Lab 8 Report

> Credit Card Clustering and Classification assignment

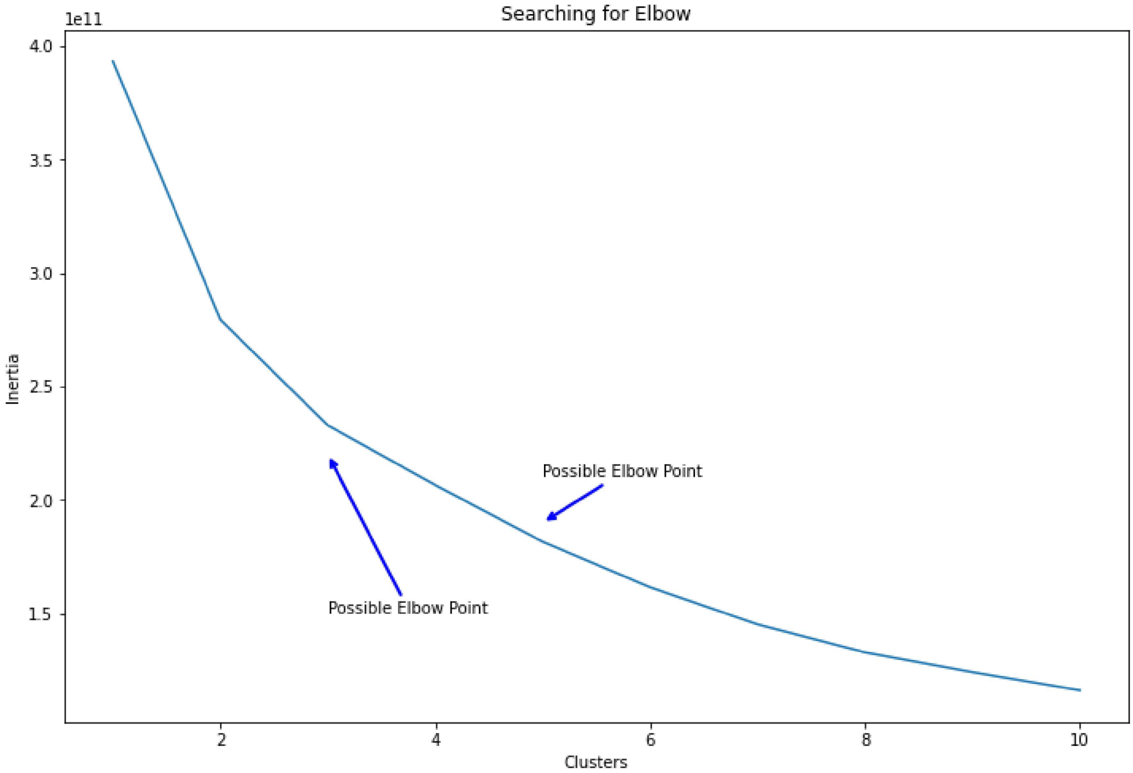
Task i

There are some corrupted values or "NaN" in the variables-"Minimum_payments" and "Credit_limit". I replace all the corrupted values to their median value of that variable. The data is then normalized to mean 0 and std 1.

From the variables, I think "Credit_limit" and "Purchases" would give the best separation because normally there are active customers who use their credit cards a lot so have a high purchase tendency. But still, there are inactive customers who don't use that much, even if they have a high credit limit. From this tendency, we could separate a few types of customer. So I expect this combination gives the best separation.

