GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

CUSTOMER SEGMENTATION USING MACHINE LEARNING

A Report for the Evaluation 3 of Project 2

Submitted by

AMAN BANDUNI

(1613112005)

In partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION OF DATA ANALYTICS

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. ILAVENDHAN A. , M.Tech, Professor

April/May-2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "**CUSTOMER SEGMENTATION USING MACHINE LEARNING**" is the bonafide work of "**AMAN BANDUNI (1613112005)**" who carried out the project work under my supervision.

SIGNATURE OF HEAD

Professor & Dean,

School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr. ILavendhan A.

Professor

School of Computing Science & Engineering

TABLE OF CONTENTS

Abstract

The emergence of many competitors and entrepreneurs has created a great deal of tension between competing businesses to find new buyers and keep old ones. As a result of the preceding, the need for exceptional customer service becomes appropriate, regardless of the size of the business. In addition, the ability of any business to understand each of its customers' needs will receive greater support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service. Each segment contains customers who share similar market features. Big data ideas and machine learning have fostered more acceptance of the automated customer segmentation approach in favour of traditional market analytics that often do not work especially when the customer base is too large. In this paper, the k-Means clustering algorithm is used for this purpose. The sklearn liabrary was developed for the k-Means algorithm (found in the appendix) and the program is trained using a two-factor dataset of 100 patterns obtained from the retail business. Features of the average number of customer purchases and the average number of monthly customer visits.

List of Figures

LIST OF Abbreviations

ACRONYM	EXPANSION
SOM	Sorting Map
IT	Information Technology
ECM	Elbow Criterion Method
SSE	Sum of Squared Errors
AVG	Average
DF	Data Frame

1) Introduction

1.1 General

Over the years, the increasing competition between businesses and the availability of large-scale historical data has resulted in the extensive use of data mining techniques to discover important and strategic information that is hidden in the information of organizations. Data mining is the process of extracting logical information from a dataset and presenting it in a human-accessible way for decision support. Data mining techniques distinguish areas such as statistics, artificial intelligence, machine learning and data systems. Data mining applications include but are not limited to bioinformatics, weather forecasting, fraud detection, financial analysis and customer segmentation. The key to this paper is to identify customer segments in the commercial business using a data mining method. Customer division is the division of the customer base of the business into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. These distinctions are based on factors that can directly or indirectly influence the market or business such as product preferences or expectations, locations, behavior and so on. The importance of customer segmentation includes, inter alia, the ability of a business to customize market plans that will be appropriate for each segment of its customers; support for business decisions based on a risky environment such as debt relations with their customers; Identification of products related to individual components and how to manage demand and supply power; reveals the interdependence and interaction between consumers, between products, or between customers and products that the business may not be aware of; the ability to predict customer decline, and which customers are most likely to have problems and raise other market research questions and provide clues to finding solutions.

Integrated proved effective for detecting subtle but subtle patterns or relationships buried in a database of unencrypted data. This mode of learning is classified under supervised learning. Integration algorithms include the k-Means algorithm, k-nearest algorithm, Sorting Map (SOM) and more. These algorithms, without prior knowledge of the data, are able to identify clusters in them by repeated comparisons of input patterns until stable qualifications in the training examples are obtained depending on the subject matter or the process. Each set contains data points that have very close similarities but vary greatly from the data points of other clusters. Integration has great applications in pattern recognition, image analysis, and bioinformatics and so on. In this paper, the k-Means clustering algorithm was applied to the customer segment. The sklearn library (Appendix) of the k-Means algorithm was developed, and the training was started using a standard Silhouette -score with two feature sets of 100 training patterns found in the retail business. After numerous indications, four stable intervals or customer segments were identified. Two factors are considered in the combination of the number of goods purchased by the customer per month and the average number of customer visits per month. From the dataset, four customers or categories are grouped and labeled as follows: cluster_metrics_1, cluster_metrics_2, cluster_metrics_3, cluster_metrics_4.

1.2 Pupose

In this series of Data Science Project, we will make one of the most important applications of machine learning - Customer Classification. For this project, we will use client components in python. Whenever you need to find your best customer, customer division is the best option. For this machine learning project, this project will provide you with a background for customer segmentation. After that we will evaluate the data from which we will build the classification model. Also, in this data science project, we will see a descriptive analysis of our data and use several types of K-means algorithm. So, follow the complete customer science project in the segment using an algorithm learning machine and run in python. Customer Classification is one of the most important applications of unreadable learning. Using merger strategies, companies can identify customer segments that allow them to identify user bases. For this machine learning project, we will use K-methods integration which is an important algorithm for integrating an unlisted dataset.

