255 Data Mining - Final Project Report

Milestone-3

Shalabh Neema 014546259

Purvi Misal 014544621

Nimit Patel 010700196

Alok Goyal 014499355

Submitted To: Prof. Carlos Rojas

Introduction

We have chosen 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. We will also apply analysis methods performed on this dataset to another dataset “ces_hybrid_real_world” and compare the results.

Motivation

Most of us have prior work experience with companies looking to sell some products or offers to the customers for the retail market.

If we observe carefully, every month there is some or the other kind of sale or promotions being run to entice the customers to buy more.

For e.g. In India, we used to have a promotion or campaign every month; given the diverse populations and culture, we always had some festival or day to celebrate and the E-commerce players looking to cash in the popularity created by these festivals would run promotions every month.

One thing which we learned throughout was, it is not easy to convince a user to buy your product. It is even harder, if you don’t even know if he/she has any interest in the product you are trying to sell and makes you wonder if you’re wasting your time and resources to the wrong audience. This came as a realization in terms of marketing and advertising costs, which always have to be justified by a solid ROI (Return on Investment) number.

Having closely worked with online retailers in day-to-day promotions and always having the curiosity to understand our userbase, data-mining presents an excellent opportunity to gain some insights into this world.

We, therefore, chose this dataset to answer the question of how the transaction history of customers/consumers can give insight into consumers’ purchasing habits and also predict the products consumers might be interested in buying in the future. This kind of information can be used to align business decisions and also to understand which consumers are most valuable to the retail store, along with other essential insights.

The data set contains transactions occurring for a UK-based non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion giftware, having many customers that are wholesalers.

Previous Work Summary on the Dataset

- K-means Clustering based on Product Description.
- Classification of Customers clusters using prediction models like Logistic Regression, KNN, Decision Trees and Random Forest etc.
- RFM Analysis.
- Cohort Analysis.
- Sales Forecasting.
- Market Basket Analysis.
- Recommendation system.
- Clustering using DBSCAN and Cure algorithm.
- Customer Segmentation based on Products.

New Approach/Methods

We have explored the following methods that have the potential to explore more in-depth analysis than the previous work done on this dataset.

- Customer Segmentation based on value. (Value based Segmentation). Here value can be revenue, completed transactions, family members etc.
- Hierarchical Clustering using a dendrogram method available at SciPy.
- Two step cluster analysis: This method identifies groupings by running pre clustering first and then by running hierarchical methods.
- Predicting nth item based on (n-1) previously purchased items using cosine similarity distance.
- Apply the analysis methods performed of the Online-Retail dataset on the ces_hybrid_real_world dataset.

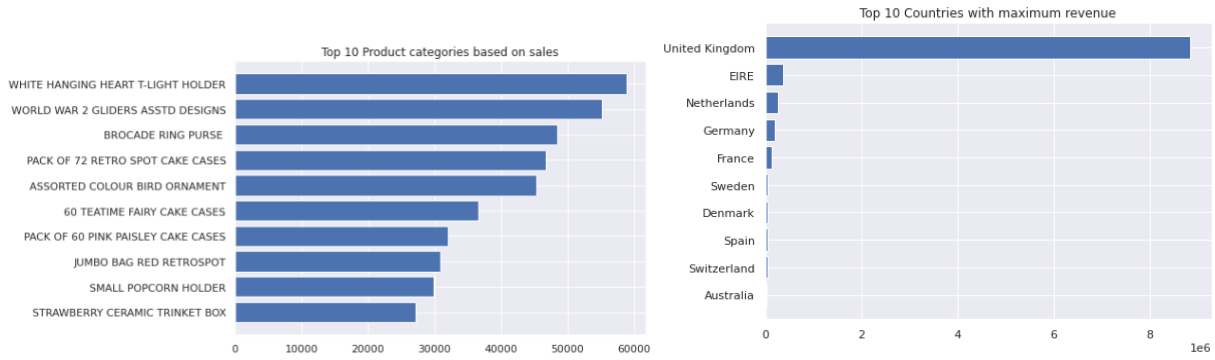
Exploratory Data Analysis:

During the Exploratory Data Analysis of our datasets, we tried to analyze consumer trends and patterns of buying and also the countries that reported having most sales and revenues for the retail brand. We also did RFM analysis, which is a common technique to determine the best customers quantitatively by computing how recently a consumer has purchased (Recency), how often they purchase (Frequency) and how much the customer spends (Monetary). This is a popular way to analyze retail datasets better.

We have tried to explore the following questions in our EDA:

- What are the top 10 Product categories based on sales?
- Who are the most valuable customers?
- In each country, which product is sold the most?
- Sales were highest on which day?
- Sales trend in the countries over time
- Top 5 Most common countries
- Top 5 Least common countries
- Total unit price sold by year
- Total quantity sold by invoice no
- Total quantity sold by Stock code
- Stock Code Feature Analysis
- Description Feature Analysis
- Customers Analysis
- Transaction Analysis based on time
- Sorting data in pairwise patterns of consumers with respect to income, geography and family size.

Sample screenshot from Jupyter Notebook



Association Analysis

We have used the Apriori Algorithm to generate association rules. Apriori Algorithm is a classic algorithm used for mining frequent item sets and devising association rules from transactional data. It takes into consideration that, that a subset of a “frequent itemset” must also be a “frequent itemset”.

There are 3 important values to observe in Association analysis

Support: An indication of how frequently the itemset appears in the dataset.
Confidence: Reliability of the inference rule found using the given support value.
Lift:
=1 means there is no correlation between the items in the inference rule.
>1 item likely to be bought together.
<1 item unlikely to be bought together.

