Boston Housing Prices

- For this project you will investigate the Boston House Price dataset collected in 1970: http://lib.stat.cmu.edu/datasets/boston
- The data consist of 506 observations and 14 non constant independent variables
- The python programming language (+Sklearn) was used to conduct this analysis
- The aim of this study is to fit a regression model that best explains the variation in medv (Neural-Network is on-going).
- My codes and results are at: https://github.com/DianeDeeDee/Python_Machine_Learning/tree/master/Housing/share

Analyze data

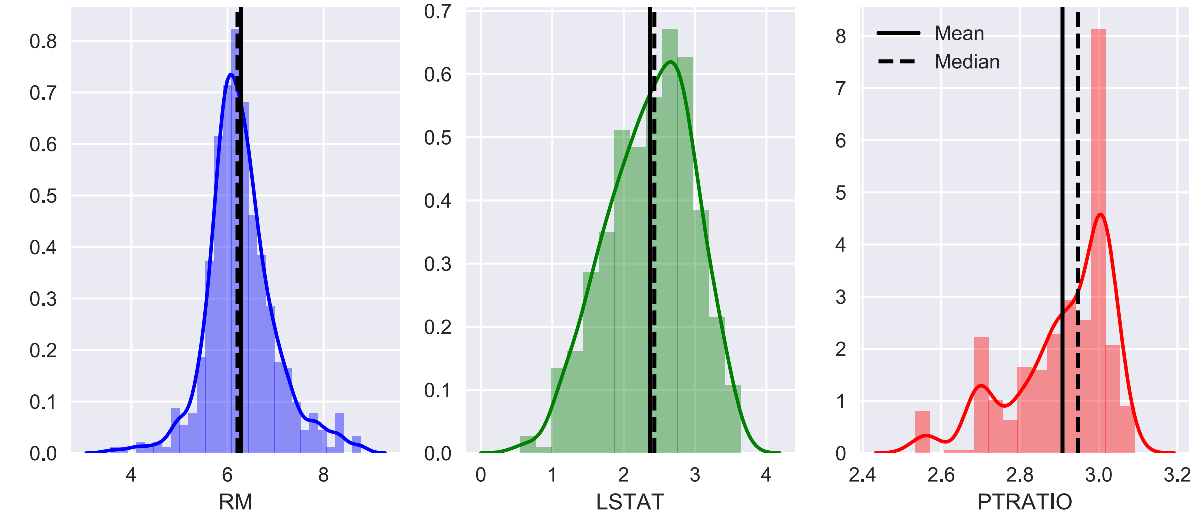
The data is loaded and summarized [data.describe()], a broad overview of the variables is in share/Data_describe.txt file. The data was quite clean and well organized, so I didn't wrote a code to process data (I used my terminal: python commands).

The correlations between the variables are possible sources of multicollinearity, and they affect variation in medv. So let's look at their distribution, and their correlation with medv but also with each other:

- Histograms for the data distributions
- The distribution of each feature versus house-price
- Attributes correlation

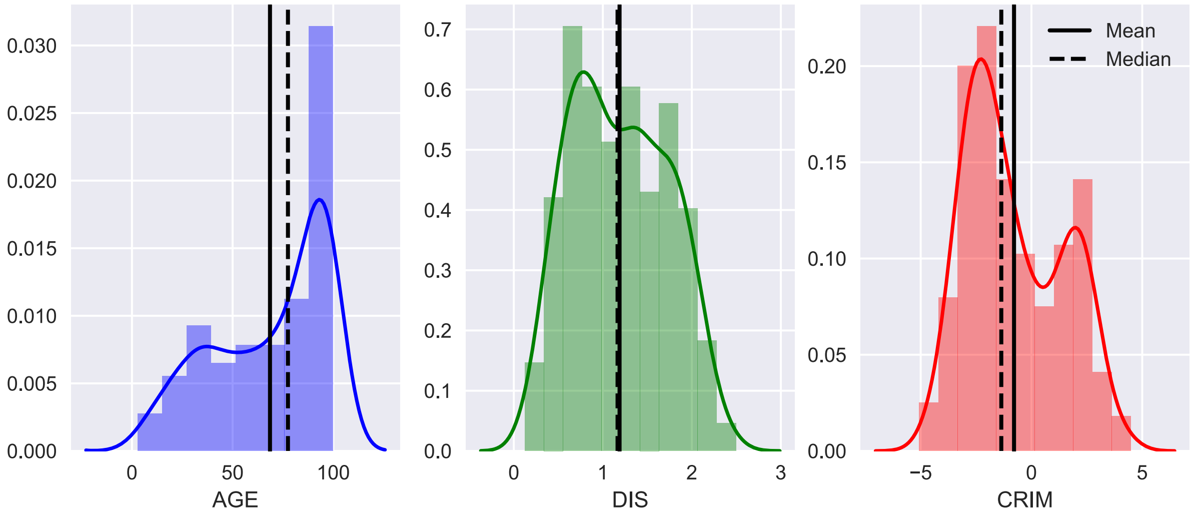
Analyze data: data distributions

- **rm – average number of rooms per dwelling**
- **lstat – lower status of the population (in %)**
- **ptratio – pupil-teacher ratio by town**