2) Literature Survey

A. Customer Classification

Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and wants of their customers, attract new customers, and thus improve their businesses. The task of identifying and meeting the needs and requirements of each customer in the business is a very difficult task. This is because customers may vary according to their needs, wants, demographics, shapes, taste and taste, features and so on. As it is, it is a bad practice to treat all customers equally in business. This challenge has led to the adoption of the concept of customer segmentation or market segmentation, where consumers are divided into subgroups or segments where members of each subcategory exhibit similar market behaviors or features. Accordingly, customer segmentation is the process of dividing the market into indigenous groups.

B. Big Data

Recently, Big Data research has gained momentum. defines big data as - a term that describes a large number of formal and informal data, which cannot be analyzed using traditional methods and algorithms. Companies include billions of data about their customers, suppliers, and operations, and millions of internally connected sensors are sent to the real world on devices such as mobile phones and cars, sensing, creating, and communicating data. the ability to improve forecasting, save money, increase efficiency and improve decision-making in various fields such as traffic control, weather forecasting, disaster prevention, finance, fraud control, business transactions, national security, education, and healthcare. Big data is seen mainly in the three Vs namely: volume, variability and speed. There are other 2Vs available - authenticity and value, thus making it 5V.

C. Data Collection

Data collection is the process of collecting and measuring information against targeted variations in an established system, enabling one to answer relevant questions and evaluate results. Data collection is part of research in all fields of study including physical and social sciences, humanities and business. The purpose of all data collection is to obtain quality evidence that allows analysis to lead to the creation of convincing and misleading answers to the questions submitted. We collected data from the UCI Machine Learning Repository.

D. Clustering data

Clustering is the process of grouping the information in the dataset based on some similarities. There are a number of algorithms which can be chosen to be applied on a dataset based on the situation provided. However, no universal clustering algorithm exists that's why it becomes important to opt for appropriate clustering techniques. In this paper, we have implemented three clustering algorithms using python sklearn library.

E. K-Mean

K- means that an algorithm is one of the most popular classification algorithm. This clustering algorithm depends on the centroid where each data point is placed in one of the overlapping K clusters pre-programmed into the algorithm, The clusters are created that correspond to the hidden pattern in the data that provides the information needed to help decide the execution process. There are many ways to make k-means assembling; we will use the elbow method.

3) Methodology

The data used in this paper were collected from the UCI Machine Learning Repository. This is a set of geographic data containing all transactions occurring between 1/1/2/10 and 9/12/2011 in an unregistered and unregistered UK broker. The company mainly sells unique gifts all together. Many of the company's customers are shopkeepers. The database contains 8 attributes. These attributes include:

“InvoiceNo: Invoice number. By default, a 6-digit aggregate number is assigned separately for each transaction. If this code starts with the letter 'c', it indicates the cancellation. ”

StockCode Code: Product (item). Name, a 5-digit number assigned only to each unique product. ”

“Definition: Product name (item). By name. ”

“Price: The value of each product (item) made. Number. "

“InvoiceDate: Invitation Date and Time. In terms of numbers, the date and time of each transaction. ”

“UnitPrice: Price is a unit. Prices, product price per unit of measurement. "

“Customer: Customer number. Name, 5-digit number assigned to each customer. ”

Country: Country name. Name, the name of the country where each customer lives. ”

In this paper several steps were taken to obtain an accurate result. It involves the addition of a feature alongside the first step of the centroids, the allocation step and the update step, which are the most common steps k-means algorithms.

A. Collect data

This is a data preparation phase. The feature usually helps to refine all data items at a standard rate to improve the performance of the clustering algorithm. Each data point changes from grade 2 to +2. Integration techniques that include Min-max, decimal and z-points The standard z-signing strategy are used to make things unequal before applying the k-Means algorithm to a dataset.

B. Customer Classification Methods

There are many ways to perform segmentation, which vary in severity, data requirements, and purpose. The following are some of the most commonly used methods, but this is not an incomplete list. There are papers that discuss artificial neural networks, particle fixation, and complex types of ensemble, but are not included due to limited exposure. In future articles, I may go into some of these alternatives, but for now, these more common methods should be sufficient.

Each subsequent section of this article will include a basic description of the method, as well as a code example for the method used. If you don't have the expertise, well, just skip the code and you'll still have to get a good handle on each of the 4 sub-sections we include in this article.

