

Business Understanding

Background:

This dataset is obtain as a freelance for the research. The problem is to find the best clustering for the dataset so, that the researcher can easily relate between the data variables.

Business Objectives:

The objective is to determine the suitable clustering for the dataset.

Business Success Criteria:

The criterion for an outcome is to improve the credit risk rating model.

Data Mining Goals:

To identify clusters that can be used to minimize default rates.

Data Understanding

Collect Initial Data

The data provided is consists of:

- dataset.xlsx: contains information about invoice no, description, customer ID, unit price. According to the dataset there are 1977 records and 7 attributes.

The data does not contain any personal information.

Description of the data:

The file containing the data required for data mining is dataset.xlsx. Below is the description of each of its fields and their data type.

Variable Name	Description	Type
InvoiceNo	Invoice code of the purchasing	Integer
StockCode	Stock code at the database of the item	Nominal
Description	Description about the customer	Nominal
Quantity	Items quantity buy	Integer
UnitPrice	Price of the item	Real

CustomerID	Identifiers of the customers	Integer
Country	Customers' country	Nominal

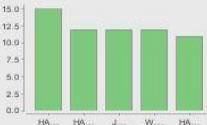
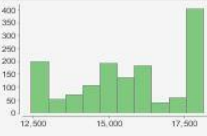
Data Preparation

Clean Data

This process was performed using the RapidMiner tool, and was executed as follows:

The dataset was imported first into a RapidMiner repository.

The dataset contains two attributes with missing values, these appear when the information to represent does not exist, and the attributes are 'Description' and 'CustomerID'.

Name	Type	Missing	Statistics	Filter (8 / 8 attributes):
InvoiceDate	Date-time	0	Earliest date Dec 1, 2010 8:26 AM Latest date Dec 1, 2010 2:33 PM Duration 0d 6h 7m 0s	<input type="text" value="Search for Attributes"/>
UnitPrice	Real	0	Min 0 Max 569.770 Average 3.805	
Country	Nominal	0	Least Netherlands (2) Most United Kingdom (1828) Values United Kingdom (1828), Norw	
Description	Nominal	4	Least YULETIDE [...] P SET (1) Most HAND WAR [...] SIGN 	
CustomerID	Integer	531	Min 12431 Max 18144 Average 15629.875 	

Showing attributes 1 - 8 Examples: 1,976 Special Attributes: 0 Regular Attributes: 8

For our analysis here we will treat the missing values as a new value 'missing =not_specified'. We have a total of 535 missing values from 2 attributes.

Considering that the operator that we will be using in our analysis to calculate the means, we use the operator "Nominal to numerical" to change the type of the attribute 'StockCode' which is non-numerical to a numerical type.