- **age – proportion of owner-occupied units built prior to 1940.**
- **dis – weighted mean of distances to five Boston employment centers.**
- **crim – per capita crime rate by town.**

- **log(y_axis) has been used**



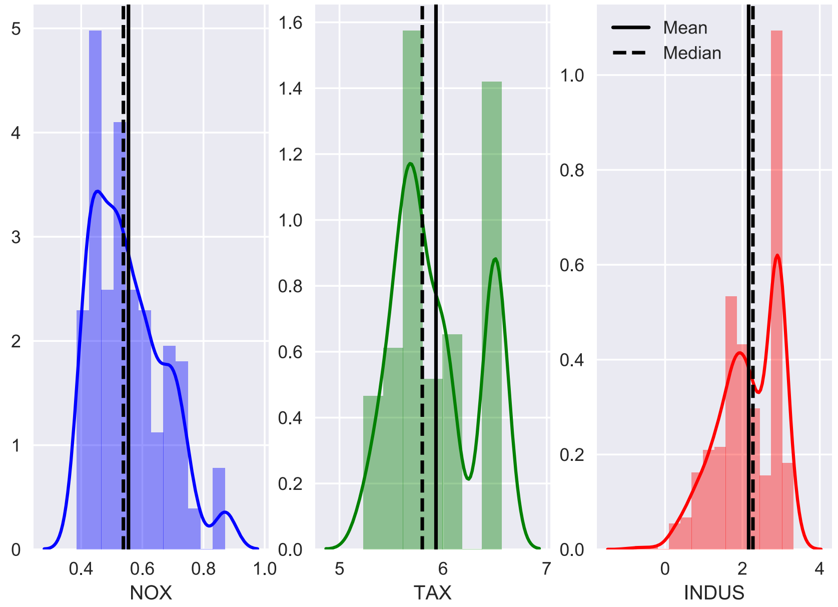
Analyze data

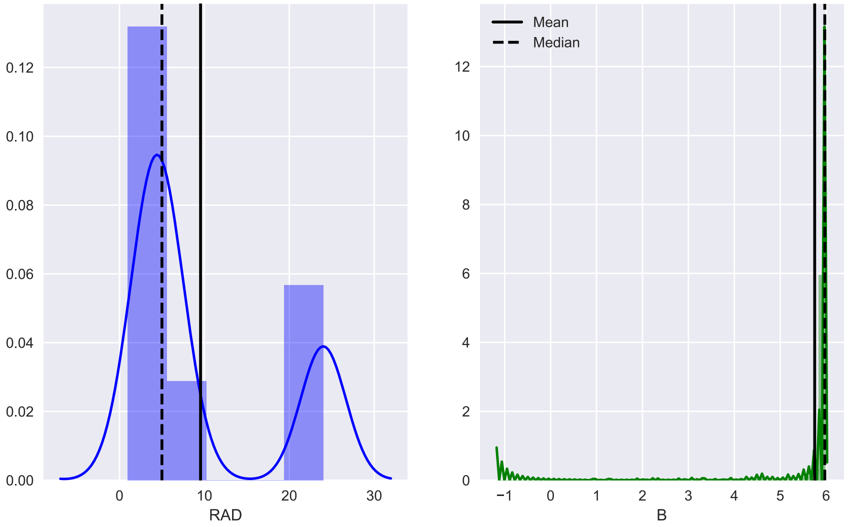
data distributions

- nox – nitrogen oxides concentration (parts per million).
- tax – full-value property-tax rate per $10,000
- indus – proportion of non-retain business acres per town.

- rad – index of accessibility to radial highways
- black - 1000(Bk – 0.63)^2, where Bk is the proportion of blacks by town.

