My Name: Yohanes Susanto

Net ID: yohanes2

IE 517 MLF F20

Module 4 Homework (Regression)

1. **Introduction**

   Housing2.csv is the dataset used for the homework

2. **Exploratory Data Analysis**

```
Size of data 506 x 27
       ATT1      ATT2      ATT3      ATT4      ATT5      ATT6      ATT7  \
0  0.038327  0.592379  0.655174  0.119839  0.652477  0.984323  0.206738
1  0.225022  0.983103  0.803619  0.836315  0.163104  0.637497  0.008760
2  0.423233  0.375808  0.271293  0.729824  0.886744  0.043703  0.457700
3  0.743370  0.929103  0.589894  0.644012  0.110490  0.774604  0.306483
4  0.378623  0.786609  0.712752  0.110274  0.762133  0.030069  0.316631

       ATT8      ATT9     ATT10  ...    NOX     RM   AGE     DIS  RAD  TAX  \
0  0.374650  0.463350  0.333610  ...  0.538  6.575  65.2  4.0900    1  296
1  0.631190  0.207978  0.880357  ...  0.469  6.421  78.9  4.9671    2  242
2  0.862450  0.901924  0.062488  ...  0.469  7.185  61.1  4.9671    2  242
3  0.880599  0.630401  0.928894  ...  0.458  6.998  45.8  6.0622    3  222
4  0.667073  0.426443  0.400557  ...  0.458  7.147  54.2  6.0622    3  222

   PTRATIO       B  LSTAT  MEDV
0     15.3  396.90   4.98  24.0
1     17.8  396.90   9.14  21.6
2     17.8  392.83   4.03  34.7
3     18.7  394.63   2.94  33.4
4     18.7  396.90   5.33  36.2
```

   The data contains 506 samples and 27 attributes.

```
ATT1       float64
ATT2       float64
ATT3       float64
ATT4       float64
ATT5       float64
ATT6       float64
ATT7       float64
ATT8       float64
ATT9       float64
ATT10      float64
ATT11      float64
ATT12      float64
ATT13      float64
CRIM       float64
ZN         float64
INDUS      float64
CHAS         int64
NOX        float64
RM         float64
AGE        float64
DIS        float64
RAD          int64
TAX          int64
PTRATIO    float64
B          float64
LSTAT      float64
MEDV       float64
dtype: object
float64
Total number of numeric columns: 27
Total number of categorical columns: 0
```

