Capstone Project

Online Retail Customer Segmentation

Name- Pranav Singhal

Batch-AlmaBetter Pro

Github- https://github.com/pranav168/Customer-Segmentation

Why Customer Segmentation?

Segmentation allows businesses to make better use of their marketing budgets, gain a competitive edge over rival companies and, importantly, demonstrate a better knowledge of your customers' needs and wants.

• Marketing efficiency – Breaking down a large customer base into more manageable pieces, making it easier to identify your target audience and launch campaigns Determine new market opportunities Better brand strategy Improve distribution strategies Customer retention.

•Determine new market opportunities – During the process of grouping your customers into clusters, you may find that you have identified a new market segment, which could in turn alter your marketing focus and strategy to fit.

- Better brand strategy – Once you have identified the key motivators for your customer, such as design or price or practical needs, you can brand your products appropriately.

- Improve distribution strategies – Identifying where customers shop and when can informatively shape product distributions strategies, such as what type of products are sold at particular outlets.

- Customer retention – Using segmentation, marketers can identify groups that require extra attention and those that churn quick, along with customers with the highest potential value.

Customer Segmentation with Machine Learning

- Developments in machine learning and deep learning have made it much easier for companies and individuals to build a high-performance customer segmentation model.

- Knowing about machine learning, and unsupervised learning, in particular, it is quite evident that the customer segmentation problem is nothing but a clustering problem. So any machine learning method that could be used for clustering problems can be applied to customer segmentation as well.

The success of a machine learning model

The success of a machine learning model, however, does not depend solely on the selection of a machine learning method. Key factors contributing to the success of the machine learning model include:

- **Data**

Data is the very prerequisite for any successful machine learning model. No matter how great your machine learning models are, you cannot get a reliable high-performance model from the prediction model without a sufficient amount of rich data.

- **Feature Engineering**

Processing raw data and making it a suitable input for the machine learning models includes **data cleaning, creating new features, and feature selection**. Feature engineering usually is the most time-consuming machine learning problem, especially when it comes to building prediction models for structured data.

- **Models**

Even though there are many machine learning methods available for certain machine learning problems, such as clustering, for example, each method has its own strengths and weaknesses. Based on our demands and requirements, we may need to choose different methods.

- **Performance Metrics**

Silhouette Score

The Silhouette Score and Silhouette Plot are used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters. This measure has a range of [-1, 1] and is a great tool to visually inspect the similarities within clusters and differences across clusters.

Inertia : actually calculates the sum of distances of all the points within a cluster from the centroid of that cluster.

Problem Description

This project our task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Objective

Online Retail Customer Segmentation

Data Description

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.

- **InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.**

- **UnitPrice: Unit price. Numeric, Product price per unit in sterling.**

- **CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.**

- **Country: Country name. Nominal, the name of the country where each customer resides.**

EDA(Count Dataset)

	InvoiceNo	InvoiceNo_count	StockCode	StockCode_count	Description	Description_count	Quantity	Quantity_count	InvoiceDate	InvoiceDate_count	UnitPrice	UnitPrice_count	CustomerID	CustomerID_count	Country	Country_count
0	573585	1114	85123A	2313	WHITE HANGING HEART T-LIGHT HOLDER	2369	1	148227	2011-10-31 14:41:00	1114	1.250000	50496	17841.000000	7983	United Kingdom	495478
1	581219	749	22423	2203	REGENCY CAKESTAND 3 TIER	2200	2	81829	2011-12-08 09:28:00	749	1.650000	38181	14911.000000	5903	Germany	9495
2	581492	731	85099B	2159	JUMBO BAG RED RETROSPOT	2159	12	61063	2011-12-09 10:03:00	731	0.850000	28497	14096.000000	5128	France	8557
3	580729	721	47566	1727	PARTY BUNTING	1727	6	40868	2011-12-05 17:24:00	721	2.950000	27768	12748.000000	4642	EIRE	8196
4	558475	705	20725	1639	LUNCH BAG RED RETROSPOT	1638	4	38484	2011-06-29 15:58:00	705	0.420000	24533	14606.000000	2782	Spain	2533
5	579777	687	84879	1502	ASSORTED COLOUR BIRD ORNAMENT	1501	3	37121	2011-11-30 15:13:00	687	4.950000	19040	15311.000000	2491	Netherlands	2371
6	581217	676	22720	1477	SET OF 3 CAKE TINS PANTRY DESIGN	1473	24	24021	2011-12-08 09:20:00	676	3.750000	18600	14646.000000	2085	Belgium	2069
7	537434	675	22197	1476	PACK OF 72 RETROSPOT CAKE CASES	1385	10	22288	2010-12-06 16:57:00	675	2.100000	17697	13089.000000	1857	Switzerland	2002
8	580730	662	21212	1385	LUNCH BAG BLACK SKULL.	1350	8	13129	2011-12-05 17:28:00	662	2.460000	17091	13263.000000	1677	Portugal	1519
9	538071	652	20727	1350	NATURAL SLATE HEART CHALKBOARD	1280	5	11757	2010-12-09 14:09:00	652	2.080000	17005	14298.000000	1640	Australia	1259

This Dataset set gives count of each and every Variable in the Original Dataset with top 10 Contributors, it help me to understand the Data Better

Important thing to note are:

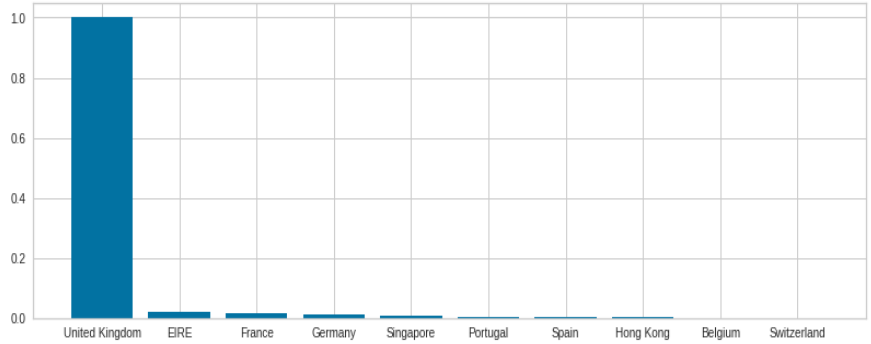
1.Only 25900 unique Invoice

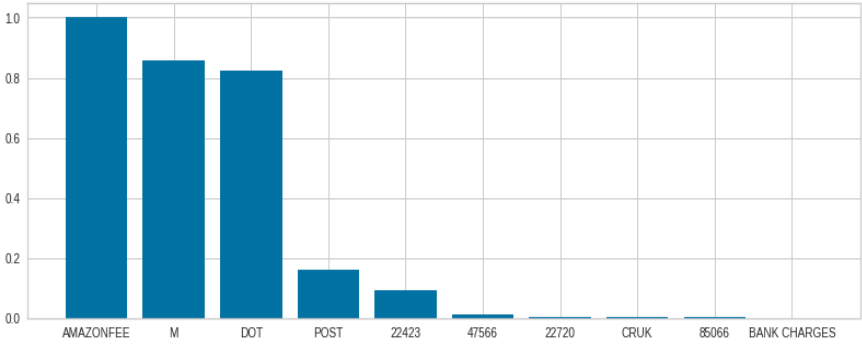
2.Only 4070 unique StokeCode(which means number of products =4070)

