Things to notice when eyeballing the data

- **relative range of different features**: All the features should have approximately the same range. It is usually preferable to normalize all the features to take values between 0 and 1. Make sure that you use only the training data to establish the parameters of preprocessing.

- **the distribution of each feature**: It is preferable that the data follows a Gaussian distribution. If not, then apply some transformation such as taking logs, to arrive at a transformed feature that follows a Gaussian.

- **are there missing values**: If yes, then either remove the corresponding samples or else, fill them with an appropriate mean value

- **are there feature which are highly correlated to each other**: This can be established by looking at the heat map between the features. Features that are highly correlated to each other do not provide independent information. Applying a PCA to a dataset will help find independent features which will make the data more manageble.

- **are there imbalanced classes in the data**: If yes, then merely having a high accuracy in the predictions will not necessarily imply that a good fit has been obtained. In this case metrics such as precision, recall, F_β-scores and ROC are much better indicators of the goodness of our fit.

In [ ]: