CUSTOMER SEGMENTATION WITH CLUSTERING

Stakeholder Report

Prepared by Matt Cocker

Contents

Introduction

The business needs to understand its customers to improve marketing efficiency and therefore profits. Customer segmentation using clustering will allow the business to identify groups of customers with similar characteristics. That knowledge can then be leveraged to enable the business to conduct targeted marketing campaigns, ensuring that marketing efforts are being concentrated where they will be most effective.

Figure 1: *Project Flow Chart*

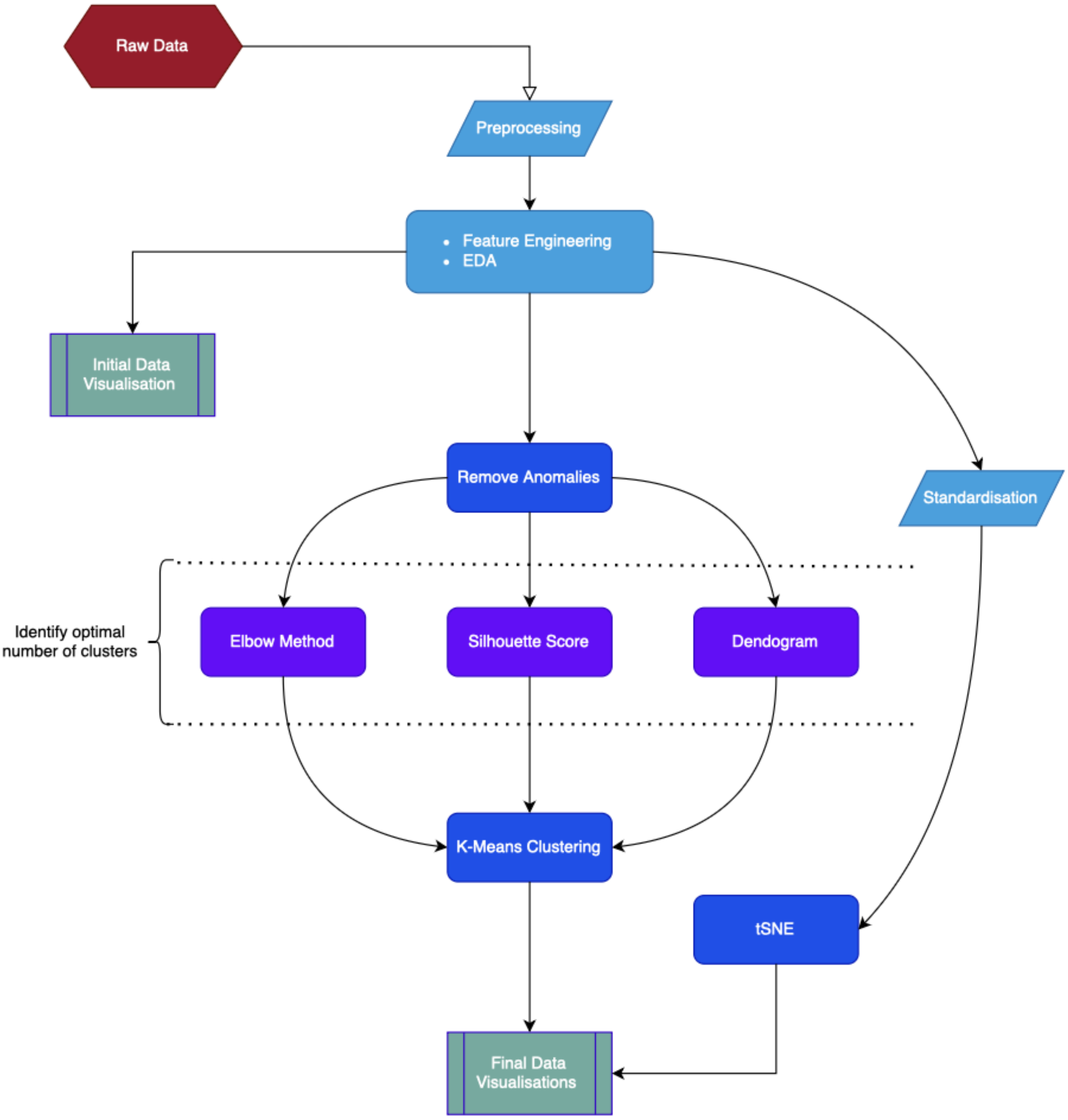


Figure 1 describes the process of performing customer segmentation. Data preprocessing ensured missing or duplicate values were dealt with. Feature engineering created a streamlined, relevant dataset for customer segmentation. The optimal number of clusters was determined, and clustering was performed. The results were visualised in box plots and a scatter plot (via dimensionality reduction), and these charts were analysed to draw conclusions.

Data

The initial dataset contained 951,668 samples, each showing data for 20 features. After missing and duplicate values were dealt with and feature engineering was performed, the dataset contained 68,300 samples across 5 features. Each sample represents a unique customer.

