# Clustering Algorithm

# Title

The aim of this task is to apply clustering algorithm to the dataset of interest- the UCI Online Retail Store Dataset. This analytical research is titled “A Clustering Study: Customer Segmentation and Behavior Analysis”

# Introduction

This analytical research explores the area of clustering algorithms using the Adult Dataset. It involves grouping similar data points into clusters to identify natural patterns and compute the features of these said clusters. The task will also involve evaluation and clustering analysis comparison in order to gain actionable insights. (Bratchell, 1989)

# Dataset Description

The principal focus for this analysis is the UK based Online Retail Store Dataset [UCI Machine Learning Repo.](https://archive.ics.uci.edu/dataset/352/online+retail)

The dataset is comprised of 1,504 instances (excluding column names) and 8 features.

This dataset provides and in-depth view of customer interactions with the online store including details that shows the behavioral patterns per customer. This is a great characteristic for a dataset with potential to make real-work relevance.

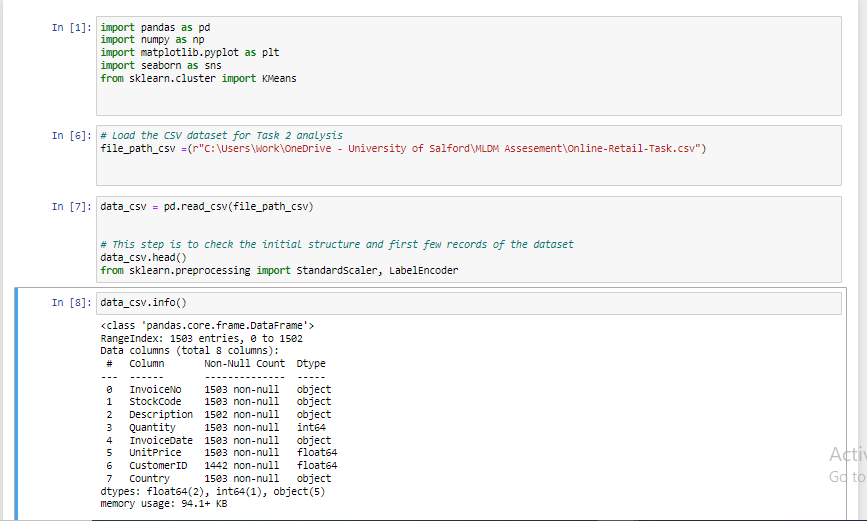
|  |  |
| --- | --- |
| **Variable** | **Variable Type** |
| Invoice No | Categorical |
| StockCode | Categorical |
| Description | Categorical |
| Quantity | Integer |
| InvoiceDate | Date |
| UnitPrice | Continous |
| CustomerID | Categorical |
| Country | Categorical |

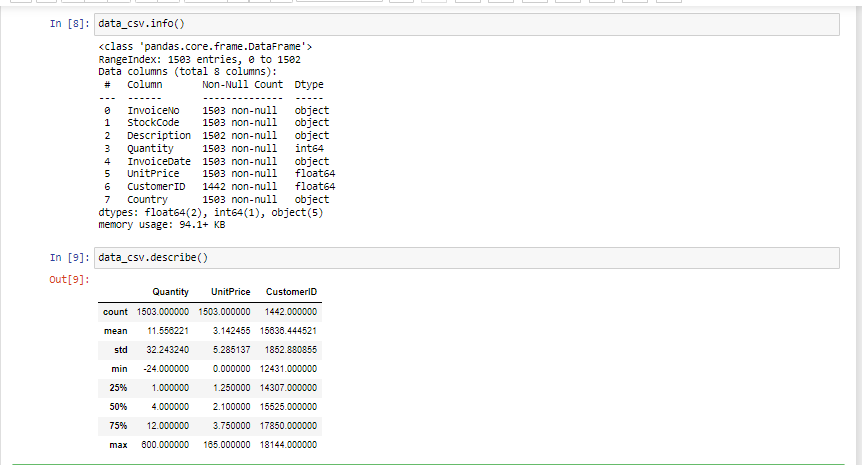
*Table 1: Variable Information of the online retail dataset from UCI.* (*Online Retail - UCI Machine Learning Repository*, n.d.)

