

Data Mining for an Online Retail Data Set

Abstract

This poster shows the data pre-processing that was completed prior to results being generating through data mining using Weka and Tableau. the following questions have been answered:
Q1: Can you predict what time of day certain items will sell best?
Q2: Are there associations with the months and countries to see when a country is more often ordering?
Q3: What combinations of items have been sold together?
Q4: What Is the Highest and Lowest Selling Months within the UK?

Description of the Data Set

Obtained from the UCI, Machine Learning Respiratory website. <https://archive.ics.uci.edu/ml/datasets/online+retail#>
File name - Online Retail.xlsx
Format – XLSX
Size – 22.6MB

“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.” [1]

- ❖ Invoice Number
- ❖ Stock Code
- ❖ Description
- ❖ Quantity
- ❖ Invoice date
- ❖ Invoice time
- ❖ Unit Price
- ❖ Customer ID
- ❖ Country

Data Pre-Processing

Calculations Performed and New Attributes within the Data

- Invoice Date – Split from the original Attribute including the time
- Invoice Time - Split from the original Attribute including the date
- Invoice Day – Used the Date field
- Invoice Month – Used the Date field
- AM/PM – Used the time field
- Part of day – using an IF formula with the time field
 - Morning
 - Afternoon
 - Evening
 - Overnight



Issues – Inserting to WEKA

The following issues had to be resolved before the data could be added to WEKA and used for data mining. String values had to be changed using an excel formula to surround them with inverted commas. Also, WEKA did not agree with some of the punctuation used within the data base and therefore the following and others had to be changed.

• String Values (Excel)

- `=""&A2&""`

GLASS STAR FROSTED T-LIGHT HOLDER
21730
Morning

➔

‘GLASS STAR FROSTED T-LIGHT HOLDER’
‘21730’
‘Morning’

• Symbols and Punctuation (Notepad++)

+
You’re
0’s

➔

AND
Youre
Os

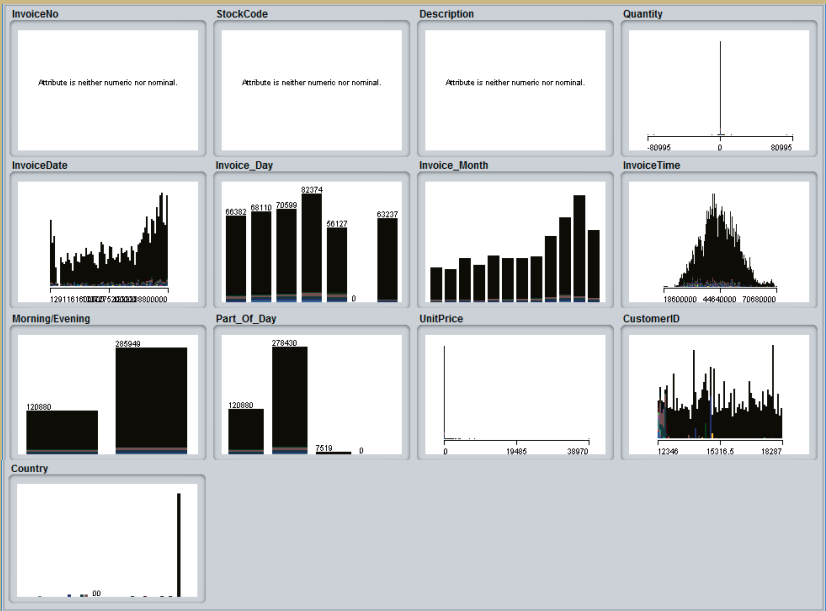
The Arff Header

The Arff header was originally different, however due to the following string values, Invoice Day, Invoice Month, Morning/Evening, Part of Day and Country, not being properly accepted within WEKA, it had to be changed to identify the values of the attributes.

```
%RELATION online retail
%
%ATTRIBUTE InvoiceNo STRING
%ATTRIBUTE StockCode STRING
%ATTRIBUTE Description STRING
%ATTRIBUTE Quantity NUMERIC
%ATTRIBUTE InvoiceDate DATE dd/MM/yyyy
%ATTRIBUTE InvoiceDay (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)
%ATTRIBUTE InvoiceMonth (January, February, March, April, May, June, July, August, September, October, November, December)
%ATTRIBUTE InvoiceTime DATE HH:mm:ss
%ATTRIBUTE MorningEvening (AM, PM)
%ATTRIBUTE PartOfDay (Morning, Afternoon, Evening, Overnight)
%ATTRIBUTE UnitPrice NUMERIC
%ATTRIBUTE CustomerID NUMERIC
%ATTRIBUTE Country (Australia, Austria, Bahrain, Belgium, Brazil, Canada, Channel_Islands, Cyprus, Czech_Republic, Denmark,
%
%DATA
```

