Bath Soap Report

1. Explore the Data
   1. Descriptive statistics (mean, median, mode, standard deviation, for each variable)
   2. Correlation analysis (find whether variables are highly correlated)
      1. If so, we can drop similar variables
2. Data Pre-Processing
   1. Scouting (Normalize the data)
   2. Handling missing values, check for missing values and decide on the imputation or remove based on the context.
3. Clustering
   1. Use KMeans clustering
   2. Hierarchical clustering
   3. DB scan (Density Based clustering)
   4. After each measure we need to analyze the cluster centroid representative point, check the size of the cluster form, avoid having very small or disproportionally large clusters, and assess tightness of the cluster, and separation between clusters.
4. Evaluation
   1. Check the cluster distinctness
   2. Check the cluster consistency
   3. Check the cluster simplicity
5. Interpretation and Recommendation
   1. For each cluster describe the main characteristics. Suggest market and promotional strategies.
   2. Highlight any cluster that shows strong brand loyalty, frequent purchases, or high susceptibility to promotions.
   3. Discuss the demographics of each cluster. Find any patterns or notable attributes.
6. Justification
   1. Choice of clustering algorithm
   2. Number of clusters
   3. Feature selection

Main Summary:

1. Do hierarchical clustering to explore the number of clusters.

2. Discuss the optimal number of clusters.

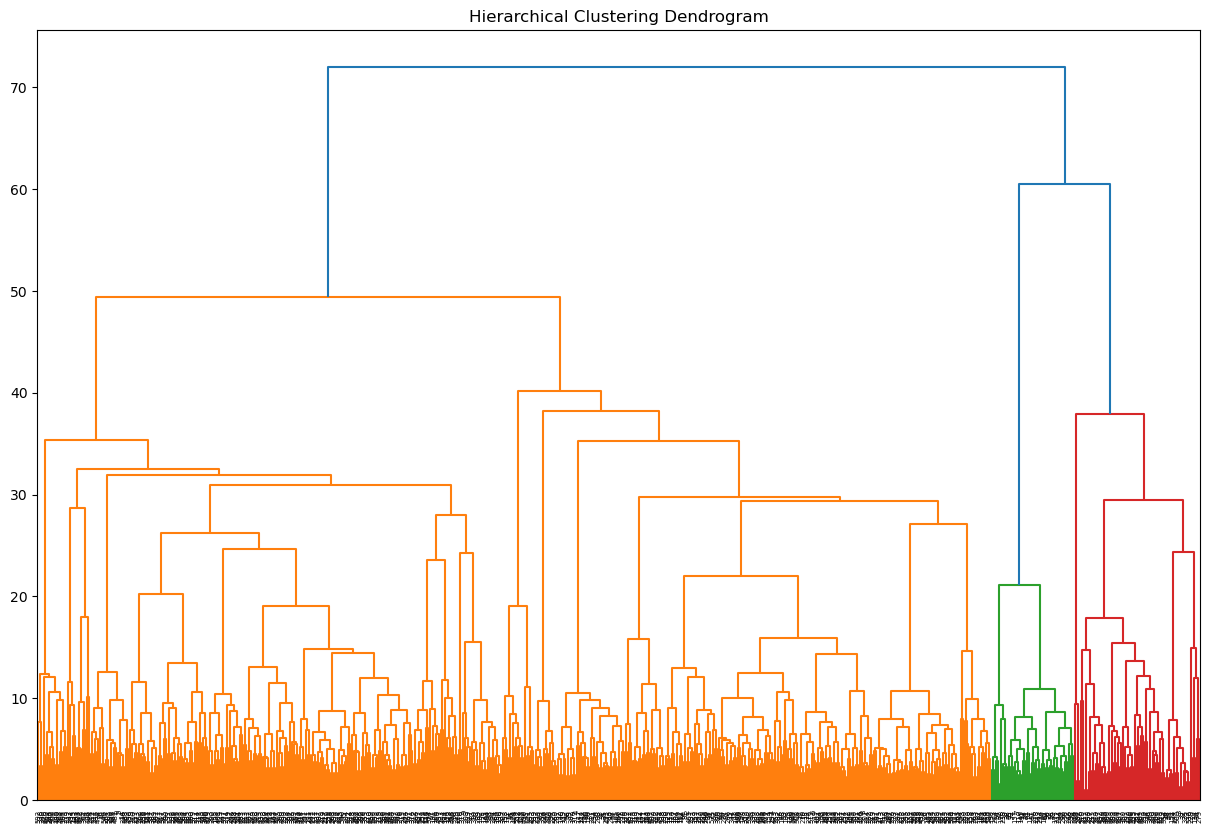
3. Perform k-means clustering on purchase behavior and basis of purchase.