- log(y_axis) has been used



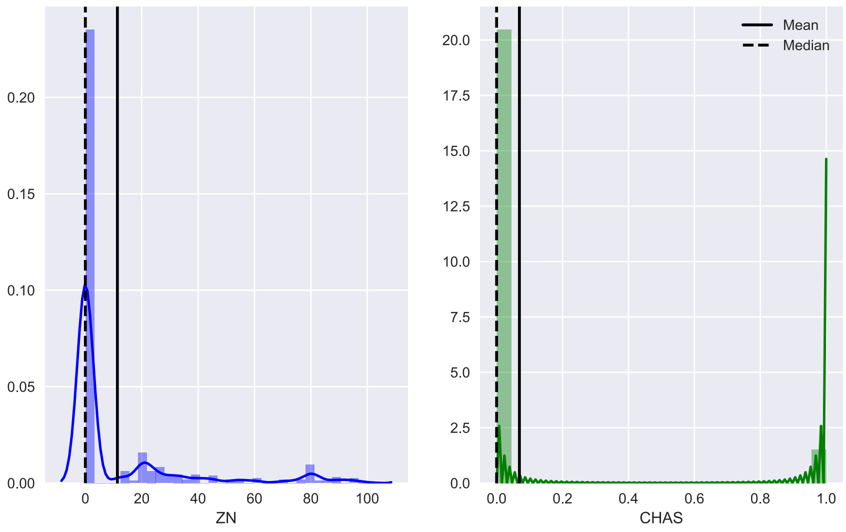


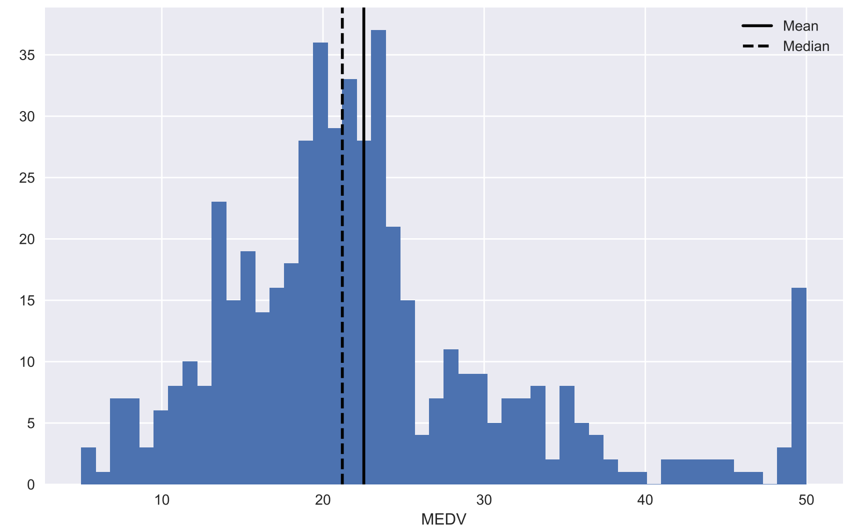
Analyze data

data distributions

- **zn – proportion of residential land zoned for lots over 25,000 sq. ft.**
- **chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)**

- **medv – median value of owner-occupied homes in $1000s.**





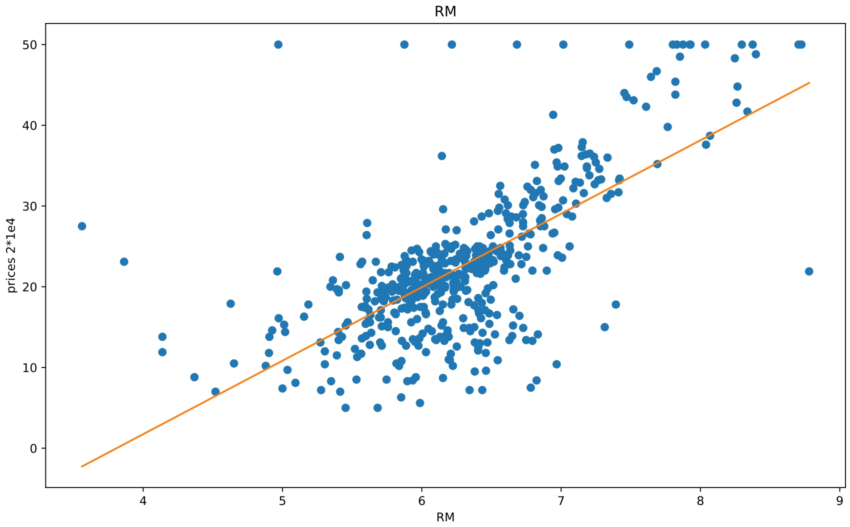
Analyze data

feature distribution versus medv

- **x_axis: feature (indep. variable)**

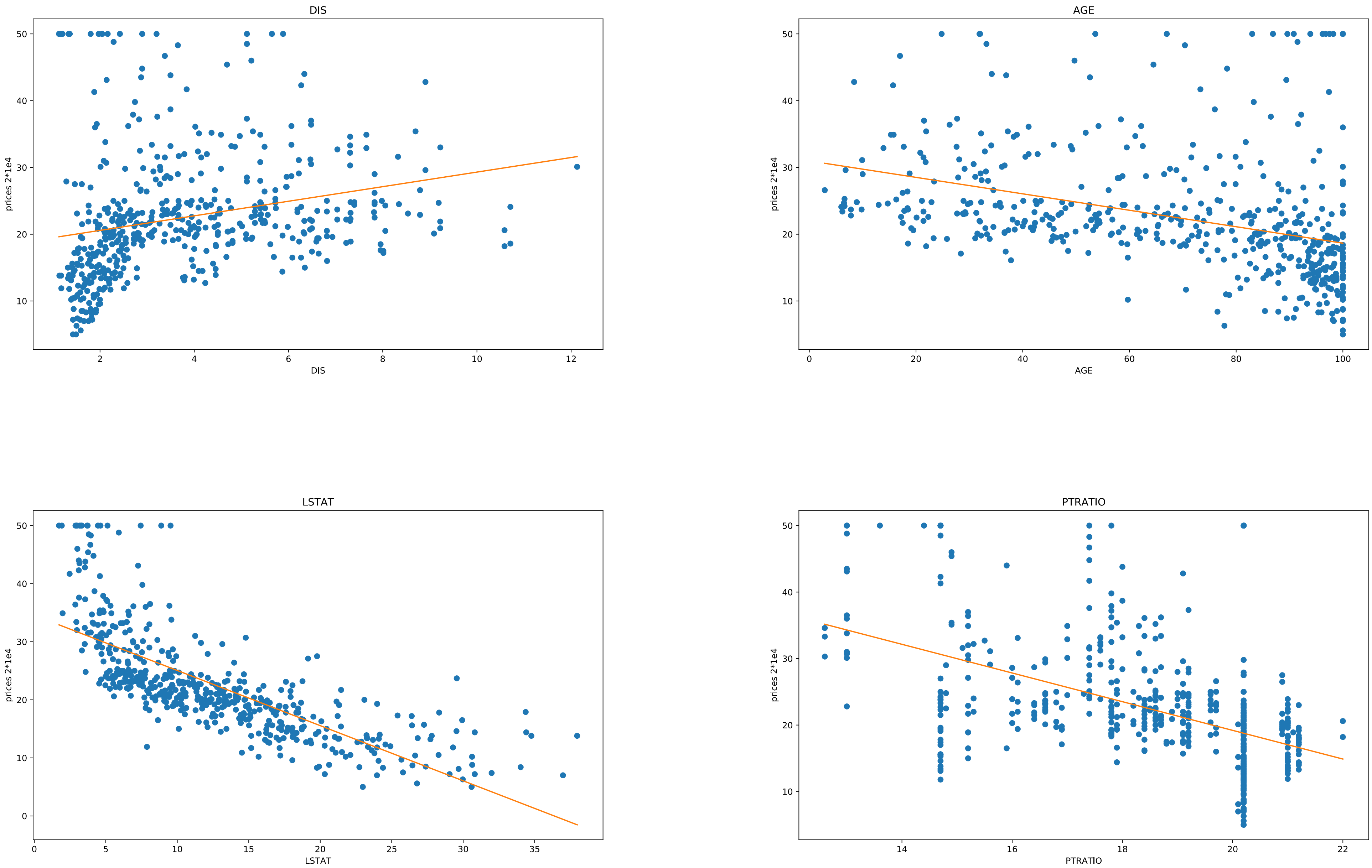
- **y_axis: medv (price) * 1e3 (unlike seen in the figures)**