Making sure that association is not random

A high support value is desired to exclude the possibility that the given association rules are not merely a coincidence. Along with it, a high confidence measure and lift greater than one should be present to reduce the possibility of a random association.

Inference:

In the Online Retail dataset, most of the transactions were from UK since the store is primarily UK based. However, it was interesting to note that still, there were no association rules when we kept the support value of 4% (0.04).
We expected to find association rules for UK with a high support value, however the results were opposite to what we expected.

On further analysis we found there were association rules for IRE (Ireland), which is the second country in terms of number of transactions at a support value of 8% (0.08).
It seems that the items being purchased in IRE are mostly related and people often buy related items together, than the items being purchased by the people in the UK. This could be due to the reason that the store carries a variety of items which people in UK readily buy, most of which might not be popular in other countries. Since the store is UK based, it must have a majority of items targeted towards the local population which the people in other countries might not be interested to buy. Therefore, given a limited selection of items available for people outside the UK, the frequency of the purchased items increases than those of other items for a given country. As the frequency increases, so does the support count, because of which we are able to find associations at twice the support threshold in case of IRE.
A similar trend was observed for Germany and France which were the next two countries in terms of number of transactions. We found association rules with a high support threshold of 26% and 24% respectively.

CES Data:

In case of CES data, we were able to find association rules at a high support value of 50% for all of the cities. This again makes sense given that CES consisted of food items consumption data which are often purchased together and there is no purchase bias as that was present in the case of online retail where the transactions were across different countries.

The associations we found for CES had a high support, confidence and lift values greater than 1 signifying that they are not random.

City: Belem

```
[64] #Building a model for each country
     #Using apriori analysis with min_support = 0.5

     #For city Belem
     frq_items = apriori(basket_Belem, min_support = 0.5, use_colnames = True)

     # Collecting the inferred rules in a dataframe
     # use association_rules with metric='confidence' and mininum threshold = 0.7 (70% of confidence or above)
     rules = association_rules(frq_items, metric ="confidence", min_threshold=0.7)

     rules_Belem = rules.sort_values(by=['confidence','lift'],ascending=[False,False])
     rules_Belem.head()
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	(butter)	(vinegar)	0.611511	0.877698	0.589928	0.964706	1.099132	0.053206	3.465228
9	(mayonnaise)	(vinegar)	0.633094	0.877698	0.597122	0.943182	1.074609	0.041457	2.152518
13	(mayonnaise, garlic)	(vinegar)	0.553957	0.877698	0.517986	0.935065	1.065361	0.031779	1.883453
11	(tomato)	(vinegar)	0.654676	0.877698	0.604317	0.923077	1.051702	0.029709	1.589928
0	(banana)	(vinegar)	0.582734	0.877698	0.517986	0.888889	1.012750	0.006521	1.100719

City: Goiania

```
#For city Goiania
frq_items = apriori(basket_Goiania, min_support = 0.5, use_colnames = True)

# Collecting the inferred rules in a dataframe
# use association_rules with metric='confidence' and mininum threshold = 0.7 (70% of confidence or above)
rules = association_rules(frq_items, metric ="confidence", min_threshold=0.7)

rules = rules.sort_values(by=['confidence'],ascending=False)
rules.head()
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
16	(egg, granulated_sugar)	(french_bread)	0.558011	0.867403	0.519337	0.930693	1.072965	0.035316	1.913181
9	(granulated_sugar)	(french_bread)	0.751381	0.867403	0.690608	0.919118	1.059620	0.038857	1.639377
27	(soy_oil, granulated_sugar)	(french_bread)	0.674033	0.867403	0.618785	0.918033	1.058369	0.034126	1.617680
24	(egg, granulated_sugar)	(soy_oil)	0.558011	0.845304	0.508287	0.910891	1.077590	0.036598	1.736034
1	(egg)	(french_bread)	0.729282	0.867403	0.662983	0.909091	1.048060	0.030402	1.458564

Customer Segmentation for Online Retail

Customer segmentation also known as Market Segmentation is the division of potential customers in a given market into discrete groups. This division and categorizing of customers if based on their buying characteristics and their habits of purchase. Customer segmentation is necessary in order to understand customers behaviors. It leverages acquired customer data like the one we have in our case, transaction data in order to divide customers into groups. There are several approaches to go about market segmentation. Two of the most relevant to our use case are as follows:

1. Needs based segmentation is based on differentiated needs and demands that customers express for a specific product or service being offered. The needs are discovered and verified through primary market research, and segments are distinguished based on different needs rather than characteristics of customers.
2. Value based segmentation differentiates customers by their economic value, grouping customers of the same value level into segments that can be distinctly targeted.

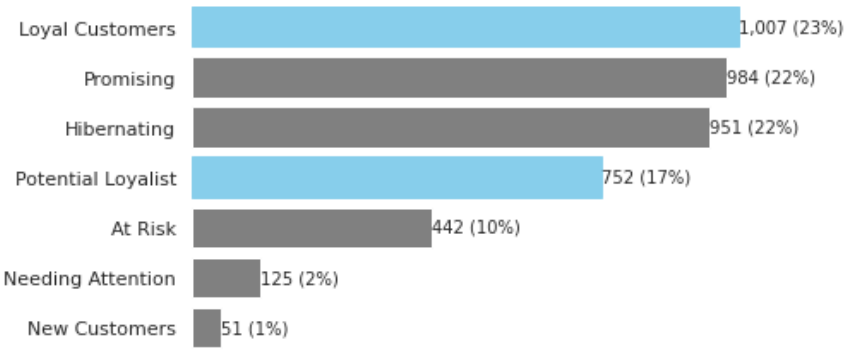
We used value-based segmentation to segment customers. To evaluate economic value of customers we used the RFM data (Recency, Frequency and Monetary Value) that we calculated earlier in Exploratory Data Analysis.

