Homework 2. Study the Ecommerce Python Notebook

Note – these words are interchangeable:

* + Customers, clients
  + Product, item
  + Order, transaction, purchase

First, describe at a high level what the author does in the ecommerce notebook. It could look like this:

* Read in and examine the data
* Look for ?? and do ??
* Build a model to do ?? using ?? (model type, data, fields)
* Examine the results using ??
* …

Second, answer questions about ecommerce notebook

* What is the size of the data file (number of records and number of fields)
* What are the three major entities in the data?
* What country has the most purchases?
* What does the C mean in front of the invoice number?
* What are the two primary things done to clean the data?
* What is the entity level of the field Description that is used for clustering?
* What is the final number of keywords the author uses for classifying the products?
* What are the rows and columns of the matrix X?
* What are the elements that fill the matrix, and what do they mean?
* What algorithm is used to cluster the products?
* How does the author select the number of product clusters, and the final clustering choice?
* How does the author manually/visually assure himself that these product clusters are meaningful?
* The author then groups purchases by customer so he can look at all the purchases for each customer. Before he aggregates all the customers’ purchases into a single record/total for each customer he first separates out the last 2 months of purchases. Why?
* He then groups the purchases by customer. What are the summary fields he makes for each of these customer records?
* What method (not library) does he use to scale the resulting customer summary values?
* What algorithm does the author use to cluster customers? How many customer clusters does he make?
* The author then looks at these customer clusters visually in what space?
* How does the author get a feeling for the characteristics of each of the customer clusters?
* For model to classify customers, what are the 6 inputs to the model?
* He then trains how many different classifier model types?
* What data does the author use to train the model? Test the model?
* How does he decide the final predicted customer category based on the outputs of these classifier models?
* How long does the python notebook take to run?