CUSTOMER SEGMENTATION USING CLUSTERING AND DATA MINING TECHNIQUES

A.NAGA RAJU, RAKOTI SIRISHA

Assistant Professor in MCA DEPT, Dantuluri Narayana Raju College, BHIMAVARAM, Andhra Pradesh
Email id: nagaraju.dnr345@gmail.com
PG Student of MCA, D.N.R COLLEGE, PG Courses (Autonomous), Bhimavaram-534202.
Email id: sirisharakoti2030@gmail.com

ABSTRACT

In today’s competitive market, it is critical to comprehend customer behavior and categorize customers according to their demographics and purchasing habits. This is an important part of consumer segmentation since it enables marketers to better customize promotional, marketing, and product development tactics to different audience groups. Customer segmentation is the practice of dividing consumers into groups based on common characteristics, spending habits, and purchasing patterns.

Customer segmentation is a technique for determining how to interact with customers in various categories to maximize the value of each client to the company. Customer segmentation can help marketers reach out to each customer in the most effective way possible. Customer segmentation study makes use of a large amount of customer data to properly identify distinct groups of consumers based on behavioral, demographic, and other criteria. K-means clustering is an unsupervised machine learning approach, as opposed to supervised ones.

When the dataset is unlabeled data, it employs this method. Unlabeled data is information that hasn't been assigned to any of the available categories or groupings. Different techniques are applied in customer segmentation to determine the optimal number of clusters, but each approach has its own drawbacks, such as the Density- based Spatial Clustering of Applications with Noise (DBSCAN) algorithm failing in the scenario of changing density clusters. Regency, Frequency, Monetary Value (RFM) research is based on historical data rather than future projections. The Hierarchical Clustering technique can never erase what has already been done, and therefore necessitates the use of labeled data. Whereas the K-means method ensures convergence, warm-starts the centroid's locations, and quickly adapts to new and form optimal cluster numbers.

.1. INTRODUCTION

1.1 Overview

When creating criteria for segmenting consumers into logical groups, such as customers who came from a given source, dwell in a specific location, or purchased a specific product or service, it's always easier to make assumptions and depend on "gut emotions." On the other hand, these high-level categorizations seldom produce the desired results. Certain clients will, without a doubt, spend more money with a firm than others.

Most loyal consumers will spend a considerable quantity of money over an extended period. Good clients will either spend a little overtime or a lot in a short amount of time. Others will not overspend or stay for a lengthy amount of time. The most successful form of customer segmentation analysis is to split consumers into groups based on projections of their full future worth to the company, to address each group or individual in the most effective way possible to maximize that future, or a lifetime, value.

1. Behavioral
2. Demographic
3. Geographic
4. Psychographic

1.2 Motivation for the project

Because the marketer's goal is to maximize the value (revenue and/or profit) from each client, knowing how each marketing action will affect the customer ahead of time is critical. In an ideal world, such "activity-centric" customer segmentation would focus on the long-term impact of a marketing action on customer lifetime value (CLV) rather than the short-term value of marketing activities. As a result, customers must be divided into groups or segments based on their CLV. [2]

In a distinct kind of customer segmentation, machine learning algorithms are utilized to find new groups. In contrast to marketer-designed segmentation models like the ones mentioned above, machine learning consumer segmentation allows advanced algorithms to reveal insights and cates that marketers could overlook on their own. Marketers that can relate their segmentation model to campaign results will see their customer groups improve over time. In these situations, the machine learning model will be able to not only fine-tune its segment definition, but also assess whether one segment is outperforming the others, therefore optimizing marketing performance.

1.3 Problem Definition And Scenarios

- To implement the customer segmentation analysis and dividing consumers into groups based on common characteristics, spending habits, and purchasing patterns
- Product Recommendation based on Customer Segmentation Engine.
- To develop a success prediction models for customer segmentation.
- Customer segmentation is a technique for determining how to interact with customers in various categories to maximize the value of each client to the company.

1.4 Organization of the report

The chapter 2 is about the literature review on the various projects that are studied and understanding of the project in more detailed way and analysis of the system that are described.

The chapter 3 is discussed on the project description in which the analysis of the existed system and how it is differ from the proposed system and analysis of the model using spiral diagram.

The chapter 4 is all about the system design where the system is explained using block, flow etc. how they are developed on the basis of simplification this gives the clear understanding of project.

The chapter 5 is mainly discussed on the system requirements of the project such as hardware and software components and their specifications.

The chapter 6 is about all modules of the project and its description and explanation that the project is divided in to module to make every module completed successfully.

The chapter 7 is mainly discussed on the implementation of the project such as hardware and software components and their specifications.

The chapter 8 is all about the result and analysis of parts that are used in our project.

The chapter 9 is all about the conclusion of the project in which how the project is concluded and what we have done in this project and also about the future enhancement of the project what will be the extension of this project for later development This the organization of thesis what we have discussed in the following chapter this brief understanding on each and every chapter.

2. LITERATURE SURVEY AND RELATED WORK

2.1 Introduction

This chapter is about analysis of the different projects that are published up to now and about the project the details that given and discussed in the analysis of project in detail.

2.2 Explore Patterns of Customer Segments

Customer behavior evolves and changes over time in real-world settings. In order to establish efficient marketing strategies, this dynamism must be considered while conducting consumer segmentation analyses and other business-related tasks. The study's major goal is to look at the patterns of structural changes in consumer groups. There hasn't been any study done on this subject yet. This is the first research to look at the influence of consumer dynamics on structural changes in segments. The goal of this study is to create a way to describe and explain this problem. Using clustering and sequential rule mining approaches, a novel strategy is suggested. A new concept and methodology for identifying distinct sequential rules are also established. The proposed strategy is tested using data from customers.[1]

2.3 Discovering New Business Opportunities

For sales and marketing departments inside major organizations, identifying and understanding new markets, clients, and partners is a vital task. Intel's Sales and Marketing Group (SMG) is experiencing similar challenges as it expands into new markets and industries and evolves its present business. To aid SMGs in sorting through millions of organizations across several locations and languages to find relevant routes, sophisticated automation that enables a fine-grained knowledge of enterprises is essential in today's challenging technological and commercial context. We show a system developed in our company that mines millions of public business

