Data Mining - CS F415
End Semester Report

Name: Gaurav Hada
ID: 2016A2PS0582H

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining involves pattern discovery, association and correlation, classification, clustering, outlier analysis.

The goal of the assignment is to implement from scratch and apply the generalized sequential pattern (GSP) algorithm to mine for sequential patterns in the online retail dataset.

Link for the dataset: http://archive.ics.uci.edu/ml/datasets/online+retail

Link to github: https://github.com/Danny-aka/Data_Mining.git

1. Data Preprocessing

a. This is 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. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

b. The preprocessing includes the understanding of the data and its data types. Further information is listed below:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

- Country: Country name. Nominal, the name of the country where each customer resides.

c. The description column is important for further analysis. Therefore, cleaning the extra spaces out of the column is important.

d. The dataset consists of duplicate rows which need to be removed from the file.

e. The information of the dataset revealed that the unit price column and quantity column have few negative values which need to be removed because these values can't be used in further processing as these might create the error for further processing.

f. The customer ID column consists of null values which approximately is 27% of the overall dataset. Therefore, these can't be removed and need to be replaced with another five-digit number which is already not present in the list.

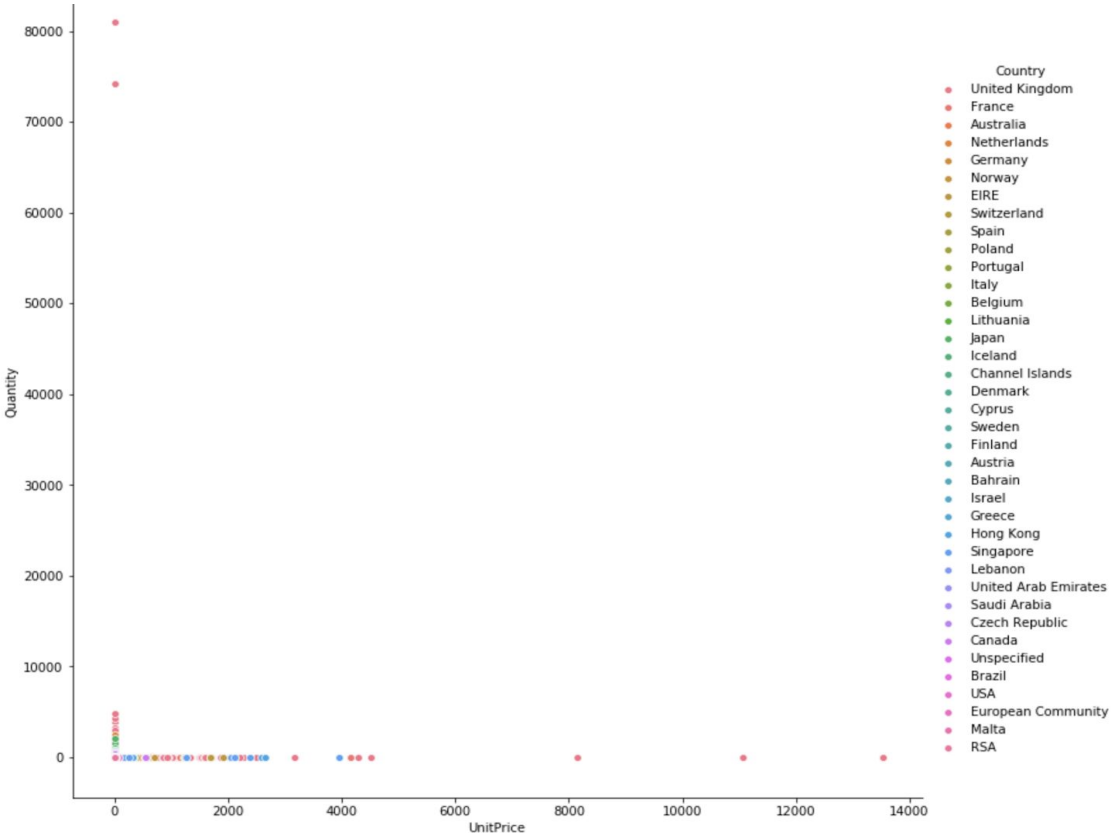
g. The resultant dataset is added with a total revenue column in order to calculate the overall revenue from the transactions.