During the feature engineering stage, 5 new features were created:

- **Frequency:** How often a customer makes purchases (in days).
- **Recency:** How recent the customer's last purchase was (in days).
- **Customer Lifetime Value (CLV):** Total net profit the business makes from the customer.
- **Average Unit Cost:** The mean cost of all items purchased by the customer.
- **Age:** The customer's age.

These features are visualised on boxplots below:

Figure 2: *Feature boxplots*

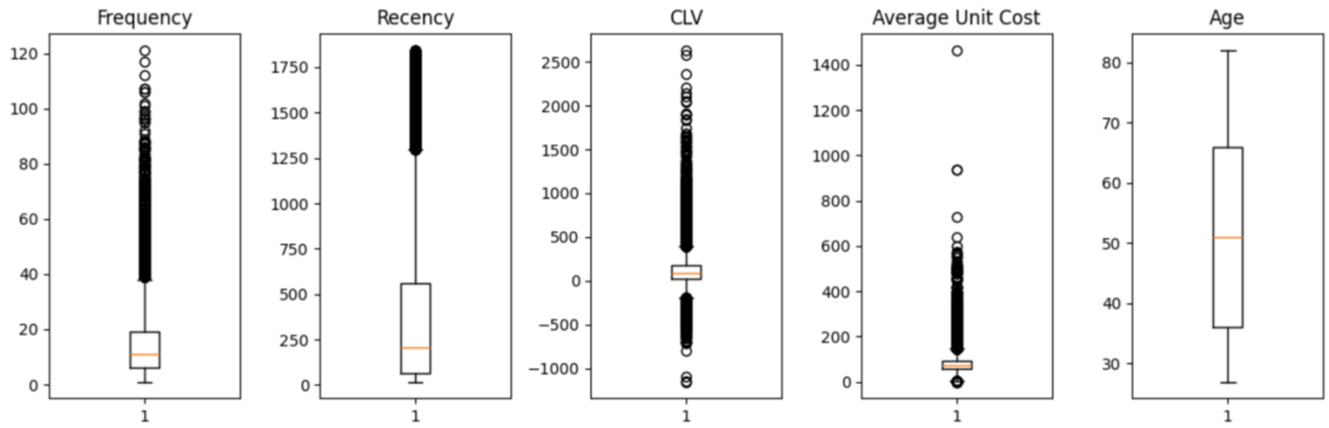


Figure 2 shows that the CLV, Average Unit Cost and Frequency features all have a large number of outliers.

Methods

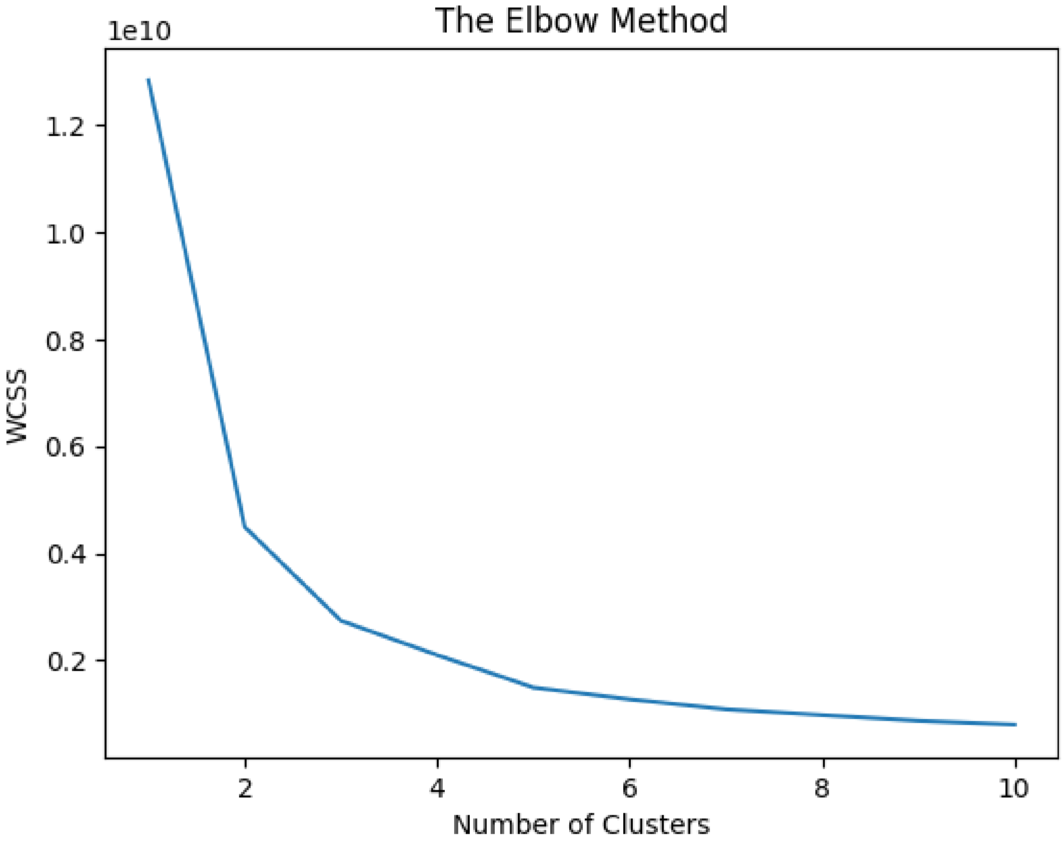
Clustering algorithms are sensitive to outliers, therefore, to ensure effective clustering, these must be removed from the dataset. Isolation Forest was used to identify outliers for its efficiency in handling large datasets.

