Customer Segmentation

0.1 Background

In the highly competitive e-commerce landscape, understanding customer behavior and preferences is crucial for personalized marketing, targeted promotions, and enhancing customer experience. Traditional segmentation methods often fall short in capturing the complex and dynamic nature of customer interactions. Therefore, leveraging machine learning techniques for clustering can provide deeper insights and more granular customer segments.

0.2 Objective

The objective of this project is to develop an unsupervised machine learning model to cluster customers based on their purchasing behavior, demographics, and engagement metrics. These clusters will help the marketing team to tailor strategies for each segment, ultimately aiming to increase customer retention, satisfaction, and lifetime value.

1. Exploratory Data Analysis

I have obtained the data set and now it is time to perform an Exploratory data Analysis (EDA) to gain insight about the dataset and prepare the data for modeling purposes.

1.1 About the Dataset

The dataset is the historical sales data that contains sales from 2010-12-01 to 2011-12-09 which is a total of 373 days. It has 8 features and 541,909 observations. The following are a short information of our feature :

- InvoiceNo : Invoice number of the customers purchase
- StockCode : Stock coed for the item purchased by the customer
- Description : Description of the product / Product name
- Quantity : Number of item purchases
- Invoice Date : Date of purchase
- Unit Price : Price of the item purchased
- Customer ID : ID of the purchasing customer
- Country : Country where the purchase was made

1.2 Data Preprocessing And Feature Engineering

1.2.1 Data Preprocessing

The data preprocessing includes the following steps:

- Cleaning and handling missing values.
- Scaling features. (Scaled features were not used as they provided worst results)

- Capping Outlier using IQR technique
- Dropping unnecessary features

1.2.2 Feature Engineering:

During the feature engineering process I use the original features to create some new relevant features that capture customer behavior and engagement. These features were obtained by grouping the original data by customer ID so that a relevant cluster identifying each customer could be obtained. A total of 3 features were derived and the following are these features:

- **Recency** : Number of days it has been since the customers last purchase
- **Purchase_Frequency** : The number of purchase the customer has made
- **Amount_Spent** : The total amount a customer has spent

2. Data distribution and Descriptive statistic Visualization

The following diagrams show the distribution and descriptive statistics of the features "Recency," "Purchase_Frequency," and "Amount_Spent" through histograms and box plots. These visualizations are crucial for understanding the characteristics of the data before and after certain preprocessing steps.

2.1 Histograms Before and After Min-Max Scaling

The first figure contains three histograms displaying the distributions of "Recency," "Purchase_Frequency," and "Amount_Spent" before applying Min-Max scaling. These histograms provide insights into the original data distribution, highlighting any skewness or outliers.

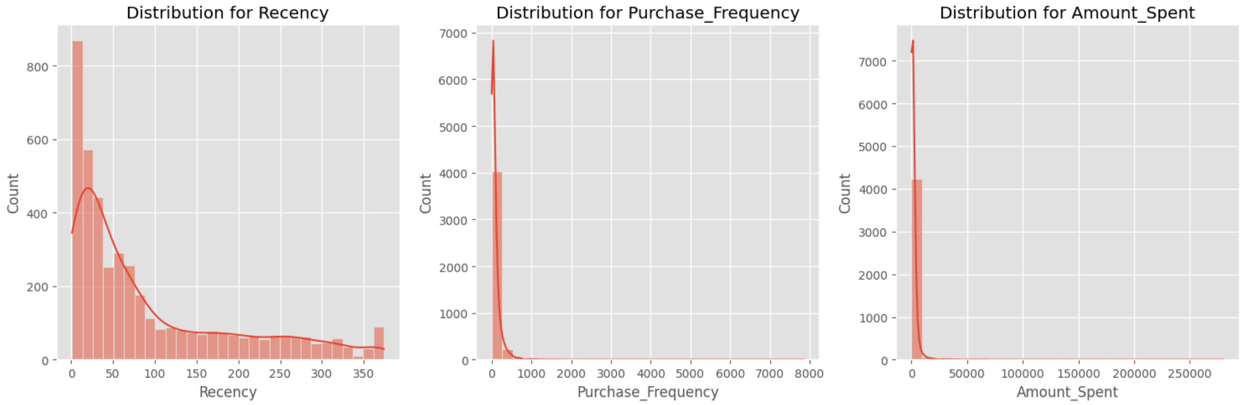

Figure 1. Histogram before scaling

The second figure contains three histograms of the same features after applying Min-Max scaling. These scaled histograms demonstrate how the data values have been transformed to a standard range, typically [0, 1], making them more suitable for machine learning algorithms that require normalized input.

