Sequential Pattern Mining of Online Retail Dataset using GSP Algorithm (End Semester Report)

Prepared under the supervision of

Prof. Manik Gupta

As part of the course

CS F415: DATA MINING

Birla Institute of Technology and Science, Pilani

By

Aman Syed
2016A7PS0046H
f20160046@hyderabad.bits-pilani.ac.in

Abstract

The following paper deals with the sequential pattern mining of an online retail dataset obtained from the UCI ML repository. The report gives insight into the data cleaning, preprocessing, exploratory data analysis, mining process and results aspect of the assignment. The objective of this assignment is to clean the dataset, develop from scratch the GSP algorithm, apply on the cleaned dataset and draw insights from the obtained results.
The dataset can be found here:
http://archive.ics.uci.edu/ml/datasets/online+retail

Link to GitHub repo: https://github.com/aman3599/data_mining_final_assignment

Introduction

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.
Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.

Data Cleaning

The UCI ML Repository describes the dataset as ‘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’

Attribute Information:

- 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.

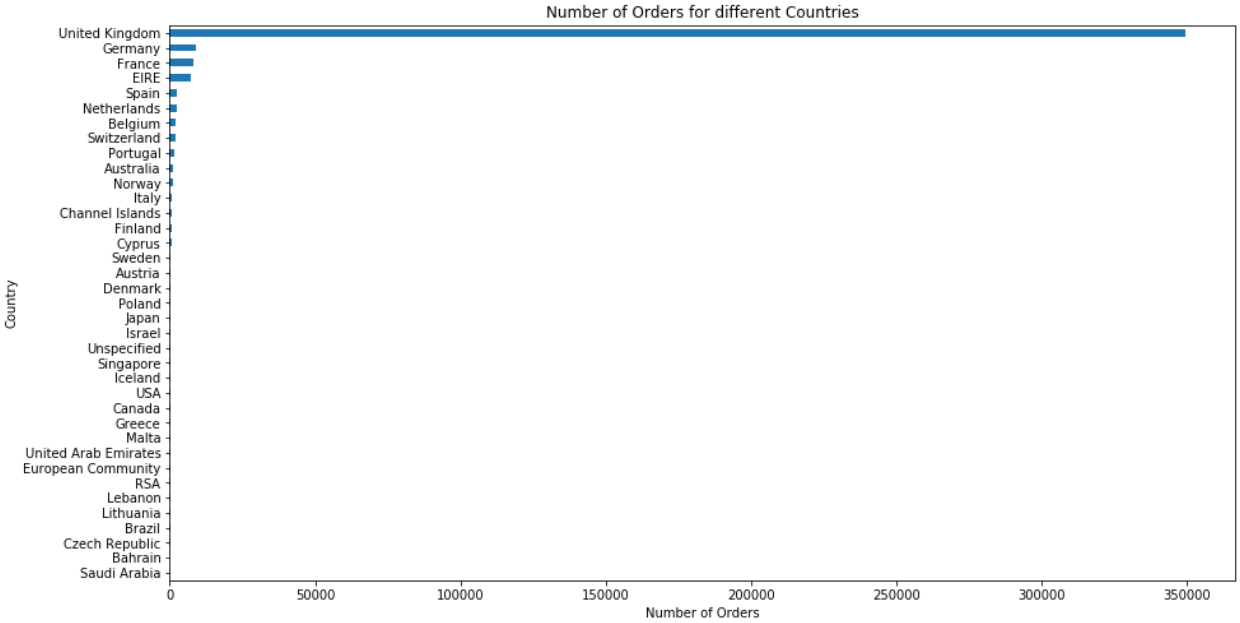
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

