# Customer Segmentation of a Wholesale Company Using Clustering

*By, Alekya Bellamkonda (002541304) Varshitha Talakanti(002537790), Saida Rao Chirumamilla(002540294)*

***Abstract :*****With the growing innovation worldwide, there is an immense increase in the conflict between various industries in terms of gaining profits. With this competition going around, the business leader needs to implement a strategic approach to attract customers that could increase their sales. For this cause, machine learning comes into play, and various algorithms are applied for disclosing the hidden patterns in the data for better decision making for the future. Customer segmentation is a way in which customers with similar backgrounds from a similar group when compared to customers from different interests. Two clustering algorithms were implemented to segment the customers and, finally, compare the results of clusters obtained from the algorithms. The best clustering algorithm is selected based on its silhouette score.**

***Keywords:***
***Customer Segmentation, feature scaling, silhouette score, Machine Learning.***

## 1. INTRODUCTION

Customer Segmentation can be a powerful means to identify unsatisfied customer needs. This technique can be used by companies to outperform the competition by developing uniquely appealing products and services.

When new business keep spreading into the market it is important for the existing businesses to apply some innovative marketing tactics to stay on top always. As the businesses base on customer orientation is increasing day by day it has become challenging forth companies to provision the requirements of each and every customer, this is where Data mining serves as very important role to resolve unknown scenarios.

Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters. This segmentation can directly or indirectly impact the marketing strategy as it opens many new paths to discover like for which segment the product will be beneficial, customizing the marketing plans according to each segment, rendering discounts for a particular segment, and interpret the customer and object relationship which has been previously unexplored by the company. Customer segmentation allows companies to visualize what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them.

### Clustering

It is one of the most common exploratory data analysis procedures used to get a foreknowledge about the structure of the data. It can be defined as the task of classifying subgroups in the data such that data points in the same subgroup (cluster) are very similar, while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering is considered an unsupervised learning method since we don't have the base truth to match the output of the clustering algorithm to the correct labels to evaluate its performance.

There are a number of clustering algorithm over which like k-means, hierarchical clustering, DBSCAN clustering etc. In this paper, we have implemented two different clustering algorithms i.e. K-means clustering algorithm and Gaussian Mixture Model clustering algorithm on dataset with two features with 200records.

### K-means:

**K-means** algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping clusters where each data point belongs to **only one group**. It tries to make the intra-cluster data points as alike as possible while also keeping the clusters as diverse as possible. It consigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less distinction we have within clusters, the more uniform the data points are within the same cluster.

Algorithm:

- Define the number of clusters *K*.
- Initialize centroids by first rearranging the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing.
- Calculate the total of the squared distance between data points and all centroids.
- Designate each data point to the nearest cluster (centroid).
- Measure the centroids for the clusters by considering the average of all data points that relate to each cluster.

### Gaussian Mixture Model:

**Gaussian Mixture Models (GMMs)** assume that there are a definite number of Gaussian distributions, and each of these distributions depicts a cluster. Therefore, a Gaussian Mixture Model tends to group the data points relating to a single distribution together.

## 2. BACKGROUND

In this section we will discuss about the dataset and various other approaches previously used to perform customer segmentation and what we did differently in our project.

The dataset we considered is composed of six important product n categories**:** Fresh**,** Milk **,** Grocery , Frozen , Detergents Paper and 'Delicatessen. This is not a normalized dataset we have normalized the dataset using MinMax scaler. We have outliers in the dataset we have detected and eliminated them by using the interquantile range. We have eliminated the points which are 1.5 above and below the interquantile range which is popularly known as turkey method.

Many people adopted to do customer segmentation by using the k-means and divided the clusters but we didn't choose that as the order of the data has an impact on the final results and difficult to predict the number of clusters.

To overcome these we have considered the silhouette coefficient which helps us to built good clusters and also helps to overcome the drawbacks of k-means.

## 3. PROPOSED SYSTEM

In this project, we will interpret a dataset comprising data on various customer's annual spending amounts (reported in *monetary units*) of various product categories for internal structure. One aim of this project is to best describe the variation in the diverse types of customers that a wholesale distributor interacts with. Doing so, we would equip the wholesaler with insight into how to best fabricate their delivery service to satisfy the needs of each customer.

