Capstone Project:

Machine Learning Using Python

Kmeans Optimal (Elbow Method) Clustering

Understanding Customer Purchase Patterns: Evidence-based Insights and Cluster Analysis for an Online Retail Store

Submitted by

Kannu Priya

Project 2

1. Problem Statement:

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

2. Project Objective:

The project objective is to analyze the customer purchase patterns for an online retail store and provide evidence-based insights. The goal is to gain a deep understanding of how customers engage with the store, identify key patterns and trends in their purchasing behavior, and provide actionable insights to optimize the store's operations and marketing strategies. By analyzing the data, the objective is to uncover valuable insights that can help improve customer satisfaction, increase sales, and enhance overall business performance.

3. Data Description:

The online_retail.csv contains 387961 rows and 8 columns.

Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

```
Number of null values in each column:
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
```

Nulls were found.

```
# Drop any rows with missing values
data = data.dropna()

null_counts = data.isnull().sum()
print("Number of null values in each column:")
print(null_counts)
```

```
Number of null values in each column:
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
```

Dropping Null values

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    406829 non-null  object
 1   StockCode    406829 non-null  object
 2   Description  406829 non-null  object
 3   Quantity     406829 non-null  int64
 4   InvoiceDate  406829 non-null  datetime64[ns]
 5   UnitPrice    406829 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      406829 non-null  object
```

4. **Data Pre-processing Steps and Inspiration:**

(i) Removed the null values.

(ii) Found outliers:

	Quantity	UnitPrice	CustomerID
count	406829.000000	406829.000000	406829.000000
mean	12.061303	3.460471	15287.690570
std	248.693370	69.315162	1713.600303
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13953.000000
50%	5.000000	1.950000	15152.000000
75%	12.000000	3.750000	16791.000000
max	80995.000000	38970.000000	18287.000000

Using Winsorization:

Winsorization replaces extreme values with the values at a specified percentile. This method helps limit the impact of outliers while still retaining some information about the original data.

```
# Handling Outliers

lower_bound = data['Quantity'].quantile(0.01)
upper_bound = data['Quantity'].quantile(0.99)
data['Quantity'] = data['Quantity'].clip(lower=lower_bound, upper=upper_bound)

lower_bound = data['UnitPrice'].quantile(0.01)
upper_bound = data['UnitPrice'].quantile(0.99)
data['UnitPrice'] = data['UnitPrice'].clip(lower=lower_bound, upper=upper_bound)
```

5. Choosing the Algorithm for the Project:

The chosen algorithm for the project is K-means clustering. K-means clustering is a popular unsupervised machine learning algorithm that partitions data points into K distinct clusters based on their similarity. It is suitable for this project as it can identify patterns and group customers based on their purchasing behavior.

6. Motivation and Reasons for Choosing the Algorithm:

K-means clustering was selected for its simplicity, efficiency, and interpretability. By applying K-means clustering, the project aims to uncover customer segments based on their purchasing patterns. This allows the retail store to gain valuable insights, such as identifying high-value customers, understanding different customer preferences, and tailoring marketing strategies accordingly.

7. Assumptions:

Some assumptions made for the project include:

- The selected features, 'Quantity' , 'UnitPrice', and 'Recency' adequately capture customer purchase patterns.

- The dataset is complete and representative of customer behavior.

- K-means clustering is appropriate for identifying distinct clusters in the data.

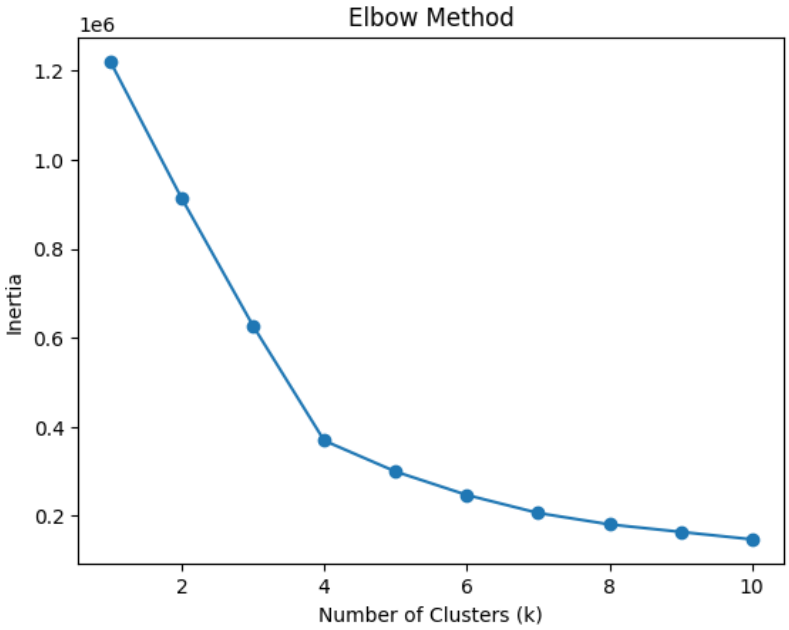
- The optimal number of clusters can be determined using the elbow method.

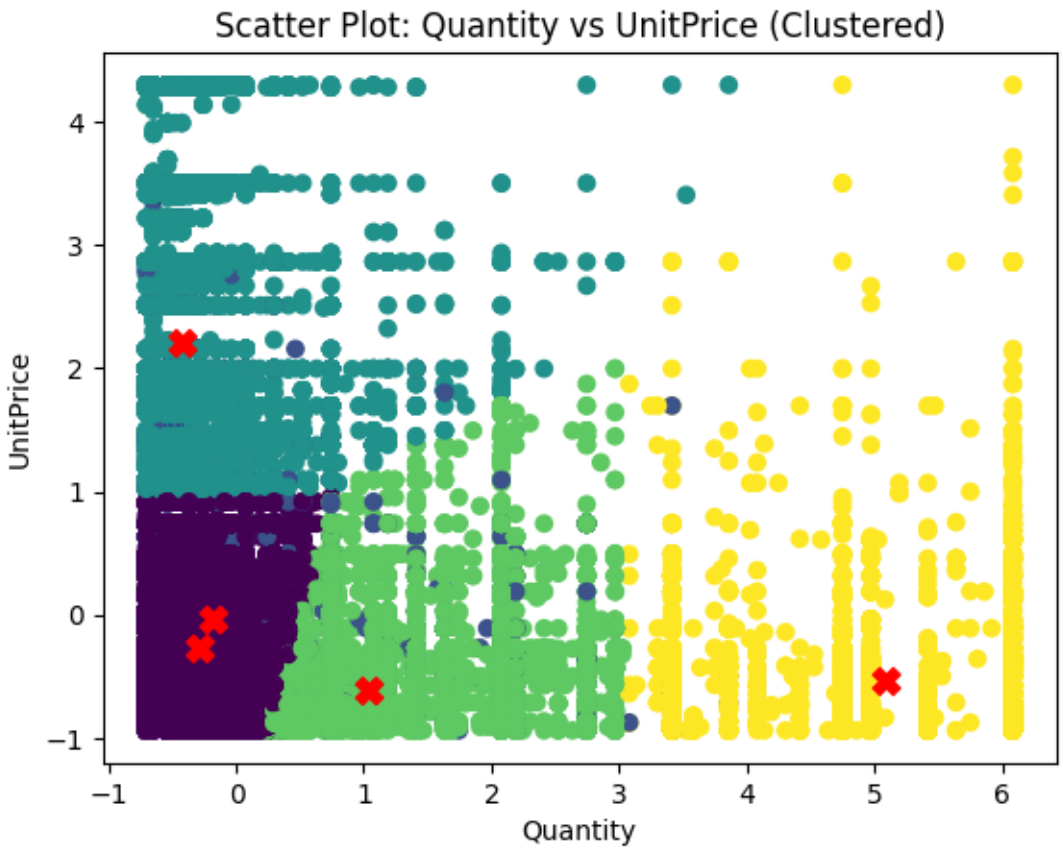
8. Model Evaluation and Techniques:

The elbow method is utilized to determine the optimal number of clusters for the data. The inertia (within-cluster sum of squares) is calculated for different values of K, and the point where the inertia starts to level off represents the optimal number of clusters. This technique helps in selecting an appropriate value of K for clustering.

9. Inferences from the Same:

The scatter plot of 'Quantity' versus 'UnitPrice' visualizes the data points before clustering. This plot helps observe the distribution and potential relationships between the two variables. The elbow curve plot aids in determining the optimal number of clusters. The project identifies that a value of K=5 appears to be a reasonable choice based on the elbow method.





10. Future Possibilities of the Project:

In the future, the project could explore additional clustering algorithms or dimensionality reduction techniques to further analyze customer purchase patterns. It could also consider incorporating additional features or external data sources to gain deeper insights into customer behavior. Further analysis and interpretation of the identified customer segments can lead to more targeted marketing strategies, personalized customer experiences, and improved business outcomes.