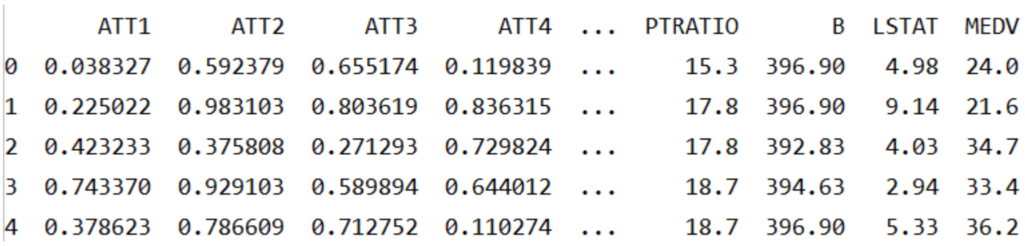
My Name (shixuef2)
IE598 MLF F19
Module 4 Homework (Regression)

Part 1 Exploratory Data Analysis

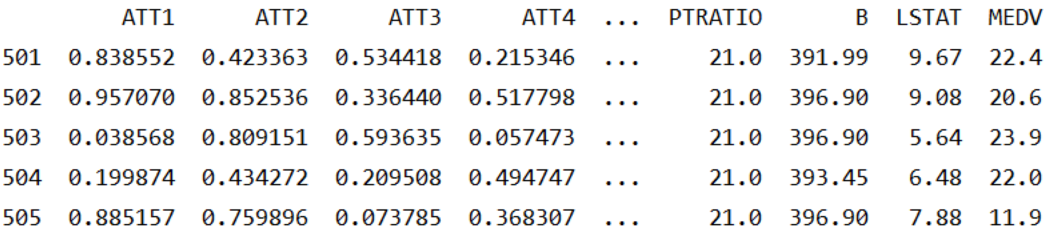
The Shape of the Data:
(452, 27)

The Head:



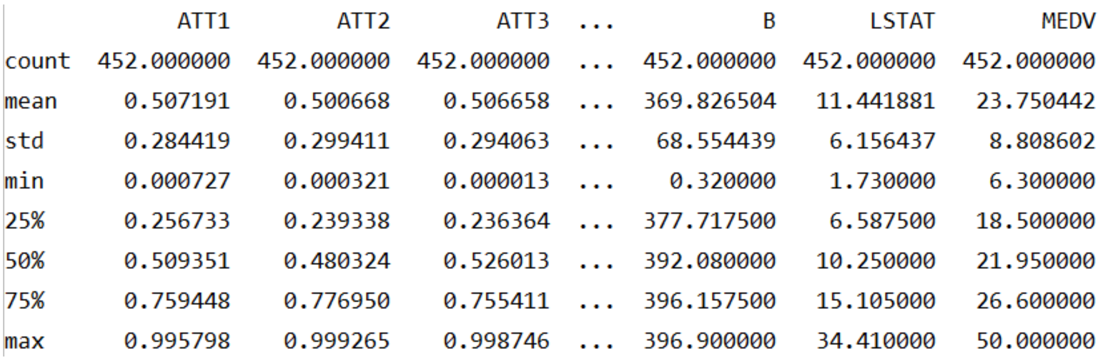
```
       ATT1      ATT2      ATT3      ATT4  ...  PTRATIO       B  LSTAT  MEDV
0  0.038327  0.592379  0.655174  0.119839  ...     15.3  396.90   4.98  24.0
1  0.225022  0.983103  0.803619  0.836315  ...     17.8  396.90   9.14  21.6
2  0.423233  0.375808  0.271293  0.729824  ...     17.8  392.83   4.03  34.7
3  0.743370  0.929103  0.589894  0.644012  ...     18.7  394.63   2.94  33.4
4  0.378623  0.786609  0.712752  0.110274  ...     18.7  396.90   5.33  36.2
```

The Tail:



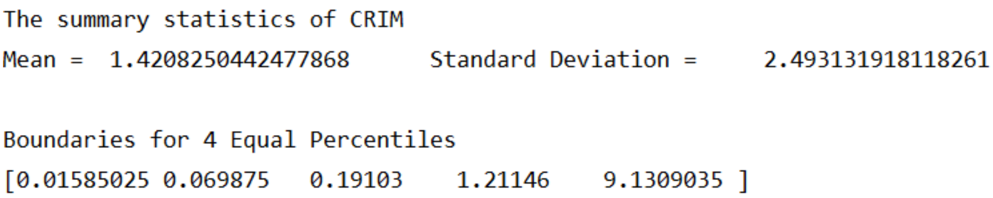
```
         ATT1      ATT2      ATT3      ATT4  ...  PTRATIO       B  LSTAT  MEDV
501  0.838552  0.423363  0.534418  0.215346  ...     21.0  391.99   9.67  22.4
502  0.957070  0.852536  0.336440  0.517798  ...     21.0  396.90   9.08  20.6
503  0.038568  0.809151  0.593635  0.057473  ...     21.0  396.90   5.64  23.9
504  0.199874  0.434272  0.209508  0.494747  ...     21.0  393.45   6.48  22.0
505  0.885157  0.759896  0.073785  0.368307  ...     21.0  396.90   7.88  11.9
```

The Summary:

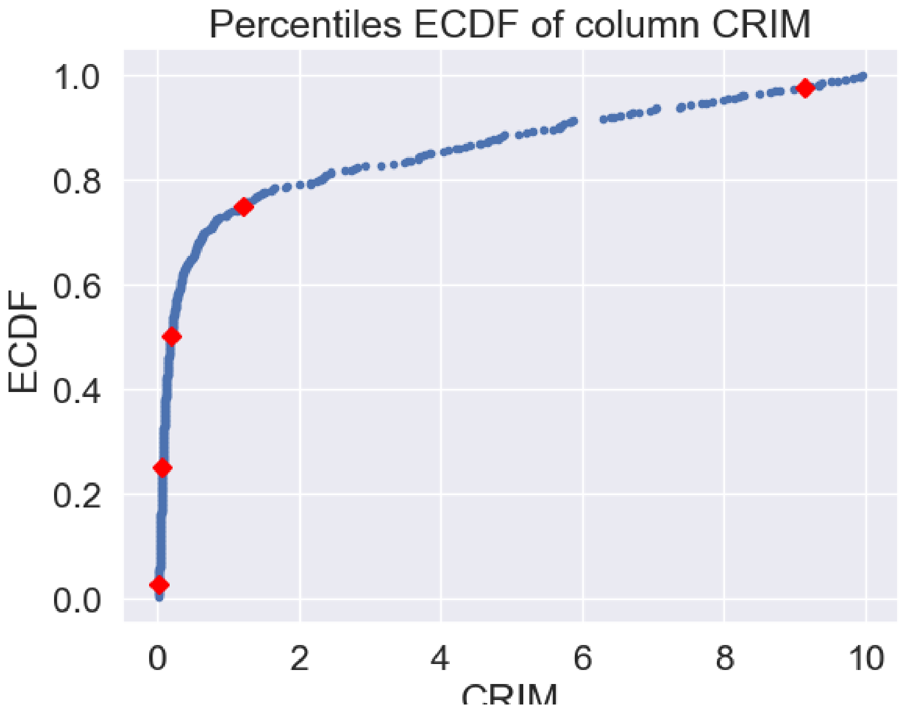


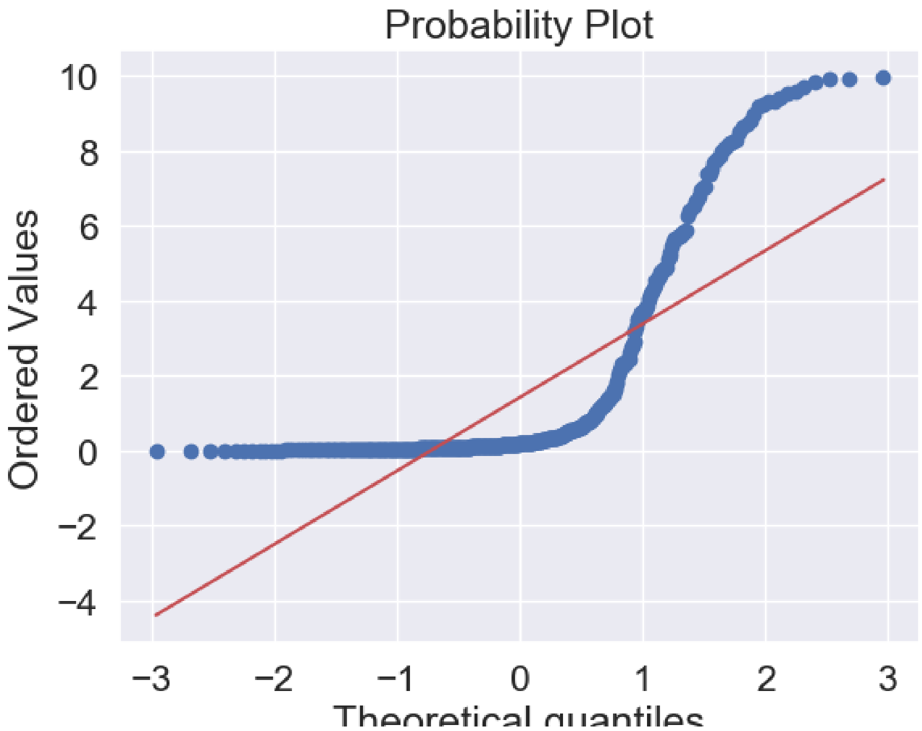
```
             ATT1        ATT2        ATT3  ...           B       LSTAT        MEDV
count  452.000000  452.000000  452.000000  ...  452.000000  452.000000  452.000000
mean     0.507191    0.500668    0.506658  ...  369.826504   11.441881   23.750442
std      0.284419    0.299411    0.294063  ...   68.554439    6.156437    8.808602
min      0.000727    0.000321    0.000013  ...    0.320000    1.730000    6.300000
25%      0.256733    0.239338    0.236364  ...  377.717500    6.587500   18.500000
50%      0.509351    0.480324    0.526013  ...  392.080000   10.250000   21.950000
75%      0.759448    0.776950    0.755411  ...  396.157500   15.105000   26.600000
max      0.995798    0.999265    0.998746  ...  396.900000   34.410000   50.000000
```

The Summary Statistics for each Feature/Target Column:
Feature 1: CRIM



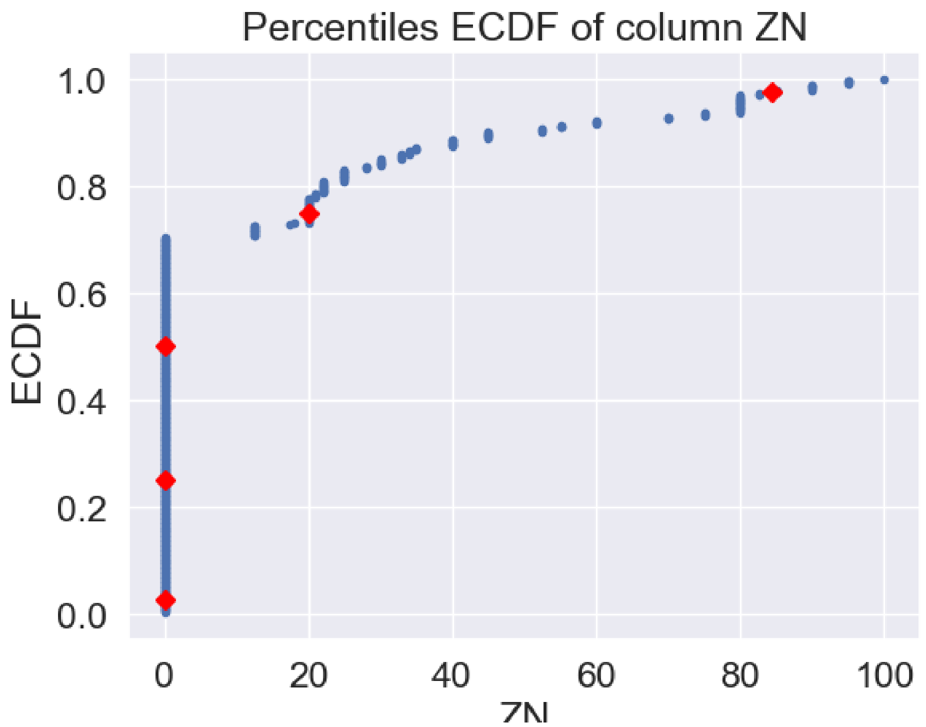
```
The summary statistics of CRIM
Mean =  1.4208250442477868      Standard Deviation =     2.493131918118261

Boundaries for 4 Equal Percentiles
[0.01585025 0.069875    0.19103     1.21146     9.1309035 ]
```