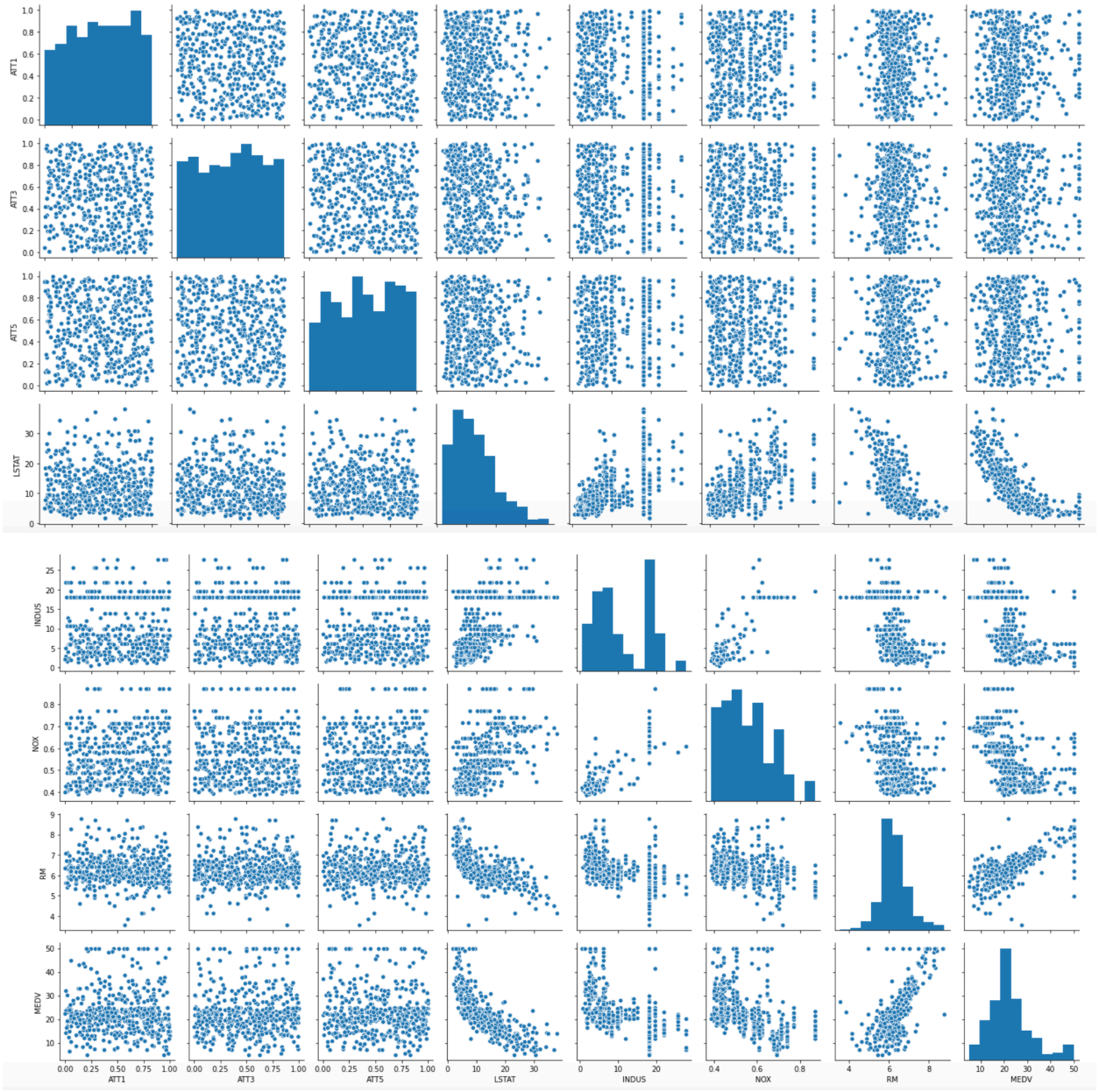
The data contains 27 numeric attributes and 0 categorical attributes

```
ATT1        0
ATT2        0
ATT3        0
ATT4        0
ATT5        0
ATT6        0
ATT7        0
ATT8        0
ATT9        0
ATT10       0
ATT11       0
ATT12       0
ATT13       0
CRIM        0
ZN          0
INDUS       0
CHAS        0
NOX         0
RM          0
AGE         0
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT       0
MEDV        0
dtype: int64
It can be seen that there are no missing values
```

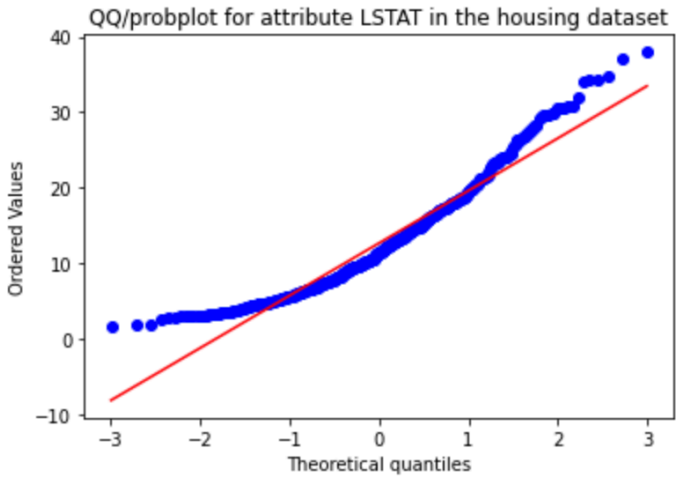
After checking for sparsity of the data, there are no missing values in the dataset.

