Question1

- **Describe and summarize the data in terms of number of data points, dimensions, target, etc**

  In this dataset, the number of observations is 506, the number of features is 13, its dimension is 506*14, the target is the average price of houses in the area. The datatype of this dataset is float.

- **Visualization: present a single grid containing plots for each feature against the target. Choose the appropriate axis for dependent vs. independent variables.**

  The figure is attached on the second to last page of Question 1.

- **Tabulate each feature along with its associated weight and present them in a table. Explain what the sign of the weight means in the third column ('INDUS') of this table. Does the sign match what you expected? Why?**

  The parameters of linear model are shown below:

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
-0.1097	0.0415	0.0109	1.9357	-17.8668	3.2808	0.0045
DIS	RAD	TAX	PTRATIO	B	LSTAT	
-1.3856	0.3668	-0.0154	-0.8941	0.0089	-0.5536	

  INDUS indicates the proportion of non-retail business acres per town. The sign of the weight of INDUS means the proportion of non-retail business acres per town has a positive influence on the target value which in our case is the average price of the house. That means more non-retail business in the area will slight increase the house price. The sign matches what I expect, because in common sense, more non-retail business will bring more people who work in the area. The demand of houses would increase, for which the price of houses would increase.

- **Test the fitted model on your test set**

  The figure of prediction is on the last page of Question 1.

- **Suggest and calculate two more error measurement metrics; justify your choice**

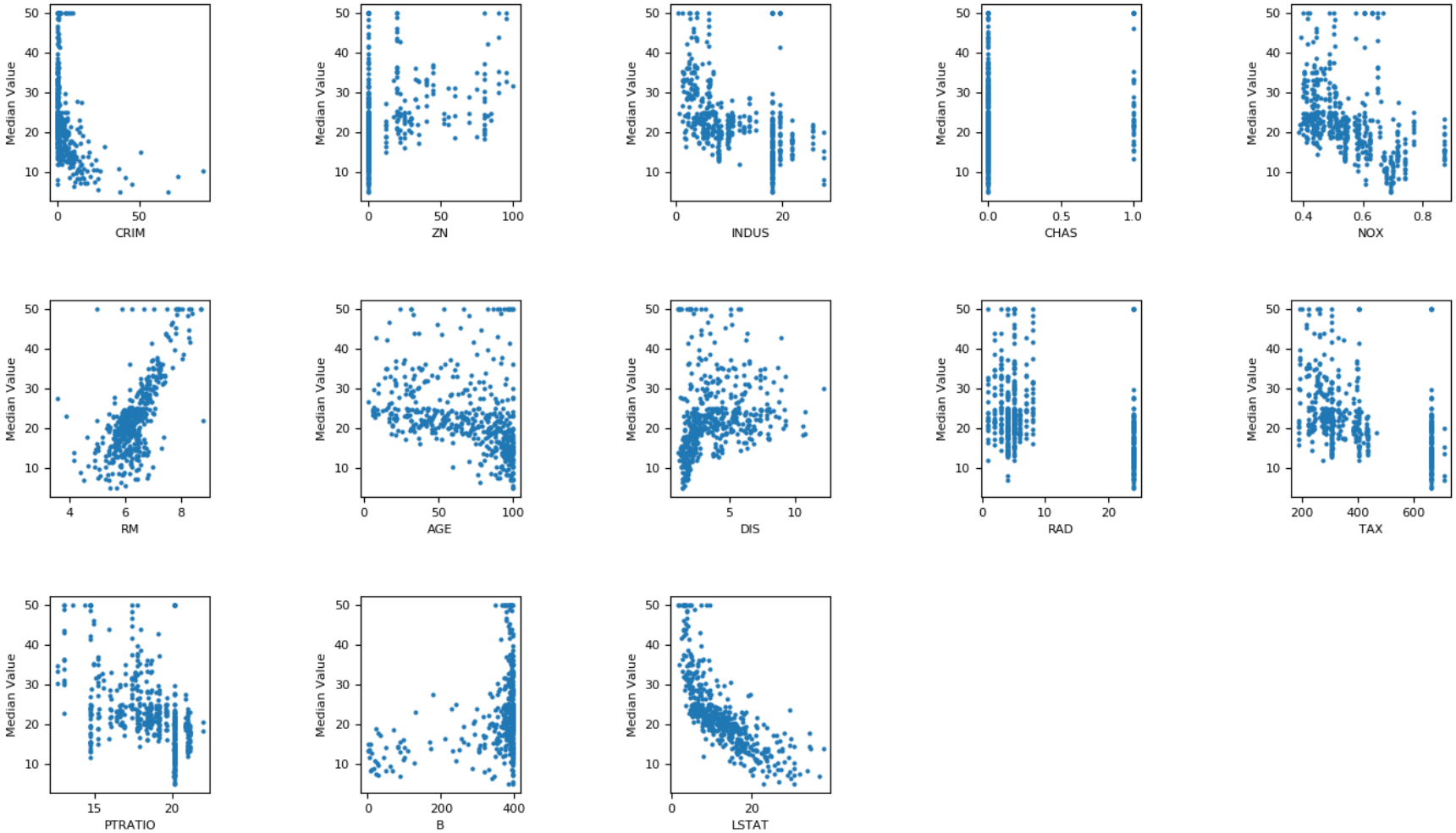
  I choose Mean Absolute Error and R2 Score. The results are blow:

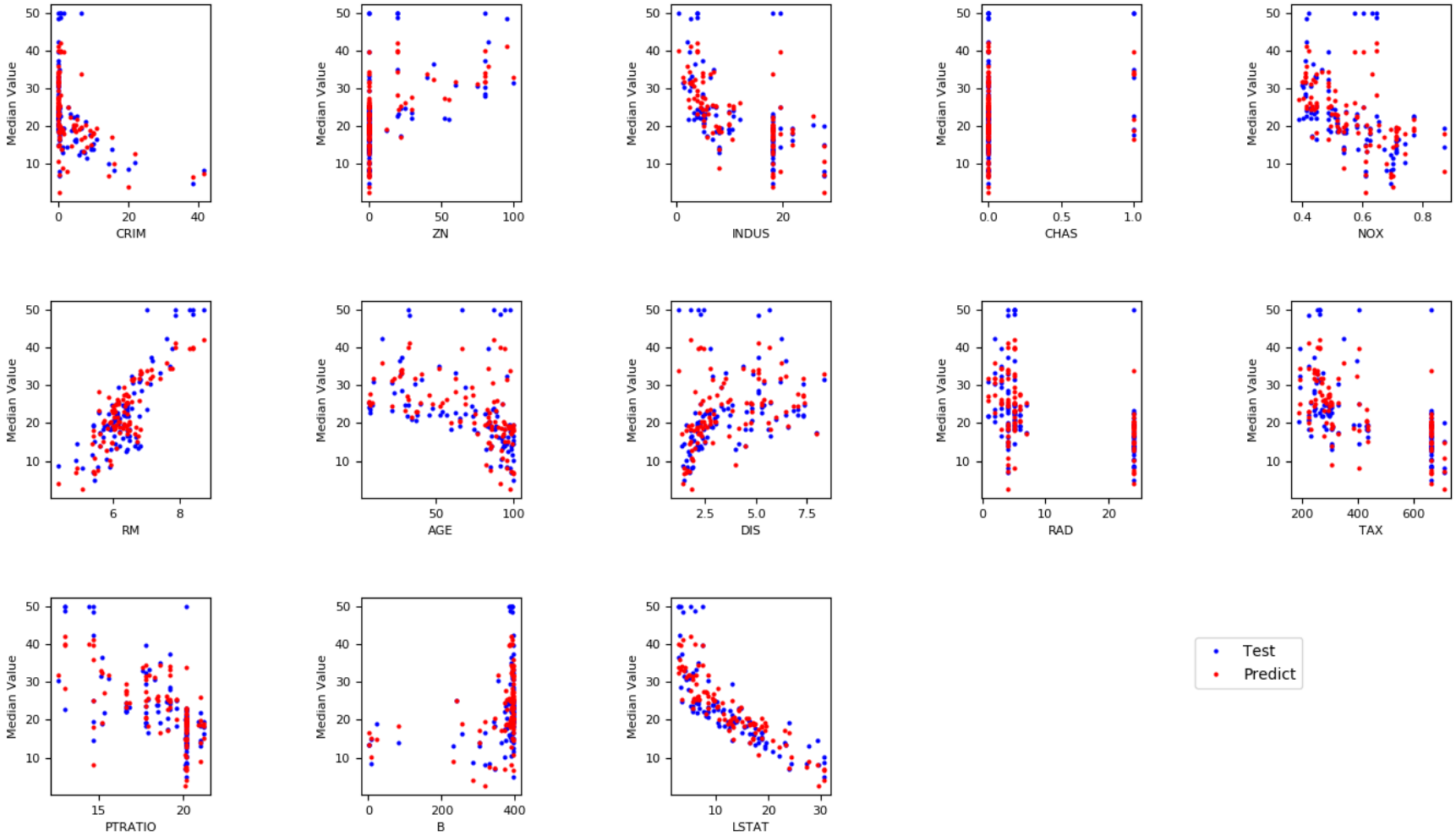
  RMS = 4.07
  MAE = 3.03
  R2 = 0.84

  R2 score and RMS are the most important two metrics in describing linear regression model. R2 Score Metric computes the coefficient of determination it's the proportion of the variance in the dependent variable that is predicted from the independent variable, which will give us information about the good of fit of a model, in another words, R2 measure

how well the regression line approximate the real data points. R2 = 0.84 means the model I built is good to predict on the test dataset. RMSE is the distance on average of data point from the fitted line measured along the vertical line.

- **Feature Selection: Based on your results, what are the most significant features that best predict the price? Justify your answer.**
  As the parameter of each feature showed on the top of this page, the feature NOX, CHAS, RM, DIS are most important features. The reason is their parameters have a large absolute value, for example, the parameter of NOX is -17.87. Thus, when there is a little change of the value of NOX, it would produce much influence on the target value.


Median Value
