CUSTOMER CLUSTERING WITH MACHINE LEARNING

Capstone Project

Ömer Faruk Kara

ISTANBUL, 2021

MEF UNIVERSITY

CUSTOMER CLUSTERING WITH MACHINE LEARNING

Capstone Project

Ömer Faruk Kara

Advisor: Asst. Prof. Dr. Tuna Çakar

ISTANBUL, 2021

MEF UNIVERSITY

Name of the project: Customer Clustering with Machine Learning
Name/Last Name of the Student: Ömer Faruk Kara
Date of Project Report Submission: 27/01/2021

I hereby state that the graduation project prepared by Your Name (Title Format) has been completed under my supervision. I accept this work as a "Graduation Project".

27/01/2021
Asst. Prof. Dr. Tuna Çakar

I hereby state that I have examined this graduation project by Your Name (Title Format) which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

Director
of
Information Technologies
Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature /Date
1. Asst. Prof. Dr. Tuna Çakar	……..
2. Director's Name	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Ömer Faruk Kara	27/01/2021	

EXECUTIVE SUMMARY

CUSTOMER CLUSTERING WITH MACHINE LEARNING

Ömer Faruk Kara

Advisor: Asst. Prof. Dr. Tuna Çakar

JANUARY 2021, 24 pages

When analyzing a company that sells in very different product ranges, you are likely to encounter different types of customers. Grouping customers correctly can set standard actions while serving them. Standardization of marketing processes leads to speed and they are easy to improve. While making this classification, the KMeans algorithm was used in Machine Learning. Inertia and Silhouette Points values were used to find the most suitable cluster number. Principal Components Analysis (PCA) was used to show customers with multidimensional features in 2 dimensions.

Key Words: Machine Learning, KMeans Algorithm, Principal Component Analysis (PCA)

ÖZET

MAKİNA ÖĞRENMESİ İLE MÜŞTERİ KÜMELEMESİ

Ömer Faruk Kara

Proje Danışmanı: Dr. Öğr. Üyesi Tuna Çakar

OCAK 2021, 24 Sayfa

Eğer çok farklı ürün gamlarında satış yapan bir şirketi analiz ediyorsanız, birbirinden farklı tipte müşteriler ile karşılaşmanız olasıdır. Eğer bu müşterileri doğru sınıflandırabilirseniz. Onlara hizmet verirken standart aksiyonlar belirleyebilir. Bu sayede hız kazanabilir ve standart aksiyonlar belirlendiğinden geliştirilebilir. Bu sınıflandırmayı yaparken Makine Öğrenmesinde Kmeans algoritması kullanıldı. En uygun sınıf sayısını bulurken Atalet ve Silüet Puanı değerlerinden faydalanıldı. Çok boyutlu özelliği olan müşterileri 2 boyutlu düzlemde gösterebilmek için de Temel Bileşenler Analizi (PCA)'nden faydalanıldı.

Anahtar Kelimeler: Makina Öğrenmesi, kmeans algoritmasi, Temel Bileşenler Analizi, PCA, Atalet Puanı, Silüet Puanı

TABLE OF CONTENTS

Academic Honesty Pledge v
EXECUTIVE SUMMARY vi
ÖZET vii
TABLE OF CONTENTS viii
TABLE OF FIGURES ix
1. INTRODUCTION 1
2. LITERATURE REVIEW ON CLUSTERING 2
2.1. What is the KMeans Algorithm? 2
2.2. What is the Silhouette Analysis? 3
2.3. What is the Elbow Method? 3
2.4. What is the Davies-Bouldin index? 4
2.5. What is Principal Component Analysis (PCA)? 4
3. THE PROJECT 5
3.1. Company Characteristics and Main Goal 5
3.2. Review of the dataset we used 5
3.3. Setting the environment 6
3.4. Python Code 7
3.5. Analyze Customer Distributions 8
3.6. Determine the cluster number 12
3.7. Principal Component Analysis 16
3.8. Analyzing common features of clusters 16
3.9. Comparison with the customer classification used in the company 17
4. RESULT 22
APPENDIX 23
REFERENCES 24

TABLE OF FIGURES

Figure 1: Example Elbow Score Graph 3
Figure 2: Entity relation diagram 5
Figure 3: Count of Customers by Customer Group 6
Figure 4: Distribution for 2 Clusters 8
Figure 5: Distribution for 5 Clusters 8
Figure 6: Distribution for 10 Clusters 9
Figure 7: Distribution for 20 Clusters 9
Figure 8: Distribution for 30 Clusters 10
Figure 9: Distribution for 50 Clusters 10
Figure 10: Distribution for 75 Clusters 11
Figure 11: Distribution for 100 Clusters 11
Figure 12: Silhouette score with cosine distance metric 12
Figure 13: Silhouette score with Euclidean distance metric 12
Figure 14: Silhouette score with l1 distance metric 13
Figure 15: Silhouette score with l2 distance metric 13
Figure 16: Silhouette score with manhattan distance metric 14
Figure 17: Silhouette score with correlation distance metric 14
Figure 18: Davies Bouldin score 15
Figure 19: Inertia - Elbow Graph 15
Figure 20: PCA for 5 Clusters 16
Figure 21: Customer Group Distribution for Cluster 0 17
Figure 22: Customer Group Distribution for Cluster 1 18
Figure 23: Customer Group Distribution for Cluster 2 19
Figure 24: Customer Group Distribution for Cluster 3 20
Figure 25: Customer Group Distribution for Cluster 4 20

1. INTRODUCTION

In today's competitive world, understanding customer behavior and categorizing customers based on their demographics and purchasing behavior is essential. It allows marketers to better adapt their marketing efforts to various audience subsets in terms of promotion, marketing, and product development strategies. We call Customer Segmentation when a market is divided into separate customer groups that share similar characteristics. Using Customer Segmentation to identify unsatisfied customer needs can be a powerful tool. Companies using this data in the world and our country can gain an advantage in the competition by developing uniquely attractive products and services. **[1]**These homogeneous groups formed in Cluster Analysis (using a mathematical model to discover similar customer groups based on finding the smallest variations among customers within each group), one of the methods used for customer segmentation, is known as "customer archetypes" or "personas". **[2]**

The goal of Cluster Analysis is to accurately segment customers to achieve more effective customer marketing through customization. The frequently used method of cluster analysis is an algorithm known as k-mean cluster analysis. Customer samples naturally emerge from the customers. **[3]**

2. LITERATURE REVIEW ON CLUSTERING

2.1. What is the KMeans Algorithm?

It is used to grouping the elements in the dataset. Each element should be in only one group. It tries to make the center of the clusters far from each other. Cluster's centroid is the arithmetic mean of all the data points that belong to that cluster.

You have to define, how many clusters you want, to use the KMeans algorithm. It begins with a random centroid and then tries to find the best center for each cluster. And assign data points to the closest centroid. **[4]**

The algorithm starts with the process of splitting n elements into k sets. After the K centroid is determined, the distances of the elements to the k centroids are calculated and the closest points to the k centroid form a cluster. Cluster elements are averaged and centroids are determined again. If the centroid has changed, it is found which centroid the points belong to according to their distance to the center, and this process continues until the centroid becomes stable.

2.2. What is the Silhouette Analysis?

It is used to study the distance between clusters. It shows a measure of how close each point in a cluster is to points in neighboring clusters and thus provides a way to visually evaluate parameters such as cluster number. **[5]**

With this value, we find out how different sets are from each other. We confirm that the closer the value between +1 and -1 is to +1, the better the number of clusters. This value is found by the equation created between the average distance between the clusters and between them.

