Solution 1) In this case, you need K-means algorithm because it works on unlabeled numerical data. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. A cluster refers to a collection of data points aggregated together because of certain similarities. So, value of K = 3 as we are asked to separate 2,000 unlabeled into 3 groups so, **three** clusters will be formed.

K-means clustering is an extensively used technique for data cluster analysis. It is easy to understand, especially if you accelerate your learning using a K-means clustering tutorial. Furthermore, it delivers training results quickly.

Solution 3)

Difference between Supervised Learning and Unsupervised learning are followings:

Supervised Learning:

Supervised learning algorithms are trained using labeled data.

Supervised learning model takes direct feedback to check if it is predicting correct output or not.

Supervised learning model predicts the output.

Supervised learning needs supervision to train the model.

Supervised learning can be categorized in Classification and Regression problems.

Supervised learning can be used for those cases where we know the input as well as corresponding outputs.

Supervised learning model produces an accurate result.

Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.

The goal of supervised learning is to train the model so that it can predict the output when it is given new data.

It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.

2) Unsupervised Learning:

In supervised learning, input data is provided to the model along with the output.

Unsupervised learning algorithms are trained using unlabeled data.

Unsupervised learning model does not take any feedback.

Unsupervised learning model finds the hidden patterns in data.

In unsupervised learning, only input data is provided to the model.

The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.

Unsupervised learning does not need any supervision to train the model.

Unsupervised Learning can be classified in Clustering and Associations problems.

Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.

Unsupervised learning model may give less accurate result as compared to supervised learning.

Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.

It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Solution 3)

speaking these are the steps. Of course, these can be different, depending on the type of problem, the data, the tools available, etc.

1. Problem definition - The first step is to understand the business problem. What is the problem you are trying to solve - what is the business context? Your client can also give you a lot of data and ask you to do something with him. In this case, you will need to review the data in more detail. Nevertheless, if the customer has a specific problem to solve, then Yo2u will have to convert the business problem into a problem of analysis. I do not know what you will understand exactly what you are going to predict. There is no point in building a fabulous model, only to realize that it is predicting exactly what the business needs.

2. Data mining - Once you have defined the problem, the next step is to explore the data and become familiar with it. This is especially important when dealing with an entirely new dataset.

3. Data Preparation - Now that you have a good understanding of the data, you will need to prepare it for modeling. You will identify and process missing values, detect outliers, transform variables, create binary variables if needed, and so on. This course is very influenced by the modeling technique. For example, regression involves a readiness measure, but it can not be compared to other comparison methods. Read also The workings of a good data seeker

4. Modeling - Once the data is ready, you can start modeling. This is usually an iterative process where you run a template, evaluate the results, fine tune your approach, run another template, evaluate the results, readjust and so on ... are happy with what you think is the best possible result given data.

5. Validation - The final model (or perhaps the two or three best models) should then be submitted to the validation process. During this process, you test the model using entirely new data, that is, data that was not used to build the model. This process ensures that your model is a good model in general

Some Basic Techniques For handling Sparce data or Missing data are followings:

I. Deletion

Unless the nature of missing data is ‘Missing completely at random’, the best avoidable method in many cases is deletion.

a. Listwise: In this case, rows containing missing variables are deleted.

In the above case, the entire observation for User A and User C will be ignored for listwise deletion

b. Pairwise: In this case, only the missing observations are ignored and analysis is done  on  variables present.

In the above case, 2 separate sample data will be analyzed, one with the combination of User, Device and Transaction and the other with the combination of User, OS and Transaction. In such a case, one won't be deleting any observation. Each of the samples will ignore the variable which has the missing value in it.

Both the above methods suffer from loss of information. Listwise deletion suffers the maximum information loss compared to Pairwise deletion. But, the problem with pairwise deletion is that even though it takes the available cases, one can’t compare analyses because the sample is different every time.

II. Imputation

a. Popular Averaging Techniques

Mean, median and mode are the most popular averaging techniques, which are used to infer missing values. Approaches ranging from global average for the variable to averages based on groups are usually considered.

For example: if you are inferring missing value for Revenue, you might assign the average defined by mean, median or mode to such missing value. You could also consider taking into account some other variables such as Gender of the User and/or the Device OS to calculate such an average to be assigned to the missing values.

Though you can get a quick estimate of the missing values, you are artificially reducing the variation in the dataset as the missing observations could have the same value. This may impact the statistical analysis of the dataset since depending on the percentage of missing observations imputed, metrics such as mean, median, correlation, etc may get affected.

The above table shows the difference in imputed missing values of Revenue arrived by taking its global mean and mean based on which OS platform it belongs to.

b. Predictive Techniques

Imputation of missing values from predictive techniques assumes that the nature of such missing observations are not observed completely at random and the variables chosen to impute such missing observations have some relationship with it, else it could yield imprecise estimates.

A predictive model could be used to impute the missing values for Device, OS, Revenues. There are various statistical methods like regression techniques, machine learning methods like SVM and/or data mining methods to impute such missing values.