model. For example, after the logistic model is generated, I always make sure to check the accuracy and other related score before I put it into practice.

The third aspect involves practical application with various methods. For instance, when conducting feature selection, I employed both SelectKBest and RFE methods, which yielded slightly different scores. Consequently, I chose the features with higher scores. Utilizing these selected features enables the formulation of a valuable strategic plan to address customer churn effectively.

# Part B - Time Series Trend Analysis and Forecasting

The dataset used for this task comprises Toyota car sales data obtained from Kaggle, spanning from January 2007 to December 2016. A time series trend analysis was conducted to uncover the overall trend, by leveraging these insights, to generate actionable recommendations for forecasting purposes.

## Task 1 Data Exploration

### Explore the dataset thoroughly

* Install pmdarima.

A screenshot of a computer

Description automatically generated

* Import relevant libraries.

Explore the data by showing its first and last few rows.

A screenshot of a computer

Description automatically generated

* Check the null value and data shape

A screenshot of a computer code

Description automatically generated

The data frame has 120 rows and 4 columns

A close-up of a computer code

Description automatically generated

* Conduct ‘DateTime’ format

Combine the ‘year’ and ‘month’ columns, and generate a new column named ‘year\_month’.

Convert the 'year\_month' column to ‘datetime’ format.

Drop the original ‘year’ and ‘month’ columns.

Sort the DataFrame by the "year\_month" column to ensure the timestamps are ascending.

Calculate the time intervals.

A screenshot of a computer

Description automatically generated

* Stationarity check

Apply the dickey-fuller test and print the result: The data is likely stationary.

A screenshot of a computer program

Description automatically generated

## Task 2 Trend Analysis

### Apply appropriate time series analysis techniques to identify and explain the overall trends

* Visualize the dataset column ‘Quantity’ over time in the line graph.

A graph with blue lines

Description automatically generated

* Calculate the moving average, plot the original data with the moving average

A screen shot of a graph

Description automatically generated

* Create an EMA for the original data with a 12-month window. Create a plot comparing original data, SMA, and EMA.

The original time series, SMA, and EMA exhibited a similar overall trend. SMA and EMA fluctuated around the original one without an above or below pattern, indicating a relatively stable trend.

The original time series data indicated the quantity of Toyota sold over time.

The SMA smoothed out the short-term fluctuations, and the EMA also smoothed out the overall fluctuation but it reacted quickly to short-term changes, giving more weight to the observations and showing more responsibility for the changes.

A screenshot of a computer

Description automatically generated

## Task 3 Seasonality Assessment

### Conduct a detailed analysis to identify and describe any seasonality patterns in the data

Perform seasonal decomposition using the seasonal decomposition method with moving averages (STL-MA).

A screenshot of a computer program

Description automatically generated

The trend represents the long-term move of the data. With a decrease from 2008 to 2009 and an increase from 2009 to 2010, after 2010, it became stable, showing a gradual upward movement.

The seasonality captures the periodic patterns from the dataset. There is a frequency every January when the quantity of sales drops significantly, indicating during the new year period, customers are unlikely to purchase a new car.

The residual reflects the random noises or irregularities after removing the trend and seasonality components. There is a large residual, suggesting some unexplained variability or unpredictability.

A graph of blue lines

Description automatically generated with medium confidence

## Task 4 Anomaly Detection

### Identify and explain any significant anomalies or outliers present in the data

I visualized and plotted the time series data by using the LOESS (STL) method.

In the residual graph, it indicated some fluctuations which are potentially outliers.

A screenshot of a graph

Description automatically generated

Highlighted the yellow points are potential anomalies. Anomalies may be extreme values that deviate significantly from other observations.

A chart with blue and yellow dots

Description automatically generated

## Task 5 Predictions and Recommendation

### Utilise time series analysis to make informed predictions with one recommendation

Perform the ACF and PACF.

