Customer Segmentation Analysis

***Abstract* –** In this paper, we explore a dataset which contains customer segmentation on attribute, campaigns, products that was certainly used for analysis to market the product according to the specific customer segment. For this we created the model through data derivation, encoding and after then applying PCA (Principal Component Analysis) to reduce the features and certainly applying K-means to cluster. We provided inference according to the exploratory data analysis (EDA) and answered the analytical questions.

I. INTRODUCTION

Customer Segmentation Analysis is an extensive evaluation of a company that gives better understanding of the target audience and how to optimize their marketing according to the trends. This paper describes how this analysis provide benefit to the company to improve or modify its products and market that product according to the customer segments.

This Analysis also provides an aid for the business by instead of spending budget on marketing of new product to each customer, a business can analyze which segment will be most likely inclined to buy the product and then certainly market that product to that specific segment

II. ANALYTICAL QUESTIONS & DATA

Our study uses customer segmentation data that have sourced from Kaggle datasets. Our main objective is to produce actionable insights that would answer our research questions. Our research would comprise of 3 analytical questions and our as follows:

1. Which Customers have not accepted promotional campaigns so that unattracted campaigns could be better targeted to boost sales?

2. Which customers have churned so that marketing team could come up with personalized offers for such customers?

3. Which group of products have low number of sales, or the products are high in demand for which deals, and discounts or adequate supplies are required, respectively, in order to increase their sales?

We have created these well formulated analytical question with the motivation that it would help the marketing team to come up with deal, discounts and personalized offer when required according to the segments.

III. DATA (MATERIALS)

A. Key Characteristics

Our data ranges from 2012-2014, has 31 features in total, 2241 records, each row represents one record, and each field is a property of said record. This dataset has nan values in only one column i.e., Income. Moreover, a combination of numerical and categorical types was observed. An extensive pre-processing will be required while imputing values instead of nan values, removing outliers and transforming columns. From the features given in the dataset, and by having domain knowledge, initial feature selection is essential to make our research interpretable.

B. Is It Suitable for Our Study?

Our Data contains customer attributes such as age, education, marital status, income etc. Furthermore, products such as wine, fruits, meat etc. and several promotions modes of purchases that will contribute to segmentation and extracting useful insights, thus answering our analytical questions stated.

C. Key Assumptions

Following are the key assumptions about the data:

1. Features are selected from the correlation matrix and will provide an aid in extracting meaningful insights.
2. The dataset is unlabeled data, therefore unsupervised learning or clustering will be applied in order to perform inference.

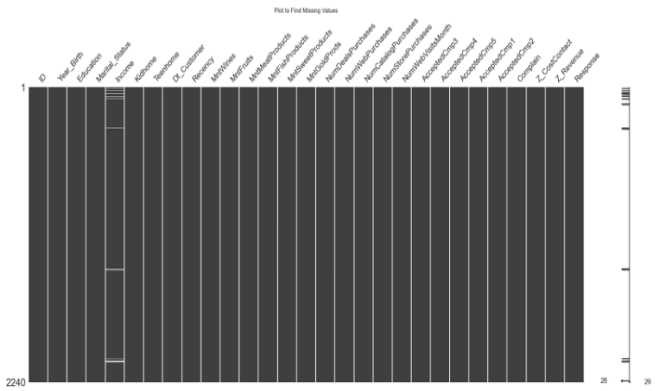
IV. ANALYSIS

A. Data Preparation

1. Dealing with Missing Data & Imputation

For the missing records in each column, we visualized by matrix plot as shown in Figure 1. From the Figure 1 we get to know that only feature named Income contains nan values that needs to be treated.

Figure 1



For imputing values, there are 3 ways such as taking mean, median or mode of a column and imputing that value in a missing row. As Income is numerical column thus, median is best practice to follow for the imputation purpose.

B. Data Derivation & Encoding

Data Derivation enables us to transform data into more useful format that is suitable for the underlying problem. We applied Feature Engineering to numerous features as it was required. Column named as *Year_Birth* was in date format; thus 'Age' was extracted, furthermore we transformed the Age feature into more categorical feature as *AgeGroup* consisting of categories such as *Young, Adult and Elderly* to visualize it with different features. Created new feature as *totalAmountSpent* from adding all the amount spent on products to get the trends for the customers with different segments. Feature called *Marital_Status* had issue of sampling error caused by low volumes of records in categories such as Yolo, Absurd, Widow due to which it was merged to Single.

