Faculty of Informatics, University of Debrecen, Hungary

Towards analyzing customers' buying pattern for better product recommendation using data mining techniques

Supervisor

Mr. Shubham Dubey

Doctoral Scholar, University of Debrecen, Hungary

Authors

Muhammad Waqas[1], Neptun(QX01OW)

Abdulrahaman Eltaweel[2], Neptun(HL8EY1)

Debrecen, 2021

'I agree to submit you for review'

Towards analyzing customers' buying pattern for better product recommendation using data mining techniques

Authors: Muhammad Waqas[1], Abdulrahaman Eltaweel[2]

Supervisor: Mr. Shubham Dubey, Doctoral scholar, Faculty of Informatics, University of Debrecen, Hungary

Abstract: This study is about customer personality analysis using data mining algorithms of clustering. It is a detailed analysis which can helps business and big brands to understand their customer better and helps them to modify their product according to the customer needs. In this paper author has conducted customer segmentation of grocery retailer data set and analyzed the data in python language. The finding will help to understand the customers' behavior such as their interest, lifestyle, spending and much more. Customer segmentation has been carried out with the help of K-Means algorithm. After the detailed analysis authors found that those who are earning more are also spending more. Based on income and total spending, customers are divided into four clusters i.e. 'Ordinary client, Elite client, Good client, Potential good client. From the mentioned groups, most of the customers fall into the Elite and Ordinary categories. The study is giving a future direction towards customer's interest mining to improve their experience. Also the future work can be done towards providing the better services in optimal cost. Several services providers which are platform as a service (PaaS) can implement the findings of this study especially data segmentation to strengthen their customer relation and service throughputs. This makes the study novel and significant for the future references.

Keywords: data mining, clustering, K-means algorithm, customer personality, data segmentation

1. Introduction

Market research is the most important phase for any company to increase its profit and analyzing the customer personality is the vital part of Market research (Akter, S. and Wamba, S. F. ,2016).

Big companies like Amazon and Alphabet spend billions of dollars on market research every year (Bowden, J. L.-H.,2009, Roberts, C. and Alpert, F. 2010). Data mining in the field of business has done miracle. A measurable portion of the business is affiliated with the analysis and deep insights (Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. 2018). Clustering, classification, association rule mining are the main area which deal with the customary data analysis (Olson, D. L., Shi, Y., & Shi, Y.,2007). Most brands don't use or store the data, so they no longer have any useful information about their products the minute they are shipped out of their factory and don't have any idea whether customers use their products correctly or not.

If brands had a deep insight into the consumers buying behavior of their products, they could improve or alter product characteristics according to consumer behavior. For example, if a coca cola company did the big data analysis of the stored data, and let's say, they get the idea that a high percentage of users only drank 70% of a given bottle, they could market smaller-sized bottles instead of bigger-sized ones.

Big data analysis does not only helps brands to save money but also save time, energy and help them to understand customer needs better. In this project we are going to analyze the customer data of a company to predict certain results such as the most frequent buyer etc. Applying big data analysis has many advantages for the business such as price optimization, getting more competitive edge, acquisition of optional customers and increased revenue. In this project we will use python to do customer segmentation. We are going to use grocery retailer data set. We will perform clustering of data on the customer's records that will give a useful insight so that we can improve revenue.

2. Related Work

Sandra C Matz and Oded Netzer conducted the research entitled" Using Big Data as a window into consumers' psychology" that predicting consumer behavior by using a big data analysis approach has created enormous opportunities for the researchers (Matz, S. C., & Netzer, O. 2017). The combination of information about 'what one does' with a great understanding of 'who one is' offers great opportunities to big brands. It not only boosts the effectiveness of marketing

campaigns but also helps customers to make better decisions (Matz, S. C., & Netzer, O. 2017, So, K. K. F., King, C., Sparks, B. A., and Wang, Y. 2016).