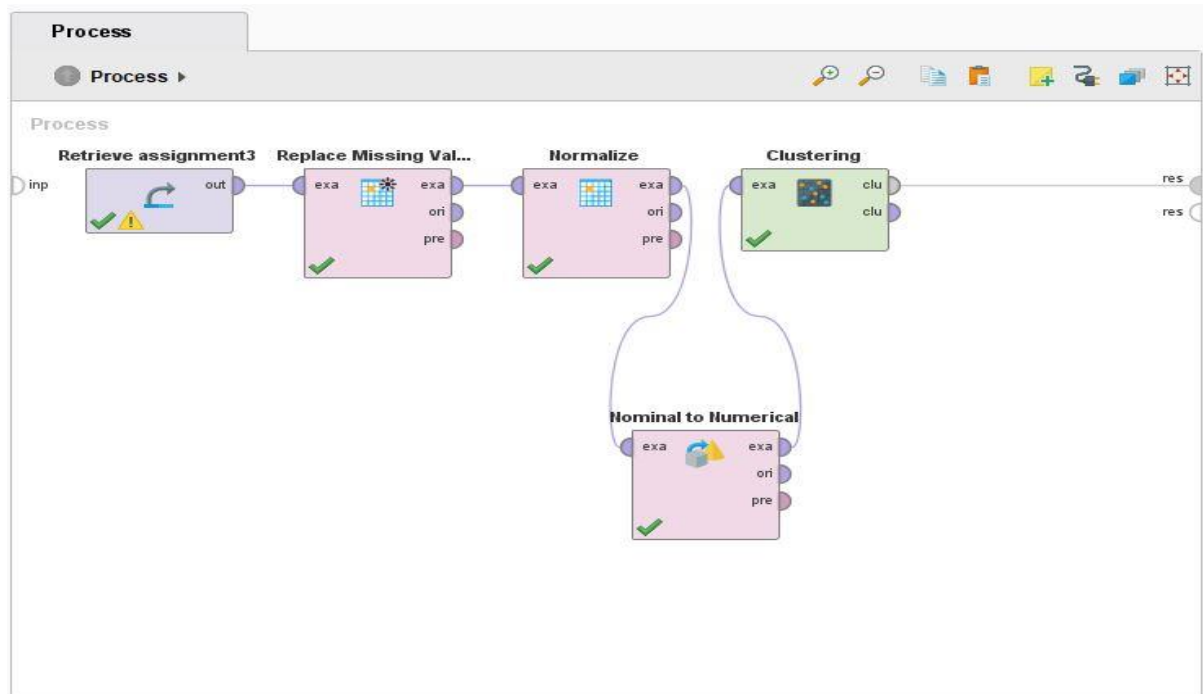
Format Data

Since the attributes of the dataset are of different units and sizes, the "Normalize" operator was used to make a fair comparison between the attributes. The attributes selected for normalization were: 'StockCode' which describes 'code for the customer's purchasing items'. In order to keep

the original distribution of the data and lessen the influence of outliers, the Z-transformation method was used.

Modeling:

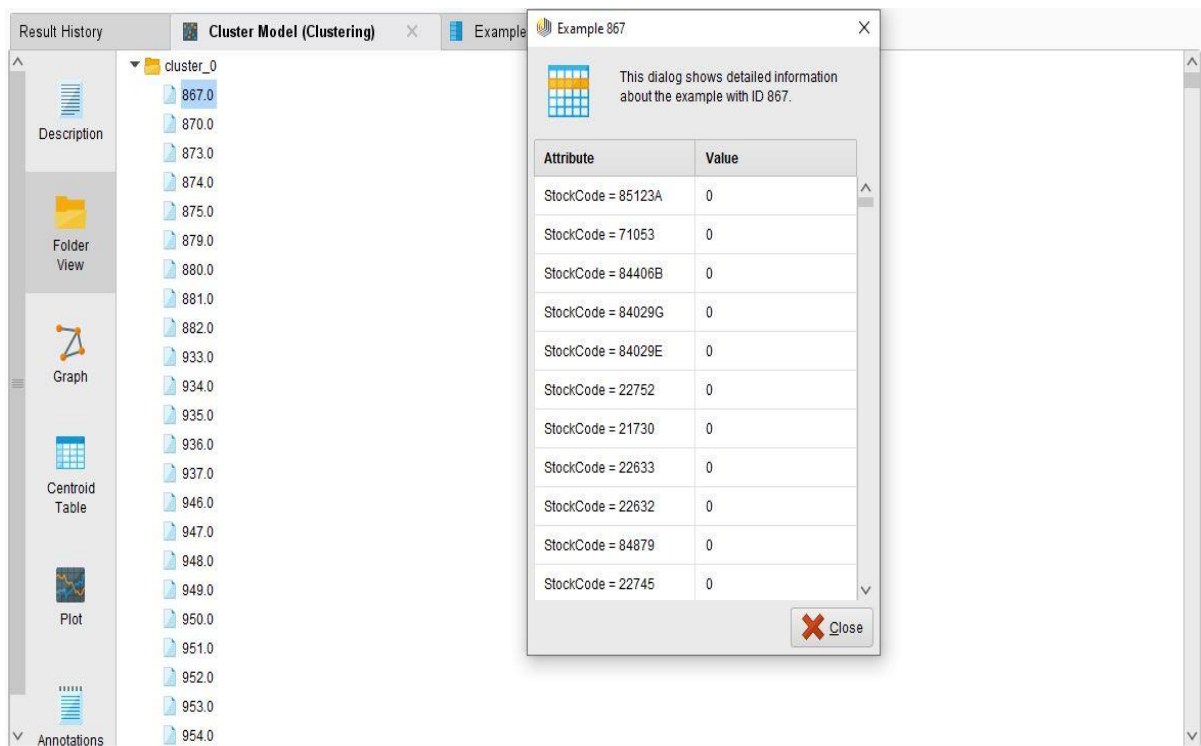
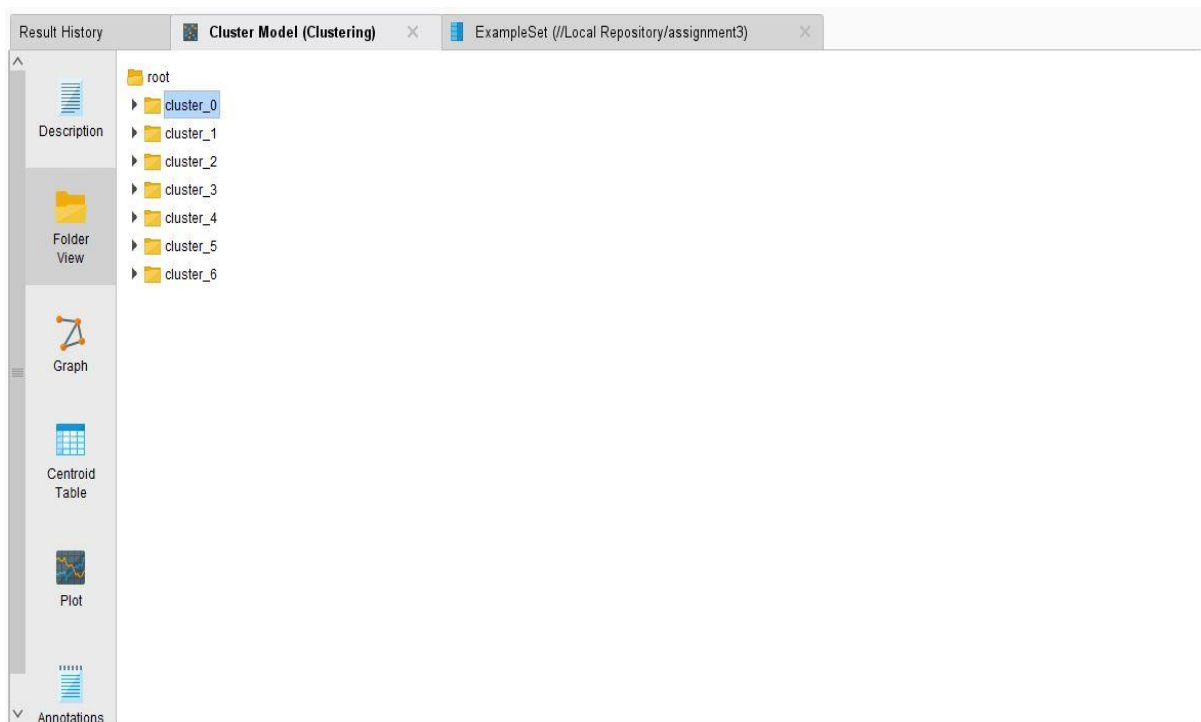
We use K-means operator to find the cluster of the group.