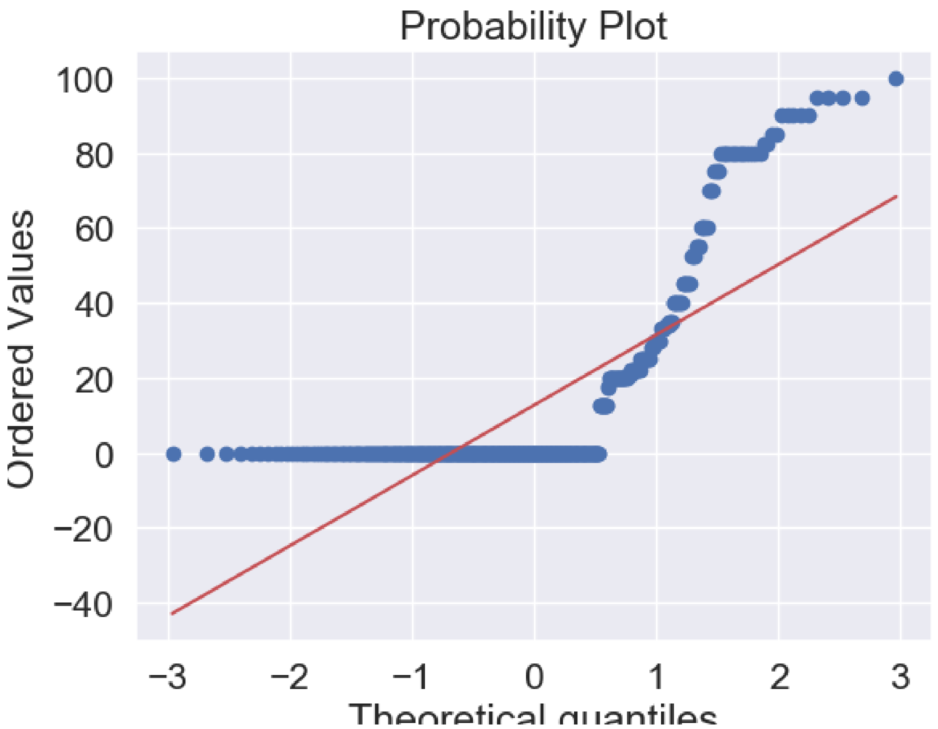




Feature 2: ZN

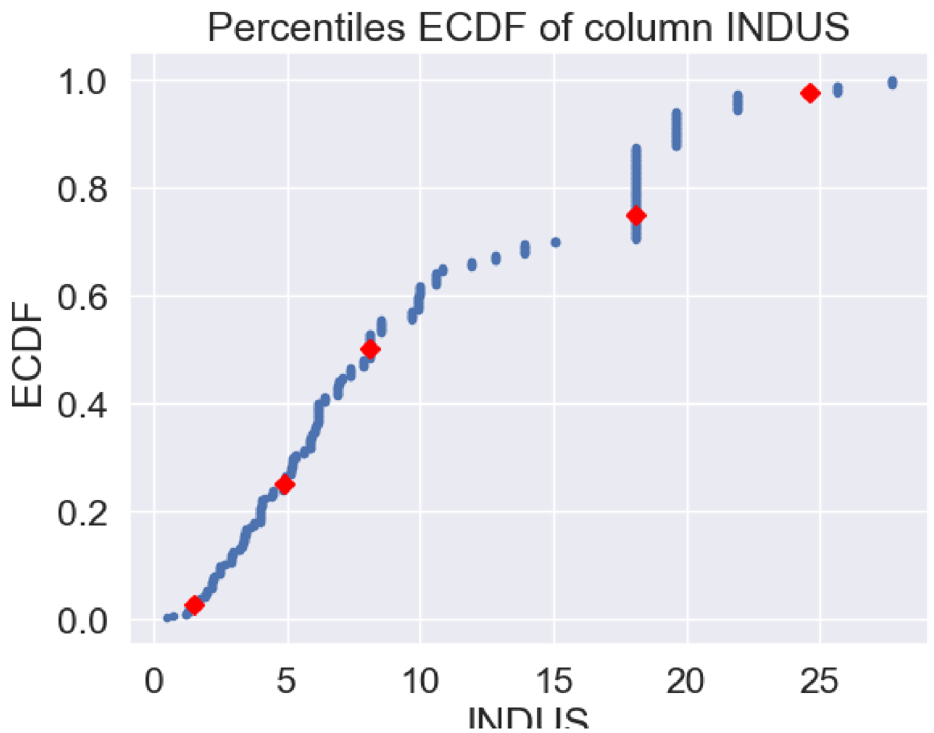
```
The summary statistics of ZN
Mean =  12.721238938053098      Standard Deviation =     24.29107567018233

Boundaries for 4 Equal Percentiles
[ 0.      0.      0.     20.     84.3125]
```

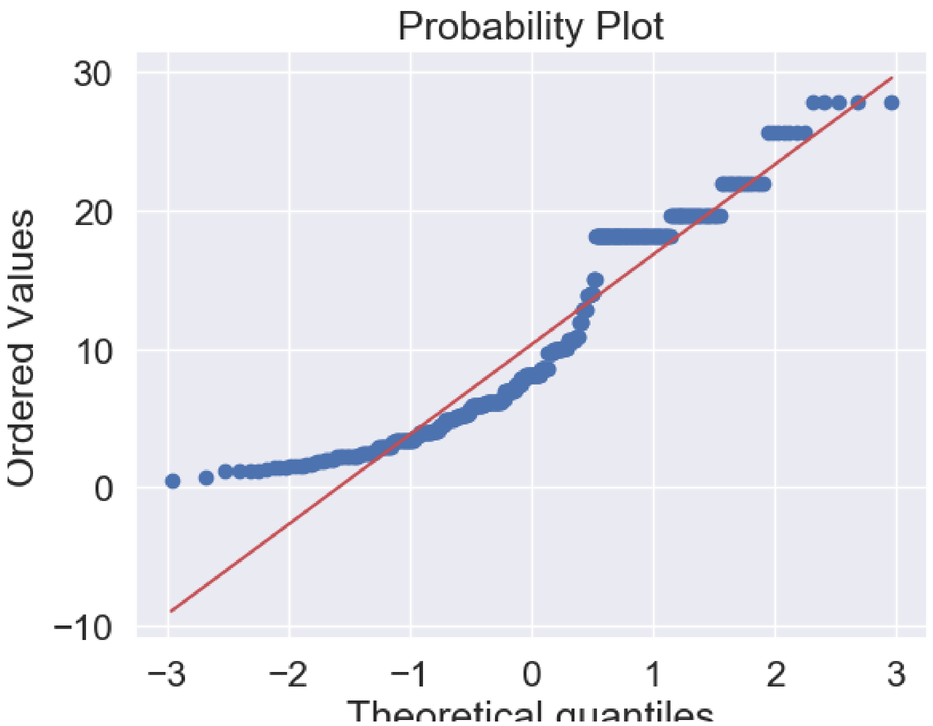




Feature 3: INDUS

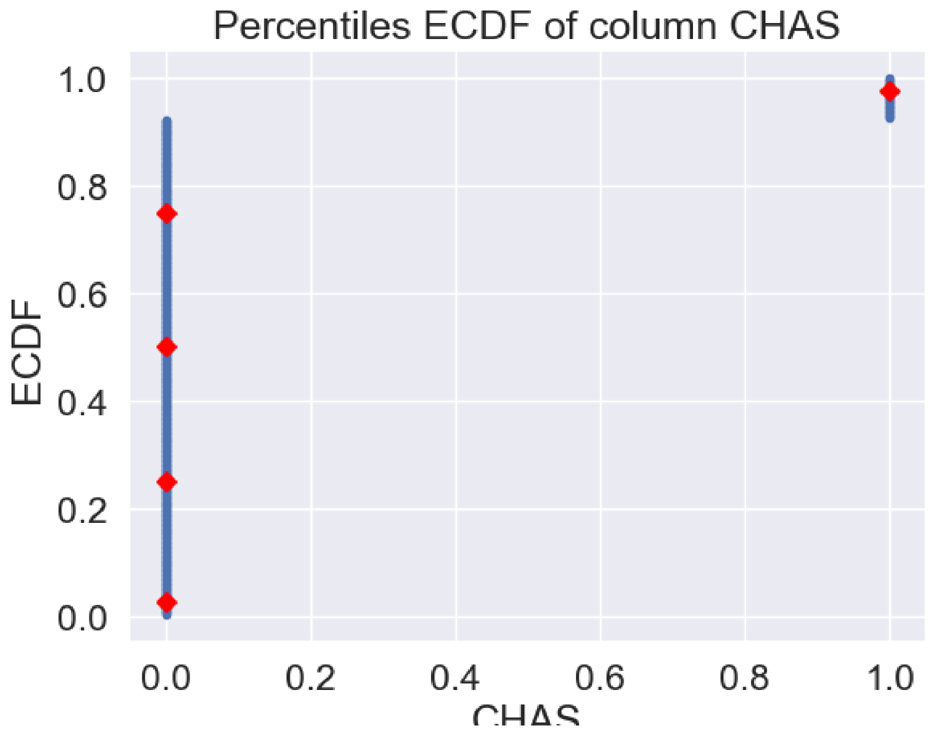
```
The summary statistics of INDUS
Mean =  10.304889380530954      Standard Deviation =     6.7895796483967095

Boundaries for 4 Equal Percentiles
[ 1.52   4.93   8.14  18.1    24.616]
```

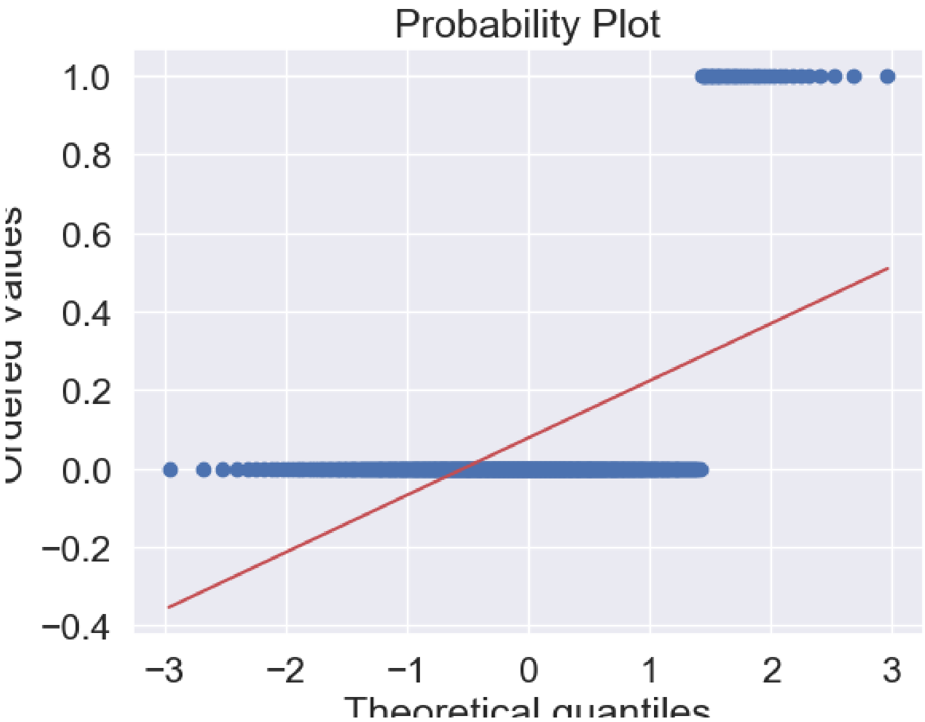

Percentiles ECDF of column INDUS


Probability Plot

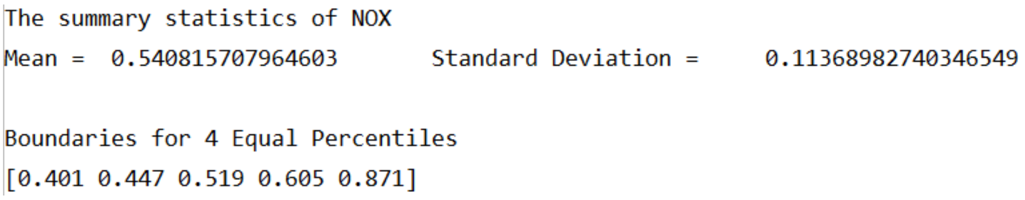
Feature 4: CHAS

```
The summary statistics of CHAS
Mean =  0.07743362831858407     Standard Deviation =     0.2672782473827653

Boundaries for 4 Equal Percentiles
[0. 0. 0. 0. 1.]
```