Web pages to provide a faceted client representation. We focus on two key characteristics of the customer that are important for discovering suitable opportunities: Industry sectors (which range from huge verticals like banking and insurance to smaller niches like manufacturing). [7]

2.4 RFM Customer Segmentation

The RFM model, which is utilized in the traditional retail business for consumer segmentation, is not ideal for an industry with different social group qualities; hence the RFMC model is established by adding the social relations parameter C. For this empirical investigation, educational e-commerce firm M was chosen, and the k-means algorithm was utilized to cluster legitimate clients of enterprise M, resulting in five unique customer groups and proving the model's usefulness. [8]

2.5 High-Dimensional Customer Segmentation

The Omni channel becomes a hot topic as a result of the rapid rise of e-commerce and clients' growing familiarity with multichannel shopping. To satisfy the present trend of client demand, several firm organizations to work on the Omni channel business issue and are progressively devoting their efforts to both online and offline business. As a result, there's no doubting that understanding online customers' purchasing patterns is critical to Omni channel success. Using the RFM (regency, frequency, monetary) model and the k-means clustering technique, consumers' information is retrieved and customers are segmented. To expand the RFM model, we divide total frequency and monetary data into weekly level data, resulting in a reduction in the number of variables associated with one week. [13]

3. EXISTING SYSTEM

- Outliers that are isolated in low-density regions are identified as outliers by DBSCAN, which merges points that are close together. Two important factors create the model's 'density': the minimum number of points required to generate dense regions min samples and the distance required to establish a neighborhood eps. Larger min samples or lower eps demand a higher density to construct a cluster.

- RFM Analysis is a marketing method that combines the three aspects of regency,, and monetary value to analyze and understand customer behavior. The RFM frequency Analysis will help businesses divide their customer base into a variety of homogenous groups, allowing them to connect with each group using a variety of targeted marketing strategies.

- A way of grouping comparable components is hierarchical clustering, often known as hierarchical cluster analysis. The endpoint is made up of several clusters, each of which is unique yet shares a lot of similarities.[2]

4. PROPOSED SYSTEM

The number of segmentation choices is practically limitless, and they are mostly determined by the quantity of consumer data we have available. It starts with the basics, such as gender, hobby, or age, and progresses to elements such as the amount of time since the user last visited the business or the amount of time spent on website.

i. Geographic: The concept of geographic consumer segmentation is straightforward; it all boils down to the user's geographic location. This may be done in several different ways. You may sort your results by city, zip code, nation, or state.

ii. Demographics: Demographic segmentation takes into account the structure, size, and movement patterns of consumers over time and space. Many businesses create and market products that are based on gender differences. Another important factor is the status of the parents. This type of information may be obtained through customer surveys.

iii. Behavioral: Customer segmentation based on past behavior can be used to forecast future behavior. Customers' favorite brands, for example, or the times of the year when they make the most purchases. The behavioral component of consumer segmentation seeks to understand not just why people buy things, but also how those reasons change over time.

iv. Psychological: The psychological segmentation of customers includes personality

traits, attitudes, and beliefs. Consumer surveys are used to collect this information, which may then be used to assess customer sentiment.

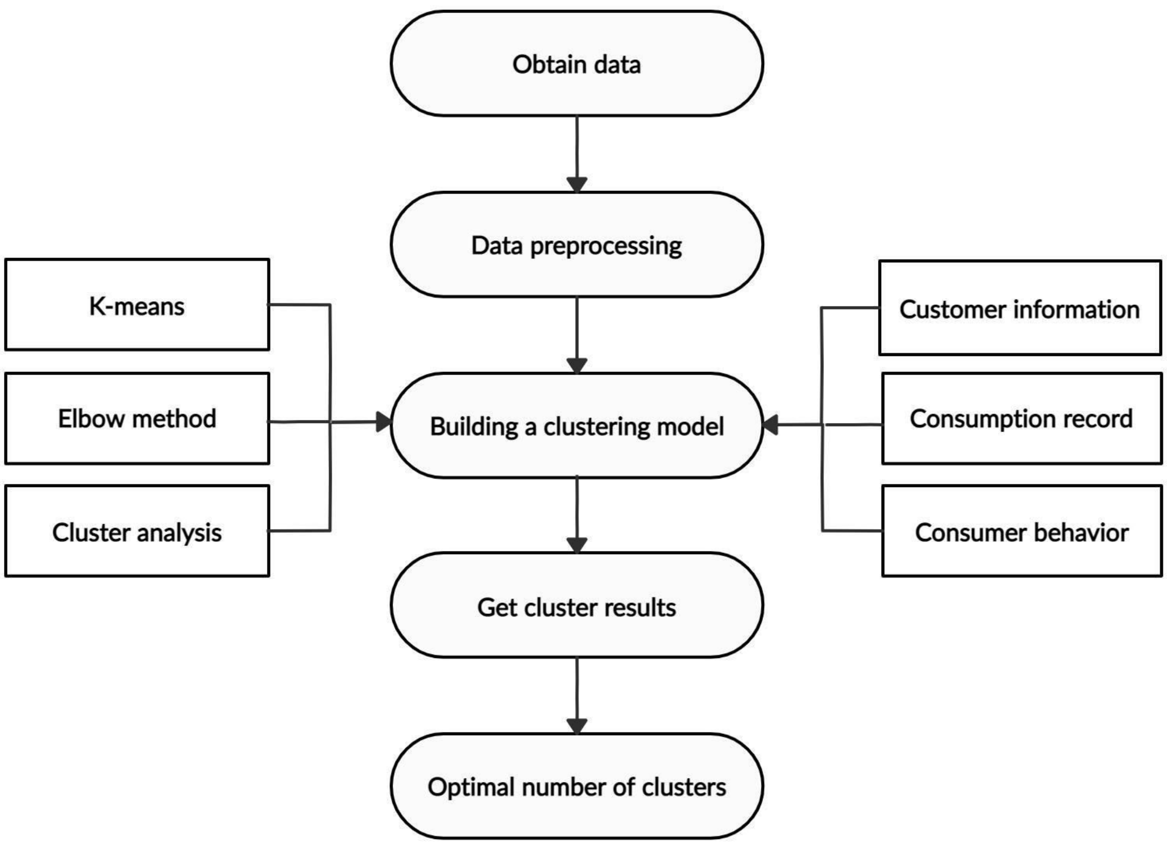


FIG 1 – SYSTEM ARCHITECTURE

5. METHODOLOGIES

SEGMENTATION

Let's have a look at how the consumer segmentation challenge is being solved.