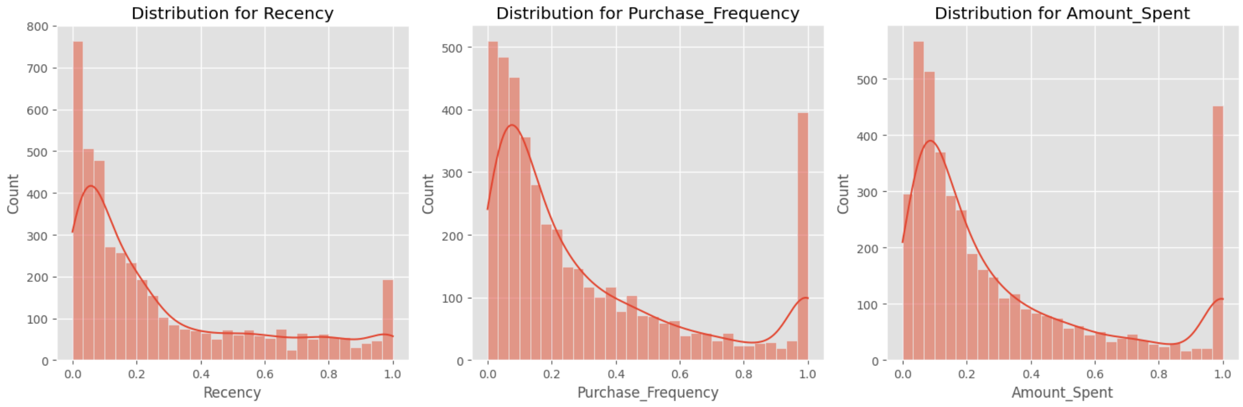

Figure 2. Histogram after min-max scaling

2.2 Boxplot Before and After Capping Outlier

The third figure presents three box plots for "Recency," "Purchase_Frequency," and "Amount_Spent" before capping. These box plots help identify outliers and understand the spread and central tendency of the data.

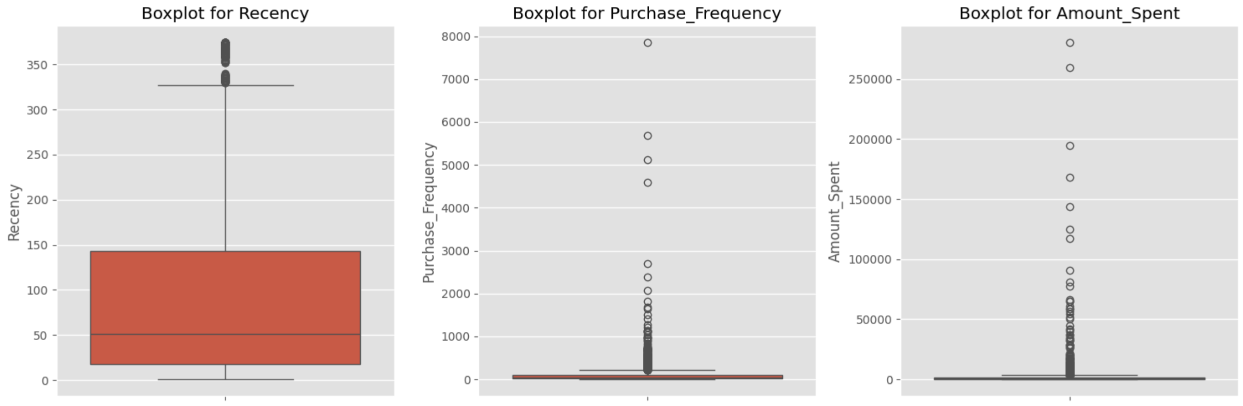

Figure 3. Boxplot before capping outliers

The fourth figure shows three box plots for the same features after applying the Interquartile Range (IQR) technique to cap the outliers. This capping helps to mitigate the effect of extreme values, providing a clearer view of the data's primary distribution.