# Data Exploration and Preprocessing

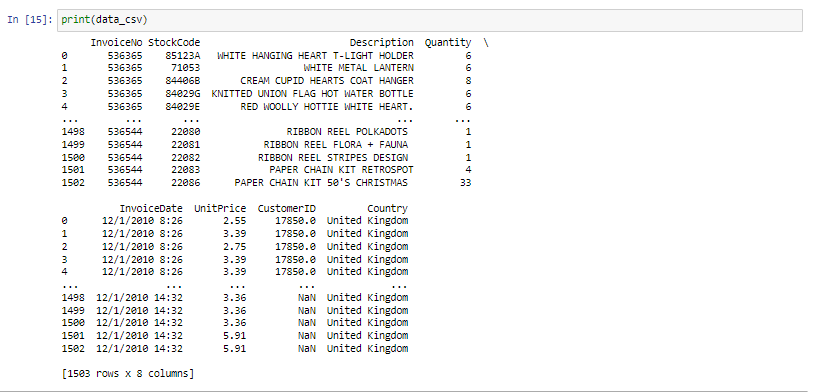
The data processing and exploration phase for this analysis is a critical step to prepare the dataset for clustering analysis.

1. Import necessary libraries, Load Data and Inspect the Data





The dataset has been uploaded from local machine and the output previewed to confirm correct schema and successful file upload. Data inspection has also been performed.

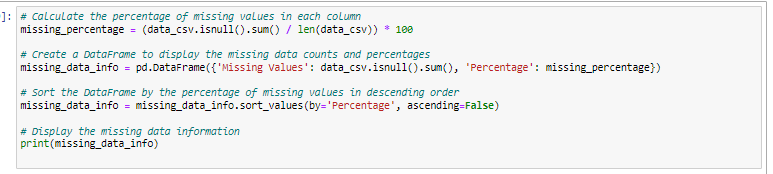


1. Handling Missing Values

The missing values within this dataset are very minimal compared to the overall data.

The dropna() method was used to resolve this because:

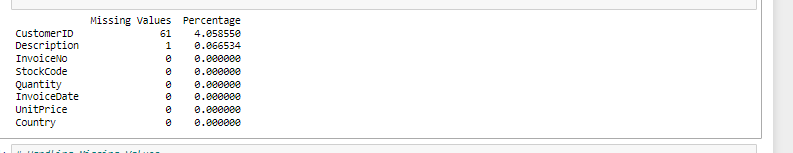
1. The total percentage of missing value
2. The variable having the missing value

This method was chosen with the strong aim to preserve the original distribution of the variables.

The summary of missing values:

-CustomerID : 61 missing values. About 4% of total data

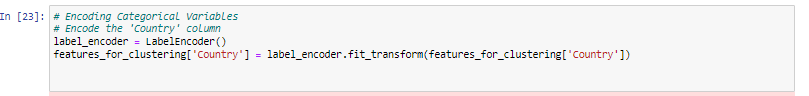
-Description : 1 missing value. About 0.067% of total data



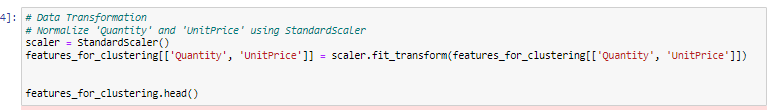


1. Data Encoding and Numerical Data Normalization

Data Encoding

This is usually performed to prepare the dataset for clustering. One Hot Encoding for Categorical Variables 'Country'. Since the algorithm would require numerical input for machine learning algorithms.

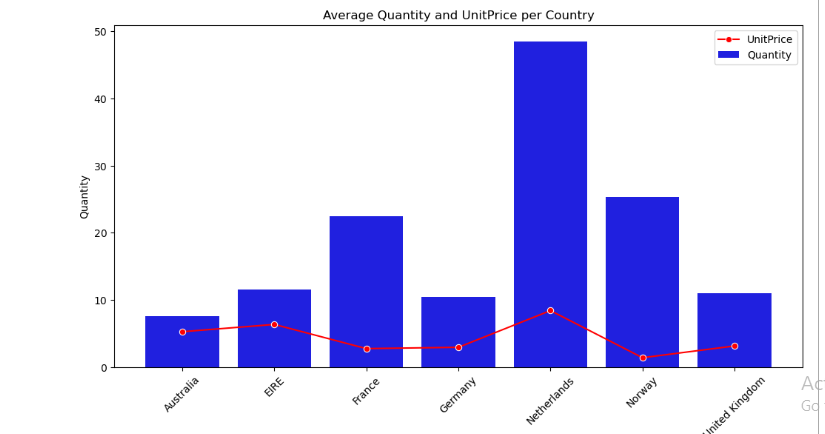
Data Normalization

Normalization is a crucial step in data pre-processing. The 'Quantity' and 'UnitPrice' features were standardized using the StandardScaler to ensure that all variables were on the same scale. Ensuring that no feature dominates the model due to its scale

