Homework #1

Due on October 26, 2019 at 11:55pm

Professor Hoda Mohammadzade

Amirhossein Yousefi

97206984

Problem 6

Section a:

In this part we use linear regression in order to predict ozone layer density.Implementation of linear regression is done by sklearn library and also we used pandas so that better understanding of data characetristics.Lets fist visualizing data for better illustration(figure1).

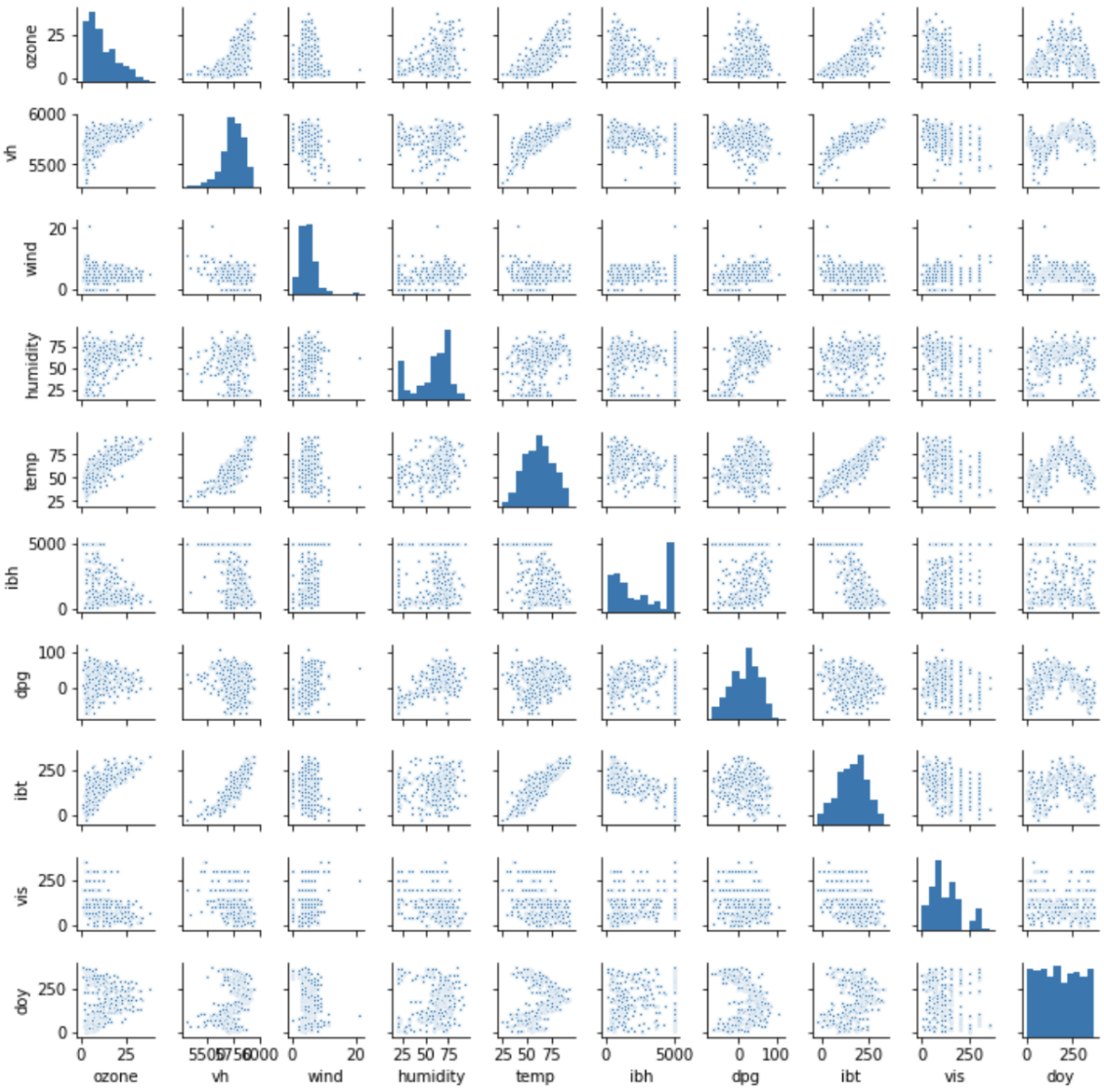


Figure 1: Ozone dataset

For better understanding relation between feature we also build correlation matrix which is in figure2.

