**QUESTION 3 – Encoding Features**

Firstly, we must identify the target column which would be the “price” column for this example. Secondly, we must clean the data which will be the most tedious task, but when this is done, it should be very easy to apply any algorithm as desired. Be sure to drop any columns that are deemed to be useless or have too many missing data. Also, you can drop rows of data if critical entries are missing, they can also be filled with the mean value of that column for example. In this example, the “Unnamed:0” column would be dropped.

Most of the Machine learning algorithms cannot handle categorical variables unless we convert them to numerical values. Many algorithm’s performances vary based on how Categorical variables are encoded. Categorical variables can be divided into two categories: Nominal (No particular order) and Ordinal (some ordered). Within this example, the “cut”, “color”, and “clarity” columns are non-numeric and needs to be converted to numerical values.

There are many ways we can encode these categorical variables as numbers and use them in an algorithm. A very popular method is to use Label Encoding. In this encoding, each category is assigned a value from 1 through N (here N is the number of categories for the feature). For example, we can do such an encoding: (Cold = 0, Warm = 1 Hot = 2 Very Hot = 3) . However, it will be wrong to say that “Hot” is “Warm” x2. One major issue with this approach is there is no relation or order between these classes as shown above, but the algorithm might consider them as some order, or there is some relationship.

A Test Set needs to be created where 80% of the data can be the Train Set and 20% of the data will be the Test Set. For this, random sampling would not be used. Stratified sampling, where the dataset is divided into homogeneous subgroups called strata, allowing the right amount of data to be chosen from each stratum that best represents the overall dataset, would be used. This is done for higher accuracy. However, to perform stratified sampling, one important feature needs to be chosen to best predict the price of the diamond. To choose this important feature, we will need to use the Standard Correlation Coefficient to determine the features that are most correlated to the price of the diamond. Whichever feature is found to be most correlated, can be categorised into smaller sub-groups, to find the distribution of the price of the diamond against those categories. The Stratified Sampling would then be performed based on those categories.

After the train and test sets are stratified, and the categorical variables are encoded to numerical values, the dataset would be fit into a model where the Root Mean Squared Error is checked, the model would be evaluated using Cross Validation, and the model would be trained. We can preview accuracy tests on the model and preview results to see how well the model fits.

There are two types of supervised machine learning algorithms: Regression and classification. The former predicts continuous value outputs while the latter predicts discrete outputs. For instance, predicting the price of a house in dollars is a regression problem whereas predicting whether a tumour is malignant or benign is a classification problem. In this example, we are predicting the price of diamonds, thus it is a regression problem. After testing the fit of the model as mentioned above, we can now apply Linear Regression to the dataset.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. If we plot the independent variable (x) on the x-axis and dependent variable (y) on the y-axis, linear regression gives us a straight line that best fits the data points. Within this example, the dependent variable (y) to be predicted, is the “price” column, and the independent variables (x), are the other features, such as “carat”, “cut”, and “color”.