1. Visual Exploration

To further understand the distribution of features, the dataset was visually analzysed.

A bar and line plot were created to view the Quantity and Unitprice interaction per country. This provides further insights into the purchasing patterns across various countries.



# Model Implementation and Result Evaluation

To perform the required clustering analysis, two algorithms will be applied and the performance evaluated.(Ikotun et al., 2023; Li et al., 2022)

|  |  |
| --- | --- |
| **Algorithm** | **Selection Justification** |
| K-Means Clustering algorithm | Simplicity and interpretability, scalability, comparison |
| Hierarchical Clustering algorithm | Scalability, comparison, hierarchy exploration |

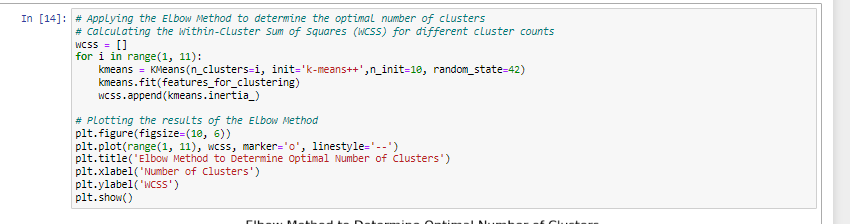
* Feature Selection for Clustering Analysis

A set of relevant features-'Quantity', 'UnitPrice', and 'Country', were selected for clustering. These features were selected based on their potential relevance to customer behavior patterns.



1. K-Means
2. Determine the optimal number of clusters visually using the WCSS Elbow Method

The elbow method was applied to determine the number of clusters using the Within-Cluster Sum of Squares (WCSS). With the WCSS, the optimal point can be observed as the point where WCSS begins to fall at slower rate. This may be referred to as the elbow point



1. Determine the optimal number of clusters mathematically using Silhouette Scoring Method

Determining the optimal number of clusters in K-Means clustering using the Elbow Method relies on visual inspection of the plot, and this makes it subjective. However, a mathematical approach to determine the optimal number of clusters is to apply the silhouette scoring method. This measures how similar an object is to its own cluster when it is compared to other clusters.

The silhouette score usually ranges from -1 to 1, where a high value infers that the object is well-matched to its own cluster while also poorly matched to neighboring clusters



1. K-Means Initialization and Cluster Analysis

Initialization of K-means clustering based on the number of clusters arrived at.

Use of init=k-means++ to initialize cluster centroids in a smart way to speed up convergence.

And finally, n\_init=10 specifies the number of times the algorithm will be run with different centroid seeds to get the best result.

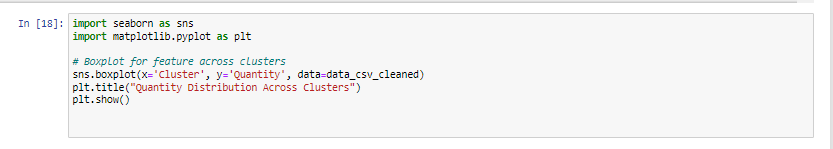
The random\_state=42 will ensure result reproducibility.



The clusters are analyzed by checking the size of each cluster (number of data points in each cluster) and calculating the mean values of features within each of those clusters.