Figure 2: Correlation matrix

At the end results for linear regression are in figure3.

	coefficents	corresponding feature
0	-0.087129	vh
1	-0.025921	wind
2	0.179676	humidity
3	0.499631	temp
4	-0.147053	ibh
5	0.005531	dpg
6	0.225382	ibt
7	-0.068683	vis
8	-0.118577	doy
0	-0.008743	intercept

(a) coefficients for linear regression

```
mean square error for train 0.28785063984112624
mean square error for test 0.33679679927074624
[-0.08712919 -0.02592093  0.17967609  0.49963127 -0.14705347  0.00553081
  0.22538202 -0.06868307 -0.11857731]
intercept is -0.008742541724202272
```

(b) error and model evaluation metrics

Figure 3: Results

Section c:
PCA is a dimension reduction method that extracts vital features corresponding to their eigen value of covriance matrix of data.For more detail figure 4 is presented.

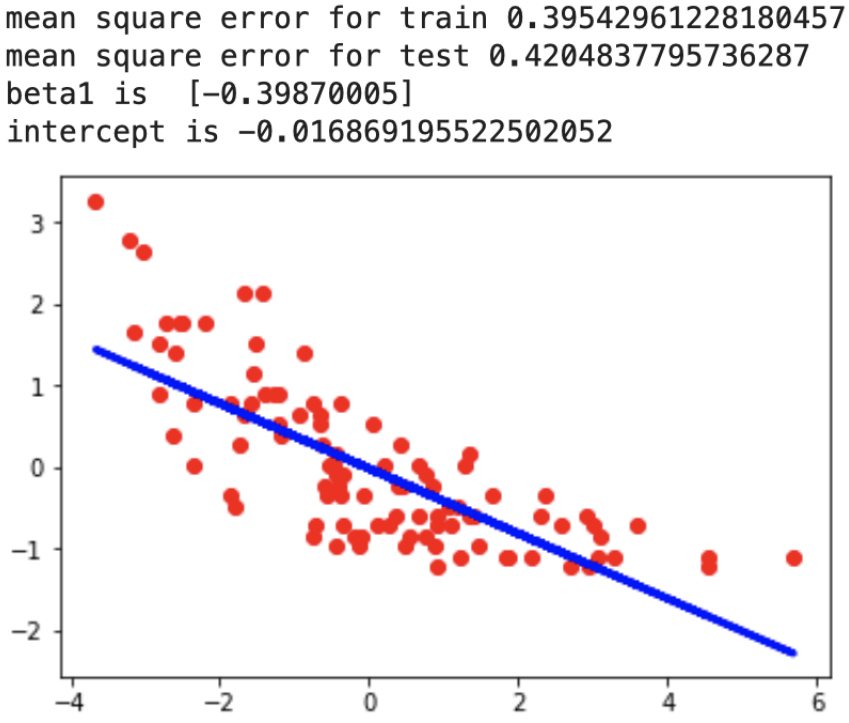
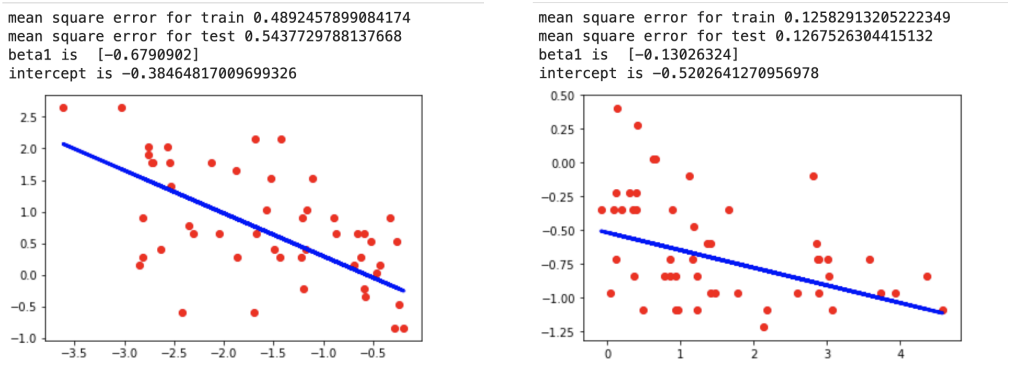


Figure 4: Ozone dataset

Section d:
At fist we have to split data as part d mentioned.Then two linear regression is done as figure5.



(a) half of the transformed data (b) another half of the transformed data

Figure 5: Results

As results shows mean squared error in (b) for half of the data is better than or less than original dataset.For all of them test error is more than train error.

Problem 7

Section a: At first data visualization is done in figure6.Then covariance matrix and correlation matrix in different types are shown in figure7.

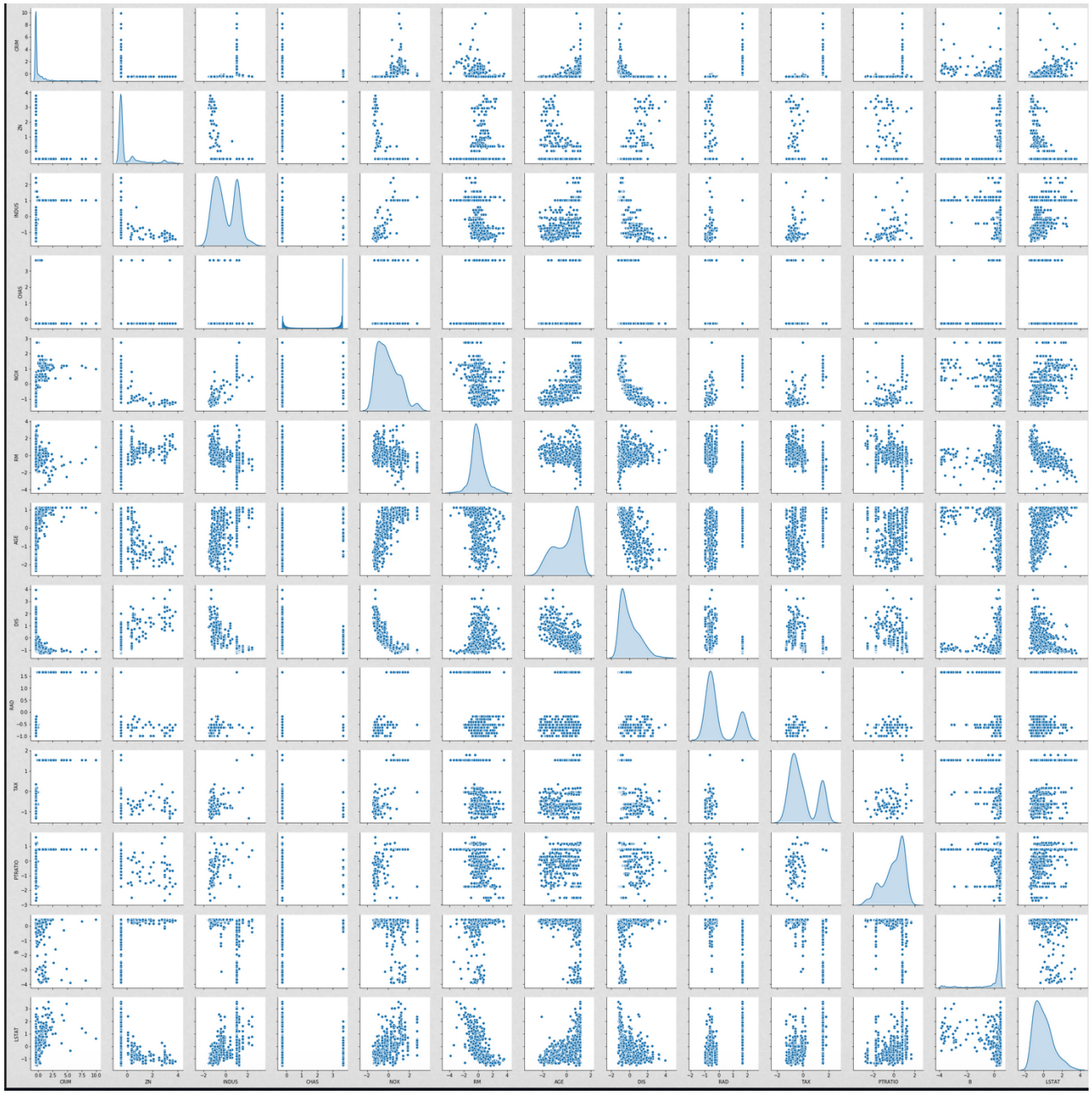
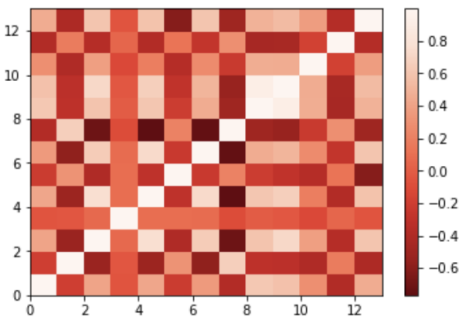