Three methods were used to identify the optimal number of clusters (*k*) to perform k-means clustering. K-means clustering was chosen instead of hierarchical clustering as it deals with large datasets more efficiently.

Elbow Method

The elbow method runs k-means clustering for multiple *k*'s. The resulting within-cluster variation (WCSS) is plotted against the number of clusters, where the optimal *k* is where WCSS is minimised subject to *k* (here representing complexity). *Figure 3* shows the optimal *k* is 5.

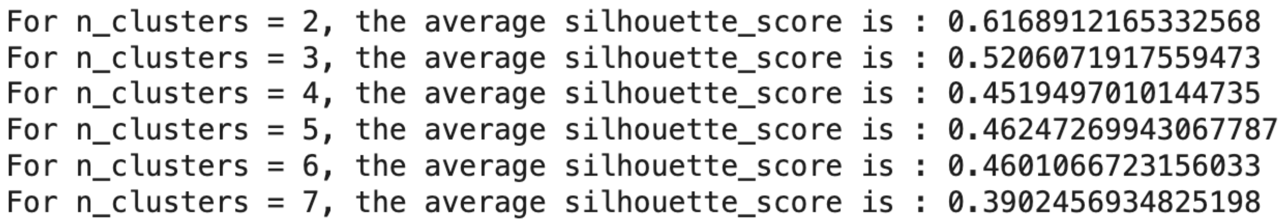
Figure 3: *WCSS vs Number of Clusters*



Silhouette Score

The Silhouette Score is a metric of how well each sample fits into its assigned cluster. The highest average score highlights the optimal number of clusters, as the higher this score, the more distinct the clusters are. The increase in silhouette score from 4 to 5 clusters is significant.

Figure 4: *Silhouette scores for selected k-values*

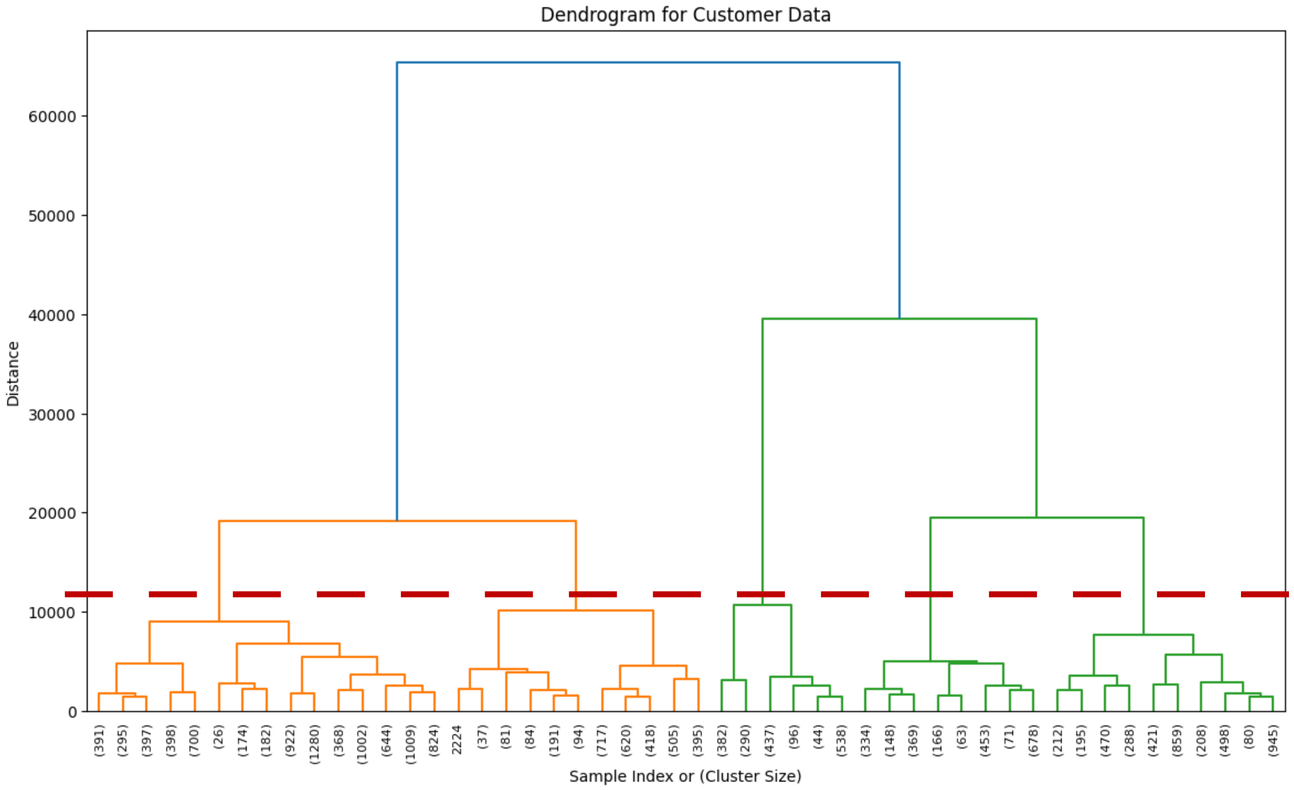


```
For n_clusters = 2, the average silhouette_score is : 0.6168912165332568
For n_clusters = 3, the average silhouette_score is : 0.5206071917559473
For n_clusters = 4, the average silhouette_score is : 0.4519497010144735
For n_clusters = 5, the average silhouette_score is : 0.46247269943067787
For n_clusters = 6, the average silhouette_score is : 0.4601066723156033
For n_clusters = 7, the average silhouette_score is : 0.3902456934825198
```

