CLASSIFYING CUSTOMERS BASED ON THEIR SHOPPING PRACTICES

Tharun Kanamneni
Master's In Data Analytics Engineering
George Mason University
Fairfax, Virginia
tkanamne@gmu.edu

Anusha Chirukuri
Master's In Data Analytics Engineering
George Mason University
Fairfax, Virginia
achiruku@gmu.edu

Bhargav Ram Nara
Master's In Data Analytics Engineering
George Mason University
Fairfax, Virginia
bnara@gmu.edu

Abstract— **In order to satisfy the demand for particular products needed by the customer, the retail industries need to understand the demand and should be able to classify the customer and their needs. In order to fulfill this, customer segmentation is necessary to divide the customer based on their needs and mark the item as high demand and low demand and this can help the retail shops to improve their stocking requirements. To solve this problem, we used various clustering and classification techniques like K-means, Random Forest, Ada Boost, K nearest neighbor and logistic regression, svc model. Based on this analysis, we develop a model that allows to anticipate the purchases that will be made by a new customer, from its first purchase. We have performed a thorough analysis of customer data, analyzed it and solved the problem of data classification by using customer's data to know the top sold products and top sales in each country which further helped to know various factors like number of repeated customers, cancelled orders, told sold products, top items returned, which day has most number of sales, which month has highest percentage of sales and products and customer clusters.**

Keywords—Customer segmentation, shopping, online market, clustering, machine learning, classification, retail industries.

I. INTRODUCTION

We have witnessed a robust increase in online retail sales from the past ten years. According to the Interactive Media in Retail Group (IMRG), online shoppers in the United Kingdom spent an estimated £50 billion in the year 2011, a more than 5000 percent increase compared with the year 2000.[2][1].This exceptional increase in online sales symbolizes that the way customers shop has radically changed. When compared to the traditional way of shopping in retail stores, online shopping has some advantages. Customer shopping processes and activities can be monitored automatically and accurately. Each customer order is normally linked with a delivery address and a billing address. The customer has an online store account with necessary contact and payment information. There are various business factors which are depending on the flow of income and expenses every day. [1]A customer usually tends to navigate to one or more websites for buying the items but if the items which are available for the customer are satisfied then the customer does not take another option for buying the item he needs.

Customer relationship management is a very important aspect in the business, and it's commonly called as CRM.[8] By using CRM, we can store the details of the customer and also, we can have a record on how many items has the customer purchased and also, we can have a record on the items sold most. This can help to get the data on segregating the customers as returning customers. However, sometimes it is hard to decide on if the customer is going to get the same brand and same item as there are new brands and items coming up every day.[7][4] So, this study aims to solve most of the problems like stocking issues, customer interests

A. *Importance of products to customers*

Customer satisfaction is very important for a company in order to sell its products and keep the client as a returning customer. Some business goes beyond customers' expectations and get their feedback in order to grow the business more and improve their standards and develop a relationship between them. Some of the other questions which arises for most of the businesses which interested us to do an enhanced research and develop a more effective approach to analyze the data are:

1. Which products / items have customers purchased together often?

2. In which sequence the products have been purchased?

3. Which types of customers are more likely to respond to a certain promotion mailing?

4. What are the sales patterns in terms of various perspectives such as products / items, regions and time (weekly, monthly, quarterly, yearly and seasonally), and so on?

II. Literature review

We have considered a UK-based and registered online retailer non-store business with around 80 members of staff. The company was established in 1981 primarily marketing/selling unique gifts for all occasions. Initially, when the business was established, the merchant relied massively on direct mailing catalogs, and the purchase orders were taken over the phone. They moved to online sales recently after creating their website. From then, the company has gathered vast amounts of data about many customers from all parts of the United Kingdom and Europe. It also uses Amazon.co.uk to market and sell its products.

The author (Morgado, A.(2018)) recommends that focus should be on potential customers rather than just referencing the customers, but whereas in the business the focus should not only be dealt with potential customers but also should be interested in customer's needs and how the customer's requirements are changing and how can we improve the products as well as service to have a successful business.

The online retailer chosen for this project is a typical one: a small business and a relatively new participant to the online retail sector. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. So, the project mainly aims to help the company to identify the types of customer groups with similar behaviors for further analysis and business strategy planning. We have done customer segmentation where we classify the customers into various unique customer groups that share similar characteristics. They provide means to identify unsatisfied customer needs. Businesses can withstand competition from its competitors by developing unique strategies to sell their products and services. Customer segmentation is done in various ways such as demographic information which provides information such as gender, age, marital status, income, education, and occupation. Geographical information, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence. Psychographics, such as social class, lifestyle, and personality traits. Performing customer segmentation help in determining relevant product pricing, in generating customized marketing campaigns, in creating an optimal distribution strategy, for extracting specific product features for deployment and in prioritizing innovative product development efforts.

III. Dataset Description

Dataset is taken from UCI Machine learning repository (https://archive.ics.uci.edu/ml/machine-learning-databases/00502/). This online Retail dataset contains all the transactions for a non-store online retail which is based in the UK between 01/12/2009 and 09/12/2011.Company primarily sells unique all-occasion giftware. Customers associated with this company are mostly wholesalers .Below is the information of all the variables in the dataset along with the datatype and description:

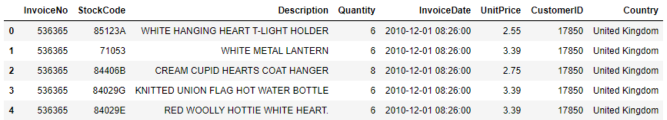
Variable Name	**Data Type**	**Description**
Invoice	Nominal	Invoice Number
Stock Code	Nominal	Product(item)
Description	Nominal	Product(item)
Quantity	Numeric	Quantity of each item
Price	Numeric	Product price per unit
Invoice Data	Numeric	Time and Date of Transaction
Address Line 1	Nominal	Delivery Address Line 1
Address Line 2	Nominal	Delivery Address Line 2
Address Line 3	Nominal	Delivery Address Line 3
Post Code	Nominal	Delivery Postcode

Country	Nominal	Delivery Country

IV. Methodology

Here, we have used the data from UCI machine learning repository, we got this data from an opensource database.

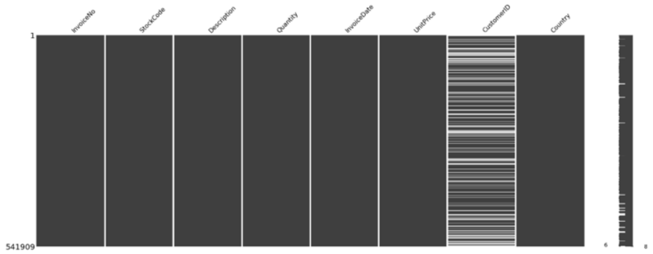
The dataset description snapshot is given below:



	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	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

A. *Data Preparation*

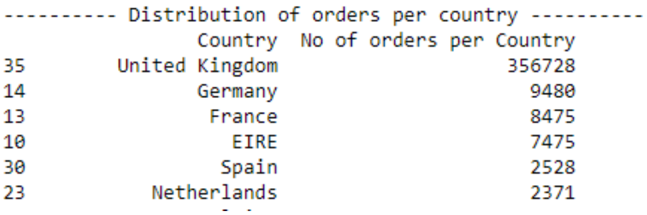
Here, we are checking the data to see if there are any null values first and then planning on dividing the dataset to obtain required columns and use them to draw required objectives. Real world data contains null values and a lot of noise. It has to be preprocessed before using it for further analysis. We have identified the null data duplicate data and removed it.



As we observe from above visualization, there are certain null values in data entries in the customer ID column which has to be removed or imputed with centralized measures. As it is not a feasible way for us to map the customer entries to customer ID. Now, we look into certain dimensions of the data and take a note of it, so that it is easy for us to drop duplicate or replicate values.

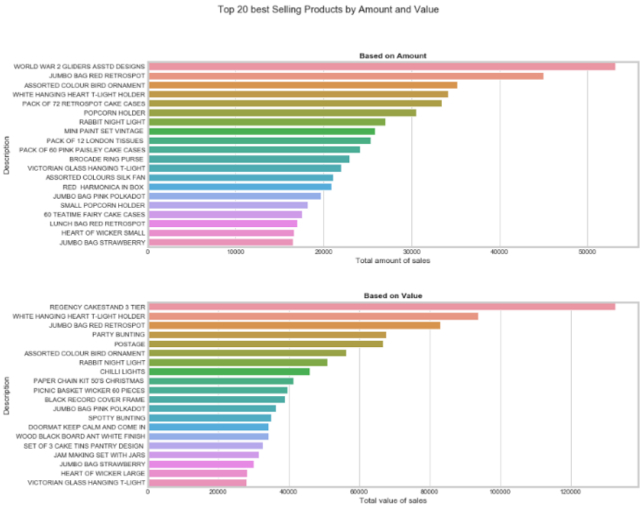
B. *Exploratory Ananlysis of the data.*

The below visualization gives information regarding the distribution of orders per country. Here, the United Kingdom is in the top.

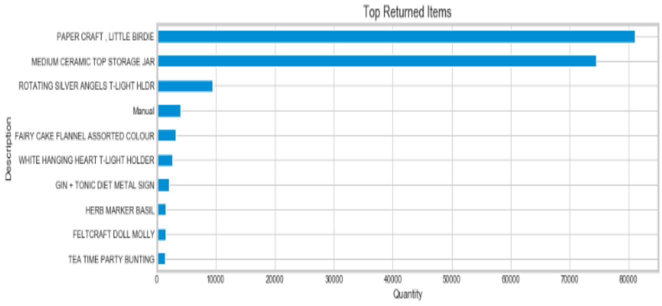


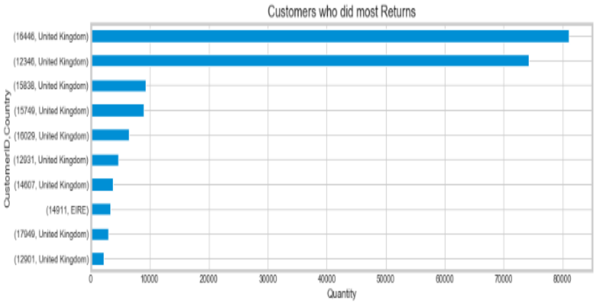
```
---------- Distribution of orders per country ----------
               Country  No of orders per Country
35      United Kingdom                    356728
14             Germany                      9480
13              France                      8475
10                EIRE                      7475
30               Spain                      2528
23         Netherlands                      2371
```

We have also performed a visualization which gives the information regarding the top 20 best-selling products based on the amount and value.

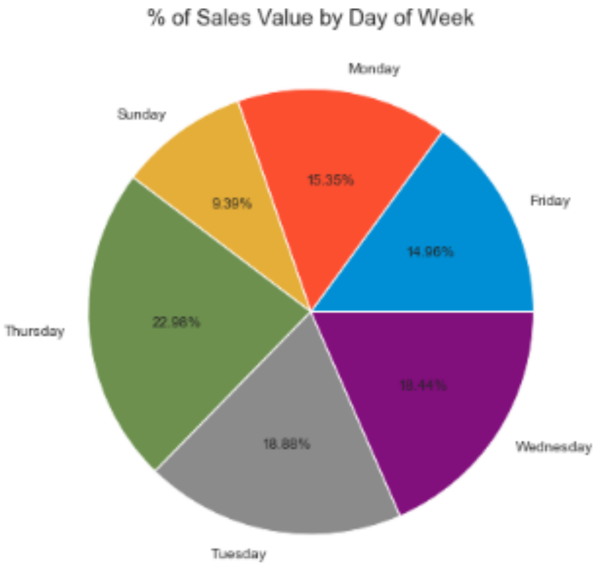


To find out most of the top returned items and customers who have done most returns we did a visualization to get the detailed analysis.

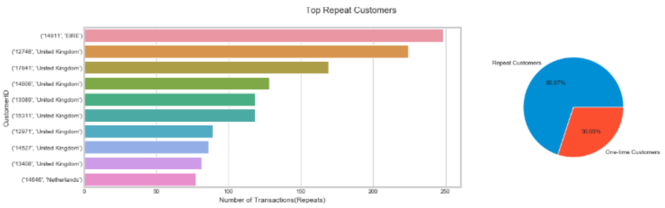




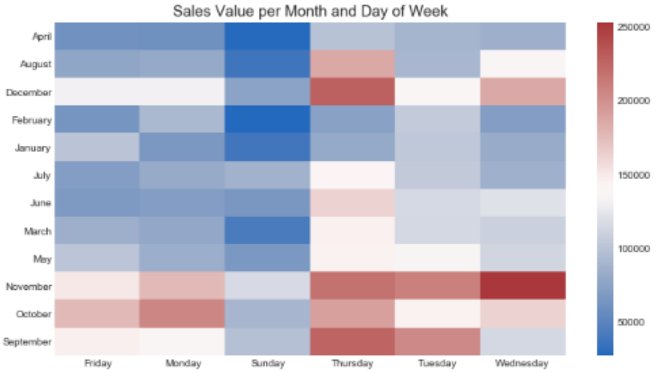
To get more detailed analysis we visualized a pie chart which gives a clear understanding of which day has more sales in a week.



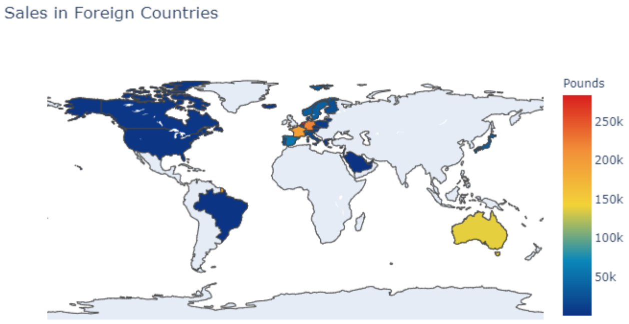
When having a business, it is required to know the count of return customers as well as one-time customers, so the below visualization gives the detailed analysis of it.



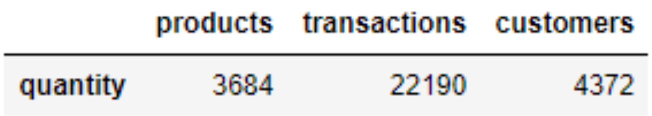
And we have also done a visualization which gives information regarding sales which are grouped by month and day.



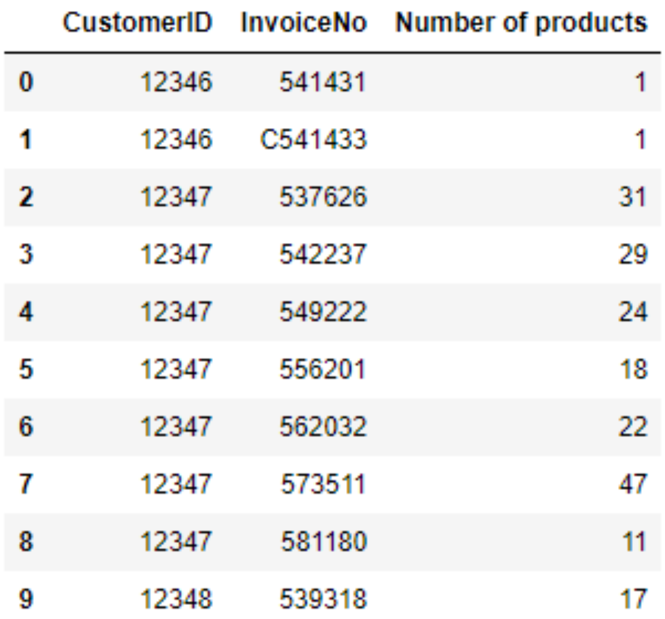
To know about the sales in each country, we have designed a visualization using the Choropleth map. From this visualization we can say that Netherlands has most sales among other foreign countries.



To understand the data in detail, we worked on exploring more about the products, transactions and customers. And also exploring transactions by grouping Customer ID. Following are the results:

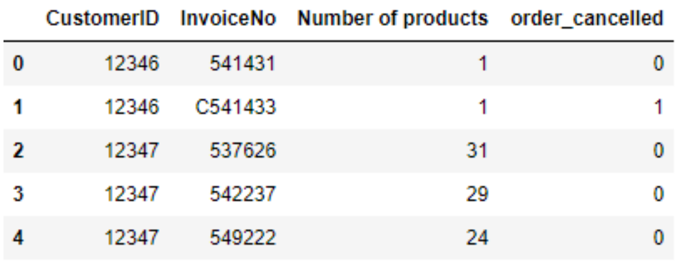


	products	transactions	customers
quantity	3684	22190	4372

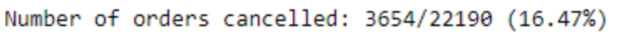


	CustomerID	InvoiceNo	Number of products
0	12346	541431	1
1	12346	C541433	1
2	12347	537626	31
3	12347	542237	29
4	12347	549222	24
5	12347	556201	18
6	12347	562032	22
7	12347	573511	47
8	12347	581180	11
9	12348	539318	17

The following tables give us understanding regarding the orders cancelled.

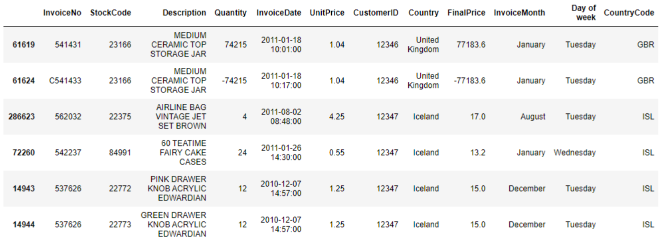


	CustomerID	InvoiceNo	Number of products	order_cancelled
0	12346	541431	1	0
1	12346	C541433	1	1
2	12347	537626	31	0
3	12347	542237	29	0
4	12347	549222	24	0



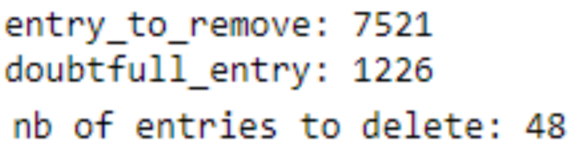
```
Number of orders cancelled: 3654/22190 (16.47%)
```

We verified if a transaction is cancelled, is there an identical transaction that is made or not. But hypothesis is failed so we can conclude that all cancelled orders need not have an identical transaction.

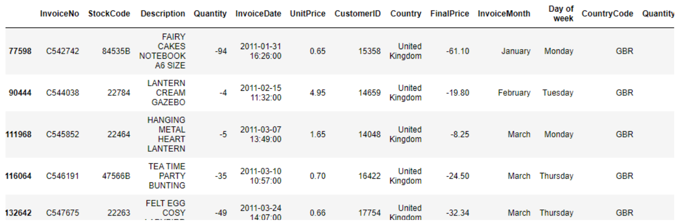


	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	FinalPrice	InvoiceMonth	Day of week	CountryCode
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom	77183.6	January	Tuesday	GBR
61624	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346	United Kingdom	-77183.6	January	Tuesday	GBR
286623	562032	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-08-02 08:48:00	4.25	12347	Iceland	17.0	August	Tuesday	ISL
72260	542237	84991	60 TEATIME FAIRY CAKE CASES	24	2011-01-26 14:30:00	0.55	12347	Iceland	13.2	January	Wednesday	ISL
14943	537626	22772	PINK DRAWER KNOB ACRYLIC EDWARDIAN	12	2010-12-07 14:57:00	1.25	12347	Iceland	15.0	December	Tuesday	ISL
14944	537626	22773	GREEN DRAWER KNOB ACRYLIC EDWARDIAN	12	2010-12-07 14:57:00	1.25	12347	Iceland	15.0	December	Tuesday	ISL

We see that the initial hypothesis is not verified so we are trying to check two scenarios. If a cancel order exists without counterpart and the other scenario is to see if there's at least one counterpart with the exact same quantity.



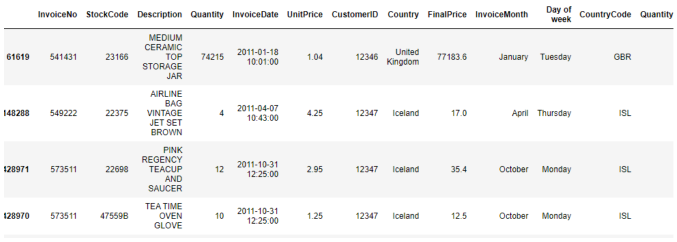
```
entry_to_remove: 7521
doubtfull_entry: 1226
 nb of entries to delete: 48
```



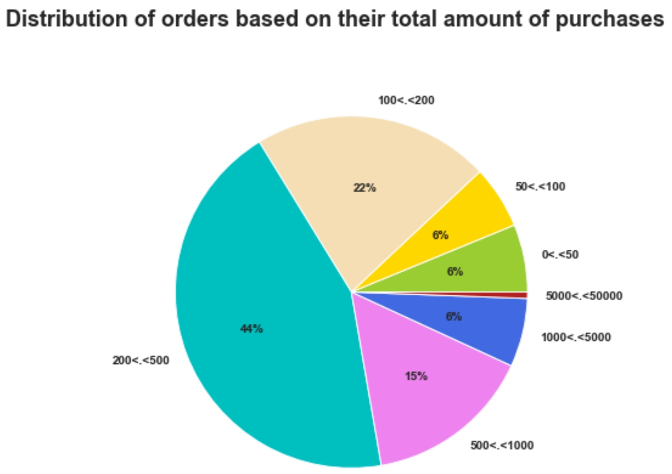
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	FinalPrice	InvoiceMonth	Day of week	CountryCode	Quantity
77598	C542742	84535B	FAIRY CAKES NOTEBOOK A6 SIZE	-94	2011-01-31 16:26:00	0.65	15358	United Kingdom	-61.10	January	Monday	GBR	
90444	C544038	22784	LANTERN CREAM GAZEBO	-4	2011-02-15 11:32:00	4.95	14659	United Kingdom	-19.80	February	Tuesday	GBR	
111968	C545852	22464	HANGING METAL HEART LANTERN	-5	2011-03-07 13:49:00	1.65	14048	United Kingdom	-8.25	March	Monday	GBR	
116064	C546191	47566B	TEA TIME PARTY BUNTING	-35	2011-03-10 10:57:00	0.70	16422	United Kingdom	-24.50	March	Thursday	GBR	
132642	C547675	22263	FELT EGG COSY	-49	2011-03-24 14:07:00	0.66	17754	United Kingdom	-32.34	March	Thursday	GBR	

C. Market Basket Analysis

To know about the basket price, we have performed an analysis which gives a more detailed basket price, an analysis for purchase amount for every single order, distribution of orders and their total amount of purchase.

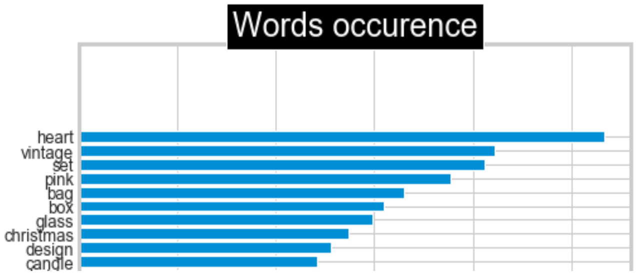


	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	FinalPrice	InvoiceMonth	Day of week	CountryCode	Quantity
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom	77183.6	January	Tuesday	GBR	
148288	549222	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-04-07 10:43:00	4.25	12347	Iceland	17.0	April	Thursday	ISL	
428971	573511	22698	PINK REGENCY TEACUP AND SAUCER	12	2011-10-31 12:25:00	2.95	12347	Iceland	35.4	October	Monday	ISL	
428970	573511	47559B	TEA TIME OVEN GLOVE	10	2011-10-31 12:25:00	1.25	12347	Iceland	12.5	October	Monday	ISL	



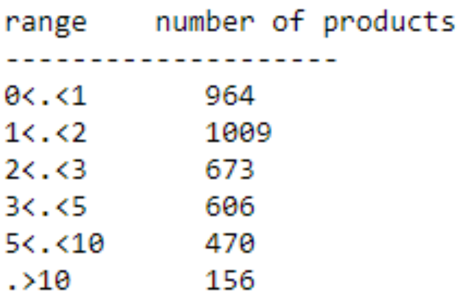
D. Product Categories

We have done word occurrence in order to form product categories based on their range.



From the above visualization, we see that "heart" is the most occurred word amongst all the other words followed by vintage.

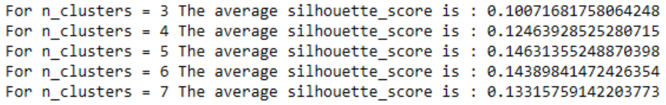
Based on the range we have categorized the products.



```
range    number of products
--------------------
0<.<1        964
1<.<2        1009
2<.<3        673
3<.<5        606
5<.<10       470
.>10         156
```

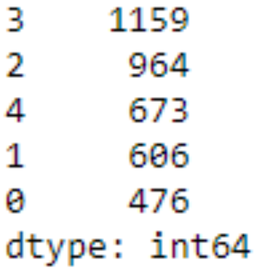
E. Creating Product Clusters Using K-Means

K-means clustering is a machine learning unsupervised approach to make inferences using various datasets. In the Online Retail dataset, if we divide our customers into only 2 segments and send out marketing material to each customer as per their cluster assignment, we may have too general of a marketing pitch. Thereby, customers may not return to our e-commerce website. On the other hand, if we divide our customers into 100 segments, we may only have a handful of customers per segment and it will be a nightmare to send out 100 variations of marketing material. So, while the choice of K is a business decision we do have techniques to guide our final decision. There are a few things left to explain before we apply the algorithm to our data. First, how do we choose how many clusters we want in our data? Well, this is exactly at the heart of the clustering problem. We do not know if there are 4 types of customers (clusters) as per our data, or 7 since this is indeed an unsupervised problem. So, part of our task is to identify the most suitable number of K clusters to segment our data. However, there is no “correct” answer here since there is no ground-truth. Indeed, choosing the value of K is often a business decision. Initially we have done clustering of products to get an average silhouette score for n clusters.



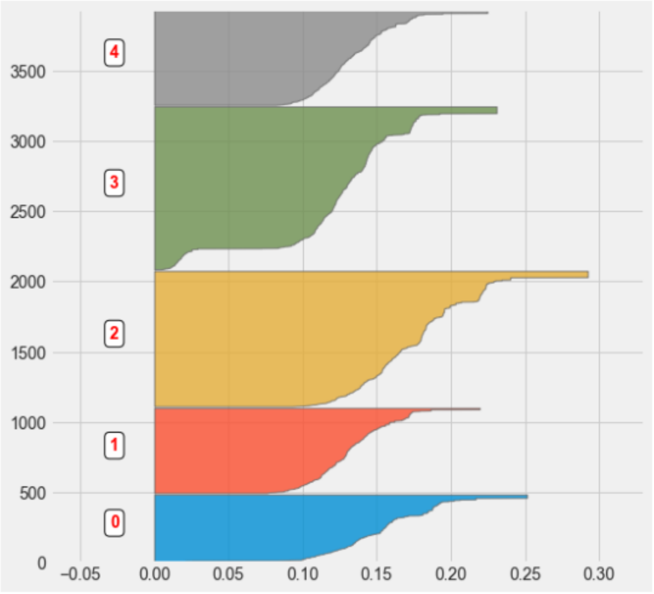
```
For n_clusters = 3 The average silhouette_score is : 0.10071681758064248
For n_clusters = 4 The average silhouette_score is : 0.12463928525280715
For n_clusters = 5 The average silhouette_score is : 0.14631355248870398
For n_clusters = 6 The average silhouette_score is : 0.14389841472426354
For n_clusters = 7 The average silhouette_score is : 0.13315759142203773
```

Then we have done characterizing the content of clusters.

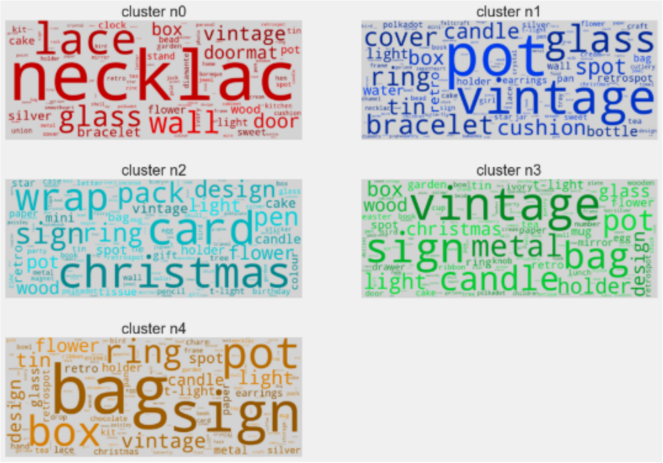


```
3     1159
2      964
4      673
1      606
0      476
dtype: int64
```

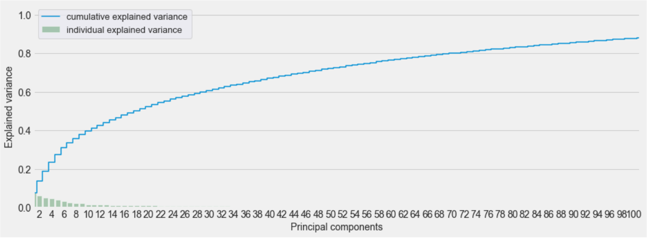
We used silhouette plot because to measure how close a cluster is from its neighboring cluster and to get the Silhouette intra cluster score.

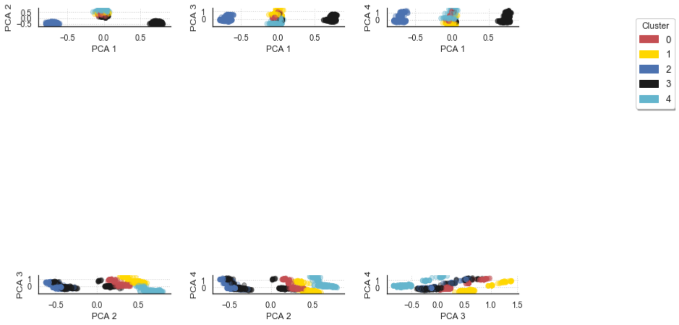


Then we have done a word cloud visualization which is used to represent the most frequently used word based on the frequency or importance. We have used the random color function for each n cluster value. And from the below results we can say that for in cluster n0 necklace and in cluster n1 pot is the most frequently used word in n2 its card, in n3 its vintage and in n4 its bag.

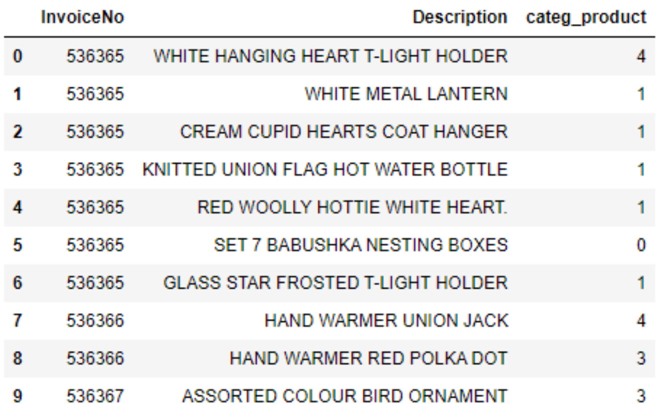


Next we have done principal component analysis which minimizes the average squared distance. While performing principal component analysis we have done some transformations to convert few observations to linearly uncorrelated variable values. The below visualization is between Principal components and then we checked for the amount of variance for each component. We need more than 100 components to explain the 90% of the variance of the data.



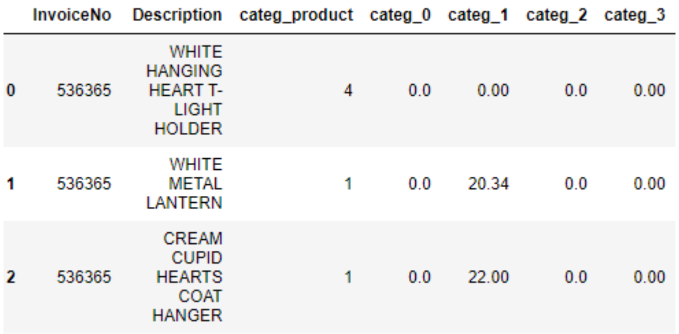


We have grouped different products in 5 clusters. For further analysis, we are introducing the information into the data frame.



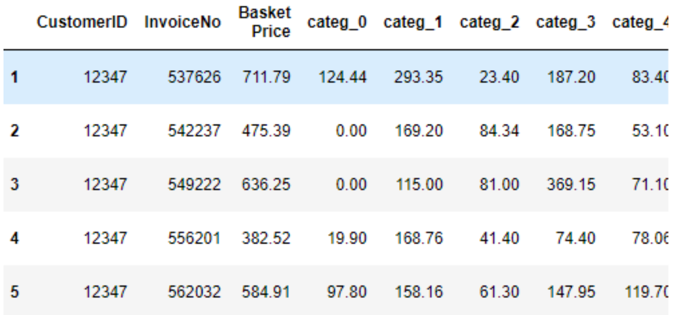
	InvoiceNo	Description	categ_product
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	4
1	536365	WHITE METAL LANTERN	1
2	536365	CREAM CUPID HEARTS COAT HANGER	1
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	1
4	536365	RED WOOLLY HOTTIE WHITE HEART.	1
5	536365	SET 7 BABUSHKA NESTING BOXES	0
6	536365	GLASS STAR FROSTED T-LIGHT HOLDER	1
7	536366	HAND WARMER UNION JACK	4
8	536366	HAND WARMER RED POLKA DOT	3
9	536367	ASSORTED COLOUR BIRD ORNAMENT	3

Secondly, we decided to create categN which contains the amount spent in each product category.



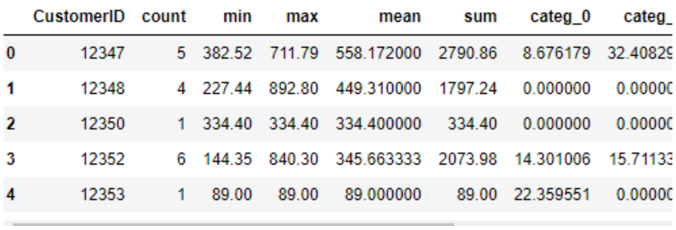
	InvoiceNo	Description	categ_product	categ_0	categ_1	categ_2	categ_3
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	4	0.0	0.00	0.0	0.00
1	536365	WHITE METAL LANTERN	1	0.0	20.34	0.0	0.00
2	536365	CREAM CUPID HEARTS COAT HANGER	1	0.0	22.00	0.0	0.00

Here we decided to collect the required information regarding the particular order and how it is distributed over the 5 categories of products.



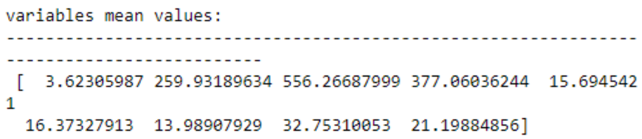
	CustomerID	InvoiceNo	Basket Price	categ_0	categ_1	categ_2	categ_3	categ_4
1	12347	537626	711.79	124.44	293.35	23.40	187.20	83.40
2	12347	542237	475.39	0.00	169.20	84.34	168.75	53.10
3	12347	549222	636.25	0.00	115.00	81.00	369.15	71.10
4	12347	556201	382.52	19.90	168.76	41.40	74.40	78.06
5	12347	562032	584.91	97.80	158.16	61.30	147.95	119.70

Then we have done splitting of data over time. The data frame basket price shows the sum of the total products that have been bought. We have also determined the number of purchases made by the user along with the minimum, maximum and average amount which has been spent during all the visits.



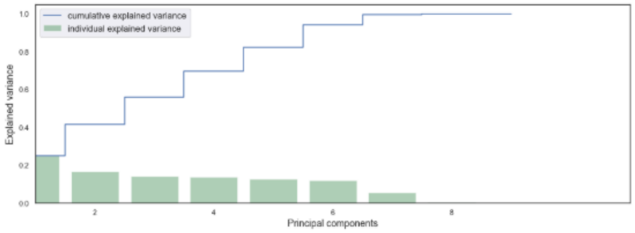
	CustomerID	count	min	max	mean	sum	categ_0	categ_
0	12347	5	382.52	711.79	558.172000	2790.86	8.676179	32.40829
1	12348	4	227.44	892.80	449.310000	1797.24	0.000000	0.00000
2	12350	1	334.40	334.40	334.400000	334.40	0.000000	0.00000
3	12352	6	144.35	840.30	345.663333	2073.98	14.301006	15.71133
4	12353	1	89.00	89.00	89.000000	89.00	22.359551	0.00000

Then we have created customers categories firstly we have done data encoding where each entry corresponds to a particular client. We have used this information to characterize the different types of customers. Then we created a matrix where we have a matrix in order to standardize the data.

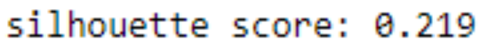


```
variables mean values:
------------------------------------------------------------------
-------------------------
 [   3.62305987 259.93189634 556.26687999 377.06036244  15.694542
1
  16.37327913  13.98907929  32.75310053  21.19884856]
```

Following we create clusters of customers using smaller base dimensions. We use this base for creating the representation of different clusters and later to verify the quality for the separation of different groups.

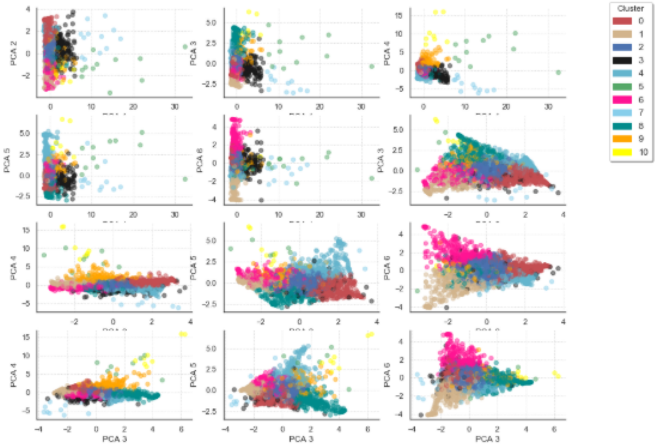


We have defined clusters of clients from the standardized matrix that was defined earlier and using the k means algorithm. We found out that using 11 clusters we get the best score.



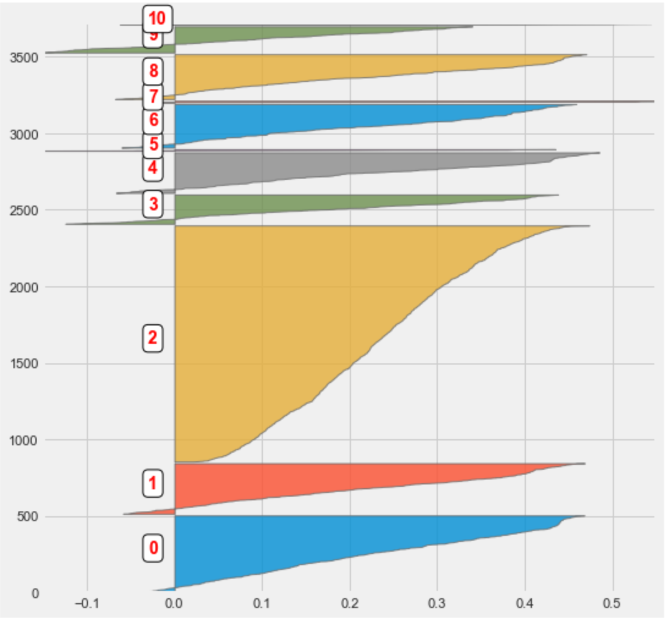
```
silhouette score: 0.219
```

We came across sizes having some disparity of different groups which have been created. Hence, we are trying to understand the quality of clustering that is done using:

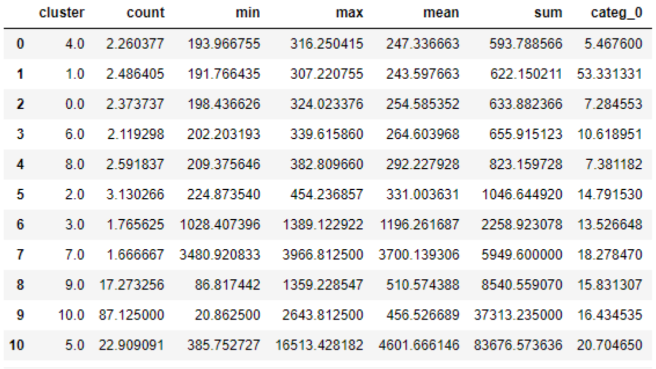


From the above visualization we see that the tiniest clusters are separated from the rest. The formed clusters are distinct from each other.

Next we tried looking at the silhouette scores for the quality of the separation.

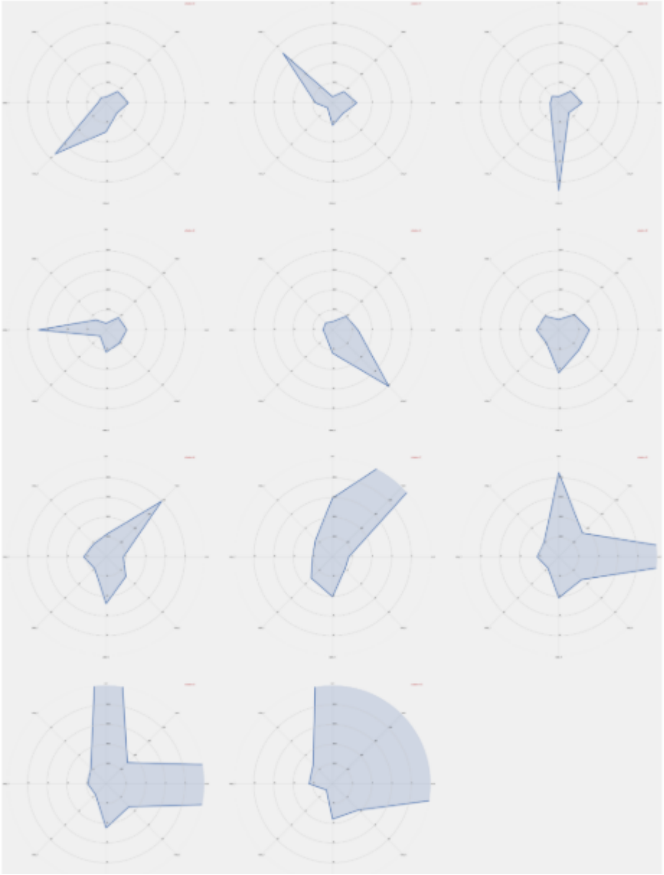


Here we verified that these clusters are disjoint. To understand habits of customers we started adding them to a different data frame where each client belongs to and averaging the data frame contents for the different group of clients. Then we reorganize the content based on the amount spent for each product category and according to the total amount spent.

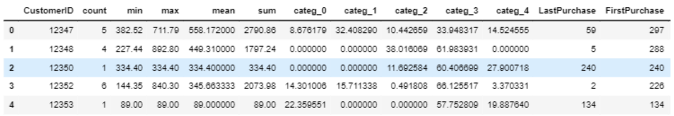


	cluster	count	min	max	mean	sum	categ_0
0	4.0	2.260377	193.966755	316.250415	247.336663	593.788566	5.467600
1	1.0	2.486405	191.766435	307.220755	243.597663	622.150211	53.331331
2	0.0	2.373737	198.436626	324.023376	254.585352	633.882366	7.284553
3	6.0	2.119298	202.203193	339.615860	264.603968	655.915123	10.618951
4	8.0	2.591837	209.375646	382.809660	292.227928	823.159728	7.381182
5	2.0	3.130266	224.873540	454.236857	331.003631	1046.644920	14.791530
6	3.0	1.765625	1028.407396	1389.122922	1196.261687	2258.923078	13.526648
7	7.0	1.666667	3480.920833	3966.812500	3700.139306	5949.600000	18.278470
8	9.0	17.273256	86.817442	1359.228547	510.574388	8540.559070	15.831307
9	10.0	87.125000	20.862500	2643.812500	456.526689	37313.235000	16.434535
10	5.0	22.909091	385.752727	16513.428182	4601.666146	83676.573636	20.704650

And finally, we did a representation of the different morphotypes, so we created a Radar Charts class so that it has a global view of the content of each cluster.



Here now we are classifying the customers into different client categories. This is being done so that the classification is possible in the first visit. Here the goal is to define for which class a client belongs to.

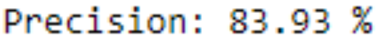


	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_4	LastPurchase	FirstPurchase
0	12347	5	382.52	711.79	558.172000	2790.86	8.676179	32.408290	10.442659	33.948317	14.524555	59	297
1	12348	4	227.44	892.80	449.310000	1797.24	0.000000	0.000000	38.016069	61.983931	0.000000	5	288
2	12350	1	334.40	334.40	334.400000	334.40	0.000000	0.000000	11.692584	60.406699	27.900718	240	240
3	12352	6	144.35	840.30	345.663333	2073.98	14.301006	15.711338	0.491808	66.125517	3.370331	2	226
4	12353	1	89.00	89.00	89.000000	89.00	22.359551	0.000000	0.000000	57.752809	19.887640	134	134

F. Running the classifiers on training data

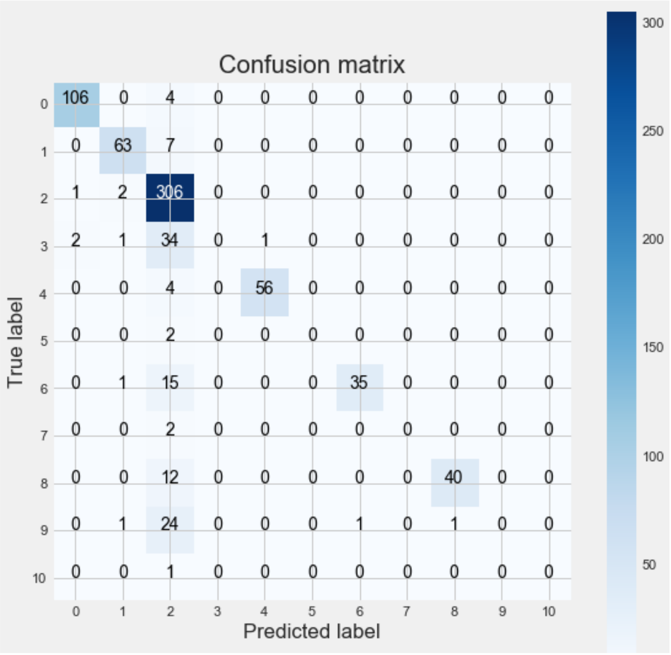
Support Vector Machine Classifier:

The first classifier which we are using is the Support Vector Machine Classifier. Here we have created an instance of a class and the call grid search function. Hyperparameters for which we seek optimal value and number of folds for cross validation are the two parameters which we have provided. After instance creation we adjusted the classifier to training data and further we test the quality of prediction. We observe that for every run there is change in the precision value.



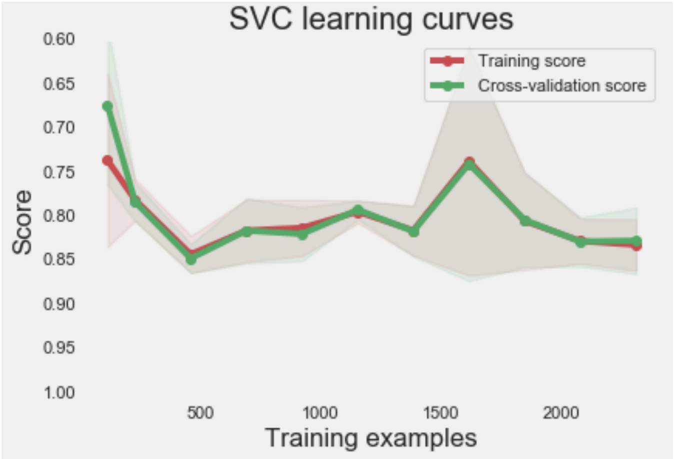
```
Precision: 83.93 %
```

We have created a confusion matrix without normalization among predicted label and true label.



Learning Curve:

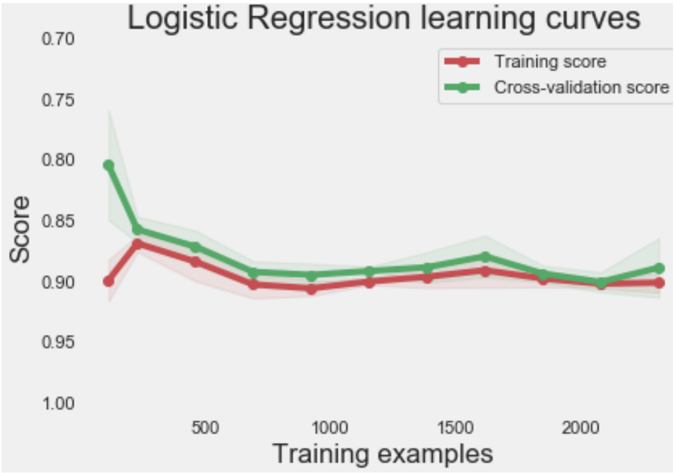
We have generated a learning curve to test the quality of a fit, to detect possible drawbacks in the model. This also would be useful to which extent the mode could benefit from a larger data sample. Below we are representing the learning curve of SVC classifiers. The below model does not support over fitting. The accuracy of the training curve is also correct in this plot.



Logistic Regression:

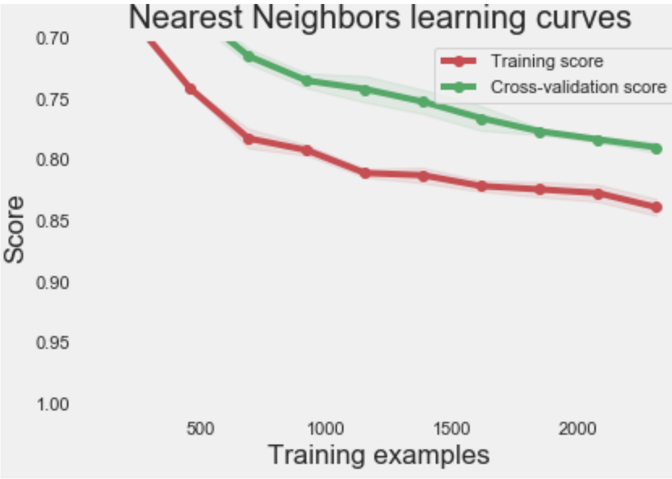
Now we are considering a logistic regression classifier. As done for the Support Vector Classifier we create an instance of a class and later adjust the model accordingly to see how the predictions are compared to the real values. And later we have plotted a curve.

Precision: 92.11 %



K - Nearest Neighbor:

Precision: 81.30 %

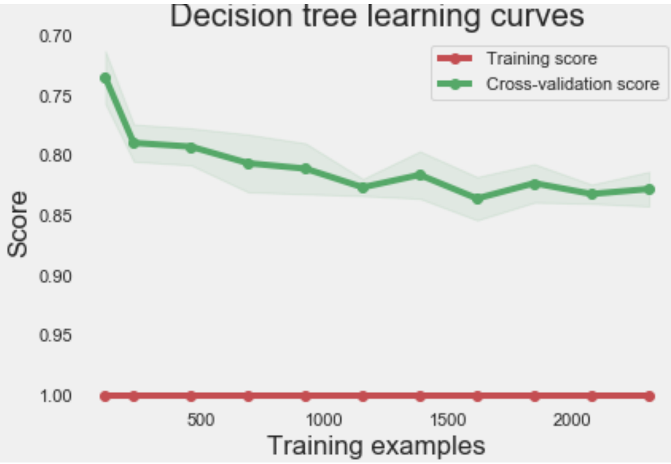


G. Decision Trees

Decision Trees (DTs) are one of the supervised learning methods used for both classification and regression. The purpose is to produce a model that predicts the value of a target variable by determining simple decision rules inferred from the data features. The main purpose that we use decision trees is for predictive modelling. It performs various operations and branches out into various outcomes and gives an end result. They are branched based on the field if it has a specific value, where a condition has to be met. They perform various rules and process information in a form of tree to get the final conclusion. By choosing the best attribute utilizing Attribute Selection Measures (ASM) to divide the records. Making the attribute a decision node and then further divide the dataset into smaller subsets. Begins tree by building the same process again and again until a final child node is achieved. The process continues until one of the conditions will match:

- All the tuples belong to the same attribute value.
- There are no more remaining attributes.
- There are no more instances.

Here we got a precision of 85.04 %



We have used this as it is easy to understand and to interpret, to understand easily.

H. Random Forest

Random forests work as follows:

- It produces a random target vector of 0s and 1s
- It creates a Random Forest classifier fitted to the random target vector
- It counts how frequently observations end up in the same terminal node

A random forest classification model consists of several decision tree classifiers {t(x,φk), k = 1,...}. Each decision tree t(x,φk) in the forest is constructed until they are fully grown. Here x is an input vector and φk is a random vector used to generate a bootstrap sample of objects from the training set D. The ideal number of trees in our random forest model was determined to be 500 by studying the Out Of-Bag (OOB) error rate in the training data. Let d be the dimensionality of the feature vector of the inputs. At each internal node of the tree, m features are selected randomly from the available d, such that m < d. m = √ d provided the highest accuracy among other common choices for m (1, 0.5 √ d, 2√ d, d). From the m chosen features, the feature that provides the most information gain is selected to split the node. Information gain can be defined as:

$$I_j = H(S_j) - \sum k\varepsilon(L,R) \, |S \; k \; j| \; |S| \; H(S \; k \; j)$$

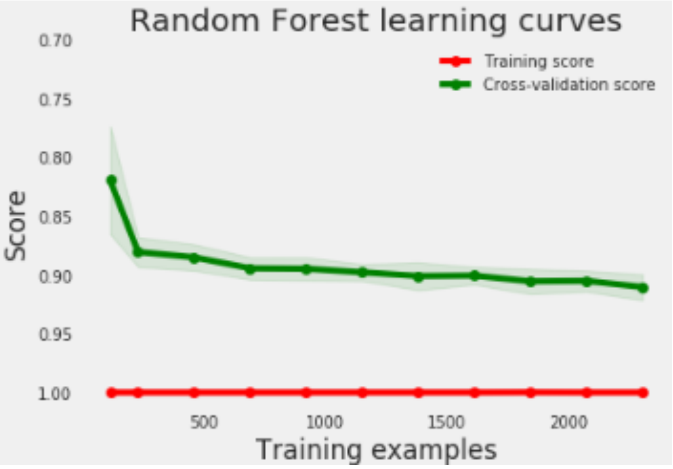
where Sj is the set of training points at node j, H(Sj) is the Shannon entropy at node j before the split, and S L j and S R j are the sets of points at the right child and left child respectively of the parent node j after the split.

The Shannon entropy can be defined as:

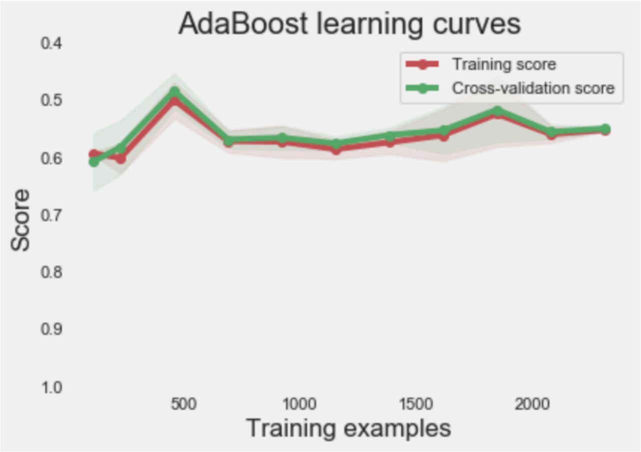
$$H(S) = -\sum c\varepsilon C \; pclog(pc)$$

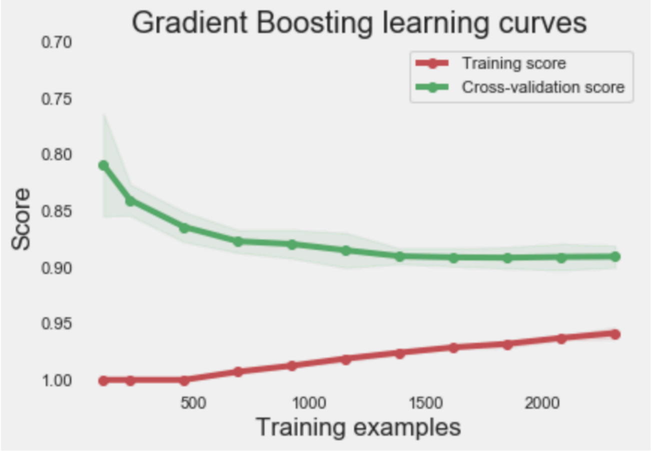
where S is the set of training points and pc is the probability of a sample being class c.

We have created a random forest classifier where we got a precision of 91.27%.

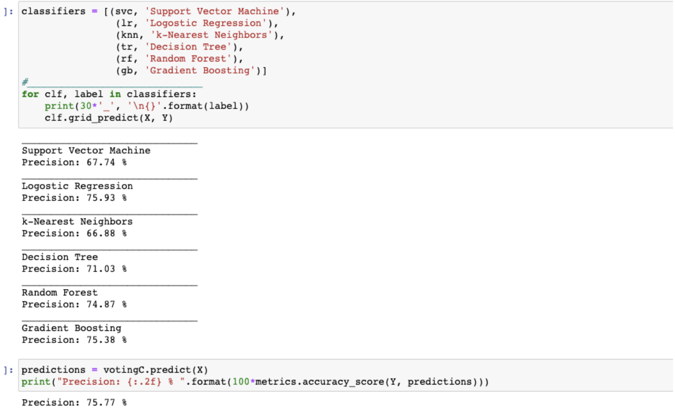


We have also done Gradient Boosting & AdaBoost Classifier and then we tried to compare all for the testing predictions. We have tried correcting the data in order to reference the time between the two datasets and to weight the variables count and sum to get the equivalence with the training set.





Testing Predictions:



```
]: classifiers = [(svc, 'Support Vector Machine'),
                   (lr, 'Logostic Regression'),
                   (knn, 'k-Nearest Neighbors'),
                   (tr, 'Decision Tree'),
                   (rf, 'Random Forest'),
                   (gb, 'Gradient Boosting')]
   #______________________________
   for clf, label in classifiers:
       print(30*'_', '\n{}'.format(label))
       clf.grid_predict(X, Y)

   ______________________________
   Support Vector Machine
   Precision: 67.74 %
   ______________________________
   Logostic Regression
   Precision: 75.93 %
   ______________________________
   k-Nearest Neighbors
   Precision: 66.88 %
   ______________________________
   Decision Tree
   Precision: 71.03 %
   ______________________________
   Random Forest
   Precision: 74.87 %
   ______________________________
   Gradient Boosting
   Precision: 75.38 %

]: predictions = votingC.predict(X)
   print("Precision: {:.2f} % ".format(100*metrics.accuracy_score(Y, predictions)))

   Precision: 75.77 %
```

We see that approximately there is 76% of customers who have been classified correctly.

V. Conclusion:

We observed that each entry in the dataset describes the purchase of a product by a particular customer and on a given date. Approximately 4300 customers appear in the dataset. Initially, we described the different products sold to the customer and also formed product categories using K-Means. During this stage, we grouped the different products into 5 main categories. Next, we classified the customers by analyzing their shopping habits for ten months. We have classified buyers into 11 major groups based on the type of products they usually buy, the number of visits they make in a year, and the amount they spend on average each time. After forming these products and customer categories, we have trained several classifiers like SVM, Logistic Regression, KNN, Decision Trees, Random Forest, Gradient Boosting to classify customers. As we have about 10 months of data, there is a highest accuracy of 76%. With more data the accuracy of the model increases.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset. The data was then processed in two steps: first, the set of data was used to define the category to which each client belongs, and then the classifier predictions were compared to this category assignment. We see that 75% of clients are awarded the right classes. The performance of the classifier can be further increased, given a large set of data.

References

[1] Index of /ml/machine-learning-databases/00502. (n.d.). Retrieved February 22, 2020, from https://archive.ics.uci.edu/ml/machine-learning-databases/00502/

[2] Chen, D. (2012). Data mining for the online retail industry: A case study of RFM model- based customer

segmentation using data mining. Journal of Database Marketing and Customer Strategy Management, 19, 197–208.

[3] D. M. J. Garbade, "Understanding K-means Clustering in Machine Learning," Medium,12-Sep-2018.[Online].Available:https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed: 06-Apr-2020]

[4] "Customer Segmentation Using K Means Clustering," KDnuggets. [Online].Available:https://www.kdnuggets.com/customer-segmentation-using-k-means-clustering.html/. [Accessed: 06-Apr-2020]

[5] "1.10. Decision Trees — scikit-learn 0.22.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. [Accessed: 06-Apr-2020]

[6] "6 Important Reasons Why Customer Support is Important | edu CBA," EDUCBA,

03-Feb-2017.[Online].Available:https://www.educba.com/customer-support/. [Accessed: 06-Apr-2020]

[7] A. V. Morgado, "The Value of Customer References to Potential Customers in Business Markets:," Journal of Creating Value, May 2018, doi:10.1177/2394964318771799.[Online].Available:https://journals.sagepub.com/doi/10.1177/2394964318771799. [Accessed: 06-Apr-2020]

[8] Ş. Ozan, "A Case Study on Customer Segmentation by using Machine Learning Methods," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, pp. 1–6, doi: 10.1109/IDAP.2018.8620892.

[9] C. Iyim, "Customer Segmentation with Machine Learning," *Medium*, 26-Feb-2020. [Online]. Available: https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84. [Accessed: 07-Apr-2020]

[10] S. Vivek, "Clustering algorithms for customer segmentation," *Medium*, 13-Aug-2018. [Online]. Available: https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac. [Accessed: 07-Apr-2020]

[11] *http://www.cs.bu.edu/~betke/papers/Joshi-etal-FG2015.pdf.* 2020.

[12] D. Poojari, "Machine Learning Basics: Descision Tree From Scratch (Part I)," Medium, 21-Feb-2020. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c. [Accessed: 08-Apr-2020]

[13] "K-Means Clustering: All You Need to Know," Byte Academy | Top Coding School For Python Fullstack Software Development, 17-Jul-2018. [Online]. Available: https://byteacademy.co/blog/k-means-clustering/. [Accessed: 08-Apr-2020]

[14] W. Koehrsen, "Random Forest in Python," Medium, 17-Jan-2018. [Online]. Available: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0. [Accessed: 11-May-2020]

[15] C. Maklin, "Decision Tree In Python," Medium, 27-Jul-2019. [Online]. Available: https://towardsdatascience.com/decision-tree-in-python-b433ae57fb93. [Accessed: 11-May-2020]

[16] C. Maklin, "AdaBoost Classifier Example In Python," Medium, 22-Jul-2019. [Online]. Available: https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464. [Accessed: 11-May-2020]