Initial Representation in WEKA

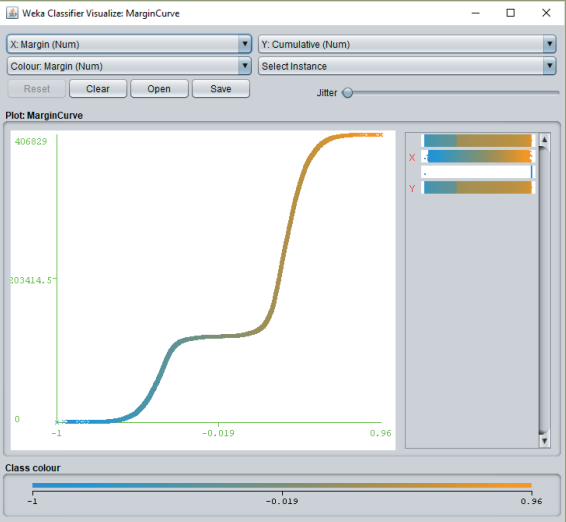
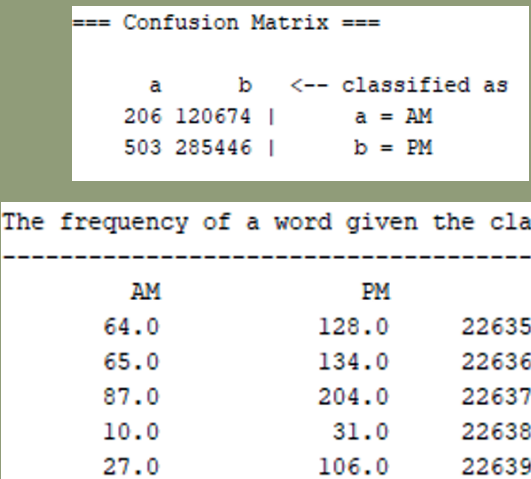
This is the representation of the data in WEKA when all the pre-processing had been completed. Some of the attributes have no visual representation, these Values are string values therefore there is no data to show within a graph.



Analysis and Results

Q1: Can you predict what time of day certain items will sell best?

The attributes that were used to answer this question was the Stock Code and Morning/Evening using the Classification rule. The overall best selling time of day was the PM, however it was difficult to identify the time of day a certain item sells best, due to the stock code being used and not a product description. Also easier if the product was known prior to the testing.



Q2: Are there associations with the months and countries to see when a country is more often ordering?

The attributes that were used for this question were, Invoice Month and Country. Both were converted from string to nominal data. The Rule that was found within the data was the United Kingdom was sold to the most during the month of November. As there was a large sum of data from the United Kingdom, the Association analysis does not recognise any other countries and their rules.

Best rules found:

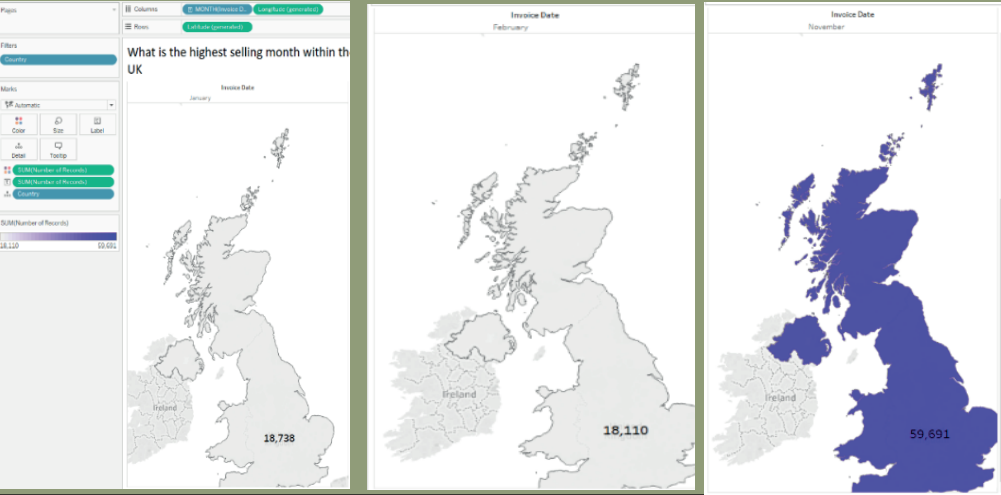
1. Invoice_Month=November 65598 ==> Country=United_Kingdom 59691 <conf:(0.91)> lift:(1.02) lev:(0) [1340] conv:(1.23)

Q3: What combinations of items have been sold together?

Many different attempts were made as this question, using a variety of attributes, however an Association Analysis could not be conducted, therefore this could not be determined.

Q4: What Is the Highest and Lowest Selling Months within the UK?

By using Tableau to answer this question and using the attributes, Invoice Date by Month and Country was filtered to only show the UK. Colour and text was shown using the Sum of the number of Records. the conclusion was that the highest selling month for the UK was November and the lowest selling was February. This supports Q2’s answer.



[1] <https://archive.ics.uci.edu/ml/datasets/online+retail#>