2.3. What is the Elbow Method?

The Elbow method is a common way of deciding the appropriate number of clusters. It is a visual way in which you can see a rapid decrease before the flat line. It is calculated as the mean square distance between each instance and its centroid. **[6]**

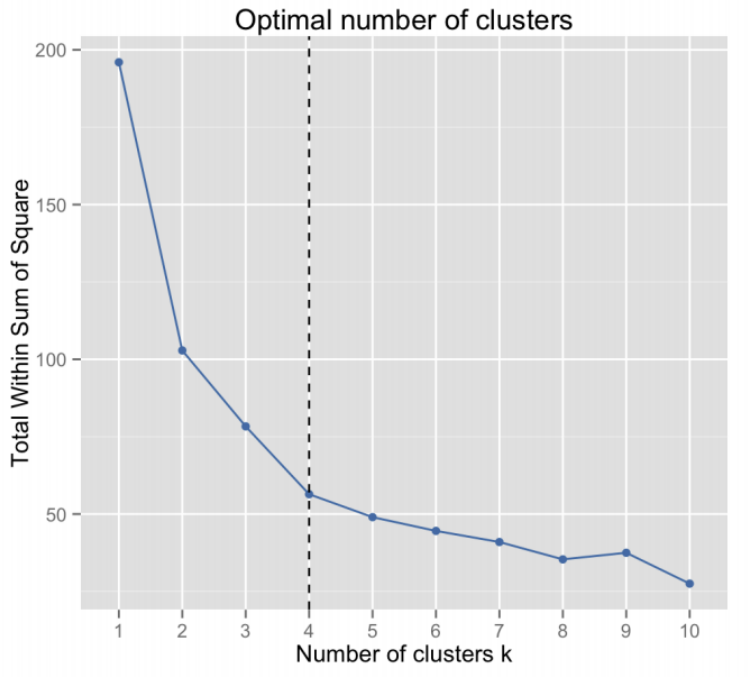

Figure 1: Example Elbow Score Graph

2.4. What is the Davies-Bouldin index?

Davies-Bouldin index is a validation metric that is often used to evaluate the optimal number of clusters to use. It is defined as a ratio between the cluster scatter and the cluster's separation and a lower value will mean that the clustering is better. **[7]**

2.5. What is Principal Component Analysis (PCA)?

Principal Component Analysis is a method to keep the data set with the highest variance in high dimensional data but to provide dimension reduction while doing this. By finding general features in multi-dimensional data, it enables to reduce of the number of dimensions and compress the data. Certain features will be lost with size reduction, but the intention is that these lost traits contain little information about the population. This method combines highly correlated variables to create a smaller set of artificial variables called "principal components" that make up the most variation in the data.

PCA is a very effective method for revealing the necessary information in the data. The basic logic behind PCA is to show multidimensional data with fewer variables by capturing the basic features in the data. **[8]**

3. THE PROJECT

3.1. Company Characteristics and Main Goal

Our example company is an international company providing hardware and fitting systems and electronic access control systems. It has customers in the furniture industry, dealers, joiners, cabinet makers, as well as architects, planners, builders.

Our purpose is to group these customers based on their purchases by the product group. Try to find customer clusters that are purchased in the same product groups.

3.2. Review of the dataset we used

It is based on the company's sales data for the last 12 months. Columns that we used are;

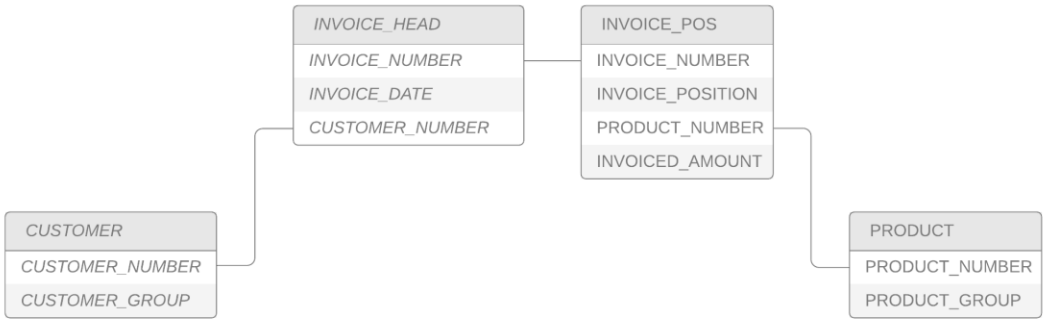

Figure 2: Entity relation diagram

As shown in Figure 2, we will use four tables to prepare the analysis data set. Our main tables are "invoicehead" and "invoicepos". They contain invoice data of the company. They are connected to the "customer master data table" to get sales group information and the "product master table" to get product group information.

There are 4726 unique customers with 269749 lines of invoice data.

After summing by 298 product groups, the result dataset has 4726 rows and 298 columns.

Most values appear to be blank, as customers do not purchase from most product groups. When compared to the total row, our data set has only 2.20847% occupancy.

Although this seems to be a very low rate, when we analyzed the data, 137 product groups were sold to less than 25 customers, 32 product groups to only 1 customer) While we can extract these product groups to increase the occupancy rate, it is beneficial not to exclude them from the data set as these product groups provide us with information about customer characteristics since we examine purchasing behavior.

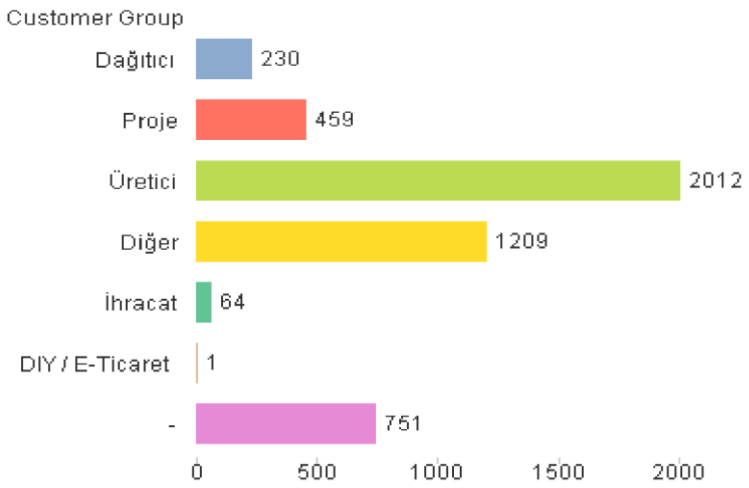

Figure 3: Count of Customers by Customer Group

As seen in Figure 3, the customers divided into six sales groups. The leading group contains 42% of the total. 15% of the total customers have no customer group selected.

These customers were divided into groups by the sales department for business strategy reasons. You can see the figure for this grouping in the table above. As you can see, 751 of 4726 customers have not been identified as members of any group for now.

After finishing our analysis we will compare our clustering with the current grouping which is actively used.

3.3. Setting the environment

- I am using python 3.8 with
- libraries
  - pandas; for reading the CSV file
  - sklearn; for KMeans Clustering
  - matplotlib; for visualization of our result.

3.4. Python Code

Before starting clustering, we first need to prepare our data set to give the best result.

Respectively;

- Replace empty values in our data with 0
- Since we read the data from a text file; Convert all values to numbers,
- Normalize the features with *StandardScaler* and *normalize* functions.

Since we don't know how many different groups we should have, we first need to find the optimal cluster number. To do this, we will run the sklearn KMeans clustering algorithm from 2 to 100 clusters one by one.

We will collect Davies-Bouldin, inertia, and silhouette scores for each cluster number and decide which one is best for our data set. find the most accurate number of clusters for our data set with silhouette elbow and davies measurements for each cluster number.

Euclidean, l2, l1, manhattan, correlation, cosine metric used to calculate the silhouette scores.

Principal Component Analysis graphs and Bar Charts were used to visualize the results.

