**Practice Problem: Segmentation**

This case requires to develop a customer segmentation to define marketing strategy.

The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months.

The file is at a customer level with 18 behavioral variables.

Following is the Data Dictionary for Credit Card dataset:

CUSTID : Identification of Credit Card holder (Categorical)  
BALANCE : Balance amount left in their account to make purchases   
BALANCE\_FREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)  
PURCHASES : Amount of purchases made from account  
ONEOFF\_PURCHASES : Maximum purchase amount done in one-go  
INSTALLMENTS\_PURCHASES : Amount of purchase done in installment  
CASH\_ADVANCE : Cash in advance given by the user  
PURCHASES\_FREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)  
ONEOFF\_PURCHASES\_FREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)  
PURCHASES\_INSTALLMENTS\_FREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)  
CASH\_ADVANCE\_FREQUENCY : How frequently the cash in advance being paid  
CASH\_ADVANCE\_TRX : Number of Transactions made with "Cash in Advanced"  
PURCHASES\_TRX : Number of purchase transactions made  
CREDIT\_LIMIT : Limit of Credit Card for user  
PAYMENTS : Amount of Payment done by user  
MINIMUM\_PAYMENTS : Minimum amount of payments made by user  
PRC\_FULL\_PAYMENT : Percent of full payment paid by user  
TENURE : Tenure of credit card service for user

For the following questions, use K-means clustering (with all above values highlighted in **yellow**).

1. Using a cluster model with 6 clusters, which cluster has the highest credit limit?
2. What is the cluster means of this cluster?
3. Does this cluster have the highest number of data points? If not, what does this suggest about this customer segment?
4. If the bank wanted to maximize their customer outreach with promotional offerings which clusters, it would target?
5. Which cluster is characterized by the largest amount of payments made to the bank? 6
6. What are the means for the cluster selected in Q5?
7. What do the means of the cluster selected in Q5 tell us about the customer profile?
8. The cubic clustering criterion (CCC) can be used to estimate the optimal number of clusters. Looking into CCC of different number of clusters, which CCC shows the most optimal number of clusters.

For the following questions, use K-means clustering (with all above values highlighted in **grey**).

1. Create a model with 3,4,5, and 6 clusters. Which cluster model is the optimal?
2. What is the cluster means of this cluster?
3. Which clusters are the top two in terms of the average number of transactions?
4. Where else these two cluster have the highest means?
5. What could tell us about these two segments?
6. Are there any areas in which these two clusters with the highest number of transactions and purchases made in installments perform worse compared to other clusters?

For the remaining questions, use hierarchical clustering (with all values except ‘CustomerID)

1. Look at difference in distance between going from a 5 cluster model to a 6 cluster model, compared to going from a 4 cluster model to a 5 cluster model. Based on your understanding of distance and length of the lines in the dendrogram, which model (5 clusters or 4 clusters) would be best to use? Why?
2. Which cluster contains the lowest amounts for PAYMENTS, MINIMUM\_PAYMENTS, PRC\_FULL\_PAYMENT?