After computing the RFM data, we segmented customers according to their RFM scores into the following categories:

(We used these categories after researching and referring to several articles and research papers on RFM analysis)

The segments we figured out from RFM scores after some research:

Segment	Description	Score
Loyal Customers	Bought recently, buy often and spend the most	[3-4][4-5]
Potential Loyalist	Recent customers with average frequency	[4-5][2-3]
New Customers	Bought most recently, but not often.	[1-2]5
Promising	Recent shoppers, but haven't spent much.	41
Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.	33
At Risk	Below average recency and frequency, purchased long ago	[1-2][3-4]
Hibernating	Haven't purchased in a long time	[1-2][1-2]



Distribution of customers into different categories

To gain further insight into customer behavior and uncover hidden characteristics and clusters of customers, we decided to use predictive modelling. To cluster customers from their customer transactions rfm data, we used the following predictive modelling algorithms:

1. K-means clustering
2. Hierarchical clustering using Dendrograms
3. Gaussian Mixture Modelling

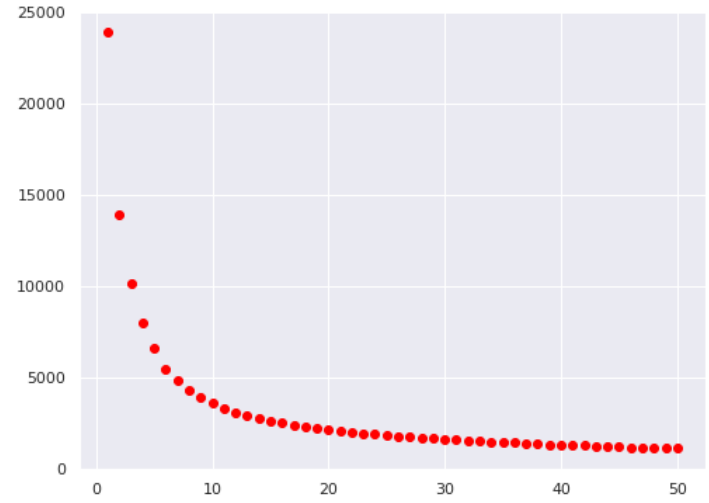
K-means Clustering:

The K-means clustering belongs to the partition based or centroid based hard clustering family of algorithms, a family of algorithms where each sample in a dataset is assigned to exactly one cluster.

Before applying K-means clustering, we had to make sure that our feature data is scaled and normalized as most clustering algorithms including K-means assume data to be normalized. To do this, we applied log transformation and used Standard Scaler of sklearn.

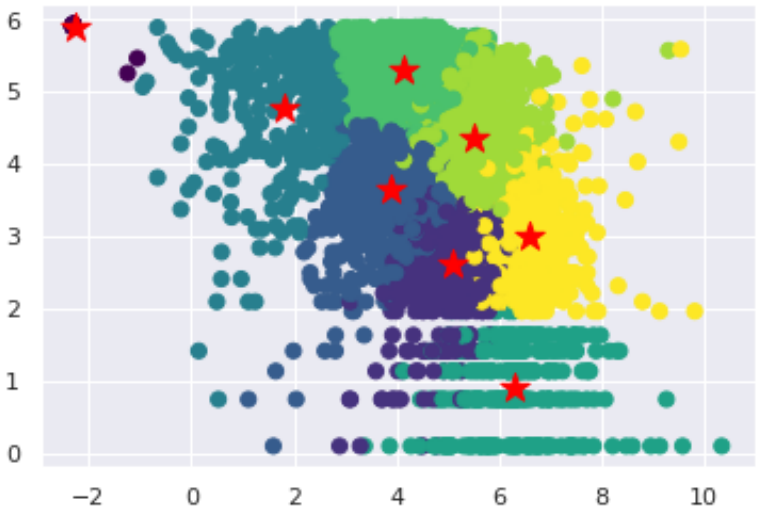
Another problem that came with K-Means was to figure out the best k for our dataset. So, to find the optimal k we used the elbow method and got the following result:

k = 8




```
The best K:  8
```

After applying K-Means clustering with k=8, we got the following clusters in our dataset:

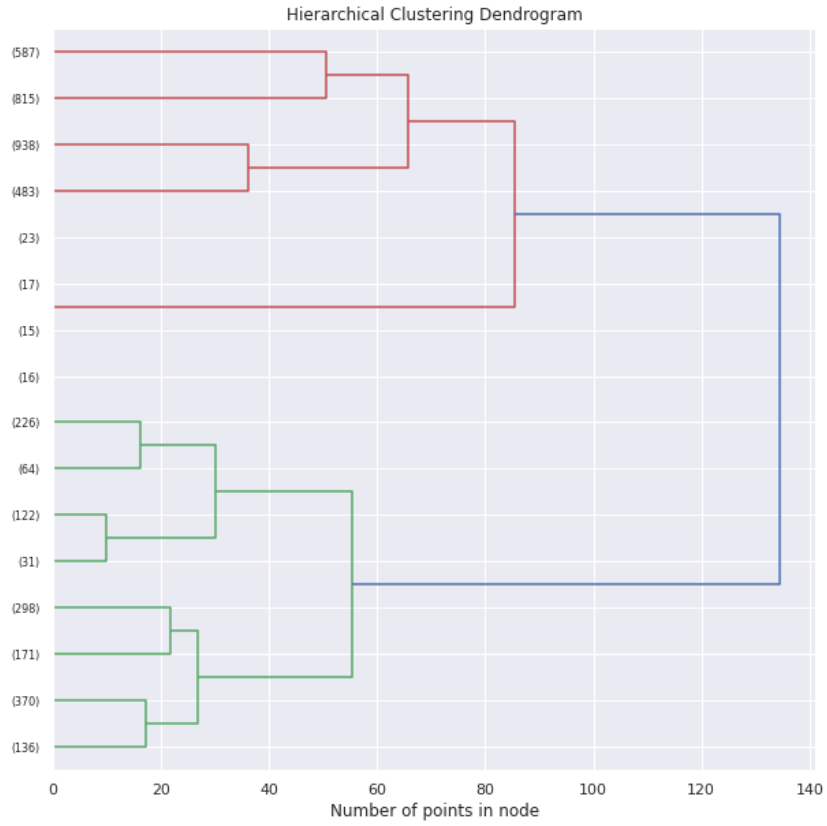


	6	5	2	1	7	4	3	0
Number of Customers	879	860	734	632	520	320	294	73

Clustering using Hierarchical Clustering with dendrograms:

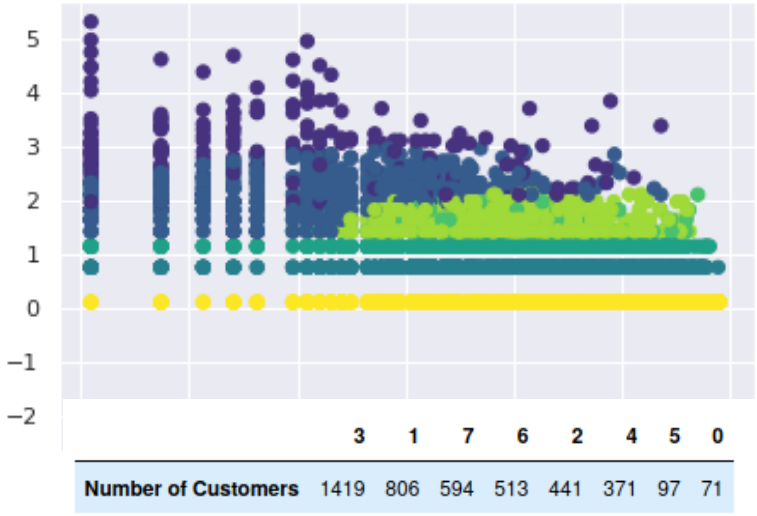
The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. So, the algorithm starts by treating each customer as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram. Even this algorithm assumes normalized and scaled data, so we use the scaled dataset from earlier. We also create a linkage matrix which contains all links between clustered classes, after which we plot the dendrogram.

The dendrogram below indicates how the clustering is done: customers are “grouped together”, starting from pairs of individual customers which are the closest to each other, and merging smaller groups into larger ones depending on which groups are closest to each other. Eventually all our data are merged into one segment as shown in the image below.



Gaussian Mixture Model Implementation

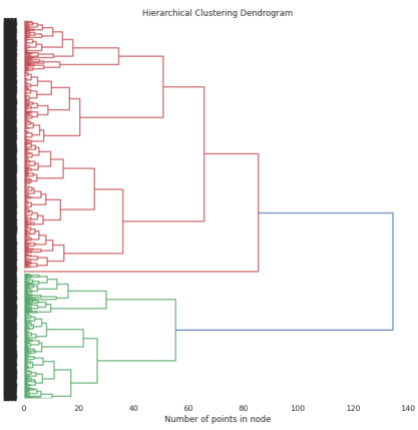
Gaussian Mixture Models are probabilistic models and use the soft clustering approach for distributing the points in different clusters. Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together. Following are the clusters obtained from GMM:



	3	1	7	6	2	4	5	0
Number of Customers	1419	806	594	513	441	371	97	71

Comparison of Clustering models used for Segmentation

K-Means was not successful in obtaining isolated clear clusters and we observe that the cluster assignments are mixed up. Another problem faced with k means is that it couldn't deal with non-circular shapes and consider one customer's probability of belonging to one or more clusters. And k means also only requires numerical data which can be a limitation for several online retail datasets.



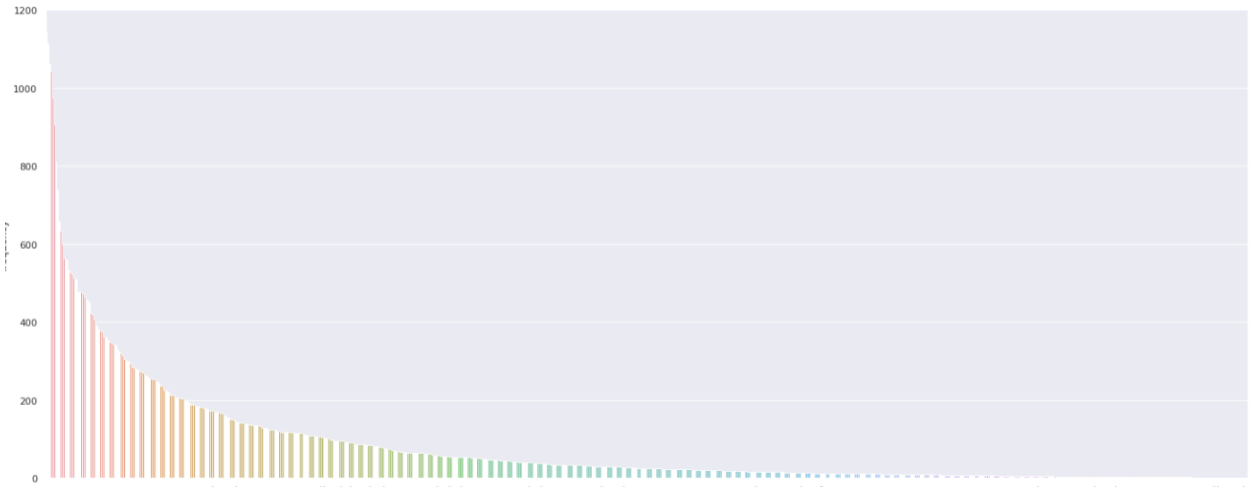
Due to this we decided to explore Hierarchical agglomerative clustering where one customer could belong to more than one cluster. Hierarchical clustering is very good with dealing with variant data, although as the number of observations (customers) we have are a lot it was difficult to visualize the dendrogram to the leaves as it would lead to something like this which is not very helpful to understand the customers and the clusters they belong to. (figure on the left)