3.5. Analyze Customer Distributions

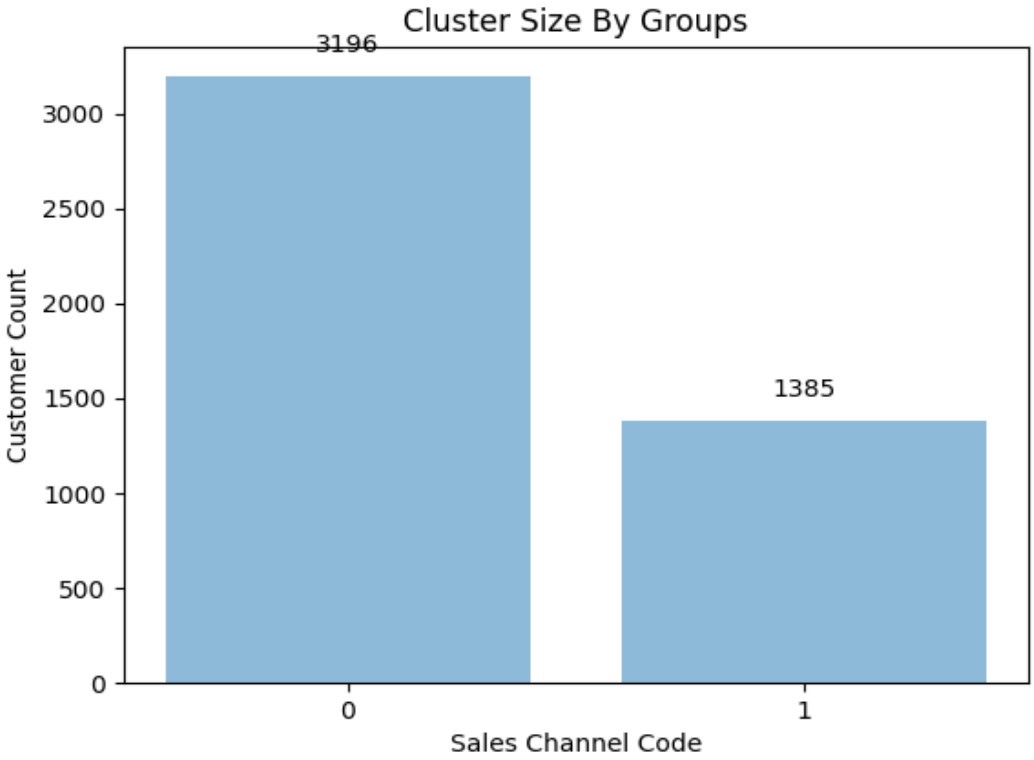

Figure 4: Distribution for 2 Clusters

Figure 4 shows the customer distribution for 2 clusters.

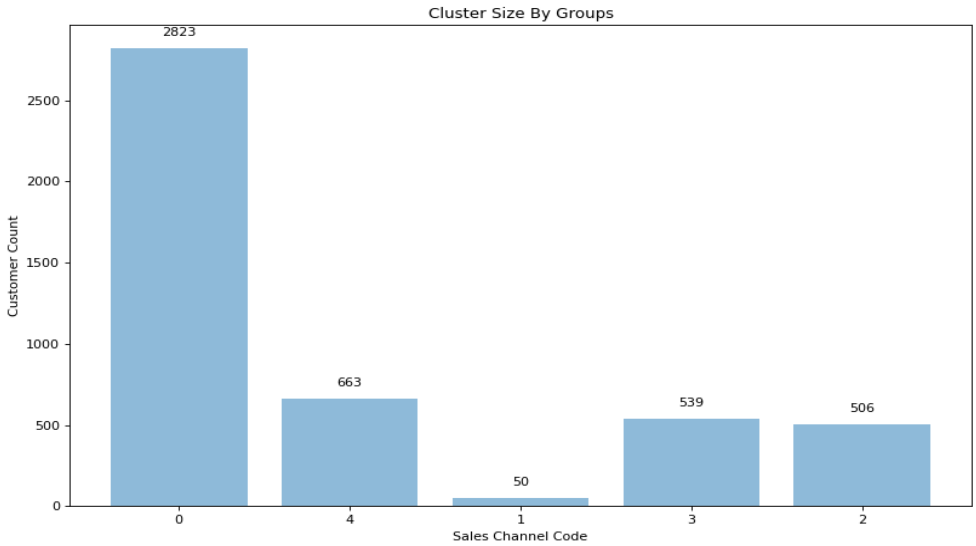

Figure 5: Distribution for 5 Clusters

Figure 5 shows the largest cluster has 2823 customers which is 60% of the total. The next 3 clusters are near size with 10% and the smallest cluster has %1.

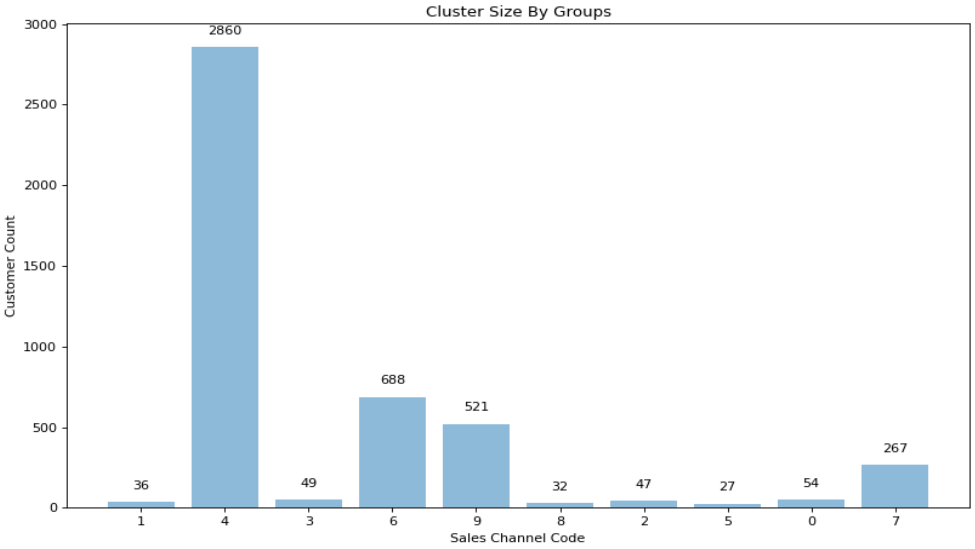

Figure 6: Distribution for 10 Clusters

Figure 6 shows the distribution with ten clusters. Like the previous figure, the largest cluster has %60 of total customers. It is an indicator that these customers' similarity is very high.

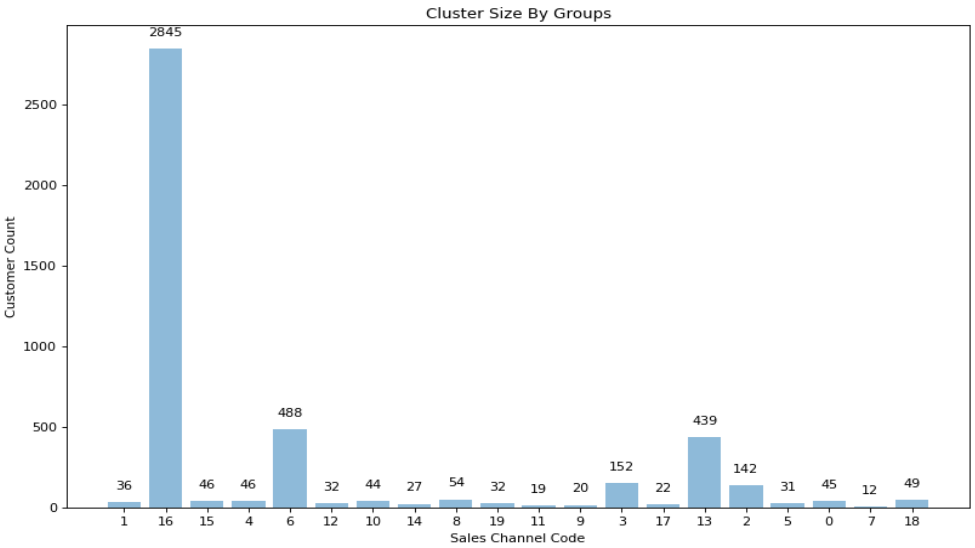

Figure 7: Distribution for 20 Clusters

Figure 7 shows that even after increasing cluster size to 20, the largest cluster has %60 of total customers.

