Missing values

How to deal with Missing values?

- Almost all datasets have missing values in it. So it's important to learn how to deal with them.
- There may be many reasons for missing data ranging from human errors during data entry, incorrect sensor readings, to software bugs in the data processing pipeline etc....
- We can easily detect what are the missing values using pandas.
- It is important to be handled as they could lead to wrong prediction or classification for any given model being used.
- We can remove those fields but Still, often there are hidden patterns in missing data points. Those patterns can provide additional insight into the problem you're trying to solve.

Missing values

Methods to deal with missing values?

- **Deleting Rows:** The most commonly used method to deal with missing 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. Complete removal of data with missing values results in a robust and highly accurate model. Deleting a particular row or a column with no specific information is better since it does not have a high weight. Loss of information and data. Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset

Missing values

Methods to deal with missing values?

- **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. Another way is to approximate it with neighbouring values. This is a better approach when the data size is small. It can prevent data loss which results in removal of the rows and columns. Imputing the approximations add variance and bias

Missing values

Methods to deal with missing values?

- **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 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. Imputing the missing variable is an improvement as long as the bias from the same is smaller than the omitted variable bias. Yields unbiased estimates of the model parameters. Bias also arises when an incomplete conditioning set is used for a categorical variable. Considered only as a proxy for the true values