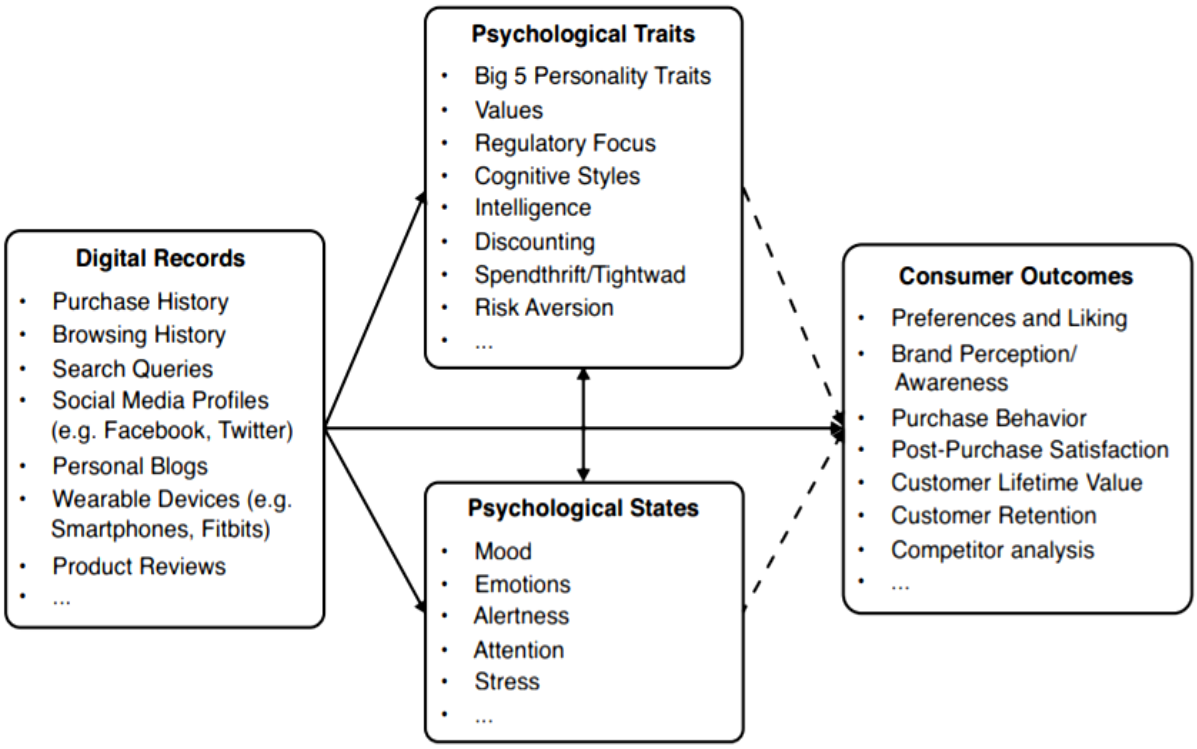

Figure 1. Leveraging Big data to infer psychological traits and states and affect customer behavior

The article titles as "Unlocking the power of big data in new product development" by Zhan, Y., Tan suggested that how big data can be used to enhance customer involvement in developing new products and they give the suggestions that it is necessary for brands to identify their truly needs by getting deep insight of consumer buying behavior (Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. , 2018).

Wedel and Kannan suggest a vital review of marketing analytics methods. They investigated different types of data (such as unstructured vs. structured); they discussed the importance of these methods, and gave the idea of current and future directions for new analytical tools for optimized marketing-mix spending (Wedel, M., & Kannan, P. K., 2016).

Van den Driest F, Sthanunathan S, Weed K, offer a discussion on how to successfully implement customer-focused methods to gain competitive advantage. They discussed the big brand Unilever,

and they gave the idea of operational characteristics to help the development of applications with great functional "insight engines" which were based on Big Data analysis van den Driest, F., Sthanunathan, S., & Weed, K.,2016).

3. Research Gap and Ideation of Solution

How can businesses get detailed insight into customer buying behavior? For example, how customers behave while deciding to buy a product which satisfies their needs. In the nutshell, how can a company improve or modify their product to increase its profit?

Customer personality analysis is a detailed analysis of consumer behavior. Customer personality analysis can help a company save billions of dollars. For example, instead of creating several marketing campaigns for a new product to every customer, a company can target the specific segment of customers by analyzing its own data. It decreases the chances of failure.