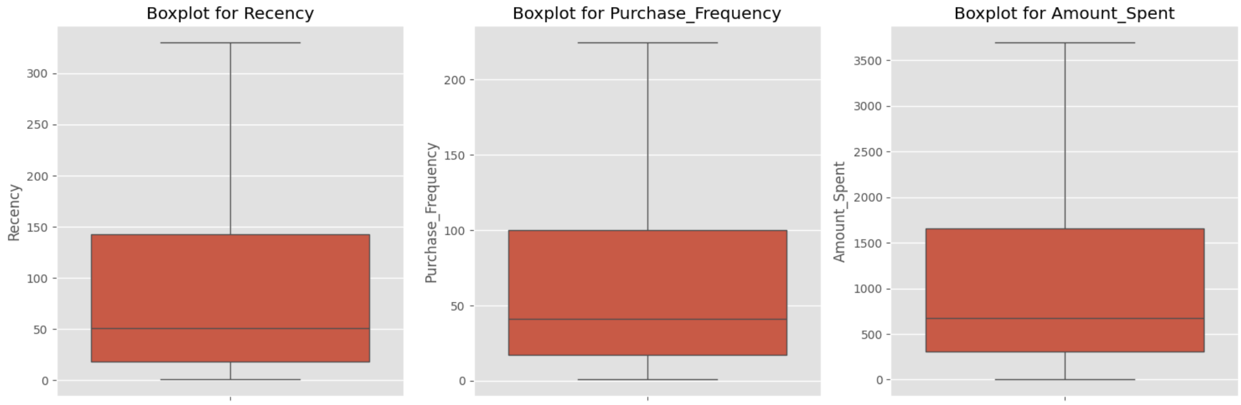

Figure 4. Boxplot after capping outliers using IQR

3. Machine Learning Approach

I will be implementing and comparing K-means, DBSCAN, and K medoid clustering to cluster our dataset. I have determined the optimal number of clusters using methods like the Elbow method, Silhouette score