	ATT1	ATT2	ATT3	ATT4	ATT5	ATT6	ATT7	ATT8	ATT9	ATT10	...	NOX	RM
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	...	506.000000	506.000000
mean	0.518457	0.500422	0.507451	0.498543	0.525487	0.508831	0.501997	0.509998	0.480159	0.501922	...	0.554695	6.284634
std	0.283847	0.298752	0.289607	0.294229	0.283387	0.282400	0.287986	0.290160	0.301086	0.294051	...	0.115878	0.702617
min	0.000727	0.000321	0.000013	0.001541	0.003970	0.000679	0.003653	0.000525	0.001093	0.000263	...	0.385000	3.561000
25%	0.272918	0.235879	0.244897	0.229861	0.283208	0.276366	0.271701	0.257320	0.208171	0.248119	...	0.449000	5.885500
50%	0.521326	0.485701	0.526013	0.506543	0.514982	0.509443	0.499804	0.508327	0.465557	0.487129	...	0.538000	6.208500
75%	0.770235	0.774921	0.750546	0.757517	0.772218	0.730899	0.756420	0.768465	0.739580	0.771559	...	0.624000	6.623500
max	0.995798	0.999265	0.998746	0.995561	0.998635	0.998194	0.999140	0.997083	0.996714	0.999321	...	0.871000	8.780000

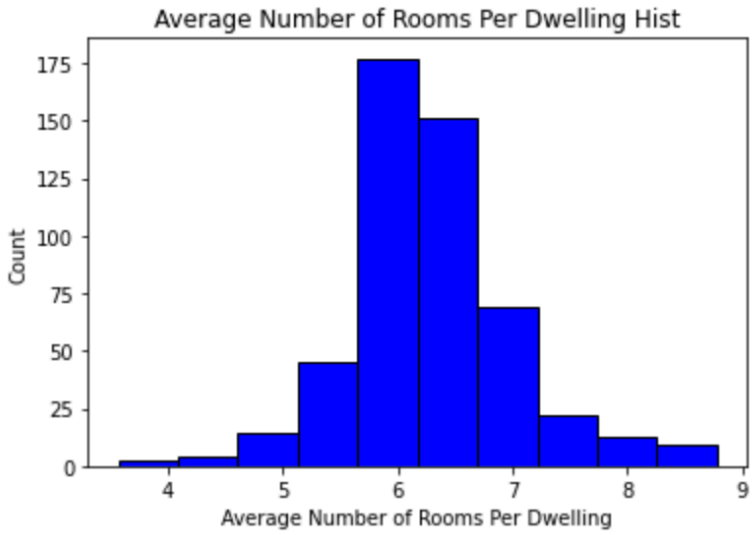
The table above shows some statistical analysis of the data set. It can be seen that most of the attributes are between 0 and 1 and some that are not.



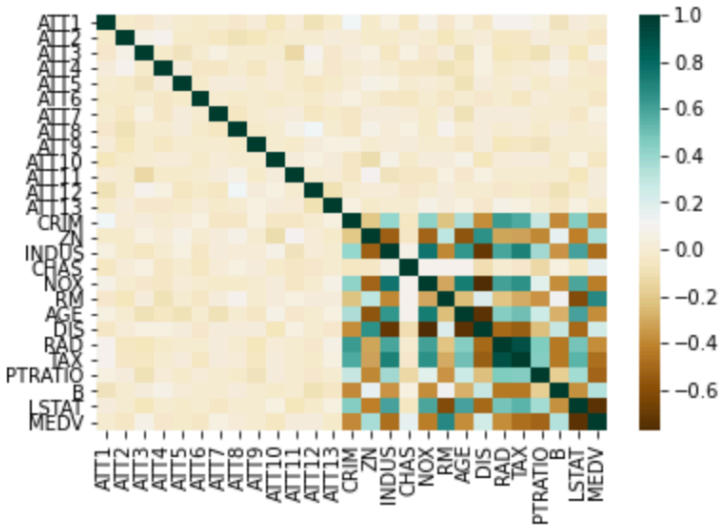
The scatter plot above shows the relationship between some attributes of the dataset.



It can be seen from this QQ plot that there are some outliers as the points that are plotted do not lie in the straight line



The histogram plot shows that the average number of rooms per dwelling has the biggest cluster in between 5.5-7 rooms which may give a clue about house median value house price



From the heat map above, it can be seen that attributes ("ATT1-13" and "CHAS") is not correlated to anything and might be irrelevant to help us run the linear regression model in this dataset.