CRIM
ZN
INDUS
CHAS
NOX
RM
AGE
DIS
RAD
TAX
PTRATIO
B
LSTAT


Median Value
CRIM
ZN
INDUS
CHAS
NOX
RM
AGE
DIS
RAD
TAX
PTRATIO
B
LSTAT
Test
Predict

Question2

- **Mathematical Derivation**

$$1.\ \frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - w^T x^{(i)}\right)^2 + \frac{\lambda}{2}\|w\|^2$$

$$= \frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - x^{(i)T} w\right)^2 + \frac{\lambda}{2}\|w\|^2$$

$$= \frac{1}{2} A\,(y - X^T w)(y - X^T w) + \frac{\lambda}{2}\|w\|^2$$

$$= \frac{1}{2}(y - X^T w)^T A (y - X^T w) + \frac{\lambda}{2}\|w\|^2$$

$$\frac{\partial L}{\partial w} = \frac{1}{2}\frac{\partial}{\partial w}\left[(y - X^T w)^T A (y - X^T w) + \lambda\|w\|^2\right]$$

$$= \frac{1}{2}\frac{\partial}{\partial w}\left\{\left[y^T - (X^T w)^T\right] A (y - X^T w) + \lambda\|w\|^2\right\}$$

$$= \frac{1}{2}\frac{\partial}{\partial w}\left\{\left[y^T - w^T X\right] A (y - X^T w) + \lambda\|w\|^2\right\}$$

$$= \frac{1}{2}\frac{\partial}{\partial w}\left\{y^T A y - y^T A X^T w - w^T X A y + w^T X A X^T w + \lambda\|w\|^2\right\}$$

$$= \frac{1}{2}\frac{\partial}{\partial w}\left\{y^T A y - 2 w^T X A y + w^T X A^T X^T w + \lambda\|w\|^2\right\}$$