C. Group Analysis

Group analysis is a unifying, or unifying, approach for consumers based on their similarities.

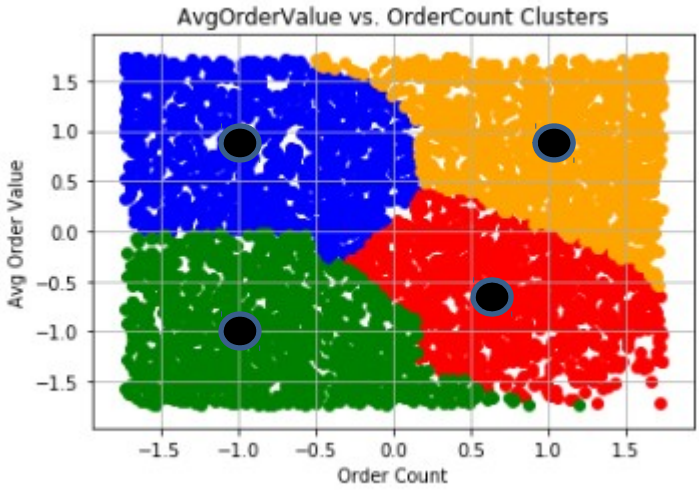
There are 2 main types of group analysis categorized into market policy: Hierarchical group analysis, and classification (Miller, 2015). In the meantime, we will discuss how to classify clusters called k-methods.

D. K-means encounter

The k-means clustering algorithm is an algorithm that is frequently used to draw insights into the formats and differences within a database. In marketing, it is often used to build customer segments and to understand the behavior of these unique segments. Let's get into building assembly models in the python environment.

E. Centroids Initiation

Selected cents or initials were selected. Figure 1 introduces the start of graduation centers. Four selected centers shown in different shapes were selected using the Forgy method. In Forgy's method of using k (in this case k = 4) data points are randomly selected as cluster centroids.



Technical Introduction

The code below was created in the Jupyter manual using Python 3.x and a few Python packages for editing, processing, analyzing, and visualizing information.

Most of the code below comes from the GitHub package of the book Hands-On Data Science for Marketing. The book is available on Amazon or OilReilly if you are a subscriber.

The open source data used in the following code comes from Cost Irvine's Machine Learning Repository.

4) Implementation

Import packages and data

To start, we import the packages needed to do our analysis and then import the xlsx (excel spreadsheet) data file. If you want to follow along with the same data, you'll need to download it from UCI. For this example, I put the xlsx file in the folder (directory) where I present Jupyter's notebook.

In [7]:
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# read-in the excel spreadsheet using pandas
df = pd.read_excel("C:\\Users\\aman banduni\\Desktop\\final project\\New try\\Online Retail.xlsx")

df.head() # take a look at the first 5 rows in the DataFrame
```

Out[7]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

As you can see, we have 8 columns of data for each row and each row represents an item purchased. This isn't that helpful yet, so let's clean and organize this data in a way that allows us to formulate more actionable insights.

Data cleaning

Below, we will remove data that is not helpful, missing, or potentially cause issues in the long run.

In [30]:
```
# Drop cancelled orders
df = df.loc[df['Quantity'] > 0]

# Drop records without CustomerID
df = df[pd.notnull(df['CustomerID'])]

# Drop incomplete month
df = df.loc[df['InvoiceDate'] < '2011-12-01']

# Calculate total sales from the Quantity and UnitPrice
df['Sales'] = df['Quantity'] * df['UnitPrice']
```

Now let's convert the data so that each record represents one customer purchase history.

In [10]:
```
# use groupby to aggregate sales by CustomerID
customer_df = df.groupby('CustomerID').agg({'Sales': sum,
                                            'InvoiceNo': lambda x: x.nunique()})