- **red line is the regression fit: `y ~ x`**



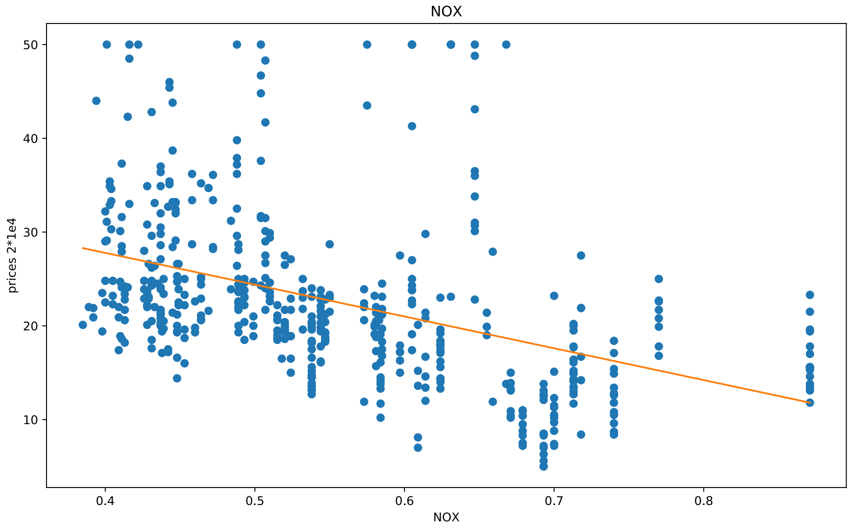
Analyze data

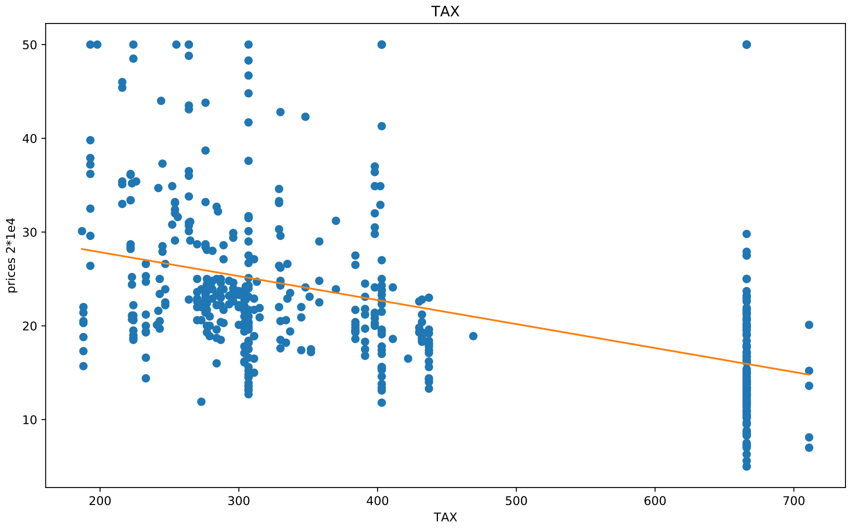
feature distribution versus medv

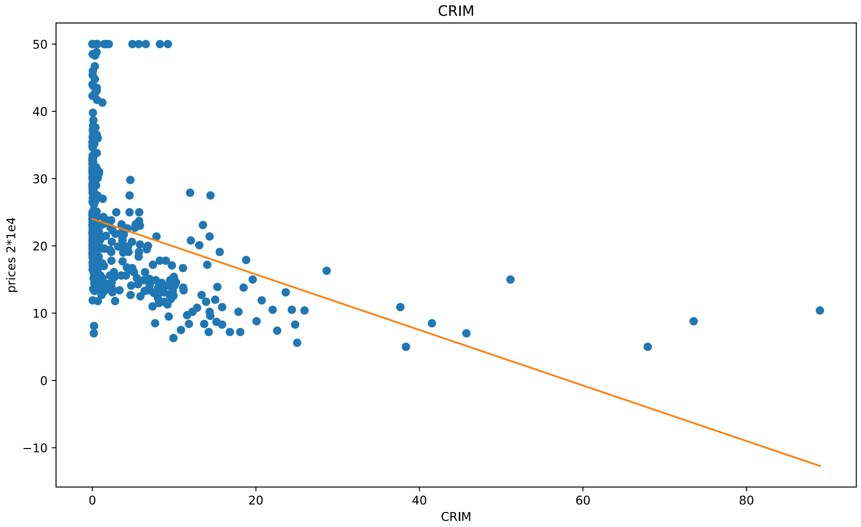


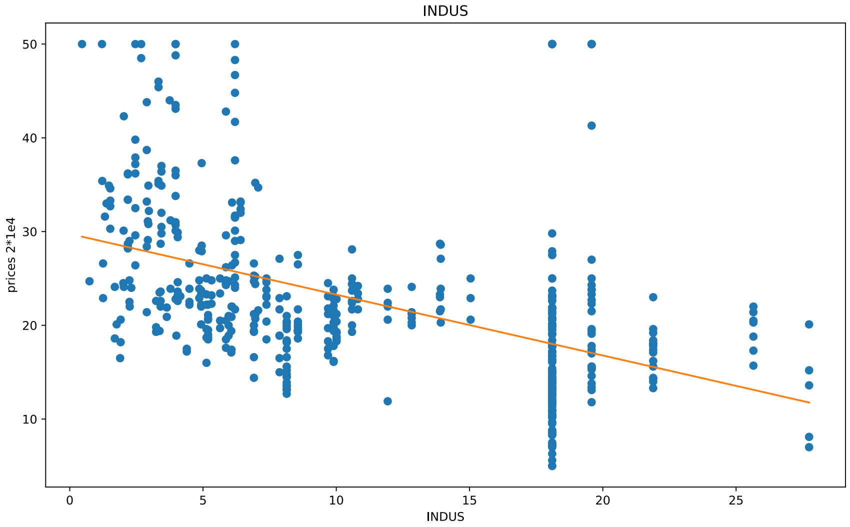