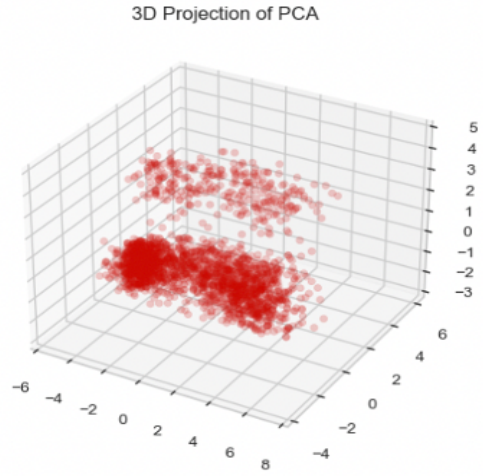
Encoding is applied when categorical variable is converted into numerical variable in order to fit in the machine learning model. In this study we are using K-means clustering, so the algorithm requires numerical variables. There were 3 categorical variables Education, Marital Status and Age Group which was converted into numerical values using one hot encoding that transforms categorical features in binary vector representation.

C. Construction of Models

1. Dimensionality Reduction (PCA)

In our data there were a lot of features, thus reduction in the number of features were required. We employed PCA a dimensionality reduction technique that reduces the number of features that are less important and retains the essential features while retaining the data integrity. PCA is sensitive to variance of features and the features having high values will be regarded as of higher importance by the PCA due to which these features will dominate the analysis, furthermore, it will be biased towards those features. To cater this issue scaling was applied.

Figure 2

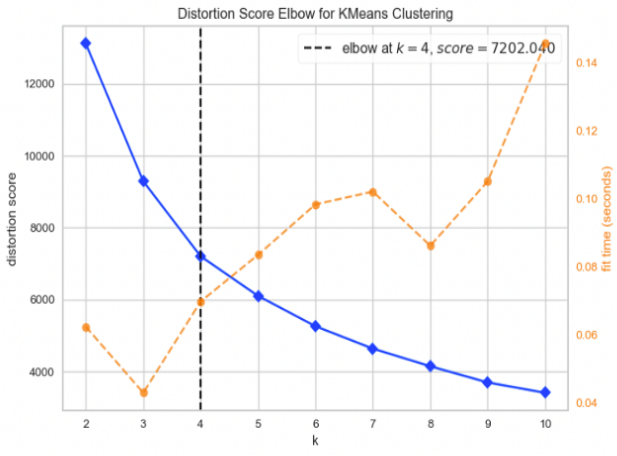


Scaling is applied on the data in order to standardize the features to have common scale. For that purpose, Standard Scaler was used that certainly subtracts mean from every feature and divides it by standard deviation. In our data many of the features were correlated as well as some were redundant so there was a need to apply PCA so that minimum information loss is observed while increasing interpretability. In fig 2 it is evident that the components were reduced to 3 dimensions

2. K-means Clustering

In this study, we will be performing clustering via K-means clustering. K-means cluster is an unsupervised learning method where it randomly selects the K (cluster) and takes Euclidean distance of the points from the centroid and assigns the point closest to that cluster. K-means finds the similarity between the points and group or cluster them together.

Figure 3

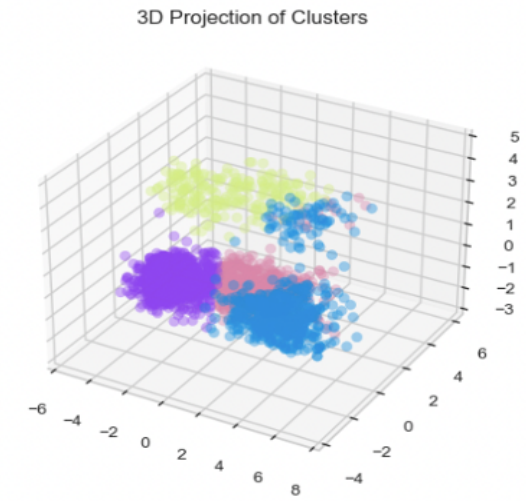


To find the optimal clusters in our data we opted Elbow method as it uses Sum of Squared Errors of points from its centroid. Fig 3 illustrates the elbow method where the distortion score the SSE score is calculated. The number of K is selected when it starts to kink in the elbow plot. Now as we got the number of clusters, we will fit our scaled data and predict on 4 number of clusters.

D. Validation of Results

Fig 4 shows how strongly clustered in the 3D projection. This depicts that selected optimal clusters. As we applied unsupervised learning, so we don't have tagged labels to the features. The purpose of our study is to study the different demographics in the clusters and provide inference according to that via exploratory data analysis (EDA).

Figure 4



V. FINDINGS

A. Which Customers have not accepted promotional campaigns so that unattracted campaigns could be better targeted to boost sales?

