**Introduction**

**Motivation:**

We 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.

**Modelling:**

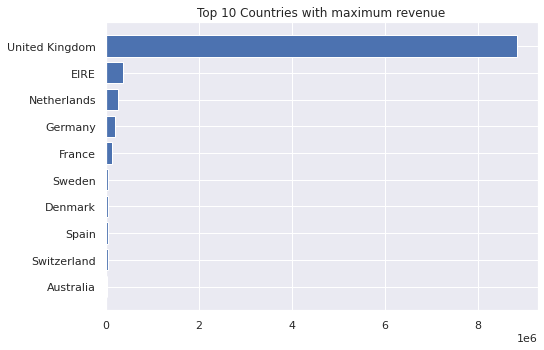
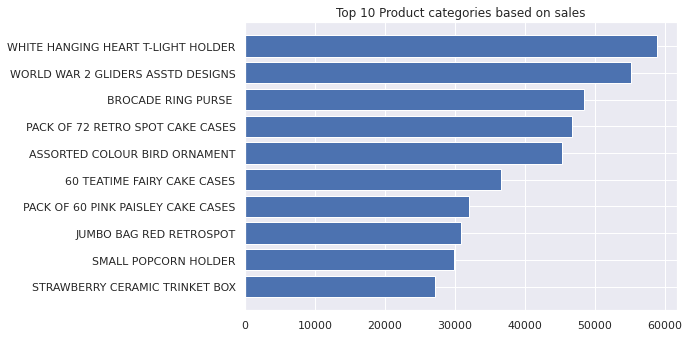
1. Predicting what a customer might buy next (or recommendation)
2. Clustering customers into groups to apply different marketing strategies on them

**Exploratory Data Analysis:**

During the Exploratory Data Analysis of our dataset, 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 tried to answer the following questions in our EDA:

* What are the top 10 Product categories based on sales?
* In each country , which product is sold the most?
* When were sales highest and lowest?
* Top 10 countries with maximum revenue
* Sale trends in countries with most purchases
* Who are the most valuable customers?
* Country with most valuable customers?
* Who is the most frequent and least frequent customer?
* Which customers spent the most and least?



**Market Basket Analysis:**

Market Basket Analysis is a technique used to uncover strength of association between pairs of products purchased together and also helps to identify patterns of co-occurrence of those products in customer transactions, i.e., give more insight into purchase behavior of buyers. It works by finding combinations of items that frequently occur together in transactions.

Rule generation is the first step in mining of frequent patterns in transactional data. An association rule is simply an implication expression of the form x -> y, where x and y are disjoint itemsets. To evaluate the “interest” of an association rule (also suggests the significance of an association rule), different metrics such as Support, Confidence and Lift are used.

We will be using 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.

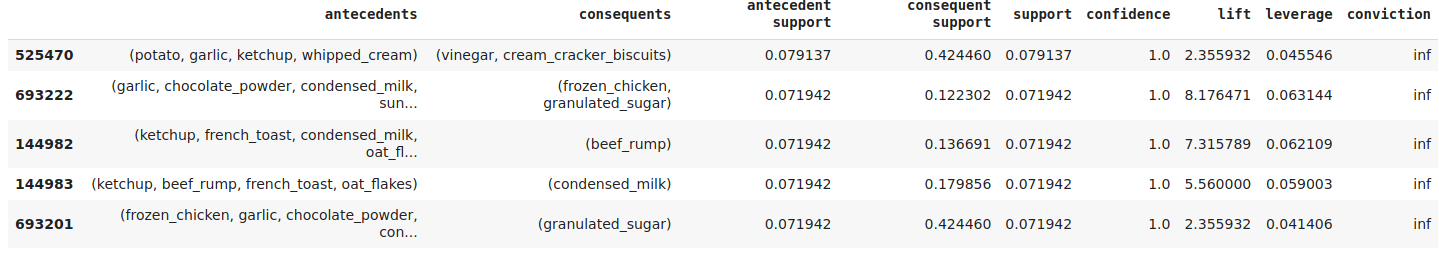
We used Market Basket Analysis using Association Mining for the Online Retail dataset as well as the CES Hybrid dataset and the following were our observations:

The most interesting association rules for Online Retail dataset with highest confidence are:

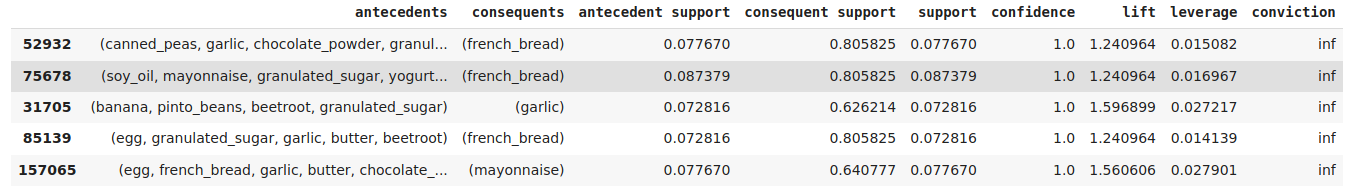
The most interesting association rules for CES Hybrid dataset with highest confidence for 2 cities:

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 itemsets in transactional data.

**Predictive Modelling:**

1. **Recommendation of items**

Using association rules can help us in recommending items based on metrics such as lift, confidence and support of the association rules. We already have association rules having confidence value above a particular threshold.

The association rules mined in the last step give us antecedents and consequents for the items and itemsets in the transactional data. Both antecedents and consequents can have multiple items and they depict the relationship between two itemsets and how much the occurrence of an antecedent can also lead to the occurrence of that consequent in a transaction.

To further prune our association rules, we will be filtering the association rules based on certain values of lift and confidence.

* Lift minimum threshold: 6

Association rules with a high lift value would mean that they occur more frequently than expected given the number of transaction and product combinations. Higher the lift value, strong the correlation between the items.

* Confidence minimum threshold: 0.8

Confidence is an indication of how often the rule has been found to be true. Higher the confidence, more interesting the association rule os.

After filtering the association rules, we would get rules with high confidence and high lift. We can simply select the given itemset for which items need to be recommended for as the antecedent to find the consequents for that itemset. This would give us the recommended items for that user transaction data. Here we would be able to predict the nth item a customer might buy when we know the n-1 items he has already bought.

**Feature Engineering:**

**Next** **steps:**

**Workload distribution for Milestone 2:**

Retail EDA

Ces hybrid preprocessing

Retail preprocessing -

Ces hybrid eda

Clustering

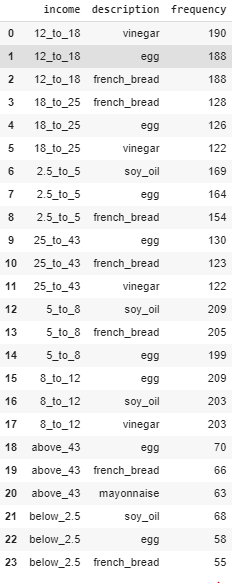
Market basket analysis

Prediction modelling

**What are we doing differently than others:**

Macro level analysis of data.  
Apriori and association rules look at the transaction level data and correlate the items present within a given transaction.

Trends and patterns are detected by this. Income, number of people in a household gives insight into the spending habits and the goods purchased by such families.   
These trends can be further used to accurately predict the goods which are likely to be bought by a customer of known income/size of family and city of residence.



Following table shows the top 3 item purchases by a family of a certain income group.