Dendrogram

Dendrograms show the merging of clusters with measurable distances, with vertical lines denoting distance between samples. On *Figure 5* below, the red line shows that 5 clusters ensures all samples in each cluster are within a reasonable distance of each other.

Figure 5: *Dendrogram of Customer Data*



Results

Principal Component Analysis (PCA) is inappropriate for visualisation, therefore t-distributed Stochastic Neighbour Embedding (tSNE) was performed to reduce the dimensions of the data, allowing it to be plotted in 2D and visualised clearly. The results are shown on *Figure 6* below:

Figure 6: *tSNE of k-means Clustering with 5 Clusters*

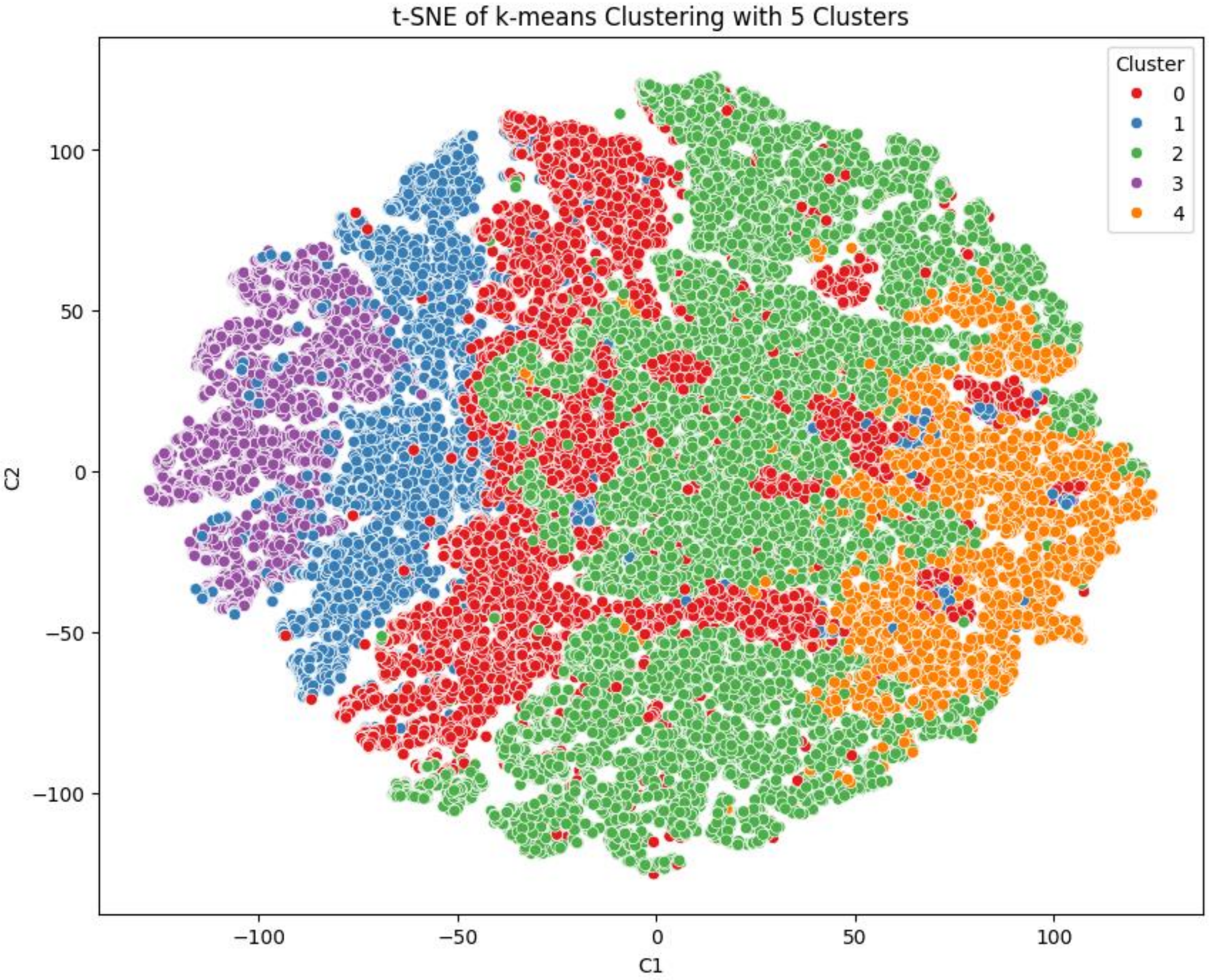


Figure 6 clearly shows 5 clusters, however there is significant overlap. Clusters 3 is distinct and tightly clustered, while Clusters 1 and 4 