4. Perform Hierarchical and k-means clustering on all variables.

Results:

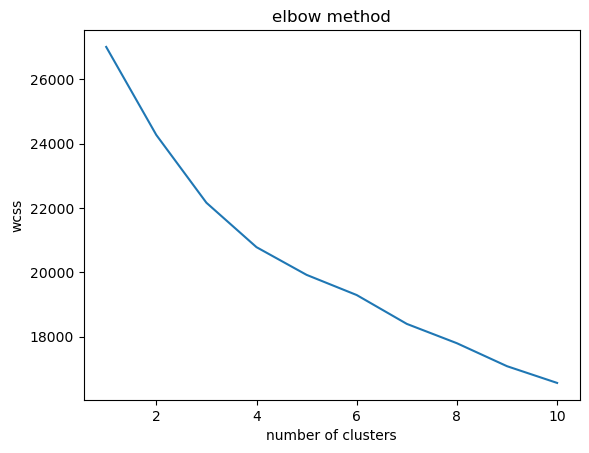
*Hierarchical Clustering Dendrogram*

We choose the longest vertical distance which is the blue line on the left. Then we draw a horizontal line across the blue line so it will intersect 2 vertical lines, suggesting 2 clusters as the optimal number.



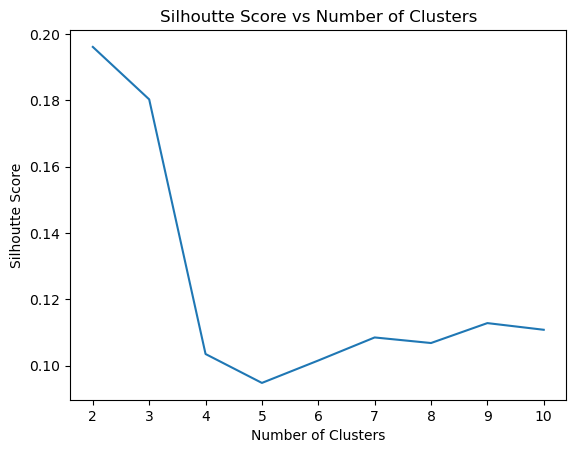
*The Elbow Method*

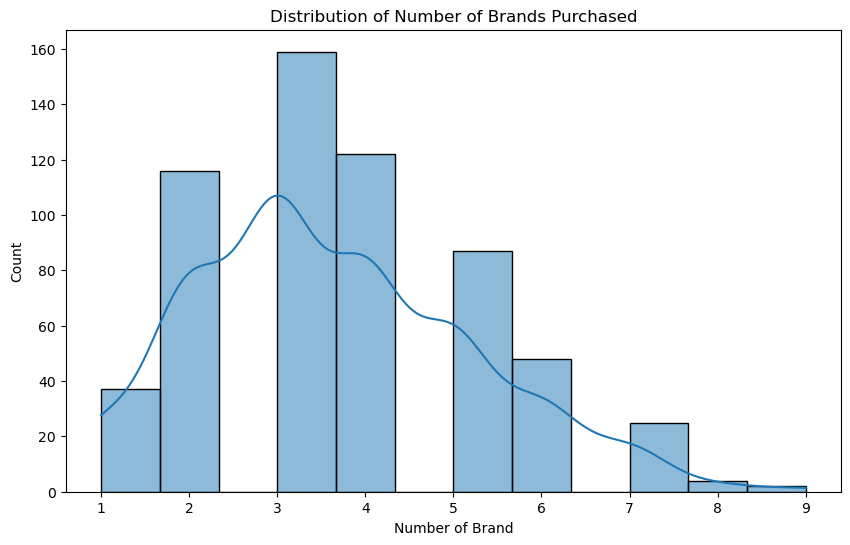
The elbow method did not give a clear indication which cluster number is the optimal number since there is no elbow in the diagram.