# Select the columns we want to use
customer_df.columns = ['TotalSales', 'OrderCount']

# create a new column 'AvgOrderValu'
customer_df['AvgOrderValue'] = customer_df['TotalSales'] / customer_df['OrderCount']

customer_df.head()
```

Out[10]:

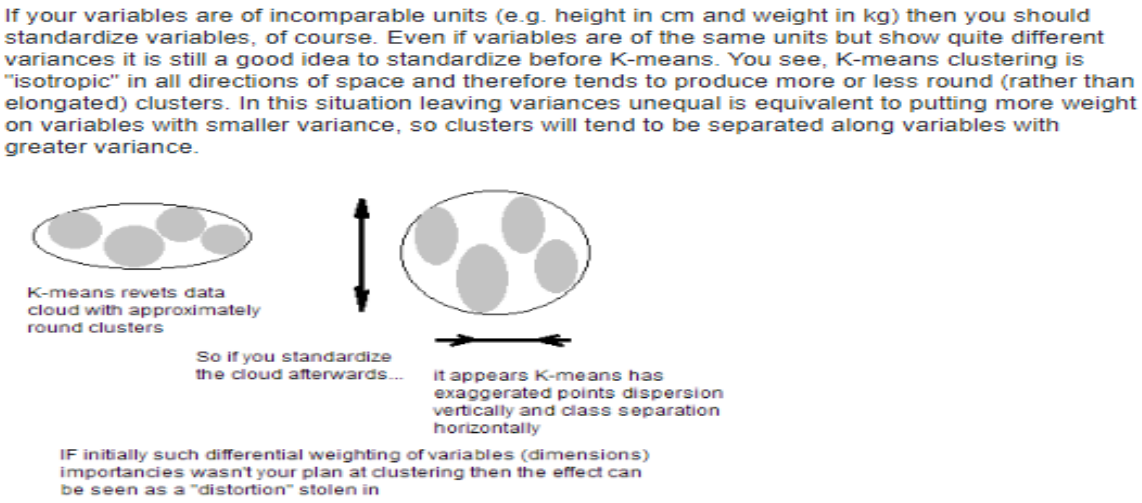
CustomerID	TotalSales	OrderCount	AvgOrderValue
12346.0	77183.60	1	77183.600000
12347.0	4085.18	6	680.863333
12348.0	1797.24	4	449.310000
12349.0	1757.55	1	1757.550000
12350.0	334.40	1	334.400000

We now have a DataFrame with complete sales, order counts, and average order price per customer. But right now we're not home.

Normalize the data

Clustering algorithms like K-means are sensitive to the scales of the data used, so we'll want to normalize the data.

Below is a screenshot from part of a StackExchange answer discussing why standardization or normalization is necessary for data used in K-means clustering. The screenshot is linked to the StackExchange question, so you can click on it and read the entirety of the discussion if you'd like more information.



In [11]:
```
rank_df = customer_df.rank(method='first')
normalized_df = (rank_df - rank_df.mean()) / rank_df.std()
normalized_df.head(10)
```

Out[11]:

CustomerID	TotalSales	OrderCount	AvgOrderValue
12346.0	1.724999	-1.731446	1.731446
12347.0	1.457445	1.064173	1.401033
12348.0	0.967466	0.573388	0.929590
12349.0	0.944096	-1.730641	1.683093
12350.0	-0.732148	-1.729835	0.331622
12352.0	1.193114	1.309162	0.169639
12353.0	-1.636352	-1.729029	-1.570269
12354.0	0.508917	-1.728223	1.612981
12355.0	-0.386422	-1.727417	0.970690
12356.0	1.268868	0.158357	1.557375

Our data is scaled between -2 and 2. Now let's get to clustering.

Select the optimal number of clusters

Alright, we're ready to run cluster analysis. But first, we need to figure out how many clusters we want to use. There are several approaches to selecting the number of clusters to use, but I'm going to cover two in this article: (1) silhouette coefficient, and (2) the elbow method.

Silhouette (Clustering)

Silhouette means how to interpret and verify consistency within data structures. This method provides a picture showing how well each item is organized. [1]

The value of a silhouette is a measure of how something is similar in its collections (combinations) compared to other clusters (divisions). The silhouette goes from –1 to +1, where a higher value indicates that an item is properly matched to its collection and compared to neighboring clusters. If multiple objects have a high value, then the integration configuration is appropriate. If most points have a value or a negative value, then the coordinate system may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as Euclidean distance or Manhattan distance.

Now that we know a whole lot more of the silhouette, let's go in and use the code to find the right number of clusters.

In [12]:
```
# Use silhouette coefficient to determine the best number of clusters
from sklearn.metrics import silhouette_score

for n_cluster in [4,5,6,7,8]:
    kmeans = KMeans(n_clusters=n_cluster).fit(
        normalized_df[['TotalSales', 'OrderCount', 'AvgOrderValue']])

    silhouette_avg = silhouette_score(
        normalized_df[['TotalSales', 'OrderCount', 'AvgOrderValue']],
        kmeans.labels_)

    print('Silhouette Score for %i Clusters: %0.4f' % (n_cluster, silhouette_avg))
```