The goal of the project is to foresee the needs of customers, get to know their interests, lifestyles, priorities and learn their spending habits so that to maximize the value of customers to the business using data mining techniques.

4. Methodology

The method is about performing the exploratory data analysis with the help of customer segmentation. Customer segmentation will be carried out with the help of K-Means algorithm. The most amazing part of consumer behavior analyzing is getting answers to the most important questions such as:

1. What are the statistical characteristics of the customers?

2. What are the spending habits of the customers?

Let's do the analysis step by step:

A. Data preparation

In this step we will upload the data and will check the missing values and try to remove the missing values. We found out that the Income column has some missing data. Let's drop the rows in the data with missing values with simple code:

```
customer = customer.dropna()
```

B. Data engineering

There is a lot of information given in the dataset related to the customers. Here we will create the new features so that we can explore the data and can get meaningful insight. Here we will create following new features:

- Age of Customers
- Months Since Enrollment
- Total Spending
- Age Groups
- Number of Children
- Marital Status

Some examples of the codes are follow:

```
data['Age'] = 2021 - data.Year_Birth
```

```
data.Marital_Status = data.Marital_Status.replace({'Together': 'Partner',
                                                   'Married': 'Partner',
                                                  'Divorced': 'Single',
                                                   'Widow': 'Single',
                                                   'Alone': 'Single',
                                                'Absurd': 'Single',
                                                   'YOLO': 'Single'})
```

```
data.loc[(data['Age'] >= 10) & (data['Age'] <= 19), 'AgeGroup'] = 'Teen'

data.loc[(data['Age'] >= 20) & (data['Age']<= 39), 'AgeGroup'] = 'Adult'

data.loc[(data['Age'] >= 40), 'AgeGroup'] = 'Senior'
```

C. Exploratory Data analysis

Here we will get deep insight of the data by exploring different features with diagram. We analysis the following:

- Marital Status
- Education Level
- Child Status
- Average Spending on product
- Age Distribution of Customers

```
plt.figure(figsize=(8,6))

sns.histplot(data["Age"])

plt.title("Age Distribution")

plt.axvline(data["Age"].median(), color="yellow", label=f"Median of Age :
{data['Age'].median()}")

plt.axvline(data["Age"].mean(), color="red", label=f"Mean of Age : {data['Age'].mean()}");
```

