Task 3:

We have data that was collected from direct marketing campaigns of a banking institution. The task is to cluster or group the customers on the basis of their attributes.

(1) Explain how you will preprocess the data and handle null values (usually, real-world datasets are never clean. Null values can be present because of many reasons ranging from a simple data entry error to a loss of data.)

(Ans 1) To preprocess the raw data, we can first split this process into three groups where the first group talks about how to handle null 'numeric' values, the second group talks about how to handle inconsistent categorical data points and the third group talks about how to make the raw data more quickly processed i.e scalable when fed into the model.

First -> For managing null numeric values(in this dataset that would be the "age" feature), we have two choices generally. The first is to delete the datapoint/row if the other values in that row are inconsistent as well. Usually, we remove the datapoint when 75% of the datapoint is useless. We always have to be wary about the data we remove because in a way we are losing information and inherently keeping the bias unaltered. So if we don't have a lot of sample data (in cases of let's say cancer dataset as it requires rare found medical expertise) then this is a good thumb rule to remember. Another choice is to replace the value with a statistical correlated value. In this case, the first choice is more apt and logical as 'age' is not a really statistical correlated value in this scenario unless given more data about the person.

Second -> To handle categorical data points we first need to encode them into a mathematical format that can be understood easily by the computer but we have to make sure that the computer doesn't interpret the mathematical format as a value of that datapoint and think that one class is more valuable than another class. To resolve this, we use dummy variables(we can make these from pandas library) which indicate the absence or presence of some categorical effect. In this dataset, we will this method on job, marital, education, default, housing, loan, contact, month and day_of_week

Third -> Last step is to scale the data so that we can feed the dataset into the model. For this dataset, the optimum choice of scaling for the upcoming clustering algorithm I am going to use will be StandardScaler. This scaling function brings all features to the same magnitude as it makes the mean=0 and variance=1 for the entire training dataset.

(2) What clustering algorithm you'll use and why? Explain what you understand about the algorithm in detail. How will you determine the number of clusters to divide the customers into?

(Ans 2) I will use density-based spatial clustering of applications with noise i.e DBSCAN algorithm for the following dataset. The main benefit of DBSCAN are that it doesn't require the user to set the number of clusters, it can capture clusters of complex shapes and it can identify points that are not part of any cluster. The only drawback is that it's slower than kmeans which is

very fast and robust for large real-world datasets but it still works well in this scenario comparatively. DBSCAN works by the idea that clusters form "dense" regions of data containing "core points" separated by relatively empty regions. There are two kinds of parameters, one is minimum samples(s) and another is the distance(d) between an arbitrary new point and core sample. So, DBSCAN classifies the customers using these two parameters that is, "if there are at least 's' points within distance 'd' then that point is a part of the core sample". This way it creates and forms new clusters and also if some points are isolated together but are less than 's', those are considered as 'noise' and not as a classifiable cluster in the dataset. If new points are given in the dataset, the same process is repeated. For this model to successfully work, we need to define and tune the values of minimum samples 's' and distance 'd' very carefully as some sensitive datapoints of a cluster may get their class changed due to any difference in the values(s and d).