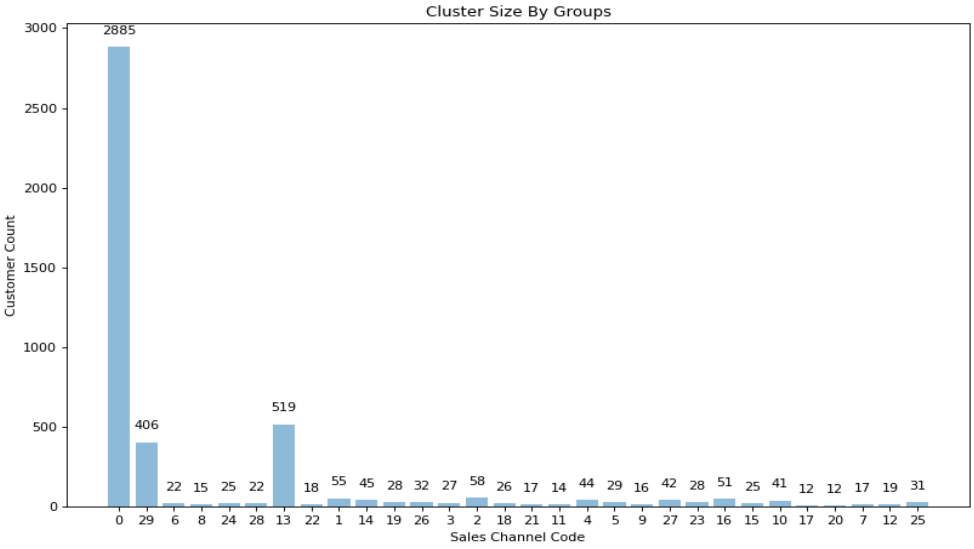

Figure 8: Distribution for 30 Clusters

Figure 8 shows that even after increasing cluster size to 20, the largest cluster still has %60 of customers.

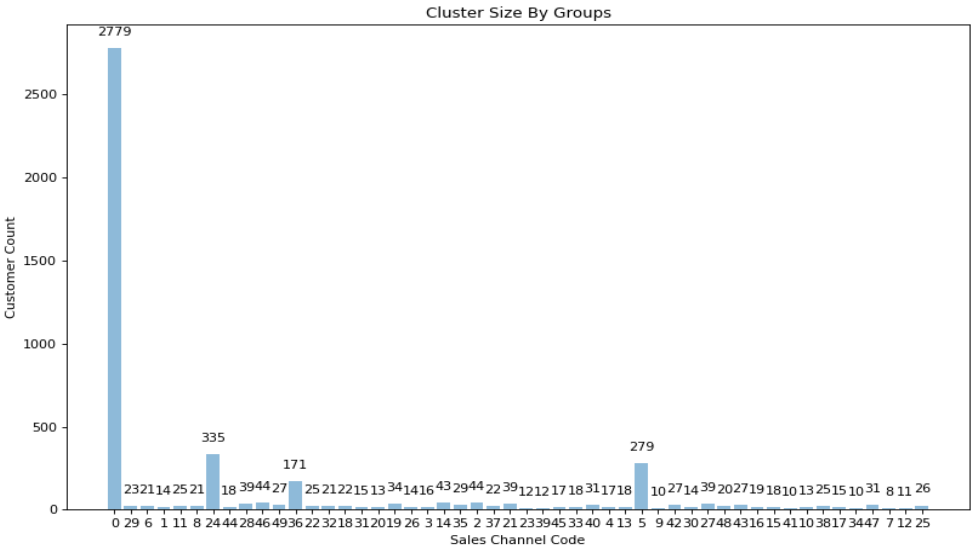

Figure 9: Distribution for 50 Clusters

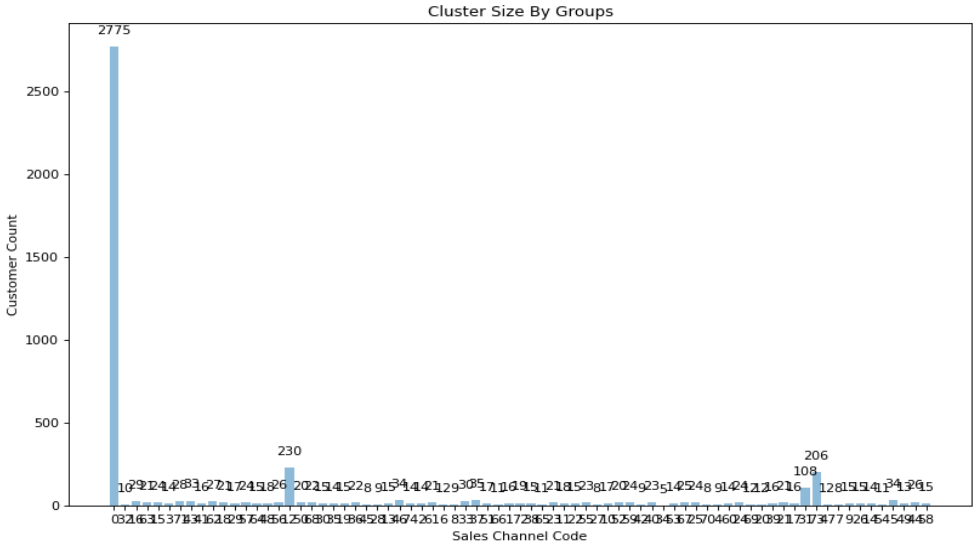

Figure 10: Distribution for 75 Clusters

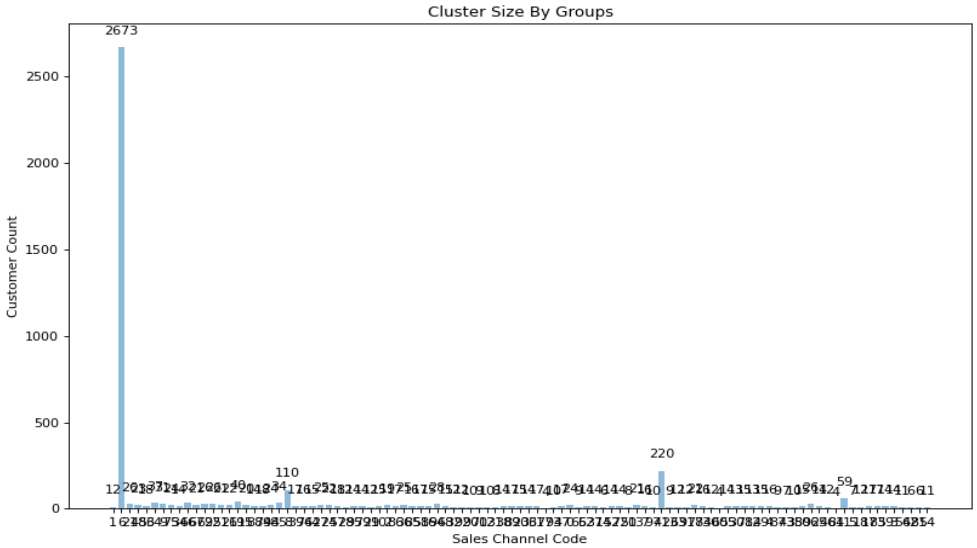

Figure 11: Distribution for 100 Clusters

Figures 9, 10, and 11 are showing distributions for cluster sizes 50, 75, and 100. They all show that %60 of customers will stay in the same cluster no matter how many clusters we choose. That is an indicator that most of the customers are having the same purchase behaviors.