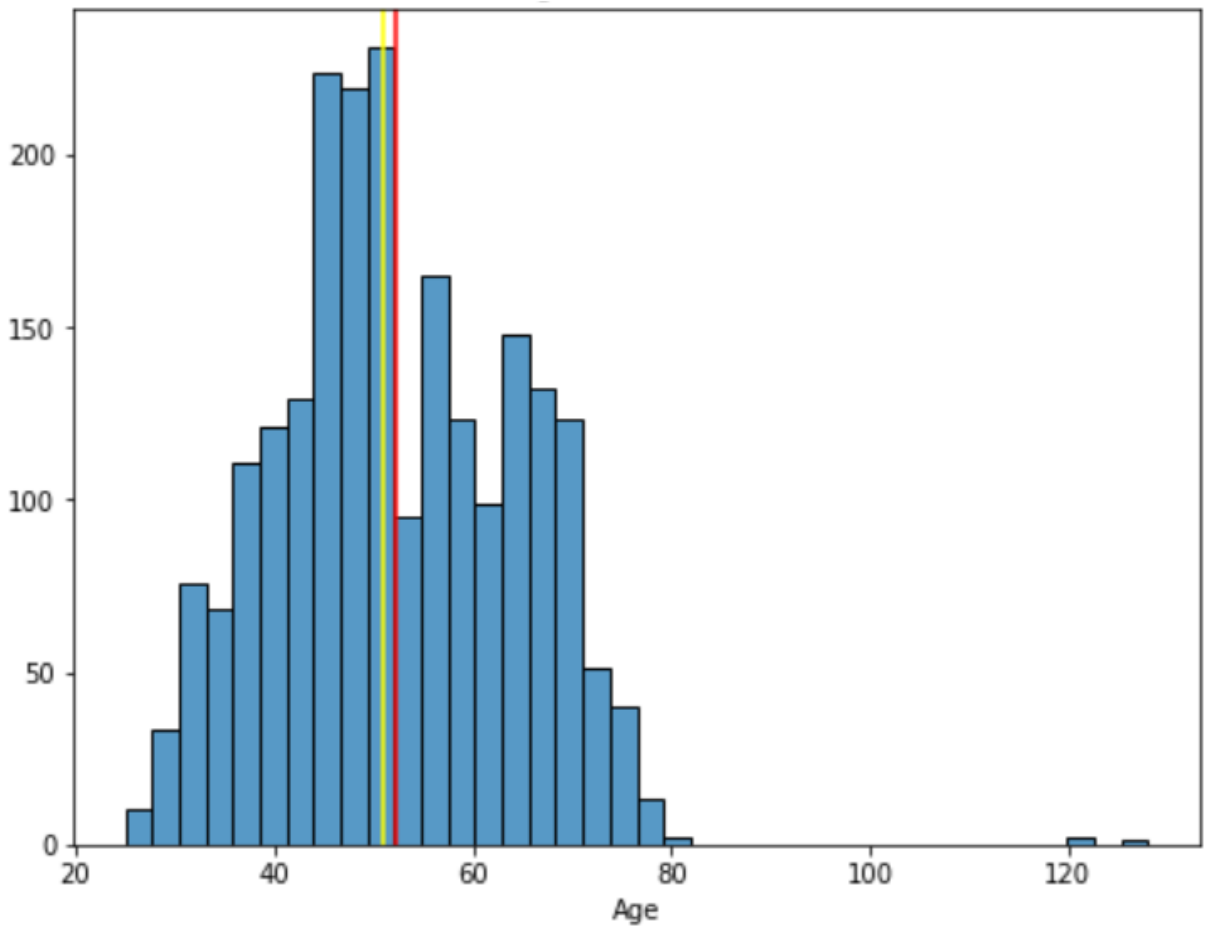

Figure 2. Age Distribution

```
labels = list(data["Education"].value_counts().index)
values = list(data["Education"].value_counts().values)
plt.figure(figsize=(5, 5))
plt.pie(values, labels=labels, autopct="%.2f")
plt.title("Education Level percentage");
```

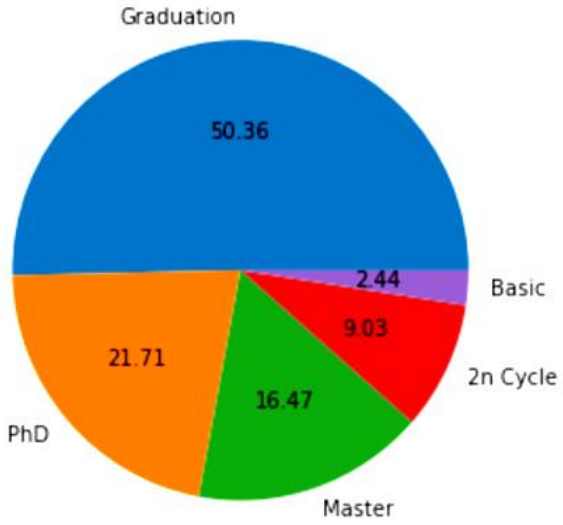

Figure 3. Customers from Several education level

```
plt.figure(figsize=(5,3))
sns.countplot(y="Marital_Status", data=data,
order=data["Marital_Status"].value_counts().index)
plt.title("Single vs Partner");
```