Analyze data

feature distribution versus medv



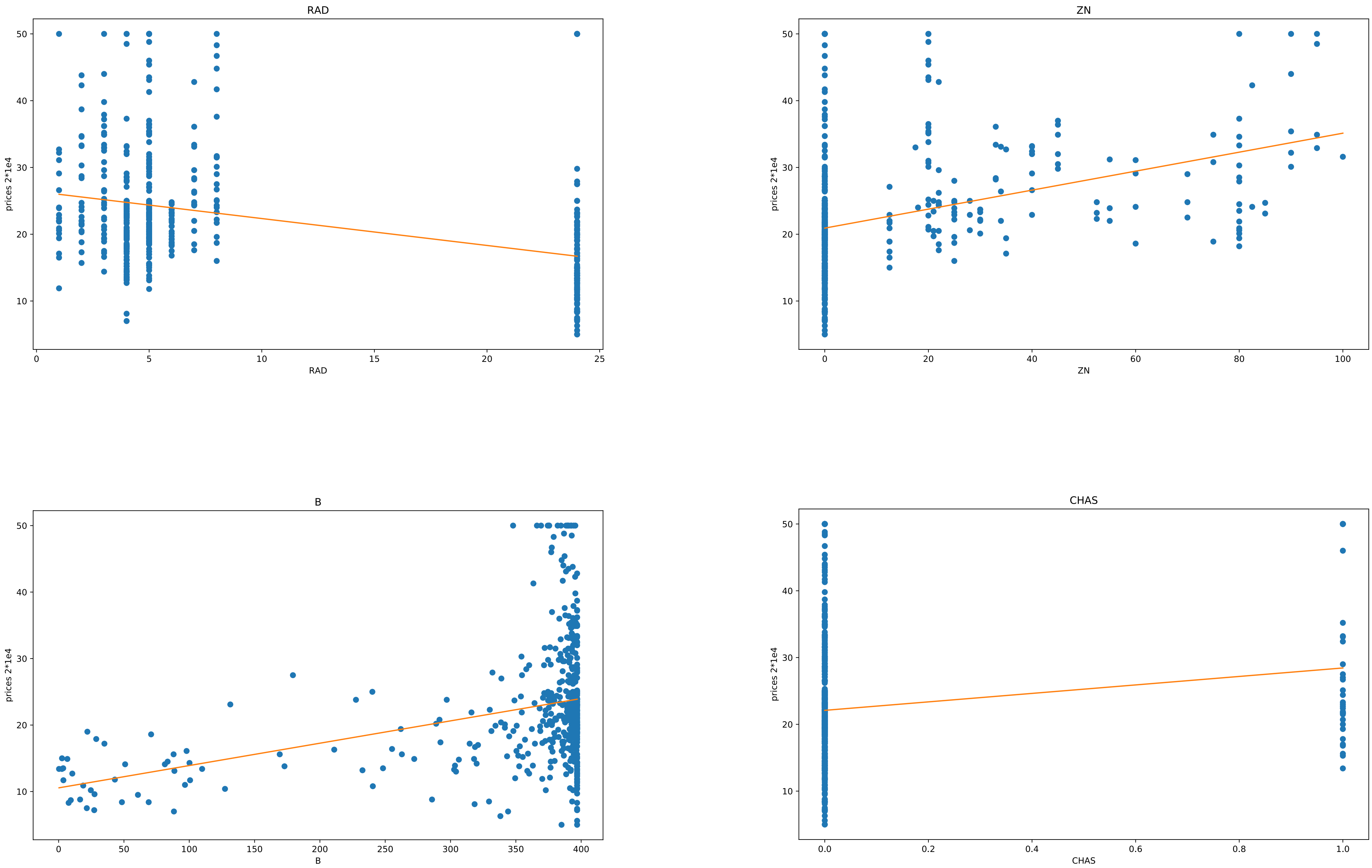






Analyze data

feature distribution versus medv



Analyze data

feature distribution versus medv

According to scatterplots, there is a good correlation between medv and RM, LSTAT.