When we examine the customer distribution, we see that the most crowded group includes around 2700 customers regardless of the number of groups. Increasing the number of groups only causes smaller groups to be divided more.

3.6. Determine the cluster number

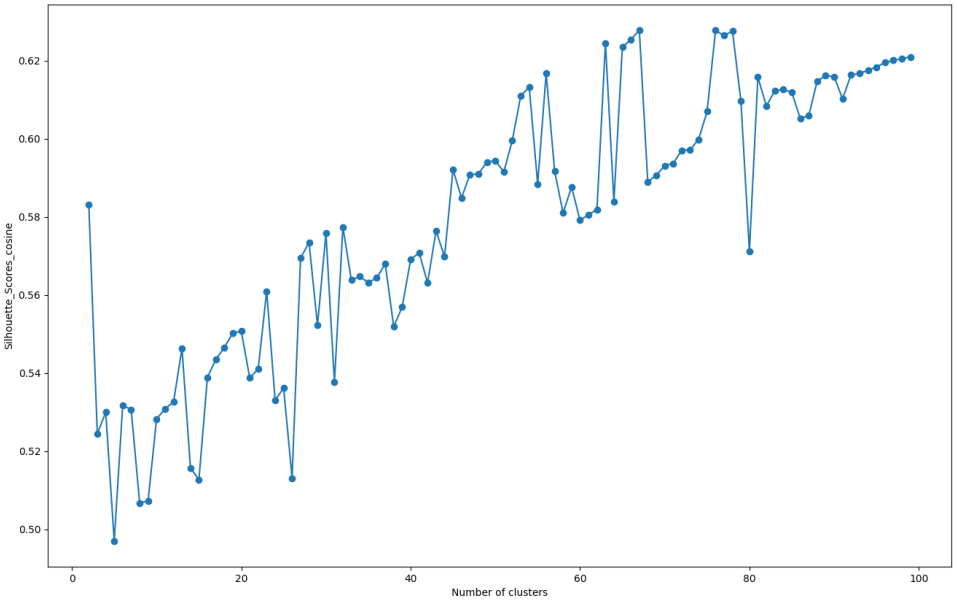

Figure 12: Silhouette score with cosine distance metric

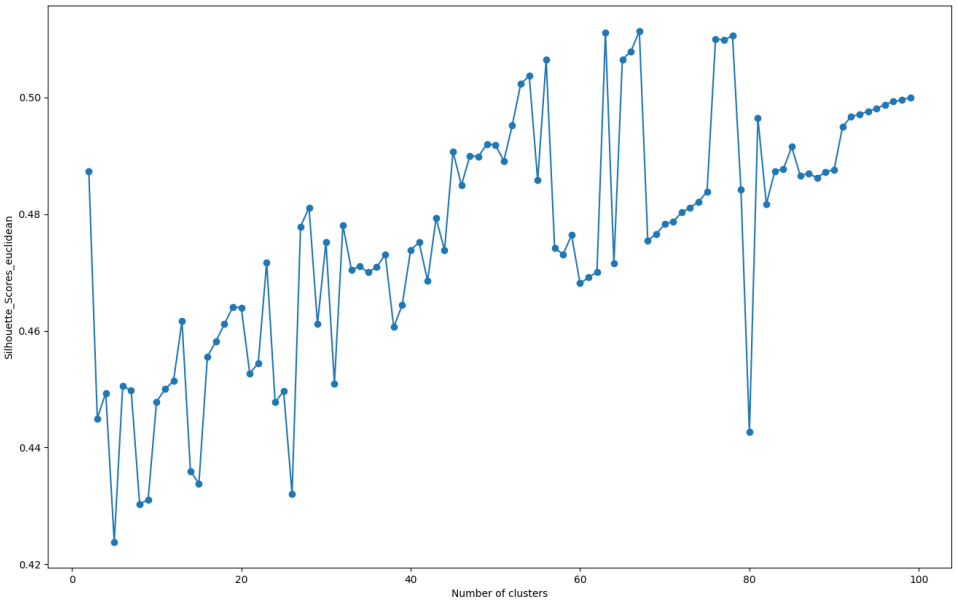

Figure 13: Silhouette score with Euclidean distance metric

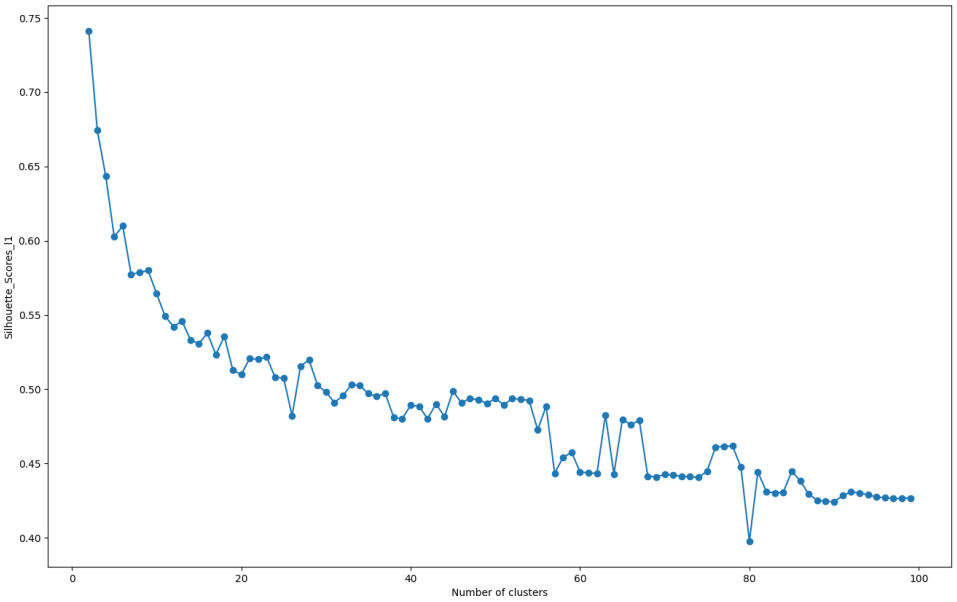

Figure 14: Silhouette score with l1 distance metric

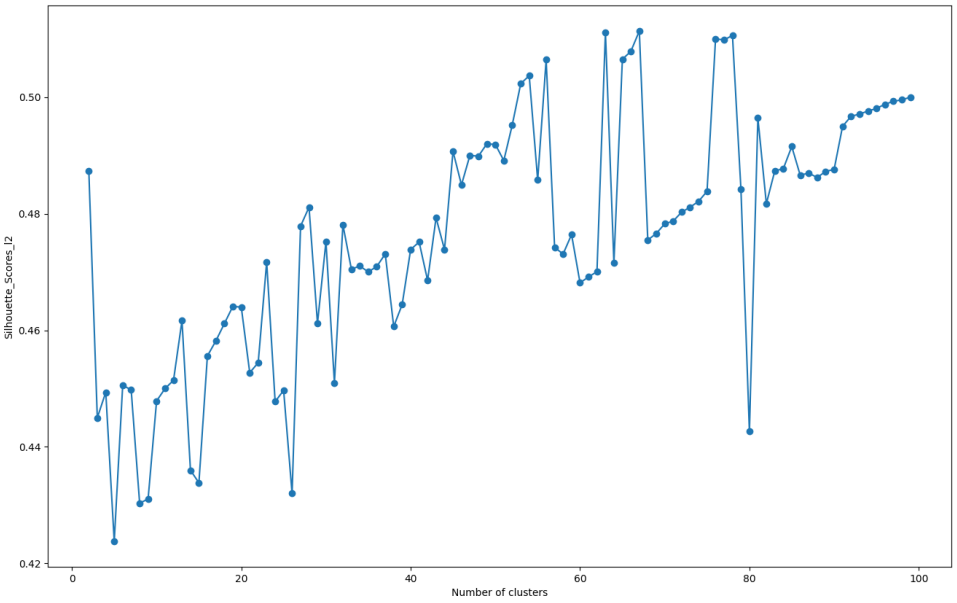

Figure 15: Silhouette score with l2 distance metric

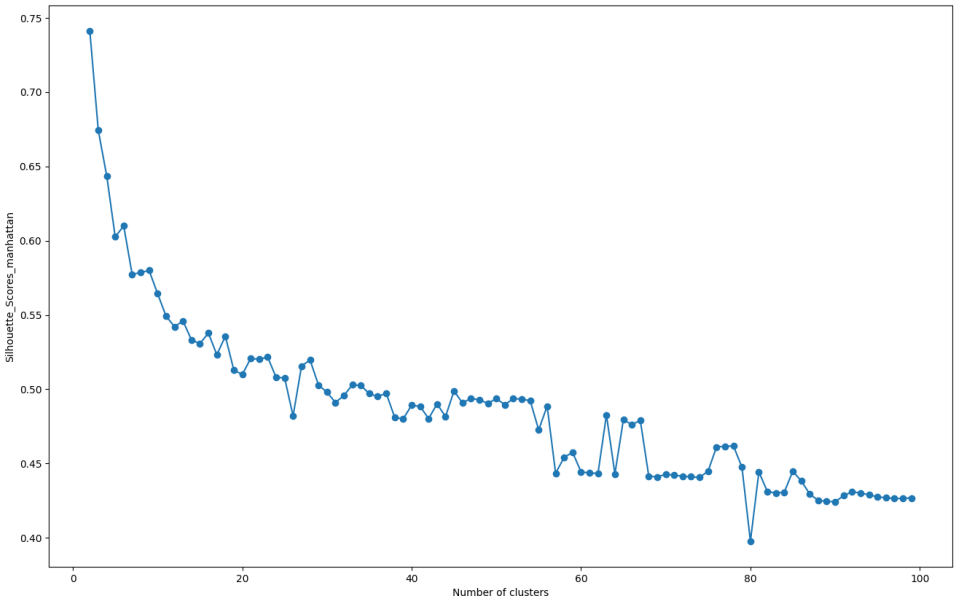

Figure 16: Silhouette score with manhattan distance metric

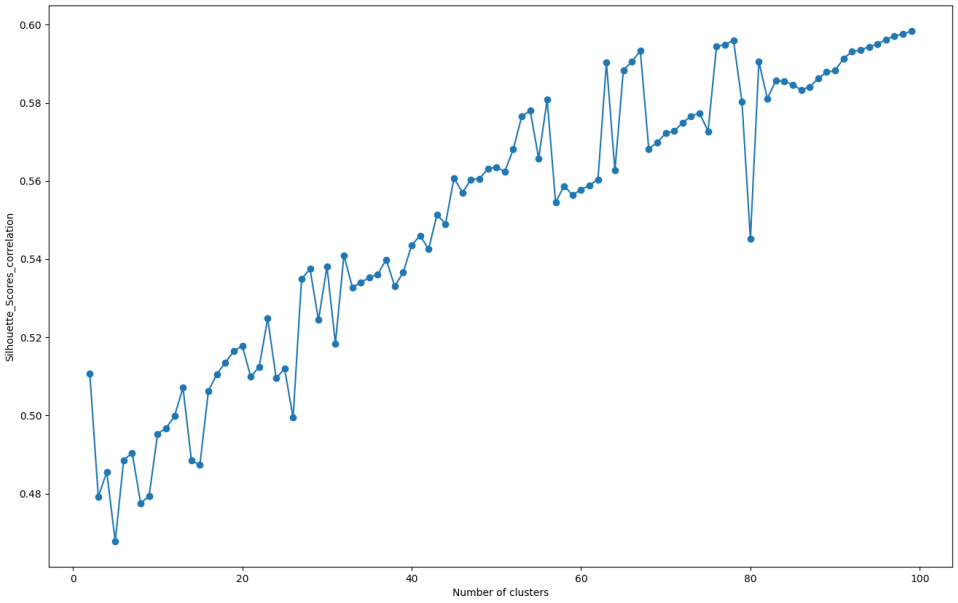