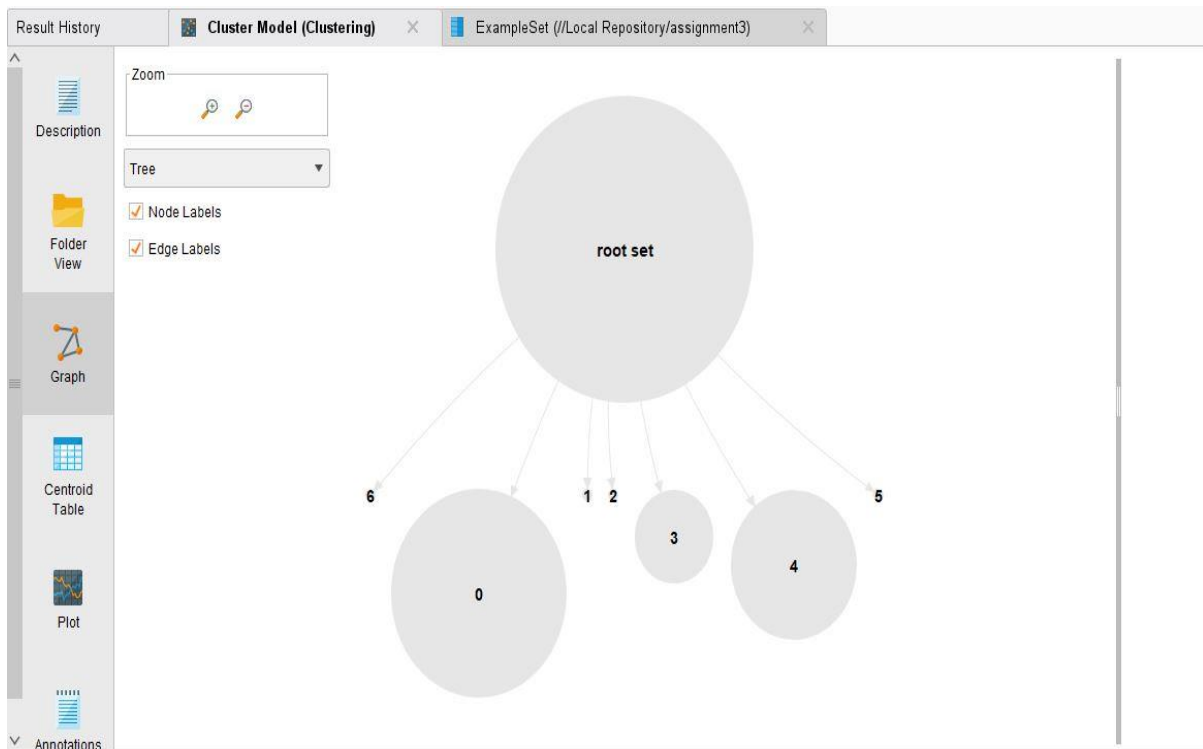
After testing the results of the model under different parameters we set the number of 'max runs' to 7 and k value is also set to 7.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
Description = BAG ...	0.001	0	0	0	0	0	0
Description = CAN...	0.001	0	0	0	0	0	0
Description = SWE...	0.001	0	0	0	0	0	0
Description = VINT...	0.001	0	0	0	0	0	0
Country = United K...	0.978	1	1	0.665	1	0.938	1
Country = France	0	0	0	0.054	0	0	0
Country = Australia	0	0	0	0.038	0	0	0
Country = Netherla...	0	0	0	0.003	0	0.031	0
Country = Germany	0	0	0	0.048	0	0	0
Country = Norway	0	0	0	0.193	0	0.031	0
Country = EIRE	0.022	0	0	0	0	0	0
InvoiceNo	0.887	0.995	-0.861	-0.608	-0.914	-0.900	-1.318
Quantity	-0.190	-0.292	16.761	0.149	-0.110	3.918	-0.292
UnitPrice	0.025	41.130	-0.161	-0.077	-0.065	-0.168	11.715
CustomerID	-0.076	0.000	-0.088	-1.495	1.003	0.034	-1.213