i. Create a business case: The first stage is to make a business case for the project. There must be an aim for everything. You don't want to become involved with this. Blindly. Otherwise, the result will be unorganized and untidy. You'll need a business case instead. It's identifying the most lucrative consumer groupingswithin the total pool of customers in this scenario.[9]

ii. Prepare the data: The data must then be prepared in the second stage. What is the size of your data set? In this instance, a hundred, thousand or ten thousand client data is preferable. This is due to the fact that you will be able to see more patternsand trends. You'll also need a set of features based on the most relevant business indicators inside the data collection. The data is then preprocessed to reduce discrepancies, which helps in better data analysis.

iii. Data analysis and exploration: Data analysis and investigation is the third stage. This is an important stage since it will allow you to discover some intriguing relationships and trends in your data. You will be able to better comprehend the client's interests and purchase patterns as a result of this, and you will be able to determine which traits are most directly associated with the consumer and, obviously, the business.[14]

iv. Clustering analysis: Clustering analysis in the context of a client is the fourth phase in the process. The use of a mathematical model to uncover groupings of similar consumers based on the tiniest differences between customers is known as segmentation clustering analysis. The purpose of cluster analysis within each group is to precisely categorize consumers so that personalization may be used to create more successful customer marketing. A mathematical approach called k- means clustering analysis is a popular cluster analysis tool. The resulting clusters aid in improved consumer modeling and analysis. There are no pre-set thresholds or standards in place for this procedure. The data, on the other hand, exposes the

customer prototypes that exist intrinsically inside the consumer base.

v. Choosing optimal hyperparameters: The fifth step is to select the besthyperparameters. "Tuning" or "hyperparameter optimization" is the process of selecting the optimal set of hyperparameters for an algorithm. Based on our past work, this is the next stage since it assists us in identifying the most accurate and rewarding client groups.[10]

vi. Visualization and interpretation: Visualization and interpretation are the final steps in the process. Now it's time to visualize and interpret your findings. Businesses can improve marketing campaigns by targeting features, launches, andproduct roadmaps when they have profitable customer profiles at their fingertips. This gives the company a much clearer picture of which customers are most likely to stick around.[15]

K-MEANS ALGORITHM

Let's look at the K-means method, which we'll use in this project. The k-means technique is an iterative method for splitting a data collection into K distinct, non- overlapping subgroups, each of which contains just one data point. It sought to make the data points in the intracluster clusters as comparable as feasible. while preserving as much space between the clusters as possible It distributes data points to clusters in such a way that the squared distances add up to one. Inside clusters, the cluster centroid is the location with the least variation, and the most homogeneous data points are found within the same cluster. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster. Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:

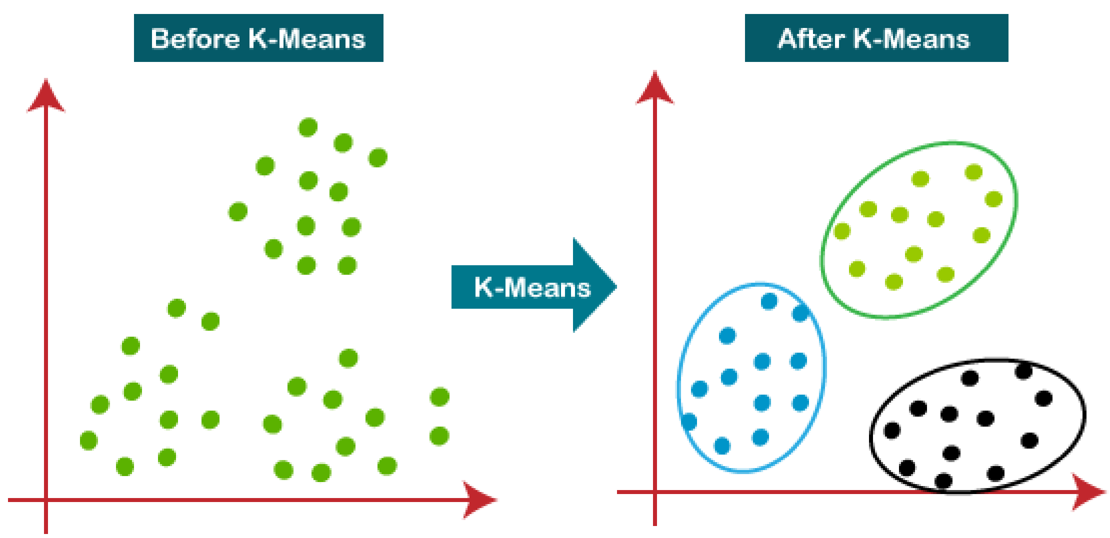


Fig. 2 K-Means Clustering Algorithm

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, but here we are discussing the most appropriate method to find the number of clusters or value of K

ELBOW METHOD

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$WCSS= \sum_{P_i \text{ in Cluster1}} \text{distance}(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} \text{distance}(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} \text{distance}(P_i\ C_3)^2$

