Assignment 2 Mini-project: Customer segmentation with clustering

Aadam Gafar, 22nd June 2025

[**Introduction 2**](#_515wz7bn3yet)

[**Methodology 2**](#_9nb10l9v2u42)

[**Analysis 2**](#_blftdmedsvfo)

[**Conclusion 4**](#_ua05fqgjl1t0)

[**References 5**](#_t8dl8ppas063)

# 

# 

# 

# 

# Introduction

It is in the best interests for a business to segment its consumers into distinct groups so they can be recipients of tailored business programmes and marketing schemes. This allows for businesses to improve market efficiency, enhance product development, increase customer satisfaction, improve retention, optimise pricing, and strategically allocate resources. In this project, we use K-Means Clustering to segment customers based on purchasing behaviour, namely: Frequency, Recency, Customer Lifetime Value (CLV), Age, and Average Unit Cost. Using data visualisations, specifically box plots, segments were assigned meaning based on the behaviour they described: "Frequent Buyers", "Older Buyers", "Low Frequency / High Spenders", and "Younger Buyers".

# Methodology

The customer data consisted of 951,669 records for analysis. Prior to segmentation, they must be cleaned and re-engineered, prioritising the most relevant behavioural features to cluster for. The chosen features were:

* Frequency: the number of records per Customer ID
* Recency: the time between today's date and a transaction's date
* Customer Lifetime Value (CLV): the total revenue per Customer ID
* Age: the time between today's date and date of birth per Customer ID
* Average Unit Cost: total revenue per quantity per Customer ID

We then aggregated data such that each record represented a singular customer's mean data per feature, as opposed to each record representing a singular transaction. Outliers were identified (but not removed) in the interest of better understanding bias in the dataset.

Then, we used an algorithm known as K-Means Clustering to group similar customers together. Its parameters were optimised using three methods: Elbow plot, Silhouette Score, and Agglomerative Hierarchical Clustering. Box plots were used to visualise the results and assign meaningful behaviours to the clusters. Finally, two dimensionality reduction techniques were used to allow cluster visualisation: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). This helped confirm the separation and clarity of the customer segments.

# Analysis

The original dataset had many features set aside in favour of six newly engineered features: Frequency, Recency, Customer Lifetime Value (CLV), Age, and Average Unit Cost. We used box plots to visualise (but not remove) outliers. This was done in the interest of better understanding any bias in the dataset. In particular, Frequency, CLV, and Average Unit Cost were observed to harbour many outlier values (see Figure 1).

| **Figure 1** |
| --- |
|  |

The data was then aggregated so as to encapsulate the mean data per Customer ID per row, as opposed to the original dataset encapsulating a single transaction per row. With this new, more relevant dataset, we were able to identify the proportion of outliers using the Inter-Quartile Range (IQR) method, producing an outlier rate of 2.49% at a threshold rate of 2 (i.e. a record with two or more outlier features is considered an outlier itself).

From here, we had the necessary dataset to commence clustering. However, in order to provide accurate results, it is important to ascertain the optimal number of clusters (k-value) to discover. This is done via several metrics: the Elbow plot, the Silhouette Score, and Agglomerative Hierarchical Clustering. These were performed, the ideal k-value chosen, and clustering commenced.

Once our dataset was labelled by cluster, box plots were visualised in order to identify any patterns in the data which would help us assign meaning to each cluster.

| **Figure 2: Box plots of each feature, segmenting by clusters** |
| --- |
|  |

To visualise our data, we used dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce our multi-dimensional dataset into two dimensions. The t-SNE visualisation, in particular, allowed us to see our labelled customer clusters with clarity and clear separation. Importantly, due to the computational cost of t-SNE reduction, it was run on a sample of 30,000 records, as opposed to the entire aggregated dataset.

| **Figure 3: PCA visualisation (left) vs. t-SNE visualisation (right)** |
| --- |
|  |

# Conclusion

After comparing the plots in **Figure 3**, we see far superior visual separation among the clusters produced by t-SNE reduction. Although it is run on a sample of the wider dataset (30,000 records), it shows delineation that outclasses PCA in regards to its ability to detect subtle patterns in multi-dimensional data.

With it, we were able to identify meaningful clusters based on our engineered features (Frequency, Recency, CLV, Age, and Average Unit Cost) and K-Means Clustering algorithm. The identified segments can now be used to inform business decisions to improve market efficiency, enhance product development, increase customer satisfaction, improve retention, optimise pricing, and strategically allocate resources.

# References

Not applicable.