Figure 6: Data visualization

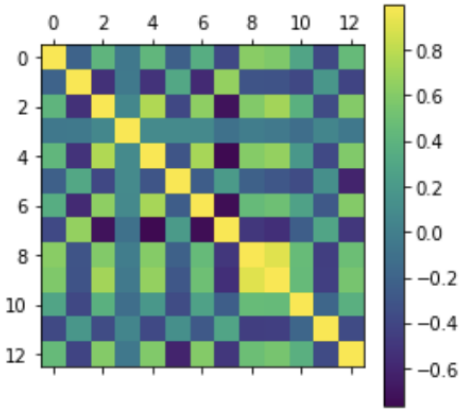
(a) Correlation matrix



(b) Correlation matrix

```
[[ 0.99802372 -0.20007304  0.40577989 -0.05578112  0.42013975 -0.21881341
   0.35203715 -0.37891975  0.62426897  0.5816126   0.28937256 -0.38430295
   0.45472104]
 [-0.20007304  0.99802372 -0.53277319 -0.04261234 -0.51558275  0.31137401
  -0.56841177  0.66309516 -0.31133133 -0.31394166 -0.39090448  0.17517344
  -0.41217838]
 [ 0.40577989 -0.53277319  0.99802372  0.06281364  0.76214225 -0.39090179
   0.64350425 -0.70662773  0.59395313  0.71933575  0.38249015 -0.35627105
   0.60260644]
 [-0.05578112 -0.04261234  0.06281364  0.99802372  0.09102256  0.09107089
   0.08634679 -0.09897978 -0.00735368 -0.03551619 -0.12127503  0.04869207
  -0.05382272]
 [ 0.42013975 -0.51558275  0.76214225  0.09102256  0.99802372 -0.30159098
   0.73002451 -0.7677099   0.61023218  0.666703    0.18855929 -0.37929955
   0.58971118]
 [-0.21881341  0.31137401 -0.39090179  0.09107089 -0.30159098  0.99802372
  -0.2397901   0.20484059 -0.20943195 -0.29147066 -0.35479892  0.12781554
  -0.61259521]
 [ 0.35203715 -0.56841177  0.64350425  0.08634679  0.73002451 -0.2397901
   0.99802372 -0.74640252  0.45512122  0.50545469  0.26099818 -0.2729934
   0.60114814]
 [-0.37891975  0.66309516 -0.70662773 -0.09897978 -0.7677099   0.20484059
  -0.74640252  0.99802372 -0.49361048 -0.5333754  -0.23201111  0.29093556
  -0.49601363]
 [ 0.62426897 -0.31133133  0.59395313 -0.00735368  0.61023218 -0.20943195
   0.45512122 -0.49361048  0.99802372  0.90842932  0.46382272 -0.44353453
   0.48771057]
 [ 0.5816126  -0.31394166  0.71933575 -0.03551619  0.666703   -0.29147066
   0.50545469 -0.5333754   0.90842932  0.99802372  0.45994226 -0.44093487
   0.54291833]
 [ 0.28937256 -0.39090448  0.38249015 -0.12127503  0.18855929 -0.35479892
   0.26099818 -0.23201111  0.46382272  0.45994226  0.99802372 -0.17703274
   0.3733051 ]
 [-0.38430295  0.17517344 -0.35627105  0.04869207 -0.37929955  0.12781554
  -0.2729934   0.29093556 -0.44353453 -0.44093487 -0.17703274  0.99802372
  -0.36536341]
 [ 0.45472104 -0.41217838  0.60260644 -0.05382272  0.58971118 -0.61259521
   0.60114814 -0.49601363  0.48771057  0.54291833  0.3733051  -0.36536341
   0.99802372]]
```