In the above formula of WCSS,

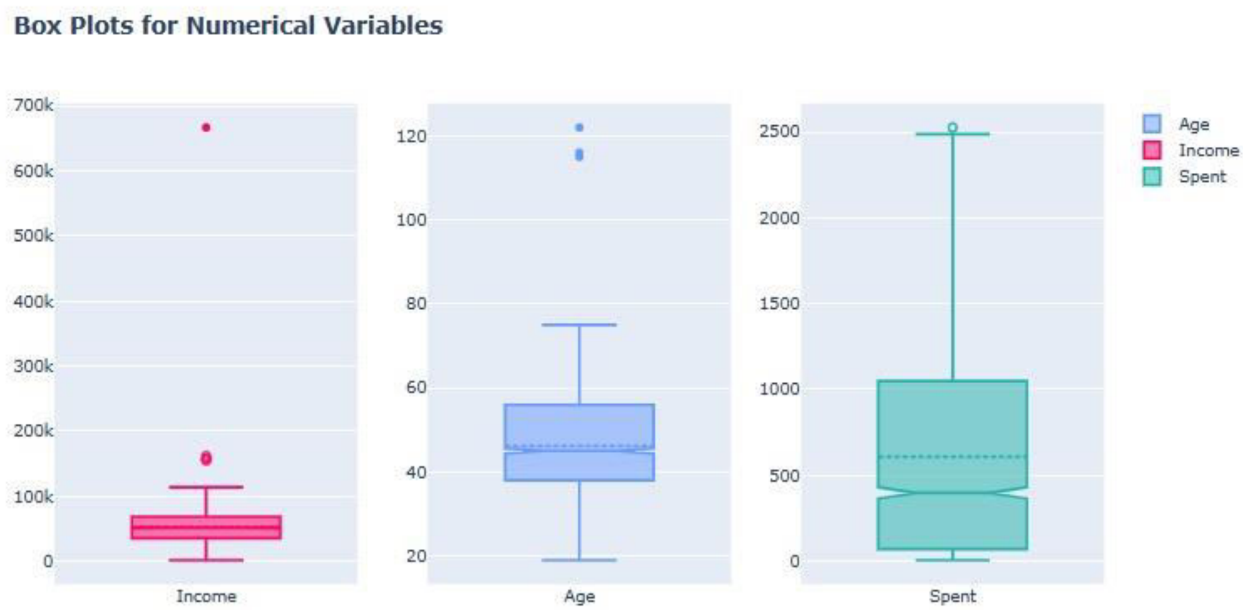
$\sum_{P_i \text{ in Cluster1}} \text{distance}(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

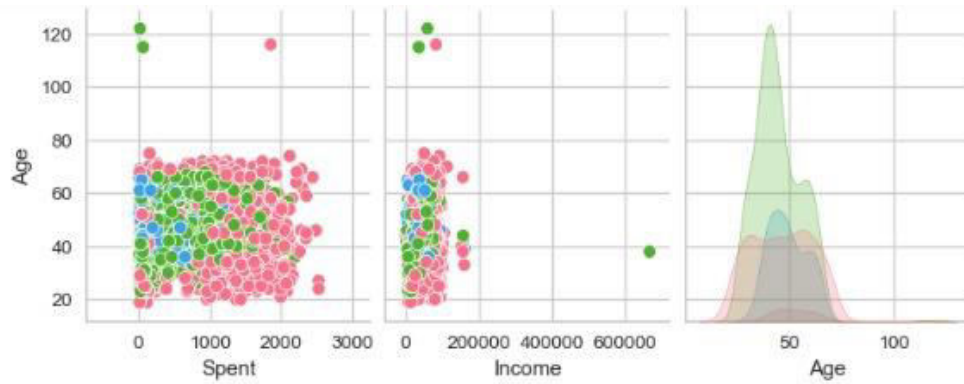
To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

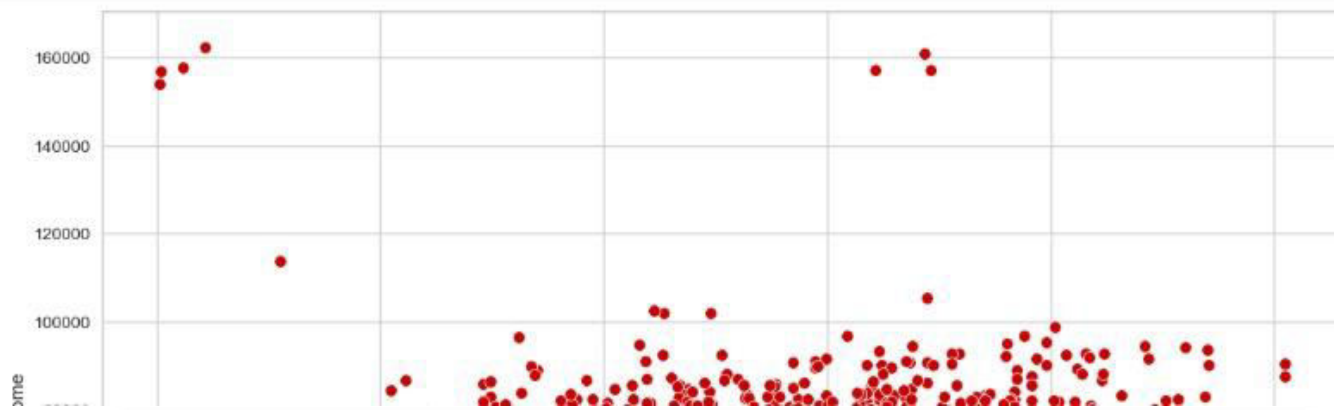
- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