```
Silhouette Score for 4 Clusters: 0.4114
Silhouette Score for 5 Clusters: 0.3773
Silhouette Score for 6 Clusters: 0.3785
Silhouette Score for 7 Clusters: 0.3913
Silhouette Score for 8 Clusters: 0.3810
```

Cluster 4 had the most complete silhouette fit, indicating that 4 could be the best number of clusters. But we'll look at that twice with the elbow way.

Elbow Criterion Method (with the Sum of Squared Errors (SSE)):

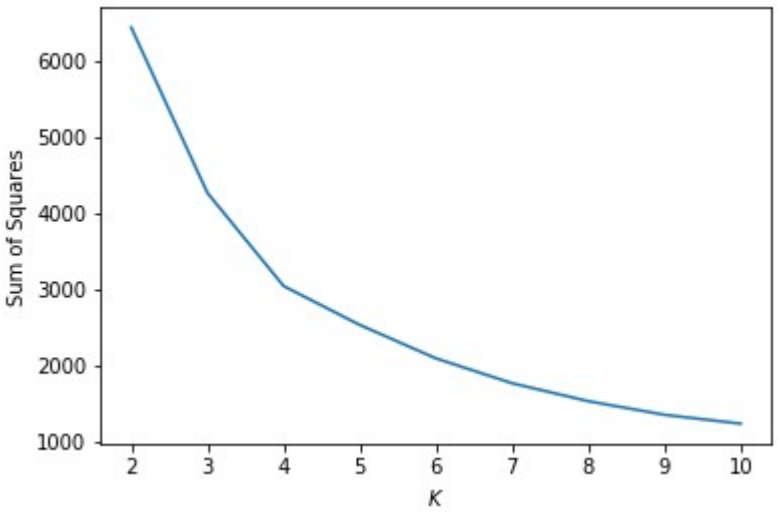
The idea behind the elbow method is to run the k-mean correlation in the given data for k values (num_clusters, e.g. k = 1 to 10), and for each k value, to calculate the sum of squared errors (SSE).

After that, adjust the SSE line for each k value. If the line graph looks like an arm - a red circle below the line of the line (as an angle), the "elbow" on the arm is the correct price (collection value). Here, we want to reduce the SSE. SSE usually drops to 0 as we go up k (and SSE is 0 where k equals the number of data points, because where each data point is its own set, and there is no error between it and its trunk).

Therefore the purpose is to select a small value of k that still has a low SSE, and the cone usually represents where it starts to have a negative return with increasing k.

Well, with the correct understanding of the elbow mechanism in hand, let's use the elbow method to see if it agrees with our previous results suggesting 4 sets.

In [13]:
```
from sklearn import cluster
import numpy as np

sse = []
krange = list(range(2,11))
X = normalized_df[['TotalSales','OrderCount','AvgOrderValue']].values
for n in krange:
    model = cluster.KMeans(n_clusters=n, random_state=3)
    model.fit_predict(X)
    cluster_assignments = model.labels_
    centers = model.cluster_centers_
    sse.append(np.sum((X - centers[cluster_assignments]) ** 2))

plt.plot(krange, sse)
plt.xlabel("$K$")
plt.ylabel("Sum of Squares")
plt.show()
```



Based on the graph above, it looks like K = 4, or 4 clusters is the correct number of clusters in this analysis. Now let's translate the customer segments provided by these components.

Interpreting Customer Segments

In [14]:
```
kmeans = KMeans(n_clusters=4).fit(normalized_df[['TotalSales', 'OrderCount', 'AvgOrderValue']])

four_cluster_df = normalized_df[['TotalSales', 'OrderCount', 'AvgOrderValue']].copy(deep=True)
four_cluster_df['Cluster'] = kmeans.labels_