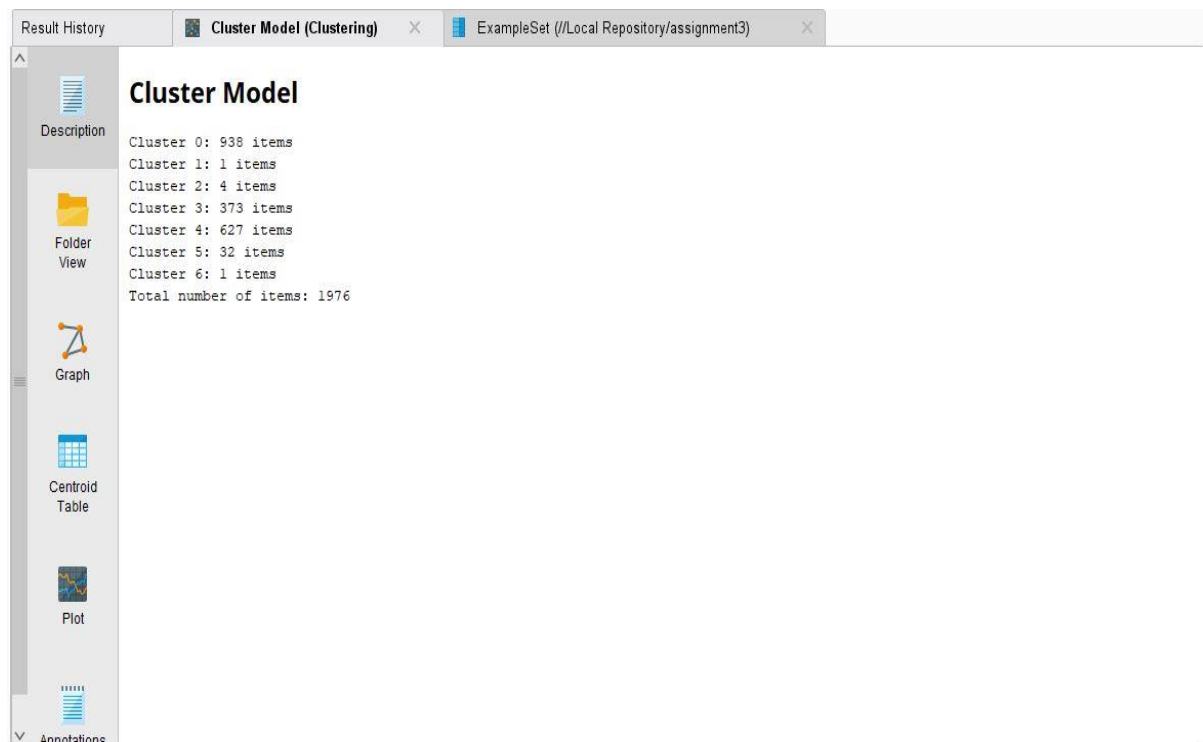
This is the centroid table of 7 clusters.



The above 2 figures show the 'Folder View' of 7 clusters having each a new folder of itself. By opening the subfolder we get an information of the cluster which is shown in the 2nd figure.



This is the Graph of a 'tree' with the clusters and each cluster is shown as big according to their clustering values.



This figure shows the results we get after running the process. It shows the total cluster we get with their respective items in them. And a total of items at the end.

Report:

Discuss what is interesting about them and describe what iterations of modeling you went through, such as experimentation with different parameter values, to generate the clusters.

The k-means operator is challenging to understand and to perform. We basically didn't went to many iterations but a lot of iterations is performed on the k-means operator single-handed to generate different values for the clusters and to get the 'k' numbers for the cluster and the 'max runs' of the cluster. We ended up with the 7 clusters with as much is the max runs.

Explain how your findings are relevant to your original question.

The question is to understand the customers' purchasing items' ability from different countries. We ended up that the United Kingdom's customers are on the top of list after the clustering.

References:

Dataset

<https://www.kaggle.com/puneetbhaya/online-retail/version/1>

k-means for rapidminer

<https://www.youtube.com/watch?v=EgXQvvbmtmM&t=413s>

k-means for orange

<https://www.youtube.com/watch?v=vgmL808eSw4&t=158s>