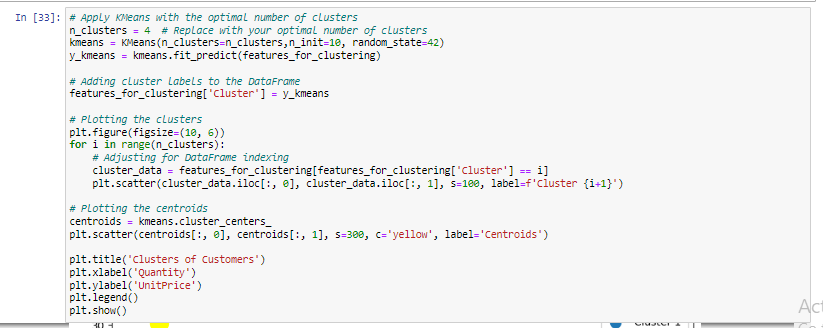
1. Quantity Boxplot

A boxplot is plotted to visually analyze the distribution of the quantity feature across clusters, aiding in understanding the spread f purchasing quantities within each cluster.

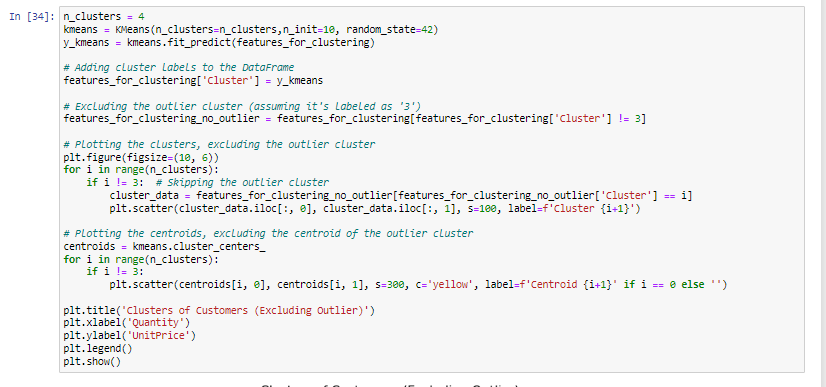


1. Cluster Visualization (*and visualization for excluded outlier cluster*)

The computed clusters are now visualized in a 2D scatter plot using 'Quantity' and 'UnitPrice' as the two axes. Each cluster is represented by a different color, and centroids are marked in yellow.



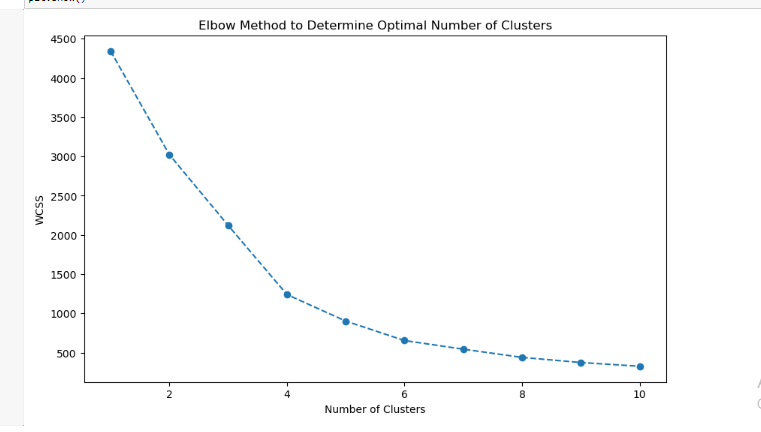
To gain broader insights, another visualization is created but with the outlier cluster excluded so as to focus on the core clusters. Nonetheless, the centroids for the remaining clusters are plotted.



ia. K-Means Results

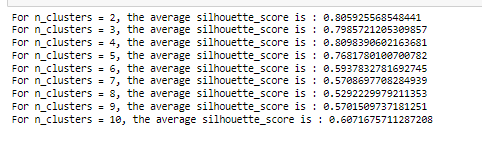
1. Determine the optimal number of clusters visually using the WCSS Elbow Method

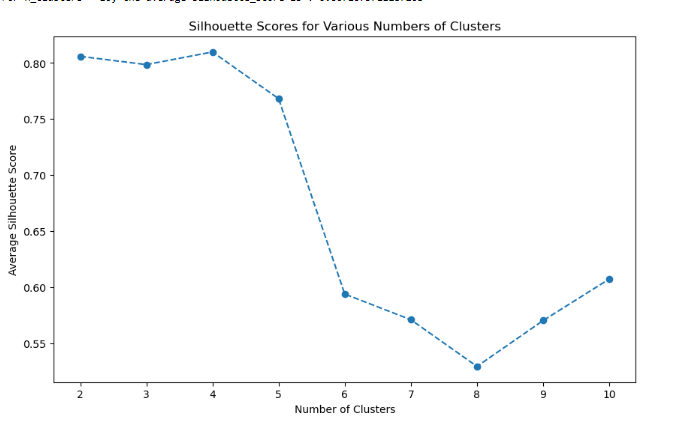
According to the visuals, the elbow point (i.e number of cluster is 4)