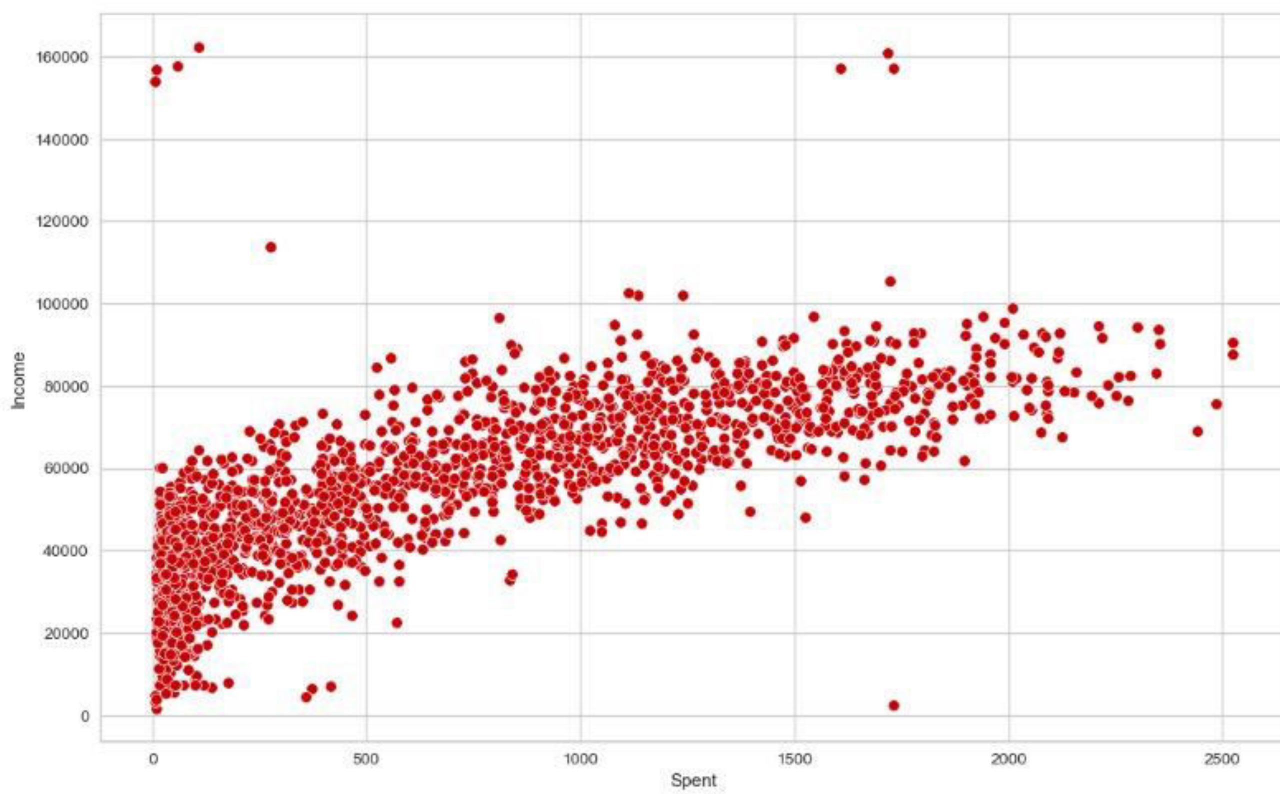
6. RESULTS AND SCREEEN SHOTS:



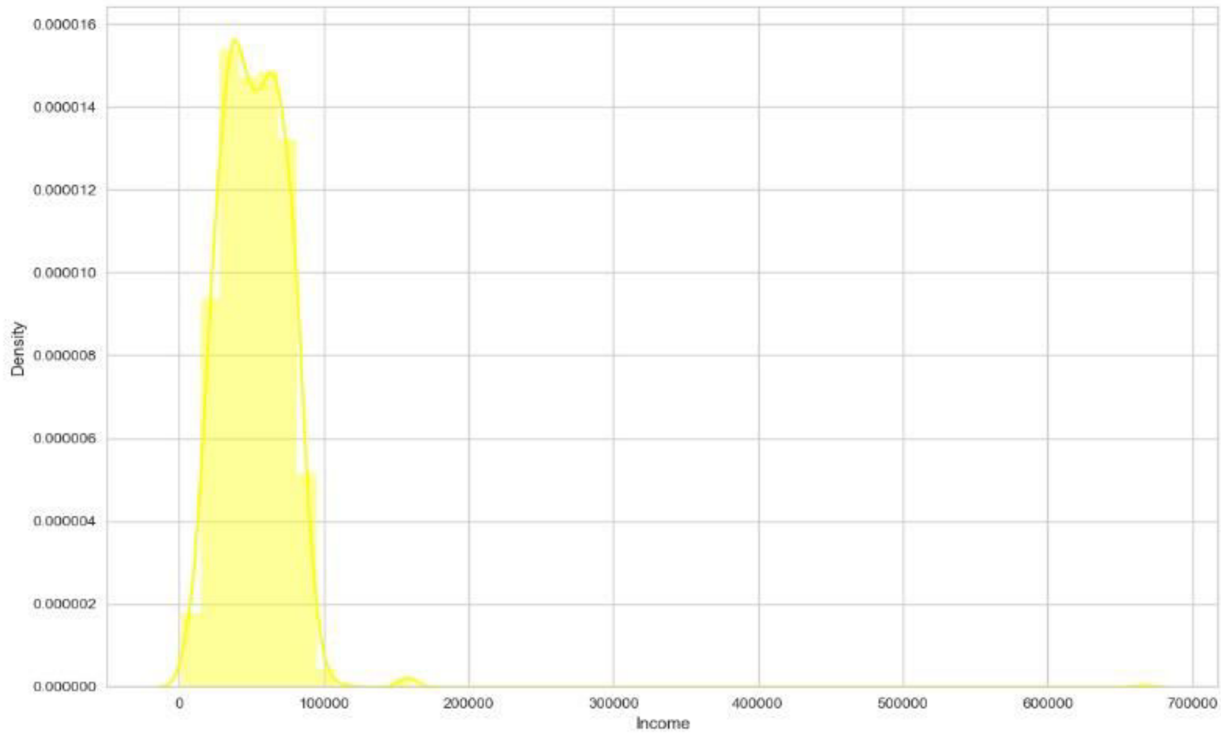


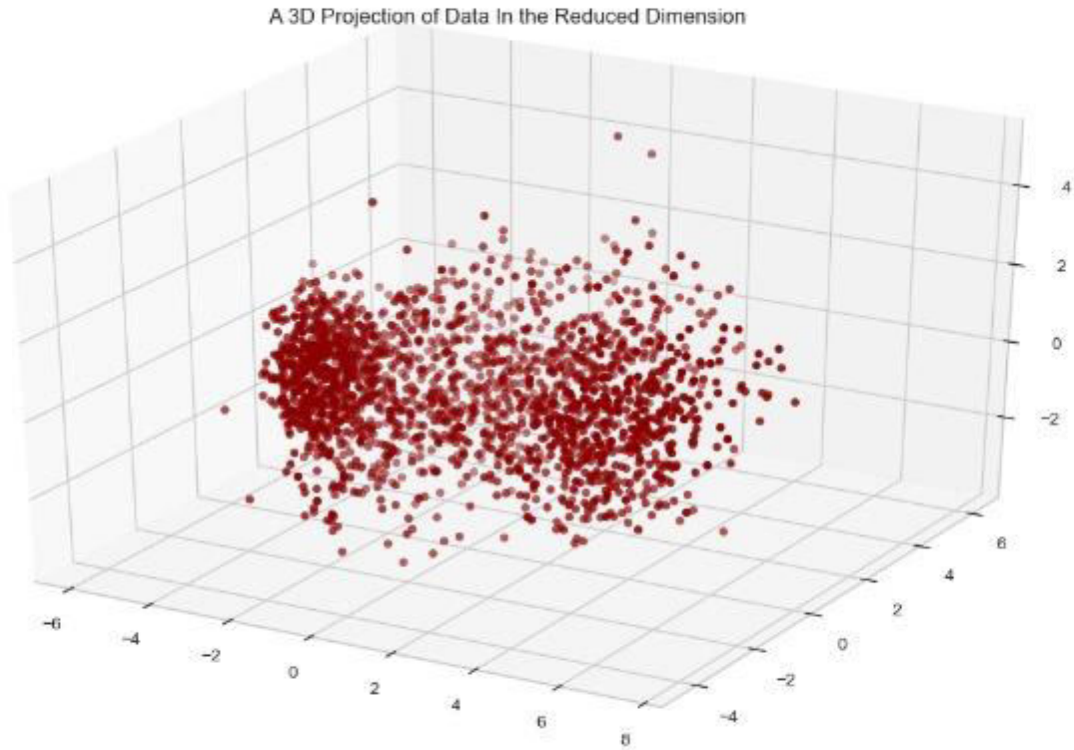

```
]: plt.figure(figsize=(13,8))
   sns.scatterplot(x=data[data['Income']<600000]['Spent'], y=data[data['Income']<600000]['Income'], color='#cc0000');
```



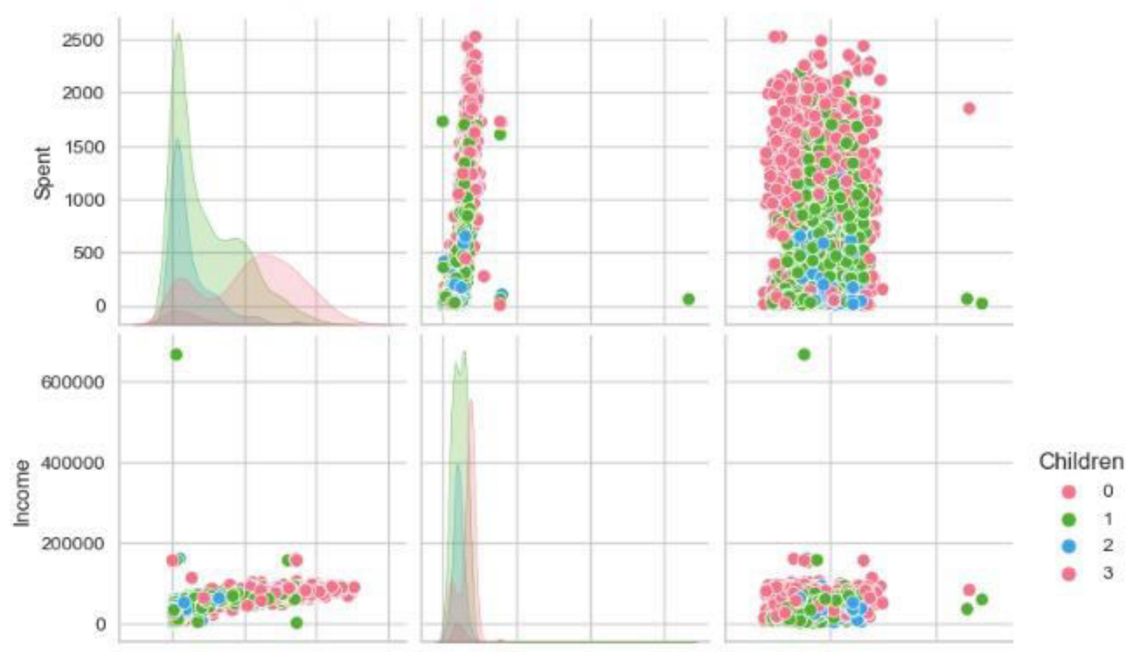


In [30]:
```
plt.figure(figsize=(13,8))
sns.distplot(data.Income, color='Yellow');
```