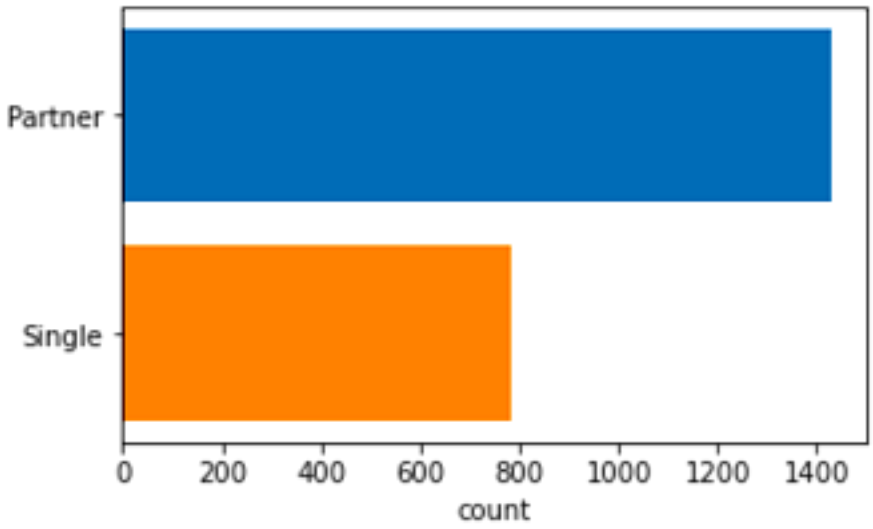

Figure 4. Martial status of customers

```
columns_pur = ["NumDealsPurchases", "NumWebPurchases", "NumCatalogPurchases",
"NumStorePurchases", "NumWebVisitsMonth"]
plt.figure(figsize=(10,10))
for i in range(1,6):
plt.subplot(2,3,i)
sns.histplot(data[columns_pur[i-1]], color="yellow", bins=14)
plt.title(columns_pur[i-1])
```

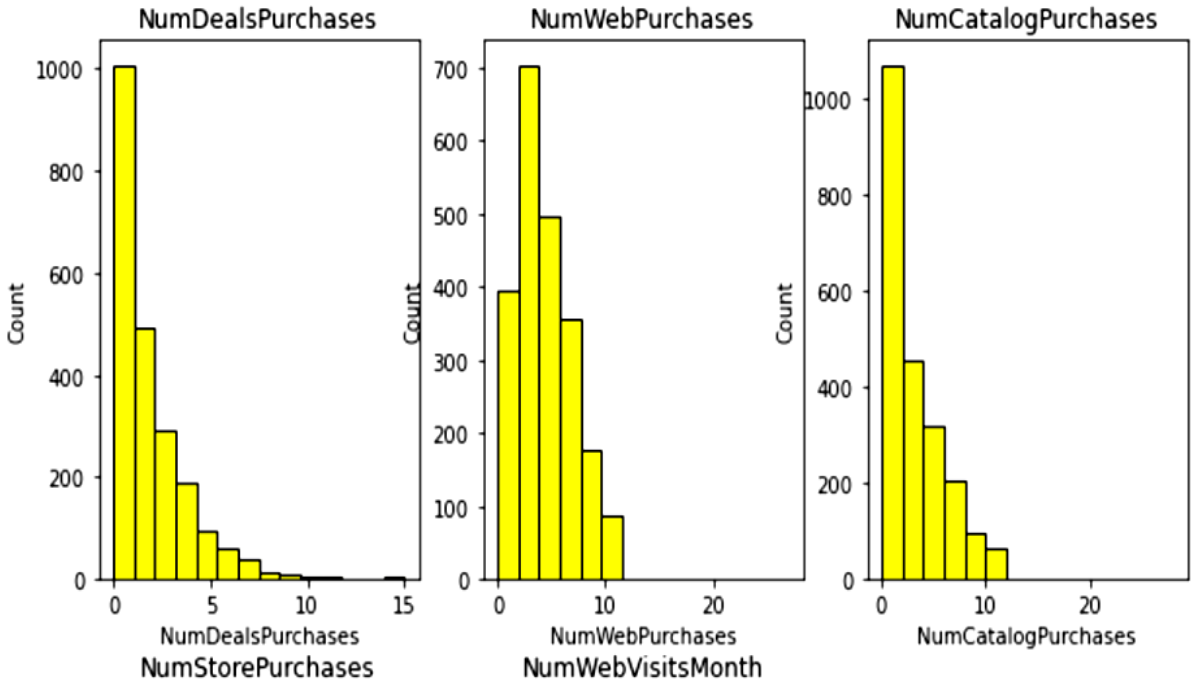

Figure 5. Customers' number of visits

D. Modeling clusters

Let's find out the different segments of the customers based on different features of the customers data using the K-Means Clusters. We dropped the unnecessary columns from the data. We have used Elbow method to find out number of cluster.

```
from sklearn.cluster import KMeans
```

```
range1 = range(2,9)

inertias = []

for n_clusters in options:

    model = KMeans(n_clusters, random_state=42).fit(col1)

inertias.append(model.inertia_)

plt.figure(figsize=(5,5))

plt.title("Number of clusters")

plt.plot(range1, inertias, '-1')

plt.xticks( fontsize=16)

plt.yticks( fontsize=16);
```

