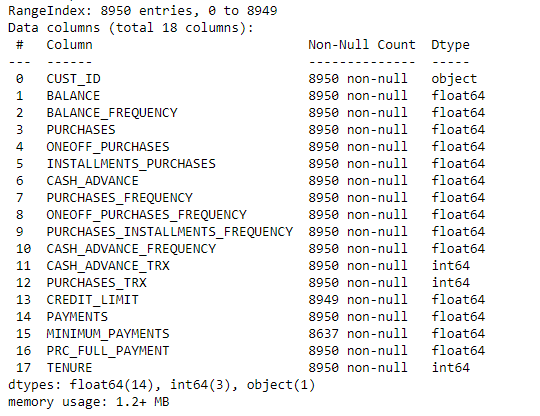
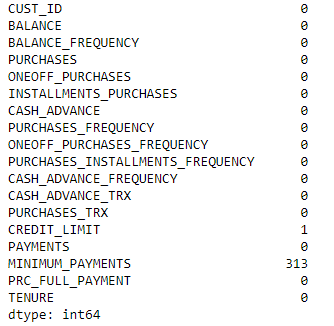
**Unsupervised Learning**

The dataset from CC GENERAL.csv has 8950 rows and 18 columns.



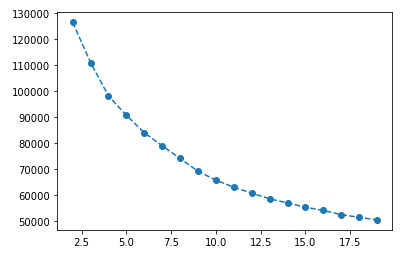
Above I have included information of the dataset. We can see that some data is missing. Below I have included count for number of values missing for each column.



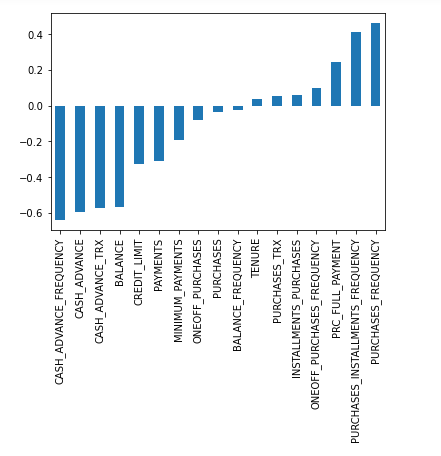
Since only 1 row has the PAYMENTS column empty, we can drop that row. For the rows missing value for MINIMUM\_PAYMENTS,

* If PAYMENTS = 0, fill with 0
* Else drop row

I then drop the CUST\_ID column and run K Means clustering algorithm for K ranging from 2 to 20 (not including 20). Plotting the sum of squared distances vs K we get,



Using the elbow method, we can see that K=4 is probably a pretty good choice. I then fit the model for K = 4. I then plot the correlation of the features with the cluster labels.



From this we can see that the features that are more negatively correlated are,

* CASH\_ADVANCE\_FREQUENCY
* CASH\_ADVANCE
* CASH\_ADVANCE\_TRX
* BALANCE

And the features that are more positively correlated are,

* PURCHASES\_FREQUENCY
* PURCHASES\_INSTALLMENTS\_FREQUENCY