1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

= a) true

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

=a) Central Limit Theorm

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

=b) modelling bounded count data

Because Poisson distribution is used for modelling the unbounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

=d) all of the above

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

=c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

= b) false

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

=b) Hypothesis

The Null hypothesis is assumed true and statistical evidence is required to reject it in favour of research or alternative hypothesis

8. Normalized data are centered at \_\_\_\_\_\_and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

=a)

9.Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

=c) because the outliers can conform to the regression relationship

10. what do you understand by term Normal distribution?

=The Normal distribution is called Guassian distribution . The normal distribution is the most commonly seen continuos distributed in nature, just the binomial distribution every event is independent of another. In te normal distribution, mean, median and mode all line up such that the center of the distribution is the mean. Because of this , exactly half of results fail to either side of the mean. The normal distribution is also identifiable by its bell shape and may sometimes are referred as bell curve

11. How do you handle missing data? What imputation techniques do you recommend?

### = 1. Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

## 2.  Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

## 3.  Assigning An Unique Category

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values

## 4.  Predicting The Missing Values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear [regression](https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/) to replace the nulls in the feature ‘age’, using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.

## 5. Using Algorithms Which Support Missing Values

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values. In the titanic dataset, taking into account the person’s age, sex, class etc, we will assume that people having same data for the mentioned features will have the same kind of fare.

Unfortunately, the SciKit Learn library for the K – Nearest Neighbour algorithm in Python does not support the presence of the missing values.

Another algorithm which can be used here is RandomForest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

**The imputation we recommend are**

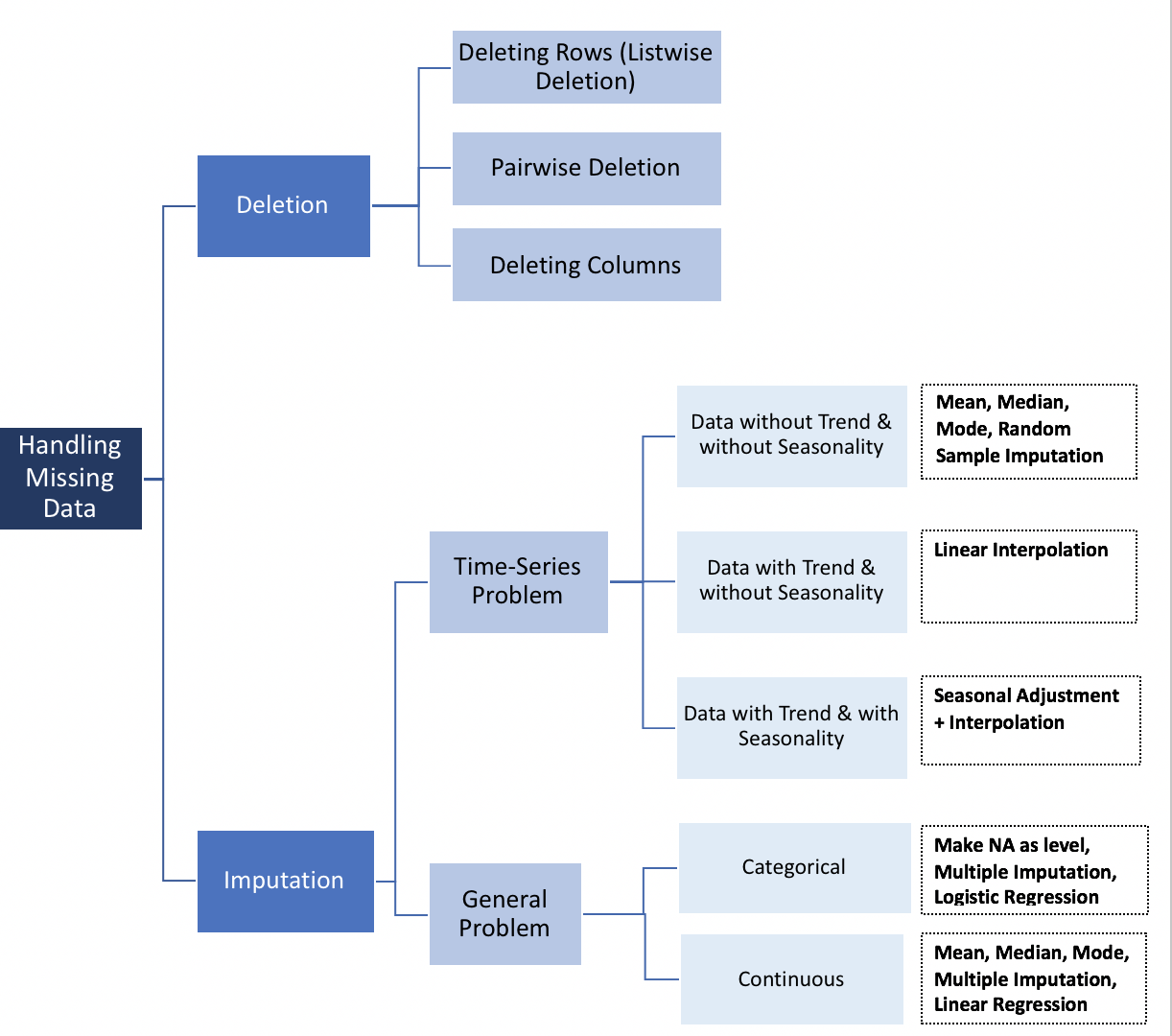
## Imputation of Categorical Variables

1. Mode imputation is one method but it will definitely introduce bias
2. Missing values can be treated as a separate category by itself. We can create another category for the missing values and use them as a different level. This is the simplest method.
3. Prediction models: Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test). We can use methods like logistic regression and ANOVA for prediction
4. Multiple Imputation

## KNN (K Nearest Neighbors)

There are other machine learning techniques like XGBoost and Random Forest for data imputation but we will be discussing KNN as it is widely used. In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors)  
The distance metric varies according to the type of data:  
1. Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine  
2. Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different.  
One of the most attractive features of the KNN algorithm is that it is simple to understand and easy to implement. The non-parametric nature of KNN gives it an edge in certain settings where the data may be highly “unusual”.  
One of the obvious drawbacks of the KNN algorithm is that it becomes time-consuming when analyzing large datasets because it searches for similar instances through the entire dataset. Furthermore, the accuracy of KNN can be severely degraded with high-dimensional data because there is little difference between the nearest and farthest neighbor.

**library**(DMwR)  
knnOutput <- **knnImputation**(mydata)In python  
from fancyimpute import KNN   
  
# Use 5 nearest rows which have a feature to fill in each row's missing features  
knnOutput = KNN(k=5).complete(mydata)



.