1. Determine the optimal number of clusters mathematically using Silhouette Scoring Method

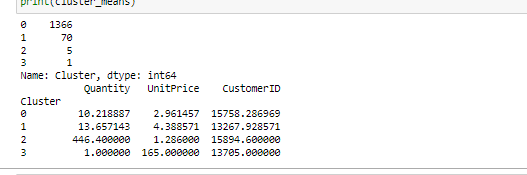
Accuracy is an important quality for this analysis, while the WCSS method provides visuals for the number of clusters; numbers would not lie. Hence this mathematical approach.





The highest average silhouette score is for 4 clusters, with a score of 0.8098. This infers that a clustering configuration with 4 clusters is the most appropriate and it indicates required interaction among the clusters. The confirmed number of clusters is 4.

1. K-Means Initialization and Cluster Analysis

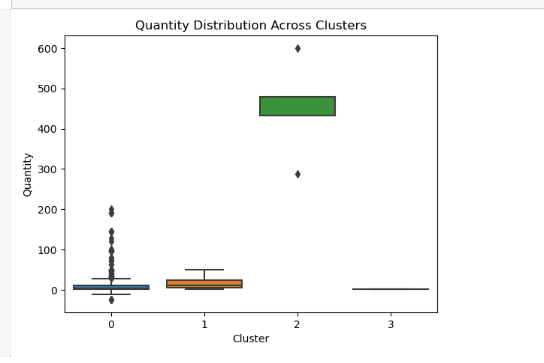


The cluster size indicates the number of transactions that belongs to each cluster.

This provides information on how datapoints (or transactions) are distributed across clusters. Lager clusters indicates that more data points show similar purchasing patterns.

The cluster mean aids in describing the characrterstics per cluster y showing the mean quantity, mean unit price, etc. From the results, we can infer that Cluster 0 represents customers with moderate quantities and unit prices, while Cluster 2 suggests a group of customers who make bulk purchases with a very high mean Quantity.

1. Quantity Boxplot



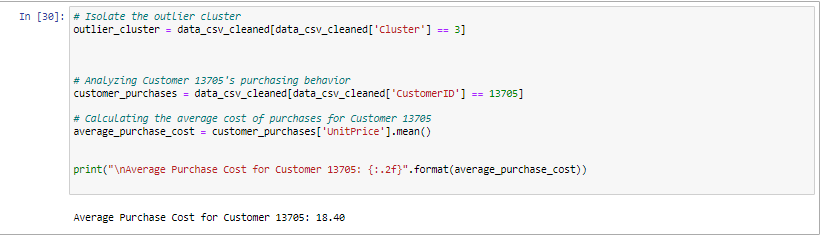
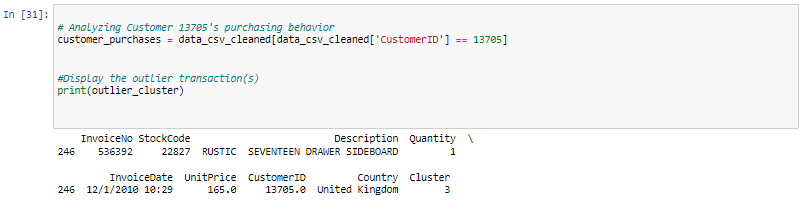
The aim of this plot is to examine the quantity distribution across clusters- businesswise, this would aid in spotting outliers (positively or negatively).

Cluster 0 (Mainstream Transactions): The largest cluster, representing regular purchasing behavior with fair quantities and unit prices.

Cluster 1 (Higher Value Transactions): A smaller cluster with slightly higher quantities and unit prices, which may be indicating a premium customer segment.

Cluster 2 (Bulk Purchases): A very small cluster characterized by large quantities and lower unit prices, most likely representing bulk or wholesale transactions.