After exploring K-Means clustering and hierarchical clustering, we decided to explore Gaussian Mixture Models to find clusters in our dataset. GMM provides a greater flexibility as clusters can have unconstrained covariances and allows probabilistic cluster assignment (soft clustering) as opposed to the previous clustering techniques explored. GMM is also successful in dealing with non-circular shaped clusters. (K-means only considers the mean to update the centroid while GMM takes into account the mean as well as the variance of the data)

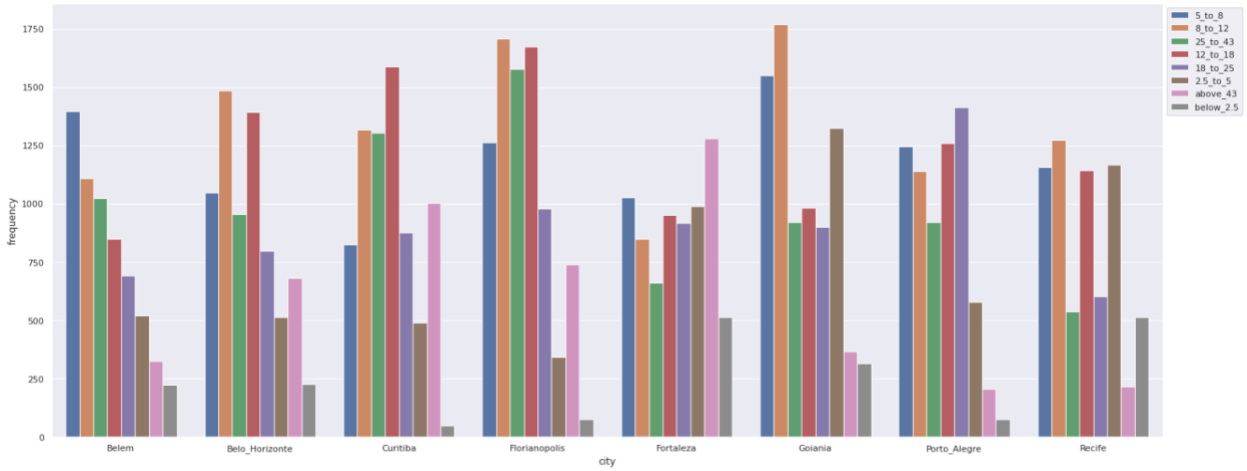
Despite all the advantages of GMM over K-Means, GMM did not provide any better results than K-Means on our dataset.

Insights into CES retail data:

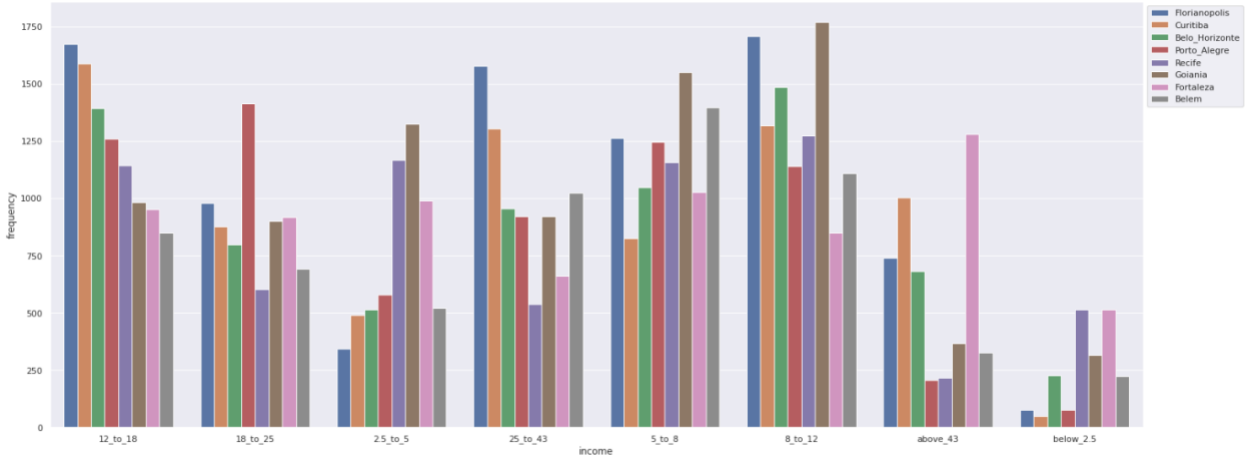
The macro level analysis of the CES data shows the distribution of merchandise sold depending upon the number of family members, city of residence and income bracket. The following graph shows the overall frequency of sale of all the merchandise across the complete dataset.



