Data Science Capstone Project

Customer Segmentation

On Ecommerce Transaction Data

Content

Introduction

E-commerce has exploded like never before, resulting in a more fierce competition among online businesses. Competition, Order Fulfillment and Customer Experience are the top challenges facing e-commerce businesses today. It’s all about the conversion – and data integration.

To successfully compete retailers must stay on top of their vast data volumes, including everything from digital images, to customer pricing and invoicing, to marketing and promotions. The explosion of digital technology and big data has changed the game for e-commerce marketers, and data-driven marketing is now a necessity for online retailers who want to retain their competitive edge.

Business Problem

This real life Europe based E-Commerce Retailer is in a competitive business of selling distinctive gifts and presents for occasions and celebrations. Most of the online retailers customers are wholesalers. It's important for the business to retain its competitive advantage over its peers with data driven solutions that give it an edge more so today than ever before, while the world reels in a pandemic.

Can the retailer's invoice data provide insights into any of the subjects listed below?

- Segment a product into cetegories to enable improved product sales
- Segment a customer into categories to enable targeted digital marketing strategies

Additionally, Below topics will also be optionally explored -

- Product sales trends for customer categories
- Customer product preferences based on Geography
- Geographic customer money expenditure trends
- Seasonal Sales increases and decreases
- Time of the day when sales are the most

Data acquisition

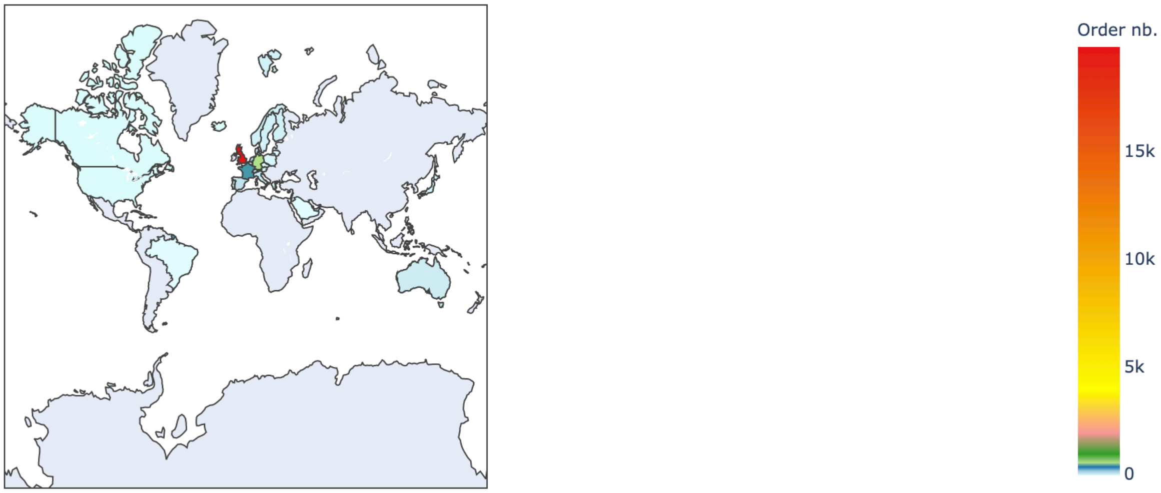
The European E-Commerce retailer's real life data is an intercontinental data set which contains all sales invoices occurring between 2010 and 2012. The data set contains a modest 8 attributes that contain the sales details for the retailer from various geographies. The features in the dataset are -

- InvoiceNo (examples: 936365, 9363656, 936367, 9363658)
- StockCode (examples: 34564D, 345754F, 567557H, 85123A)
- Description (examples : WOOD S/43 CABINET ANT FINISH, PINK GIRLY TOOL SET, SET OF 6 FUNKY BEAKERS)
- Quantity (examples: 5,10,30,100)
- InvoiceDate (examples: 11/1/2010 9:53, 12/2/2010 9:59, 12/3/2010 9:59)
- UnitPrice (examples: 4.95, 3.29,15.95, 8.99)
- CustomerID (examples: 137850 , 778546, 556455, 787853)
- Country (examples: India, Nepal, Sri Lanka, Bhutan) The scope to analyze this dataset is substantial and can cover areas such as time series trends, customer clustering, classification and more.

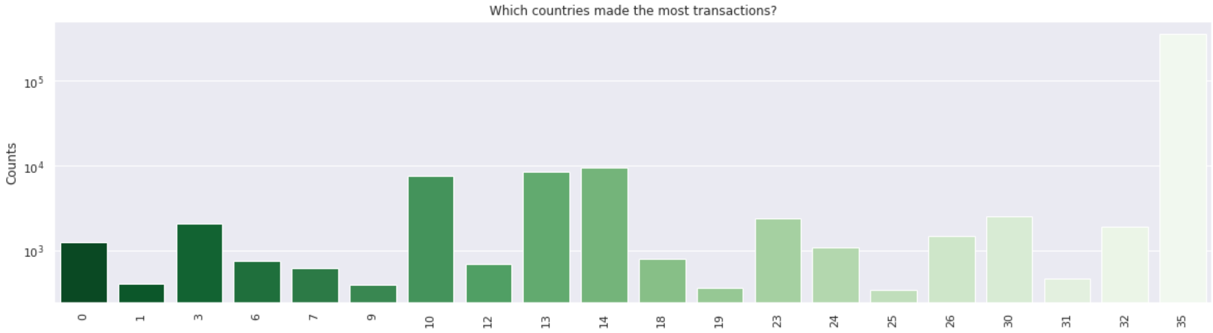
For the Capstone project we will limit the analysis to Customer Segmentation and Purchase Prediction. It remains to be seen which of these classification model best serves in distinguishing the customers into various segments - Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Decision Tree, or if any of the cutting edge models such as Gradient Boosting Trees, AdaBoost Classifier, or Deep Learning will provide boosted accuracies.

Exploratory Data Analysis

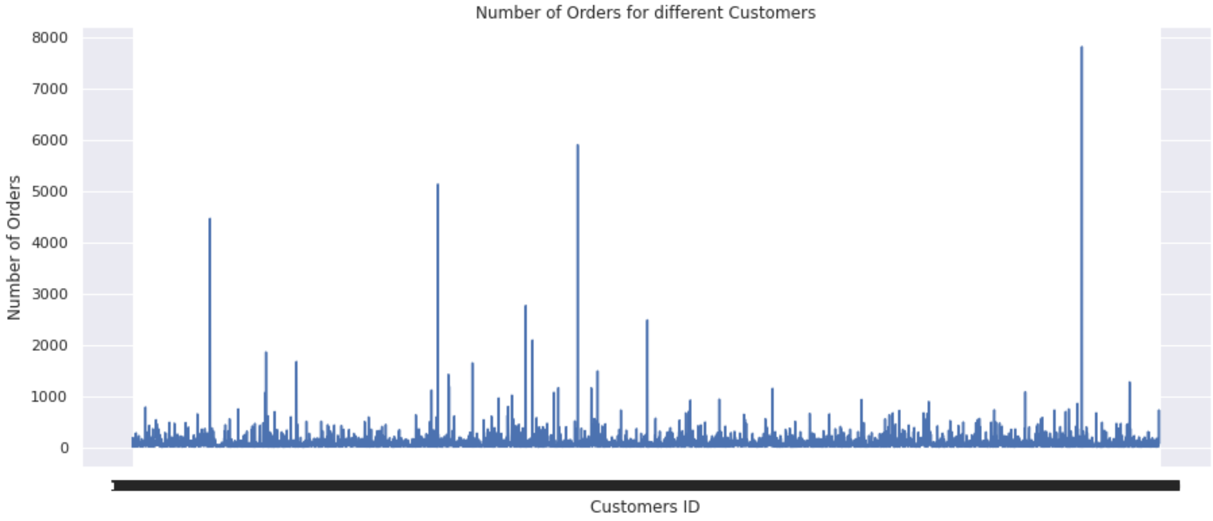
Choropleth of number of orders placed per geographic location. Europe appears to be the place from where most orders are placed.