ACF plot indicates the data is stationary and seasonable and the cut-off after a certain lag suggests the order of the AR model. PACF plot can determine the order of the MA model.

A screen shot of a graph

Description automatically generated

A graph with blue dots and lines

Description automatically generated

Split the data into training and testing.

A close-up of a computer code

Description automatically generated

Fit the SARIMA model and print the result.

A screenshot of a computer

Description automatically generated

p-value less than 0.05 is considered statistically significant.

Ljung-Box (Q): 0.42, Heteroskedasticity (H): 0.38, the low values for Ljung-Box and Heteroskedasticity suggest minimal autocorrelation, meaning the residuals are not independent, indicating a lack of fit.

Jarque-Bera (JB): 44.47, the high value of Jarque-Bera means a potential departure from normality.

The positive Skew equalling 0.51 and the high Kurtosis equalling 5.97, suggest a rightward skew with a longer tail on the right side and a sharper peak compared to a normal distribution. This implies the data have more extreme values than would be expected under normal distribution.

When applying the model to real practice, it’s vital to take these factors into consideration.

Visualize the model by showing the plot graph.

A screenshot of a graph

Description automatically generated

Based on the above graph, the model managed to detect the overall trend.

Exogenous variables are additional factors that are external to the model but can influence the behaviour of the target variable. So I added the exogenous variables and ran the model again.

A screenshot of a computer program

Description automatically generated

A black text with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

Visualize the model in the plot graph

A screen shot of a graph

Description automatically generated

After incorporating the exogenous variable into the model, there appears to be no discernible change in the graph. The exogenous variables exhibit no significant impact on the time series forecasting, which means the exogenous does not provide any additional explanatory power to the model.

# Part C - Business Decision-Making using Un-Supervised Machine Learning Methods (Clustering)

The wine consumption dataset was downloaded from Kaggle. In this task, I applied unsupervised machine learning techniques to segment customers based on their characteristics and preferences. I implemented algorithms to cluster data samples and allocated dimension-reduction methods for feature selection to enhance accuracy. In the end, I summarized the key aspects of the segmentation analysis.

## Task 1 Data Preparation

### Data Cleaning and Pre-processing: Addressing Missing Values and Outliers

1. Import related libraries and display a few rows

A screenshot of a computer

Description automatically generated

1. Check the missing values. In the ‘Income’ column, there are 24 null values.

A screenshot of a computer screen

Description automatically generated

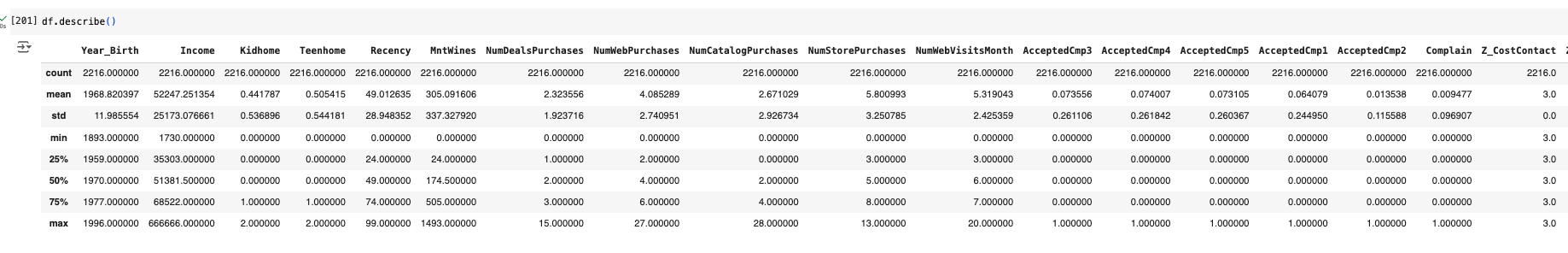
1. Use the mean of Income to fill null value in the Income
2. A screenshot of a computer

   Description automatically generated
3. Drop the columns that will not be used in this task. Since I will focus on wine consumption in this case, I dropped the columns containing other product information and Checked the new dataset shape, 2240 rows and 22 columns.

