Data preprocessing

How operations like imputation , feature selection, normalization etc. changes across the training and the testing data

- **Normalization** - should do it seperately
  - Normalization is a technique we use to normalize data and keep it center by subracting mean and dividing it with variance of total dataset
  - Training dataset and testing data set are two different datasets with some common features.
  - Training dataset is used to train your model using the past data so that it can predict what will happen with the features we have mentioned . Test data is real data .
  - Code snippet : `dataset = dataset - dataset.mean() / dataset.std()`
  - Here we can see we are using mean and variance that we get from data that exist , If we use whole dataset (train + test) to get mean and variance which will be used for train dataset , it means we are using some extra data (test data) on training of the train dataset
  - So , we have to do feature normalization seperately on train dataset and test dataset
  - This way we can avoid extra data interference in traing of our model
  - After training , we can normalize test data using mean and variiance of test data
  - Normalization makes it easy to compute the matrices as it makes all the values near to each other and small

- **Imputation** - should do it seperately
  - Imputation means dealing with missing data
  - There may be some missing values in provided dataset before applying any algorithm on the data you have to make sure the data is correctly labelled(data type etc)
  - There are many methods you can use to impute the data
  - some are :
    1. Mean imputation
    2. median Imputation
    3. mode imputation
  - Here similar to normalization we have to use parameter derived from whole dataset like mean ,median ,mode
  - These parameters are used to replace the missing value in data
  - If we use train + test datasets to get parameters for traing the dataset ,again we are using extra innappriate and unneccesary data , it gives you wrong parameters
  - So ,split the data and then impute the data
  - Though the process of imputation doesn't change but it cannot be used together

- **Feature Selection** -
  - Feature selection is nothing but selecting best features (approriate) out of many features mentioned in dataset
  - Sometimes using less features will give you best results than using many features to train the model ,it reduces the risk of overfitting
  - We select appropriate features using diffrenet methods and parameters
  - We use the data to get the parameters which decide what features we should use
  - one method could be calculating correlationa and using which are highly correlated , If we use test dataset also then we may not get correct parameters
  - There's no need to o features selection on test datase t , we can directly use the features selected from train data set