Task ii

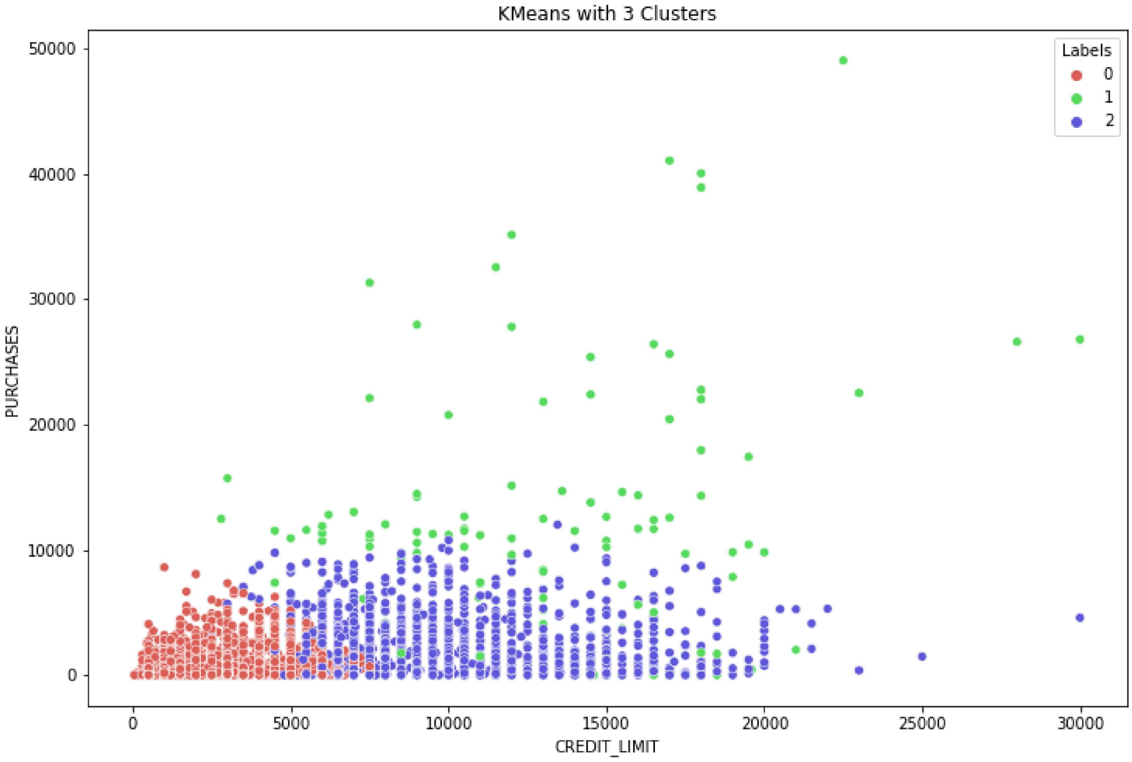
I first plot the "inertia" using K-means method for $k = 1$ to $k = 11$. The plot is shown below:

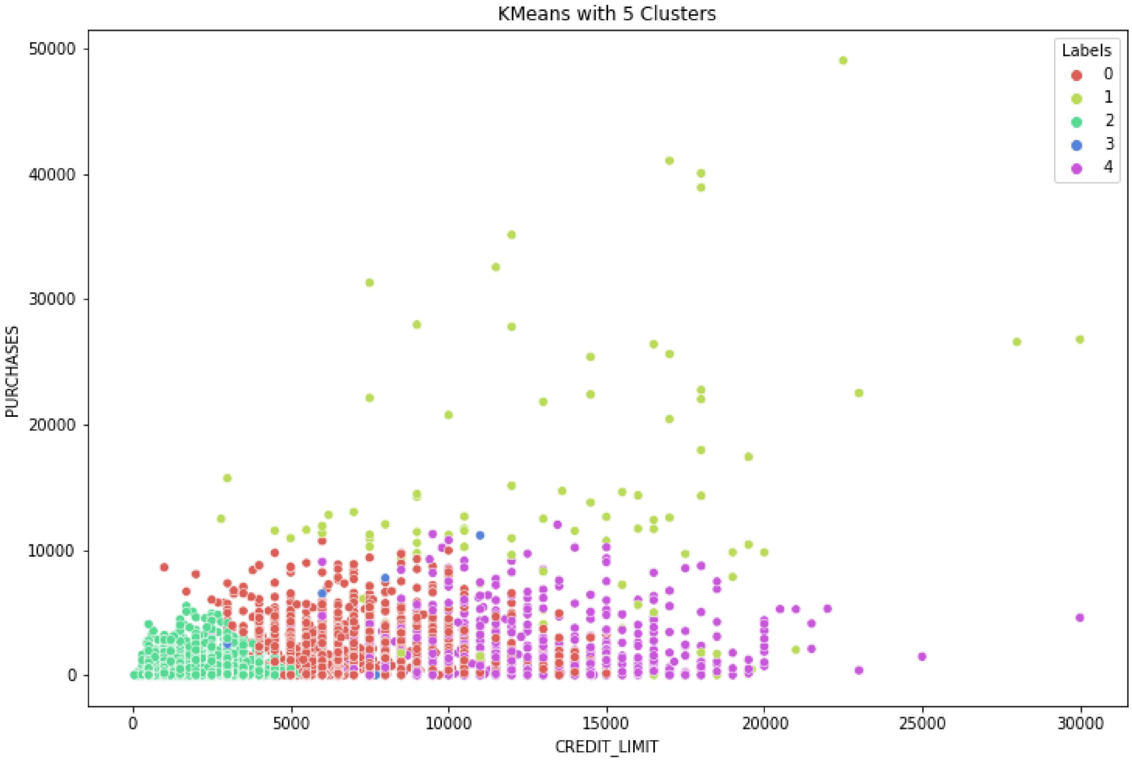


The possible "elbow" point shows that good candidates for *k* value are *3* and *5*.

Then I plot the combination of 'credit limit' and 'purchases' based on task 1. The plots below shows the clustering on this combination when k =3 and k =5.

Task iii

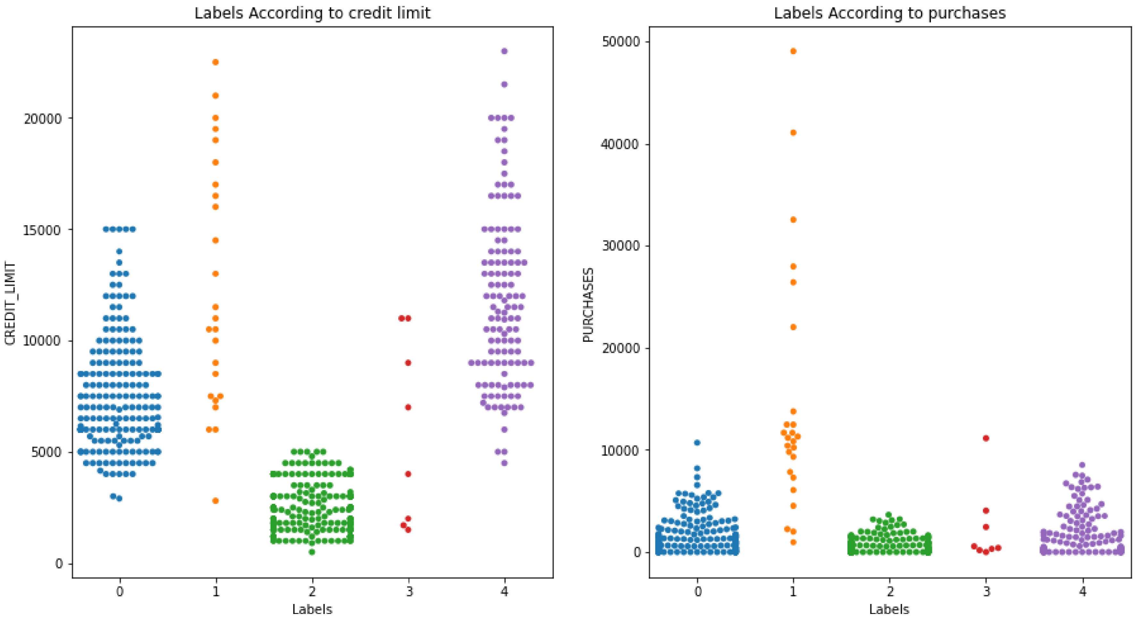




We see that k=3 and k=5 both give good separations. However, it is safer to say there are 5 clusters, because we only plot 2 variables now. Things could be more complex in reality. So I choose k=5 be the best value.