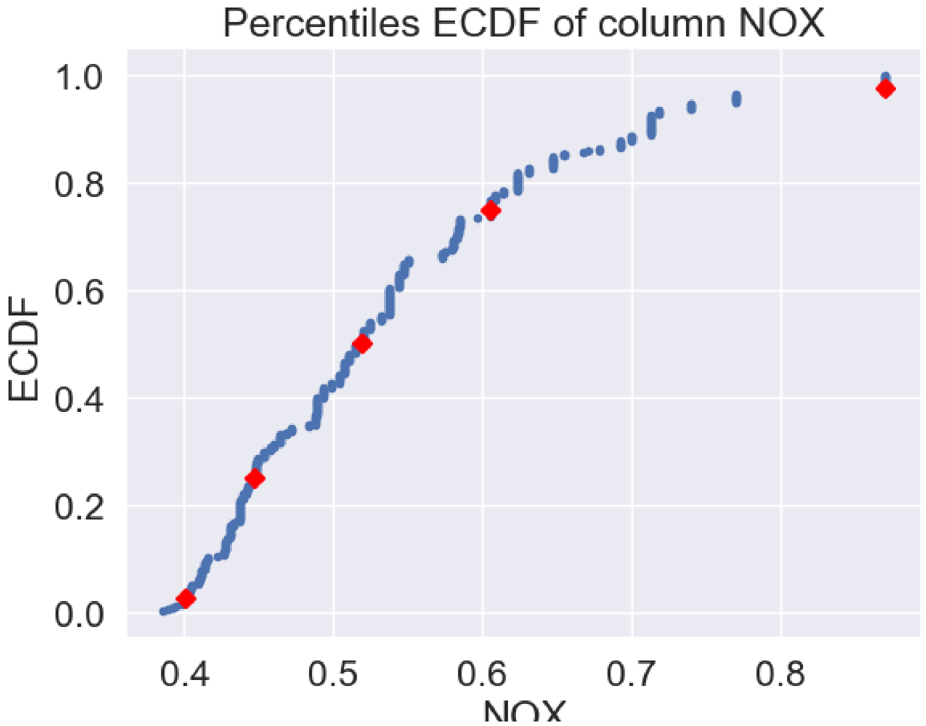


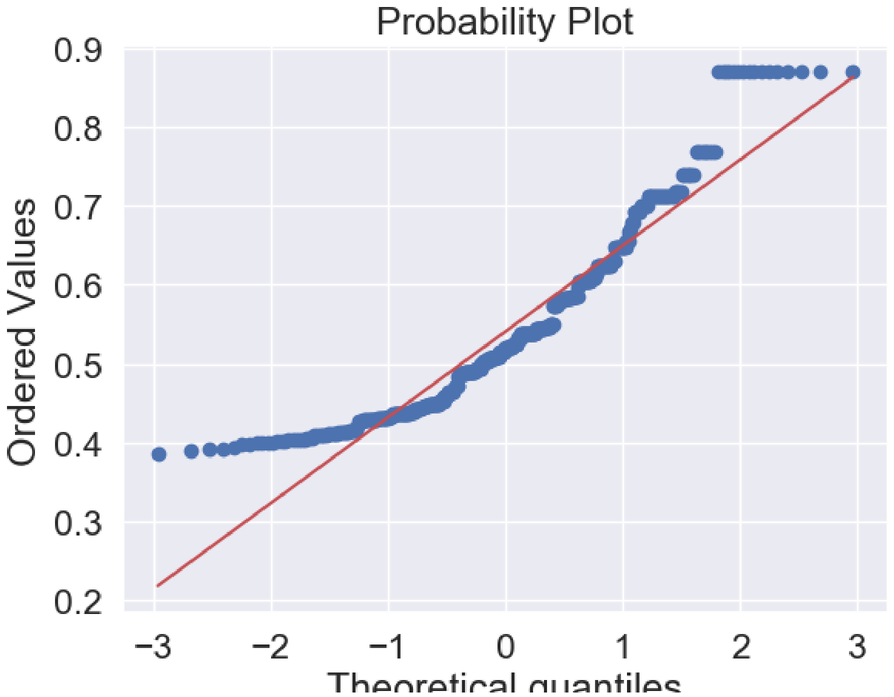


Feature 5: NOX

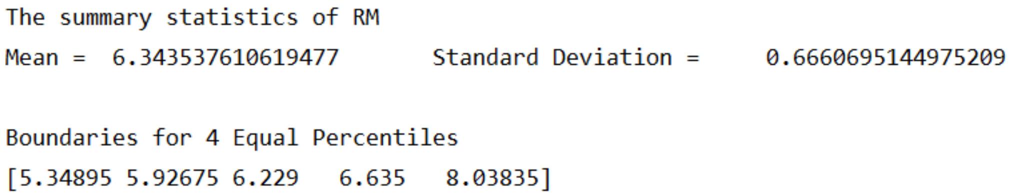


```
The summary statistics of NOX
Mean =  0.540815707964603       Standard Deviation =     0.11368982740346549

Boundaries for 4 Equal Percentiles
[0.401 0.447 0.519 0.605 0.871]
```

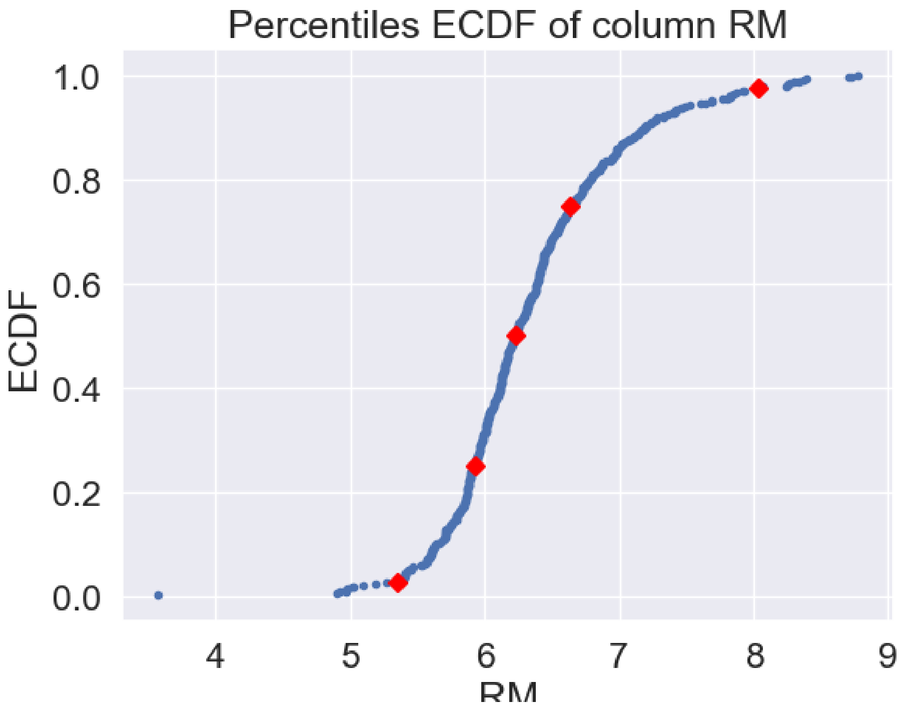


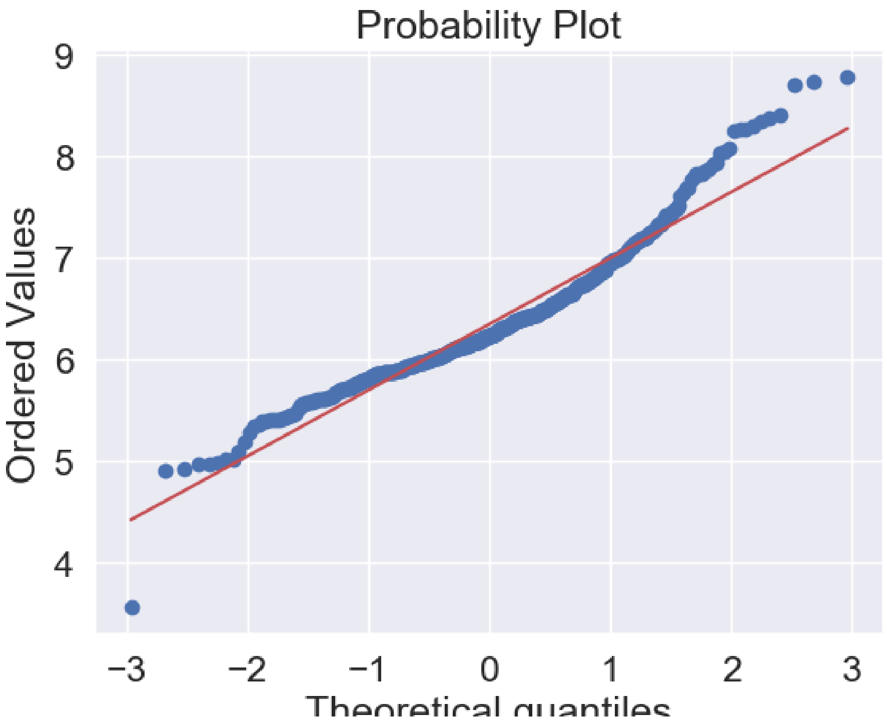


Feature 6: RM

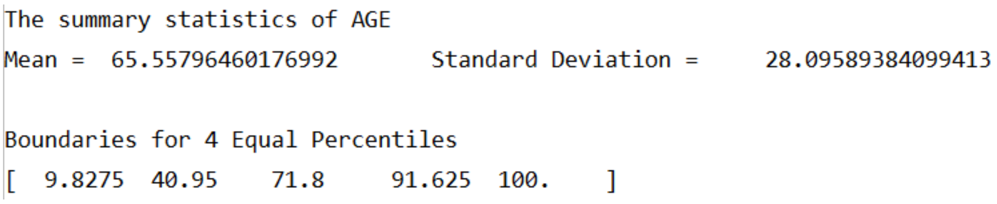


```
The summary statistics of RM
Mean =  6.343537610619477       Standard Deviation =     0.6660695144975209

Boundaries for 4 Equal Percentiles
[5.34895 5.92675 6.229   6.635   8.03835]
```

