**Clustering Grocery Items**

**Abstract:**

Online shops often sell tons of different items and this can become very messy very quickly! Data science can be extremely useful to automatically organize the products in categories so that they can be easily found by the customers.

The goal of this project is to look at user purchase history and create categories of items that are likely to be bought together and, therefore, should belong to the same section.

Company XYZ is an online grocery store. In the current version of the website, they have manually grouped the items into a few categories based on their experience.

However, they now have a lot of data about user purchase history. Therefore, they would like to put the data into use!

In this project we would like to answer the following questions:

The company founder wants to meet with some of the best customers to go through a focus group with them. We will identify the ID of the following customers to the founder:

* The customer who bought the most items overall in her lifetime
* For each item, the customer who bought that product the most
* Cluster items based on user co-purchase history. That is, create clusters of products that have the highest probability of being bought together. We will replace the old/manually created categories with these new ones. Each item can belong to just one cluster.

**Existing System:**

The existing system does not provide a way of grouping customers and hence identifying natural clusters is difficult.

**Disadvantages of Existing System:**

The limitations of available systems are not sufficient to deal with the complex data. In this section, we present some of the limitations that are present in the existing system.

* The system uses DBMS and hence can return records based on the filters.
* The system also requires data extensive data preprocessing and Exploratory Data Analysis(EDA) inorder to perform feature engineering.

**Proposed System:**

We aim to implement K-Means, Hierarchical clustering and others and also fine tune the parameters of the model. These models would be trained on a data set which will be engineered carefully after performing the feature engineering.

**Advantages:**

* Load and explore the dataset and generate ideas for data preparation and model selection.
* Perform Exploratory Data Analysis to find correlations.
* Visualize clusters produced by the algorithms

**Software requirements:**

Operating System : Windows 7 , Windows 8, (or higher versions)

Language : Python 3.5 and other libraries likes numpy, pandas, matplotlib, seaborn and scikitlearn.

Mozilla Firefox(or any browser)

**Hardware requirements:**

Processor : Pentium 3,Pentium 4 and higher

RAM : 2GB/4GB RAM and higher

Hard disk : 40GB and higher