are relatively distinct, but have some significant variance. Clusters 0 and 2 show are not particularly distinct or tightly clustered, and overlap significantly.

As such, characteristics displayed by customers in Cluster 3 are distinct and can be used for an effective targeted marketing campaign, and the same could be said for customers in Clusters 1 and 4, with the caveat that it will be ineffective for some customers (particularly Cluster 1). It may be difficult to draw distinctions between customers in Clusters 0 and 2; as such marketing campaigns to those groups may be ineffective.

Clearly it is impossible to tell what the prominent features of these clusters are. As such, boxplots of each feature showing how the 5 clusters are distributed for that feature are shown below:

Figure 7:

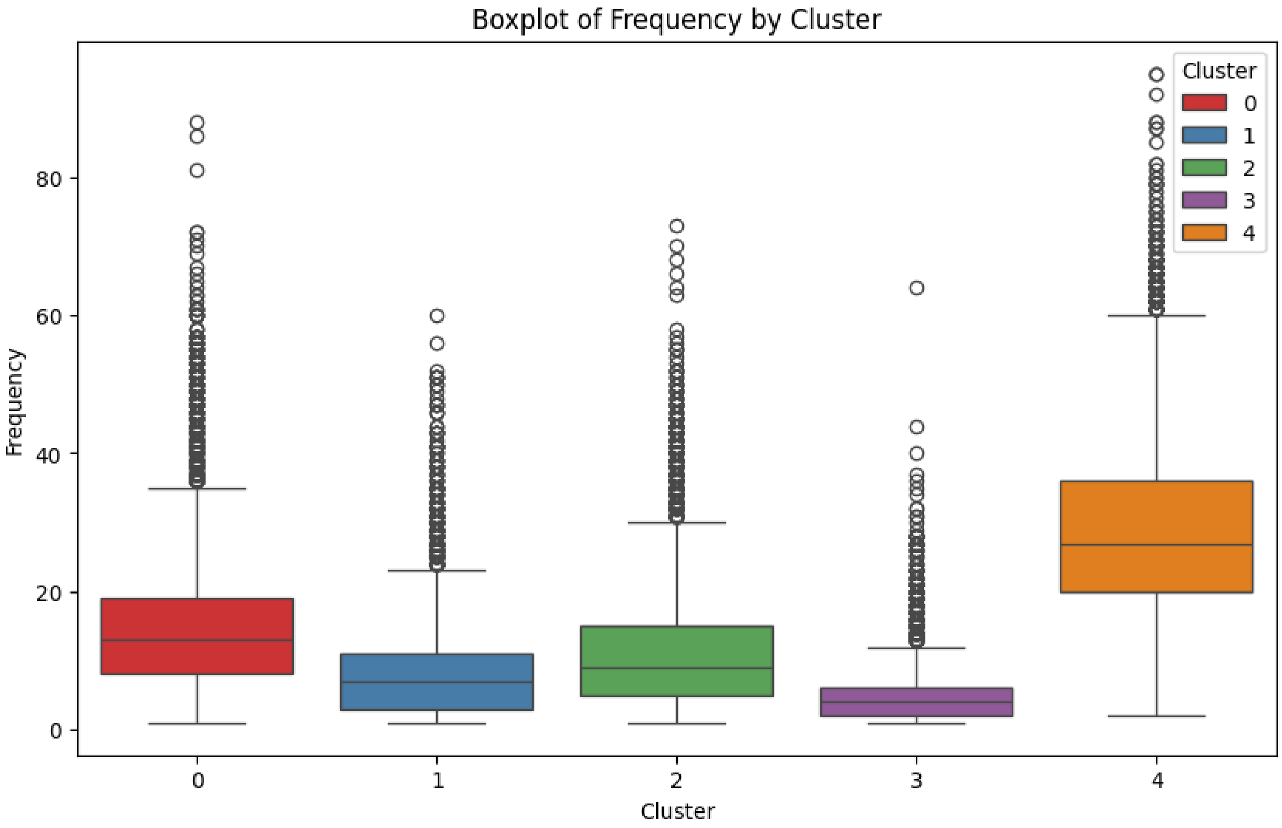


Figure 8:

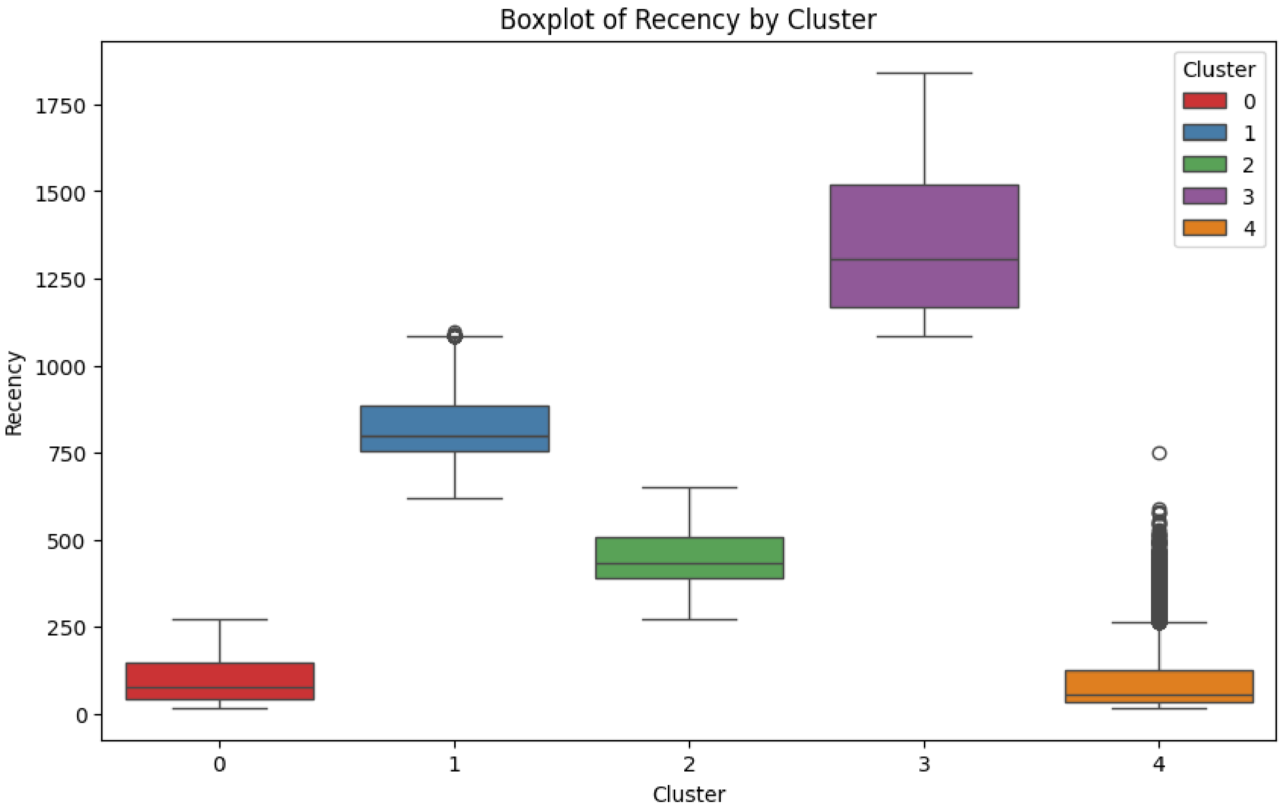


Figure 9:

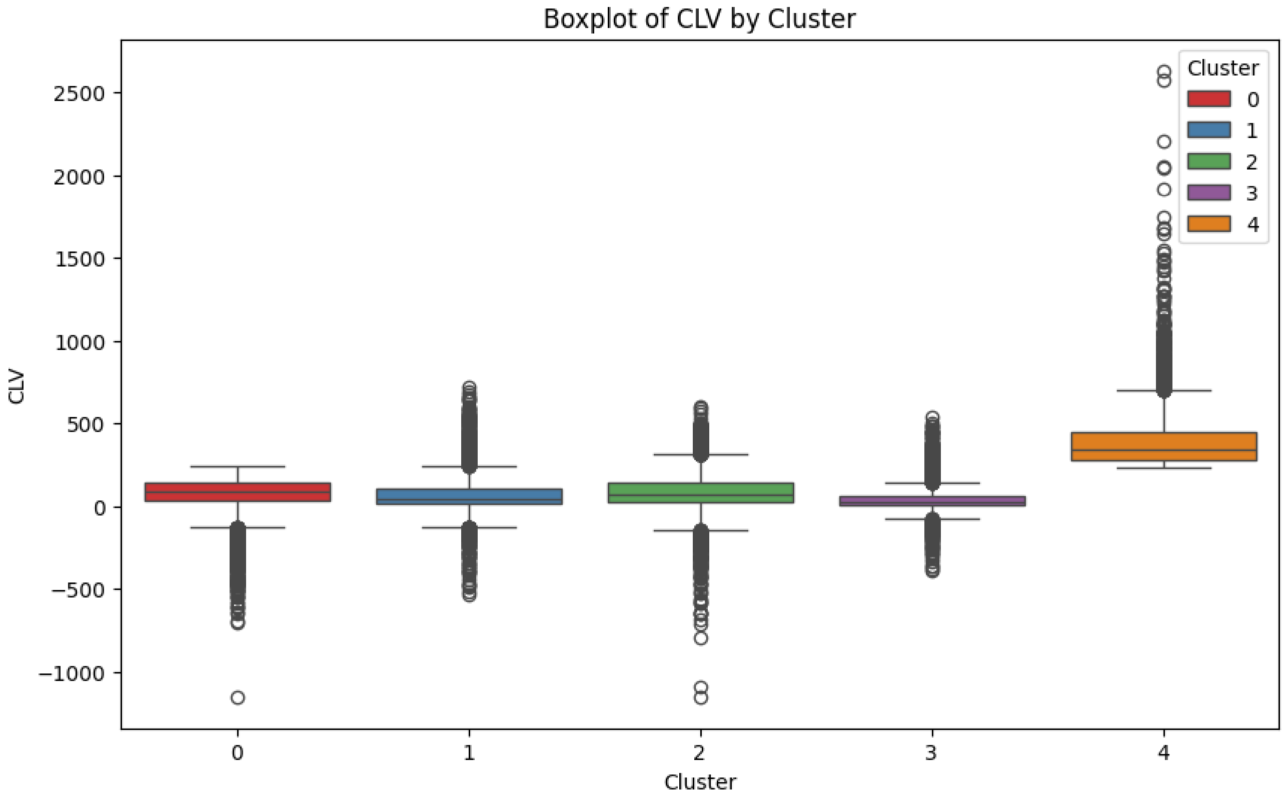


Figure 10:

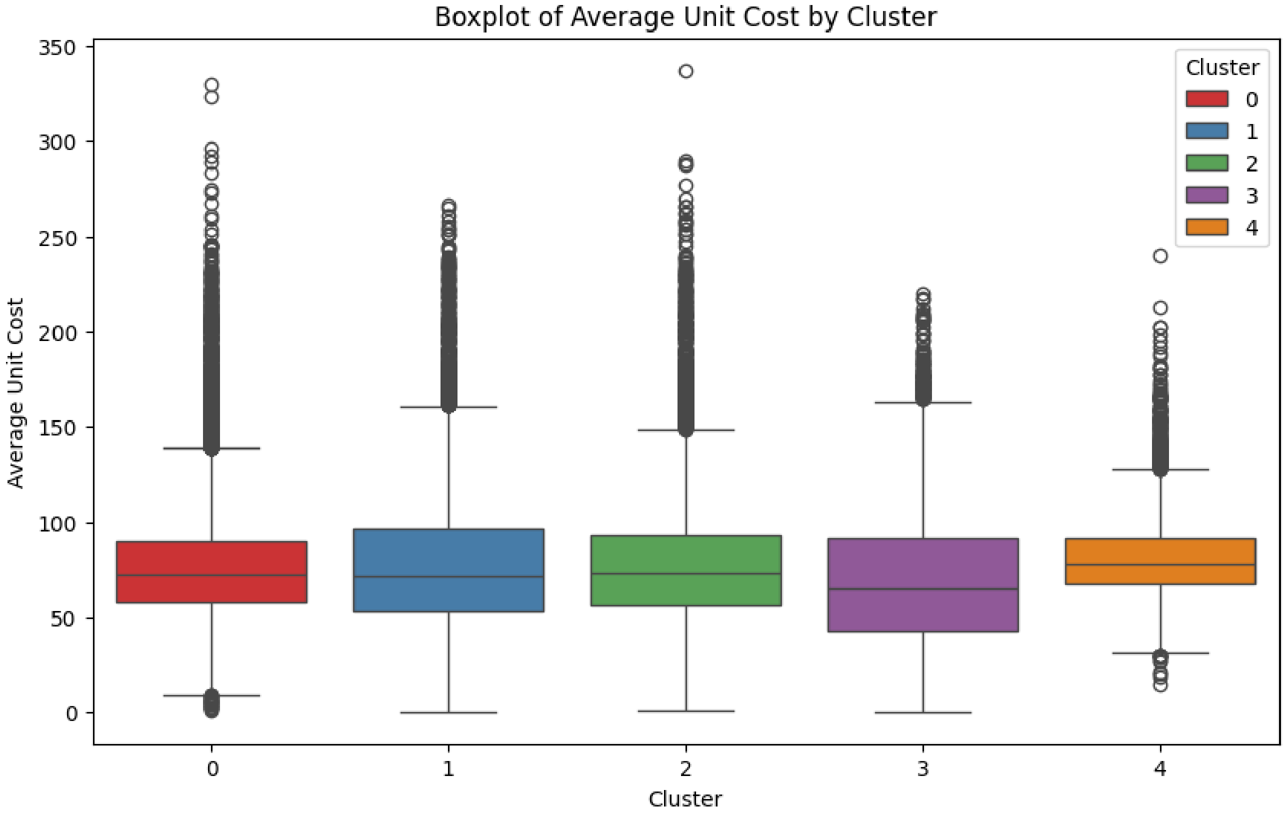


Figure 11:

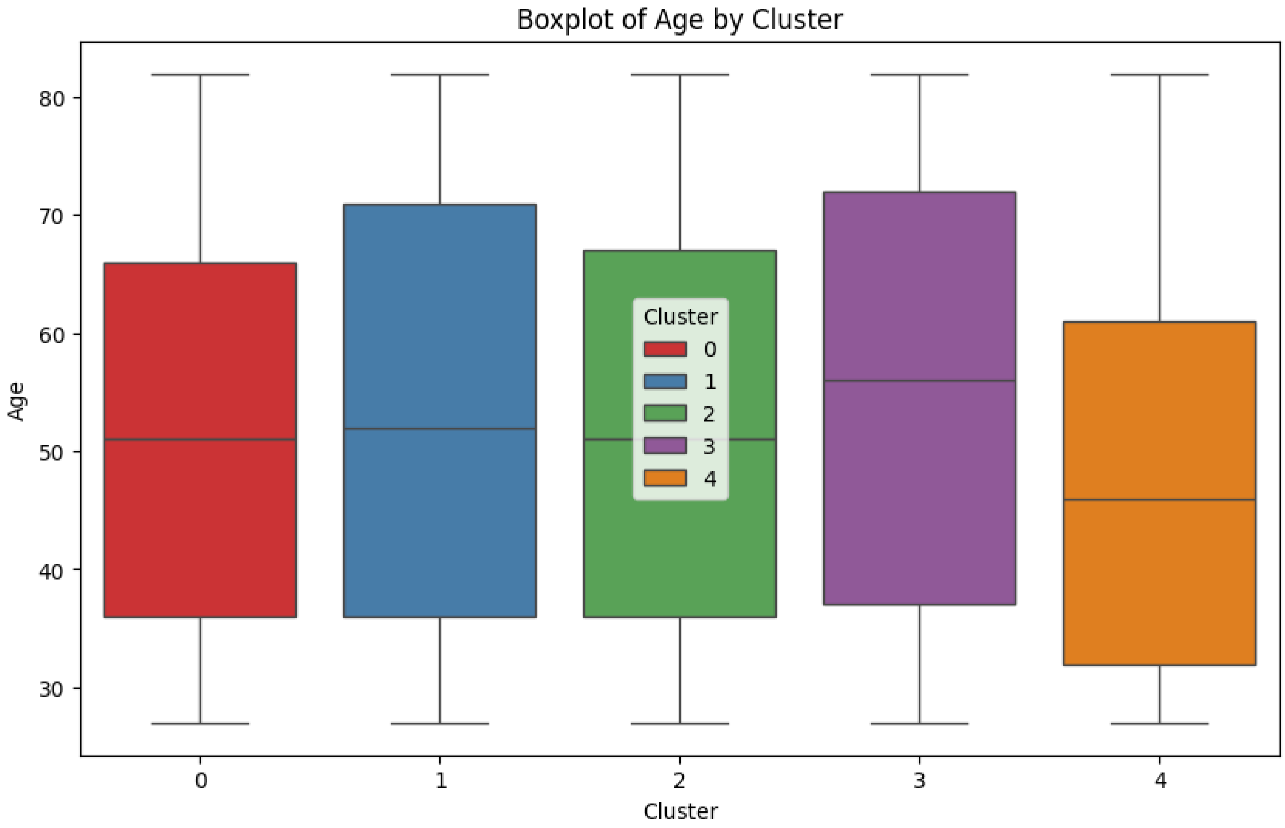


Figure 7 shows that Cluster 3 is made up of customers who last made purchases from the business around 4 years ago. It is unlikely that these customers will make another purchase. The same can be said for Cluster 1, which is made up of customers who last made a purchase around 2 years ago. Our analysis of *Figure 6* concluded that Clusters 0 and 2 lack distinction and the boxplots would support that, with the median Age, Frequency and Average Unit Cost of these clusters being broadly similar.

This leaves Cluster 4, with a higher median CLV and Average Unit Cost. These customers have made purchases recently but infrequently. These are clearly younger, high value customers, and therefore it is important to retain them.

Conclusions & Next Steps

It is clear that marketing efforts should focus on customers in Cluster 4; targeted marketing to encourage these customers to make more frequent purchases would be an effective way to increase the profitability of the business.

In terms of next steps, it may be prudent to identify anomalies in the dataset; clustering algorithms are sensitive to outliers, therefore a clearer output could potentially be generated from a dataset with fewer anomalies. Another clustering algorithm, such as hierarchical clustering, could be tested.

However, it is most likely that different features, such as average spend, would need to be selected to generate tighter, more distinct clusters. Moreover, removing customers who have not made a purchase within the last 2 years would streamline the dataset, enabling different patterns to emerge, whilst it is likely that those customers would not make another purchase.

References

Tripathi, S. et. al., (2018)., Approaches to Clustering in Customer Segmentation., International Journal of Engineering & Technology., doi:10.14419/ijet.v7i3.12.16505.