Performing Data Mining on Retail Data

Problem:

Experiment with conducting Machine Learning and Data Mining related to consumer segmentation from given retail dataset. Many retailers have made great use of such consumer segmentation algorithms in suggesting products to its users. For this purpose we use the following basket dataset from UCI: http://archive.ics.uci.edu/ml/datasets/Online+Retail

Then install the required APIs including: pip install pandas. Here We explore the use of Recency, Frequency and Monetary (R_F_M) analysis on the above retail dataset. The goal of this analysis is to determine quantitatively which customers are the best ones by examining recency of a customer's purchase has purchased, the purchase frequency, and the customer's monetary spends. We perform R_F_M analysis because 80% of the business comes from 20% of your customers.

The R_F_M will be used for business intelligence on consumer segmentation through answering questions like:

- Who are my best customers?
- Which customers are at the verge of churning?
- Who has the potential to be converted in more profitable customers?
- Who are lost customers that you don't need to pay much attention to?
- Which customers you must retain?
- Who are your loyal customers?
- Which group of customers is most likely to respond to your current campaign?

More on understanding the metrics used for R_F_M analysis:

https://www.putler.com/rfm-analysis/

Proposed Solution:

RFM Analysis is done for successful customer segmentation. The provided UCI dataset has records of customer purchases in Excel format. It consists of the following:

- Invoice Number
- Stock Code
- Description
- Quantity
- Invoice Date
- Unit Price
- Customer ID
- Country

We are to calculate and analyze the Recency, Frequency and Monetary spending of a customer (RFM) in order to group customers based on their transaction history.

Counting the number of Invoices issued per customer will give us the frequency. Calculating the number of days passed since the last issued Invoice per customer will give us the Recency and lastly, the sum total of all the spending per customer will give us the required Monetary value.

Since it is evident that none of the calculations can be made without the Customer ID, we remove all entries with a null Customer ID as our first step for data pre-processing.

Secondly, we set our date of study, i.e. today. It is one day more than the last recorded invoice date in the dataset. Today minus the last recorded invoice date per customer ID gives us the Recency.

We calculate the total price for each Invoice first. Then we group all the invoices per customer. The number of invoices per customer becomes the Frequency and the sum of its total prices is the Money spent per customer.

After achieving the RFM values, we evaluate and assign each customer RFM scores. These scores are in a range from 0 to 5. The scoring is done by dividing the entire population into 6 equal quintiles, i.e. 16.67% each. High frequency and monetary values, say of the people in the last quintile are given a score of 5 each whereas low recency values like those in the first quintile is given a high score of 5 and vice versa.

After getting the RFM scores, we now move towards segmenting.

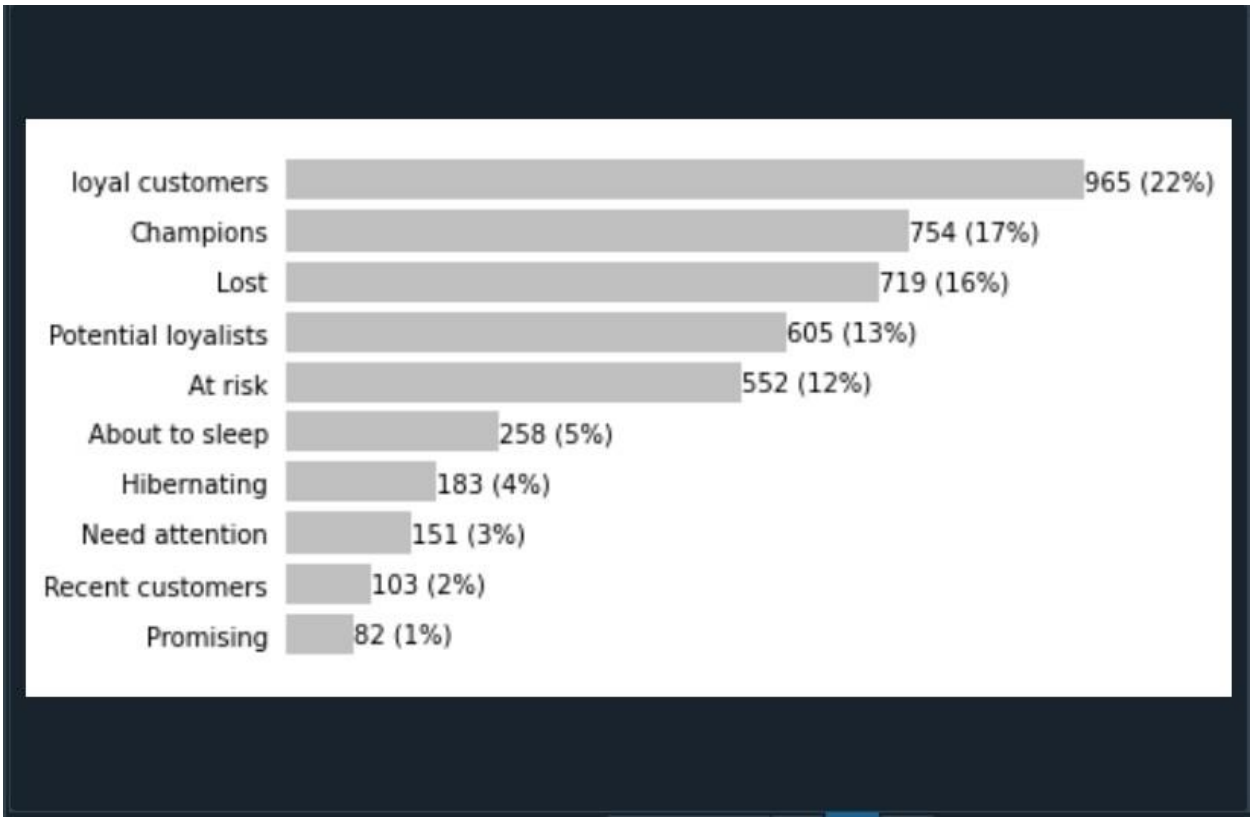
Following the Putler-RFM-Analysis, we do the following:

Customer Segment	Recency Score Range	Frequency & Monetary Combined Score Range
Champions	4-5	4-5
Loyal Customers	2-5	3-5
Potential Loyalist	3-5	1-3
Recent Customers	4-5	0-1
Promising	3-4	0-1
Customers Needing Attention	2-3	2-3
About To Sleep	2-3	0-2
At Risk	0-2	2-5
Can’t Lose Them	0-1	4-5
Hibernating	1-2	1-2

Lost	0-2	0-2

We combine the frequency and monetary scores by taking their average as they fall in the same line, i.e. customer spendings and on the other axis we plot the recency.

Lastly, we have calculated the number of customers in each of these eleven segments and plotted the retrieved data in a graph.



This graph along with the table stating which segment each customer belongs to answers all the questions mentioned in the problem statement above.