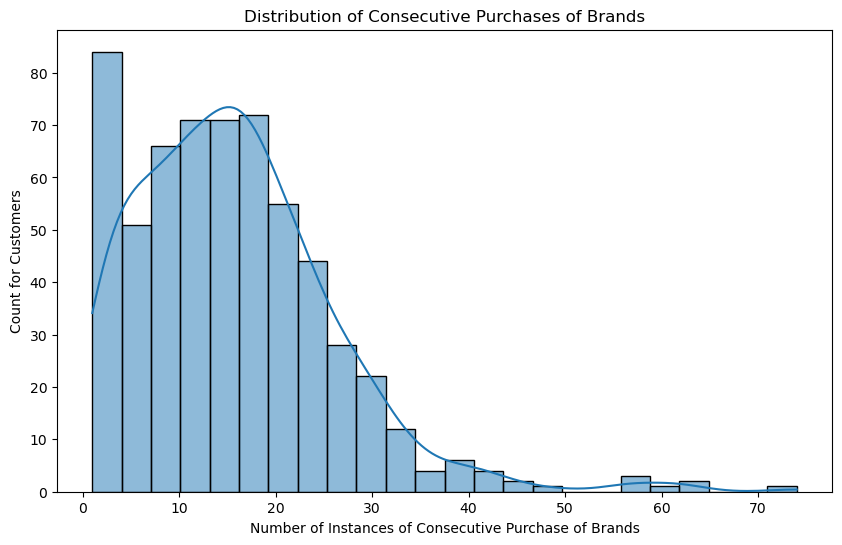
*The Silhouette Method*

For the silhouette method when we have cluster number 2, we have the highest y value. That indicates that we should use 2 as the optimal cluster number.





The majority of customers purchased 3-4 brands. They are moderately loyal compared to the 1-2 brand’s extreme loyalty group. 5-9 is not loyal.

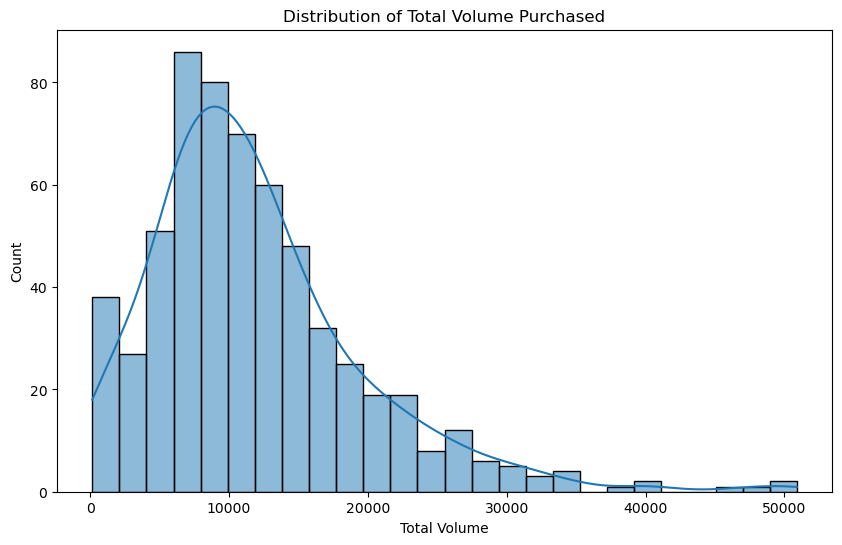


The right side shows more loyalty.

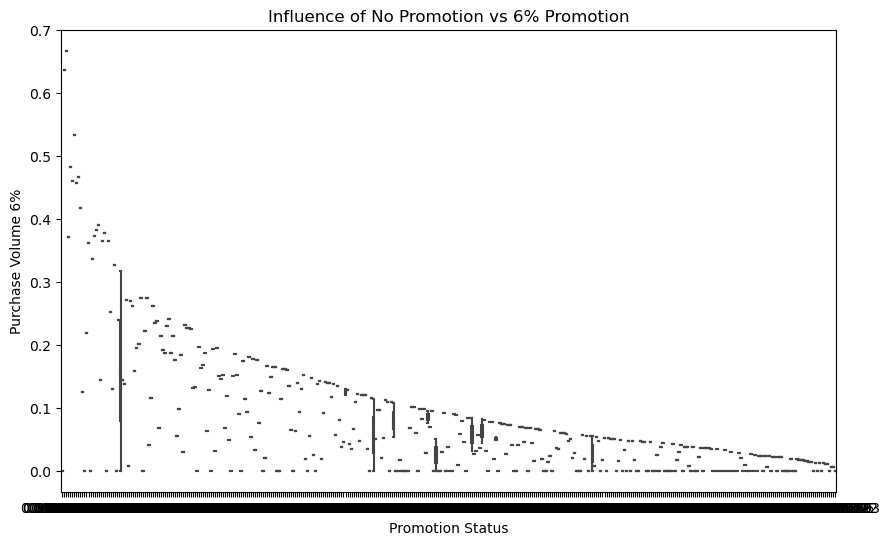
0-10 group switches brands pretty often. They are disloyal.