Cluster 3 (Outlier Transaction): A single transaction with a very high unit price, standing out as an outlier. Further analysis will be carried out on this cluster. To reiterate, this cluster is an outlier because of it’s position far above the whisker box which already has the highest mean quantity

* Further Outlier Examination

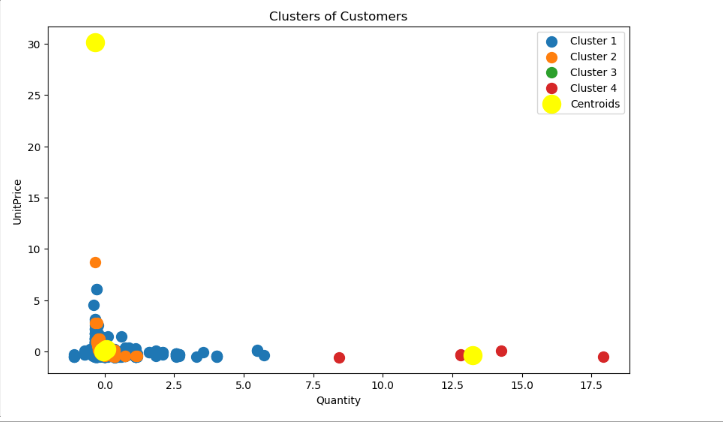
From further examination of the dataset:

-The outlier seems to be a one-off unit purchase

-The item “RUSTIC SEVENTEEN DRAWER SIDEBOARD” seems to be an expensive product as only one unit of the is purchased in the dataset. But this outlier transaction deviates from the norm.

-Customer 13705 is not a "one-time-buy-customer". The average cost of his/her purchases is 18.39

- Decision: To maintain a focus on regular customer behavior and ensure representativeness in the across analysis, the decision has been made to exclude this outlier transaction from further analysis. This approach should prevent skewing the results and will allow for more accurate insights into general purchasing trends.

ei. Cluster Visualization (*and visualization for excluded outlier cluster*)

-Cluster 1 (Blue): This cluster indicates customers with low quantity and low unit price. This cluster has many points close to the origin. This blue group may represent a group of customers that make smaller, less expensive purchases.

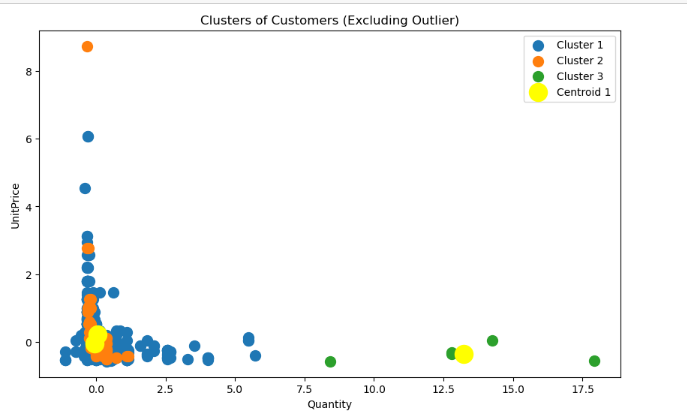
-Cluster 2 (Orange): This cluster spreads out more along the quantity axis, however it still remains low on the unit price axis. These customers are buying in larger quantities but still at lower unit prices. Possibly, this may be indicative of bulk purchasing of less expensive items

-Cluster 3 (Green): This could represent customers purchasing premium items but they are purchased in smaller quantities. The green cluster is sparse and it seems to have higher unit prices on average than clusters 1 and 2, but the quantity is still relatively low. Perhaps, they are buying at discounts.

-Cluster 4 (Red): This group could be representative of high value customers, making large and expensive purchases, but less frequent. Represented by the red cluster, this group has the fewest points, scattered mostly at higher quantities and higher unit prices.

-Centroids (Yellow): These are points that represent the average (mean) quantity and unit price for each cluster. The centroids are essential as they are used to define the center of each cluster; summarizing the typical behavior within each cluster.

eii. Cluster Visualization Excluding Outlier



Excluding the outlier though simplifies the plot but do not offer a significant variation. Nonetheless, the exclusion of the outlier may potentially show much accurate behavior in each cluster.