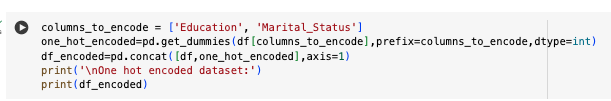
A screenshot of a computer

Description automatically generated

1. Check the descriptive statistics of the dataset.



1. Perform one hot encode to convert the categorical variables ‘Education’ and ‘Marital\_Status’ to numerical format.



A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

1. Use the Z-score to detect the outliers and drop the outliers.

Set up the threshold to 3. Drop the categorical columns ‘Education’ and ‘Marital\_Status’. Then conduct the z-score method to find and drop the outliers. 785 rows got dropped.

A white rectangular object with black text

Description automatically generated



A screenshot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

### Exploring Data with Techniques and Insights

1. Subplot to detect the distribution of the variables

The skewness in ‘MntWines’ is 1.31 indicating that the distribution is right skewed, meaning it has a longer right tail, and the Kurtosis is 1.11 indicating a sharper peak compared to a normal distribution. In addition, ‘Year\_Birth’ and ‘Income’ are flatter with lighter tails compared to a normal distribution.

The analysis concludes that the range of wine consumption is extensive, with s significant portion of individuals not purchasing wine. While the age distribution appears to be normal, there is evidence suggesting that the younger generations buy less wine compared to older age groups. Furthermore, wine consumption seems to be popular among individuals with income from 25,000 to 75,000.

A screenshot of a computer screen

Description automatically generated

1. Bar plot and box plot visualization

The number of graduation levels occupies the highest proportion, with the amount of PhD and master’s staying at a similar quantity. The basic level is the smallest group.

A screenshot of a graph

Description automatically generated

In most education classes, 25% of individuals consume low amounts of wine. The predominant consumption range for wine falls between 0-500, which accounts for 50% of the total consumption. Across different education levels, some individuals consume wine in the range of 1000 to 1400, which is considerably high. Due to insufficient data from the Basic education level, the boxplot in that category is not well represented.

A screen shot of a graph

Description automatically generated

1. Scatter plot to find out the relationship between variables

Across all age groups, wine consumption exhibits a high density between 0 and 400, tapering off gradually from 400 to 800. However, there is a noticeable decrease in wine consumption among the younger generation. Individuals born after 1990 appear to consume wine less frequently compared to older people.

A screen shot of a graph

Description automatically generated

The graph indicates that there is not a strong correlation between wine consumption and recency. Individuals who purchase wine frequently don’t necessarily buy large quantities, while those who have not made purchases for a while may buy substantial amounts. This discrepancy could be attributed to various factors such as brand preferences, length of residence in a particular area, or other influencing factors.

A screen shot of a chart

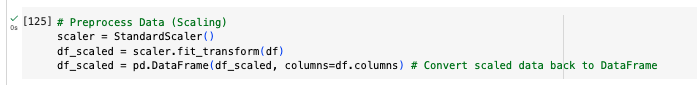
Description automatically generated

## Task 2 Unsupervised Algorithm Implementation

### Select two unsupervised segmentation algorithms, justify, explain, and demonstrate

I would apply the K Means clustering model and the agglomerative clustering model to the wine consumption dataset.

Standardize the data.



1. Perform the Agglomerative clustering model. The dendrogram represents two clusters.

A screen shot of a computer program

Description automatically generated

A graph of a city

Description automatically generated with medium confidence

Visualize the segmentation. There is a lot of overlap in the 2D graph. I will perform the PCA in the later task, the overlapping issue will be solved.

A screen shot of a computer screen

Description automatically generated

1. Perform the K Means clustering model.

Implement the Elbow Method. There is an elbow point at the number 2.