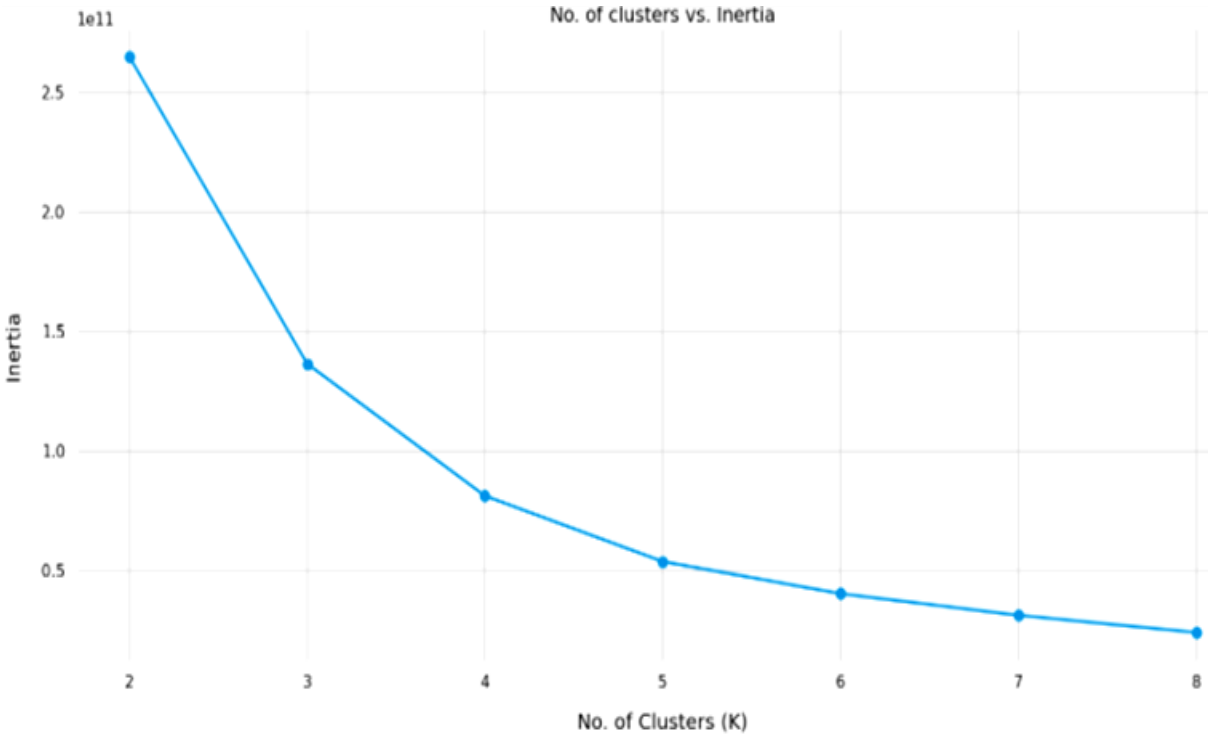

Figure 6. Cluster vs Inertia

```
model = KMeans(n_clusters=4, init='k-means++', random_state=52).fit(X)

preds = model.predict(X)

data_kmeans = X.copy()

data_kmeans['clusters'] = preds
```

Based on the above discussion, researchers have segmented the customers into 4 clusters, as the inertia value do not decrease much after 4 clusters.