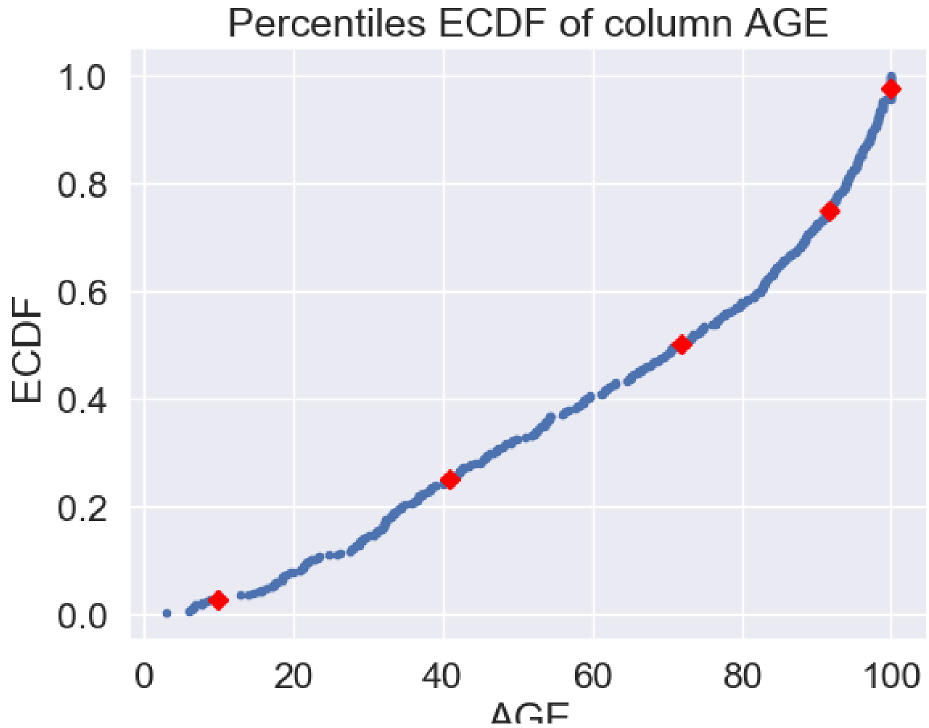


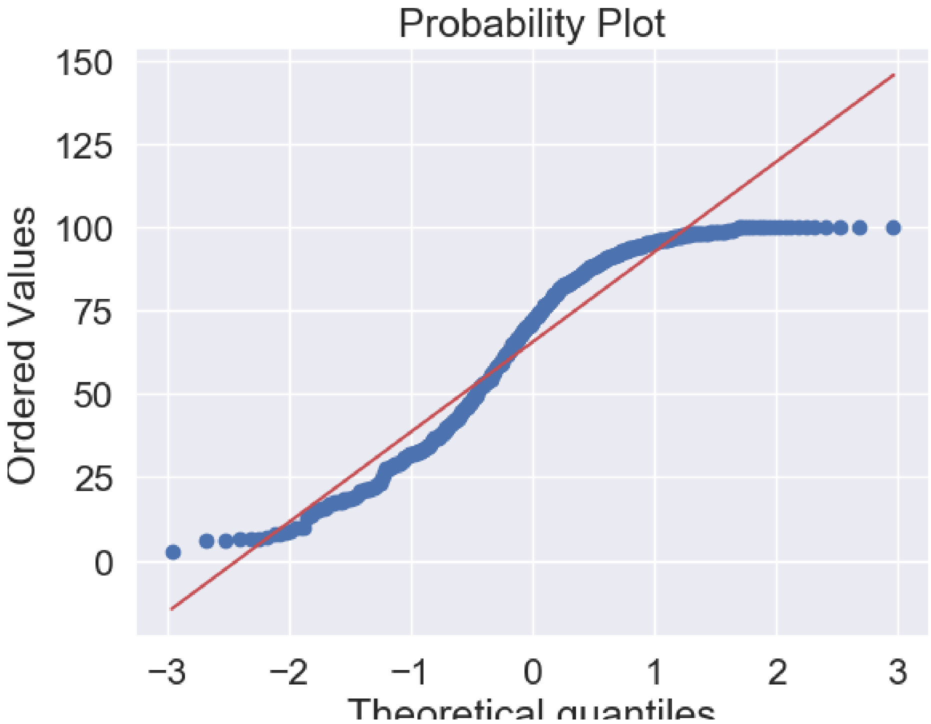


Feature 7: AGE



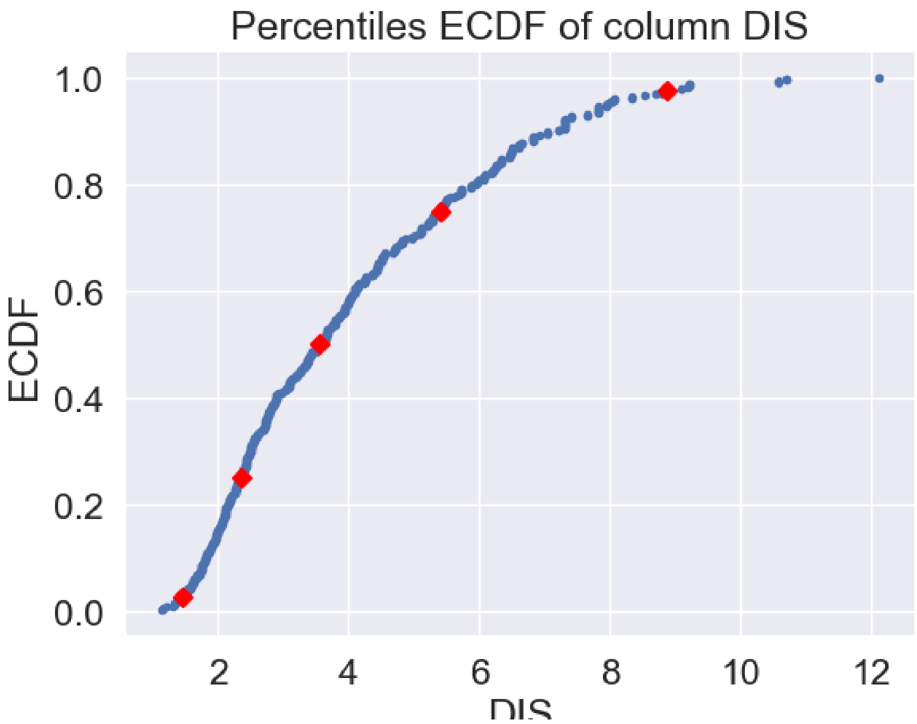
```
The summary statistics of AGE
Mean =  65.55796460176992       Standard Deviation =     28.09589384099413


Boundaries for 4 Equal Percentiles
[  9.8275  40.95    71.8      91.625  100.     ]
```

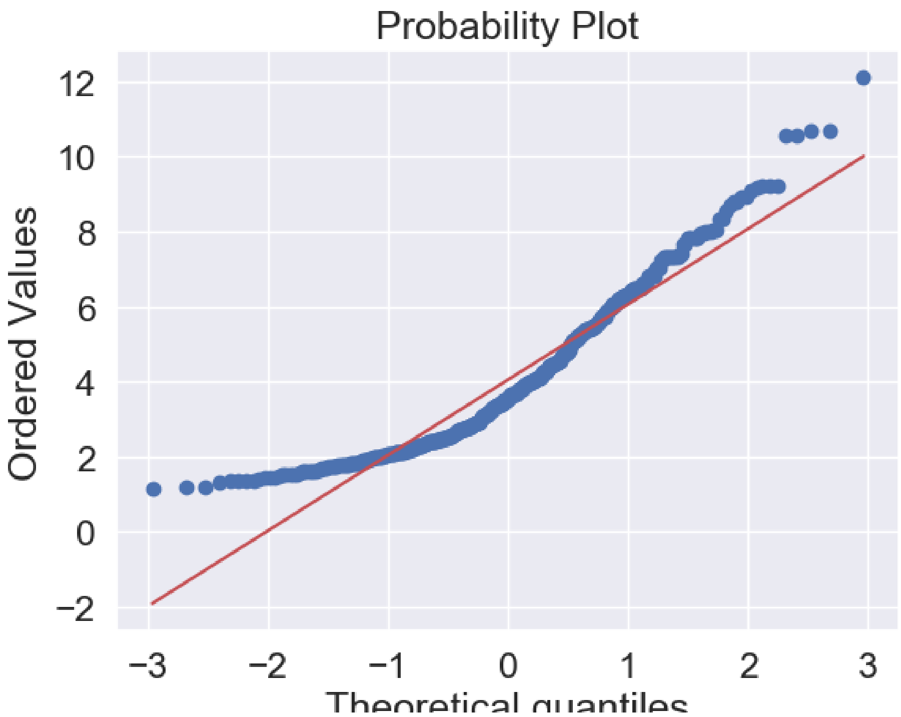




Feature 8: DIS

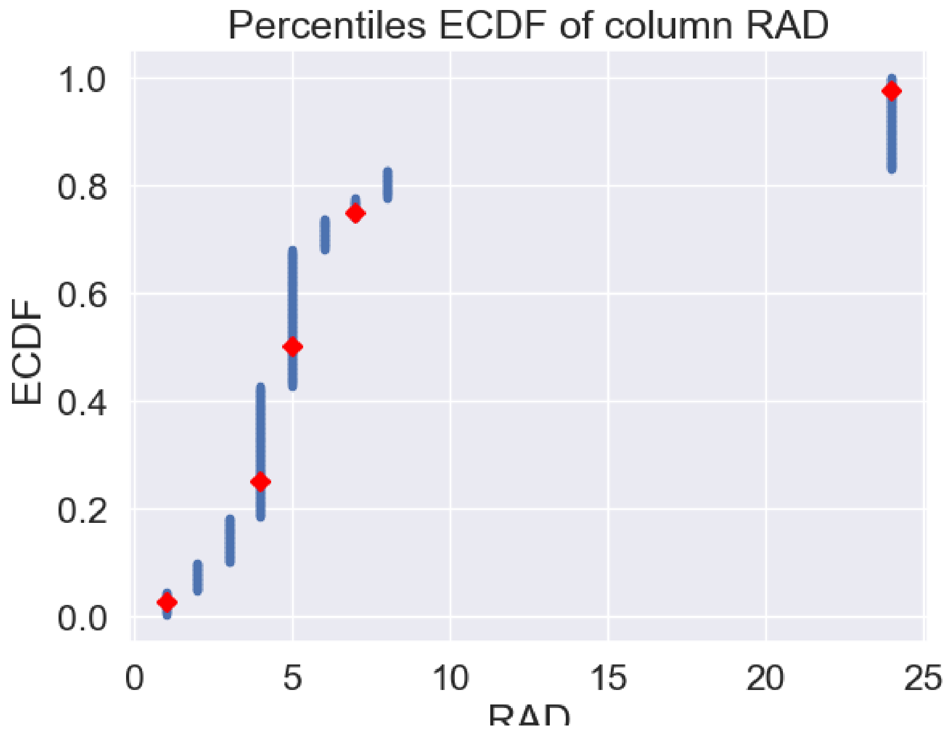
```
The summary statistics of DIS
Mean =  4.0435703539822985      Standard Deviation =     2.0881782846436736

Boundaries for 4 Equal Percentiles
[1.4563775 2.35475   3.5504    5.4011    8.875185 ]
```

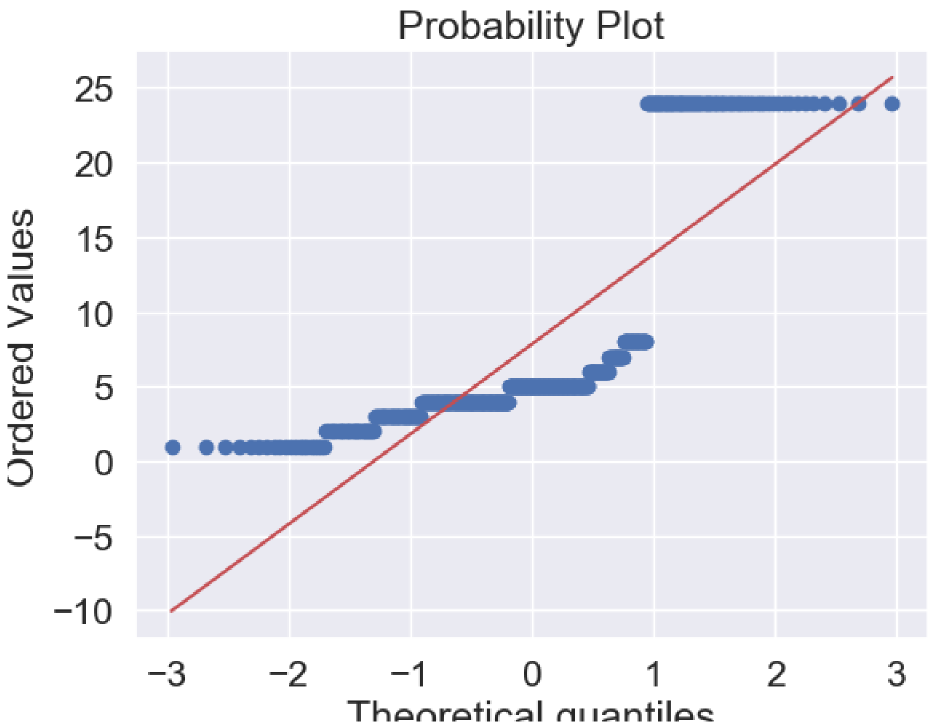




Feature 9: RAD

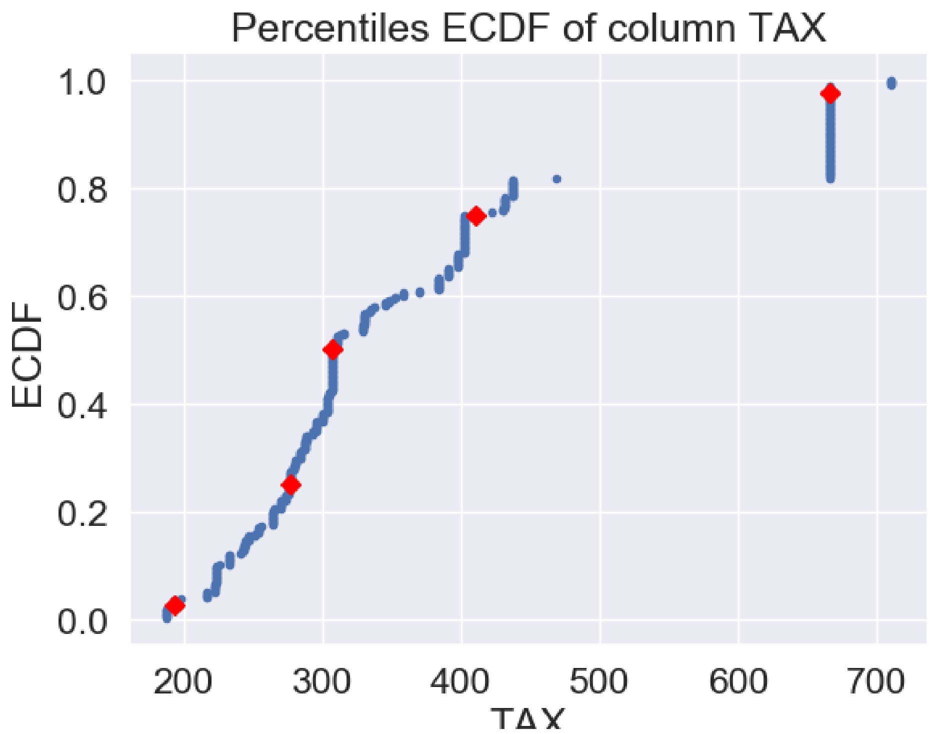
```
The summary statistics of RAD
Mean =  7.823008849557522       Standard Deviation =     7.535144898801841

Boundaries for 4 Equal Percentiles
[ 1.  4.  5.  7. 24.]
```