There is no strongly marked correlation between mdev and the other features.

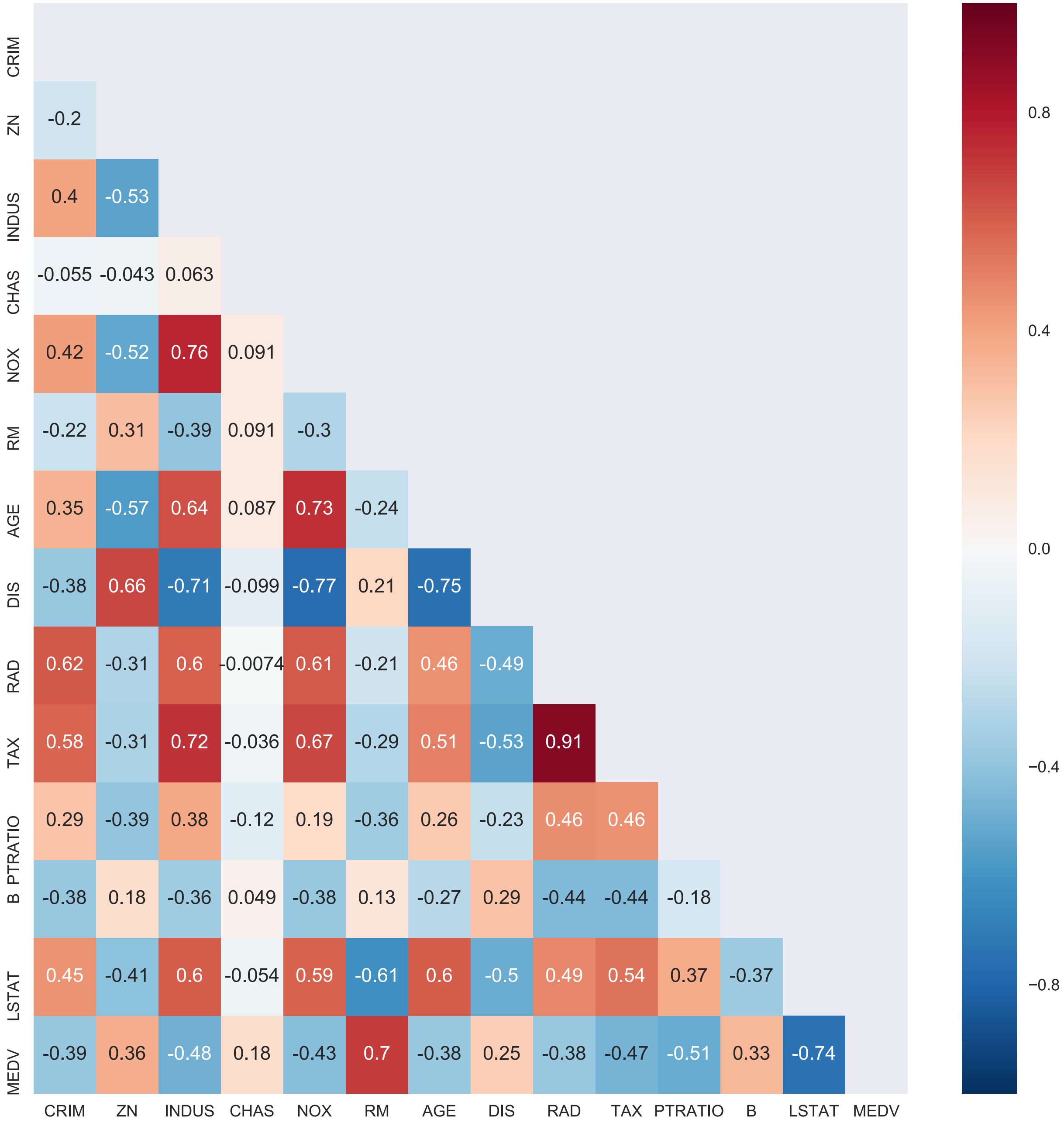
Analyze data

Attributes correlation

The correlation matrix was used for correlation calculation between features:

It can be seen that positive and negative correlations are in red (rad,tax) and blue(lstat,medv), respectively:

Charles River area, greater number of rooms and the presence of the black community are factors which increase housing price. The presence of lower class people, high crime rate, distance to the main employment centers and high teachers/student ratios decrease housing price. The last one is quite surprising…



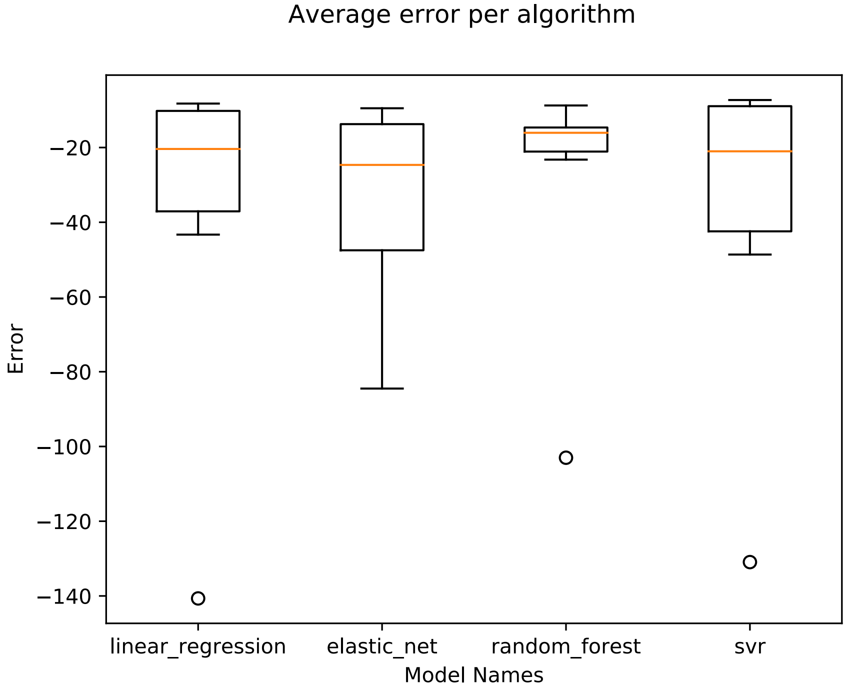
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM														
ZN	-0.2													
INDUS	0.4	-0.53												
CHAS	-0.055	-0.043	0.063											
NOX	0.42	-0.52	0.76	0.091										
RM	-0.22	0.31	-0.39	0.091	-0.3									
AGE	0.35	-0.57	0.64	0.087	0.73	-0.24								
DIS	-0.38	0.66	-0.71	-0.099	-0.77	0.21	-0.75							
RAD	0.62	-0.31	0.6	-0.0074	0.61	-0.21	0.46	-0.49						
TAX	0.58	-0.31	0.72	-0.036	0.67	-0.29	0.51	-0.53	0.91					
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46				
B	-0.38	0.18	-0.36	0.049	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18			
LSTAT	0.45	-0.41	0.6	-0.054	0.59	-0.61	0.6	-0.5	0.49	0.54	0.37	-0.37		
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.7	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	

Algorithms evaluation: CrossValidation.py

Now we need a model to fit the data. It would be wise to pick more than one model and evaluate each model estimator performance.

I used Cross validation (n_split = 10, scoring='neg_mean_squared_error', all original features) to evaluate estimator performance of 4 models:

Model	Average accuracy from CV =10	Error on average acc (+/-)
Linear Regression	-33.16	5.76
Support Vector Regression	-33.37	5.78
Random Forest	-25.13	5.01
Elastic Net	-33.83	5.82



From the above result, we can see that Random forest and Linear Regression have the least average error (& high average accuracy), indicating that both models have a good performance to evaluate Boston house-price. The numbers are in neg. because of the scoring option.