From the analysis we can segment the customers into 4 groups based on their income and total spending:

Ordinary client: The one's with highest earnings and highest spending

Elite client: The one's with high earnings and high spending

Good client: The one's having low salary and less spending

Potential good client*:* The one's having lowest salary and least spending

```
data_kmeans.clusters = data_kmeans.clusters.replace({1: 'Ordinary client',

                                                     2: 'Elite client',

                                                     3: 'Good client',

                                                0: 'Potential good client'})


data['clusters'] = data_kmeans.clusters
```

E. Data exploring clusters based

Let's explore the data again based on the modelled clusters to identify the spending habits of the customers. Here we will analyze the following:

- Spending Habits by Clusters
- Purchasing Habits by Clusters

```
plt.figure(figsize=(10,10))

pl = sns.scatterplot(data = data,x=data["Total_spending"],
y=data["Income"],hue=data["clusters"], palette= pal)

pl.set_title("Cluster's Based On Income And Spending")

plt.legend()

plt.show()
```

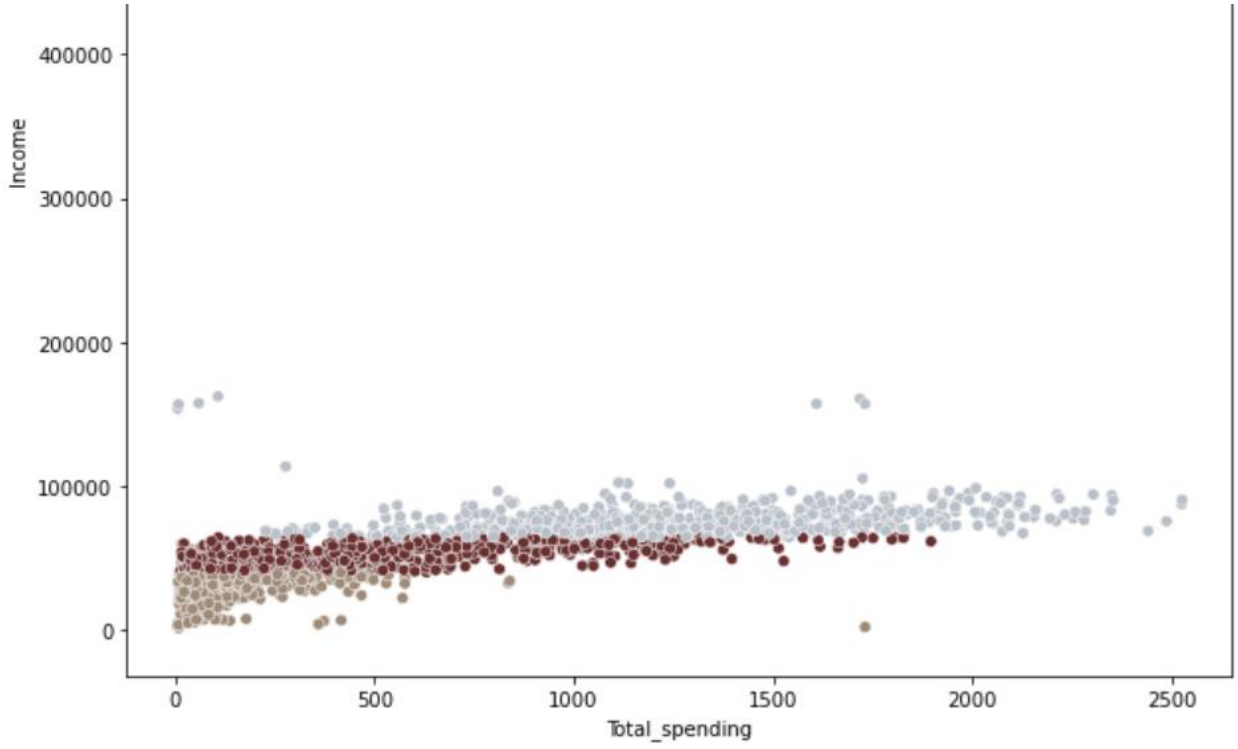

Figure 8. Customers' spending vs Income

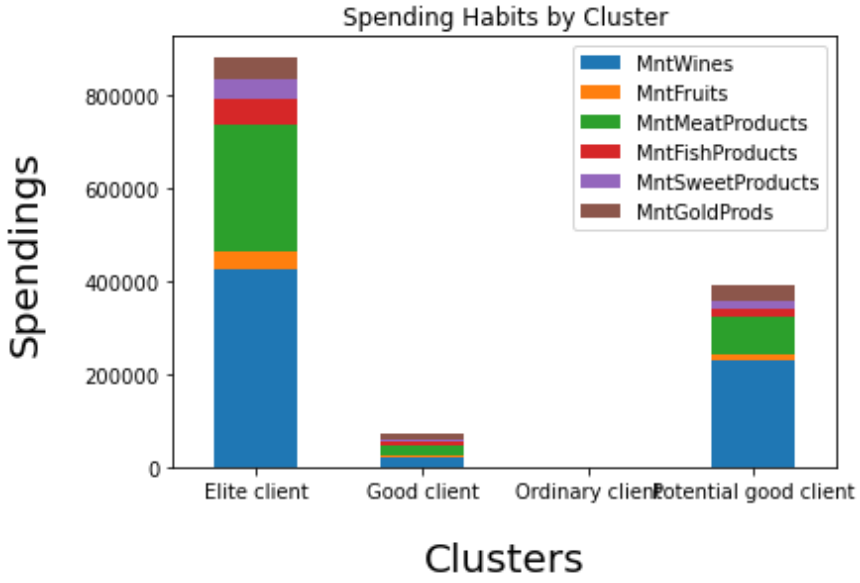

Figure 7. Spending habits by clusters

Insights: Customers from all the segments have spent most of their money on Wine and Meat products

5. Conclusions

Customers' interest mining is a crucial concern when it comes in business and management. This study was aiming to analyze the customers' buying patterns and their personal life. So that, the recommendation can provides best service with least effort and in minimal cost. This will give keep both customers and service providers in win-win conditions. According to the output and overall analysis conducted on this data science project on customer personality analysis with Python, we can conclude that most of the customers are university graduates and living with their partners. Middle-Aged Adults are spending on average, more than other age groups. Most of the customers are earning between 25000 and 85000 USD. Wine and Meat products are very famous among the customers. Based on income and total spending, customers are divided into four clusters i.e. 'Ordinary client, Elite client, Good client, Potential good client. Figure 7 depicts that the most of the customers fall into the Elite and Ordinary categories. Those who are earning more