Following all the exploratory data analysis, we then split the dataset into 26 attributes and 1 target. The target is the Median Value and the attributes are the rest of the attributes in the dataset.

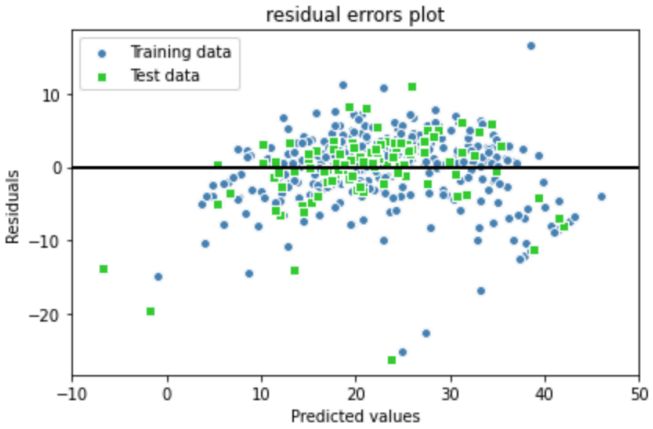
3. **Linear Regression**

After running linear regression with Scikit, the followings results was gathered.

```
Slope: 1.369
Intercept: 37.264

MSE train: 20.613, test: 26.630
R^2 train: 0.763, test: 0.637
```

The residual errors plot for the linear regression model is attached below

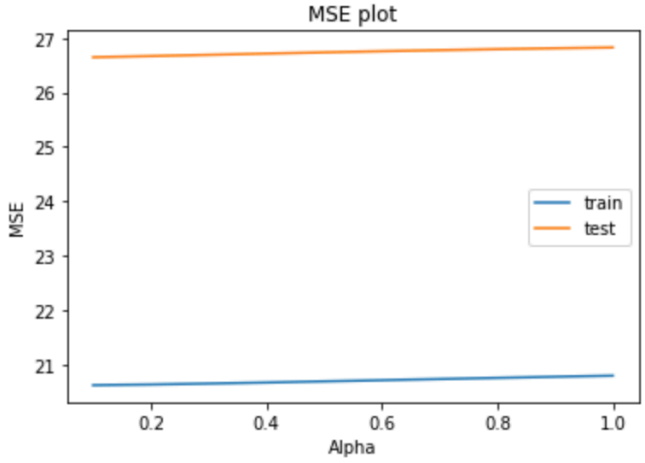


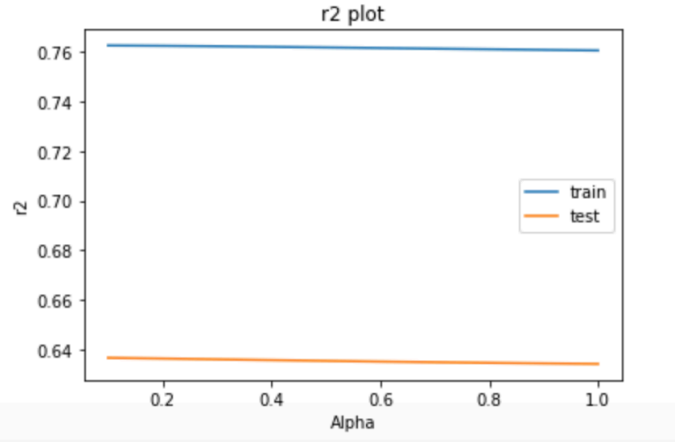
The result below shows the result when 10-fold linear regression model to the dataset

```
[ 0.72739998  0.38778035 -1.44162871  0.625988    0.52145024  0.73970376
  0.40406165 -0.14028481 -0.787782    0.47863144]
Average 10-Fold CV Score: 0.15153198952678143
```

4.

