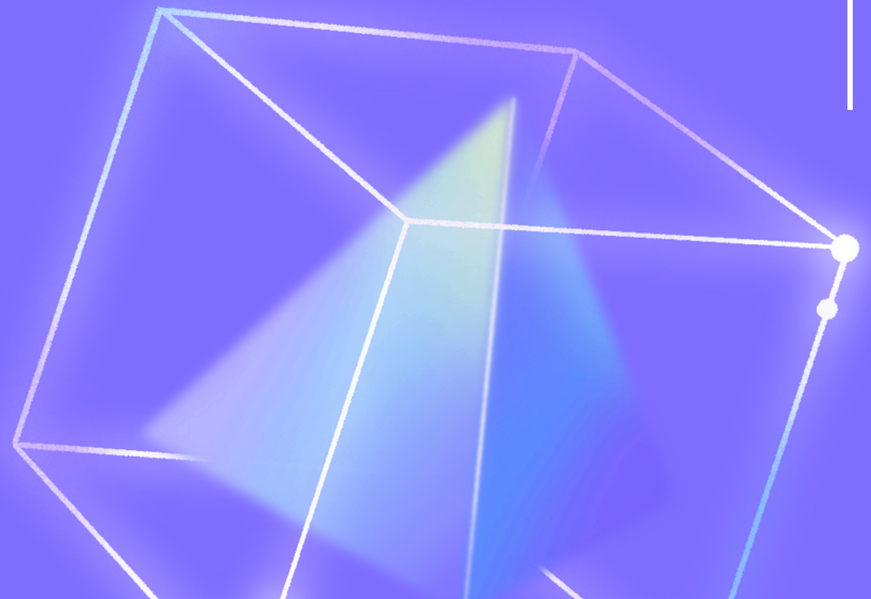
Customer Segmentation using K -Means Clusterning Machine Learning Technique

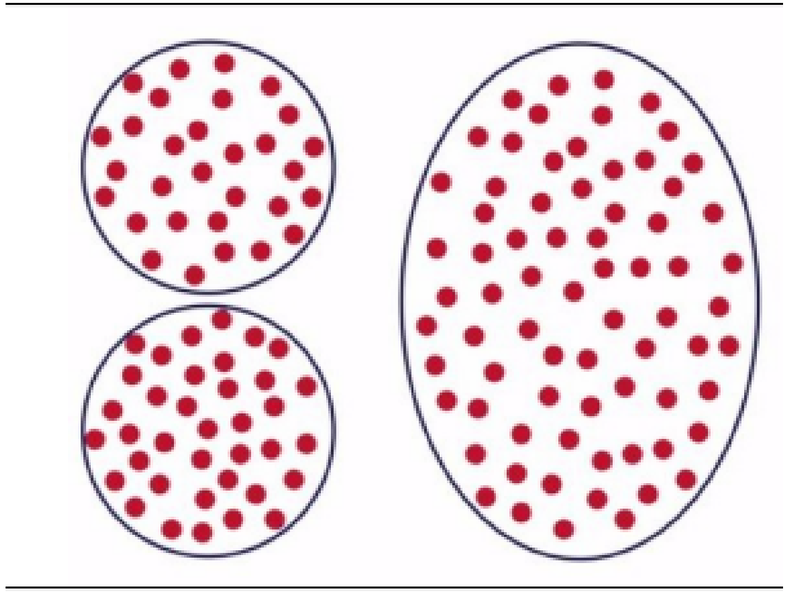
PROJECT OUTLINE

- Introduction
- Problem Statement
- Methodology
- Data Preprocessing
- Data Modelling
- Conclusion

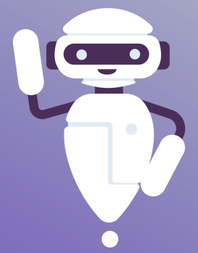


INTRODUCTION

If training machine learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. The most important unsupervised learning technique is clustering, which creates different groups or clusters of the given set of inputs and is also able to put any new input in the appropriate cluster.



