Documentation

First phase

First of all, as suggested, we went for finding those attributes : frequency (F), recency (R), lifetime/tenure (T), and monetary value (M). As we could see in the dataset every record represent one item of the shopping list of certain customer.

The simplest thing we could do is first of all add a new column, namely "AmountSpent" which is derived as a product of quantity and unit price.

(F) We derived the frequency on the following way. We created lambda function that takes CustomerId. After that we have taken counted how much unique InvoiceNo's were assigned to given CustomerId and simply returned that number.

(R) Also, created function that served as a lambda function. Also this function takes CustomerId. We took the subset of main dataset with just given CustomerId. Now we did some data wrangling regarding InvoiceDate of that subset and finally sorted id. Using datetime library we calculated number of days that have passed from the *last* date in sorted subset.

(T) Literally, same thing applied as in (R), we just taken number of days from the *first* date in the sorted subset.

(M) Once again, function that takes CustomerId, and first takes subset where CustomerId is given Id. Now iterating through the given subset we just added AmountSpent values and returned the sum.

After that we created new dataset where columns were only : Id, Frequency, Recency, Lifetime, MonetaryValue. Also, to avoid repetition of this procedure, we saved this dataframe and exported it to new '.csv' file. (The work described here is summarized and executed through 'Starting Notebook for Assigment.ipynb')

Second phase

Most of the work in the second phase was mainly based around finding and extracting outliers. We want for the most concrete practice. We assumed outlier was a record where his attribute was not in interval [Q1-1.5*IQR, Q3+1.5*IQR], where IQR is so called inter quantile range, or in practice Q3 - Q1.

In addition, Q1 and Q3 and first and third quantiles. Now applied this logic to every column, and finally we got clean data that we wanted. After we cleared the data, we also tried with the heatmap of the correlation. Unfortunately, nothing special occurred. Only thing that was extractable is that Frequency and MonetaryValue are highly correlated.

But this is something that we would expect without any fancy heatmaps, since the more you come, more likely is that you will spend more money.(The work described here is summarized and executed through 'Exploratory Analysis and possible additional filtration .ipynb')

Third phase

Now, we are approaching the last phase - model building. Since we were suggested to apply k means clustering, we are faced with problem of calculating right "k". We tackled that problem of using elbow method. Python has very nice library that is used for determining optimal "k" - it's called yellowbrick. And based on the data it turns out that optimal k is 4. So we proceeded with that k. What is necessary for every k means clustering is data preparation. We have checked wheter the data is skewed or not. If it is we would apply log transformation to make it look more Gaussian. Also, one important thing, since k means is working with Euclidean distance (generally this doesn't have to be the case) we have to normalize each column, i.e. make their mean equal 0 and standard deviation 1. And that is what we did. After that we expanded the dataset with these new predictions/labels. After grouping the data we see some interesting results. Cluster 0 has by far highest monetary value and frequency, or in other words they spend a lot and come often - we want as many of these as possible. Cluster 1 customers visited us long time ago, come after that reasonable amount of times, but didn't spend a lot of money. If we motivate them in some way, they might start visiting us frequently. Cluster 2 visited us last time as customers from Cluster 1, but not as nearly as frequent as them, nor did they spend same amount of time. But interesting thing is that they started to buy very recently, so they can't be "blamed" for not performing so well from the sellers perspective. And the Cluster 3 is the worst one - they don't buy often and they don't spend a lot of money. Let's now try to summerize in couple of words.

- Cluster 0 - loyal, spend a lot of money (Active Loyals)
- Cluster 1 - semi loyal, not spending a lot of money (Habitual Loyals)
- Cluster 2 - questionable loyalty since they started buying recently, not spending a lot of money (Situational Loyals)
- Cluster 3 - not loyal, not spending a lot of money (Active Disloyals)