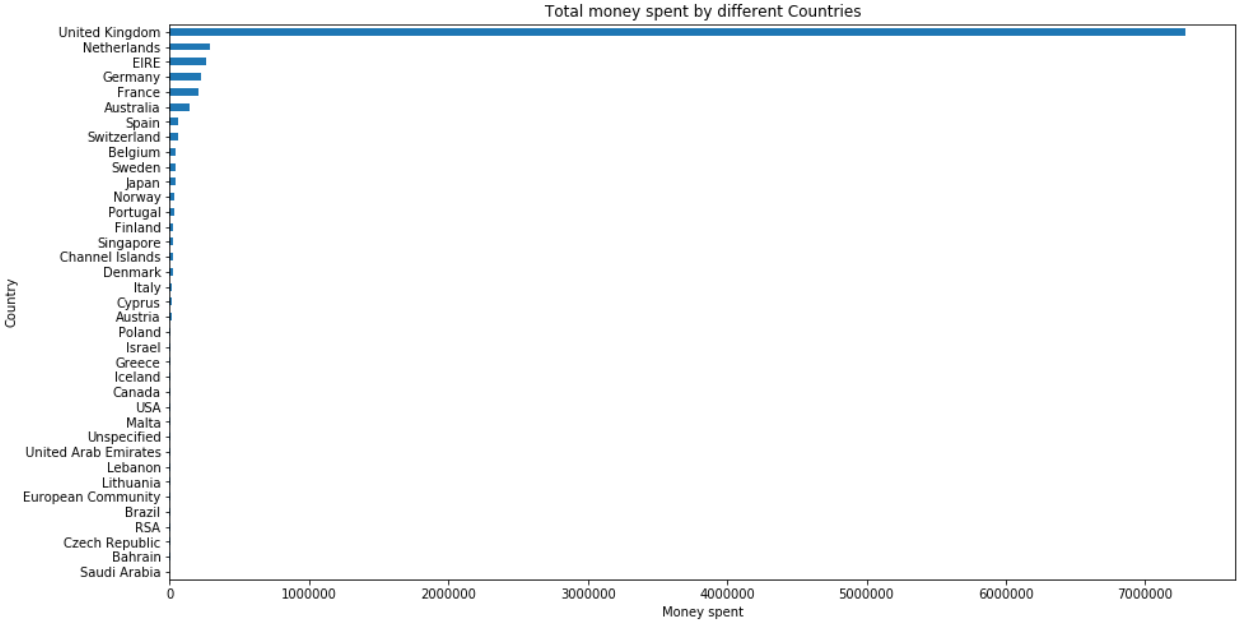
I followed the following steps while cleaning the data

1. CustomerID contains a lot of null values. Drop the transactions with null values.
2. The Quantity column contains a few negative values, which were removed as these could not be used for analysis.
3. The Description column was smoothened by removing commas and extra spaces in the description.
4. Duplicate transactions were removed from the dataset.
5. A new column giving the total order amount was added for the purposes of exploratory data analysis.
6. InvoiceNo starting with ‘C’ refers to cancelled orders. All such orders were removed.
7. Before GSP, the unwanted columns of 'InvoiceDate', 'UnitPrice', 'OrderAmount', 'CustomerID' and ‘'Country' were removed.
8. Data Columns were rearranged in a more understandable and relevant order.

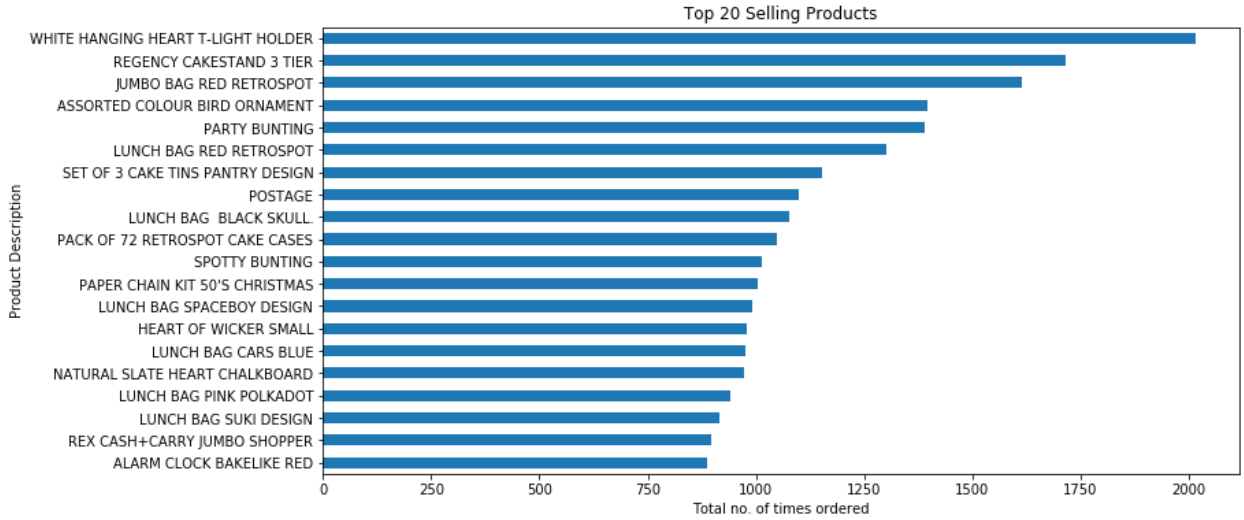
Data Preprocessing (Visualisation)