Algorithms evaluation: CrossValidation.py

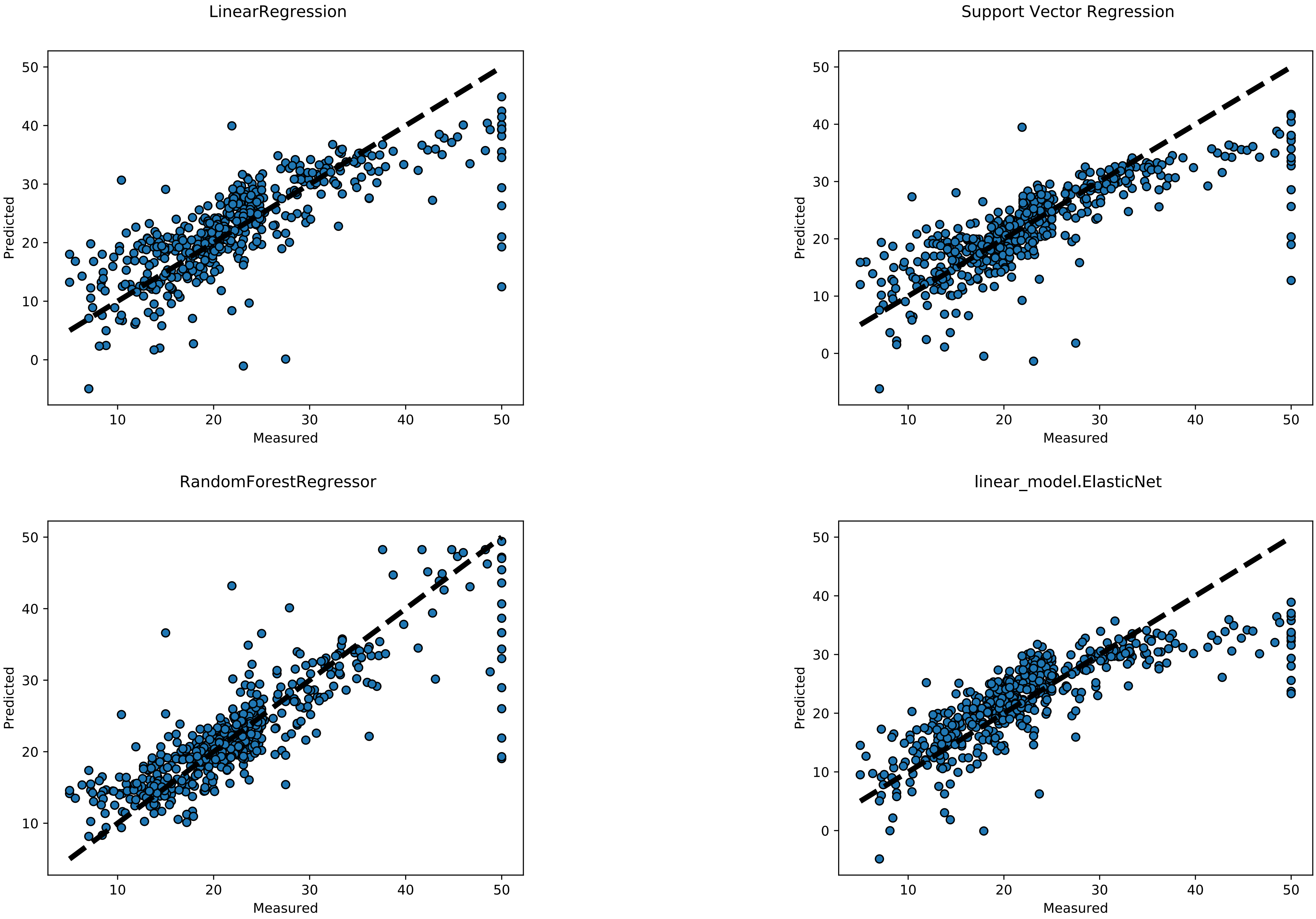
As an example, we looked at the medv predictions for houses with different number of rooms (rm). As we can see in the following table, the house price increased with the number of rooms when rm > 5. The house gets bigger so it gets more expensive. I expected the same correlation for the region of rm < 5…

number of rooms	medv predictions Linear Regression
1 to 5	10.84
6	19.94
7	29.04
8	38.15

We also visualized medv prediction versus medv prevision for each model. From the plots (the next slide), we can see that some models are more or less sensitive to predict a certain range of medv (house-price).

Ex: SVR model doesn't predict very well medv greater than ~35, but this model is very sensitive for medv predictions in [15 , 35]. This is why, it's better to do the predictions with more than one model. We already can say that, a model such as Random Forest and Neural Networks would be more appropriate when it comes to make prediction with medv multiple control regions.

Algorithms evaluation: CrossValidation.py



Data fitting

Linear regression: multiple_linear_regression.py

We Split the dataset into the Training set and Test set: test_size = 0.2. The model is fitted to the training dataset.

We will drop the variables based on their AIC scores (the predictions are made with 95% CL p-value < 0.05, etc.). We also check that Adj R^2 and R^2 values are in [0,1]. No further variable can be removed when we start sacrificing accuracy. The detailed results are listed in LR_results.txt. The most important OLS Regression Results are listed in the following table:

medv ~	R^2	Adj. R^2
sum (all 13 variables)	0,954	0,952
crim + nox + age + chas + rm + dis + ptratio + b +lstat	0,955	0,954
crim + chas + rm + dis + ptratio + b +lstat	0,957	0,957
chas + rm + dis + ptratio + b +lstat	0,957	0,956

No further variable can be removed when we start sacrificing accuracy, e.g:

medv ~ 16.71 - 0.06 crim + 2.83 chas + 5.80 rm - 0.46 dis - 0.60 ptratio + 0.01 b - 0.48 lstat

#The first 5 predictions for medv: [31.34 25.51 31.85 30.23 29.52]

Residual: R^2 = 0.957, Adj. R-squared = 0.957

Data fitting

Linear regression: multiple_linear_regression.py

The graph showing fitted values with residuals is: share/True_verus_Predicted_Values.pdf

I should have done it for each feature…

The next step would be to fit data using the other 3 models. It would be when I finish neural network.

Model applicability:

Once we came this far, we need to discuss whether the constructed model should or should not be used in a real-world setting. This is very important for the business side of the job. To do so, there are few questions to answering:

1) How data from 1978 can be relevant to today? It's about the size (Big Data) too!

2) How much the features in 78's data are sufficient to describe a today home? Which are the additional features (climate change, public transport access, …)?

3) How much the model will describe today home prices? (demographics, pb has changes since 78's)

4) Use Big Data to try the model for other type of regions

5) Deep learning is changing the way we do business thanks to more accurate insights.