10-30 group made more consecutive purchases, thus switching brands less often and being more loyal.

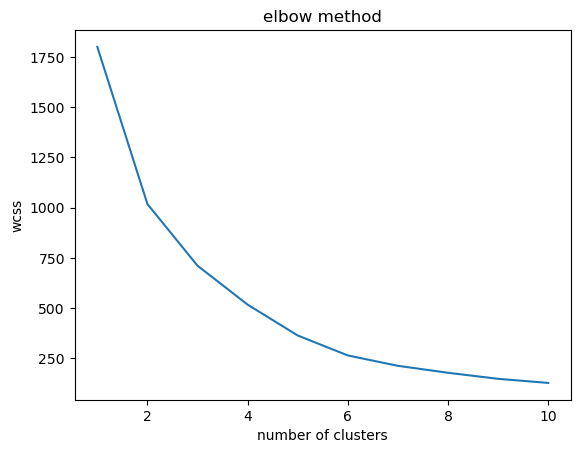
30+ are loyal customers that keep buying the same brand again and again consecutively.



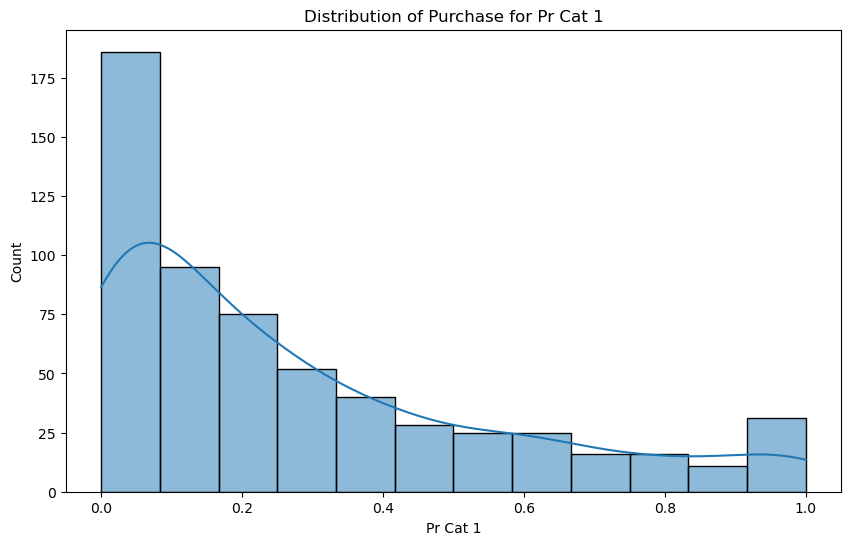
Higher volume is more loyal compared to low volume.



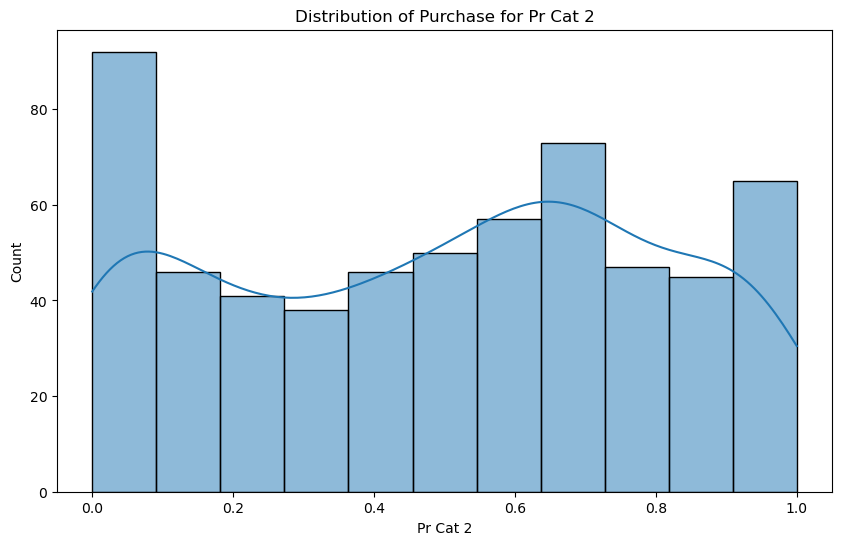
A significant number of data points clustered around lower purchase volumes. Especially for customers in the no promotion category. This might suggest a sizable portion of customers tend to make smaller purchases regardless whether there is a promotion or not.