METHEDOLOGY

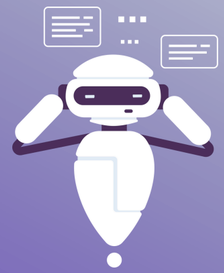


Preprocess Data

Data cleaning
Normalization/Standardization,F
Feature Selection

Model Development

Model selection based on Silhouette Score Method



Model Evaluation

Model evaluation based on Silhouette Score and elbo curve method

DATA SET

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

PREPROCESSING

- R (Recency): Number of days since last purchase
- F (Frequency): Number of tracsactions
- M (Monetary): Total amount of transactions (revenue contributed)

FINDING OPTIMAL NUMBERS OF CLUSTERS USING SSD

Silhouette Analysis

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

p is the mean distance to the points in the nearest cluster that the data point is not a part of

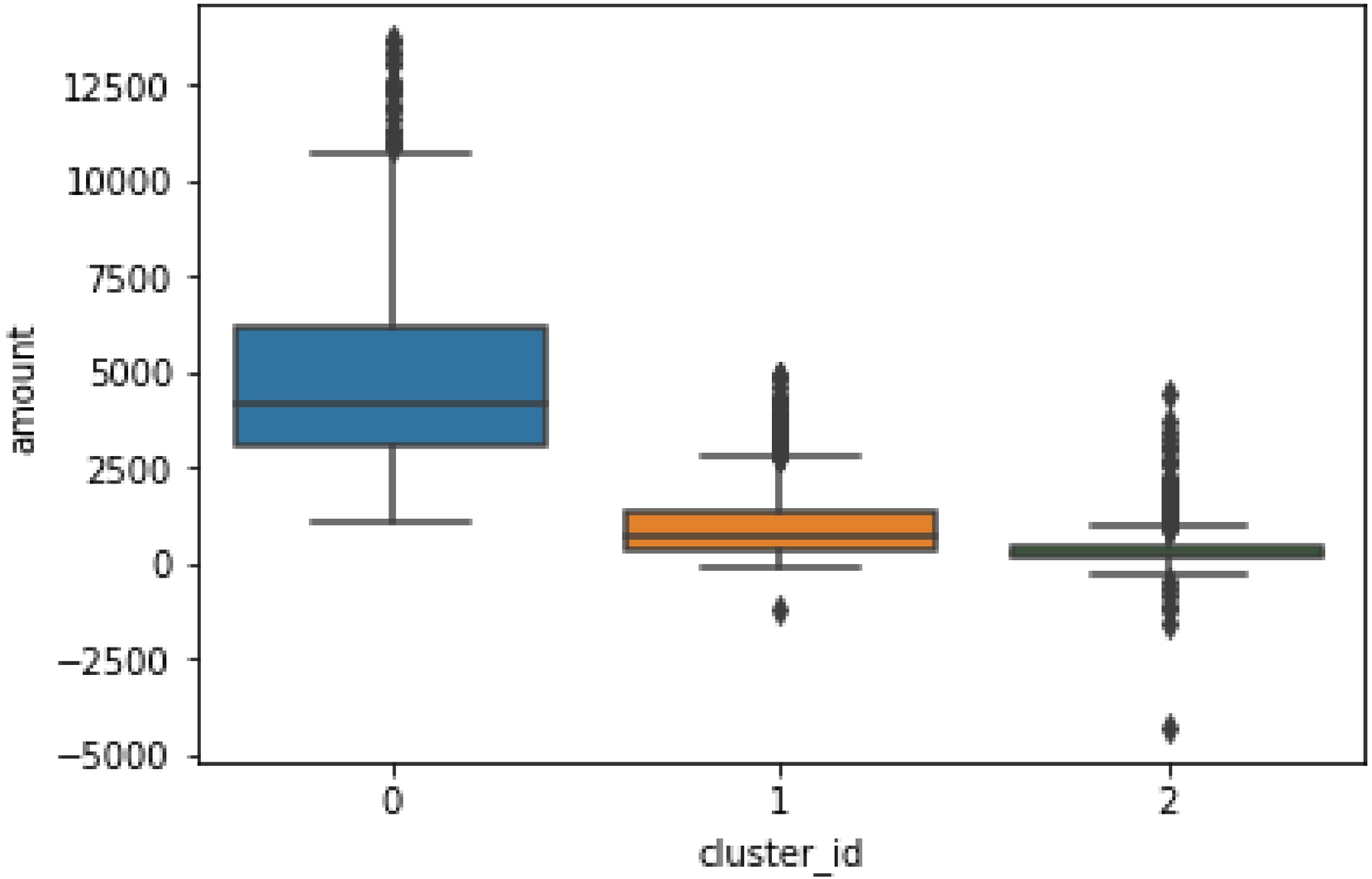
q is the mean intra-cluster distance to all the points in its own cluster.

- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

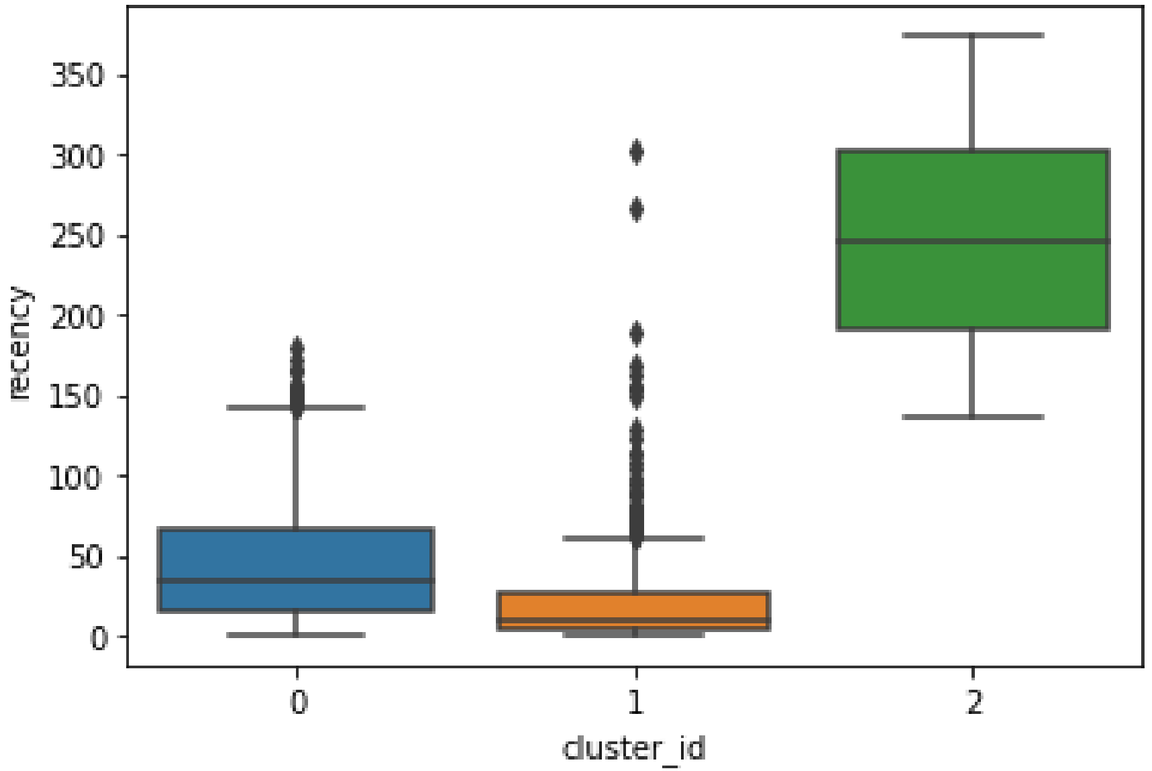
```
For n_clusters=2, the silhouette score is 0.5415858652525395
For n_clusters=3, the silhouette score is 0.5084896296141937
For n_clusters=4, the silhouette score is 0.4814786837400834
For n_clusters=5, the silhouette score is 0.4658529685822305
For n_clusters=6, the silhouette score is 0.41707960376211345
For n_clusters=7, the silhouette score is 0.4158077420309644
For n_clusters=8, the silhouette score is 0.4059904161107271
```

