* not all datasets have complete data – some missing data e.g. number of bedrooms in housing dataset
* data imputation - filling missing data with values such as the mean/median value of the rest of the dataset
* Should not write own code to split data into test and train
  + Rely on scikitlearn to do this for you using its functions
* Stratified sampling – ensures your sample is representative of the real situation / data by using the same proportions of different data types as there is in the real situation / context