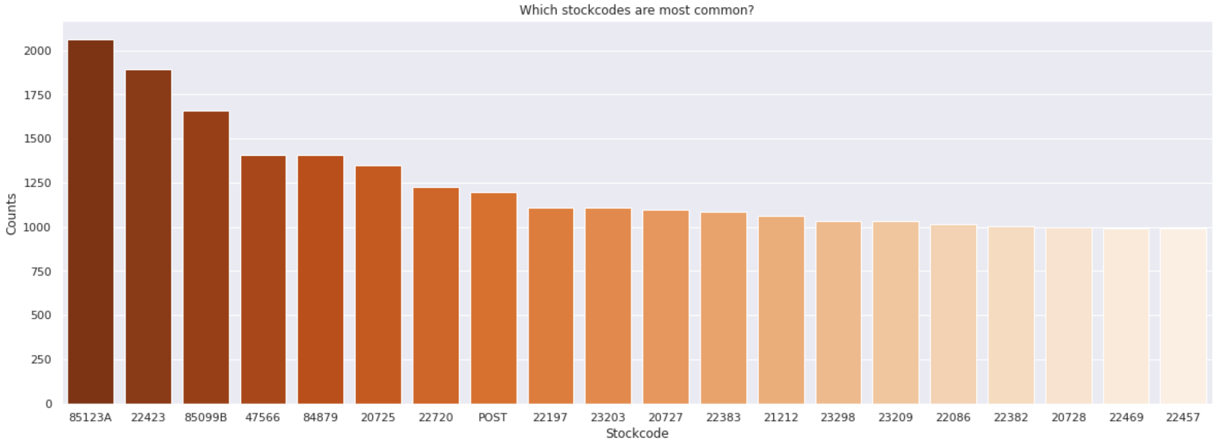
View the countries which drive the most traffic and sales to the ecommerice site.

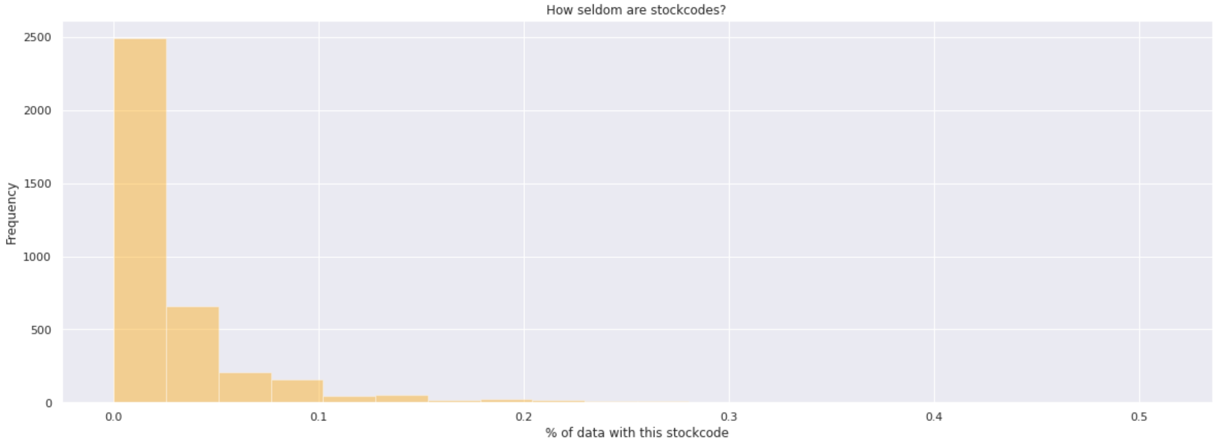


Money Spent by different Customers

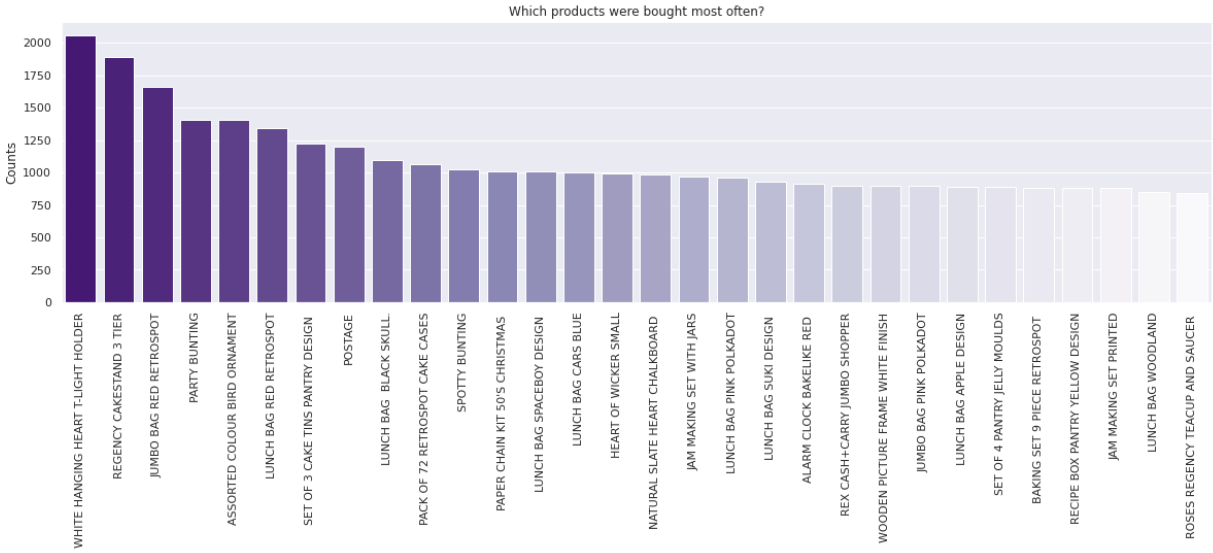


Visualize the frequent Stockcodes purchased by customers.

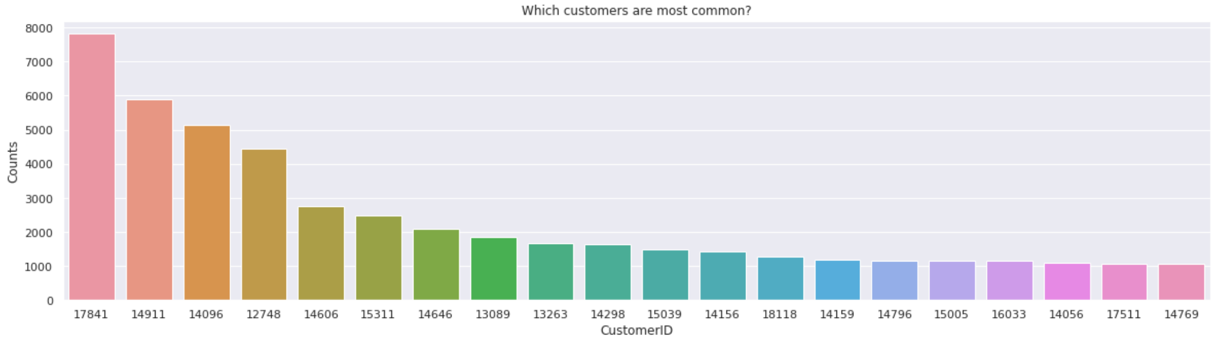




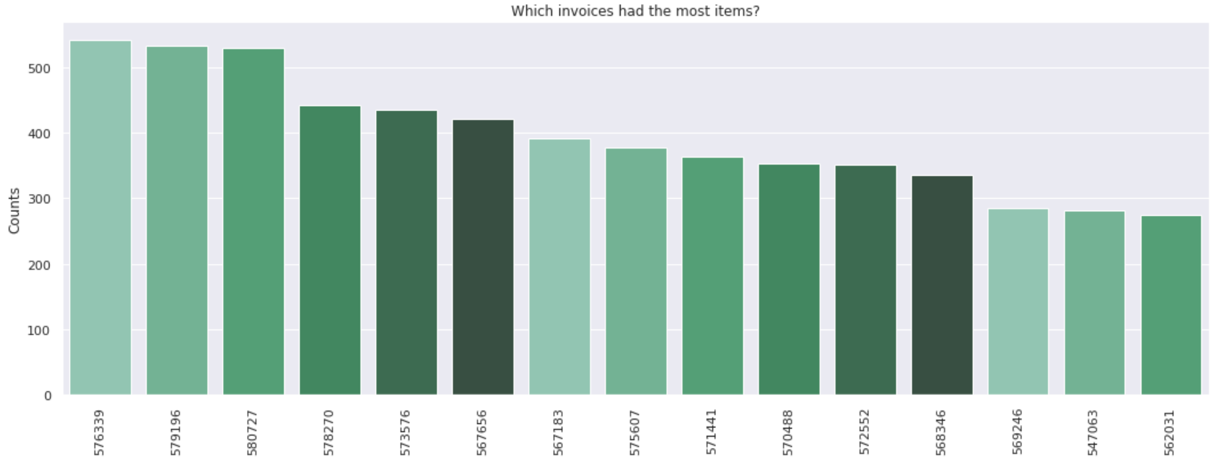
Visualize popular products among customers.



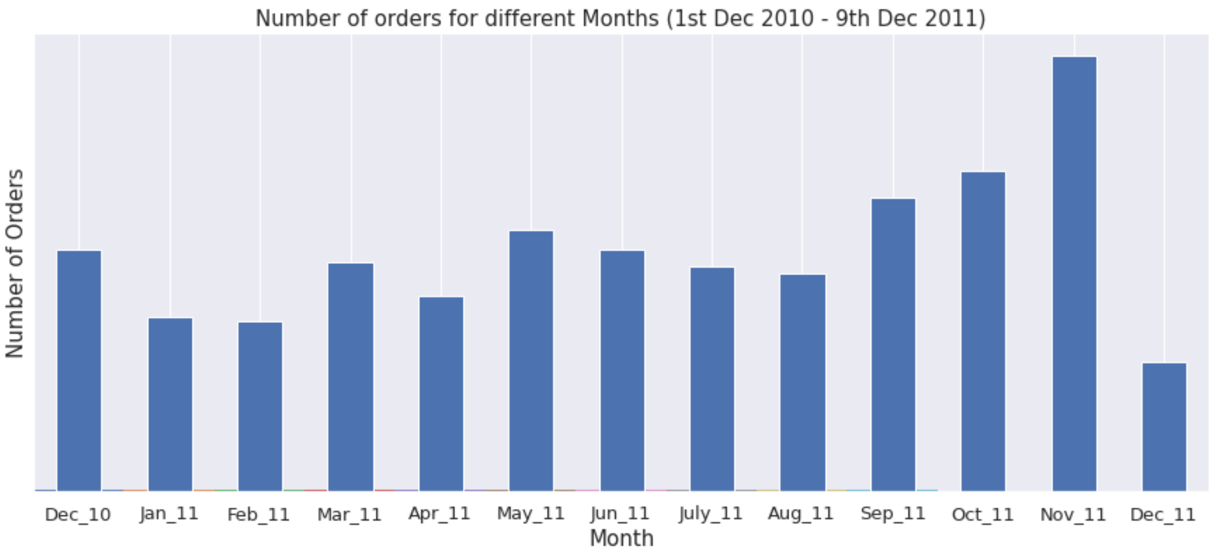
View the customers who are most loyal.

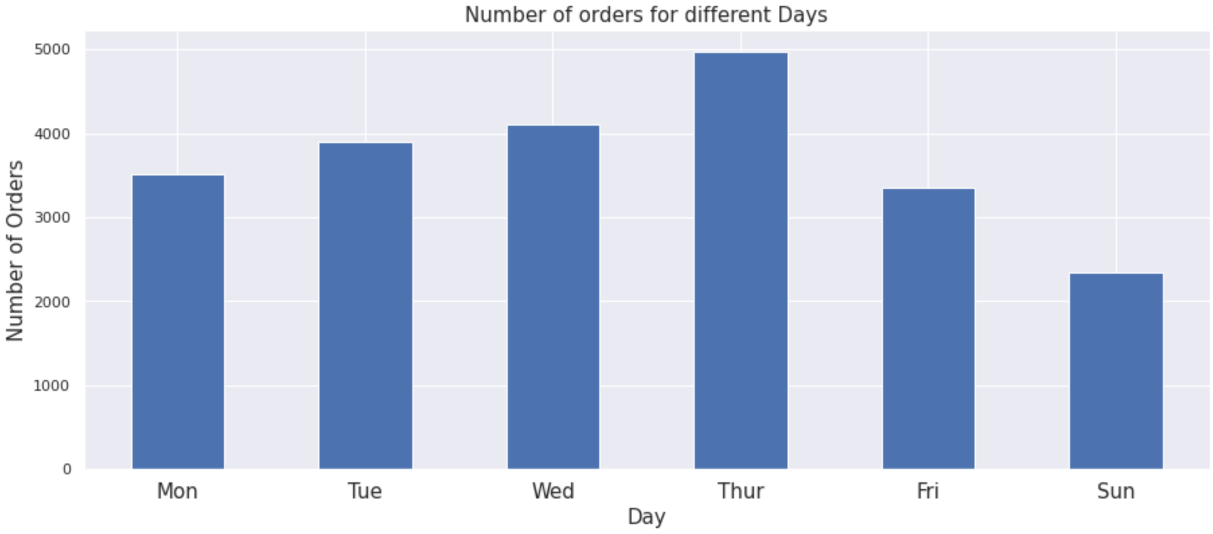


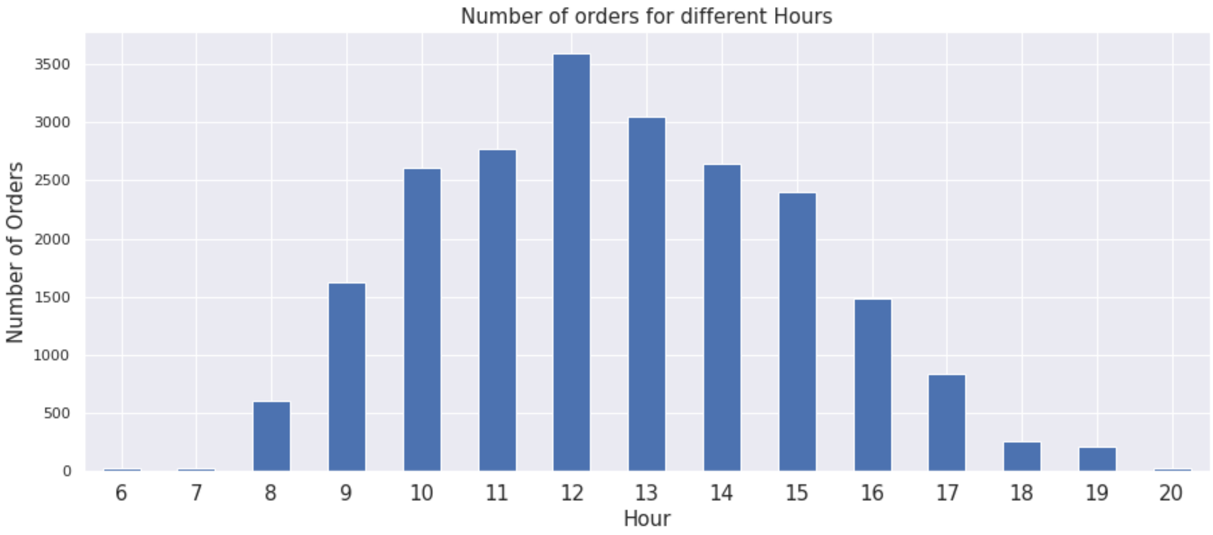
Invoice Analysis.



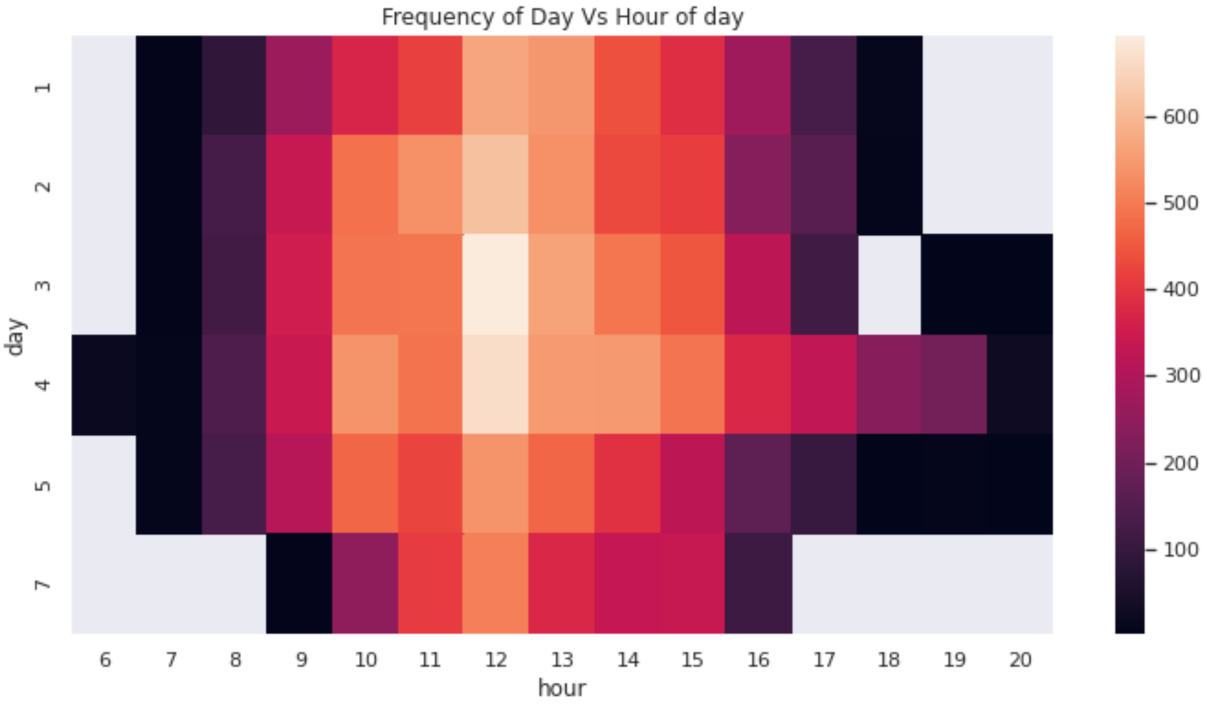
Number of orders per month, week and day.


Number of orders for different Months (1st Dec 2010 - 9th Dec 2011)


Number of orders for different Days


Number of orders for different Hours

Heatmap of popular time periods when orders are placed.



Data Preprocessing

- We couldn’t impute or replace the missing values and we couldn’t keep data without the value in the customer id columns since we want to classify the customers. So we dropped the lines with missing values on the customer ID column.
- More than 90% of the data is coming from UK
- The stock code values aren't only numerical, there are speciales values like D which means Discount
- The InvoiceNo aren't also only numerical since there is a C before the other numbers for every negative value in the quantity column, this could mean that the order was canceled. It appears that more than 16% of the transactions were canceled which is significant. These are specific operations which doesn't characterize our customers so we just dropped these transactions from our dataset.
- It seems that the customer can also cancel just a part of the transaction which is logical so we need to take this into account for later.
- It was tempting to replace the null values by the most common one but it might

be a special discount or something else so it was left as it is.

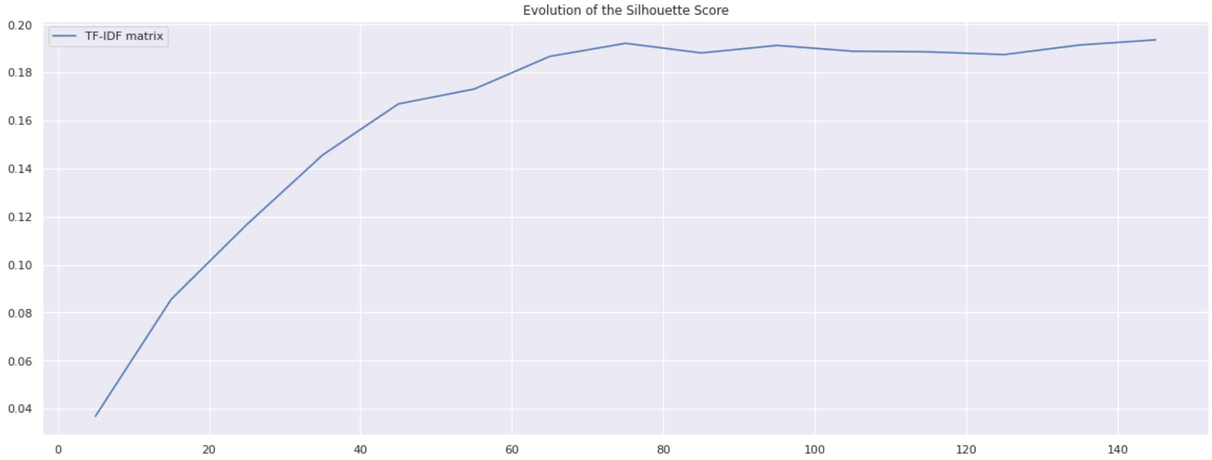
- Removed the items that got completely canceled in order to harmonize the future clusters and not have too much special values.
- We have implemented here the RFM principle to classify the customers in this database. RFM stands for Recency, Frequency and Monetary. It is a customer segmentation technique that uses past purchase behavior to divide customers into groups.
- Finally we'll set a score for each customer in the database.

Predictive ML Modeling

Clustering Products into Categories

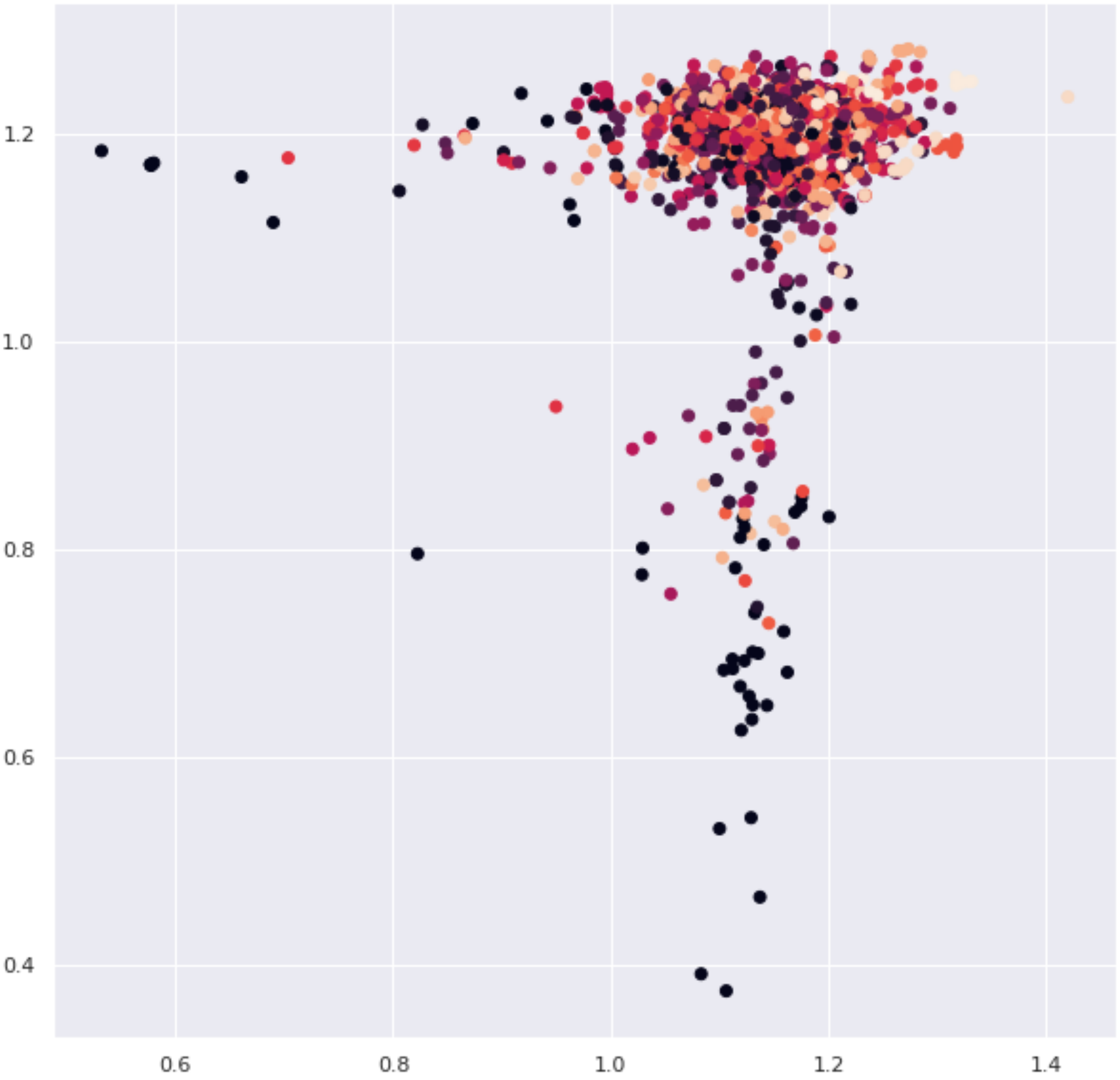
KMeans was used to cluster products into various categories based on the text description each product has in Description column.

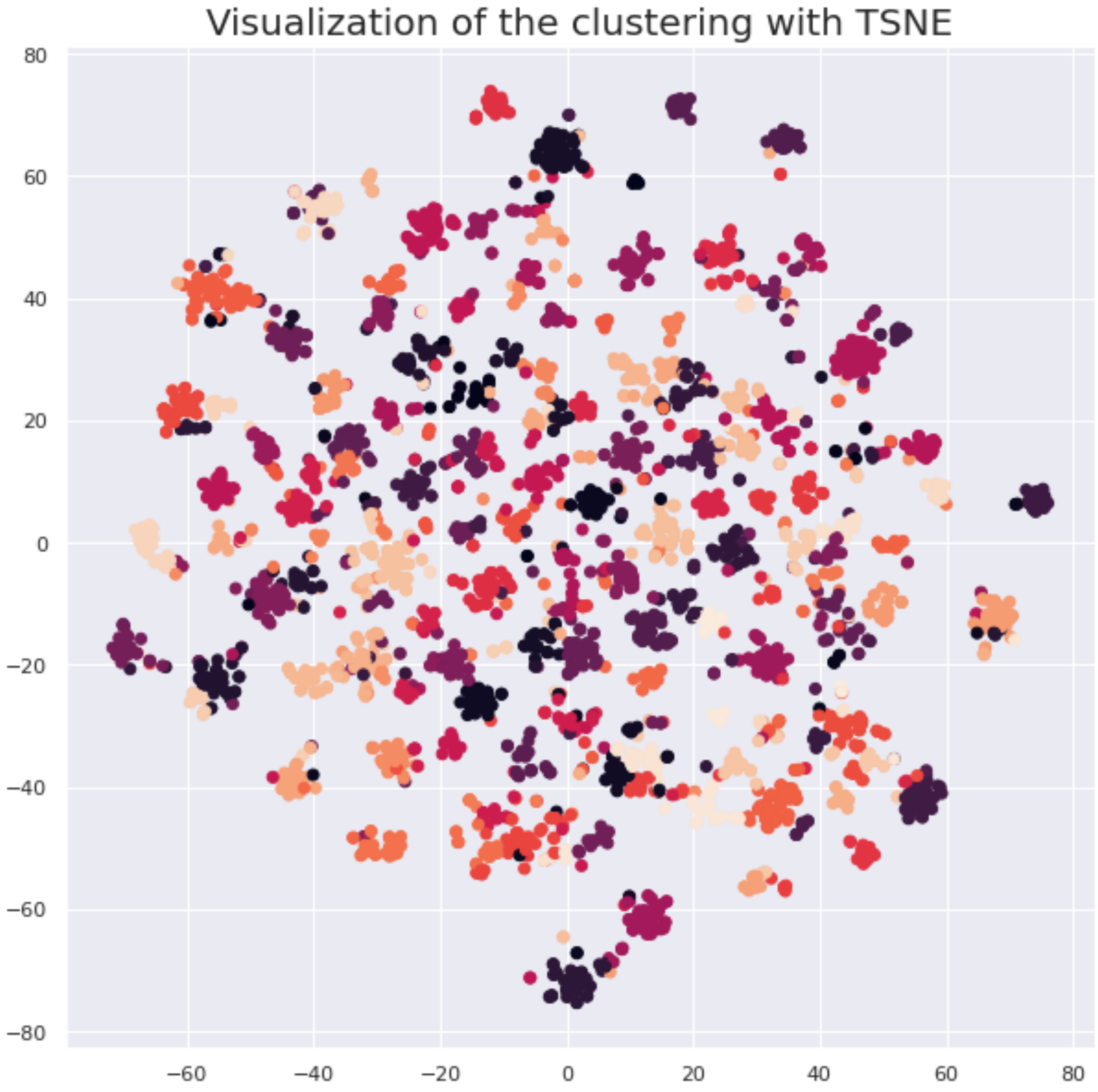
Below graph was built based on different values of K to find optimum K for the model.



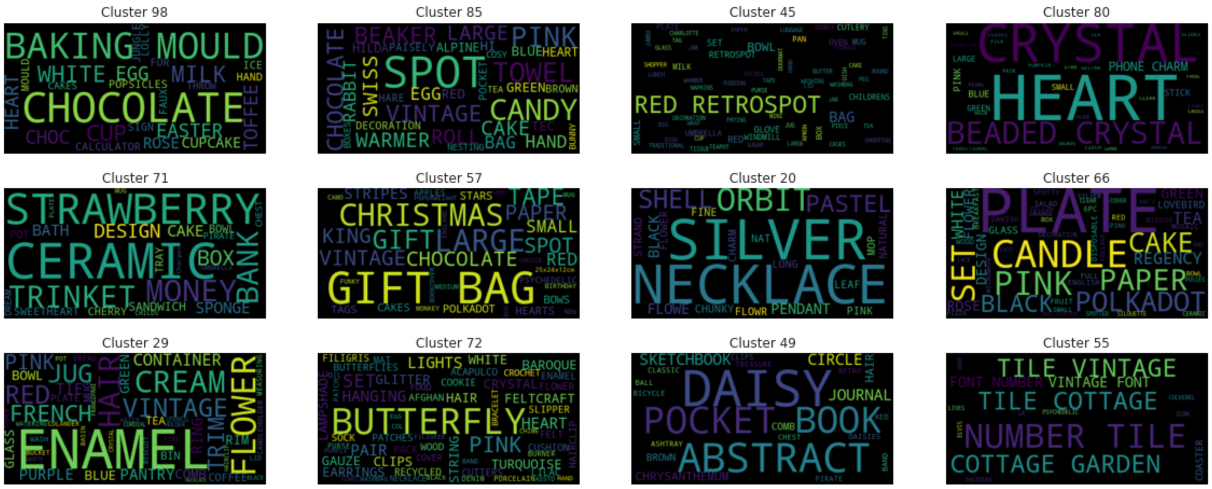
Best performance was at n_clusters = 135