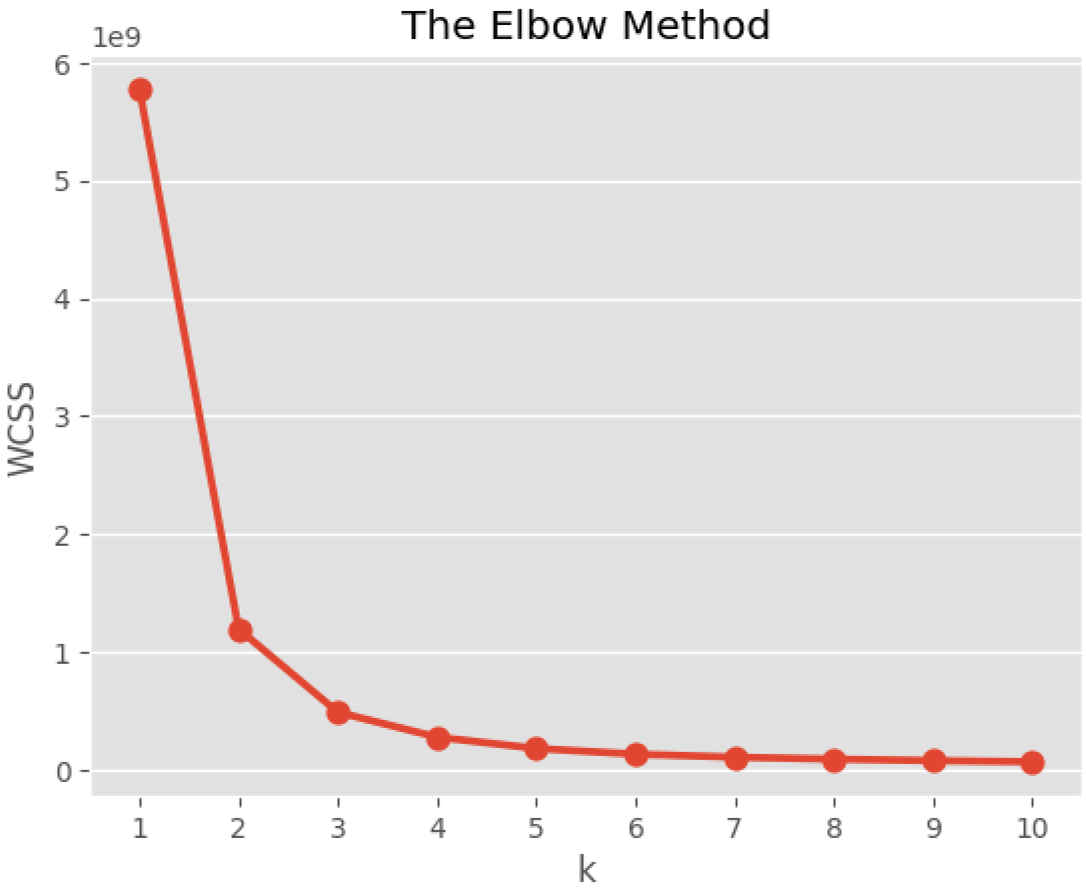

Figure 5. The Elbow Method

4. Evaluation

The following are the Silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Index for our models:

4.1 Metrics Overview:

Silhouette Score:

The Silhouette score is a metric used to evaluate the quality of a clustering algorithm. It provides a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, with higher values indicating better-defined clusters.

Silhouette Coefficient Calculation:

- **a(i):** The average distance from the i^{th} point to all other points in the same cluster. This represents how well the point is clustered with its own cluster (lower is better).
- **b(i):** The average distance from the i^{th} point to all points in the nearest cluster that it is not a part of. This represents how well the point is separated from the nearest other cluster (higher is better).

The Silhouette score for a single point is given by:

$$s(i) = \frac{b(i) - a(i)}{max(a(i),b(i))}$$

Overall Silhouette Score: The average of the Silhouette scores for all points in the dataset. A higher average score indicates better-defined and well-separated clusters.

Davies-Bouldin Index (DBI):

The Davies-Bouldin Index (DBI) is another metric used to evaluate the quality of a clustering algorithm. It is an internal evaluation scheme, where a lower value indicates a better clustering result. Specifically, the DBI is a measure of the average "similarity" ratio of each cluster with the cluster that is most similar to it.

- **Inter-cluster Distance:** It measures the distance between clusters. Smaller distances between clusters indicate that the clusters are not well-separated.

- **Intra-cluster Distance:** It measures the distance within each cluster. Smaller intra-cluster distances indicate that the clusters are more compact.

The DBI takes the ratio of intra-cluster distances to inter-cluster distances and then averages these ratios over all clusters. A lower DBI value implies:

- **Tighter Clusters:** Each cluster is more compact (low intra-cluster distance).
- **Better Separation:** Clusters are well-separated from each other (high inter-cluster distance).

Calinski-Harabasz Index (CH):

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is another metric used to evaluate the quality of a clustering model. This index assesses the ratio of the sum of between-cluster dispersion and within-cluster dispersion for all clusters.