$$= \frac{1}{2}\left(-2XAy + 2(XA^T X^T)^T w + 2\lambda w\right)$$

$$= -XAy + XA^T X^T w + \lambda w$$

$$= -XAy + (X^T A X + \lambda I) w$$

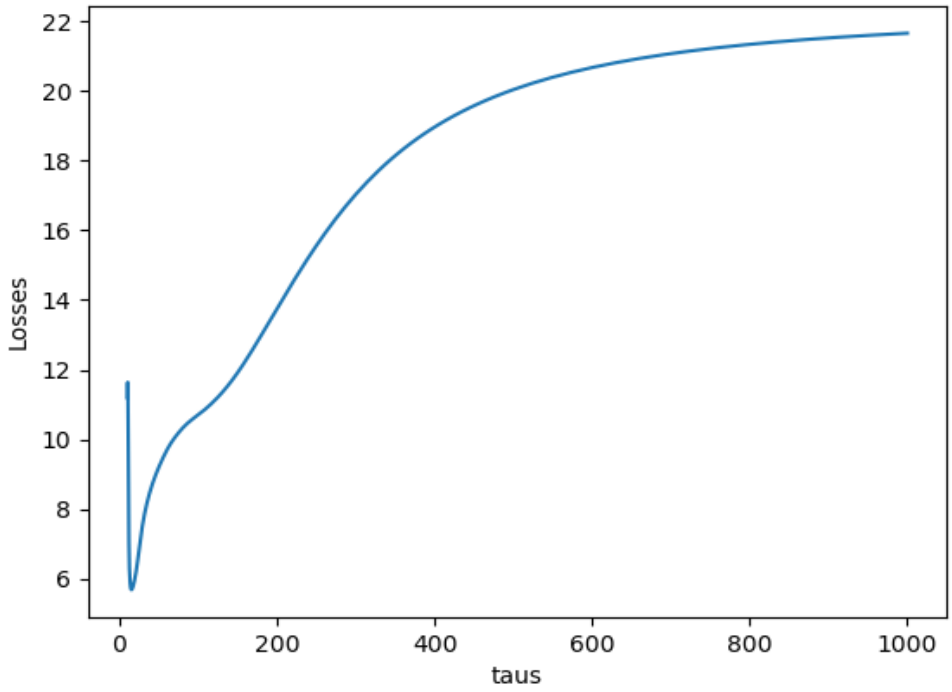
$$\frac{\partial L}{\partial w} = 0 \longrightarrow -XAy + (X^T A X + \lambda I) w = 0$$

$$\Rightarrow w = (X^T A X + \lambda I)^{-1} X^T A y$$

The code of this question is in the file q2.py.

- **Use k-fold cross-validation to compute the average loss for different values of τ in the range [10,1000] when performing regression on the Boston Houses dataset. Plot these loss values for each choice of τ.**

Below is the figure of the relationship between taus and losses I got.



