**Part I**  
**Explain how you would handle missing data in categorical and numerical variables?**

R’s representation for blank or missing data is NA. Addressing of missing values early, can save some unnecessary work, and a lot of headaches. **Summary** function of the data helps to spot potential problem of missing data.

Fundamentally, there are two things can be done with the missing variables:

1. **drop** the rows with missing values: When missing data represents a fairly small fraction of the dataset, it’s probably safe just to drop these customers from your analysis
2. **Convert the missing values** to a meaningful value.

**For missing data in categorical variables**: The most straightforward solution is just to create a new category for the variable called missing.

**For missing values in numeric data:** replace the missing values with the expected, or mean based on the other input variables and this method of can also be applied to categorical data, as well.

**When values are missing systematically**: convert the numeric data into categorical data, and then

Use the above methods. Grouping approach can work well, especially if the relationship between the variables is no monotonic (where one variable is strictly not depending on the other variable). It does require the selection of good cuts.

**Part II**  
Give few data transformation techniques and cases where you would be applying them.  
Briefly explain the **log transformation** and when it should be used.

The need for data transformation can also depend on which modeling method you

plan to use. Some of the data transformation techniques are:

1) Converting continuous variables to discrete: Used where the exact value matters less than whether they fall

into a certain range. It brings out the relation between it and other variables.

2) normalizing variables: Normalization is useful when absolute quantities are less meaningful than relative ones. Normalizing by mean and standard deviation is most meaningful when the data distribution is roughly symmetric.

3) Log transformations: useful for skewed and wide distributions

Usage of log in other words linearize relationships among variables in regression model. For example, if the dependent variable is a multiplicative rather than additive function of the independent variables, or if the relationship between dependent and independent variables is linear in terms of percentage changes rather than absolute changes, then applying a log transformation to one or more variables may be appropriate.

**Log transformation for skewed and wide distributions:**

For the purposes of modeling, which logarithm you use—natural logarithm, log base 10, or log base 2—is generally

not critical. In regression, for example, the choice of logarithm affects the magnitude of the coefficient that corresponds to the logged variable, but it doesn’t affect the value of the outcome. It’s also generally a good idea to log transform data with values that range over several orders of magnitude—first, because modeling techniques often have a difficult time with very wide data ranges; and second, because such data often comes from multiplicative processes, so log units are in some sense more natural.