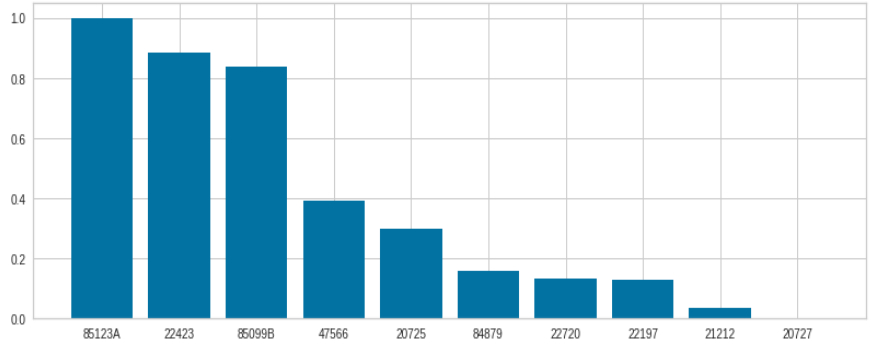
3.There are only 4372 Unique Customers (thus we should look a way to group data as per customers, thus we will have to treat the customers with Nan values as similar.

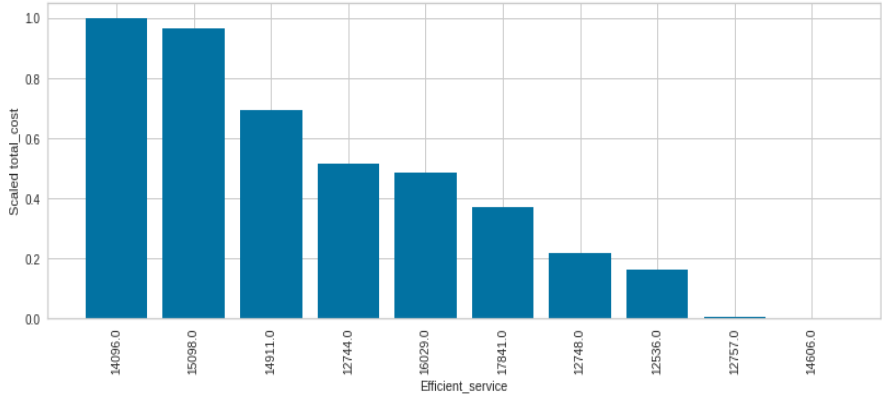
4.only 38 Country of which 10 country represents almost entire data, and UK alone explain 90% Data thus we will group country 'UK' as 1 and 'Others' as 0.

5.Also well be droping Description Coulmns as we already have Stockcode for product mapping also their are few Nan in Description Coulmns.









Bar Chart

Chart 1: total_cost_vs_country, chart 2: total_cost_vs_Efficient_service, chart 3: units_vs_Popular_service, chart 4: total_cost_vs_Valueable_CustomerID

Feature Engineering

```
df['Order Status']=df['InvoiceNo'].apply(lambda x: np.where(str(x)[0]=='C',0,1))          #As there are some Canceled Data in the Transaction we'll be marking them as '0'.
print(df['Order Status'].value_counts())                                                  #Counting the Number of Canceled Orders.
try:
  df.drop('Description',axis=1,inplace=True)
  df.dropna(inplace=True)
except:
  pass
df['Country']=df['Country'].apply(lambda x: np.where(x=='United Kingdom','1','0'))
df.head()
```

Extracting useful information

```
print(type(df['InvoiceDate'][0]))                                                         #Time stamp format
max_date=df['InvoiceDate'].max()
print(max_date)                                                                           #Lets take is date as a refrence for the Last Transaction, thus we will now compute the Last date of each transaction
print(df['InvoiceDate'].min())
df['Days Before Last Trans']=df['InvoiceDate'].apply(lambda x: (max_date-x).days )
```

Making New Features

```
df['Amount']=df['Quantity']*df['UnitPrice']                                               #Multiplying Qnt and Price per unit to get Amount, and then Droping UnitPrice Column, Using Quantity i will make another Feature later
try:
  df.drop(['UnitPrice'],axis=1,inplace=True)
except:
  pass
df.head()
```

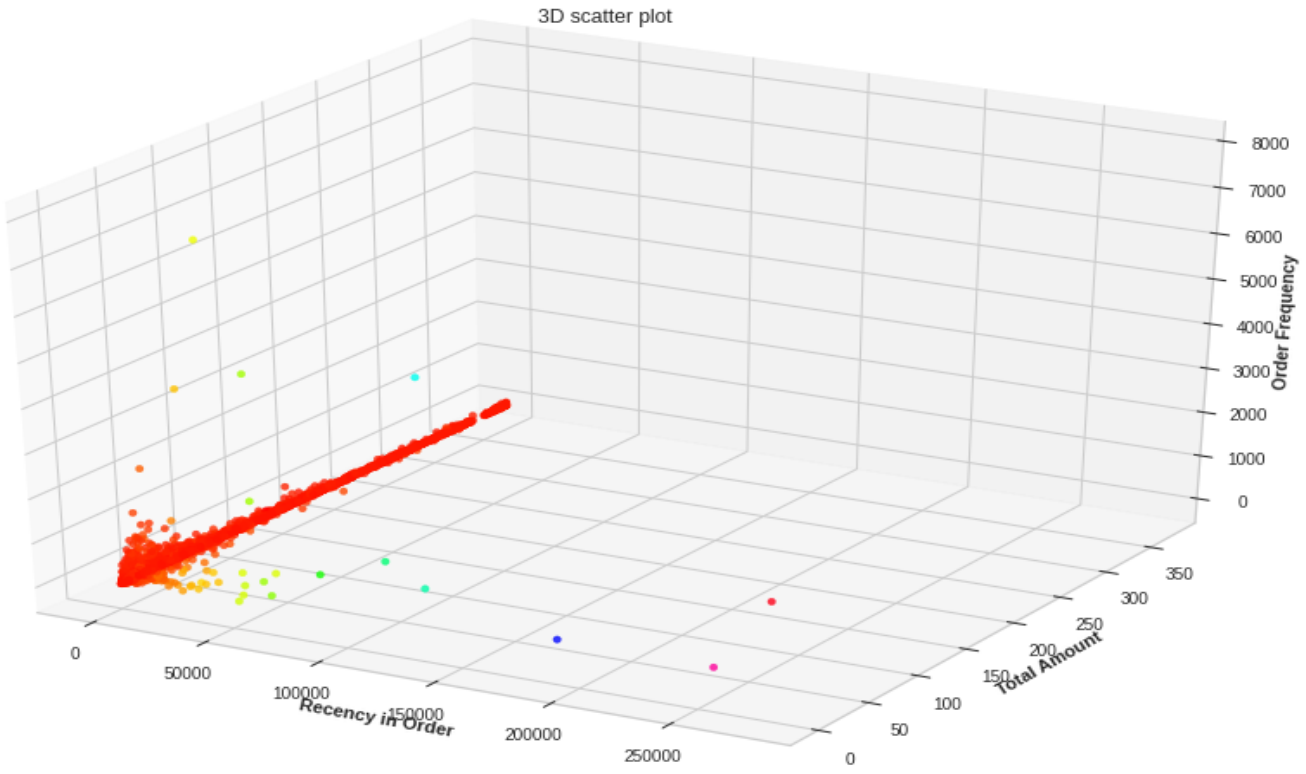
One hot encoding

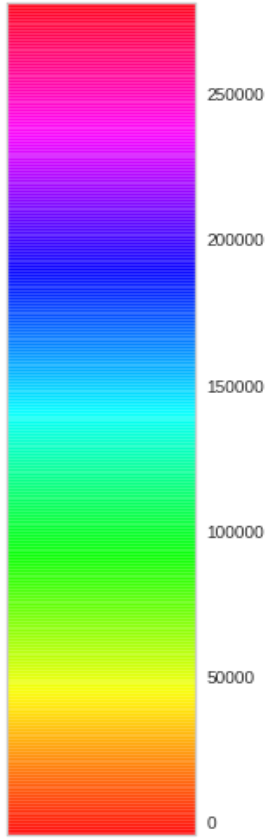
```
one_hot_entity=['Order_Status']
column_one_hot=['Order Status']                                    #One hot encoding of Order Status so as to get the gist of number of successful, canceled orders
count=0
for i in column_one_hot:
  temp_df=pd.get_dummies(df[i], prefix=one_hot_entity[count])
  count+=1
  df=pd.concat([df, temp_df], axis=1)
try:
  df.drop(['Order Status','StockCode'],axis=1,inplace=True)        #Dropping Order Status and StockCode as they are of no use now
except:
  pass
df.head()
```

Preparing the Data Set

```
df1=df.groupby('CustomerID')['Country'].max()                      # Our Features are now ready and thus i will be making the Final Dataset
df2=df.groupby('CustomerID')['Order_Status_0'].sum()
df3=df.groupby('CustomerID')['Order_Status_1'].sum()
df4=df.groupby('CustomerID')['Days Before Last Trans'].min()
df5=df.groupby('CustomerID')['Amount'].mean()
df6=df.groupby('CustomerID')['Amount'].sum()
df7=df.groupby('CustomerID')['Quantity'].mean()
columns=['Order_Status_0','Order_Status_1','Days Before Last Trans','Avg Amount','Total Amount','Avg_Quantity']
data_frames=[df2,df3,df4,df5,df6,df7]
final_df=pd.DataFrame(df1)
count=0
for i in columns:
  final_df[i]=data_frames[count]
  count+=1
```

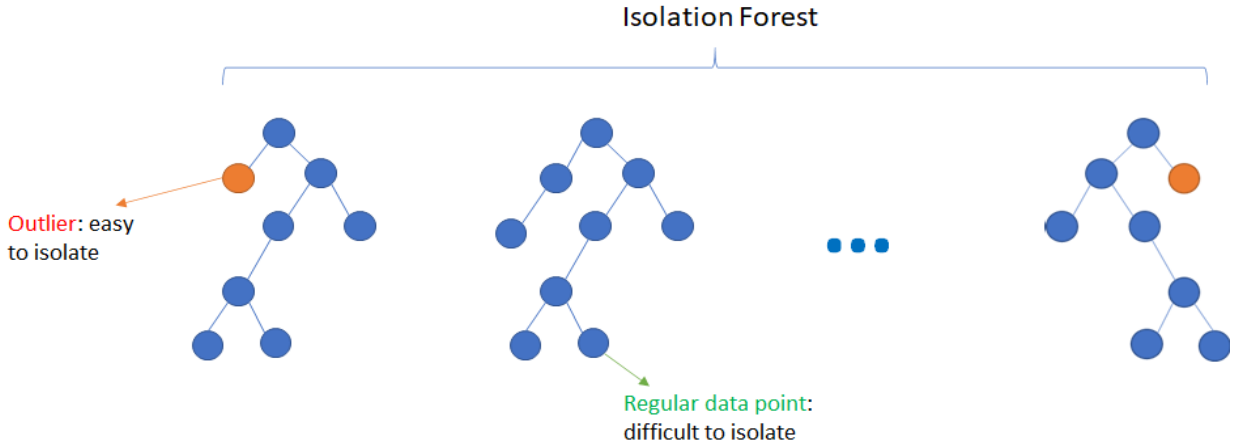
Doing this we have converted our humongous dataset contain 5Lakh 41 Thousand observations to just 4300.



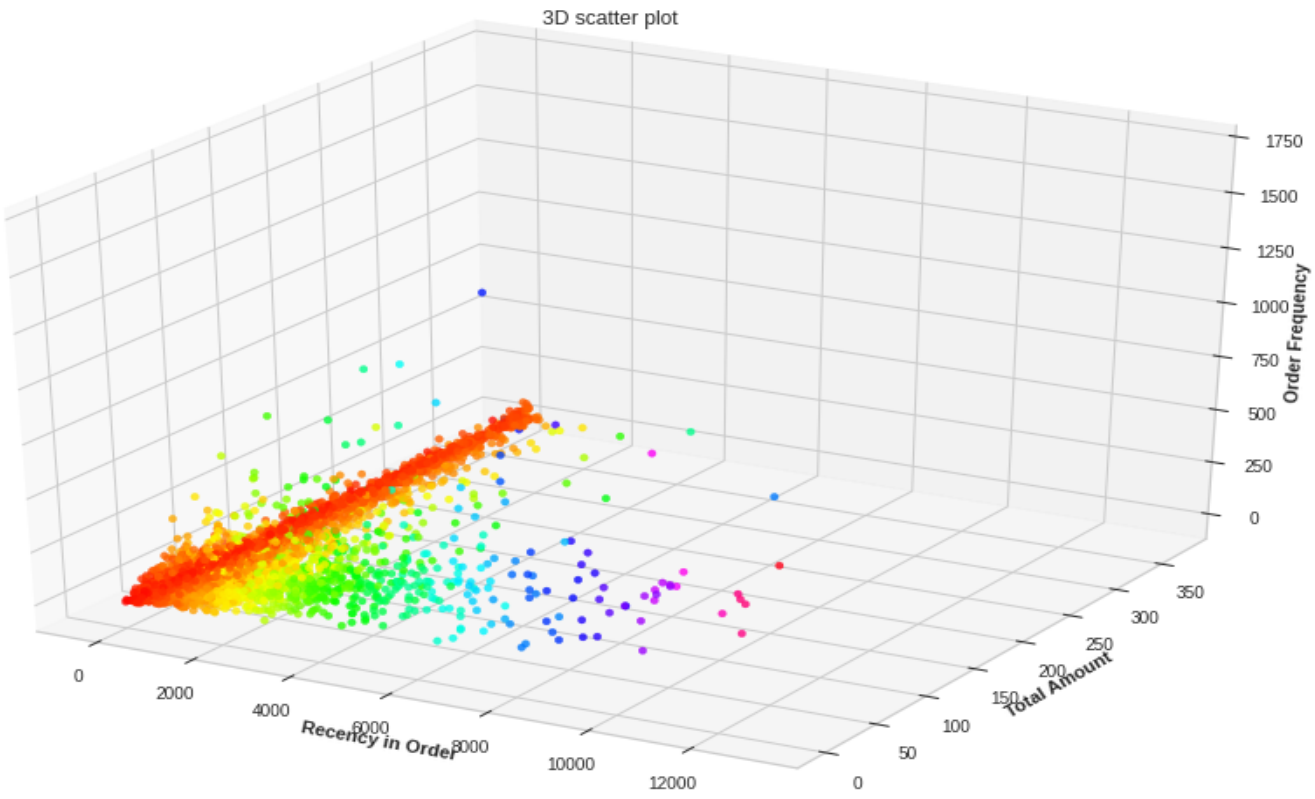


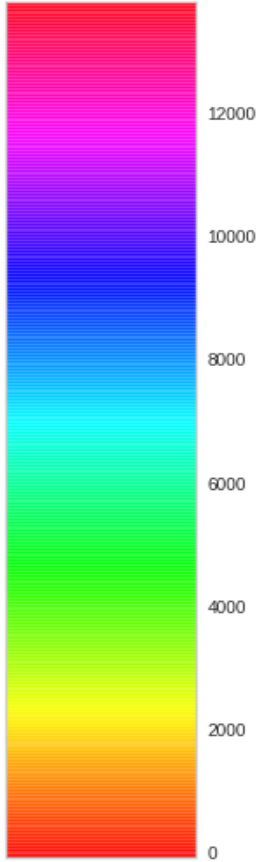
3D Scatter Plot is Really Very Conserve which means there are quite a few anomalies in our Data lets remove them

Outlier Detection with Isolation Forest

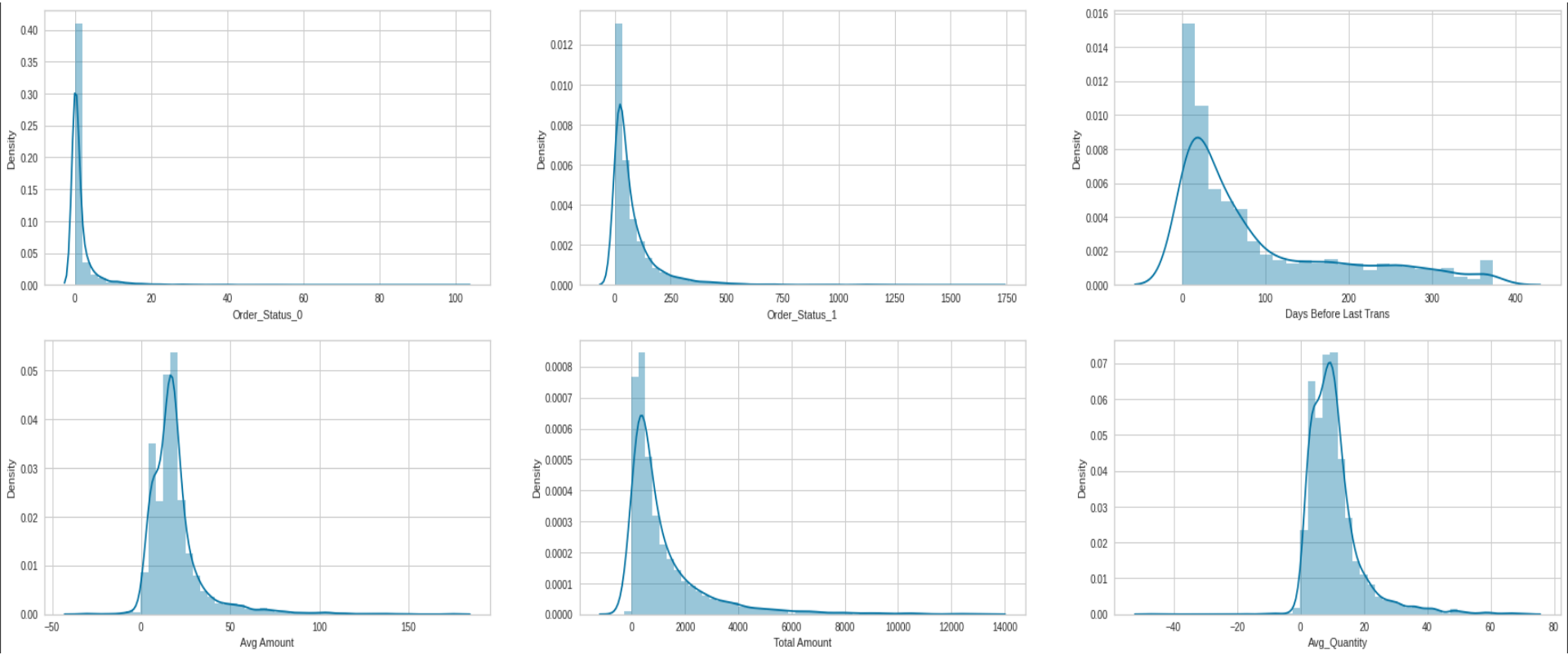


Lets do Anomaly detection now, as we have already removed our Wholesalers from the data these customers can be safely termed as Anomalies based on multi-Variate analysis.

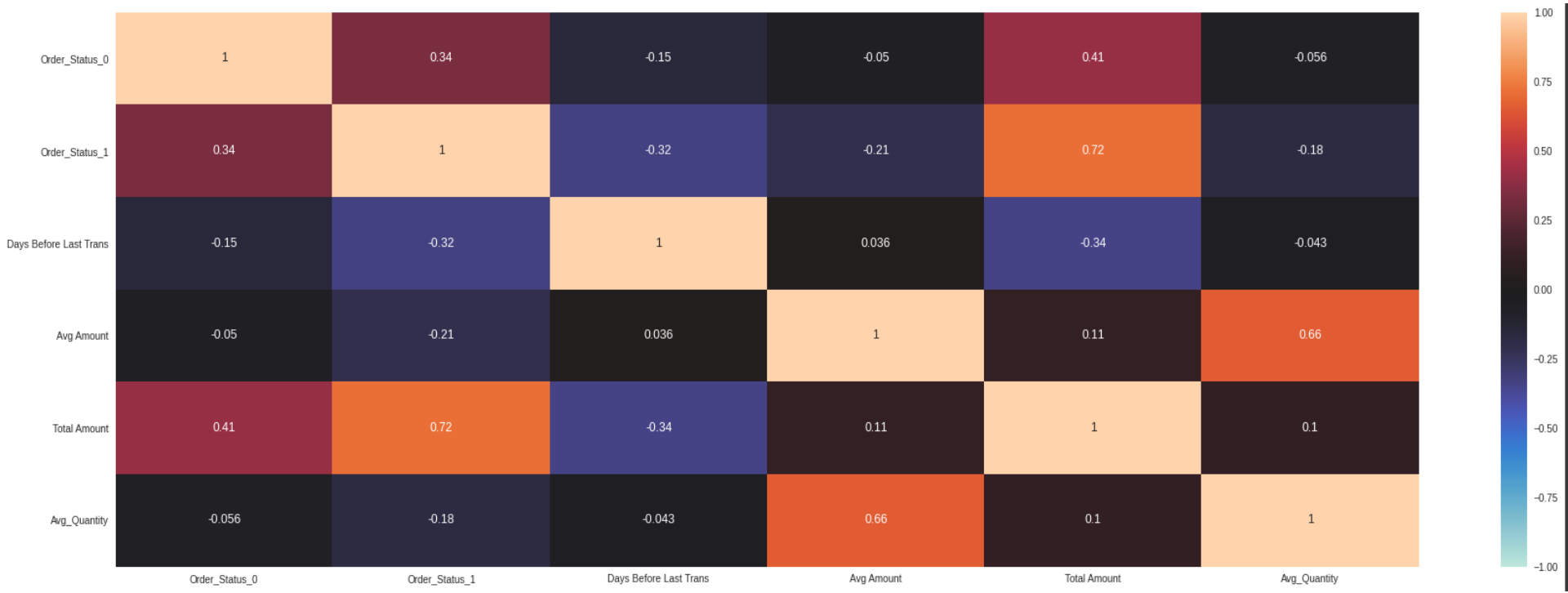




3D Scatter Plot now looks good



Distribution of our processed Data Frame

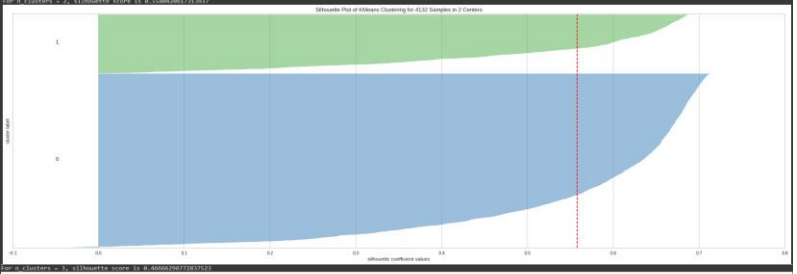


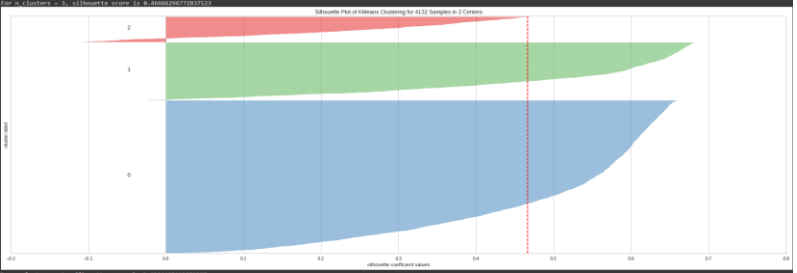
Looking at Correlation chart it can be seen that there is vey high correlation between some of the independent variables thus it is quite possible that we will be using PCA for feature extraction and dimensionality reduction

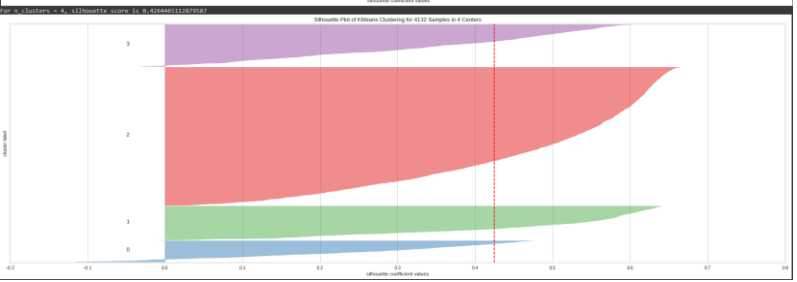
Let's make our first basic KMeans model using all the all the features and Various Values of K(number of clusters)

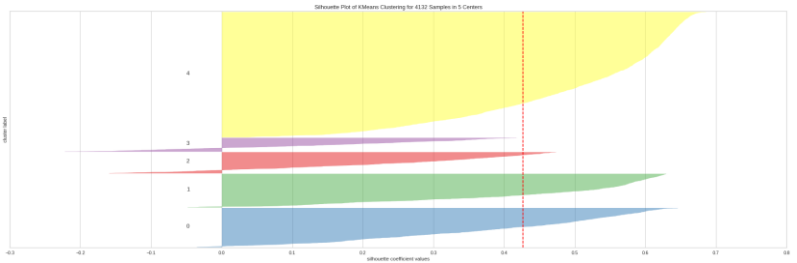
```
minmax = MinMaxScaler(feature_range=(0, 1))                                    # i will not consider contries as a basis of segmentation in model, beacuse if we want that we can do it here itself, in model it will create a bias.
X = minmax.fit_transform(final_df[['Order_Status_0','Order_Status_1','Days Before Last Trans','Avg Amount','Total Amount','Avg_Quantity']])
def silhouette_score_analysis(n):
  for n_clusters in range(2,n):
      km = KMeans (n_clusters=n_clusters, max_iter=100,tol=0.01)
      preds = km.fit_predict(X)
      centers = km.cluster_centers_
      score = silhouette_score(X, preds, metric='euclidean')
      print ("For n_clusters = {}, silhouette score is {}".format(n_clusters, score))
      visualizer = SilhouetteVisualizer(km)
      visualizer.fit(X)
      visualizer.poof()
silhouette_score_analysis(6)                                                   # More than 5 Clusters would Not seem Logical for Buisness Point of view.
```

Let's See the ideal value of K for this model by plotting silhouette score





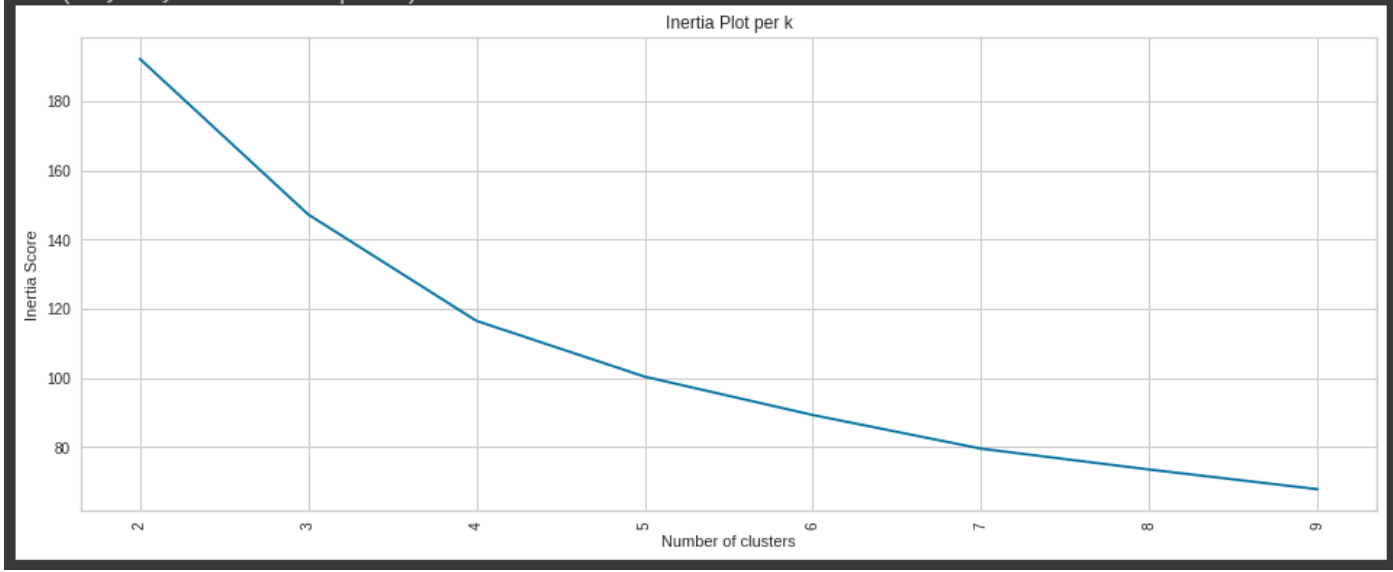




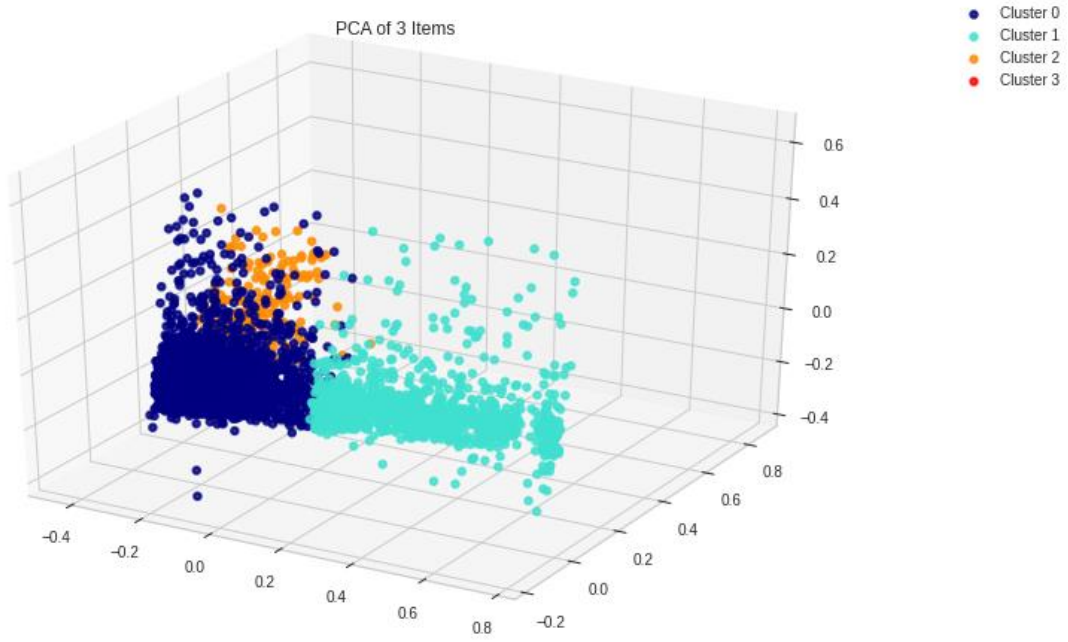

```
For n_clusters = 2, silhouette score
is 0.5580420617313937

For n_clusters = 3, silhouette score
is 0.46666296772837523

For n_clusters = 4, silhouette score
is 0.4264465112879587

For n_clusters = 5, silhouette score
is 0.433987882359444
```

For 2 cluster the value of silhouette score is maximum, lets make inertia plot to validate this.

```
The innertia for : 2 Clusters is: 192.2443540843187
The innertia for : 3 Clusters is: 147.3060843866467
The innertia for : 4 Clusters is: 116.55536377919431
The innertia for : 5 Clusters is: 100.42517516359318
The innertia for : 6 Clusters is: 89.33647021991288
The innertia for : 7 Clusters is: 79.61208491433041
The innertia for : 8 Clusters is: 73.57988547829648
The innertia for : 9 Clusters is: 67.89062892962464
Text(0.5, 1.0, 'Inertia Plot per k')
```



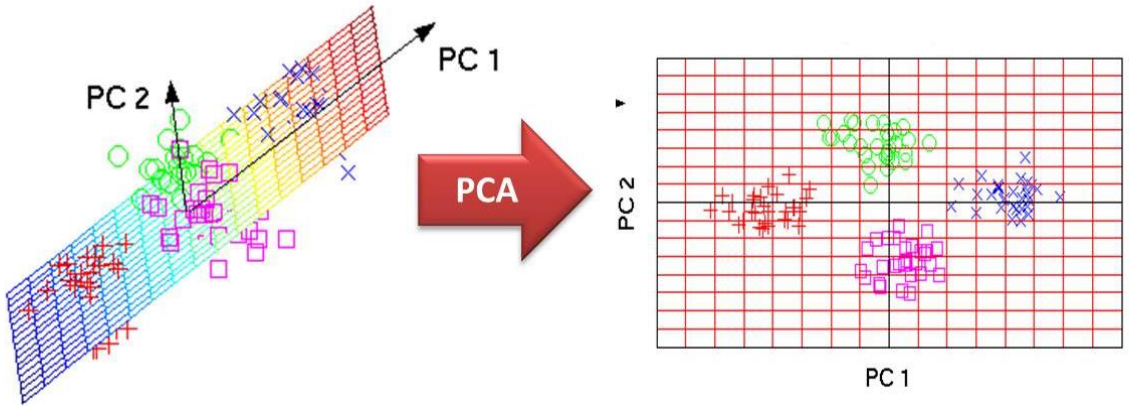
This elbow curve depicts a value of 3 or 4 which means, these two plots give conflicting value for no. of clusters, thus will try principal component analysis to get watch if there is any change, before which I will plot this model on a 3D scatter plot with clusters=3.



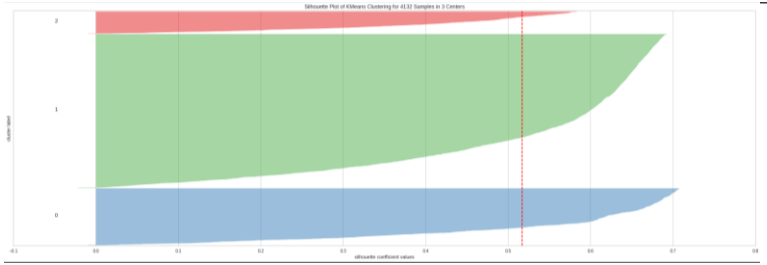
3D Scatterplot of our Base Model we will try to improve this further using PCA

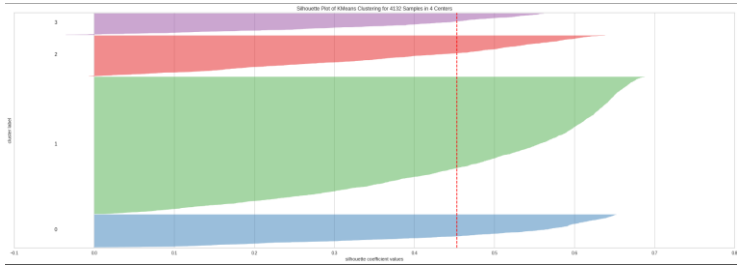
Principal Components Analysis

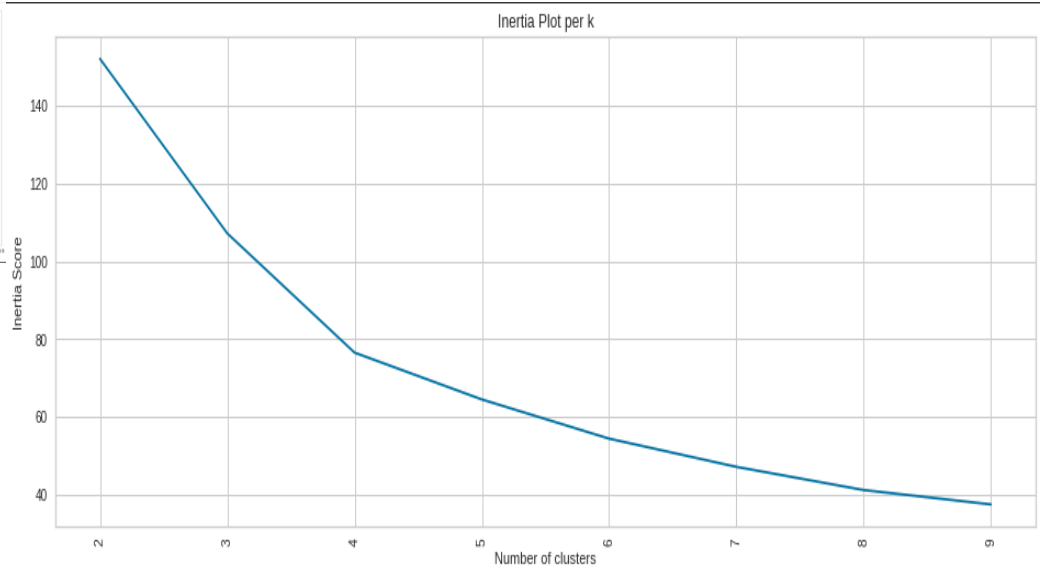
The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first i-1 vectors. Here, a best-fitting line is defined as one that minimizes the average squared distance from the points to the line. These directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated. Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.



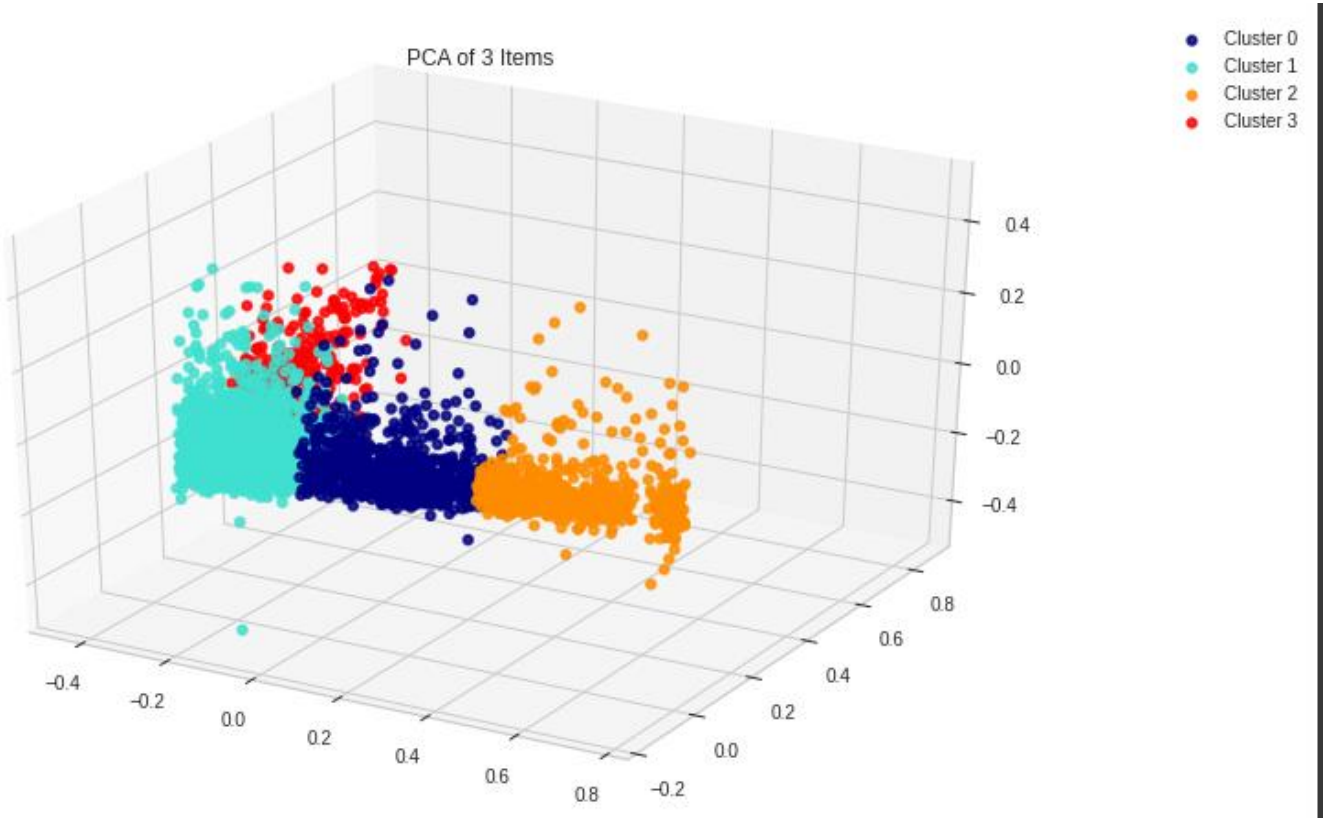
After Using PCA





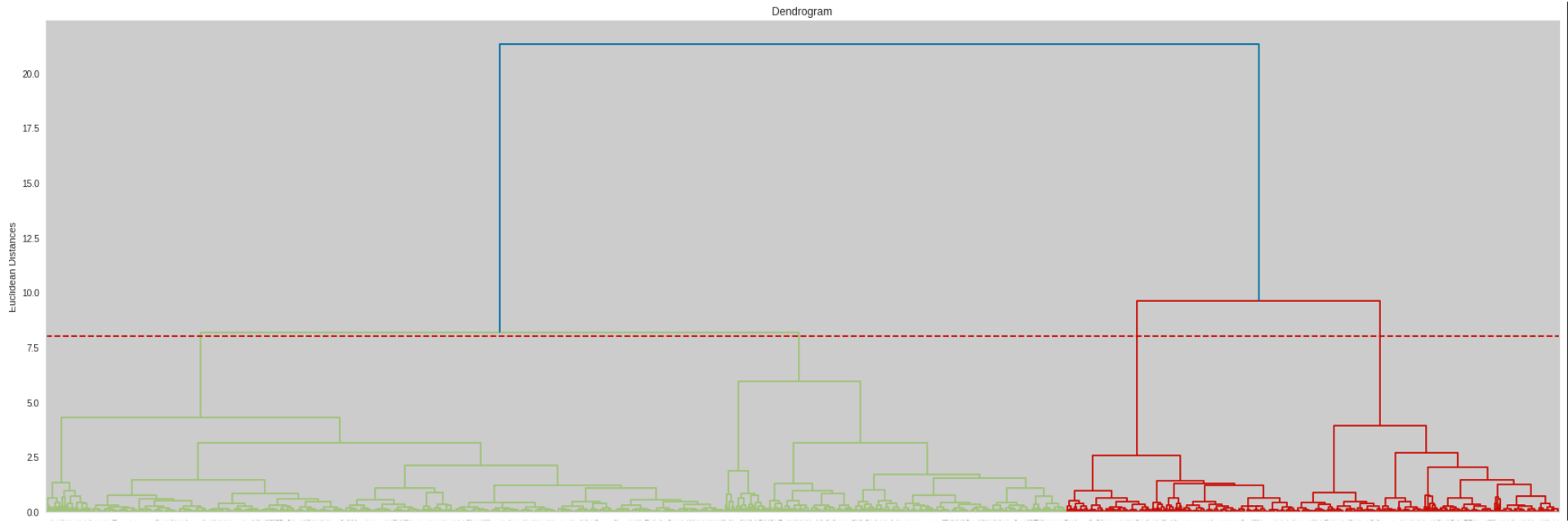


Silhouette curve gives a value of 3 or 4, and elbow curves 4 as a value thus I will take my clusters =4.



3D Scatterplot Improved Model using 4 clusters

Hierarchical Clustering and plotting a Dendrogram



As per Dendrogram either 2 or 4 clusters should exist, we will go with 4 and try to explain each segment if every thing will make sense we will stick to it, else it we'll take the number of cluster as 2

Following are the Customer Segments

- Customer Segment 1: Active Customers with medium buying frequency and Avg Transaction Values.
- Customer Segment 2: Passive Customers with low buying Frequency and medium transaction Values
- Customer Segment 3: Passive Customers with low buying Frequency and transaction Values
- Customer Segment 4: Active Customers with high Order frequency and Transaction Value.
- Customer Segment 5(wholesalers): Active Customers with low order frequency, high Total Amounts and Really high Quantities Order

All the Segments looks good thus we are good to go.

THANK YOU