Firstly, we will commence exploring the data through visualizations and systems to recognize how each feature is related to the others. We will realize a statistical description of the dataset, consider the relevance of each feature, and select some sample data points from the dataset, which will track us throughout the project.

The dataset is comprised of six important product categories: **'Fresh', 'Milk'**, **'Grocery'**, **'Frozen'**, **'Detergents_Paper'**, and **'Delicatessen'**.

After data exploration, we will now select samples from the dataset. To get a better perception of the customers and how their data will change through the analysis, it would be best to choose a few sample data points and explore them in more particular.

Following that, we will perform the feature relevance task. One appealing thought to consider is if one (or more) of the six product categories is relevant for understanding customer purchasing. That is to say, is it possible to determine whether customers purchasing some amount of one type of product will necessarily purchase some proportional amount of another category of products? We can make this determination quite easily by training a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the extracted feature.

To get a better comprehension of the dataset, we will then create a scatter matrix of each of the six product features present in the data. The scatter matrix might show a correlation between that feature and another feature in the data.

Now to preprocess the data, we will create a better representation of customers by performing scaling on the data and detecting (and optionally removing) outliers. Preprocessing data is oftentimes a critical step in assuring that results obtained from the analysis are significant and meaningful.

Feature Scaling

If data is not normally distributed, especially if the mean and median vary significantly (indicating a large skew), it is most usually suitable to apply a non-linear scaling particularly for financial data. One way to achieve this scaling is by using a Box-Cox test, which calculates the best power transformation of the data that reduces skewness. A more straightforward approach that can work in most cases would be applying the natural logarithm.

Outlier Detection

Detecting outliers in the data is very important in the data preprocessing step of any analysis. The presence of outliers can often skew results that take into consideration these data points. There are many "rules of thumb" for what composes an outlier in a dataset. An *outlier step* is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal.

## Feature Transformation

We will use principal component analysis (PCA) to conclude the underlying structure of the wholesale customer data. Since using PCA on a dataset calculates the dimensions which best maximize diversity, we will find which composite combinations of features best represent customers.

### PCA

Now that the data has been scaled to a more normal distribution and has had any necessary outliers removed, we can apply PCA to the good_data to discover which dimensions about the data best maximize the variance of features involved. In addition to finding these dimensions, PCA will also report the *explained variance ratio* of each dimension — how much variance within the data is explained by that dimension alone. Note that a component (dimension) from PCA can be considered a new "feature" of the space. However, it is a composition of the original features present in the data.

While using principal component analysis, one of the chief objectives is to lessen the dimensionality of the data — in effect, decreasing the complexity of the problem. **Dimensionality reduction** comes at a cost: Fewer dimensions practiced implies more limited total variance in the data. Because of this, the *cumulative explained variance ratio* is significant for knowing how many dimensions are necessary; additionally, if a considerable amount of variance is defined by only two or three dimensions.

Now clustering comes into the picture. When the number of clusters is not known *a priori*, there is no guarantee that a given number of clusters best segments the data, since it is unclear what structure exists in the data — if any. However, we can quantify the "goodness" of clustering by calculating each data point's *silhouette coefficient*. The silhouette coefficient for a data point measures how similar it is to its designated cluster from -1 (dissimilar) to 1 (similar). Estimating the *mean* silhouette coefficient provides for a simple scoring method of a given clustering among the two clustering methods(K-means & Gaussian mixture model).

### Cluster Visualization

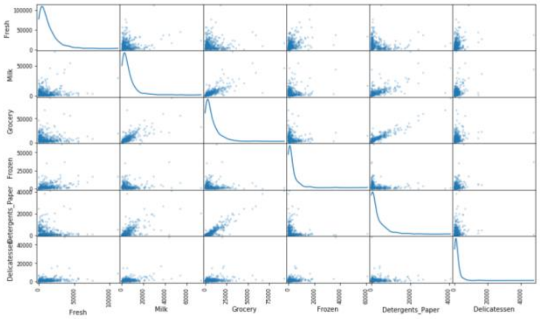
Once we have chosen the optimal number of clusters for your clustering algorithm using the scoring metric(silhouette score), we will visualize the results.