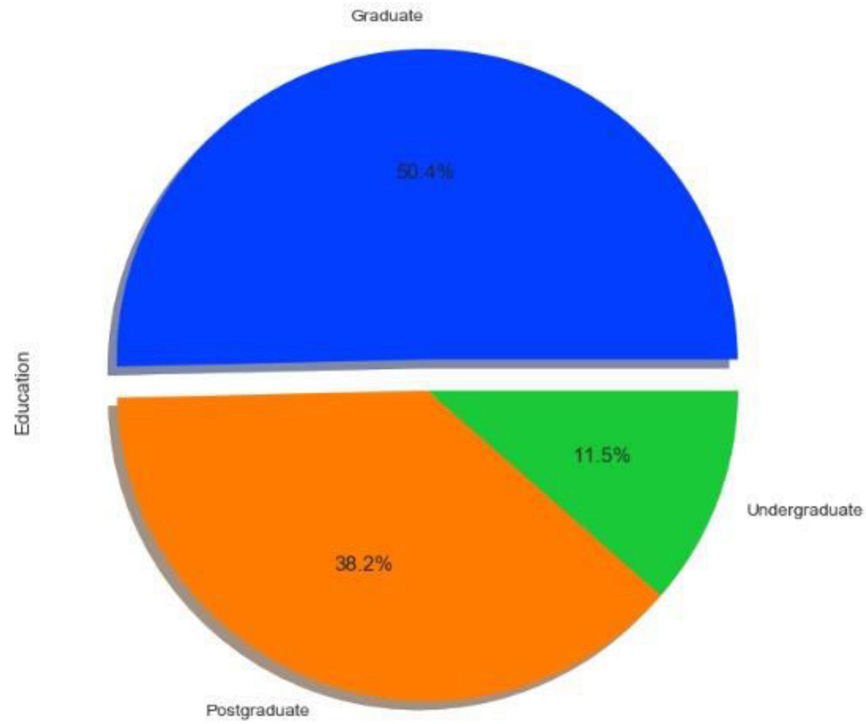




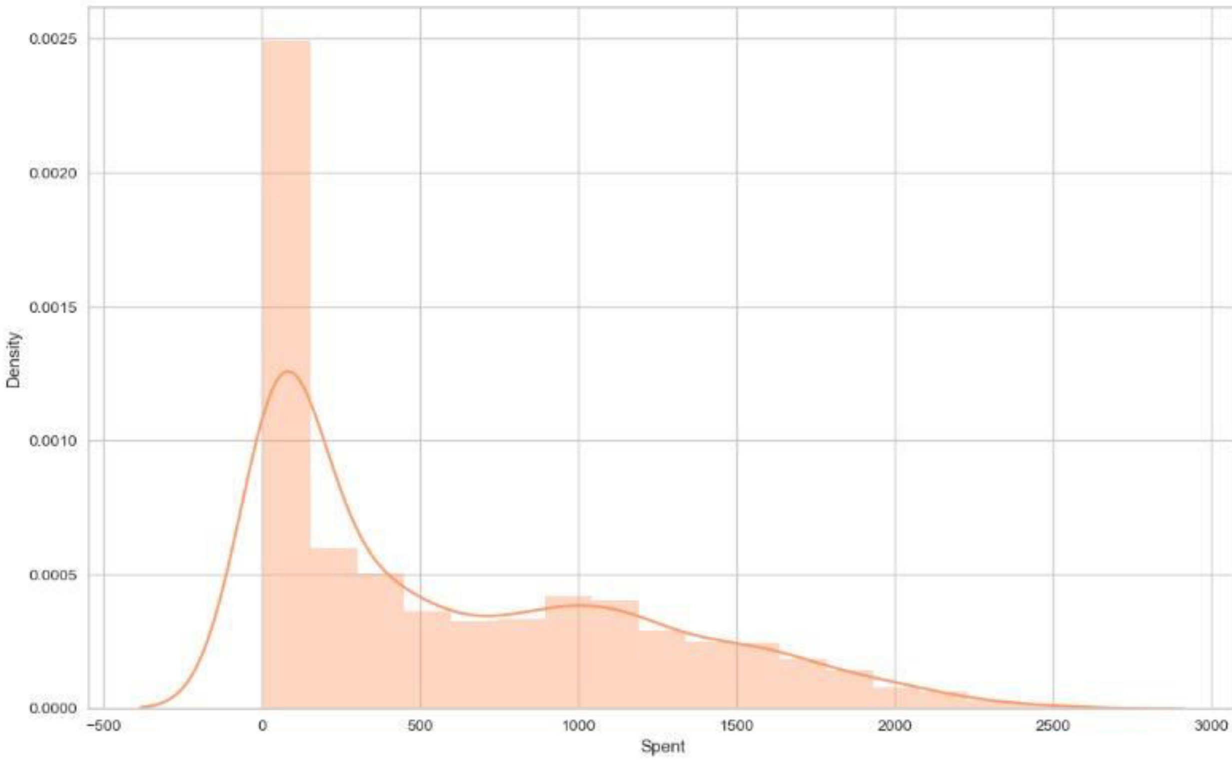

```
In [24]: sns.pairplot(data , vars=['Spent','Income','Age'] , hue='Children', palette='husl');
```



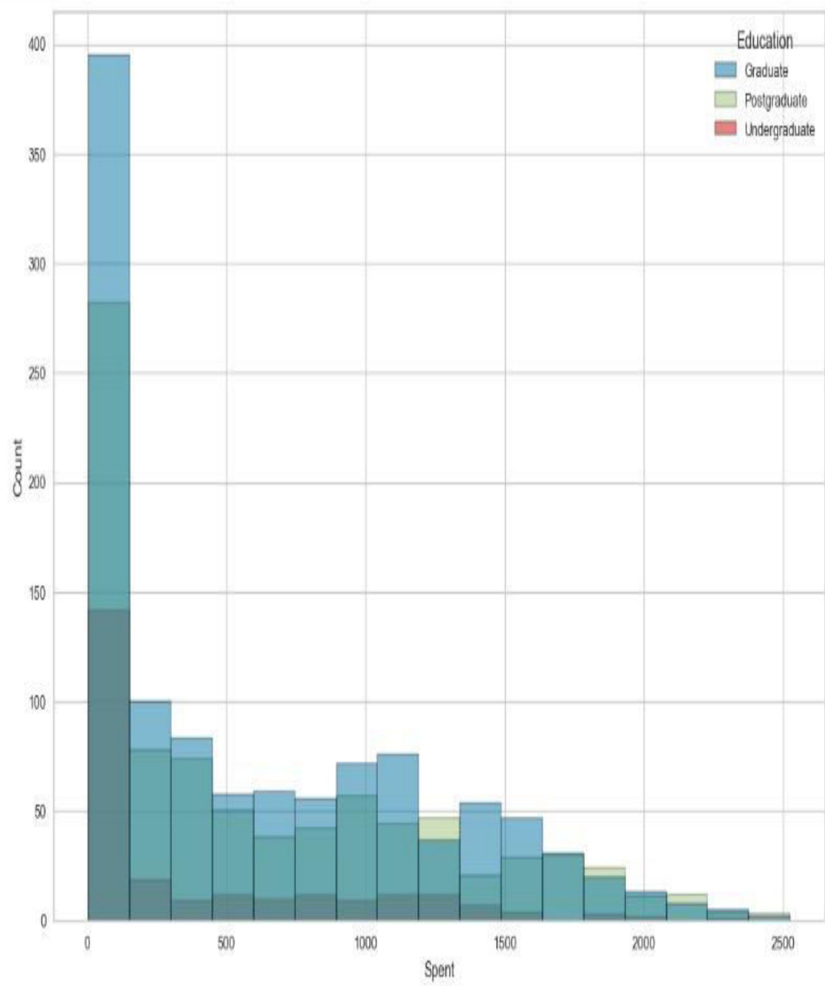
```
In [28]: data['Education'].value_counts().plot.pie(explode=[0.1,0,0], autopct='%1.1f%%', shadow=True, figsize=(8,8), colors=sns.color_pal
```

