# Customer Segmentation using K-Means Algorithm

**Udaybhan Rathore**

ROLL NUMBER- 19328

Deparment- DSE

Course – DSE309
EMAIL-ID -

udaybhan19@iiserb.ac.in

## *Abstract—*

**The technique of separating consumers into distinct categories depending on specific characteristics is known as customer segmentation. But why segment customers? Well, every business is based on manufacturing specific products, and for every product developed there is a defined group of customers. Targeting those defined customers will assure guaranteed sales to the organizations. like all the other essential processes used in business strategies, customer segmentation is also an important step. This integration helps organizations to communicate with a specific group of customers based on their current interests and needs. This project applies customer segmentation to the customer data from a company ulabox and derives conclusions and data driven ideas based on it.**

**Keywords—clustering, k-means, un-supervised learning, segmentation, cluster.**

## *INTRODUCTION —*

The sole purpose of every firm is to make profit. More so, it is much easier for industries to make profit when they understand their customers' needs which usually leads to their satisfaction. This strategy which fulfils the expectations of customers is called customer satisfaction. It requires the understanding of consumer behaviour, profit or loss, as well as sales, of a firm. Over the years, many technologies have been developed to make this happen. Cluster analysis, also called segmentation analysis is a technique that explores data in order to find groups, patterns and structures.

The aim of this project is to indicate what clusters of consumers emerge based all product categories using the K-means clustering method which is one of the most widely used unsupervised clustering methods. In the next section, the data background is discussed. Furthermore, the methodology and results are discussed, and finally, a conclusion is eventually made based on the research findings.

## *DATA BACKGROUND —*

The data used in this project was sourced from GitHub (The Ulabox Online Supermarket Dataset 2017, accessed from https://www.github.com/ulabox/datasets ). It includes data about Ulabox- an accomplished and well-known online supermarket in Spain which earns about one million euros per month and has achieved a customer satisfaction rate of over 90%. The dataset consists of 30,000 rows and 14 columns. The rows include different samples of about 10,000 customers who have made purchase severally on the website.

Below, we describe the columns in detail.

• **Customer:** The unique anonymized identifier of the customer.

• **Order:** The order ID of the customer

• **Total items:** The total number of goods purchased in the order.

• **Discount:** The total percentage of discount. For example, a customer is assigned a discount percentage of 10 if they save €10 in an order of €100. This means he must pay €90.

• **Weekday:** The day of the week the order was purchased, from Monday (1) to Sunday (7).

• **Hour:** The hour of purchase, i.e. from 00 to 23.

• **Food%:** The percent amount of money spent on non-perishable foods like snacks, pasta, spice and canned food.

• **Fresh%:** The percentage amount spent on fresh and frozen food like fruits, frozen pizza, chicken, etc.

• **Drinks%:** The percentage amount spent on beverages of all kinds like water, wine, milk

• **Home%:** The percentage amount spent on home products such as toilet paper, home appliances.

• **Beauty%:** The percentage amount spent on body cleaning items and makeup, like shaving cream, shampoo and cosmetic.

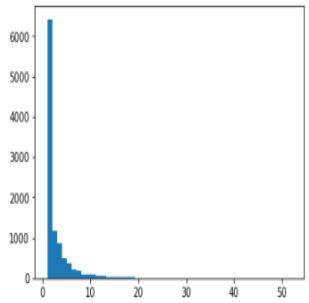
• **Health%:** The percentage amount spent on medicinal solutions which do not require medical prescriptions to be sold. Examples include diet pills, condoms, etc.

• **Baby%:** The percentage amount spent on baby items such as baby food and diapers.

• **Pets%:** The percentage amount spent on pet related items like sanitary sand, food, etc.

# *METHODOLGY —*

## A. *Data Clearning and data preparation:*

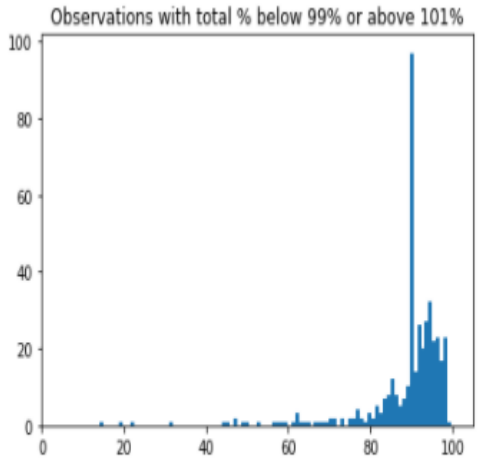


```
There are 10239 unique customers in the dataset of 30000 orders


1    4479
2    1944
3    1180
4     862
5     492
Name: order, dtype: int64
```

The number of orders per customer has an exponential distribution, with about 44% of customers having just one order in the dataset. When I add a column summing all the product categories, the summary of the numeric data raises two issues:

1. We can see that there are some samples in the dataset that do not have total=100%.
2. There are negative discounts and discounts of 100%




```
There are 416 potential outliers in the dataset of 30000 samples
```

Since there are only 416 potential outliers in Total%. Outliers can skew the dataset. They should be removed to have a normalized dataset for k-means clustering. And after visualizing the data we are replacing Food%, Fresh% etc. of customers with the average among all the orders.

## B. *Principal component analysis:*

As the data contain more than two variables, we need to reduce the dimensionality in order to plot a scatter plot. This can be done using principal component analysis (PCA) algorithm. The combination of PCA (Principal Component Analysis) and K-Means is always the first choice for most companies to cluster their data. This is because the combination is quick, simple, and can scale to large datasets. The main idea behind PCA is to lower the dimension of data features (data columns) to optimize the clustering result to avoid the 'curse of dimensionality' (too many columns will have data noises).

.

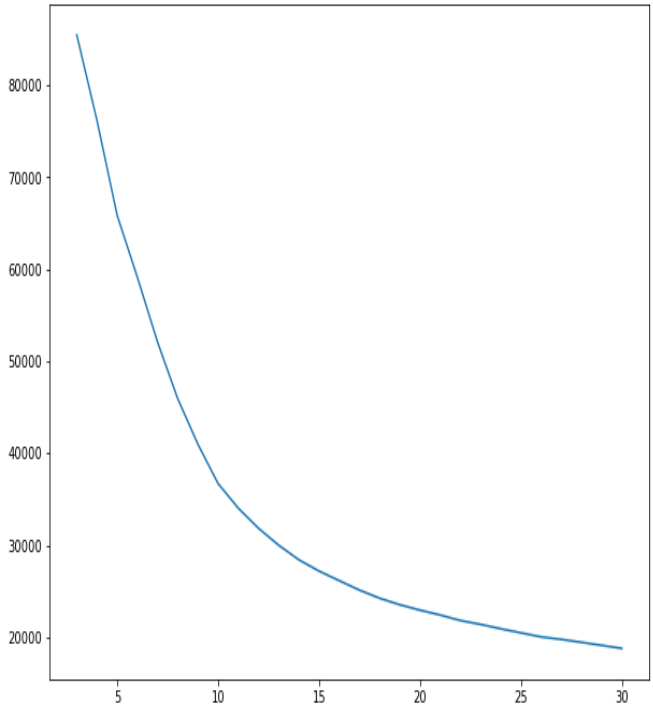
## C. *k-means clustering:*

k-means clustering model is one the vastly used model for clustering. Being unsupervised learning algorithm, it has many applications. The K-Means algorithm is used to measure the distance between each data point and the group center (the closer the distance from the point to the cluster center within a cluster the better). The critical part is to identify the K (number of groups) before the clustering

calculation. It is impossible for us to tell how many possible customer groups of Ulabox so we need to identify the optimized K number using different methods. We will use Elbow method and Average silhouette method.

### *ELBOW METHOD:*

**Elbow** method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick *k* at the spot where SSE starts to flatten out and forming an elbow. We'll use the dataset and evaluate SSE for different values of *k* and see where the curve might form an elbow and flatten out.




The graph above shows that k= 10 is not a bad choice. Sometimes it's still hard to figure out a good number of clusters to use because the curve is monotonically decreasing and may not show any elbow or has an obvious point where the curve starts flattening out.

### *AVERAGE SILHOUETTE METHOD:*

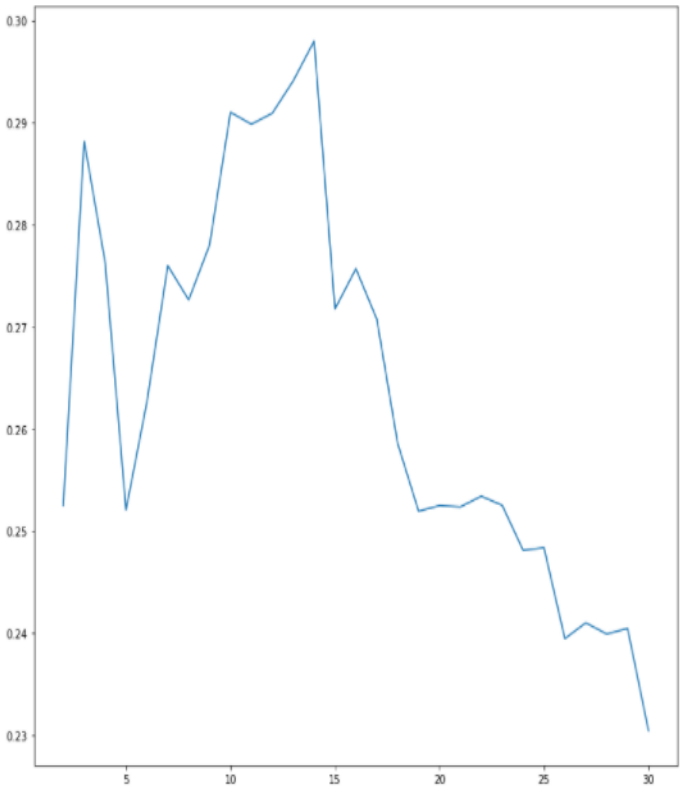
**Silhouette analysis** can be used to determine the degree of separation between clusters. For each sample:

- Compute the average distance from all data points in the same cluster (ai).
- Compute the average distance from all data points in the closest cluster (bi).
- Compute the coefficient:

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

- The coefficient can take values in the interval [-1, 1].
- If it is 0 –> the sample is very close to the neighbouring clusters.
- If it is 1 –> the sample is far away from the neighbouring clusters.
- If it is -1 –> the sample is assigned to the wrong clusters.

Therefore, we want the coefficients to be as big as possible and close to 1 to have a good cluster. We'll use here dataset again and find out from the below figure that around 10 number of clusters its shows highest value of average silhouette.




## *RESULTS AND DISCUSSION—*

In the figure 5 the number of product categories represented in individual orders is distributed fairly normally. As show in the figure product category drinks (4) show the highest number of orders placed.

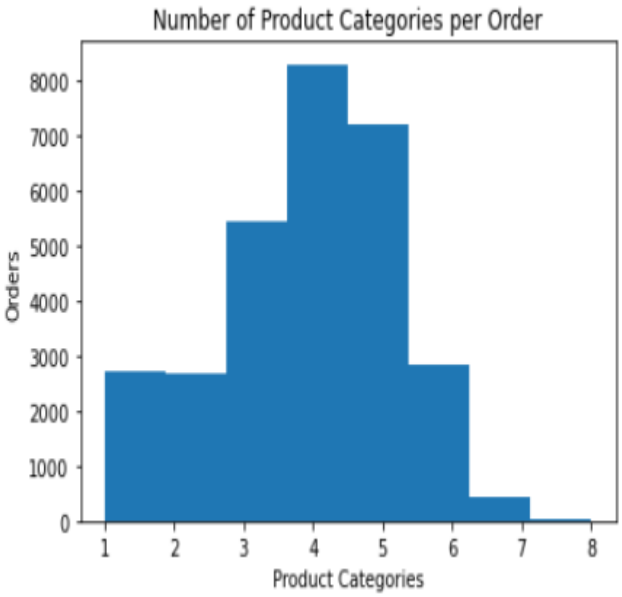



Figure 5

From the figure 6 we can see that the most orders occur on Monday (1) and Sunday (7) while orders are much lower on Friday and Saturday. The most popular time for ordering is between 10-11pm, followed by 10am-noon.

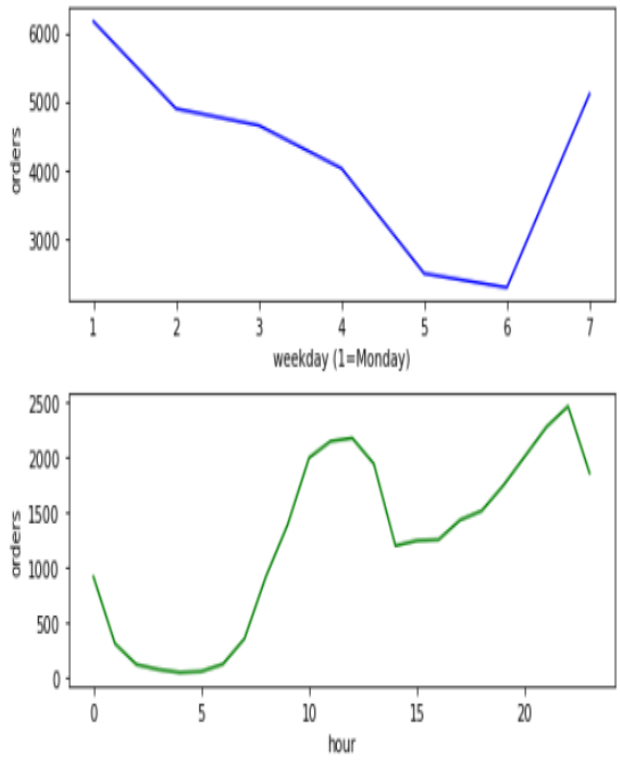



Figure 6

From the figure 7, we can see that this heatmap shows that most of the late evening orders are placed on Sunday. Overall, Sunday evening and Monday midday are the most common times for orders to be placed.

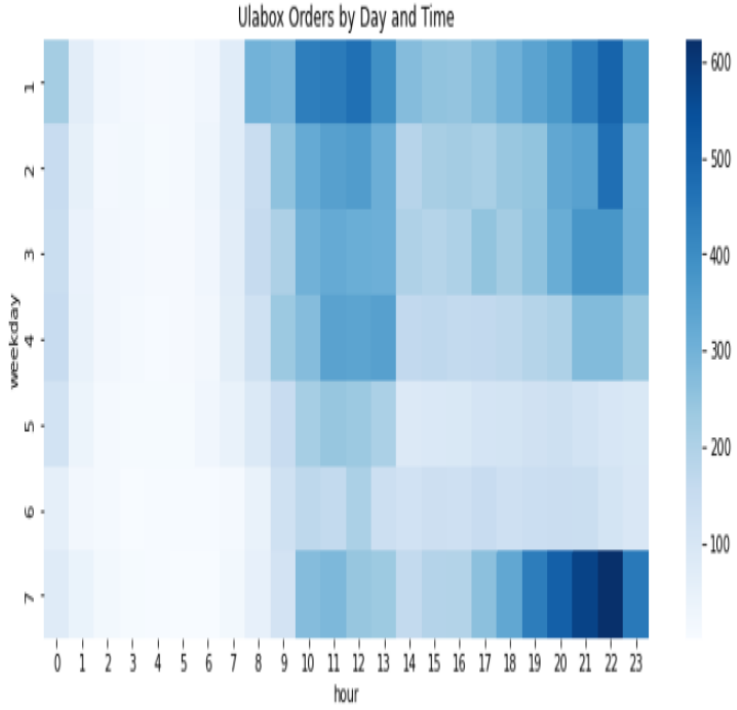



Figure 7

### *CLUSTER INTERPRETATION:*

From the Elbow and Average Silhouette Method we figure it out that there could be around 10 clusters in the data.