The data set after preprocessing can be used for data visualization as the errors are eliminated in the previous step which could have intervened in the process and affected the results at a further stage.

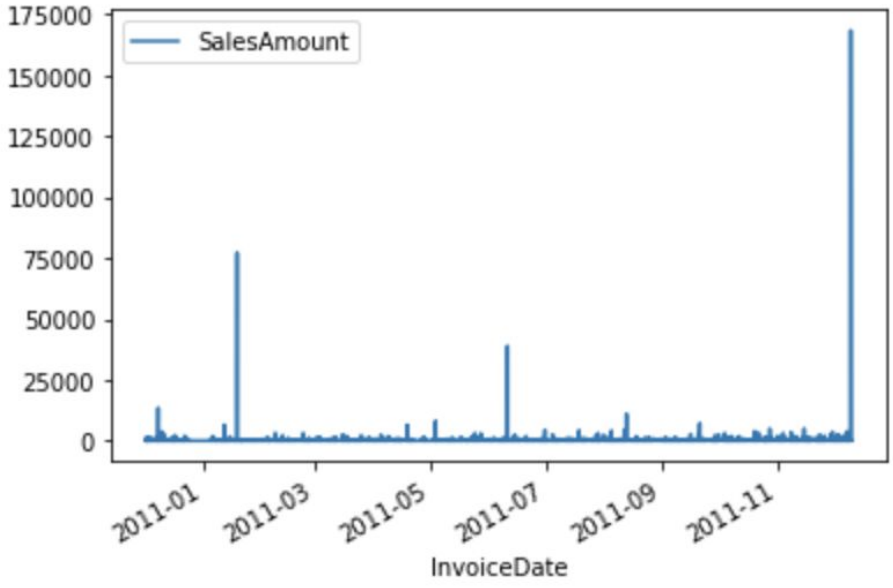
2. Data Visualization

The data visualization consists of extracting important insights from the data in graphical, tabular, chart format. It helps us grab meaning insight from the dataset for further implementation of the plans.

a. Relationship between Unit Price and Quantity for a different country in the dataset



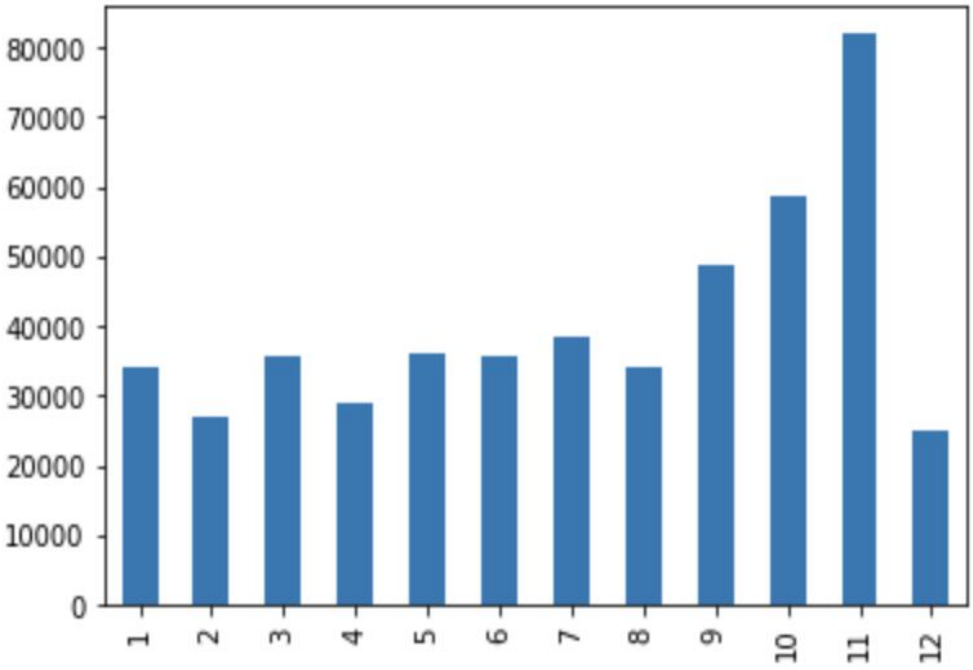
b. Relationship between Invoice month and Sales Amount