A screen shot of a graph

Description automatically generated

Generate the K mean Silhouette plot and print the Silhouette score. The score is quite low, which indicates that there is a lot of overlapping. The Silhouette score for 2 clusters indicates the highest score.

A screenshot of a computer program

Description automatically generated

A screenshot of a graph

Description automatically generated

A graph of different colors

Description automatically generated with medium confidence

A graph of different colors

Description automatically generated

Visualize the segmentation with 2 clusters, as the Silhouette score for this configuration is the highest. While the segmentation is discernible, there remains some overlap in the middle, albeit less than in the Agglomerative clustering model.

A screenshot of a computer screen

Description automatically generated

### Apply Principal Component Analysis (PCA) to reduce the dimensionality with justification

Examine the range of components in the PCA and visualize the explained variance ratio across different component numbers. An elbow point is noted after 6 components, although other components also contribute to variance, their contributions are not significant. Hence, 6 components are selected.

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

Perform PCA with the 6 components.

A computer code with green and black text

Description automatically generated

Perform the Elbow method and there is an elbow point on 4 clusters.

A screen shot of a graph

Description automatically generated

Apply K means clustering method on the dimension-reduced dataset and choose the cluster number of 4. At last, visualize the segmentation in a scatter plot. This time the zone between each cluster is getting slightly clearer.

A screen shot of a computer screen

Description automatically generated

In addition, I performed the PCA with 4 components on the agglomerative clustering model. I found the overlapping issue is getting better.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

### Provide two examples of visualisation techniques

The Cluster scatter plot and the Silhouette plot would be suitable for visualizing the segmented data.

1. Silhouette plot

I managed to create 2 variables X1 and X2, representing the two different columns in the scaled dataset.

A close-up of a white background

Description automatically generated

I did an Elbow method on the X1 data and found there is an elbow point on 3 clusters.

A screen shot of a graph

Description automatically generated

I performed the Silhouette plot on the X1 dataset and found the number of 3 clusters yielded the highest score, consistent with the result obtained from the elbow method.

­A screenshot of a computer program

Description automatically generated

A graph of different colors

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

1. Cluster scatter plot

I used the variable X2 to perform this task.

I conducted the Elbow method and I found there is an elbow point on 3 clusters.

A screen shot of a graph

Description automatically generated

I conducted the K Means cluster and visualized the result in the cluster scatter plot.

A screen shot of a computer screen

Description automatically generated

## Task 3 Conclusion

### Summarise the two key aspects of the segmentation analysis

The purpose of segmentation analysis is to group untitled data based on shared characteristics or patterns. In the tasks above of unsupervised machine learning, I found that determining the method to employ for segmentation and understating the features to reduce dimensions are two key aspects.

Firstly, it involves determining the method to employ for segmentation, such as Hierarchical clustering or K-means clustering. Each method possesses its own set of advantages and weaknesses. K-means is straightforward to use, but at times, it can be challenging to determine the optimal number of clusters. Moreover, its result may change upon repetition of the running. On the other hand, Hierarchical clustering offers a visual representation, slowing for an intuitive understanding of the number and distribution of clusters defined by the model. In practice, initially, I found the agglomerative model challenging to grasp, particularly when dealing with variables with large ranges causing numerous branches at the bottom. In such cases, early preparation, particularly in removing outliers and scaling, becomes crucial as it can effectively prevent issues and yield more efficient results.

In addition, some datasets might be excessively large, leading to numerous overlaps and complexities in segmenting results. Utilizing dimensionality reduction techniques can mitigate data redundancy, enhance the quality, and make the model outcome more efficient. In the previous task, I initially constructed a segmentation model for the dataset and encountered significant entanglement, making segmentation challenging. Nonetheless, upon applying the PCA method to select the components that contain higher variances, and put the selected features into the model again, the model became clearer and more conducive.