Michelle Kim
July 16th 2023

CIS 053 Final

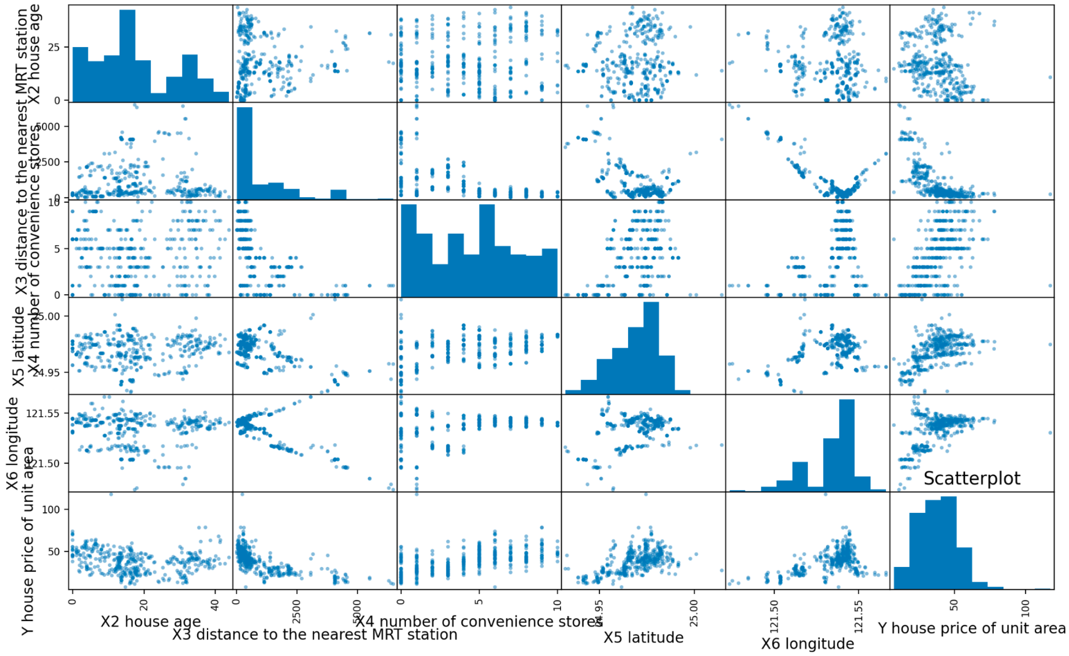
a. EDA

Descriptive Statistics:

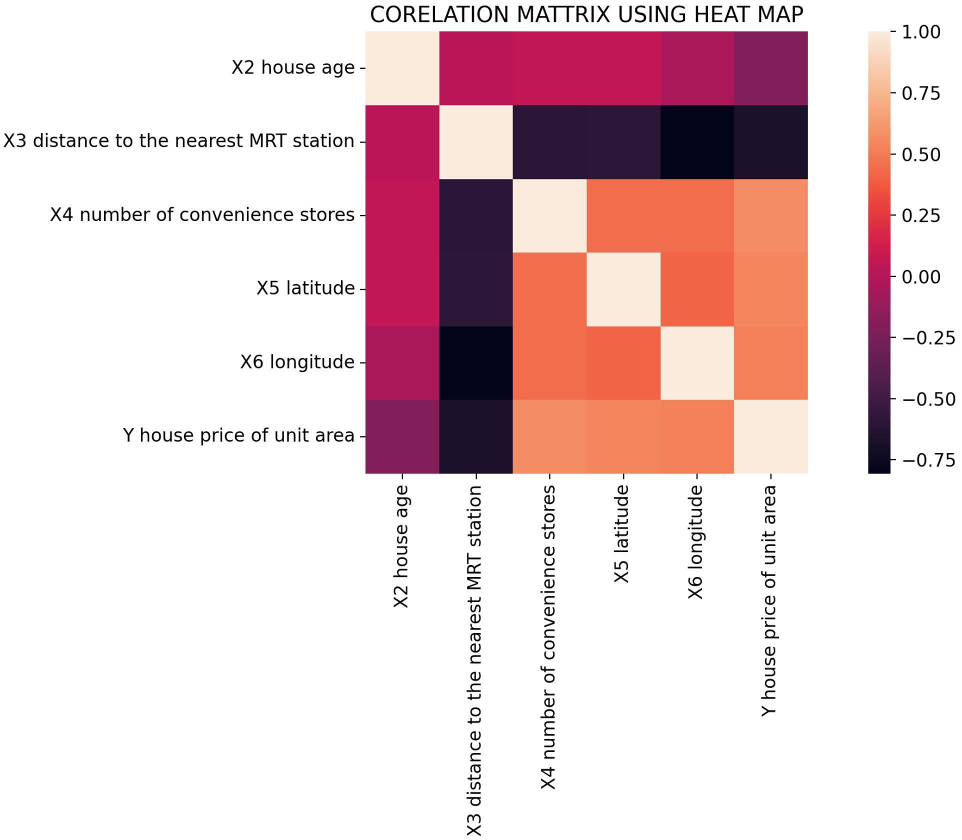
	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores
count	414.0	414.0	414.0
mean	17.7	1083.9	4.1
std	11.4	1262.1	2.9
min	0.0	23.4	0.0
25%	9.0	289.3	1.0
50%	16.1	492.2	4.0
75%	28.1	1454.3	6.0
max	43.8	6488.0	10.0

	X5 latitude	X6 longitude	Y house price of unit area
count	4.1e+02	4.1e+02	414.0
mean	2.5e+01	1.2e+02	38.0
std	1.2e-02	1.5e-02	13.6
min	2.5e+01	1.2e+02	7.6
25%	2.5e+01	1.2e+02	27.7
50%	2.5e+01	1.2e+02	38.5
75%	2.5e+01	1.2e+02	46.6
max	2.5e+01	1.2e+02	117.5

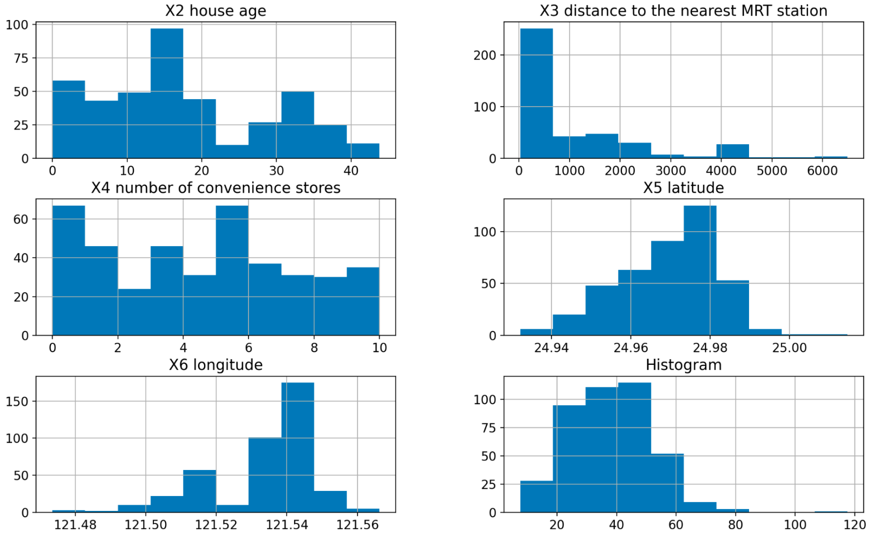
I removed X1 transaction date feature because it had a small range and little relevance to the target feature house price of unit area.



- The scatter plot shows that Y has a roughly linear relationship with all of the features. However, house age(top row) seems to have little linear correlation with other features so it may be a weakly correlated feature.