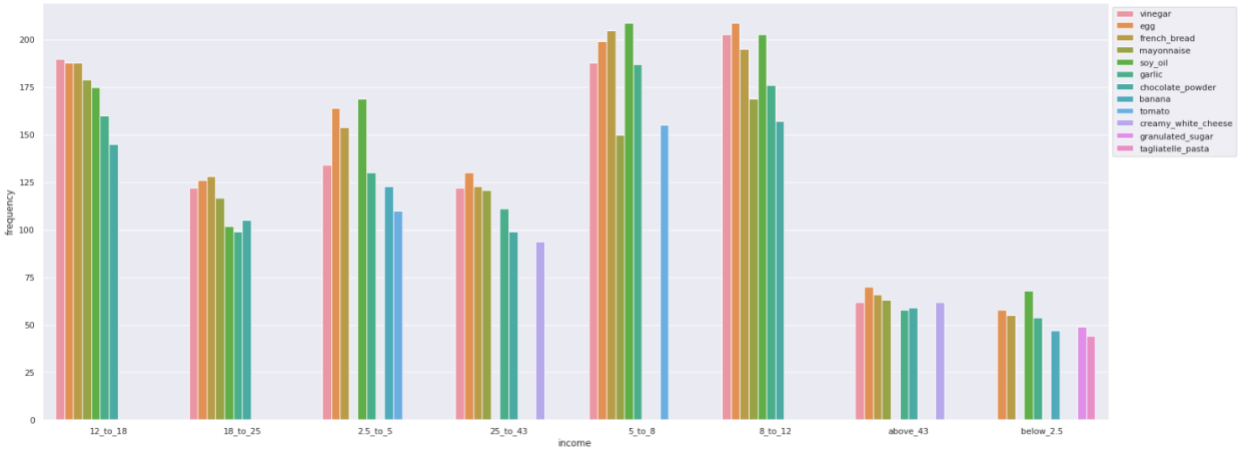
The following graph shows the distribution of various income groups in each of the cities. This can help in targeting high/low value goods in cities with higher/lower income family members.



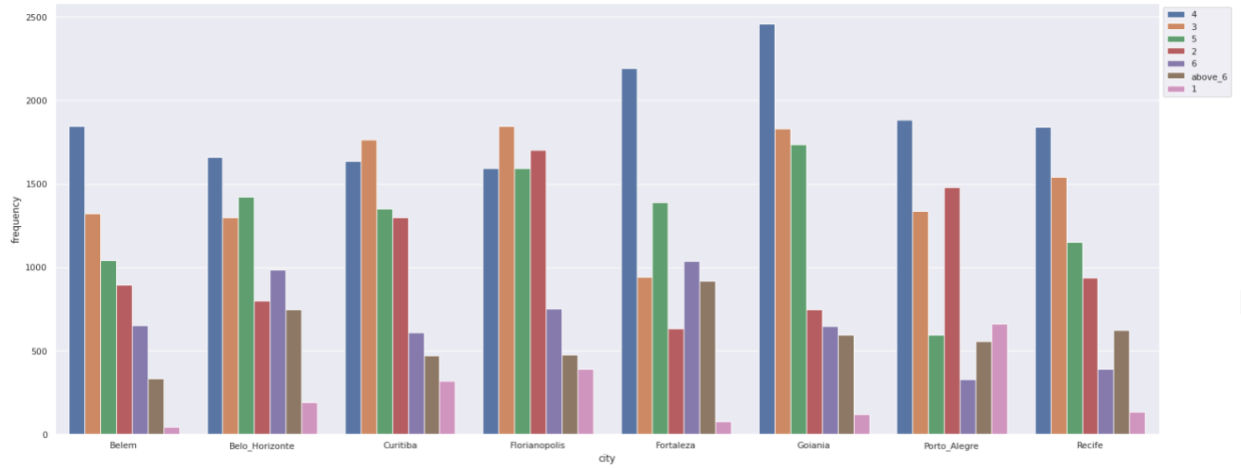
The following graph is like the above graph, wherein, the proportionate share of a family of a certain income group can be determined.



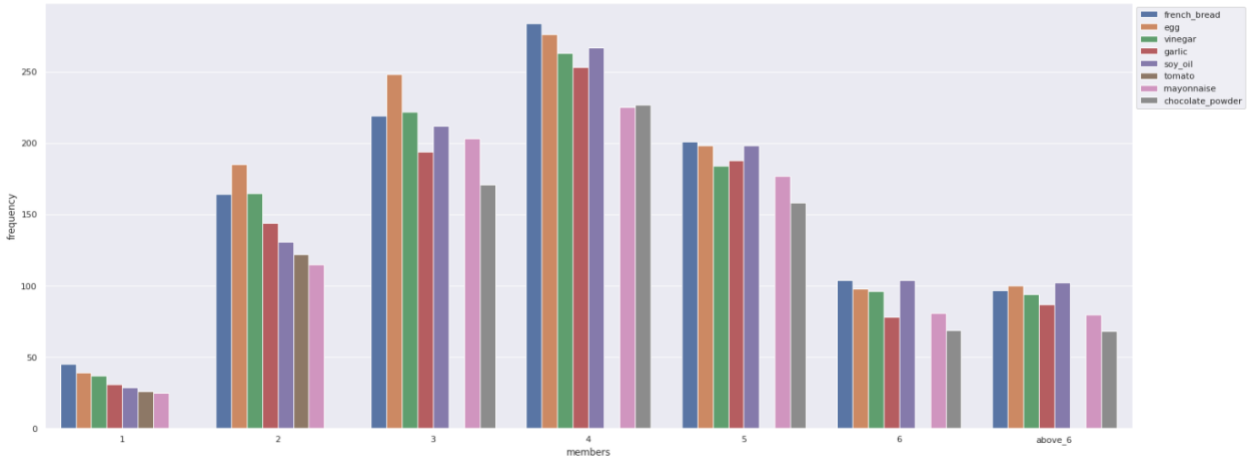
The following graph shows the top items purchased by each of the income level of family member



