**255 Data Mining - Project Report**

**Milestone-2**

​Shalabh Neema 014546259​

Purvi Misal 014544621

Nimit Patel 010700196

Alok Goyal 014499355

**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 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 so as 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 gift-ware, 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 has 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 transaction , family members etc.
* Hierarchical Clustering using dendrogram method available at SciPy.
* Two step cluster analysis: This method identifies groupings by running pre clustering first and then by running hierarchical methods.
* Recommendation System based on collaborating filtering to recommend nth item based on (n-1) previously purchased items.
* Apply the analysis methods performed of the Online-Retail dataset on the ces\_hybrid\_real\_world dataset and compare the result.

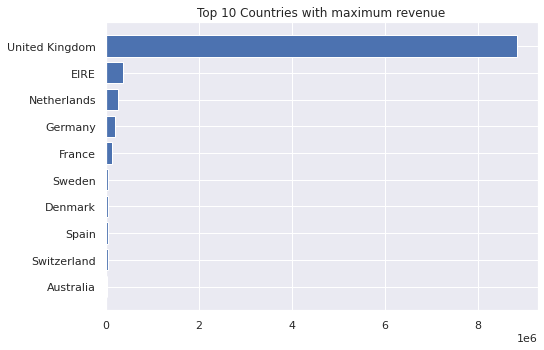
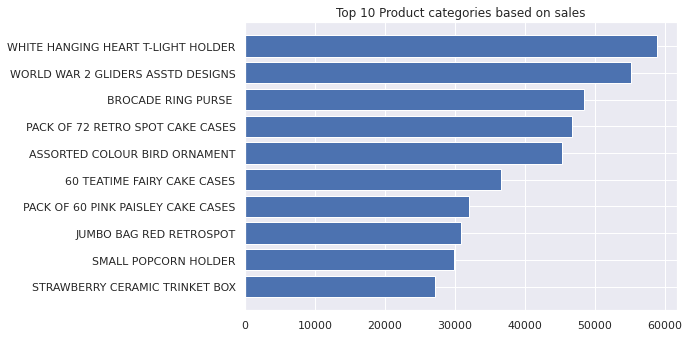
**Exploratory Data Analysis:**

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 pattern of consumers with respect to income ,geography and family size.

Sample screenshot from Jupyter Notebook



**Association Analysis**

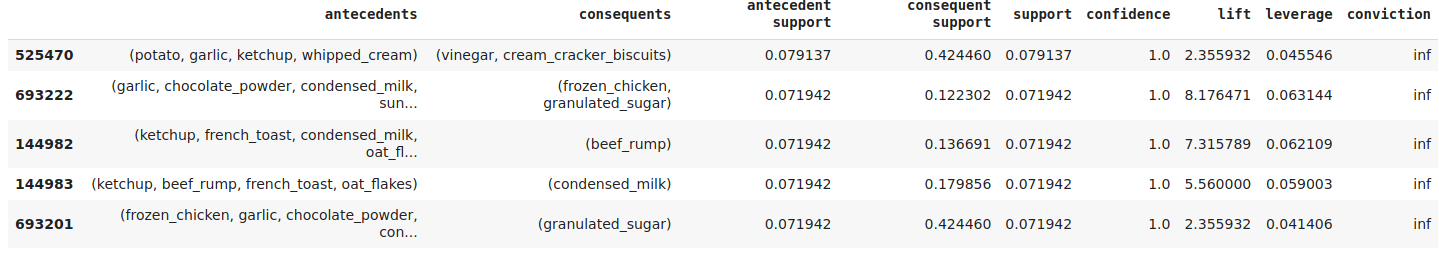
We have the Apriori Algorithm to generate association rules. Apriori Algorithm is a classic algorithm used for mining frequent itemsets and devising association rules from transactional data. It takes into consideration that, that a subset of a “frequent itemset” must also be a “frequent itemset”. The value of “frequent itemset” > than a threshold value(i.e. support).

Support used here is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule has been found to be true.

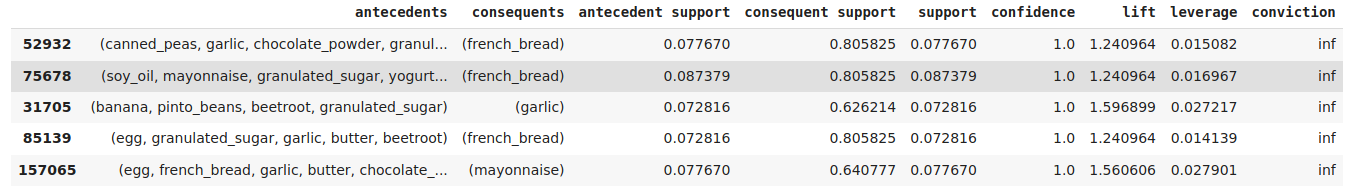
**Making sure that association is not random**

For association mining, we took our metric as Confidence and minimum threshold as 0.7, i.e, association rules having confidence 70% or more only, will be considered. This will make sure that any associations that occurred by chance or at random will not be taken into consideration.

For City Belem:



For City Belo Horizonte:



The association rules generated suggest the high correlation and co-occurrence of items or item sets in transactional data.

**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) In-progress

Comparison of Analysis between 2 datasets. (Nimit) In-progress

Recommendation System (Shalabh) In-progress

Value based Segmentation (Purvi) In-progress

Hierarchical Clustering using dendrogram (Nimit, Purvi)

Two step cluster analysis (Shalabh , Alok)