We then with 10 clusters we compared with some metrics. A common approach is to compare the average value for each column and cluster. We form the data frame's by sorting the cluster-data with each label and interpretate the data frames and then we see the statistic of each data frames. The statistical data frames illustrate the average value across clusters (labels). Starting from the main segments (labels that range from 0 to 9), some interesting patterns are found. It is obvious that customers are segmented by preferred categories (average percentage of the order value).

After examine we statistical data we identify that Cluster 0, which accounts for about 66.96800% of his total spent in the shop, buy most in drinks therefore we classified that cluster to drink lovers. Similarly, in different data frame of different cluster we found some interesting patters and we label that cluster we that patters such as pets' lovers, new parent, loyal, home decorators etc. I have completely provided the information about how I have labelled the cluster data in DSE_19328_Udaybhan_Project.ipynb.

## *Conclusion —*

Kmeans clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of kmeans is to group data points into distinct non-overlapping.

| class | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| beauty_concious | 425.0 | 1.101176 | 0.452064 | 1.0 | 1.0 | 1.0 | 1.0 | 5.0 |
| drink_lovers | 1294.0 | 2.442813 | 1.740129 | 1.0 | 1.0 | 2.0 | 3.0 | 14.0 |
| fresh_regulars | 1877.0 | 3.010655 | 2.021445 | 1.0 | 1.0 | 2.0 | 4.0 | 10.0 |
| grocery_regular | 3091.0 | 2.642834 | 1.794426 | 1.0 | 1.0 | 2.0 | 4.0 | 9.0 |
| grocery_shoppers | 389.0 | 1.213368 | 0.530732 | 1.0 | 1.0 | 1.0 | 1.0 | 5.0 |
| health_concious | 125.0 | 1.376000 | 0.867551 | 1.0 | 1.0 | 1.0 | 1.0 | 6.0 |
| home_decorators | 1017.0 | 1.637168 | 1.077287 | 1.0 | 1.0 | 1.0 | 2.0 | 9.0 |
| new_parents | 1354.0 | 1.904727 | 1.322026 | 1.0 | 1.0 | 1.0 | 2.0 | 11.0 |
| pet_lovers | 74.0 | 1.729730 | 1.126241 | 1.0 | 1.0 | 1.0 | 2.0 | 6.0 |
| royals | 593.0 | 12.706577 | 4.704610 | 7.0 | 10.0 | 12.0 | 15.0 | 52.0 |

Figure8

In figure 8, we have grouped the final data into number of orders. And we conclude that in royal's class the minimum number of orders itself is 7 therefore they are the most valuable customers. Regulars, drink lovers and new parents have pretty impressive order counts as well. Similarly, we can perform different operation to find out the information of each customer and provide them different offers to gain trust of each customer.

But it should be noted that the clustering is based on the order level but not the customer level. For interpreting the real customer segments, it is better to retrieve new variables, such as the SKU quantity and absolute shopping value instead of value percentage% for each customer. This is because the % cannot be added up for each customer's order. However, the process of clustering is the same. The most important part is how to turn insights into an implementable marketing plan. For each cluster, a clear combination of discount + product category + key products + promotion time + promotion message should be identified.

Here, we conclude over report.

## *References*


1) Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics
2) I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.
3) T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence
4) https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
5) https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
6) https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
7) https://github.com/ulabox/datasets
8) https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html