- **Between-Cluster Dispersion (B):** This measures the variance between different clusters. Higher values indicate that clusters are well-separated from each other.
- **Within-Cluster Dispersion (W):** This measures the variance within each cluster. Lower values indicate that clusters are more compact.

$$CH = \frac{B}{W} \times \frac{(N-k)}{(k-1)}$$

where,

N = The total number of data points
k = The number of cluster

4.2 Model Scores:

The following are the scores for each of the models I used for clustering:

- **K means Clustering:**

```
Silhouette score: 0.6483413204987358
Davies-Bouldin Index: 0.48547272199943725
Calinski-Harabasz Index: 23263.69839505305
```

- **DBSCAN:**

```
Silhouette score: 0.5124535393925451
Davies-Bouldin Index: 0.3059262753930306
Calinski-Harabasz Index: 55.81098659050135
```

- **K-Medoids :**

```
Silhouette score: 0.6392076721049347
Davies-Bouldin Index: 0.5002676602413151
Calinski-Harabasz Index: 22700.2056197247
```

4.3 Conclusion

While DBSCAN has the lowest Davies-Bouldin Index, which suggests good separation and compactness, its lower Silhouette and Calinski-Harabasz scores indicate that it might not be the best overall performer for this dataset.

K-Means shows the best performance across multiple metrics, making it the most suitable model for your clustering problem based on the provided evaluation criteria.

5. Cluster Quality and Interpretation through Visualization

I will now access cluster quality and interpretability of the cluster through various visualizations with regards to the machine learning cluster. I have used the clustering results from K-means clustering as this model provided the best performance.

5.1 Cluster Visualization

The following are the visualization of our data set to determine the nature of each of our clusters.