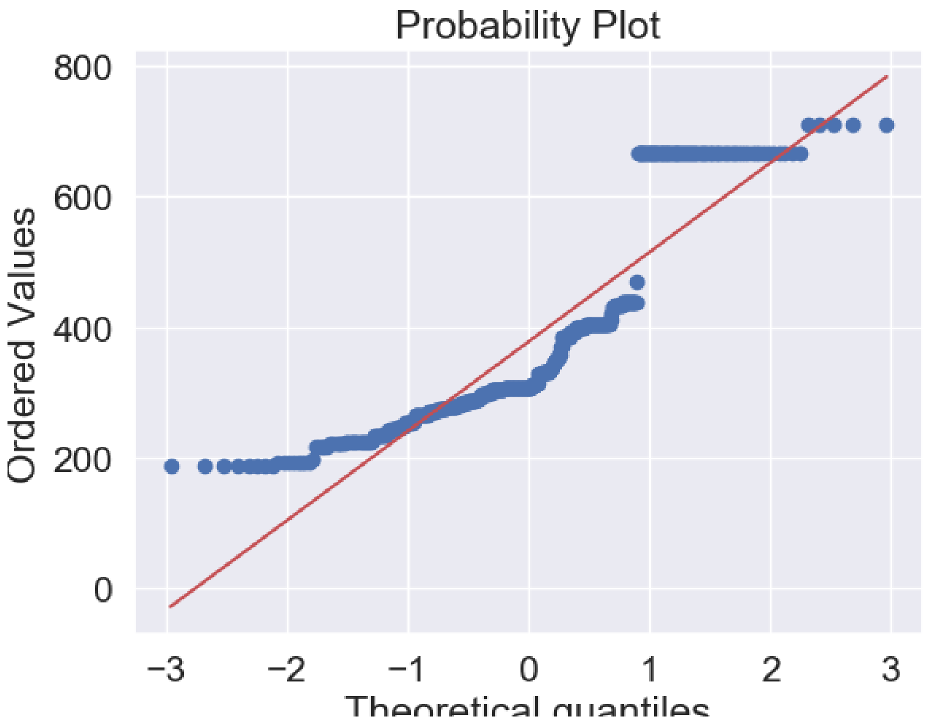




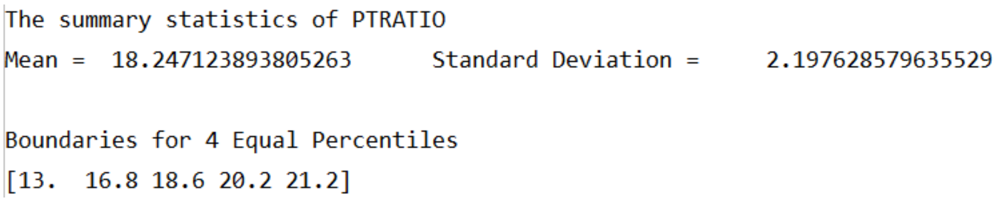
Feature 10: TAX

```
The summary statistics of TAX
Mean =  377.4424778761062       Standard Deviation =     151.1600826442304

Boundaries for 4 Equal Percentiles
[193.   276.75 307.    411.    666.   ]
```

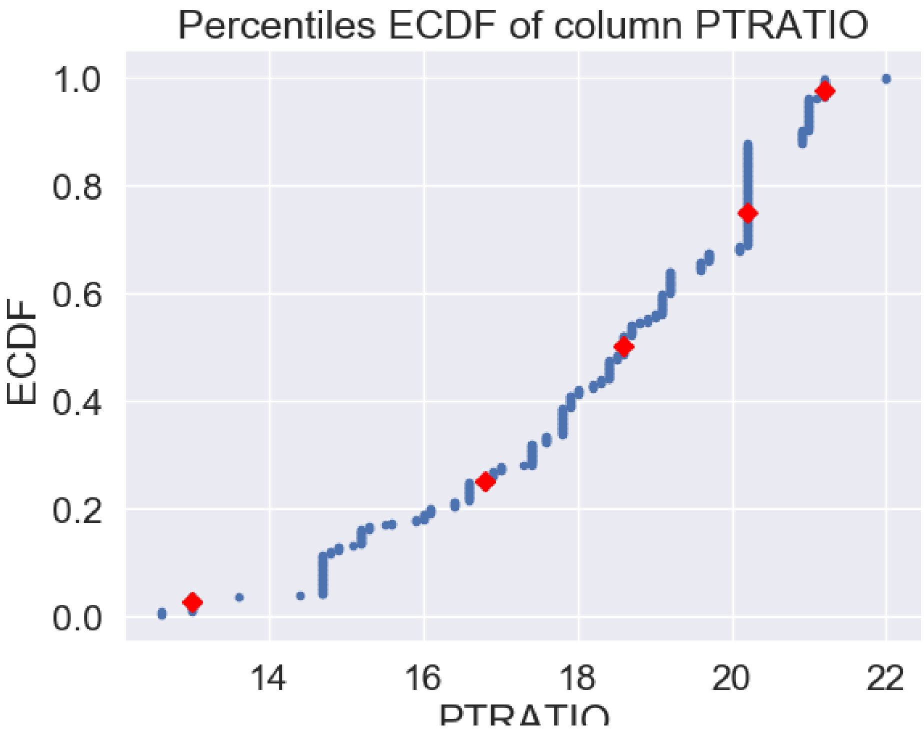


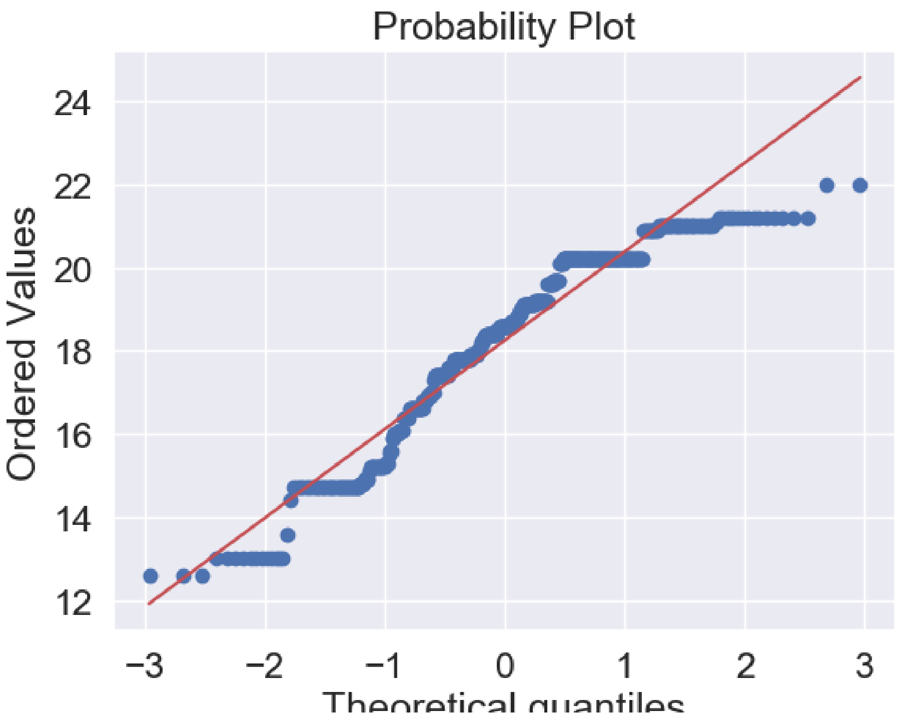


Feature 11: PTRATIO



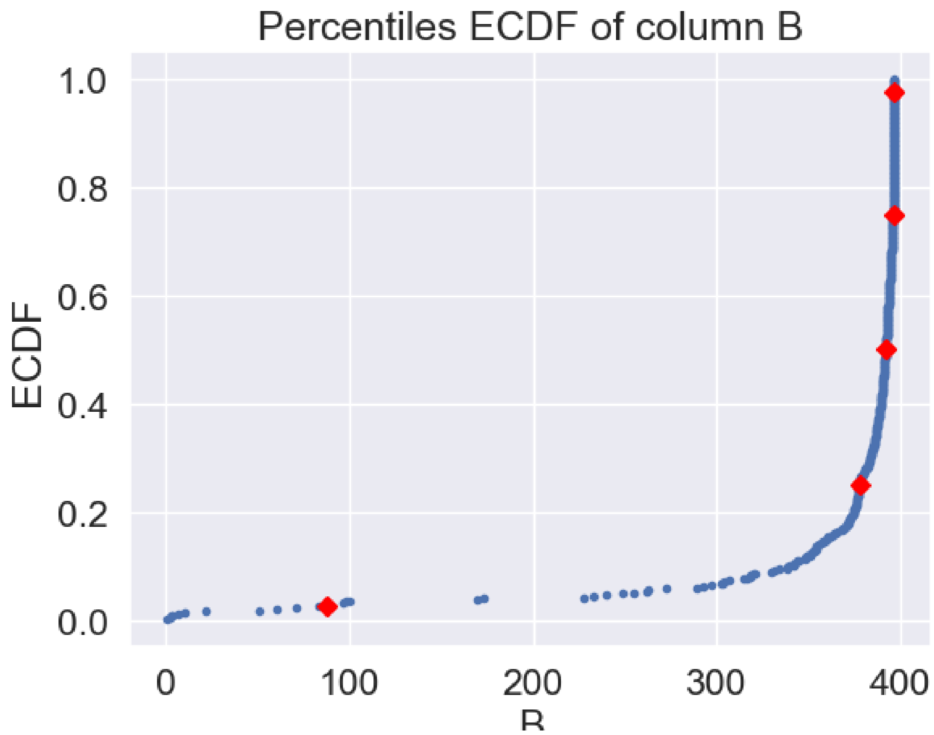
```
The summary statistics of PTRATIO
Mean =  18.247123893805263      Standard Deviation =     2.197628579635529

Boundaries for 4 Equal Percentiles
[13.   16.8 18.6 20.2 21.2]
```

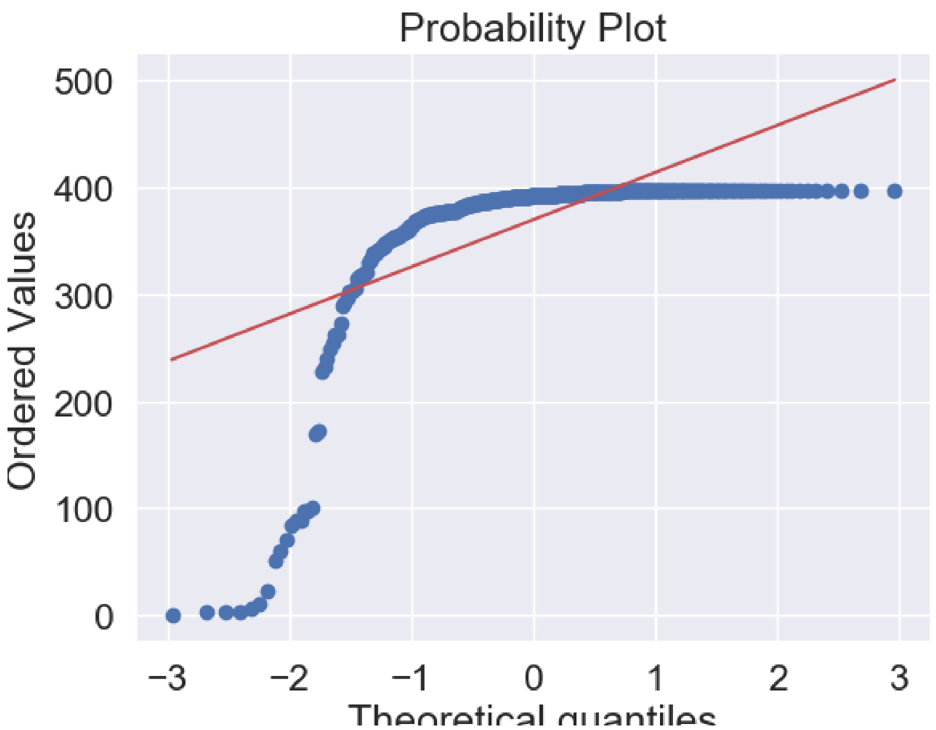




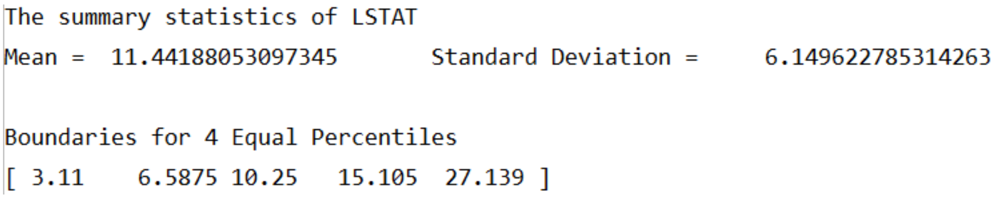
Feature 12: B

```
The summary statistics of B
Mean =  369.8265044247781       Standard Deviation =     68.47856213335493

Boundaries for 4 Equal Percentiles
[ 88.1805 377.7175 392.08   396.1575 396.9    ]
```