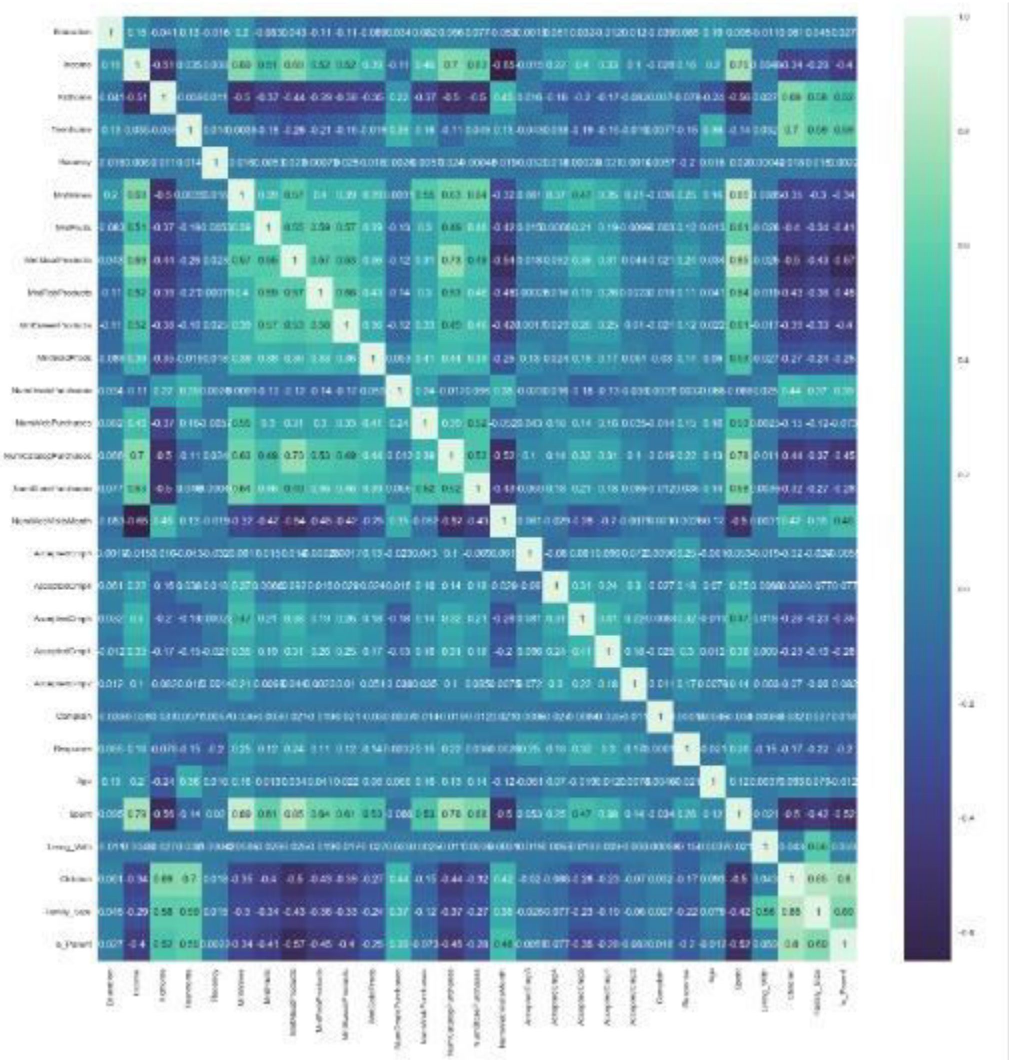


In [31]:
```
plt.figure(figsize=(13,8))
sns.distplot(data.Spent, color='#ff9966');
```



In [27]:
```
plt.figure(figsize=(13,8))
sns.histplot(x=data['Spent'], hue=data['Education']);
```





7. CONCLUSION AND FUTURE SCOPE

CONCLUSION

Our contribution to this research paper is to highlight and solve the challenges of product recommendation based on the customer segmentation engine. Data science can use Customer segmentation to build a stronger relationship with their customers. It enables them to make educated retention choices, develop new features, and position their product strategically in the market.

FUTURE SCOPE

As part of their customer segmentation strategy, retailers with extensive, multi-category offers must show their things in such a way that target customers may search and pick from those offerings. In the first piece of this dissertation, a product segmentation approach is described. The proposed approach offers merchants a methodology for identifying customer-centric, cross-category product segments from large numbers of goods across multiple categories, where products within a segment are bought by the same type of customers. The research also looks at the relationship between the recommended product segmentation approach and a customer segmentation method. Because the approaches are so closely linked, the product and customer groups inferred by each are likely to be identical.

8. REFERENCES

[1] Elham Akhond Zadeh Noughabi;Amir Albadvi;Behrouz Homayoun Far "How Can We Explore Patterns of Customer Segments' Structural Changes? A Sequential Rule Mining Approach" In Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration, 26 October 2015

[2] Wayne Xin Zhao;Sui Li;Yulan He;Edward Y. Chang;Ji-Rong Wen;Xiaoming Li"Connecting Social Media to E-Commerce: Cold-Start Product Recommendation Using Microblogging Information" IEEE Transactions on Knowledge and Data Engineering,volume: 28, pages: 1147-1159, 17 December 2015

[3] Xiaojun Chen; Yixiang Fang; Min Yang; Feiping Nie; Zhou Zhao; Joshua Zhexue Huang "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data" IEEE Transactions on Knowledge and Data Engineering, volume: 30, page 559-572,26 October 2017

[4] Qinglu Gao;De Xia;Yangyan Shi;Ji Quan"Policies Adoption for Supply Disruption Mitigation Based on Customer Segmentation" IEEE Access, volume: 7, page 47329 - 47338, 29 March 2019

[5] Fuxiang Liu"3D Block Matching Algorithm in Concealed Image Recognition and E- Commerce Customer Segmentation" IEEE Sensors Journal, volume: 20, pp. 11761 - 11769, 19 August 2019

[6] Caroline Gobbo Sá Cavalcante;Diego Castro Fettermann"Recommendations for Product Development of Intelligent Products" IEEE Latin America Transactions, volume: 17, pages: 1645-1652, October 2019

[7] Itay Lieder; Meirav Segal;Eran Avidan;Asaf Cohen;Tom Hope "Learning a Faceted Customer Segmentation for Discovering new Business Opportunities at Intel", In Proceedings of 2019 IEEE International Conference on Big Data (Big Data),24 Dec. 2019