From this 5 clusters plot, we can see the 5 separated groups:

- Label 0: medium credit limit, less spending
- Label 1: medium to high credit limit, high spending
- Label 2: less credit limit, less spending
- Label 4: high credit limit, meidum to less spending
- Label 3: only a few present, could be inactive users or noisy data



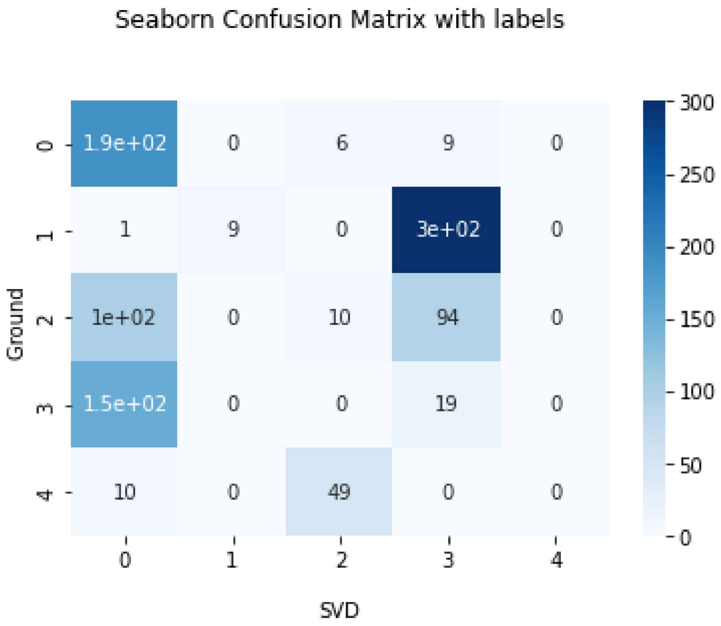
Task iv

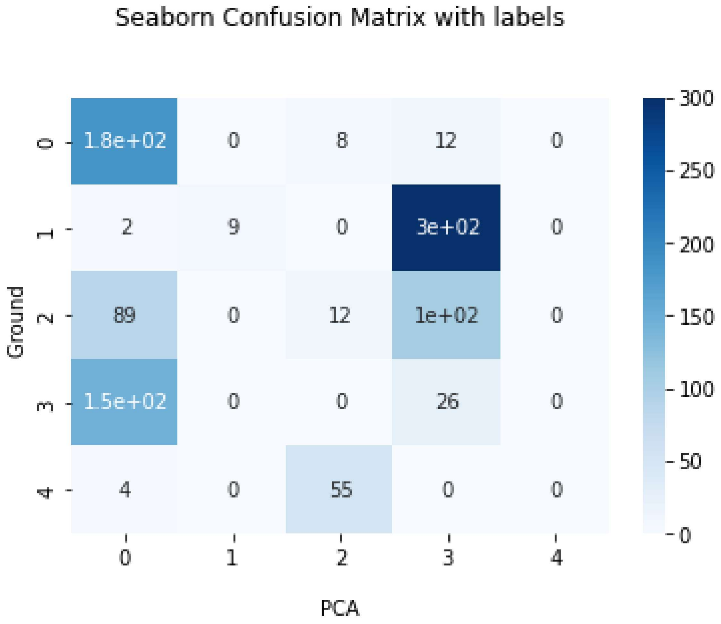
I use r value of 5

Task v

The coordinates for the 5 cluster centers are transformed to svd and pca, using the same labels assigned from K-means. Then I assign the nearest cluster centers to label the test data.

Then using full coordinates as ground value, the two confusion matrices for svd and pca are computed and plotted as below





It looks like the most affected class is class 1: which is users with medium to high credit limit, and moderate spending.

The least affected class is class 4, which is users with medium to high credit limit, and high spending.