ACP with 135 clusters



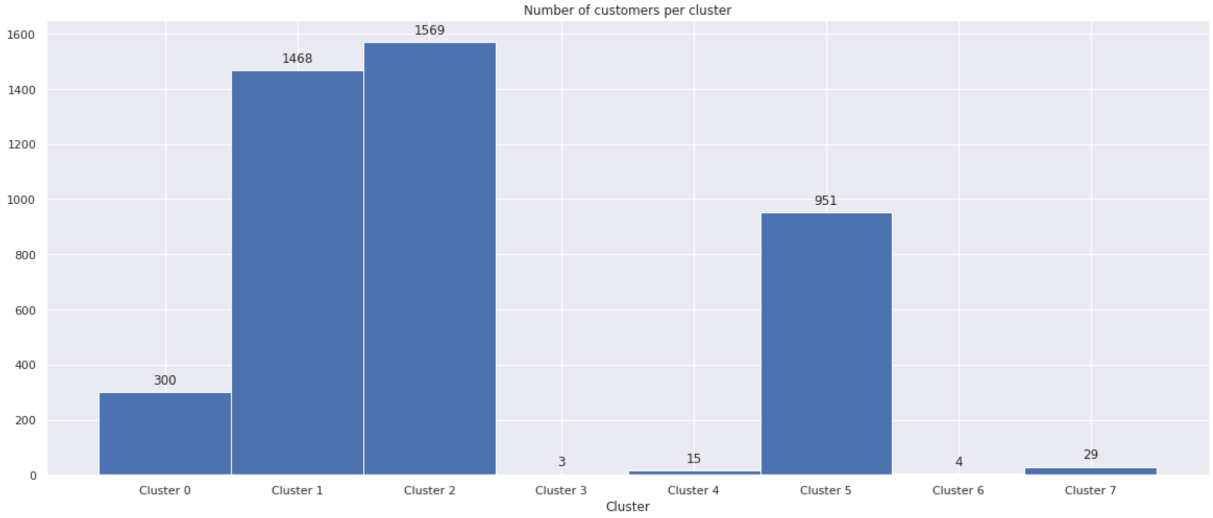


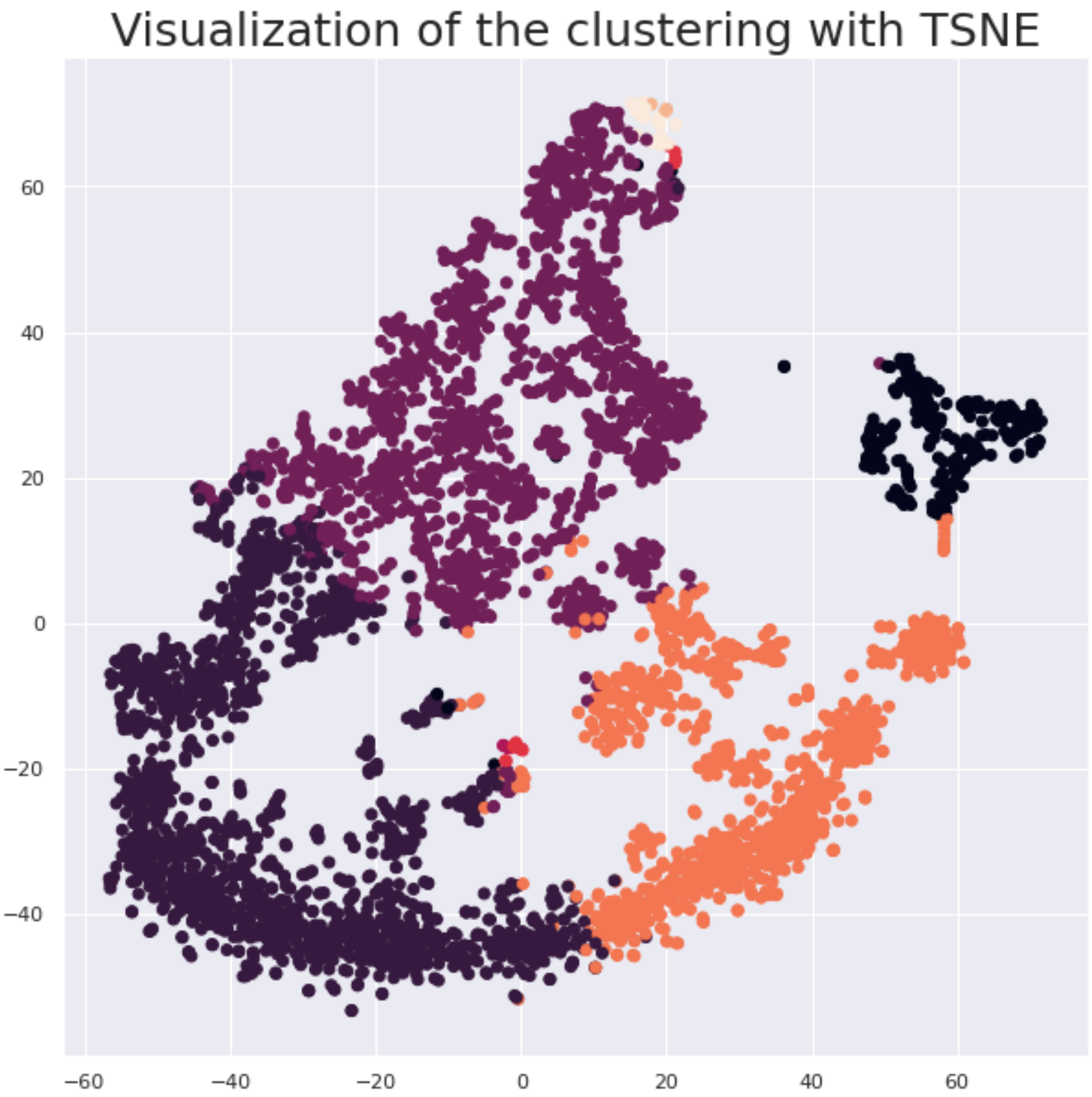
Word Cloud of most popular product description keywords among the product clusters



Customer Segmentation - Clustering customers into appropriate groups

Kmeans was used for Customer segmentation as well. We want to have at least 5, 6 clusters so we didn't take 2 or 3 clusters even though they have the highest silhouette scores, 8 clusters was fit the best here.





Results and Discussion

Lets see the results per cluster.

Cluster 0

These customers seems to be good since they have good RFM scores, the 4 most represented categories are (111, 211, 322, 222). They seem to be normal customers.

Key figures:

- Min Basket Price: 10.86
- Mean Basket Price: 30.60
- Max Basket Price: 68.57
- Quantity: 10.00
- UnitPrice: 2.87
- QuantityCanceled: 0.04
- TotalPrice: 17.09
- Frequency 11.134050
- Recency 24.574626

TOP 10 bought products :

- WHITE HANGING HEART T-LIGHT HOLDER: 1345
- JUMBO BAG RED RETROSPOT: 1079
- REGENCY CAKESTAND 3 TIER: 960
- ASSORTED COLOUR BIRD ORNAMENT: 926
- PARTY BUNTING: 924
- LUNCH BAG RED RETROSPOT: 898
- LUNCH BAG BLACK SKULL: 753
- SET OF 3 CAKE TINS PANTRY DESIGN: 725
- LUNCH BAG CARS BLUE: 679
- LUNCH BAG PINK POLKADOT: 676

Cluster 1

This cluster represents almost lost customers. The weird part about them is that there are some months when they didn't shop at all, it looks like a pattern.

Key figures:

- Min Basket Price: 20.83
- Mean Basket Price: 33.77
- Max Basket Price: 26.43

- Quantity: 9.06
- UnitPrice: 2.68
- QuantityCanceled: 0.02
- TotalPrice: 13.77
- Frequency 3.065758
- Recency 36.131902

TOP 10 bought products :

- PAPER CHAIN KIT 50'S CHRISTMAS: 267
- BAKING SET 9 PIECE RETROSPOT: 263
- WHITE HANGING HEART T-LIGHT HOLDER: 250
- ASSORTED COLOUR BIRD ORNAMENT: 247
- REX CASH+CARRY JUMBO SHOPPER: 223
- HOT WATER BOTTLE KEEP CALM: 215
- REGENCY CAKESTAND 3 TIER: 208
- RABBIT NIGHT LIGHT: 200
- GARDENERS KNEELING PAD KEEP CALM: 194
- SPOTTY BUNTING: 193

Cluster 2

The cluster 2 represents the best customers with a high recency which have around 60 visits, a lot of quantity bought on average, a high moneraty value and also a high frequency around 60 visits. These customers must be taken care.

- Min Basket Price : 13
- Mean Basket Price : 513
- Max Basket Price : 3812
- Quantity 117.083422
- UnitPrice 2.830180
- QuantityCanceled 0.069282
- TotalPrice 258.683970
- frequency 67.812655
- min_recency 1.679039

Cluster 3

This cluster is full of lost customers. Indeed, as we can see in the month histogramm there are almost no invoices after july. We can see that there are in december but it's december of the past year. So this cluster is pretty bad, they don't want to have new customers in there. Furthermore they are cheap customers since the mean basket price is 28.91$.

Key figures:

- Min Basket Price: 24.20
- Mean Basket Price: 28.91
- Max Basket Price: 34.52
- Quantity: 8.25
- UnitPrice: 3.29
- QuantityCanceled: 0.04
- TotalPrice: 15.20
- Frequency 2.606359
- Recency 237.013433

TOP 10 bought products :

- WHITE HANGING HEART T-LIGHT HOLDER: 227
- REGENCY CAKESTAND 3 TIER: 182
- PARTY BUNTING: 137
- ASSORTED COLOUR BIRD ORNAMENT: 125
- REX CASH+CARRY JUMBO SHOPPER: 103
- SET OF 3 CAKE TINS PANTRY DESIGN: 100
- NATURAL SLATE HEART CHALKBOARD: 100
- JAM MAKING SET WITH JARS: 99
- HEART OF WICKER SMALL: 98
- HEART OF WICKER LARGE: 86

Cluster 4

This cluster is quiete heterogeneous since there are 17 best customers, 6 lost cheap customers and so on. They do have a high mean basket price of 505 but it's mostly due to the mean quantity they buy (130) because the mean unit price is very low (3.26)

For the time features, what is interesting is that these customers shop less on weekend and they shopped more at the end of the year.