a. Ridge Regression

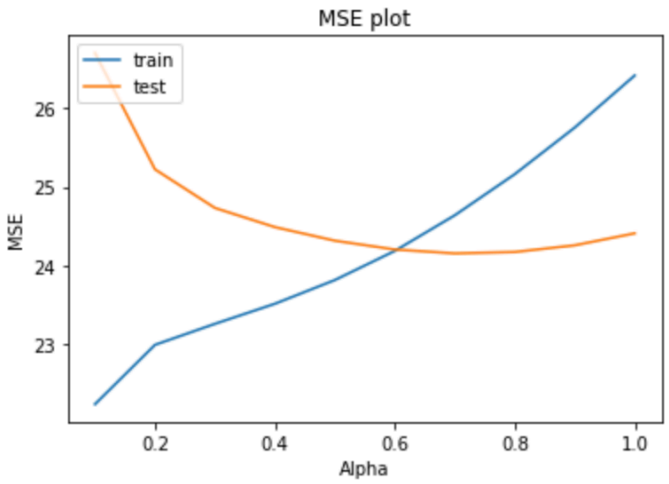


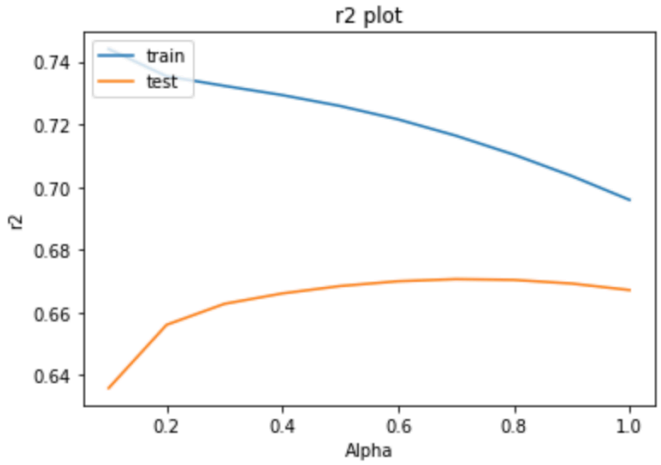


From my analysis of both the MSE plot and r2 plot above, the model runs best when alpha is .1 in MSE and 1 in r2. Thus, I took alpha = .1 and calculated the coefficients of each attribute and got the following result

```
[ 2.88102433e+00 -2.64844380e-01  5.44526750e-01 -1.02903995e-01
 -7.92138103e-01 -6.25099143e-01  6.69322048e-01 -2.70237882e-01
 -1.35944155e-01 -1.00762310e+00 -9.71536144e-01  4.54526414e-01
 -1.50937706e-01 -1.27005165e-01  3.27536567e-02  3.45425757e-02
  2.70828309e+00 -1.57200685e+01  4.48657220e+00 -7.46572094e-03
 -1.45212968e+00  2.54231693e-01 -1.06486381e-02 -8.77099347e-01
  1.31259305e-02 -5.03657959e-01]
Y_test MSE: 26.64726904455334
Y_test r2: 0.6366303783033143
```

b. Lasso Regression





From another analysis that I made from both the MSE plot and r2 plot above, the model runs best when alpha is .7 in MSE and .1 in r2. Thus, I took alpha = .7 and calculated the coefficients of each attribute and got the following result

```
[ 0.          -0.           0.          -0.          -0.          -0.
  0.           0.          -0.          -0.          -0.           0.
 -0.          -0.0864539    0.03087581 -0.           0.          -0.
  2.53457636  0.00580573 -0.80667813  0.22959913 -0.01271975 -0.72093109
  0.01188698 -0.68780861]
Y_test MSE: 24.154355317323876
Y_test r2: 0.6706244478821253
```

5. Conclusion

To conclude, the report paper, I strongly believe that Lasso Regression is the best method to predict the regression of the target. As we can see that a lot of the attribute coefficients are zeros which meant that they're not relevant/correlated to the target and thus it was zeroed out. It means that these attributes may be noises. We can see

that attributes “ATT1-13” and “CHAS” have almost 0 correlations with the other attributes and it can be assumed that these attributes are not needed for the regression model. However, using Lasso regression with ideal alpha = .7, we can see that there are additional attributes such as “NOX” and “RM” are irrelevant to the problem. Thus linear regression is the best method to solve this issue with the ideal alpha of .7 that gives the MSE and r2 errors of the target variable to 24.154355317323876 and 0.6706244478821253 respectively.

6. **Appendix**

   https://github.com/yohanesusanto/IE517_HW4