## 4. EXPERIMENTAL RESULTS

In this part, we would provide our experimental results for the dataset we used.

In the Feature relevance component, Delicatessen was the feature that attempted to predict the feature relevance. The prediction score was -2.25.

The scatter plot for better understanding of the dataset, a scatter matrix of each of the six product features present in the data is construct which looks like the image below.



We ran test to see if there is any correlations in the dataset we found out that **Grocery - Detergents_Paper** show the maximum correlation of 0.92, followed by **Grocery - Milk** of 0.73.

For feature scaling, After applying a natural logarithm scaling to the data, the distribution of each feature should appear much more regular. For any pairs of features, we may have identified earlier as being correlated, we have observed that correlation is still present.

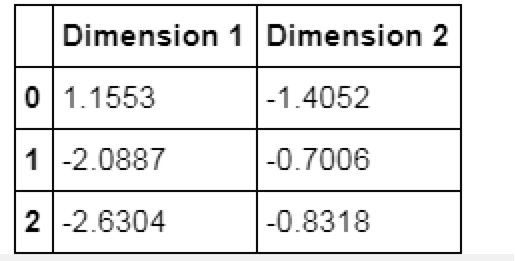
We tried to identify the outliers and successfully found out Data points as outliers for more than 1 feature:

- 154: Delicatessen, Milk and Grocery.
- 128: Delicatessen and Fresh.
- 75: Detergents Paper and Grocery.
- 66: Delicatessen and Fresh
- 65: Frozen and Fresh

Few observations while performing PCA were:

- Variance explained in total by first and second principal component is 70.68%
- Variance explained in total by first four principal components is 93.11%

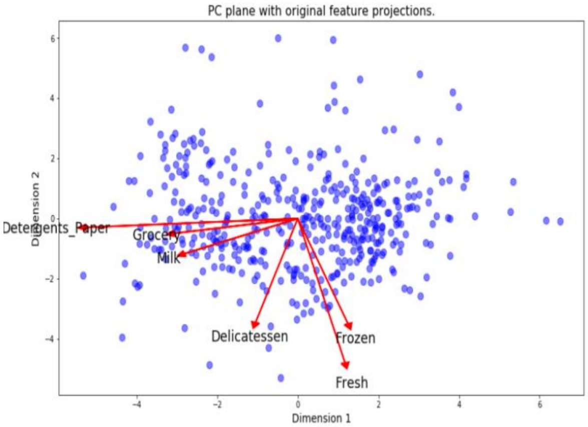
The log-transformed sample data has changed after having a PCA transformation applied to it, using only two dimensions. We have observed how the values for the first two dimensions remain unchanged when compared to a PCA transformation in six dimensions.



| | Dimension 1 | Dimension 2 |
|---|---|---|
| 0 | 1.1553 | -1.4052 |
| 1 | -2.0887 | -0.7006 |
| 2 | -2.6304 | -0.8318 |

A bi-plot is a scatter plot where each data point is represented by its scores along the principal components. A bi-

plot can help us interpret the reduced dimensions of the data, and discover relationships between the principal components and original features.



the original feature projections (in red), it is easier to interpret the relative position of each data point in the scatter plot. For instance, a point the lower right corner of the figure will likely correspond to a customer that spends a lot on 'Milk', 'Grocery' and 'Detergents Paper', but not so much on the other product categories.

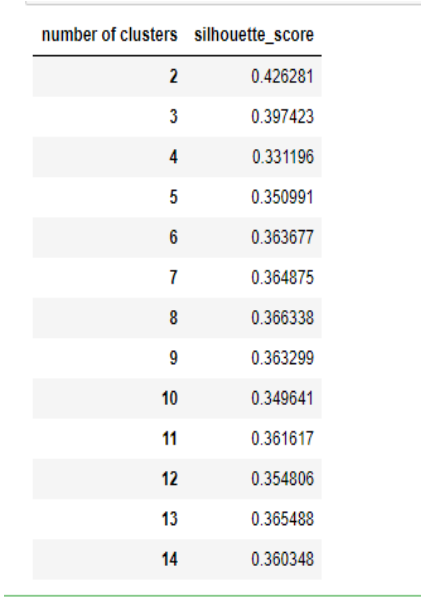
To perform clustering we have used:

1. K-means

2. Gaussian Mixture

After performing the k-means the

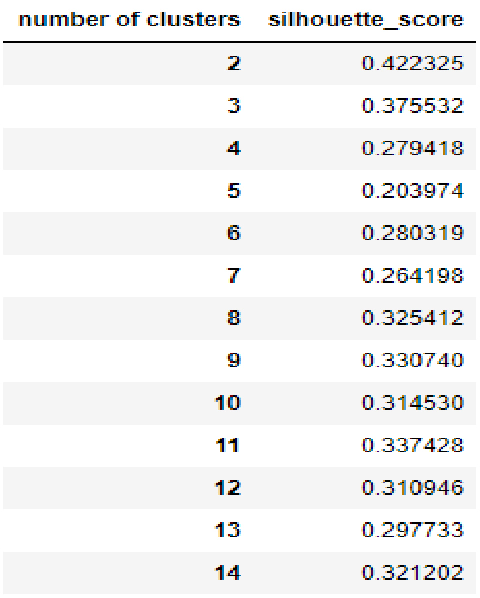
Table shows the results



| number of clusters | silhouette_score |
|---|---|
| 2 | 0.426281 |
| 3 | 0.397423 |
| 4 | 0.331196 |
| 5 | 0.350991 |
| 6 | 0.363677 |
| 7 | 0.364875 |
| 8 | 0.366338 |
| 9 | 0.363299 |
| 10 | 0.349641 |
| 11 | 0.361617 |
| 12 | 0.354806 |
| 13 | 0.365488 |
| 14 | 0.360348 |

After performing the Gaussian Mixture the

Table shows the results



| number of clusters | silhouette_score |
|---|---|
| 2 | 0.422325 |
| 3 | 0.375532 |
| 4 | 0.279418 |
| 5 | 0.203974 |
| 6 | 0.280319 |
| 7 | 0.264198 |
| 8 | 0.325412 |
| 9 | 0.330740 |
| 10 | 0.314530 |
| 11 | 0.337428 |
| 12 | 0.310946 |
| 13 | 0.297733 |
| 14 | 0.321202 |

We Considered Silhouette Coefficient as the metric to quantify our clusters.

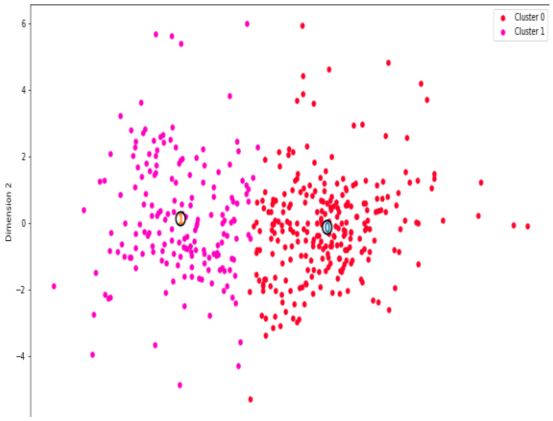
For Gaussian Mixture the best result was for 2 clusters

For K-means the best result was for 2 clusters

| MODELS | NUMBER OF CLUSTERS | SILHOUETTE SCORE |
|---|---|---|
| K-MEANS | 2 | 0.426281 |
| GAUSSIAN MIXTURE | 2 | 0.422325 |

Finally, after performing all the above tasks, we were successfully able to separate 2 clusters which represents two types of customers shopping behaviour.

From the fig below, the red group represents one cluster and the other group represents one cluster.



## 5. REFERENCES

1. Jiwai Han , Micheline Kamber , Data Mining and Techniques, 2nd Edition , Simon FraserUniversity.
2. Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi.(2012).The Survey of data

mining Applications & Future Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June2012.

3. Sensor Network and Automation (IMSNA) , 2013 2nd International Symposium on DOI : 10.1109/IMSNA . 2013.6743300. IEEE Publications.
4. Publication Year: 2012 , Page(s): 295 – 299
5. wikipedia.org/wiki/Supervised_learning
6. Bhavika Tekwani "Amazon review classification" from GitHub