In our analysis, we got to know from the figure 5 that our valuable customers who have spent the most were from cluster 0. The least spent were from cluster 1, Cluster 2 were the second most spent and the Cluster 3 had average spent.

Figure 5

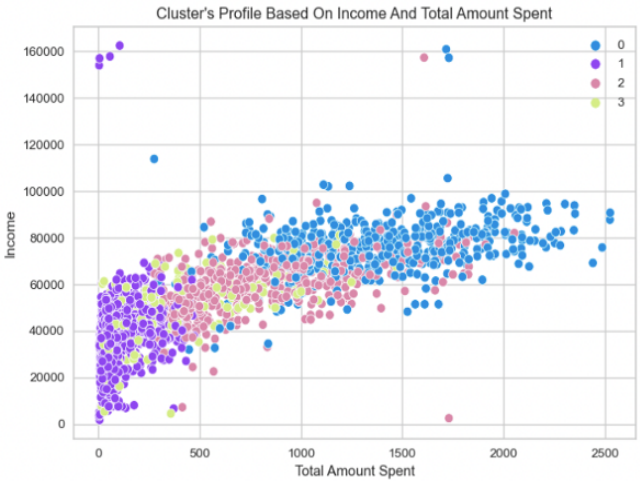
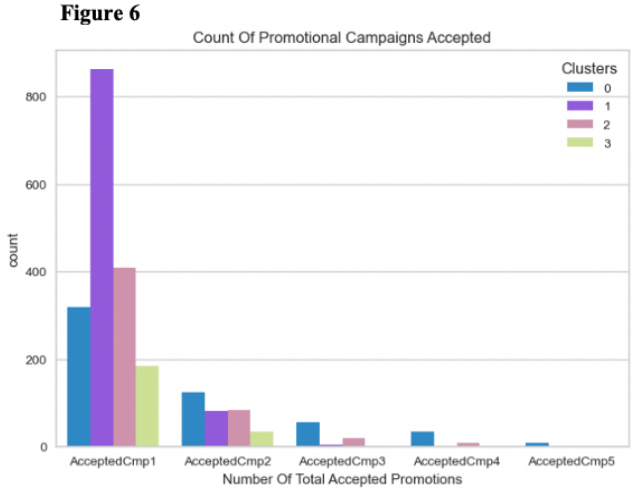


Fig 6 exhibits the total accepted promotional campaigns by all the customers in the Clusters. The plot shows that 1st promotional campaign was the successful from all the 5 campaigns and 2nd campaign onward it has seen rapidly decreasing all the way to 5th promotional campaign. Surprisingly, it shows that Customers from cluster 1 who spent the least were the one who accepted the 1st promotional campaign whole heartedly and the Cluster 0 who were our valuable customers were not that interested.

Figure 6

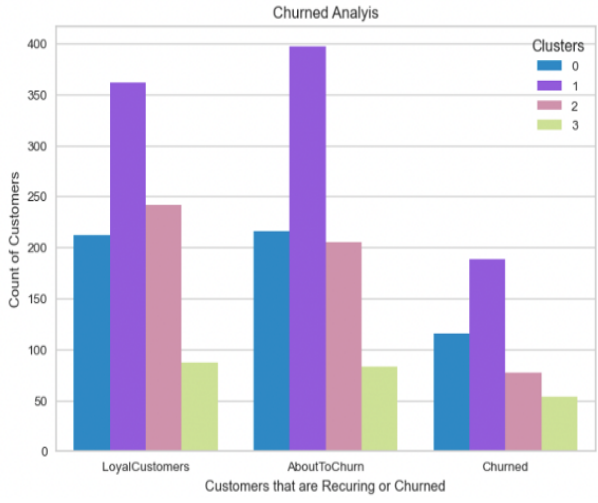


We can speculate by looking at the plot that the 1st promotional campaign might have the products that were inexpensive, therefore Cluster 1 customers were interested to buy. Now after all, we can infer that customers from Cluster 0 and Cluster 2 were only interested but in few numbers when we talk about campaigns such as 3 and 4 while campaign 5 only had few members of Cluster 0. So, the marketing team should reevaluate the short comings of the campaigns 2,3,4,5 and make them attractive in terms of price, products etc. to boost sales of the company.

B. Which customers have churned so that marketing team could come up with personalized offers for such customers?

For this analysis, we calculated the churned customers through the recency of a customer of buying a product. We divided the churned analysis in 3 groups by defining the range of days. For Loyal customers we opted between 0-30 days, about to churn group was ranging between more than 30 less than 50, thus after 50 days customers were grouped as churned.

Figure 7



From the Fig 7, it indicates that customers from Cluster 0 have higher percentage of churning and about to churn after Cluster 1.

