1. What exactly is a feature? Give an example to illustrate your point.

Ans: **feature is a column in a data set. For example in a data consisting age and income columns, these both are features. One of them is a dependent feature and the other one is an independent feature.**

1. What are the various circumstances in which feature construction is required?

Ans: **feature construction is required when an existing column can provide you a separate feature for further exploratory data analytics.**

1. Describe how nominal variables are encoded.

Ans: **nominal encoding is achieved when there is no ranking in the categorical variable such as colors. We usually create k-1 features for nominal encoding and pd.get\_dummies can be used for this encoding.**

1. Describe how numeric features are converted to categorical features.

Ans:  **we can create bins or ranges to convert numeric features into categorical features such as age between 10 and 25 is 1 cateogry and the age between 26 and 45 is 2nd category.**

1. Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach?

Ans: **we defined different set of features and try to fit the model on each subset of features until we get the best results out of the model. The model is evaluated against the same performance matrix such as R2 or adjusted R2. It is computationally expensive but it can give you a best subset of features for which model performs the best.**

1. When is a feature considered irrelevant? What can be said to quantify it?

Ans: **feature is said irrelevant when there is no correlation between the independent feature and the dependent feature. In other words, the feature is unable to explain the variation in the target variable. For quantification or visualization: df.corr() method can be used or pairplot can be drawn to ascertain the relationship between variables.**

1. When is a feature considered redundant? What criteria are used to identify features that could be redundant?

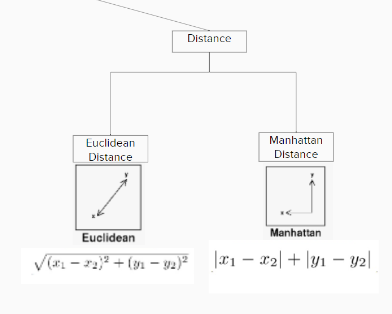
Ans: **redundant features are those which are corelated with other features and give same information about the target variable. They can be identified with the help of corr() method.**

1. What are the various distance measurements used to determine feature similarity?

Ans: **the most commonly used one is Euclidean distance to determine the feature similarity. Other ones are Cosine similarity or Manhattan distance. Euclidean is the shortest path between two points.**

1. State difference between Euclidean and Manhattan distances?

Ans: **below are the formulas to calculate each of these types of distance. Euclidean distance is the shortest path between two points and takes the square value of the distance between points whereas Manhattan distance takes the absolute values of the distance.**



1. Distinguish between feature transformation and feature selection.

Ans: **Feature selection is just selecting the relevant features from the original set of feature without making any changes in the values but feature transformation involves transforming the features in order to make them relevant to the target variable.**

11. Make brief notes on any two of the following:

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

3. The width of the silhouette: **this is used to measure how an object is similar to its own cluster compared to the neighbouring cluster. The value ranges between [1.-1] and the higher the value the more it has chances to belong to its own clusters.**

4. Receiver operating characteristic curve

Ans: **ROC curve is used to determine the threshold value for the classification problems. We compare the values of FPR and TPR and based on that we decide the optimal value of threshold value. Additionally, the involvement of domain expertise is also required to assist in deciding the value of threshold. For evaluation, we can use the confusion matrix as well with ROC.**