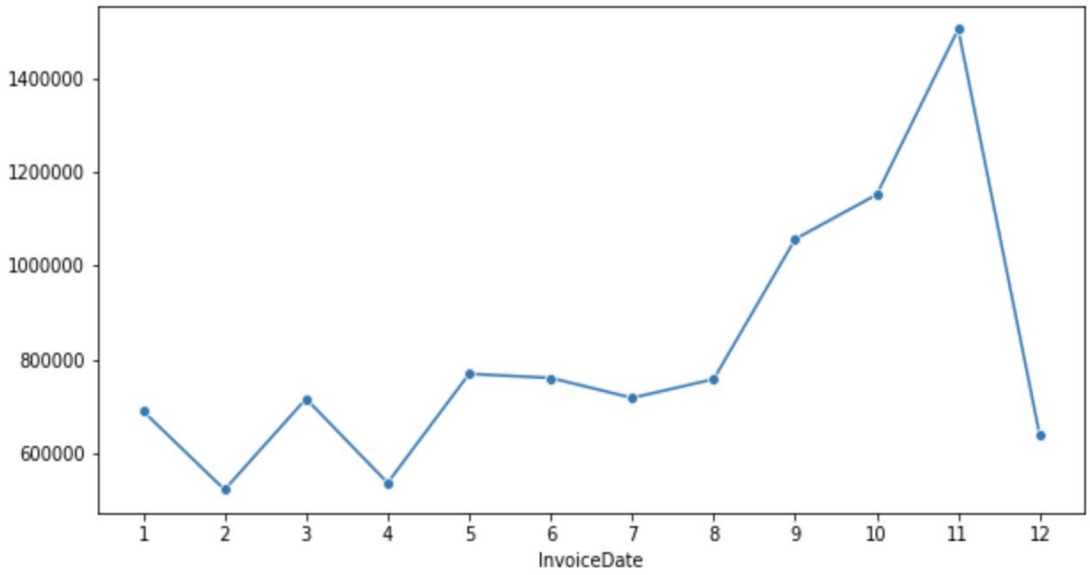
c. Countries with the highest transaction in the dataset:

```
United Kingdom    91.45%
Germany            1.72%
France              1.6%
EIRE                1.5%
Spain              0.47%
Netherlands        0.45%
Belgium            0.39%
Switzerland        0.37%
Portugal           0.28%
Australia          0.23%
Name: Country, dtype: object
```

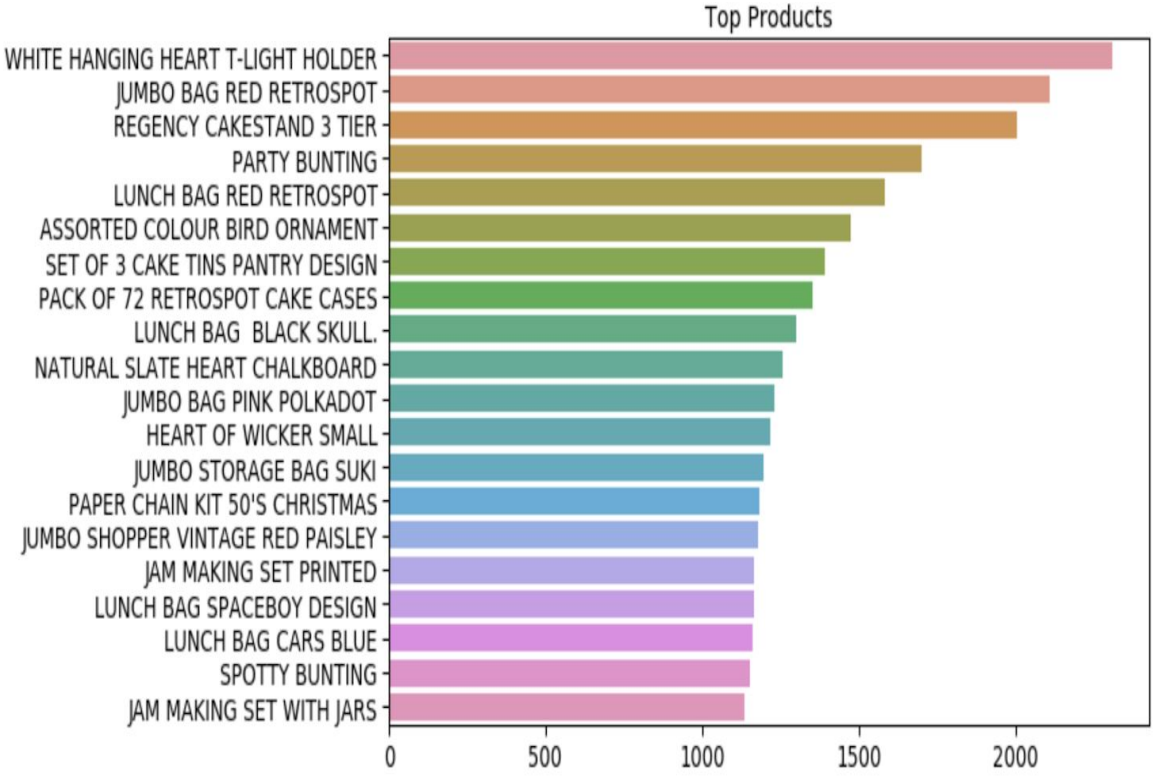
d. Count of transaction in different months of the data set:



e. Visualization of the top-grossing months in the dataset:



f. Visualization of some top products from the list of products from the description column:



The above visualization helps us gather important relationship between various aspects of the dataset.

3. Generalised Sequential Algorithm:

GSP algorithm (*Generalized Sequential Pattern* algorithm) is an algorithm used for sequence mining. The algorithms for solving sequence mining problems are mostly based on the *apriori* (level-wise) algorithm. One way to use the level-wise paradigm is to first discover all the frequent items in a level-wise fashion. It simply means counting the occurrences of all singleton elements in the database. Then, the transactions are filtered by removing the non-frequent items. At the end of this step, each transaction consists of only the frequent elements it originally contained. This modified database becomes an input to the GSP algorithm. This process requires one pass over the whole database.

4. Results and Discussion:

- Few of the frequent itemset from the list of items with K=1 are as given below:

```
({'JUMBO BAG PAISLEY PARK'}),
({'LUNCH BAG PAISLEY PARK'}),
({'VINTAGE DOILY TRAVEL SEWING KIT'
({'PLAYING CARDS KEEP CALM & CARRY
({'LOVE HOT WATER BOTTLE'}),
```

- Few of the frequent itemset from the list of items with K=2 are as given below:

```
[frozenset({'JUMBO BAG RED RETROSPOT', 'JUMBO BAG VINTAGE DOILY'}),
 frozenset({'SET OF 12 FAIRY CAKE BAKING CASES',
            'SET OF 12 MINI LOAF BAKING CASES'}),
 frozenset({"JUMBO BAG 50'S CHRISTMAS", 'JUMBO BAG VINTAGE CHRISTMAS'}),
 frozenset({'REGENCY CAKESTAND 3 TIER', 'SET OF 3 REGENCY CAKE TINS'}),
 frozenset({'GARDENERS KNEELING PAD CUP OF TEA',
            'GARDENERS KNEELING PAD KEEP CALM'}),
```

- Few of the frequent itemset from the list of items with K=3 are as given below:

```
[frozenset({'LUNCH BAG  BLACK SKULL.',
            'LUNCH BAG RED RETROSPOT',
            'LUNCH BAG SUKI DESIGN'}),
 frozenset({'PINK REGENCY TEACUP AND SAUCER',
            'REGENCY CAKESTAND 3 TIER',
            'ROSES REGENCY TEACUP AND SAUCER'}),
 frozenset({'GREEN REGENCY TEACUP AND SAUCER',
            'PINK REGENCY TEACUP AND SAUCER',
            'ROSES REGENCY TEACUP AND SAUCER'}),
```

- The frequent itemset from the list of items with K=4 are as given below:

```
[frozenset({'GREEN REGENCY TEACUP AND SAUCER',
            'PINK REGENCY TEACUP AND SAUCER',
            'REGENCY CAKESTAND 3 TIER',
            'ROSES REGENCY TEACUP AND SAUCER'})]
```

- The items should be suggested as a set of 3 or higher with the chosen threshold support as the lower frequent itemset have a large number of frequent itemset which will create difficulty in the assortment of the products.