(c) Covariance matrix



(d) Correlation matrix

Figure 7: covariance and correlation matrix

Section b:

At the end coefficients and results are shown.

	coefficents	corresponding feature
0	-0.113455	CRIM
1	0.112760	ZN
2	0.008460	INDUS
3	0.069352	CHAS
4	-0.204503	NOX
5	0.294815	RM
6	-0.030561	AGE
7	-0.343492	DIS
8	0.229247	RAD
9	-0.202888	TAX
10	-0.239573	PTRATIO
11	0.067644	B
12	-0.377926	LSTAT
0	-0.008317	intercept

(a) coefficients

```
('regression score is', 0.7645451026942549)
('mean square error for train', 0.23594979171921113)
('mean square error for test', 0.3215157723496526)
('coefficients for linear regresion are', array([-0.11345494,  0.11276007,  0.00846006,  0.06935244, -0.20450349,
        0.29481534, -0.03056082, -0.34349212,  0.22924672, -0.20288752,
       -0.23957313,  0.06764365, -0.37792619]))
('intercept is', -0.008317428965594752)
```

(b) error and model evaluation metrics

Figure 8: Results

For more information about features, figure9 is shown which contains measure about importance of features which is p-value. which is p-value

Dep. Variable:	MEDV	**R-squared:**	0.765
Model:	OLS	**Adj. R-squared:**	0.756
Method:	Least Squares	**F-statistic:**	84.92
Date:	Sun, 20 Oct 2019	**Prob (F-statistic):**	2.76e-98
Time:	22:53:18	**Log-Likelihood:**	-246.69
No. Observations:	354	**AIC:**	521.4
Df Residuals:	340	**BIC:**	575.6
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0083	0.027	-0.313	0.754	-0.061	0.044
CRIM	-0.1135	0.036	-3.185	0.002	-0.184	-0.043
ZN	0.1128	0.040	2.834	0.005	0.035	0.191
INDUS	0.0085	0.051	0.166	0.868	-0.092	0.109
CHAS	0.0694	0.028	2.483	0.014	0.014	0.124
NOX	-0.2045	0.057	-3.618	0.000	-0.316	-0.093
RM	0.2948	0.037	7.867	0.000	0.221	0.369
AGE	-0.0306	0.048	-0.634	0.527	-0.125	0.064
DIS	-0.3435	0.054	-6.402	0.000	-0.449	-0.238
RAD	0.2292	0.073	3.152	0.002	0.086	0.372
TAX	-0.2029	0.078	-2.586	0.010	-0.357	-0.049
PTRATIO	-0.2396	0.035	-6.803	0.000	-0.309	-0.170
B	0.0676	0.032	2.099	0.037	0.004	0.131
LSTAT	-0.3779	0.047	-8.068	0.000	-0.470	-0.286

Omnibus:	133.612	**Durbin-Watson:**	2.019
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	634.086
Skew:	1.547	**Prob(JB):**	2.04e-138
Kurtosis:	8.781	**Cond. No.**	9.72

Figure 9: Data visualization