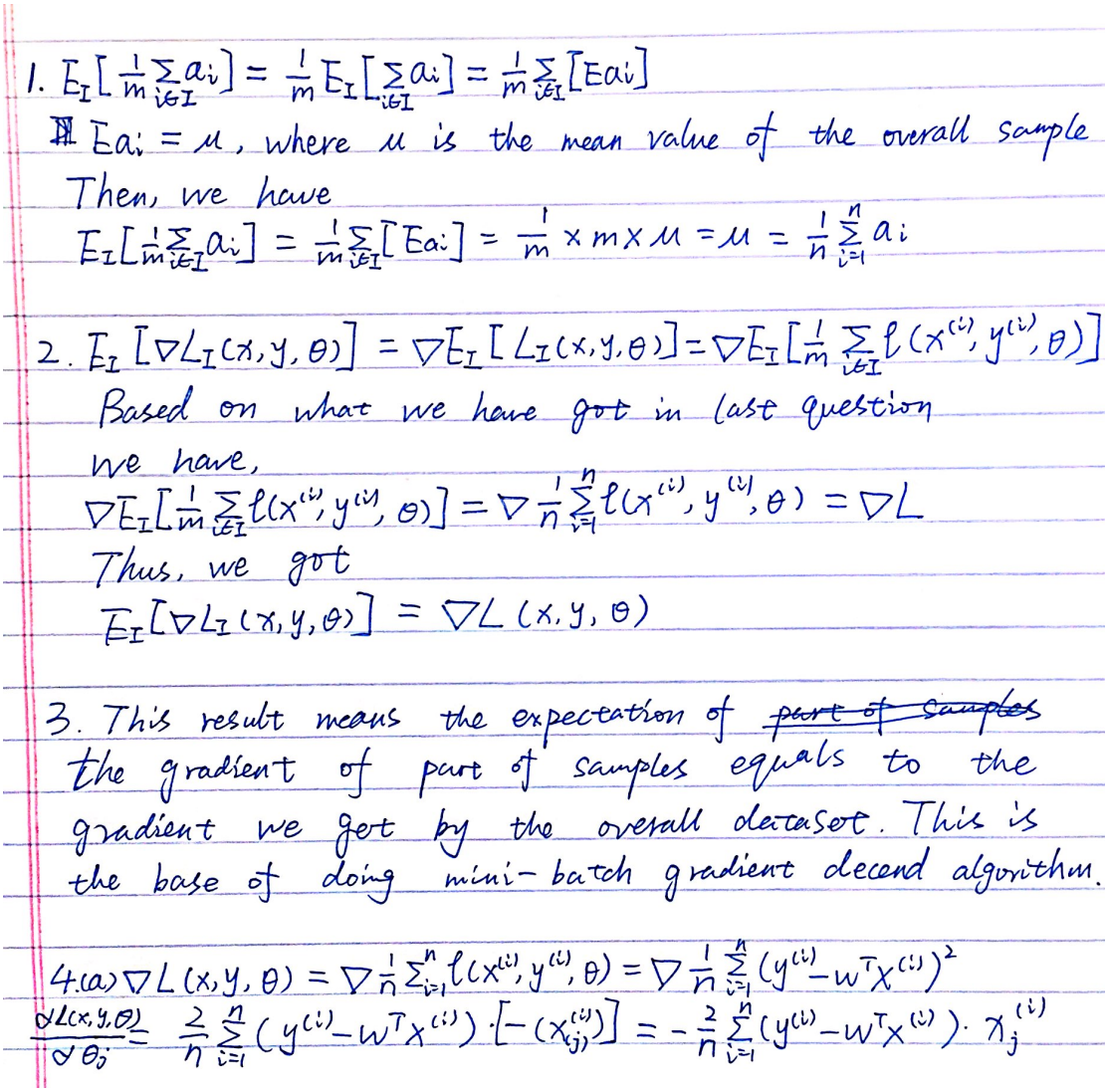
The minimum loss is 5.69

- **How does this algorithm behave when $\tau \to \infty$? When $\tau \to 0$?**

When tau is close to 0, the algorithm will cause overfitting problem, the loss is high. When tau is close to infinite, the algorithm will cause underfitting problem, every weight parameter will be equal to each other, the model will be like linear regression model.

Question3

- **Mathematical Derivation**



The code is in the file q3.py

- **Randomly initialize the weight parameters for your model from a N (0, I) distribution. Compare the value you have computed to the true gradient, ∇L, using both the squared distance metric and cosine similarity. Which is a more meaningful measure in this case and why?**

The result I got of the squared distance metric is 2717494.27, and the cosine similarity is 0.99997. In our case, the cosine similarity is more meaningful, because we do not care about the value of the gradient in light that the learning rate is set by ourselves, we only care about the direction of the gradient. Like in my result, the cosine similarity is equal to one, which means the two gradient vectors I got are parallel to each other, so that the directions of these two gradients are the same. Although these two gradients' vectors are so far to each other in the space, once they have the same direction, they will have the same result when doing gradient descent.

- **Plot the figure of the relationship between log(m) and log(variance)**