Figure 17: Silhouette score with correlation distance metric

In figures 12 to 17 silhouette scores for each cluster, with different distance metrics. Silhouette score can be a maximum of 1 and the greater values mean more efficient clustering.

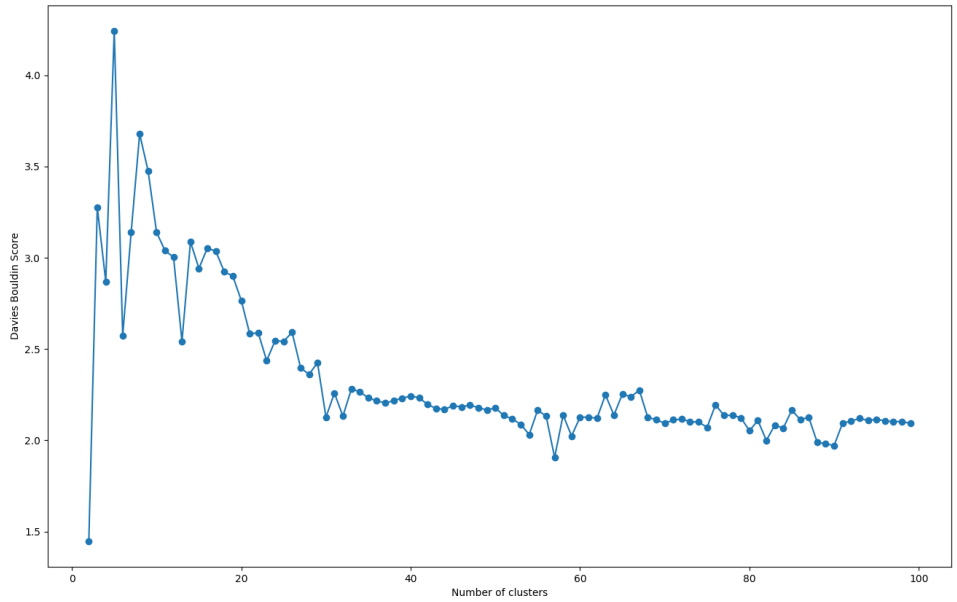

Figure 18: Davies Bouldin score

In Figure 18 Davies Bouldin scores for our dataset for each cluster number. A greater score means better clustering and after 10 clusters the score drops.

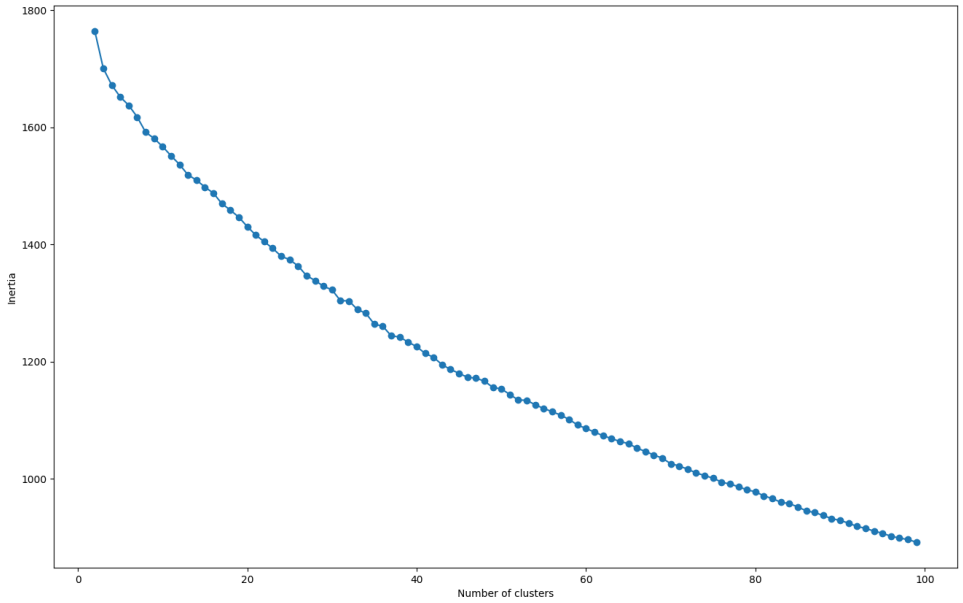

Figure 19: Inertia - Elbow Graph

We have different silhouette graphs generated by different distance metrics. We expect high value on Davies Bouldin scores and flattening in inertia scoring. Since it is not flattened enough in the inertia graph;

Let's choose 5, which is the number of groups with the highest Davies Bouldin score.