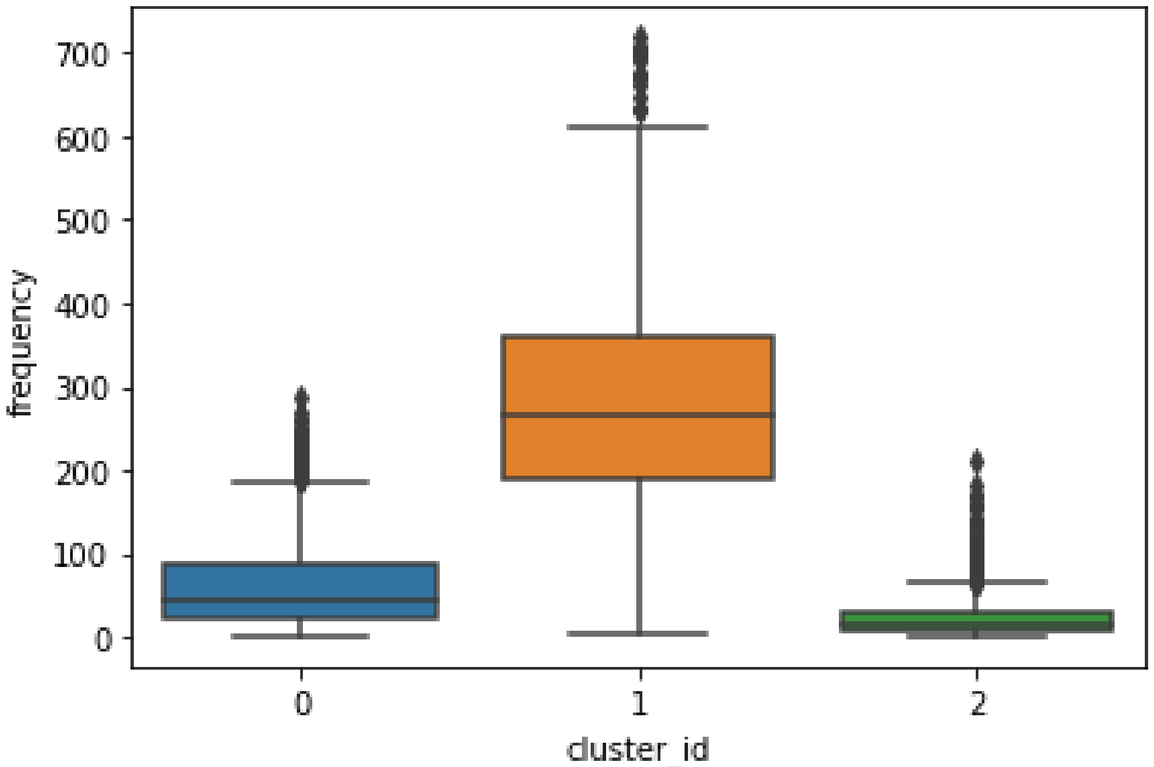
OUTLIER DETECTION

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a485086d8>
```



OUTLIER DETECTION





PRACTICAL CONSIDERATIONS IN K-MEANS ALGORITHM

1.) THE CHOICE OF INITIAL CLUSTER CENTRE HAS AN IMPACT ON THE FINAL CLUSTER COMPOSITION

2.) CHOOSING THE NUMBER OF CLUSTERS K IN ADVANCE

3.) IMPACT OF OUTLIERS

4.) STANDARDISATION OF DATA

5.) NON-APPLICABILITY WITH THE CATEGORICAL DATA

THANK YOU