- Min Basket Price : 247
- Mean Basket Price : 505
- Max Basket Price : 1023
- Quantity 130.299145
- UnitPrice 3.264359
- QuantityCanceled 2.332590
- TotalPrice 184.308595

TOP 10 products bought :

- JUMBO BAG RED RETROSPOT : 38
- BLACK RECORD COVER FRAME : 31
- RECORD FRAME 7" SINGLE SIZE : 28
- REGENCY CAKESTAND 3 TIER : 25
- WORLD WAR 2 GLIDERS ASSTD DESIGNS : 24
- WHITE HANGING HEART T-LIGHT HOLDER : 24
- PARTY BUNTING : 24
- LUNCH BOX I LOVE LONDON : 23
- RED HARMONICA IN BOX : 23
- CHILLI LIGHTS : 23

Cluster 5

The cluster 5 contains 3 customers which are very much alike. Indeed, they bought only once or twice a few items at a huge quantity. It might be some profesionnals which bought it at discount and will sell back the commodity. Even if they have a high monetary value they're not very interesting and we could consider them as lost customers.

- Min Basket Price : 3368

- Mean Basket Price : 3697
- Max Basket Price : 3533
- Quantity 2213.777778
- UnitPrice 2.386667
- QuantityCanceled 0.000000
- TotalPrice 3890.091111
- Frequency 1.666667
- Min_recency 210.888889

Cluster 6

What is very specific about this cluster is that there are no customers from UK, it's only foreign countries (Germany, France, Belgium, Italy and Finland). This cluster is also heterogeneous in terms of RFM since the 2 most represented categories are Best customer and Lost cheap customer. The average basket is very low (33) comparing the ones above but I guess that the more customers we have in a cluster and the more the average customer will be represented which doesn't spent 500$ per transactions like the ones above.

October and november have the most invoices which isn't surpring approaching Christmas.

Key figures:

- Min Basket Price: 20.58
- Mean Basket Price: 33.55
- Max Basket Price: 59.31
- Quantity: 13.785663
- UnitPrice: 2.884687
- QuantityCanceled: 0.057975
- TotalPrice: 23.749951
- Frequency 7.865563
- Recency 46.622343

TOP 10 bought products :

- ROUND SNACK BOXES SET OF4 WOODLAND: 233
- REGENCY CAKESTAND 3 TIER: 161
- PLASTERS IN TIN WOODLAND ANIMALS: 150
- ROUND SNACK BOXES SET OF 4 FRUITS: 146
- RED TOADSTOOL LED NIGHT LIGHT: 144
- PLASTERS IN TIN CIRCUS PARADE: 141
- SPACEBOY LUNCH BOX: 137
- RABBIT NIGHT LIGHT: 120
- PLASTERS IN TIN SPACEBOY: 120
- WOODLAND CHARLOTTE BAG: 111

Cluster 7

The cluster 7 contains 19 customers who are considered as best customers since they by the most, very frequently (75) and recently. The difference with cluster 2 is that they cluster 7's customers buy more frequently (75 vs 60) but have a lower monetary value (58000 vs 249000). They have a mean basket price lower than the other clusters.

- Min Basket Price : 10
- Mean Basket Price : 138
- Max Basket Price : 648
- Quantity 23.257769
- UnitPrice 2.615444
- QuantityCanceled 0.109129
- TotalPrice 34.916436
- Frequency 121.570291
- Recency 2.599109

TOP 10 products bought :

- REGENCY CAKESTAND 3 TIER: 136
- JUMBO BAG RED RETROSPOT: 135
- WHITE HANGING HEART T-LIGHT HOLDER: 121
- CHILLI LIGHTS: 102
- PAPER BUNTING RETROSPOT: 97
- LUNCH BAG BLACK SKULL: 95
- GUMBALL COAT RACK: 93
- LUNCH BAG RED RETROSPOT: 91

- JUMBO BAG PINK POLKADOT: 84
- LUNCH BAG CARS BLUE: 81

Conclusion

Let's quickly classify the clusters in terms of importance :

Cluster 2:

High frequency with a lot of quantity (mean basket price of 513) bought on average and high monetary value (VIP clients)

- The cluster 2 represents the best customers with a high recency which have around 60 visits, a lot of quantity bought on average, a high moneraty value and also a high frequency around 60 visits.
- These customers must be taken care.

Cluster 7 :

Very high purchase frequency with a mean basket price of 150 but good monetary value.

- The cluster 7 contains 19 customers who are considered as best customers since they by the most, very frequently (75) and recently.
- The difference with cluster 2 is that they cluster 7's customers buy more frequently (75 vs 60) but have a lower monetary value (58000 vs 249000).
- They have a mean basket price lower than the other clusters.

Cluster 4:

Very high basket price (huge quantity of products bought on average)

- This cluster is quiete heterogeneous since there are 17 best customers, 6 lost cheap customers and so on.
- They do have a high mean basket price of 505 but it's mostly due to the mean quantity they buy (130) because the mean unit price is very low (3.26)
- For the time features, what is interesting is that these customers shop less on

weekend and they shopped more at the end of the year.

Cluster 0:

Good average customers

- These customers seems to be good since they have good RFM scores, the 4 most represented categories are (111, 211, 322, 222). They seem to be normal customers.

Cluster 6:

Good foreign customers

- What is very specific about this cluster is that there are no customers from UK, it's only foreign countries (Germany, France, Belgium, Italy and Finland).
- This cluster is also heterogeneous in terms of RFM since the 2 most represented categories are Best customer and Lost cheap customer.
- The average basket is very low (33) comparing the ones above, but the more customers we have in a cluster and the more the average customer will be represented which doesn't spent 500$ per transactions like the ones above.
- October and november have the most invoices which isn't surpring approaching Christmas.

Cluster 1:

This cluster represents almost lost customers.

- The strange part about them is that there are some months when they didn't shop at all, it looks like a pattern.

Cluster 5:

Highest monetary value but only one or two purchases over the year

- The cluster 5 contains 3 customers which are very much alike. Indeed, they bought only once or twice a few items at a huge quantity.
- It might be some profesionnals which bought it at discount and will sell back the commodity.
- Even if they have a high monetary value they're not very interesting and we could consider them as lost customers.

Cluster 3:

Lost customers

- This cluster is full of lost customers.
- Indeed, as we can see in the month histogramm there are almost no invoices after july.
- We can see that there are in december but it's december of the past year. * So this cluster is pretty bad, they don't want to have new customers in there.
- Furthermore they are cheap customers since the mean basket price is 28.91$