3.7. Principal Component Analysis

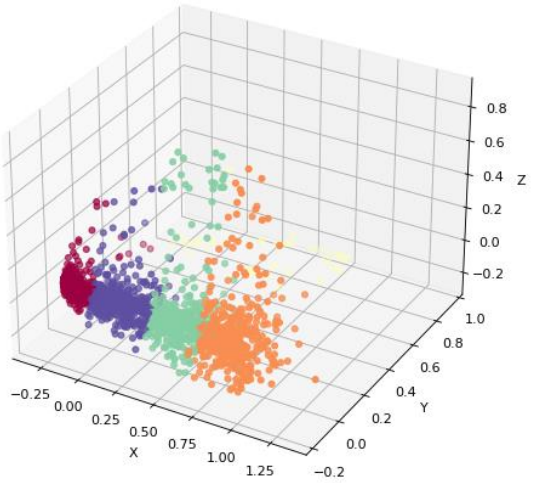

Figure 20: PCA for 5 Clusters

3.8. Analyzing common features of clusters

Most purchased Product Groups for cluster 0

- TRKK030 *
- TRMD100040 **
- TRMK020
- TRMD090020030
- TRKD060020 ***

Most purchased Product Groups for cluster 1

- TRMD100040 **
- TRMD030020020 *****
- TRMD030010020 ****
- TRKK030 *
- TRKD060020 ***

Most purchased Product Groups for cluster 2

- TRAY020
- TRMD030020020 *****

- TRMD100040 **
- TRMD040050
- TRKD060020 ***

Most purchased Product Groups for cluster 3

- TRKK030 *
- TRMD100040 **
- TRKD060020 ***
- TRMD030010020 ****
- TRMD070180

Most purchased Product Groups for cluster 4

- TRMD100040
- TRMD030010020 ****
- TRKK030 *
- TRMD030020020 *****
- TRMD090020030

Although this company has nearly 300 product groups, it can be interpreted that some product groups are used frequently and perhaps they are not determinative.

3.9. Comparison with the customer classification used in the company

Customer Groups defined by the business for cluster 0

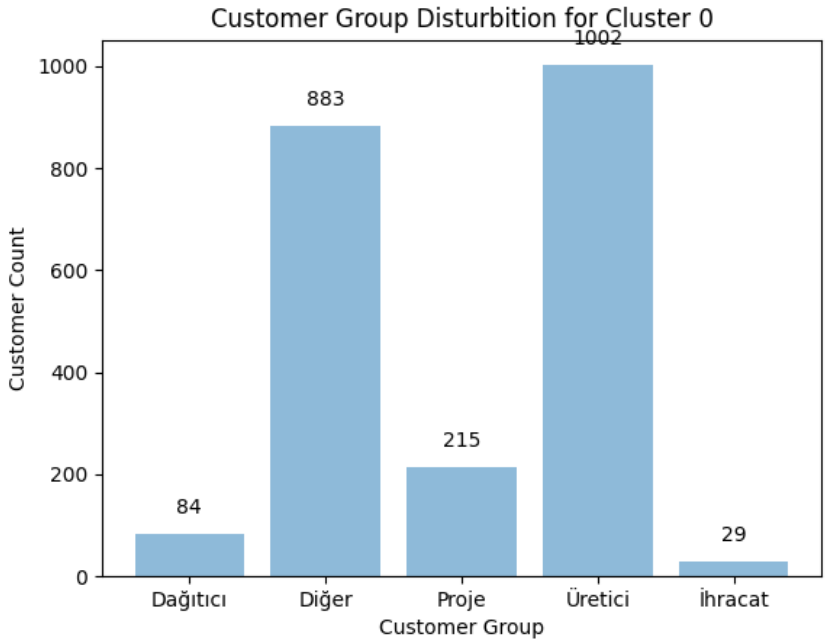

Figure 21: Customer Group Distribution for Cluster 0

In Figure 21 we can see it contains customers from all five customer groups.

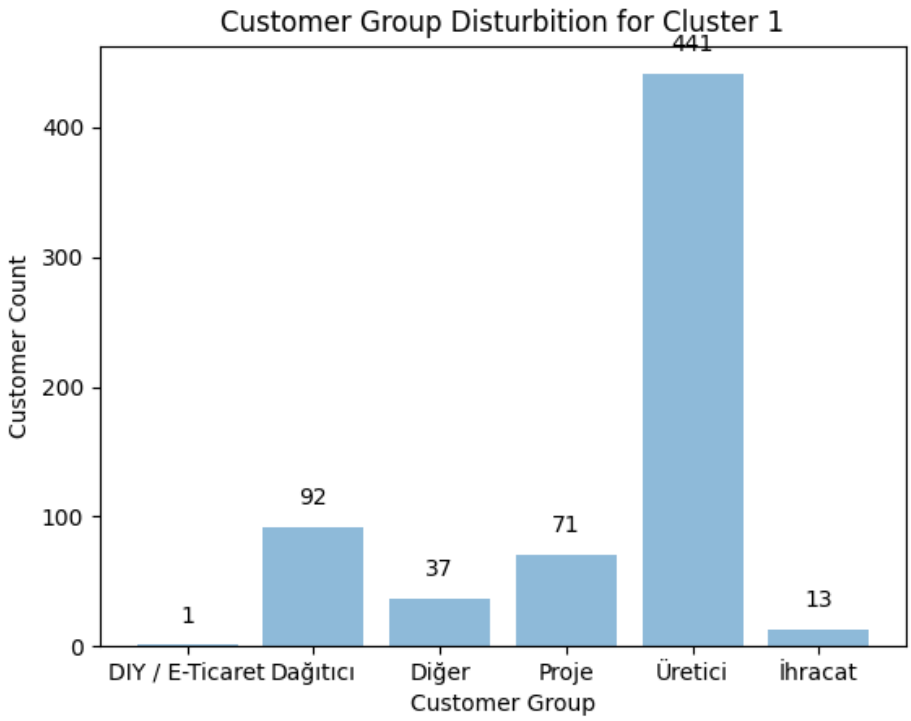

Figure 22: Customer Group Distribution for Cluster 1

In Figure 22 we can see Cluster 1 contains customers from all five customer groups with the addition of only DIY customer.