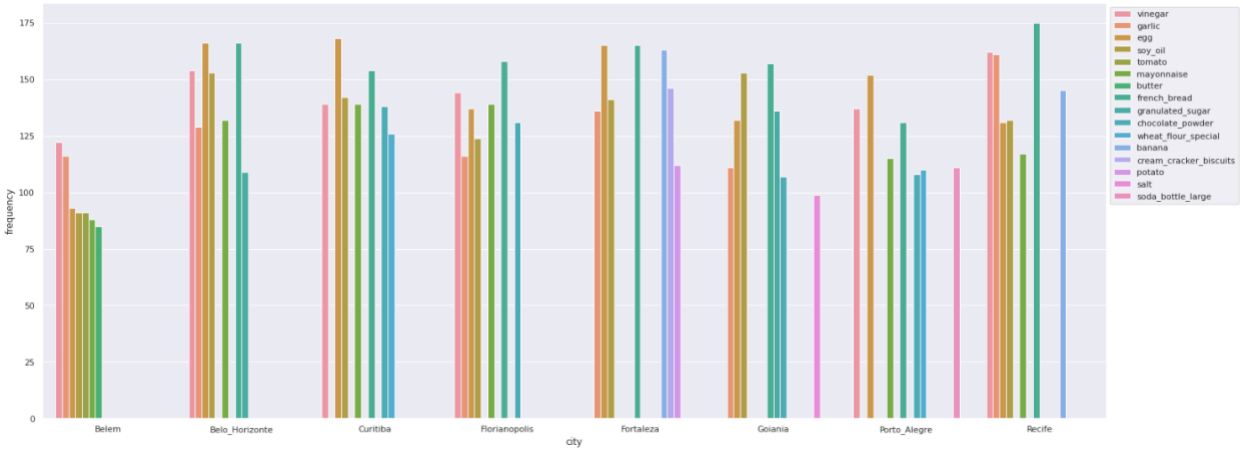
The following graph shows the distribution of the number of family members across all the cities.



The following graph shows the top purchased item according to the number of members in the family.



This graph shows the top items purchased in each of the cities.



N_{th}-Item Prediction

Prediction of the next item by customers based on the previous purchased products can be an important factor for customer behavior analysis.

Here we are trying to find the next nth product that may be purchased by customers based on (n-1) previously purchased items.

We have initially evaluated the cosine similarity between the all the unique products and generate the matrix of similarity scores. Next, we evaluate the similar items for each n-1 items and merged them in the dictionary. If the same product comes twice in the dictionary, then we will eliminate the entry that has a lower similarity score. Next, we will sort the products based on similarity scores and display the top 10 products.

Below are the sample predictions on both the datasets.

In Online Retail Dataset, for the customers having following items

['12 ASS ZINC CHRISTMAS DECORATIONS',

'12 COLOURED PARTY BALLOONS',

'12 DAISY PEGS IN WOOD BOX'],

```
('SAVE THE PLANET COTTON TOTE BAG', 0.91047830359767 97)
('MAGNETS PACK OF 4 VINTAGE COLLAGE', 0.8900696841088993)
('MAGIC DRAWING SLATE DINOSAUR', 0.8812436242995388)
('EASTER CRAFT IVY WREATH WITH CHICK', 0.8572010949657745)
('WOODLAND PARTY BAG + STICKER SET', 0.8337442805462084)
('MINI HIGHLIGHTER PENS', 0.8244774418426081)
('MAGIC DRAWING SLATE SPACEBOY', 0.8129907558665912)
('MAGIC DRAWING SLATE DOLLY GIRL', 0.8127932554909505)
('MAGNETS PACK OF 4 CHILDHOOD MEMORY', 0.770718701402017)
('LUNCH BAG SUKI  DESIGN', 0.7628788346557364)
```

Top 10 similar products.

In CES_hybrid dataset, for the customers having following items

'Alphabet_pasta','

Amazon_papaya'

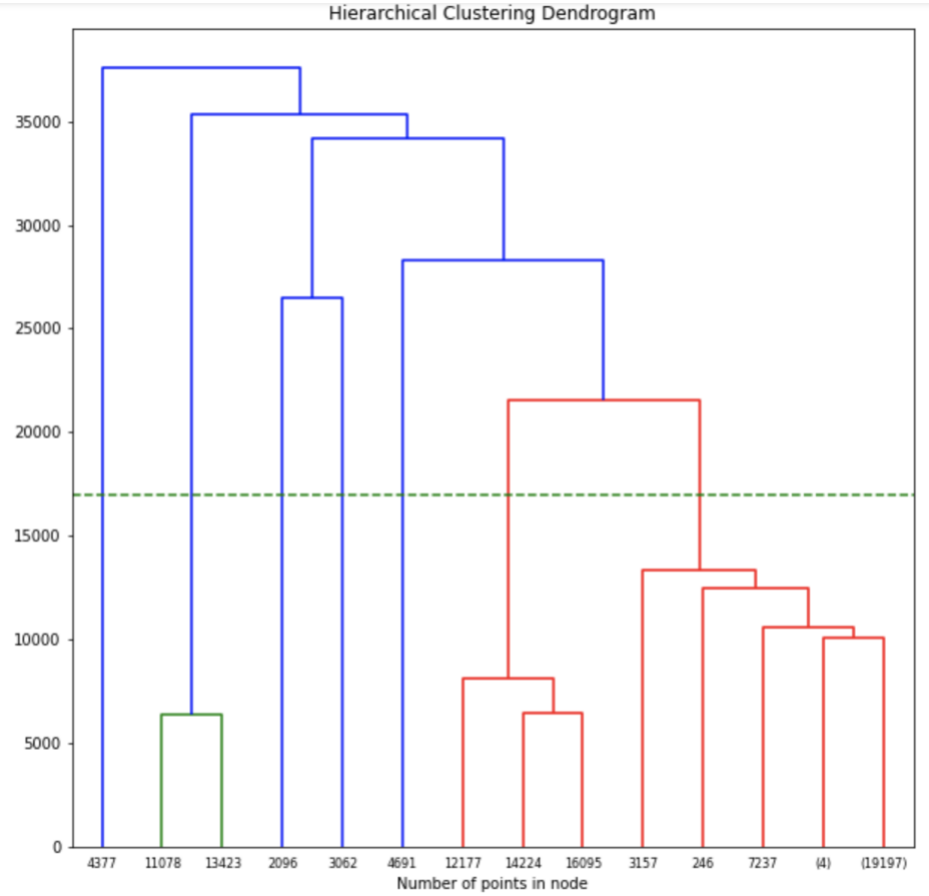
```
('french_bread', 0.4463091570732047)
('egg', 0.44041832481001403)
('chocolate_powder', 0.4310899208108515)
('vinegar', 0.4305120879429061)
('mayonnaise', 0.4163955911684075)
('garlic', 0.40961649233592456)
('canned_peas', 0.3987599228786232)
('cauliflower', 0.3947078178725582)
('soy_oil', 0.38967659810337474)
('cabbage', 0.3889812577077757)
```