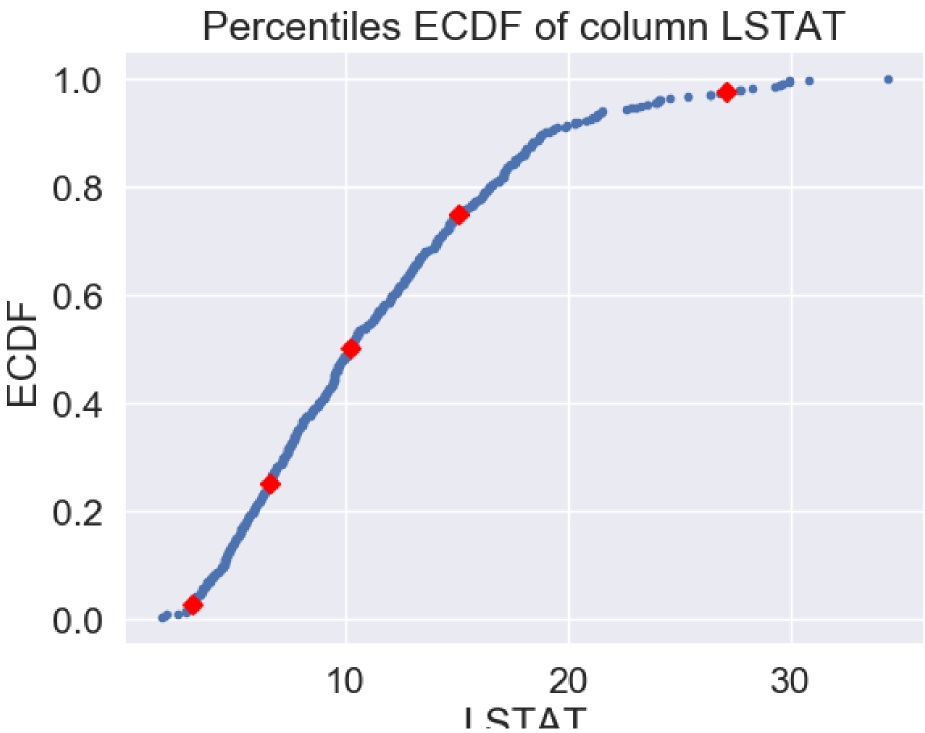


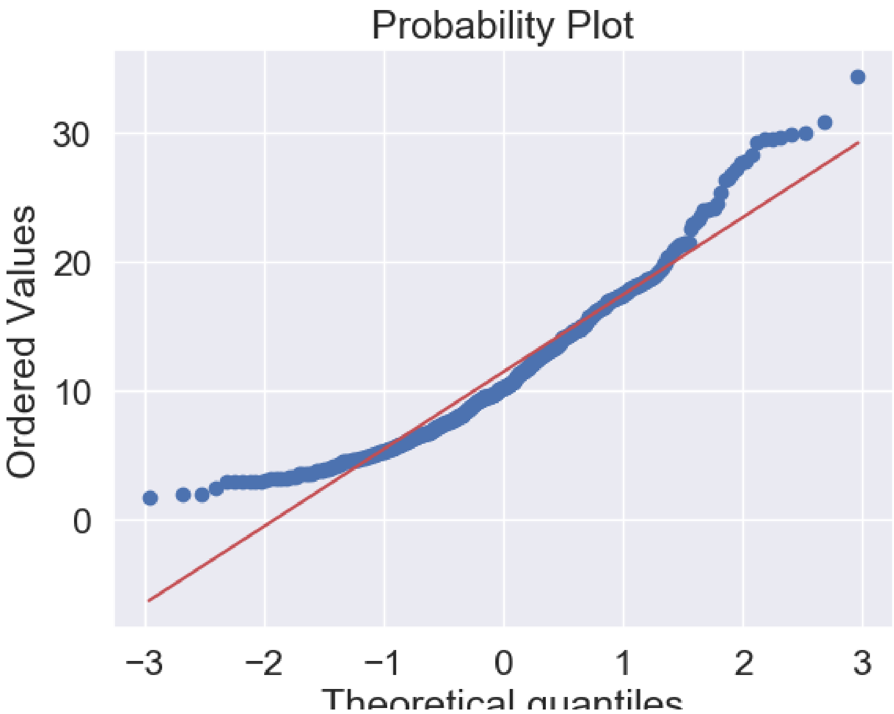


Feature 13: LSTAT

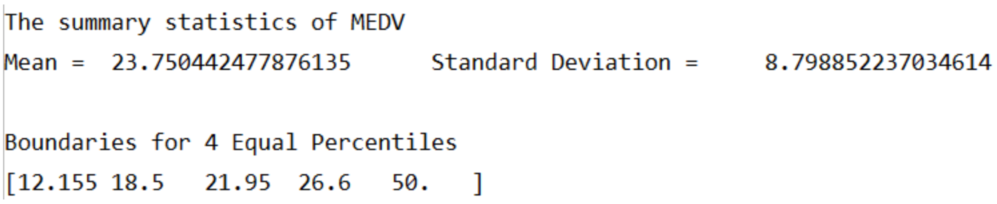


```
The summary statistics of LSTAT
Mean =  11.44188053097345       Standard Deviation =     6.149622785314263


Boundaries for 4 Equal Percentiles
[ 3.11    6.5875 10.25   15.105   27.139 ]
```

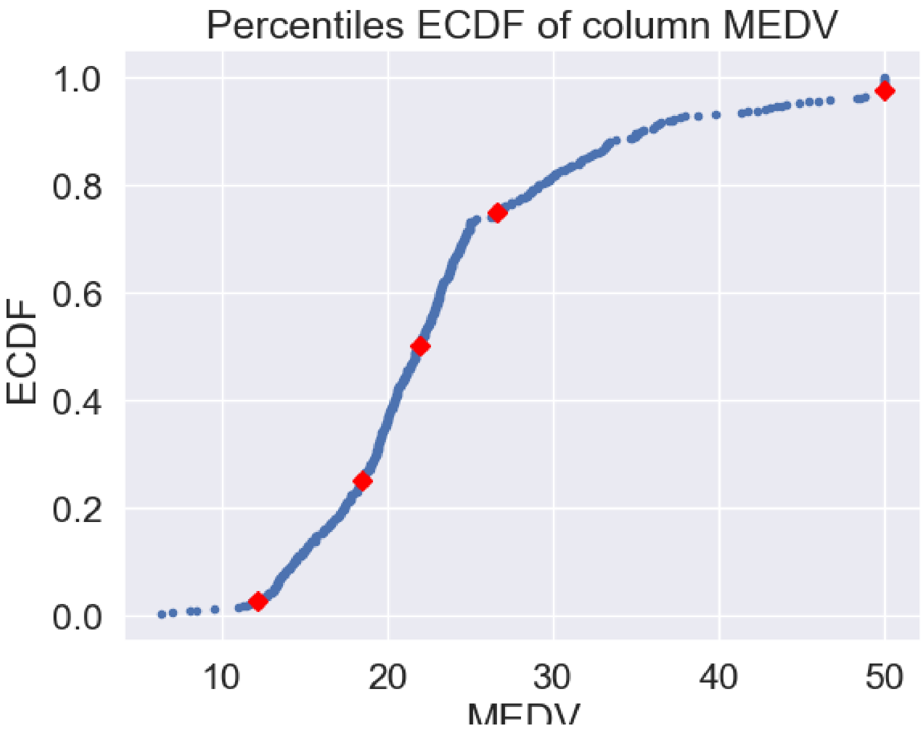


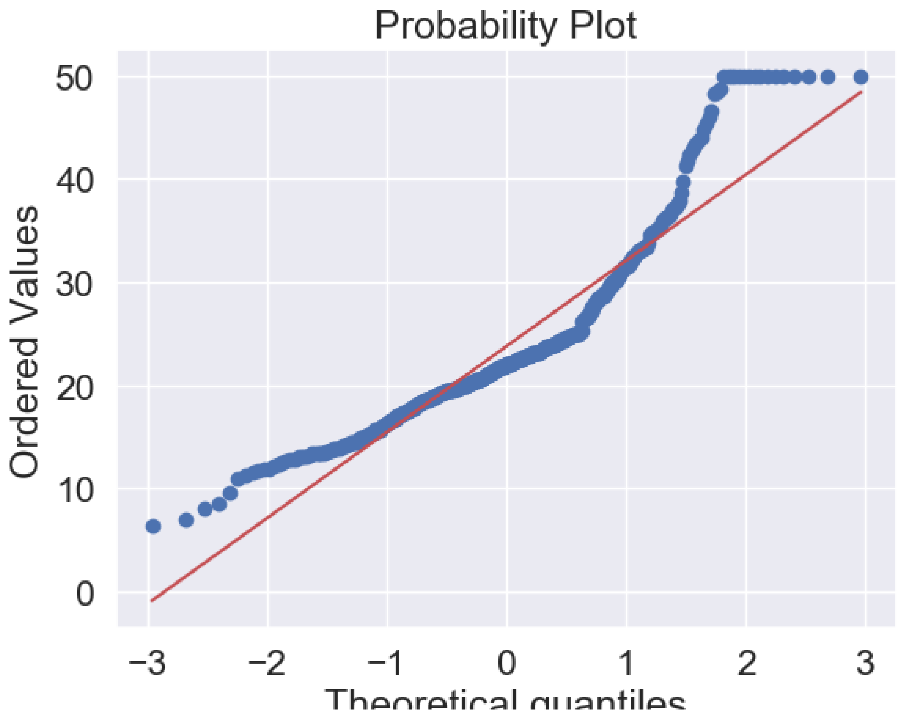


Target: MEDV



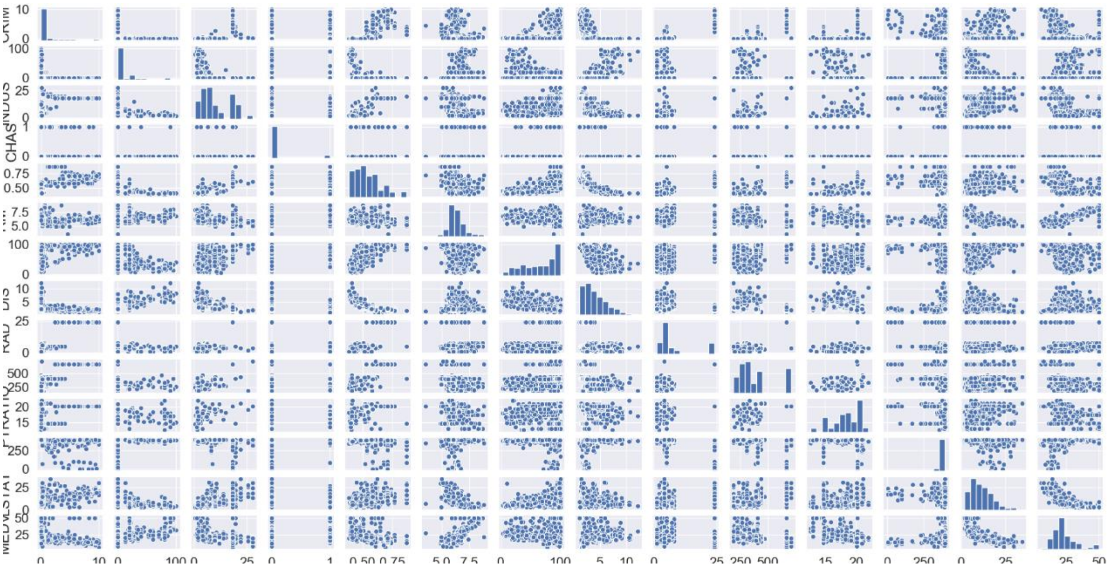
```
The summary statistics of MEDV
Mean =   23.750442477876135      Standard Deviation =     8.798852237034614

Boundaries for 4 Equal Percentiles
[12.155 18.5    21.95   26.6    50.    ]
```



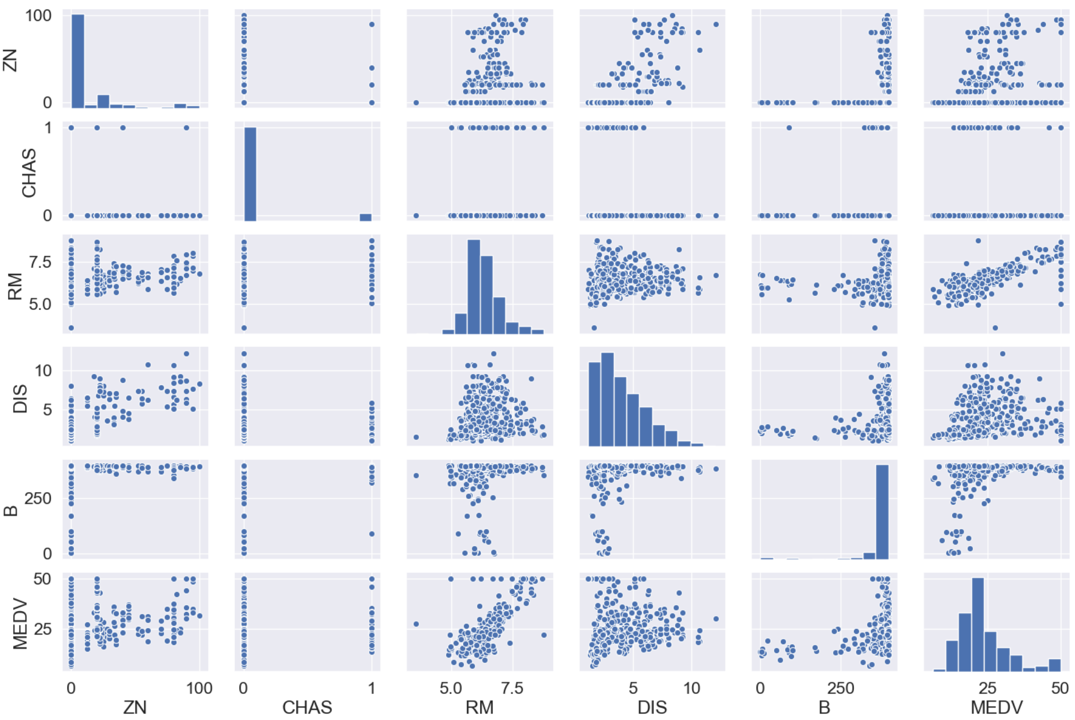


The Graphical Summary of the Relationships:

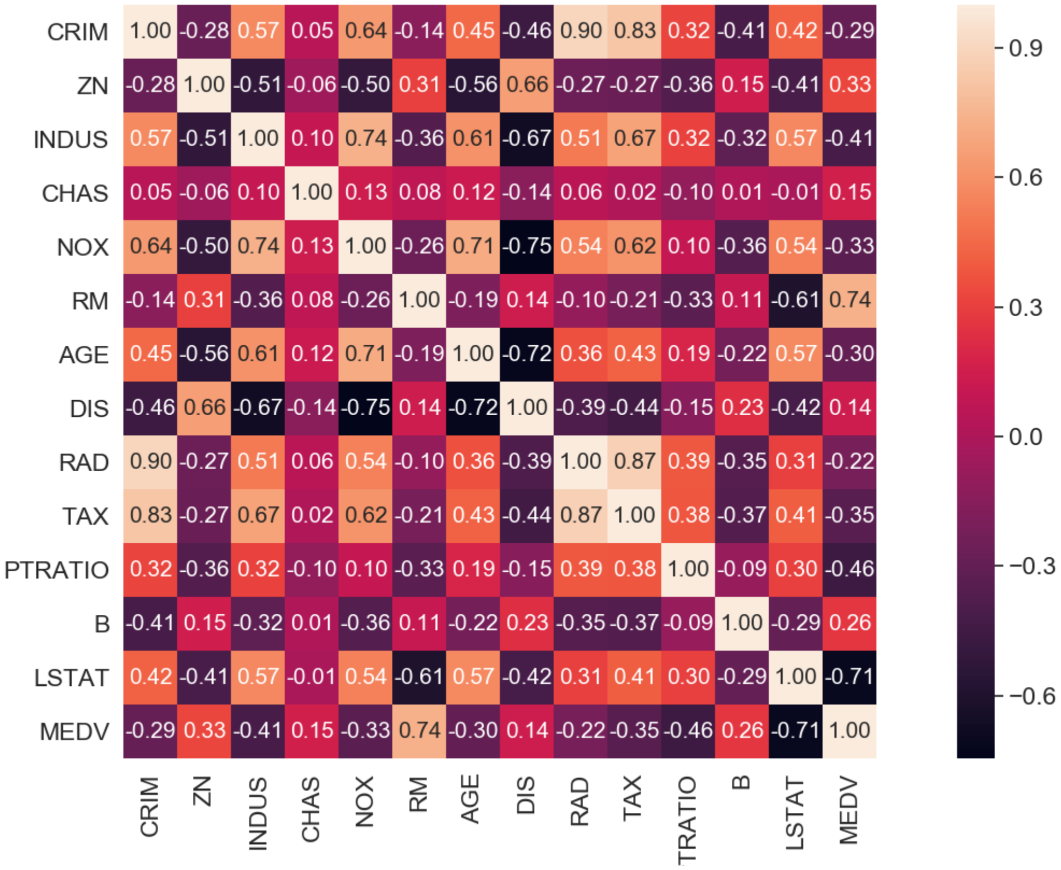
Scatter Plot with all features and target:



Scatter Plot with several features that are most highly correlated with the target:



Heat Map:

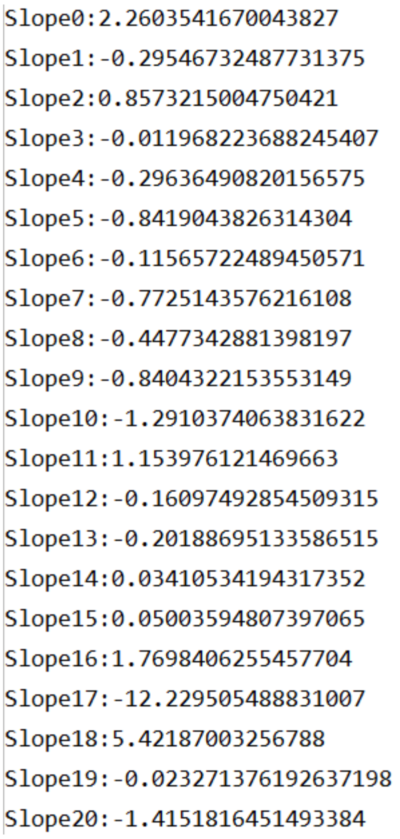


	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	TRATIO	B	LSTAT	MEDV
CRIM	1.00	-0.28	0.57	0.05	0.64	-0.14	0.45	-0.46	0.90	0.83	0.32	-0.41	0.42	-0.29
ZN	-0.28	1.00	-0.51	-0.06	-0.50	0.31	-0.56	0.66	-0.27	-0.27	-0.36	0.15	-0.41	0.33
INDUS	0.57	-0.51	1.00	0.10	0.74	-0.36	0.61	-0.67	0.51	0.67	0.32	-0.32	0.57	-0.41
CHAS	0.05	-0.06	0.10	1.00	0.13	0.08	0.12	-0.14	0.06	0.02	-0.10	0.01	-0.01	0.15
NOX	0.64	-0.50	0.74	0.13	1.00	-0.26	0.71	-0.75	0.54	0.62	0.10	-0.36	0.54	-0.33
RM	-0.14	0.31	-0.36	0.08	-0.26	1.00	-0.19	0.14	-0.10	-0.21	-0.33	0.11	-0.61	0.74
AGE	0.45	-0.56	0.61	0.12	0.71	-0.19	1.00	-0.72	0.36	0.43	0.19	-0.22	0.57	-0.30
DIS	-0.46	0.66	-0.67	-0.14	-0.75	0.14	-0.72	1.00	-0.39	-0.44	-0.15	0.23	-0.42	0.14
RAD	0.90	-0.27	0.51	0.06	0.54	-0.10	0.36	-0.39	1.00	0.87	0.39	-0.35	0.31	-0.22
TAX	0.83	-0.27	0.67	0.02	0.62	-0.21	0.43	-0.44	0.87	1.00	0.38	-0.37	0.41	-0.35
PTRATIO	0.32	-0.36	0.32	-0.10	0.10	-0.33	0.19	-0.15	0.39	0.38	1.00	-0.09	0.30	-0.46
B	-0.41	0.15	-0.32	0.01	-0.36	0.11	-0.22	0.23	-0.35	-0.37	-0.09	1.00	-0.29	0.26
LSTAT	0.42	-0.41	0.57	-0.01	0.54	-0.61	0.57	-0.42	0.31	0.41	0.30	-0.29	1.00	-0.71
MEDV	-0.29	0.33	-0.41	0.15	-0.33	0.74	-0.30	0.14	-0.22	-0.35	-0.46	0.26	-0.71	1.00

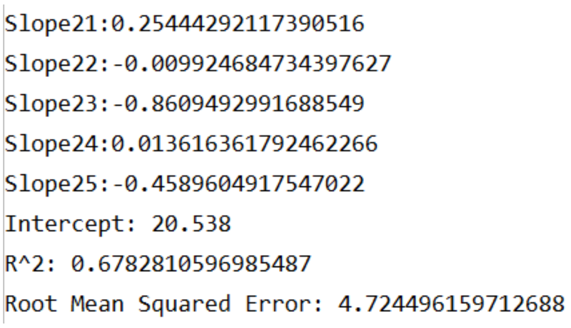
Part 2 Linear Regression

Basic Information:

Coefficients:

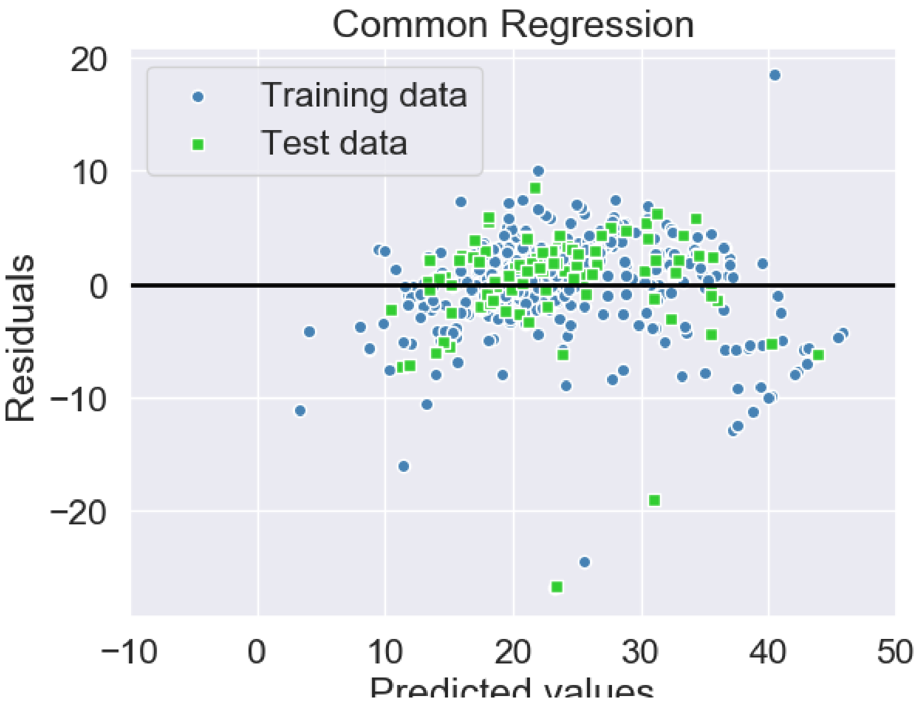


```
Slope0:2.2603541670043827
Slope1:-0.29546732487731375
Slope2:0.8573215004750421
Slope3:-0.011968223688245407
Slope4:-0.29636490820156575
Slope5:-0.8419043826314304
Slope6:-0.11565722489450571
Slope7:-0.7725143576216108
Slope8:-0.4477342881398197
Slope9:-0.8404322153553149
Slope10:-1.2910374063831622
Slope11:1.153976121469663
Slope12:-0.16097492854509315
Slope13:-0.20188695133586515
Slope14:0.03410534194317352
Slope15:0.05003594807397065
Slope16:1.7698406255457704
Slope17:-12.229505488831007
Slope18:5.42187003256788
Slope19:-0.023271376192637198
Slope20:-1.4151816451493384
```



```
Slope21:0.25444292117390516
Slope22:-0.009924684734397627
Slope23:-0.8609492991688549
Slope24:0.013616361792462266
Slope25:-0.4589604917547022
Intercept: 20.538
R^2: 0.6782810596985487
Root Mean Squared Error: 4.724496159712688
```

Residual Errors Plot:

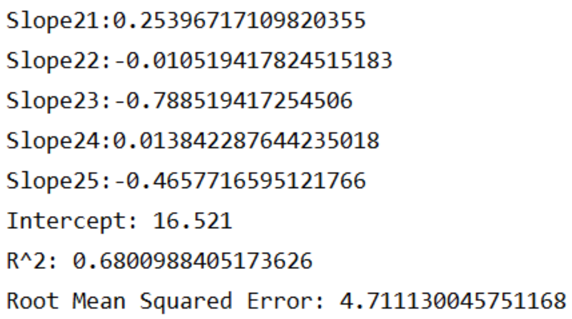


Part 3.1 Ridge Regression

Basic Information (alpha=1):

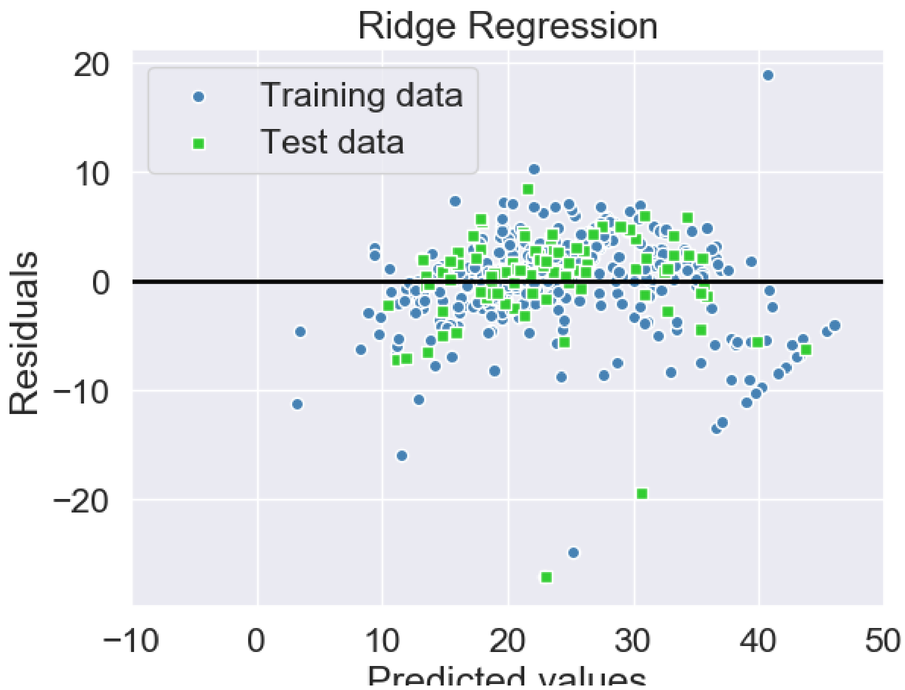
Coefficients:

```
Slope0:2.1028140568415723
Slope1:-0.4008879923555803
Slope2:0.8015166712534275
Slope3:0.0022278178814300535
Slope4:-0.40949966730792353
Slope5:-0.8718283170681906
Slope6:-0.16866412859340024
Slope7:-0.683763268778345
Slope8:-0.40758099893026484
Slope9:-0.7843351192524964
Slope10:-1.3210754802359475
Slope11:1.0520236315792806
Slope12:-0.18324476745319004
Slope13:-0.2637469024378047
Slope14:0.03641580121322092
Slope15:0.025320199647566287
Slope16:1.6461769494504923
Slope17:-6.274457623086177
Slope18:5.455181162110687
Slope19:-0.027709323551092038
Slope20:-1.334963475362354
```



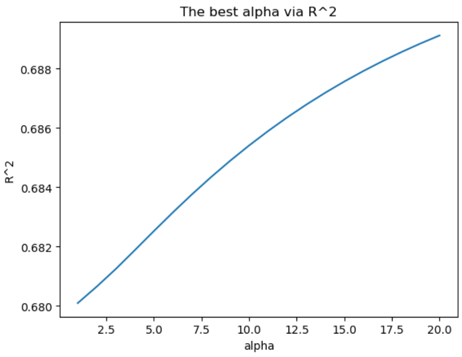
```
Slope21:0.25396717109820355
Slope22:-0.010519417824515183
Slope23:-0.788519417254506
Slope24:0.013842287644235018
Slope25:-0.4657716595121766
Intercept: 16.521
R^2: 0.6800988405173626
Root Mean Squared Error: 4.711130045751168
```

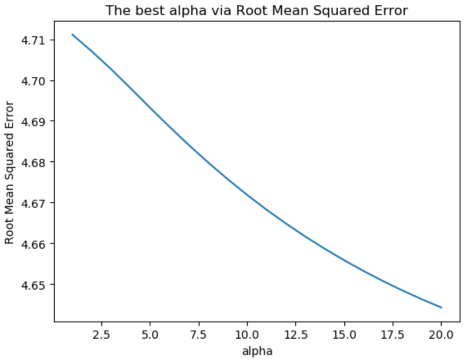
Residual Errors Plot:



The Best Alpha for Ridge Regression:

We choose alpha from 1 to 20, and make the plots of each one's R^2 & RMSE.