four_cluster_df.head(10)
```

Out[14]:

CustomerID	TotalSales	OrderCount	AvgOrderValue	Cluster
12346.0	1.724999	-1.731446	1.731446	0
12347.0	1.457445	1.064173	1.401033	2
12348.0	0.967466	0.573388	0.929590	2
12349.0	0.944096	-1.730641	1.683093	0
12350.0	-0.732148	-1.729835	0.331622	0
12352.0	1.193114	1.309162	0.169639	2
12353.0	-1.636352	-1.729029	-1.570269	3
12354.0	0.508917	-1.728223	1.612981	0
12355.0	-0.386422	-1.727417	0.970690	0
12356.0	1.268868	0.158357	1.557375	2

Now let's combine the metrics of the integration and see what we can gather from the standard data for each cluster.

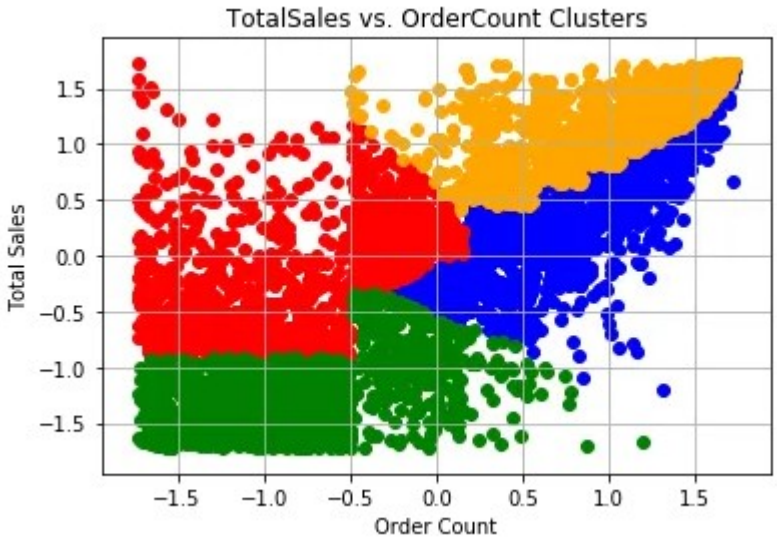
In [15]:
```
cluster1_metrics = kmeans.cluster_centers_[0]
cluster2_metrics = kmeans.cluster_centers_[1]
cluster3_metrics = kmeans.cluster_centers_[2]
cluster4_metrics = kmeans.cluster_centers_[3]

data = [cluster1_metrics, cluster2_metrics, cluster3_metrics, cluster4_metrics]
cluster_center_df = pd.DataFrame(data)

cluster_center_df.columns = four_cluster_df.columns[0:3]
cluster_center_df
```

Out[15]:

	TotalSales	OrderCount	AvgOrderValue
0	-0.134009	-0.847864	0.794788
1	0.243849	0.740432	-0.640950
2	1.205607	1.000922	0.876846
3	-1.234513	-0.784431	-1.055606

In the following section, we will visualize the clustering by adding different columns to the x and y axes. Let's see what we say.

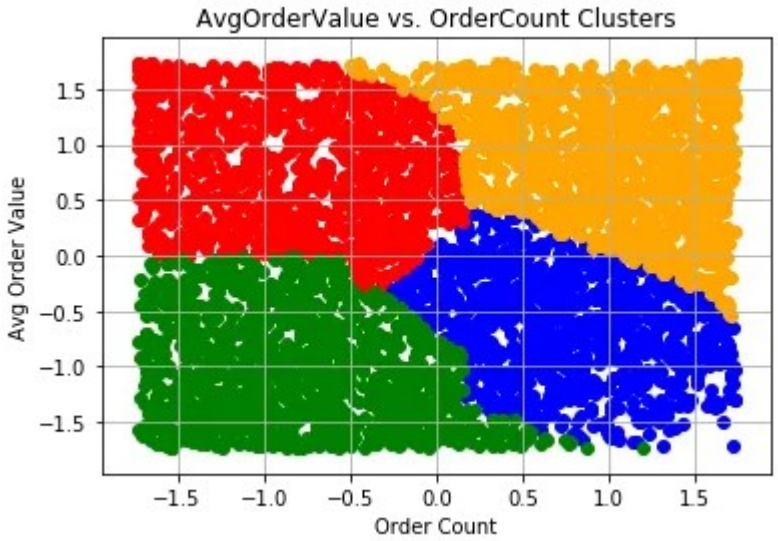
In [28]:
```
plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['TotalSales'],
    c='blue')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['TotalSales'],
    c='red')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['TotalSales'],
    c='orange')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['TotalSales'],
    c='green')

plt.title('TotalSales vs. OrderCount Clusters')
plt.xlabel('Order Count')
plt.ylabel('Total Sales')

plt.grid()
plt.show()
```



Green customers have the lowest price and lowest order count, which means they are the lowest bidder. On the other hand, orange customers have the highest total SALE and highest order count, indicating that they are the highest

priced customers.

In [29]:
```
plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['AvgOrderValue'],
    c='blue')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['AvgOrderValue'],
    c='red')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['AvgOrderValue'],
    c='orange')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['OrderCount'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['AvgOrderValue'],
    c='green')