are also spending more.Sweets and Fruits need some effective marketing. Company needs to run promotions for these products in order to increase the revenue from these products.

References:

1. Akter, S. and Wamba, S. F. (2016), "Big data analytics in E-commerce: a systematic review and agenda for future research", Electronic Markets, Vol. 26 No. 2, pp. 173-94.
2. Alexander III, L., Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. European Journal of Personality, 34(5), 632-648. (https://journals.sagepub.com/doi/abs/10.1002/per.2305)
3. Bosch-Sijtsema, P., & Bosch, J. (2015). User involvement throughout the innovation process in high-tech industries. Journal of Product Innovation Management, 32(5), 793–807.
4. Bowden, J. L.-H. (2009), "The process of customer engagement: A conceptual framework", Journal of Marketing Theory and Practice, Vol. 17 No. 1, pp. 63-74.
5. Matz, S. C., & Netzer, O. (2017). Using big data as a window into consumers' psychology. Current opinion in behavioral sciences, 18, 7-12.
6. Olson, D. L., Shi, Y., & Shi, Y. (2007). Introduction to business data mining (Vol. 10, pp. 2250-2254). New York: McGraw-Hill/Irwin.
7. Roberts, C. and Alpert, F. (2010), "Total customer engagement: designing and aligning key strategic elements to achieve growth", Journal of Product & Brand Management, Vol. 19 No. 3, pp. 198-209
8. So, K. K. F., King, C., Sparks, B. A., and Wang, Y. (2016), "Enhancing customer relationships with retail service brands: The role of customer engagement", Journal of Service Management, Vol. 27 No. 2, pp. 170-93.
9. Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best

customer profile cluster. In IOP Conference Series: Materials Science and Engineering (Vol. 336, No. 1, p. 012017). IOP Publishing.

10. van den Driest, F., Sthanunathan, S., & Weed, K. (2016). Building an insights engine. Harvard business review, 94(9), 15.
11. Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), 97-121.
12. Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. (2018). Unlocking the power of big data in new product development. Annals of Operations Research, 270(1), 577-595.