1. The company receives the highest number of orders from customers in the UK (since it is a UK-based company). Therefore, the TOP 5 countries (including UK) that place the highest number of orders are as below:
    a. United Kingdom
    b. Germany
    c. France
    d. Ireland (EIRE)
    e. Spain
2. The TOP 5 countries (including the UK) that spend the most money on purchases are as below:
    a. United Kingdom
    b. Netherlands
    c. Ireland (EIRE)
    d. Germany
    e. France





3. Top 20 selling products are as shown below:



Generalised Sequential Pattern (GSP)

Pseudocode:

```
F1 = the set of frequent 1-sequence
   k=2,
   do while Fk-1 != Null;
       Generate candidate sets Ck (set of candidate k-sequences);
       For all input sequences s in the database D
       do
           Increment count of all a in Ck if s supports a
       End do
       Fk = {a ∈ Ck such that its frequency exceeds the threshold}
       k = k+1;
   End do
   Result = Set of all frequent sequences is the union of all Fk's
```

The following are the results obtained from the application of the algorithm.

Results obtained:

Note - Support is taken to be 1%, since there only 1-frequent itemsets above 5%.

1. Some of the 1-frequent itemsets:

```
[frozenset({'JUMBO BAG PAISLEY PARK'}),
 frozenset({'LUNCH BAG PAISLEY PARK'}),
 frozenset({'VINTAGE DOILY JUMBO BAG RED'}),
 frozenset({'TRADITIONAL NAUGHTS & CROSSES'}),
 frozenset({'TRADITIONAL PICK UP STICKS GAME'}),
 frozenset({'VINTAGE DOILY TRAVEL SEWING KIT'}),
```

2. Some of the 2-frequent itemsets:

```
[frozenset({'HAND WARMER OWL DESIGN', 'HAND WARMER RED LOVE HEART'}),
 frozenset({'HOT WATER BOTTLE I AM SO POORLY', 'HOT WATER BOTTLE KEEP CALM'}),
 frozenset({'HOT WATER BOTTLE KEEP CALM', 'LOVE HOT WATER BOTTLE'}),
 frozenset({'HOT WATER BOTTLE KEEP CALM',
            'HOT WATER BOTTLE TEA AND SYMPATHY'}),
 frozenset({'CHOCOLATE HOT WATER BOTTLE', 'HOT WATER BOTTLE KEEP CALM'}),
 frozenset({'JUMBO BAG RED RETROSPOT', 'JUMBO BAG VINTAGE DOILY'}),
```

3. Some of the 3-frequent itemsets:

```
[frozenset({'LUNCH BAG APPLE DESIGN',
            'LUNCH BAG RED RETROSPOT',
            'LUNCH BAG WOODLAND'}),
 frozenset({'LUNCH BAG APPLE DESIGN',
            'LUNCH BAG PINK POLKADOT',
            'LUNCH BAG RED RETROSPOT'}),
 frozenset({'LUNCH BAG  BLACK SKULL.',
            'LUNCH BAG APPLE DESIGN',
            'LUNCH BAG RED RETROSPOT'}),
```

4. The 4-frequent itemsets:

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

~ END ~