Boston Housing Prices Dataset

You should do this assignment in Jupyter Notebook. Try to explain all your choices in comments.

You can get the data with scikit-learn with

```
from sklearn.datasets import load_boston
boston_data = load_boston()
```

1) Load the data into pandas DataFrame.
2) How many examples are there in the dataset?
3) What is the number of features?
4) Plot each feature against ‘Median Value’ (housing price). What 3 features do you think are best suited to predict ‘Median Value’ with a linear model? Why these 3?
5) What are the statistics for the chosen variables (min, max, mean, std deviation)?
6) Split the data into the training and testing sets with train_test_split (testing set should contain 10% of all data).
7) Apply data normalization to the training set.
8) Fit the linear regression model on the training set. Report mean squared error and coefficient of determination R^2 on the training and testing sets (don’t forget to apply normalization to the testing set).
9) Try to add more features, report which features you added and what influence it had on the model performance.
10) Use 3 features you selected in the task 4 and fit regression with polynomial features with degree = 2, 5, 7. Report the model performance on the training and testing sets.
11) Do task 10 without data normalization. Report results.
12) Make some more features from existing features (for example, transform existing features, multiply feature with another feature, use nonlinear transform (log, exp), etc.), train regression and report results.

Useful functions:

Pandas: describe
Scikit-learn: train_test_split, LinearRegression, score

Linear Regression (Numpy)

1) Implement linear regression algorithm for one feature with numpy
2) Report run time of naive vs numpy implementations.
3) Bonus task:
    a) Implement linear regression algorithm for multiple features with numpy