Conclusion

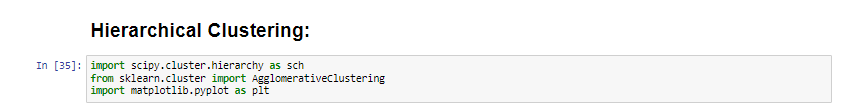
An overview on this scatter plot shows that a business could infer different customer behaviors and potentially build marketing strategies or stock decisions accordingly. For instance, customers in Cluster 4 might be targeted for luxury or high-value items, while those in Cluster 1 might be more price-sensitive.

1. Hierarchical Clustering
2. Import Libraries

Import necessary libraries:

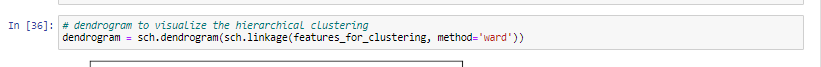
-. scipy.cluster.hierarchy used for generating the dendrogram to visualize the hierarchical clustering.

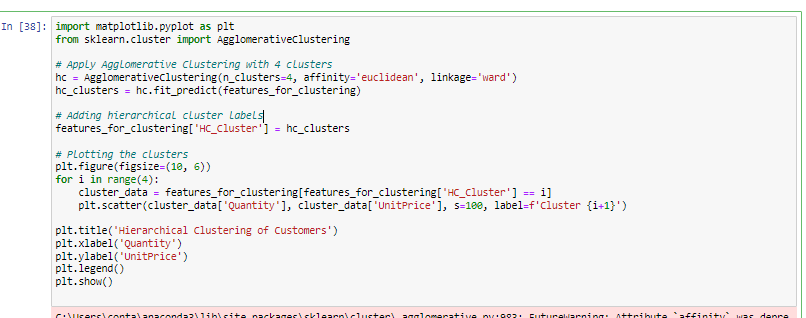
-The AgglomerativeClustering class from sklearn.cluster is used to perform the actual clustering.



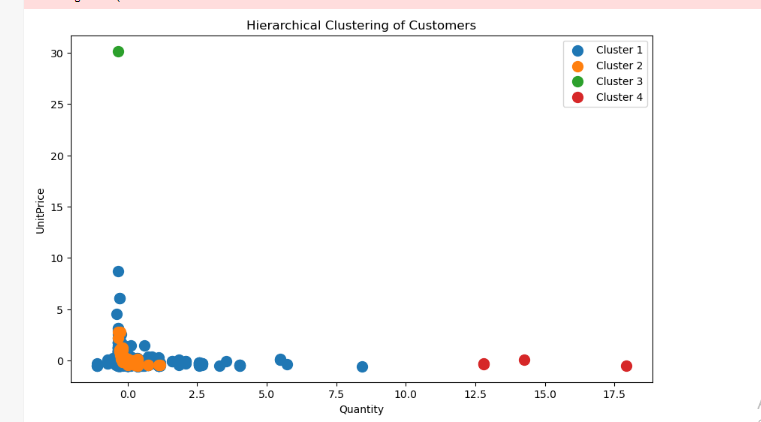
1. Dendrogram Visualization

sch.dendrogram() and sch.linkage() functions are used in this type of clustering analysis. They apply hierarchical clustering with the Ward's method. Ward's method is an agglomerative clustering method.

The dendrogram aids in deciding the number of clusters by visualizing the point at which clusters are joined together.

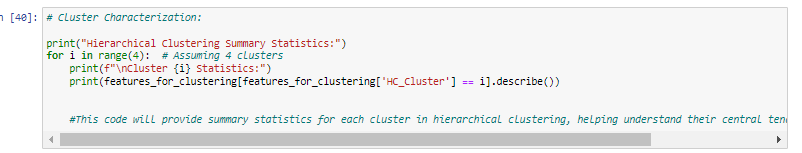
1. Applying Agglomerative Clustering
2. Plotting Clusters

To visualize the clusters based on 'Quantity' and 'UnitPrice', a scatter plot is created.



1. Cluster Characterization

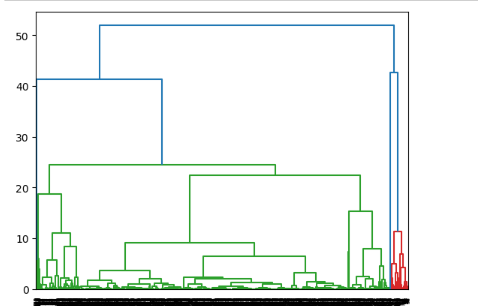
Characterization of the cluster involves the summary statistics for each cluster. This includes count, mean, standard deviation, min, quartiles



Iia. Hierarchical Clustering Results

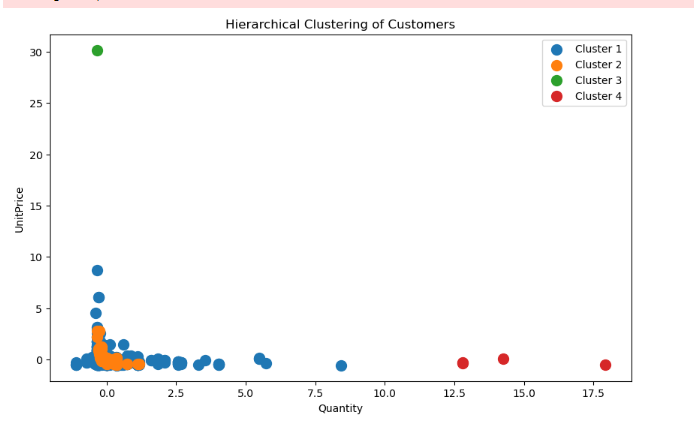
1. Dendrogram Visualization Result

The visualization shows the hierarchical clustering process with the merge of clusters being represented by the vertical line, while the height of the merges indicates the dissimilarity between the clusters.



1. Plot Clusters

This plot visualizes the final clustering. Each point representing a customer and their respective cluster based on 'Quantity' and 'UnitPrice', inferring how the customers are distributed within each cluster and the relative distances between the said clusters.



1. Cluster Characterization

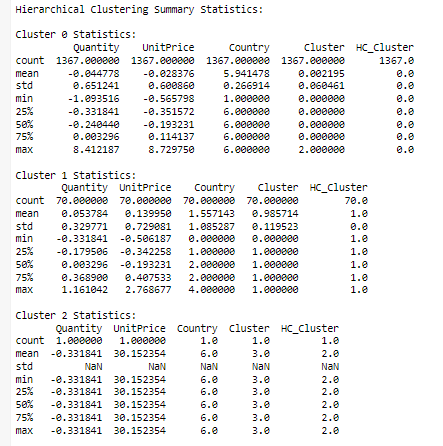
From the apparent statistical results from each of the clusters, these results infer that:

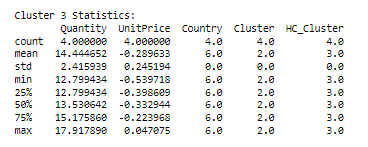
-Cluster 0 represents the "typical" customer with small, less expensive purchases.

-Cluster 1 has customers who spend more both in terms of quantity and unit price, but with greater variation in spending pattern.

-Cluster 2 is not really a cluster but a single outlier with a very high unit price.

-Cluster 3 represents a very small group of customers who are bulk purchasers.





Conclusion:

The hierarchical clustering analysis has segmented customers into four different groups, providing insights into their purchasing behavior and power- determining how much and how many they buy.

In real-world scenario, this segmentation will provide the base for targeted strategies to cater for different customer clusters. It is a strategic marketing tool indeed.

# General Conclusion

In this analysis, two different clustering methods were applied, K-Means and Hierarchical Clustering, to segment customer transaction data from the UCI Online Retail dataset. The main aim was to identify relevant customer segments based on purchasing behavior, which can shed valuable light for business strategies, marketing and decision-making.

-K-Means Clustering:

The K-Means clustering method successfully segmented the dataset into four different clusters. These clusters exhibited clear separation in scatterplots, and the centroids were also well-defined. Each of the clusters had unique characteristics in terms of mean quantities, unit prices, and customer IDs, providing actionable insights into customer behavior.

-Hierarchical Clustering:

Hierarchical Clustering was also applied which resulted in four clusters. However, it was observed that the separation and interpretability of these clusters were less pronounced compared to K-Means. While this method provided an alternative perspective on clustering, it did not offer the same level of clarity and separation as K-Means did.

# Recommendation

Based on the just concluded clustering analysis, the K-means is the recommended choice .K-means provided separation and interpretability, making it efficient in identifying customer’s behavioral patterns.