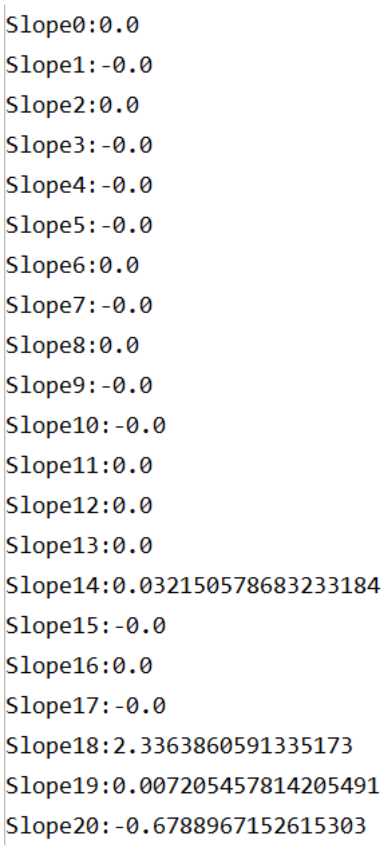


The best alpha has to have the smallest RMSE as well as the highest R^2, so it is alpha = 20. The R^2 is 0.689, and the RMSE is 4.644.

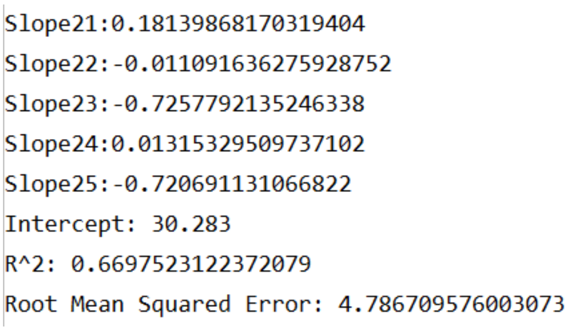
Part 3.2 Lasso Regression

Basic Information (alpha=1):

Coefficients:

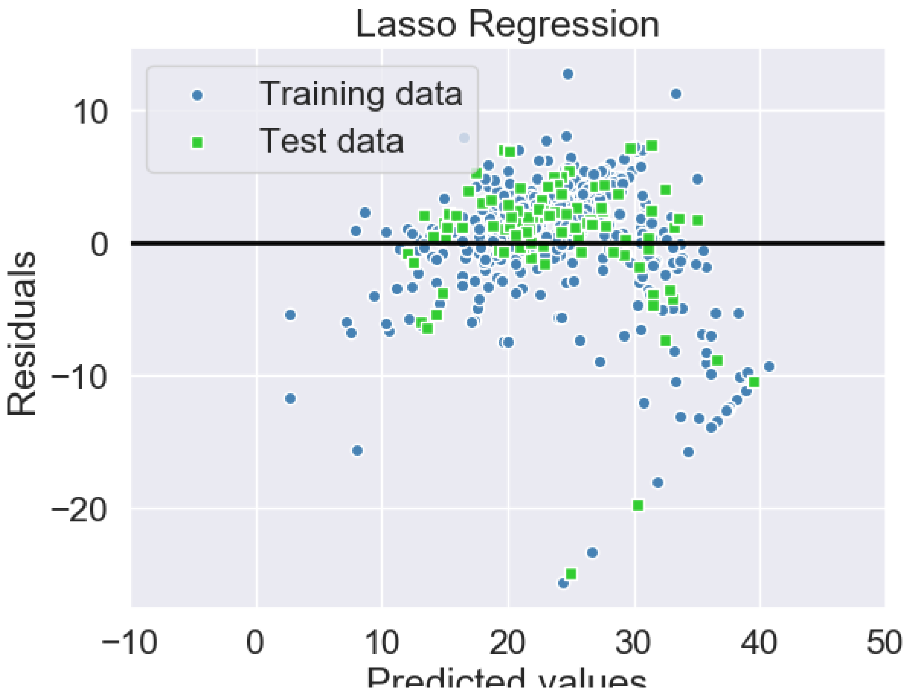


```
Slope0:0.0
Slope1:-0.0
Slope2:0.0
Slope3:-0.0
Slope4:-0.0
Slope5:-0.0
Slope6:0.0
Slope7:-0.0
Slope8:0.0
Slope9:-0.0
Slope10:-0.0
Slope11:0.0
Slope12:0.0
Slope13:0.0
Slope14:0.032150578683233184
Slope15:-0.0
Slope16:0.0
Slope17:-0.0
Slope18:2.3363860591335173
Slope19:0.007205457814205491
Slope20:-0.6788967152615303
```

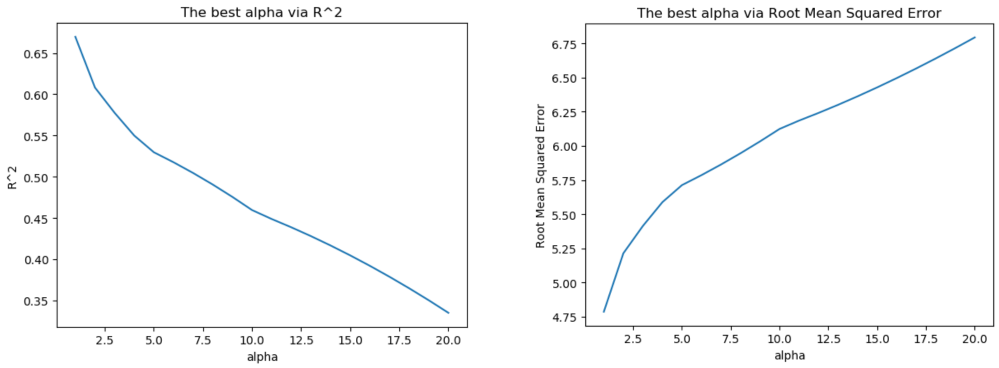


```
Slope21:0.18139868170319404
Slope22:-0.011091636275928752
Slope23:-0.7257792135246338
Slope24:0.01315329509737102
Slope25:-0.720691131066822
Intercept: 30.283
R^2: 0.6697523122372079
Root Mean Squared Error: 4.786709576003073
```

Residual Errors Plot:



The Best Alpha for Lasso Regression:



The best alpha has to have the smallest RMSE as well as the highest R^2, so it is alpha = 1. The R^2 is 0.670, and the RMSE is 4.787.

Part 4 Conclusions

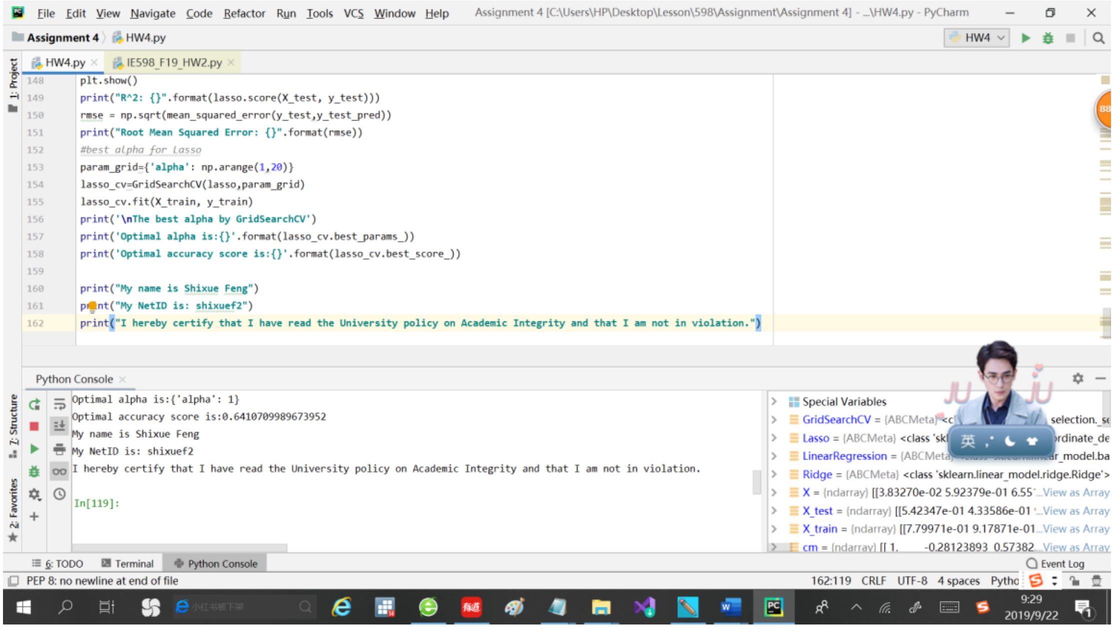
	R^2	RMSE
Linear Regression	0.678	4.724
Ridge Regression (best)	0.689	4.644
Lasso Regression (best)	0.670	4.787

According to the R^2 and RMSE from various regressions, we can know that Ridge Regression can best explain the data with the largest R^2 and the smallest RMSE compared with others. However, the Lasso Regression don’t perform well as our expectation. I think we can try to find the reason from the coefficient form. Lasso Regression sets too many coefficients as 0. As a result, this model may explain the data inefficiently and present more errors.

Part 5 Appendix

Link to my code:
https://github.com/fengzixue96/IE598_F19_HW4/blob/master/IE598_F19_HW4.py

The screenshot:



File Edit View Navigate Code Refactor Run Tools VCS Window Help
Assignment 4 [C:\Users\HP\Desktop\Lesson\598\Assignment\Assignment 4] - ...\HW4.py - PyCharm
Assignment 4 〉 HW4.py
HW4
HW4.py
IE598_F19_HW2.py
1: Project
7: Structure
2: Favorites
Python Console
6: TODO
Terminal
Python Console
Event Log
PEP 8: no newline at end of file
162:119 CRLF UTF-8 4 spaces Pytho
9:29
2019/9/22

```python
plt.show()
print("R^2: {}".format(lasso.score(X_test, y_test)))
rmse = np.sqrt(mean_squared_error(y_test,y_test_pred))
print("Root Mean Squared Error: {}".format(rmse))
#best alpha for Lasso
param_grid={'alpha': np.arange(1,20)}
lasso_cv=GridSearchCV(lasso,param_grid)
lasso_cv.fit(X_train, y_train)
print('\nThe best alpha by GridSearchCV')
print('Optimal alpha is:{}'.format(lasso_cv.best_params_))
print('Optimal accuracy score is:{}'.format(lasso_cv.best_score_))

print("My name is Shixue Feng")
print("My NetID is: shixuef2")
print("I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.")
```

```
Optimal alpha is:{'alpha': 1}
Optimal accuracy score is:0.6410709989673952
My name is Shixue Feng
My NetID is: shixuef2
I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.

In[119]:
```

```
Special Variables
GridSearchCV = {ABCMeta} <c                selection._se
Lasso = {ABCMeta} <class 'skl              rdinate_de
LinearRegression = {ABCMeta}               r_model.ba
Ridge = {ABCMeta} <class 'sklearn.linear_model.ridge.Ridge'>
X = {ndarray} [[3.83270e-02 5.92379e-01 6.55'...View as Array
X_test = {ndarray} [[5.42347e-01 4.33586e-01 '...View as Array
X_train = {ndarray} [[7.79971e-01 9.17871e-01...View as Array
cm = {ndarray} [[ 1.        -0.28123893  0.57382...View as Array
```