- - The correlation heat map shows that X3 has a strong negative correlation with every feature and X4, X5, and X6 has a moderately strong positive correlation with every other feature. X2 House age is the most weakly correlated with the other features.

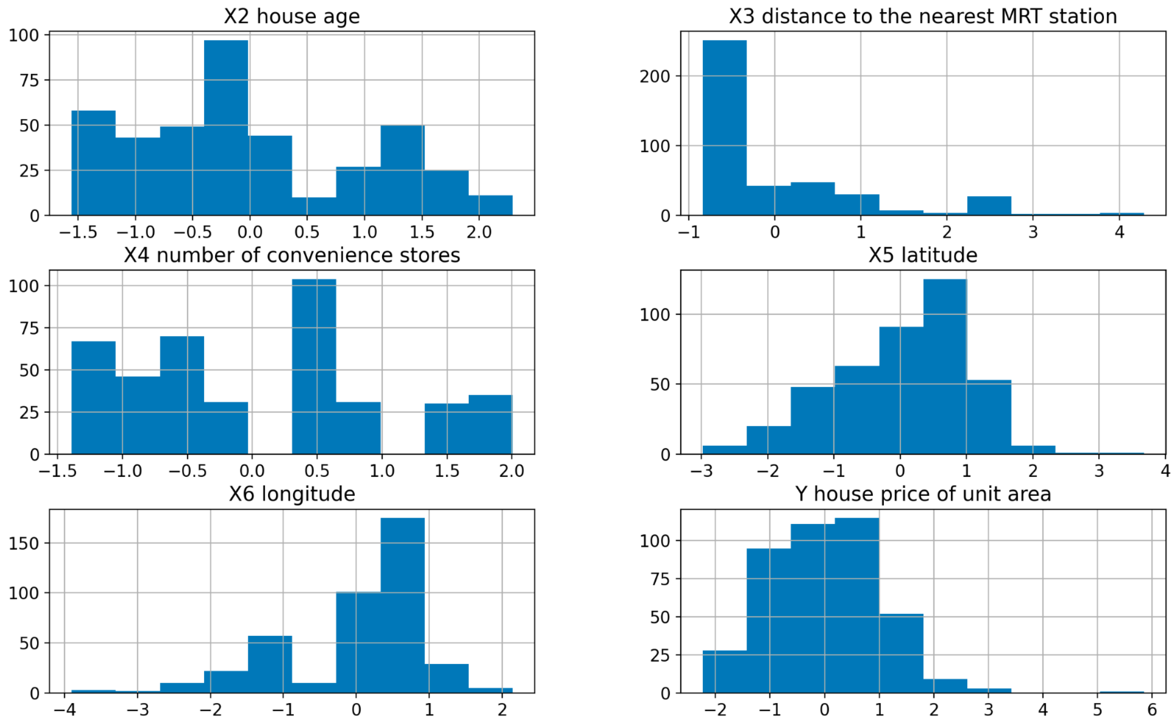


-

- X5 and X6 have mostly Gaussian histograms while X3 has a right skew. X2 and X4 are spread out without a very clear peak.

b. I choose to standardize the data because the data is 1) roughly Gaussian, follows a bell shape (see histograms) and 2) the features have different units of measurement such as age, distance, number, etc.

Standardized Data Histogram:



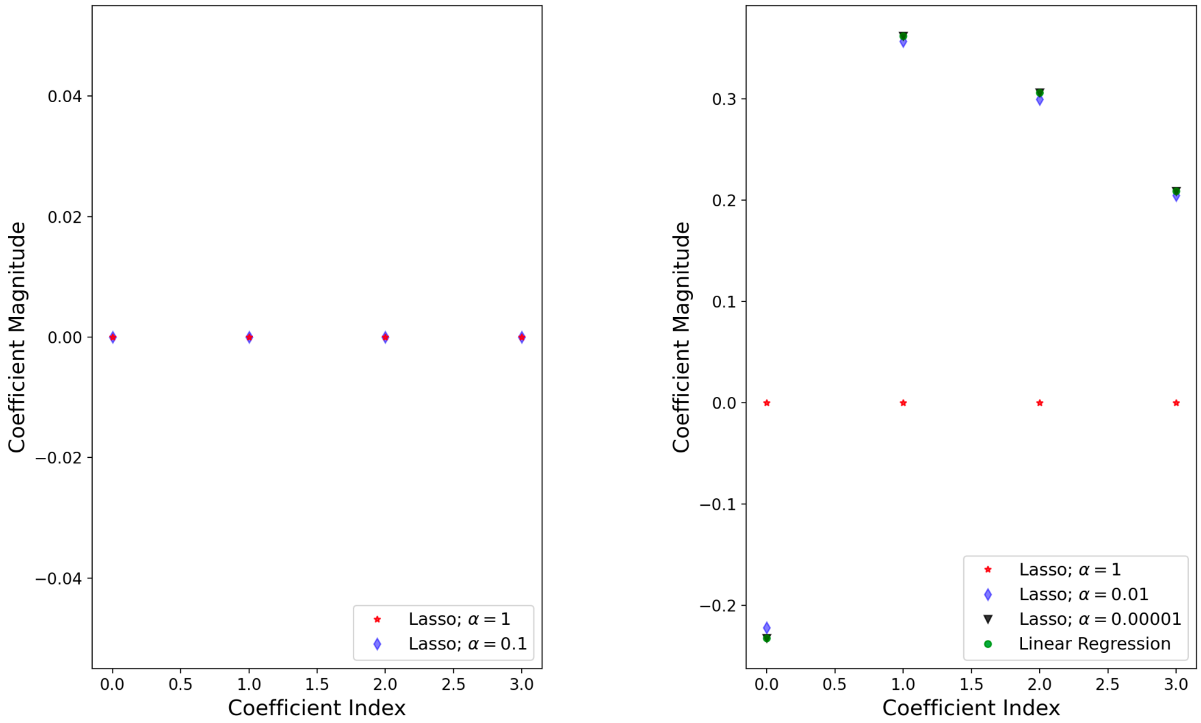
c. MLR Model

```
Num Features: 1
Selected Features: [False False False  True False]
Feature Ranking: [4 5 3 1 2]
Model Score with selected features is:  0.29845095856969295
Num Features: 2
Selected Features: [False False False  True  True]
Feature Ranking: [3 4 2 1 1]
Model Score with selected features is:  0.40529602245610863
Num Features: 3
Selected Features: [False False  True  True  True]
Feature Ranking: [2 3 1 1 1]
Model Score with selected features is:  0.48095106332608073
Num Features: 4
Selected Features: [ True False  True  True  True]
```

```
Feature Ranking: [1 2 1 1 1]
Model Score with selected features is:  0.5347208212325543
Num Features: 5
Selected Features: [ True  True  True  True  True]
Feature Ranking: [1 1 1 1 1]
Model Score with selected features is:  0.5711617064827457
```