This is alarming as the Cluster 0 as we have discussed were the star customers therefore, marketing team shall come up with personalized offers to target our star customers. Customers from Clusters 2 and 3 have also churned but in lower proportions whereas, Cluster 1 who have the spent the least have churned the most. Notably, from the plot it depicts that our most valuable consumers i.e., Cluster 0 were 3rd on the bar when considering loyalty in terms of recency of buying. This churned analysis would be good for the company to focus on star customers in terms of strategizing in creating personalized offers and sending them through emails, SMS etc. so that it enhances sales. Cluster 2 and 3 are also valuable to the company as they have shown in the table 1, spent high and average, respectively. They should also be focused as when forming offers as Cluster 2 and 3 have shown loyalty to the company.

C. Which group of products have low number of sales, or the products are high in demand for which deals, and discounts or adequate supplies are required, respectively, in order to increase their sales?

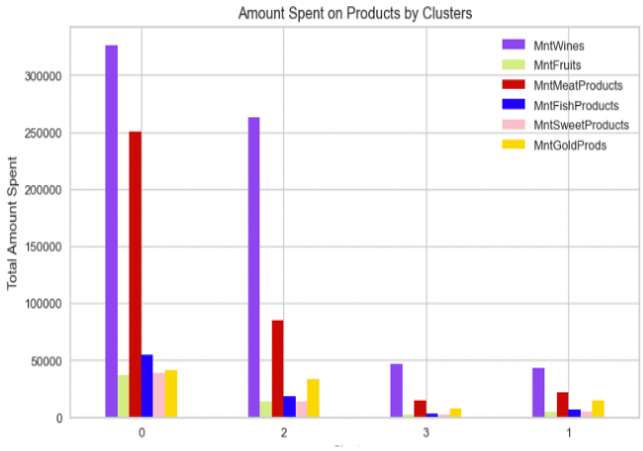
In our analysis, from figure 8 we can infer that product i.e., wine has the overwhelming response as maximum sale followed by meat and then the sales rapidly decline for other 3 products (fish, sweet and gold). From the table 1 we can also see that customers that are from Cluster 0 and are majority elderly and not a parent, are buying wines and meat in greater amount.

Table 1

Cluster 0	*Cluster 1*	*Cluster 2*	*Cluster 3*
Not a Parent	*Majority Parents*	*Parents*	*Parents*
-	*Majority Teen home*	*Majority Teen home*	*Majority Teen*
Majority Elderly	*Adults*	*Adults*	*Majority Elderly*
High Spent, High Income	*Low Spent, High Income*	*High Spent, High Income*	*Average Spent, Average Income*

Also, Cluster 2 follows the same patters as customers are adults and parents and are buying wines and meat in substantial amount. Meanwhile, Cluster 1 and 3 being the customers who didn't spend in considerable amount also bought wine and meat the most while neglecting other products. Thus, it infers that wine and meat products are high in demand and the company is required to store adequate supplies so there is no disruption in supplies of these hot selling products. Meanwhile, products such as fish, sweet and gold needs deals and discount so that customer feel these attractive to buy these products more often.

Figure 8



VI. REFLECTIONS, FURTHER WORK

We believe our study was exhaustive enough to answer all our analytical questions. But we also believe that other than some details in the features of data could've made the research more effective. For instance, if we would've known the prices of the products, we could rectify that the promotional campaigns accepted by non-valuable customers was due to in expensive products listed in the campaign. Furthermore, prices of products would have justified that why the fruit, gold and sweet products are not bought frequently.

VII. WORDCOUNT

Section	Expected	Actual
Abstract	150	74
Introduction	300	101
Analytical questions and data	300	142
Data (Materials)	300	186
Analysis	1000	653
Findings, reflections and further work	600	744
Total	2650	1900

VIII. References

[1] Jonker, Jedid-Jah, Nanda Piersma, and Dirk Van den Poel. "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability." Expert Systems with Applications 27, no. 2 (2004): 159-168.

[2] ‘Cooil, Bruce, Lerzan Aksoy, and Timothy L. Keiningham. "Approaches to customer segmentation." Journal of Relationship Marketing 6, no. 3-4 (2008): 9-39.

[3] Cuadros, Alvaro Julio, and Victoria Eugenia Domínguez. "Customer segmentation model based on value generation for marketing strategies formulation." Estudios Gerenciales 30, no. 130 (2014): 25-30.

[4] Teichert, Thorsten, Edlira Shehu, and Iwan von Wartburg. "Customer segmentation revisited: The case of the airline industry." Transportation Research Part A: Policy and Practice 42, no. 1 (2008): 227-242.