Two-Step Clustering

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods

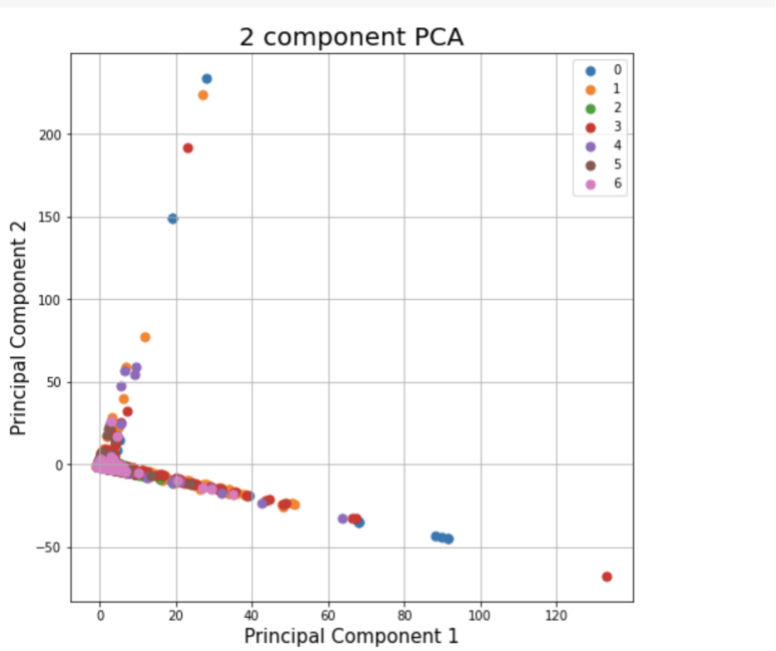
The k-means algorithm takes numeric data as input, and generates crispy partitions (i.e., every object only belongs to one cluster) as the output. It has been shown to be a robust clustering method in practice. Therefore, the k-means algorithm is applied in second clustering step to cluster data sets. K-Means starts by randomly selecting or by specifically picking k objects as the centroids of k clusters. Then K-Means iteratively assigns the objects to the closest centroid based on the distance measure and updates the mean of objects in this cluster as the new centroid until reaching a stopping criterion. This stopping criterion could be either non-changing clusters or a predefined number of iterations.

Here we initially grouped data by Invoice and Description for the analysis of evaluating the best number of clusters. Hierarchical Clustering has been performed to evaluate the best number of clusters possible based on above grouped data.



Here we can clearly visualize the above dendrogram that the number of clusters should be 7 for this dataset. (The line is cutting 7 vertical lines).

After having the cluster number fixed, we applied the both phases of Two Step clustering. After remapping the clusters to the original dataset of more than 4 lakh rows, we plotted the graph utilizing PCA.



Clusters are not clearly isolated as we were expecting from the Two step clustering. There are many possible directions that can be further explored. Such as, encoding of categorical variables based on co-occurrence can be done rather than simply Label encoding the variables.

The main advantage of two steps clustering analysis is it can handle huge amounts of data. Cluster center may be slightly distorted with the two phases, but it is efficient for large datasets.

When we tried, above data of 4 lakh rows in Hierarchical clustering, kernel crashed and we were unable to get any result.

We referred to the following research paper for the implementation of two phases.

http://www2.tku.edu.tw/~tkjse/13-1/02-IE435.pdf

Workload distribution

Online-Retail EDA (Purvi, Shalabh, Alok). Completed

Ces_hybrid_Preprocessing. (Nimit) Completed

Online Retail preprocessing (Shalabh, Alok). Completed

Market Basket Analysis, ces_hybrid_data (Purvi, Nimit) Completed

Market Basket Analysis, Online Retail (Alok, Shalabh) Completed

Nth Item Prediction of both datasets (Shalabh, Alok) Completed

Value based Segmentation (Purvi, Nimit) Completed

Hierarchical Clustering using dendrogram (Nimit, Purvi) Completed

Gaussian Clustering (Purvi, Nimit) Completed

Two step cluster analysis (Shalabh, Alok) Completed