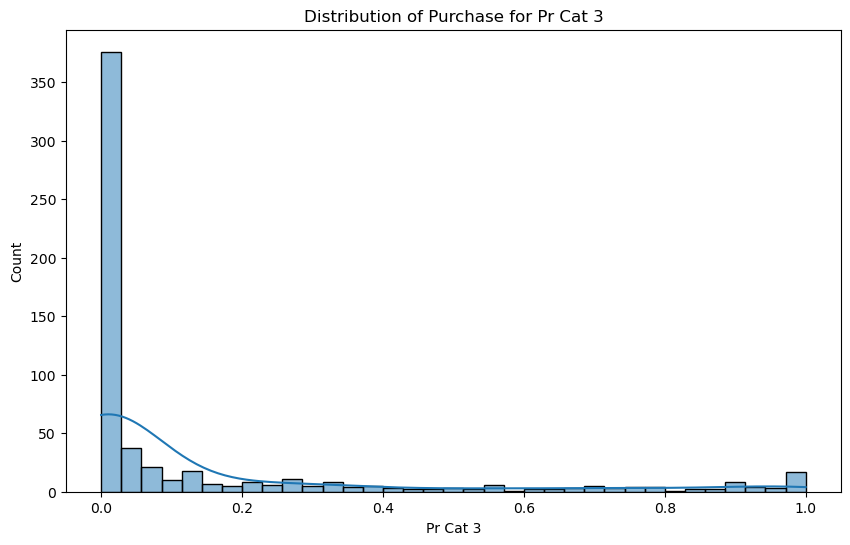
2 is the optimal cluster that we should use.



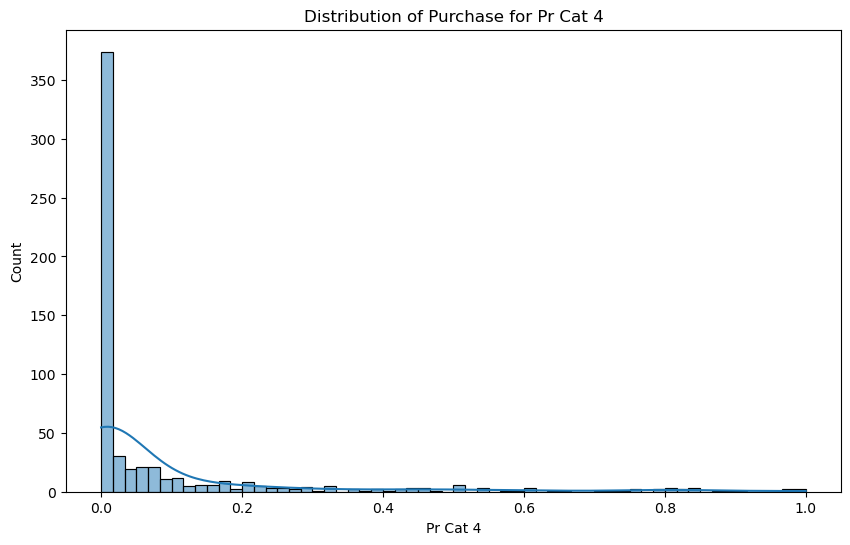
The majority of customers buying in the low-price range in this category. As price goes up we see the number of customers decrease, the only outlier is the last one (1.0) because more customers are buying in that particular price point, representing high loyalty for those customers.



This is difficult to interpret as the number of customers goes down and up in a fluctuated way. There is no clear indication without knowing what the category represents.



For this particular category, the majority of customers purchase on the lower range of the graph. This represents high customer loyalty as the price goes up in that category the number of customer count decreases dramatically which indicates low customer loyalty.



For this particular category, the majority of customers purchase on the lower range of the graph. This represents high customer loyalty as the price goes up in that category the number of customer count decreases dramatically which indicates low customer loyalty.