What to do when we encounter missing data?

1. First determine how much data is missing. So if a column has more than say 10% missing values then do we even want to consider it for analysis? Usually no. If a row has more than 10% missing values we may want to eliminate that data point. But if the number of missing values is below 10% (or another threshold) then we can replace the missing values with one of the methods below.

2. Replace the missing value with the mean value of the column, or the median, or the mode - this is a very common method (with the mean)

3. Treat the column as a target variable, learn a regression model (like linear regression), and use the predictions to determine the missing values - also called imputation

4. If a row has a missing value for a given column then make copies of that row and for each copy insert one of the known values from the other rows.

Let's look at an example for method 4.

f1	f2
1	4
2	X
6	3

The method 4 is to create a new dataset that looks like this:

f1	f2
1	4
2	4
2	3
6	3