plt.title('AvgOrderValue vs. OrderCount Clusters')
plt.xlabel('Order Count')
plt.ylabel('Avg Order Value')

plt.grid()
plt.show()
```



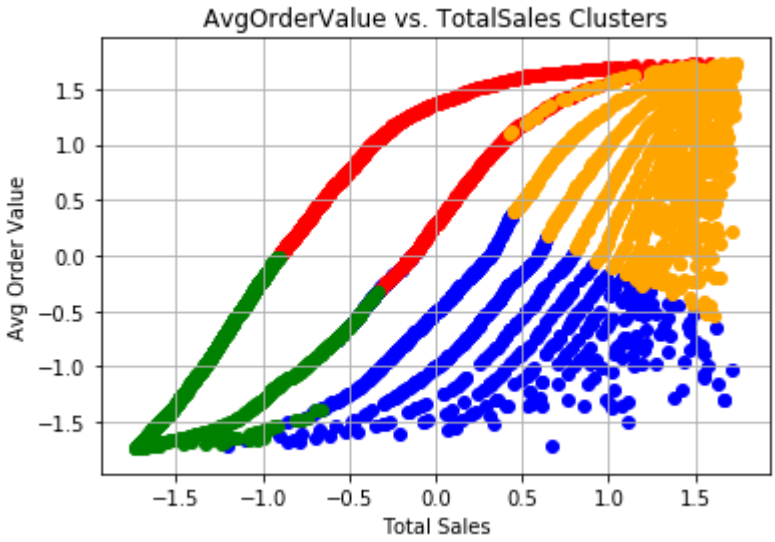
In this structure, we consider the average order value vs the order value. Once again, green buyers are the lowest price and the customers in the orange are the highest prices.

You can look at it this way. You can target customers in red graphics and try to find ways to increase their order count through email reminders or SMS notifications directed to other identification features. Maybe you can email them

a discount if they come back within 30 days. Ideally, you can provide a delayed coupon (which will be used at some point) at checkout.

Similarly, with customers who are in the blue segment, you may want to try other sales and marketing strategies for the cart. Probably the fastest offer, based on market basket analysis (see section on market basket analysis below).

In [19]:
```
plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['TotalSales'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 0]['AvgOrderValue'],
    c='blue')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['TotalSales'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 1]['AvgOrderValue'],
    c='red')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['TotalSales'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 2]['AvgOrderValue'],
    c='orange')

plt.scatter(
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['TotalSales'],
    four_cluster_df.loc[four_cluster_df['Cluster'] == 3]['AvgOrderValue'],
    c='green')

plt.title('AvgOrderValue vs. TotalSales Clusters')
plt.xlabel('Total Sales')
plt.ylabel('Avg Order Value')

plt.grid()
plt.show()
```



In this building, it has a median value and order compared to the total retail price. This structure also strengthens the previous 2 sites in identifying the orange group as the highest value customers, the green as the lowest priced customers, and the blue and red as the high potential customers.

From a growth perspective, I focus my attention on the blue and red collection. I try to better understand each encounter and their intelligent behavior on site to identify which team to focus on first and introduce a few test cycles.

The Best-selling item by segment

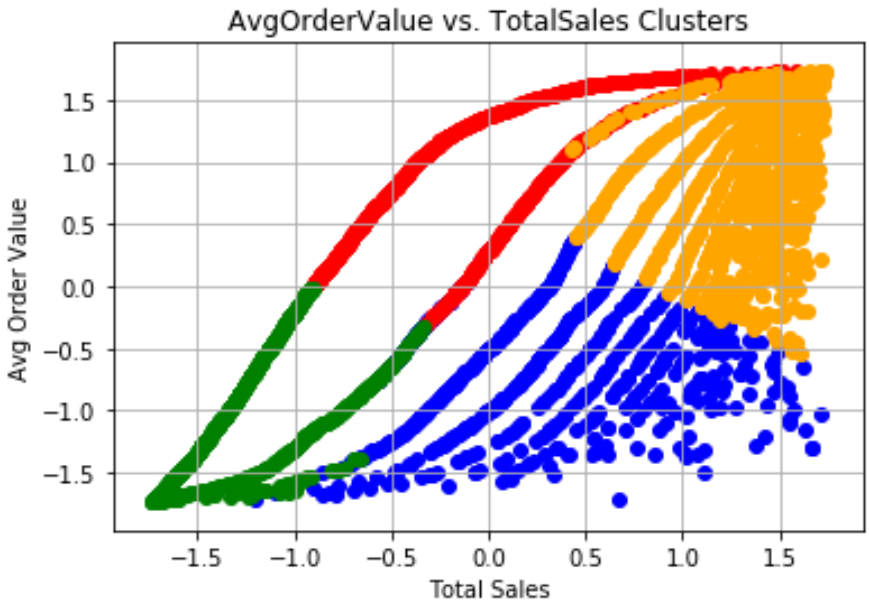
We know we have 4 categories and we know how much they spend on each purchase, their total usage, and the number of their orders. The next thing we can do is to help us better understand customer segments to find out which items are best sold in each segment.