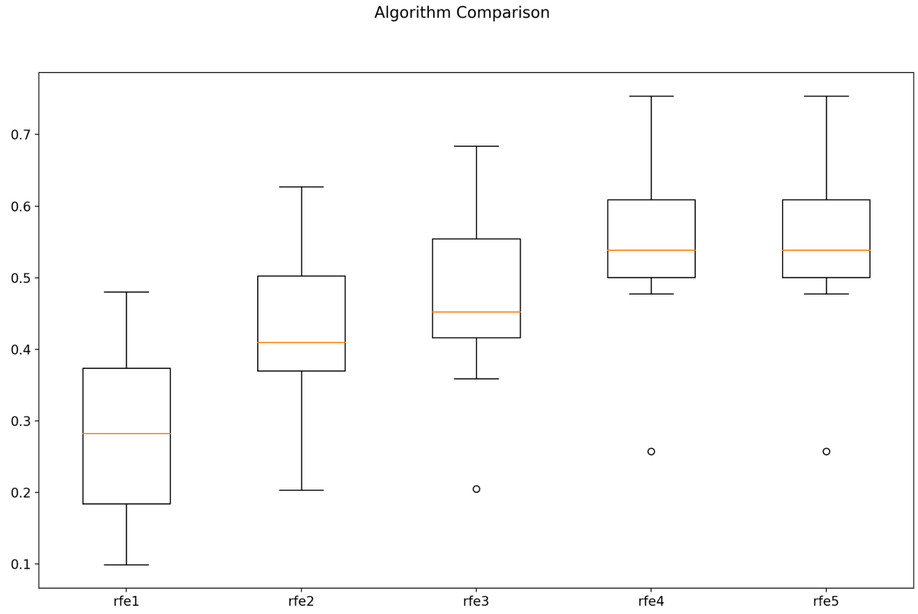
d. regularized version of the regression model and compare with the results of step c

```
linear regression train score: 0.5563145979243255
linear regression test score: 0.6038048023415713
ridge regression train score low alpha: 0.5563145970530934
ridge regression test score low alpha: 0.6038089813453422
ridge regression train score high alpha: 0.5379363261069321
ridge regression test score high alpha: 0.5984294805725647
training score: 0.0
test score:  -0.000634330627290014
number of features used:  0

training score for alpha=0.5: 0.0
test score for alpha =0.5:  -0.000634330627290014
number of features used: for alpha =0.5: 0

training score for alpha=0.01: 0.5554509831034635
test score for alpha =0.01:  0.6066758613370471
number of features used: for alpha =0.01: 4

training score for alpha=0.0001: 0.5563145031313953
test score for alpha =0.0001:  0.6038441485928245
number of features used: for alpha =0.0001: 5

LR training score: 0.5563145979243255
LR test score:  0.6038048023415713
```



- The regularized model score is definitely higher with the LR test score (num features = 5) being 0.603 and the score of the RFE (num features = 5) being 0.571. Thus, I conclude that regularization was effective and necessary for this dataset.

e. Use cross-validaton to compare the results of all models and choose the best model based on test performance measure.



Cross validation shows that num_features = 4 is the best model. It is significantly better than all the other rfe models (rfe 1- 3) or very similar (rfe 5).

F. Justify your answer based on your understanding of how cross-validation works and in the context of bias-variance trade off.

- Cross validation relates to the bias variance trade off because it reduces both bias and variance.
- For example, k-fold cross validation, which I used in this project, divides the data into k parts with 1 part used as the test set and k -1 parts used as the training set. The technique rotates through each part so that each part is used as the test set once and the results of each test are compared.
- Unlike to a technique which divides the data into a train and test set once, the k-fold method allows all of the data to be used as both a train and test set.
- This allows for a more accurate assessment of the models that consequently has lower variance and lower bias.
- The results of the cross validation showed that 4 features is the optimal amount.
- Rfe 4 uses four features and it has a low bias and high variance as a result. The rfe 4 box plot has a high median score indicating a low mean squared error which indicates low bias. Additionally, it has a smaller range indicating lower variability.
- The cross validation box plots confirm that until 4 features the median score significantly increases but adding a 5th feature has little to no effect on the model's performance.

g. Explain if you find any discrepancy between the results in the cross-validation step and steps c and d.

- There were no extreme differences between the cross validation, regularization, and Multiple Linear Regression (MLR) steps.
- However, from the MLR to the regularization step, the model became more accurate because the regularization step achieved a higher score for all models at all numbers of features.
- Furthermore, the cross validation step illustrated that there was no meaningful gain between 4 and 5 features better than the other two steps. This is evident through how the MLR and regularization steps gave a single model score for each number of features, but the cross validation step generated a distribution of model scores (box plot). Thus, the MLR step made there seem to be a slight difference in score between 4 features and 5 features of ~0.036. The cross-validation step showed that 4 features and 5 features had basically the same distribution and were effectively the same.