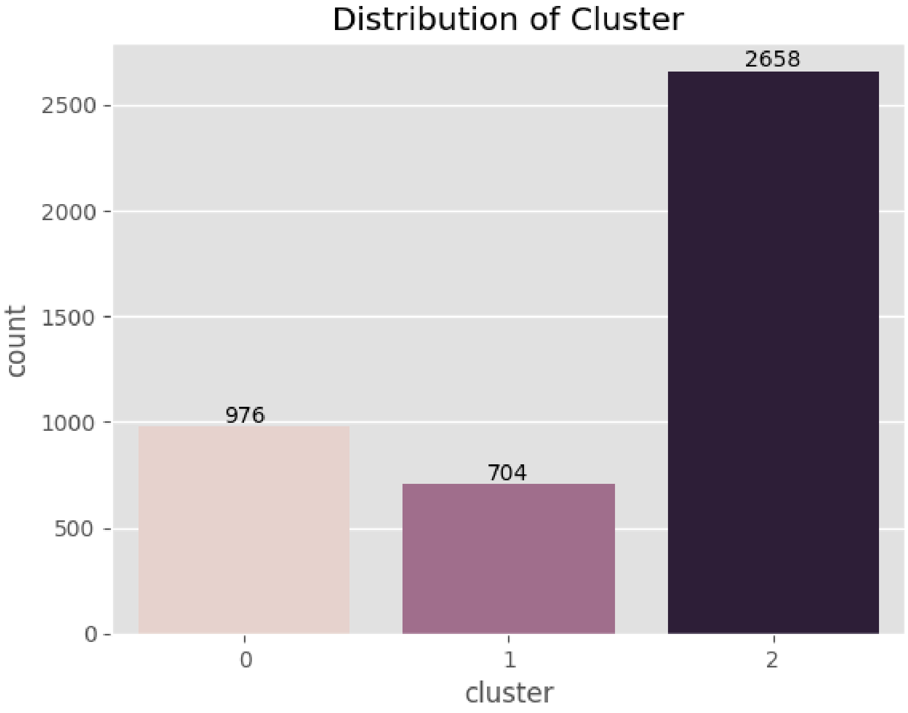

Figure 6. Cluster Countplot

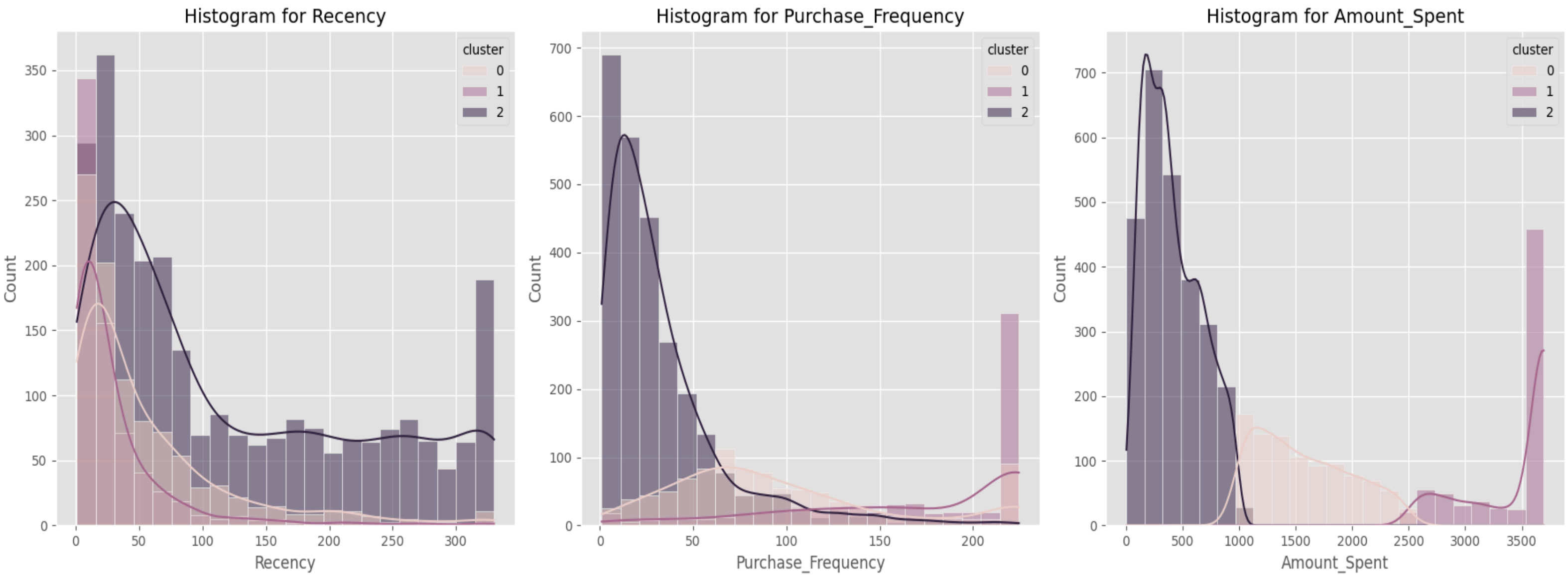

Figure 7. Data distribution based on cluster

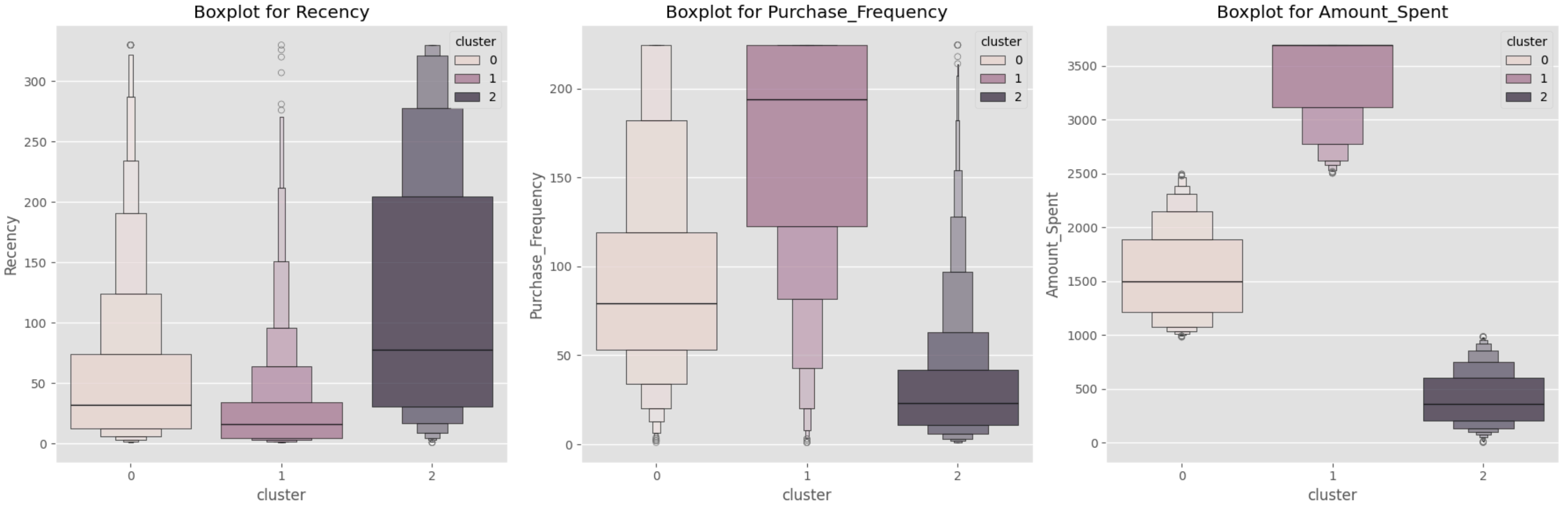

Figure 8. Boxenplot for cluster

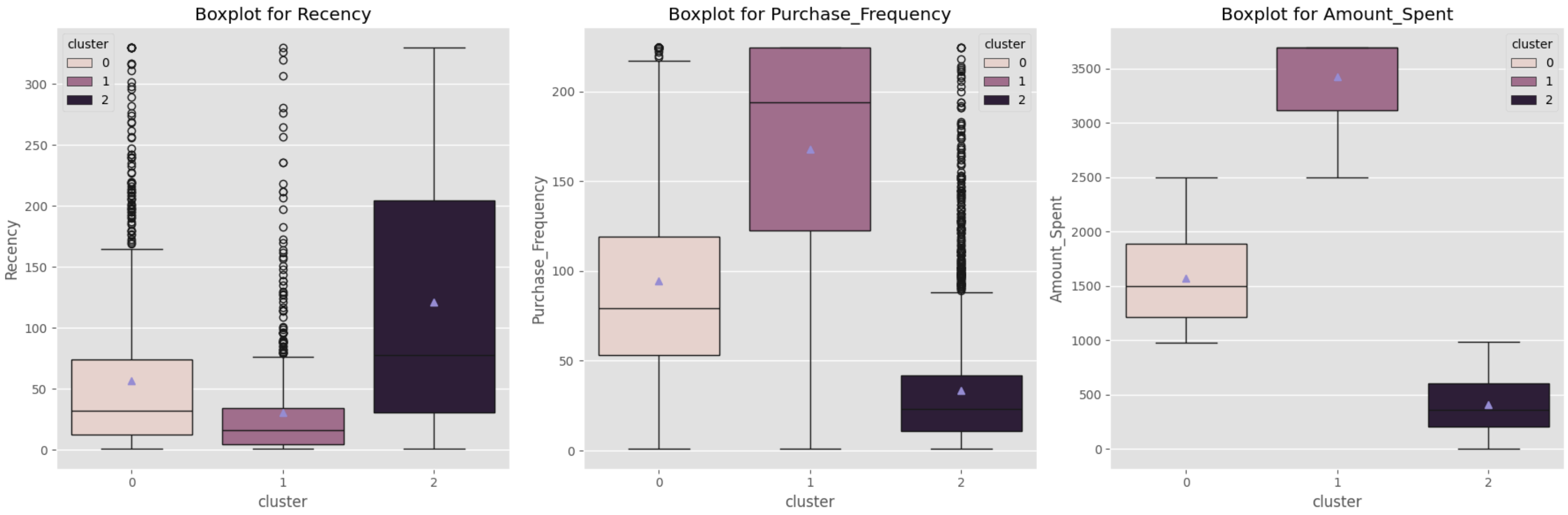

Figure 9. Boxplot for cluster with mean

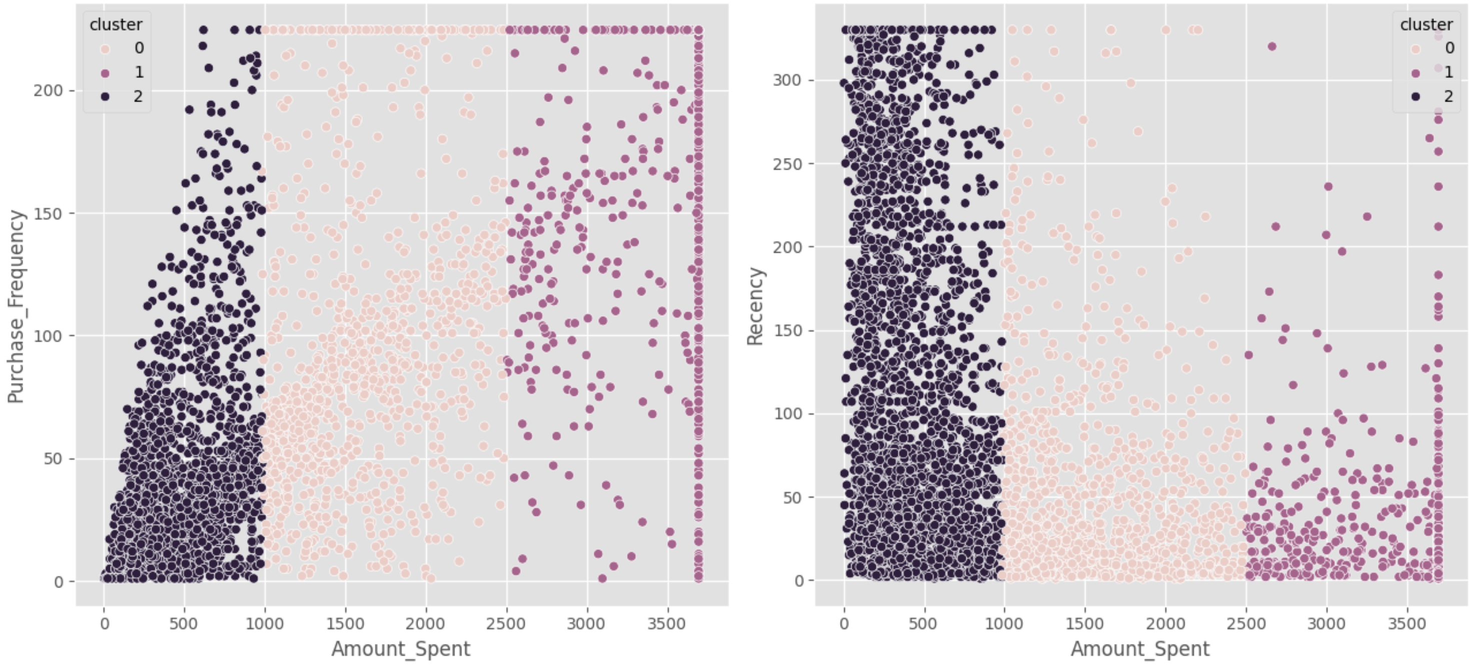

Figure 10. Scatter plot with our cluster

5.2 Cluster Observation

The following observations can be made about each of the cluster:

1. **Cluster 0 (Neutral Customers)**

- There are 976 unique customers in this cluster

- This group of customers are in the middle of heard in Recency, frequency and amount spent
- The customer in this cluster have spent on average a total of $1500 a year
- On average the customer of this cluster made a purchase 55 days ago from the final date of record
- Throughout the year the customers in this cluster makes 90 different transaction on average

2. **Cluster 1 (Loyal Customers)**

- There are 704 unique customers in this cluster
- This group of customers are mosts loyal ones with low Recency and high frequency and amount spent
- The customer in this cluster have spent on average a more than of $3400 a year
- On average the customer of this cluster made a purchase 30 days ago from the final date of record
- Throughout the year the customers in this cluster makes 170 different transaction on average

3. **Cluster 3 (Infrequent Customers)**

- There are 2664 unique customers in this cluster
- This group of customers are in the lower end with high Recency, and low frequency and amount spent
- The customer in this cluster on avg spend less than $500 a year
- On average the customer of this cluster made a purchase 125 days ago from the final date of record
- Throughout the year the customers in this cluster makes 30 different transaction on average