```
In [23]: high_value_cluster = four_cluster_df.loc[four_cluster_df['Cluster'] == 2]

pd.DataFrame(df.loc[df['CustomerID'].isin(high_value_cluster.index)].groupby(
    'Description').count()['StockCode'].sort_values(ascending=False).head())
```

Description	StockCode
JUMBO BAG RED RETROSPOT	1129
REGENCY CAKESTAND 3 TIER	1080
WHITE HANGING HEART T-LIGHT HOLDER	1062
LUNCH BAG RED RETROSPOT	924
PARTY BUNTING	859

Based on this information, we now know that the Jumbo Bag Red Retrospot is the best-selling item by our most expensive team. With that information available, we can make recommendations for other potential customers in this section.

5) Result



Here, the result states that the orange cluster as the highest value customers, green as the lowest value customers, and the blue and red as high opportunity customers.

Description	StockCode
JUMBO BAG RED RETROSPOT	1129
REGENCY CAKESTAND 3 TIER	1080
WHITE HANGING HEART T-LIGHT HOLDER	1062
LUNCH BAG RED RETROSPOT	924
PARTY BUNTING	859

Result also concludes that the Jumbo Bag Red Retrospot is the best-selling item.

6) Conclusion

As our dataset was unlabelled, in this paper we have opted for internal clustering validation rather than external clustering validation, which depends on some external data like labels. Internal cluster validation can be used for choosing clustering algorithm which best suits the dataset and can correctly cluster data into its opposite cluster.

Customer segmentation can have a positive impact on a business if done properly.

Hence we can give special discounts or gift vouchers to the people of orange clusters to retain them for long and for people in blue and red cluster we can give discounts and do advertisement of highly selling objects to attract them, and for the low value people which are in green clusters, we can arrange feedback column to know what we can change to attract them as well.

Based on the above information, we now know that the Jumbo Bag Red Retrospot is the best-selling item by our most expensive team. With that information available, we can make recommendations for other potential customers in this section.

7) Hardware and software package Specifications

Hardware necessities

Hardware choice is essential to the standard and potency of any software package. In Hardware choice, size and power necessities are necessary.

Customer isolation will be with success run on the system with AN i3 processor with a minimum of four GB RAM and disc drive with 500GB and fifteen.6 inches to observe system performance. (Printer is needed for text output).

• Pentium processor -------- two GHz or on top of
• RAM capability -------- four GB
• Hard Disk -------- five hundred GB

Software necessities

One of the foremost troublesome tasks is, software package choice, as long because the would like for the program is thought to search out out if a specific software package package fits the wants. once the primary choice of alternatives safety is needed to urge the need for a few software package compared to the opposite candidates. This section initial summarizes the application's question so proposes an in depth comparison.

• Operating System :: Windows seven or ten
• Software: Jupyter Notebook
• Databases::Excel sheets
• Python Libraries
• Packages

8) References

- Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited
- Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.
- Hill, K. (2012, February 16). How Target found out that a teenage girl was pregnant prior to her father's pregnancy. Forbes.com. Retrieved from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl- was pregnant- before her father-in-law /
- Hong, T., Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. Expert System Applications, 39 (2), 2127-2131.
- Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited
- Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.
- Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.
- Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011
- By Jerry W Thomas. 2007. Accessed at: www.decisionanalyst.com on July 12, 2015.
- T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. AnuManufactured cluster analysis using a new algorithm from structured and unstructured

data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.

- McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015.
- Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf July 14, 2015.
- A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.
- Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814
- Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom Customer Classification Based on Group Analysis of K-methods", JIRCCE, Year: 2015.
- Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm”, IJCSI,Year: 2011.