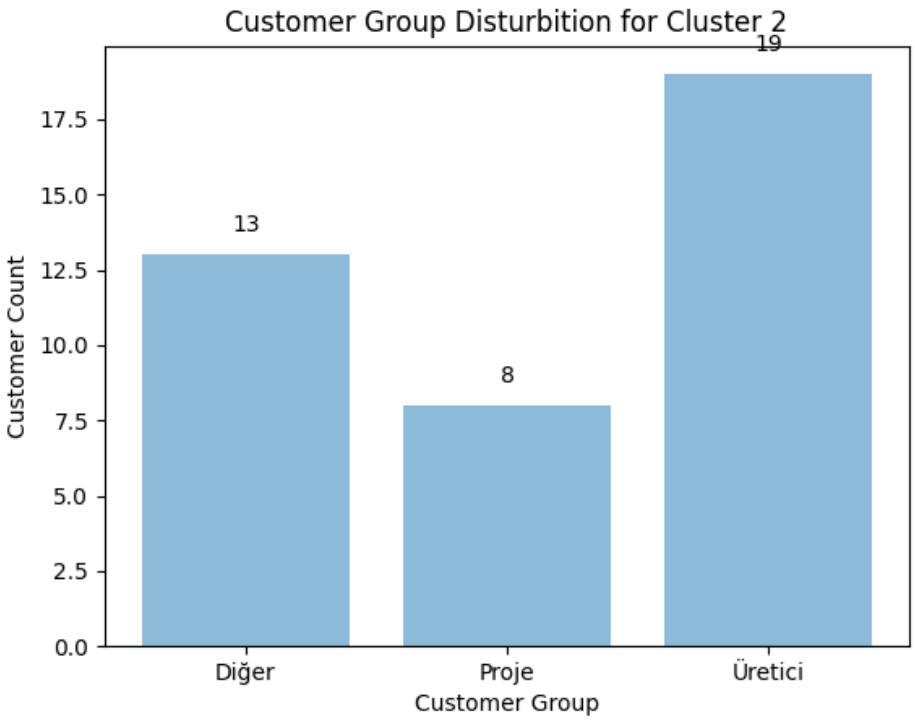

Figure 23: Customer Group Distribution for Cluster 2

Figure 23 shows that Cluster 2 has only customers with 3 different groups.

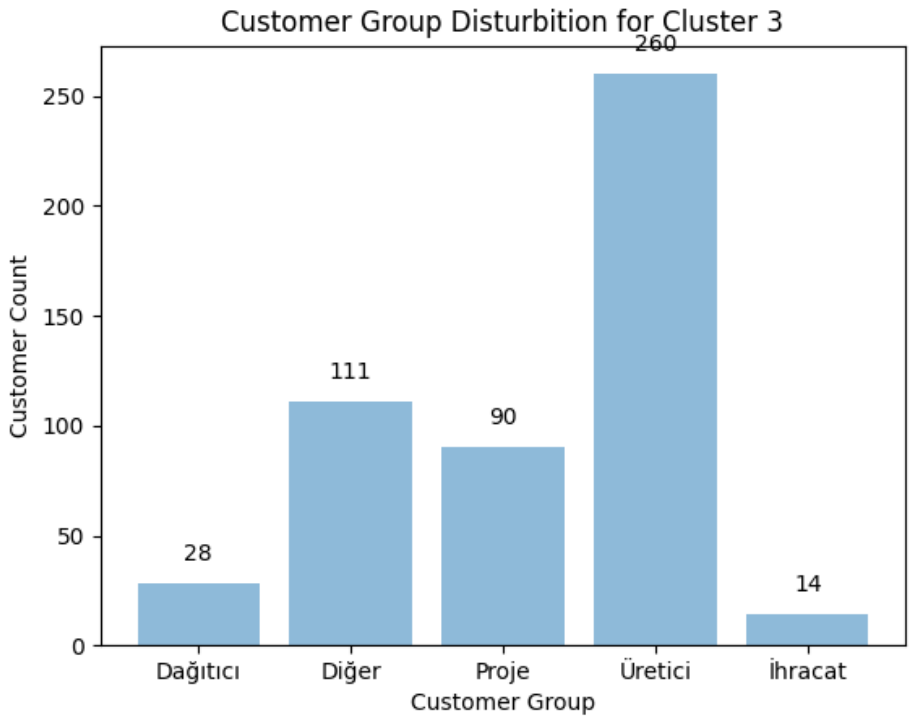

Figure 24: Customer Group Distribution for Cluster 3

Figure 24 shows Cluster 3 is similar to Cluster 0 but having fewer customers.

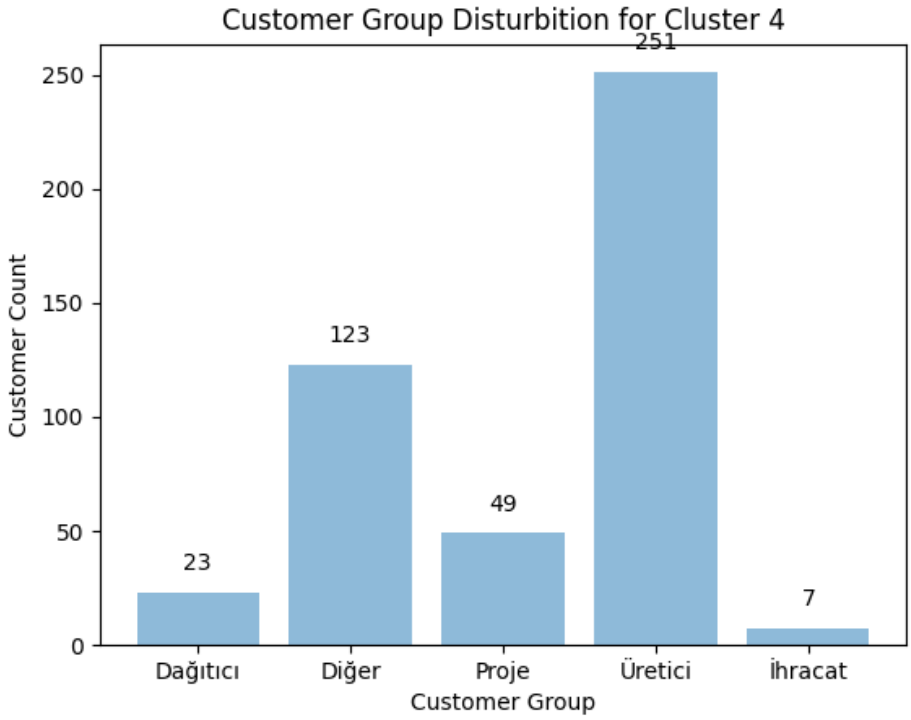

Figure 25: Customer Group Distribution for Cluster 4

Figure 25 shows the distribution of customer groups in Cluster 4 which is similar to other clusters.

Almost every cluster has all the customer groups, which shows they are not defined by their purchase behaviors. It depends on two different customer segments,

Most of the customers are manufacturers and they are subgroup by their production areas like kitchen furniture or living room furniture. But our result shows us even if they produce different products their purchased product groups are similar.

The second main group is distributors which purchase a product to sell producers themselves. So we don’t expect them to have different purchase behavior than producers.

4. RESULT

Although there are many product groups defined in the company, there is no close distribution among them, most of the sales are in certain product groups and even many product groups have not been sold to any customer. Correct customer classification may not be possible in such modeling because of the product groups and sales strategy. As seen in the results, more than half of the customers remain in the same group despite the increase in the number of groups.

What can be done to create a better model;

By removing common and widely used product groups, the differences between customers can be made clear.

As another suggestion, manufacturers can be evaluated among themselves and compared according to their basic production areas.

With the Pareto method, customers can be selected from customers that make up 80% of the total turnover, and customer-based controls can be analyzed.

As a result, it will give more efficient results for this company to classify customers and determine actions by performing RFM analysis rather than analyzing on a product group basis.

APPENDIX

The codes are stored in a public Google Drive folder.

https://drive.google.com/drive/folders/1AOMbddO6l7vjTnKFynmF_mYD53FG_yOK?usp=sharing

REFERENCES

[1] https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac

[2] https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3

[3] https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis#:~:text=The%20clusters%20that%20result%20assist,their%20wants%2C%20needs%20and%20preferences.&text=Rather%2C%20the%20data%20itself%20reveals,within%20the%20population%20of%20customers.

[4] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[5] https://medium.com/@jyotiyadav99111/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308

[6] https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering

[7] https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index

[8] https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial