*I certify that I have <u>personally</u> done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment.* ARM.

1   Part 1: Exploratory Data Analysis

Figure 1 is a correlation matrix between the numerical features found in the dataset. Inspecting the matrix a little closer we can see that the most correlated feature price compared to area and price compared to bathrooms. This makes sense. If there is more total area or more bathrooms in a house chance are the price is also higher. The least correlated feature is the total number of stories in the house compared to the number of parking spots. No correlation between the features makes sense because the number of stories has no effect on the number of parking spots and vice versa. Figure 2 shows the distribution of the categorical variables that are found in the dataset. A trend that is noticed is that many of the houses don't have air conditioning or heated water. It will be interesting to compare the prices of those houses with houses that do have those features. Other features that may affect the prices of housing are whether the house is on the main road or whether the house has guest rooms or not.
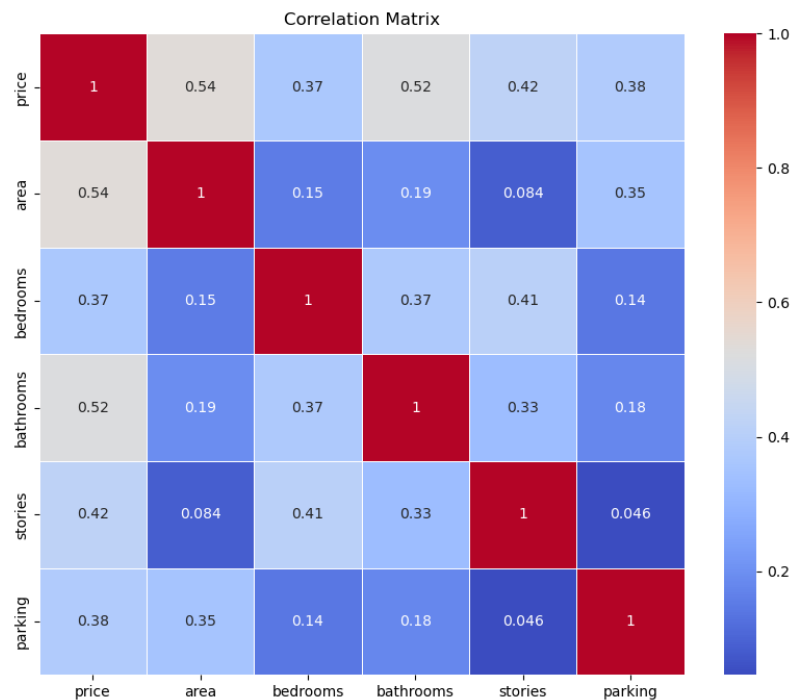

Figure 1. Covariance matrix between features present in the dataset.
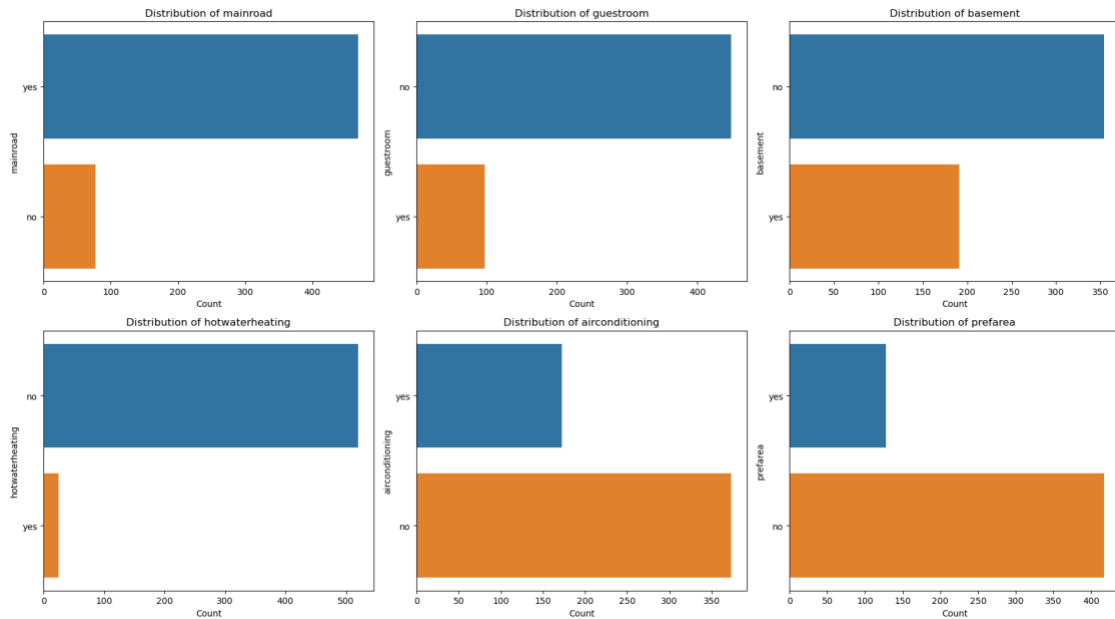
Figure 2. Distribution of categorical variables in the dataset.

## 2   Part 2: Plot of MSE on Validation Data

Figure 3 shows the plot between MSE and the regularization parameter lambda. The plot shows the relationship between the two variables that is when lambda increase, the MSE also increase. This makes sense because having a higher regularization parameter can lead to more bias in the model and underfitting therefore leading to higher values of MSE. The optimal lambda value is found at the minimum of the MSE on the plot which is indicated with the red dotted line, around 0.001.
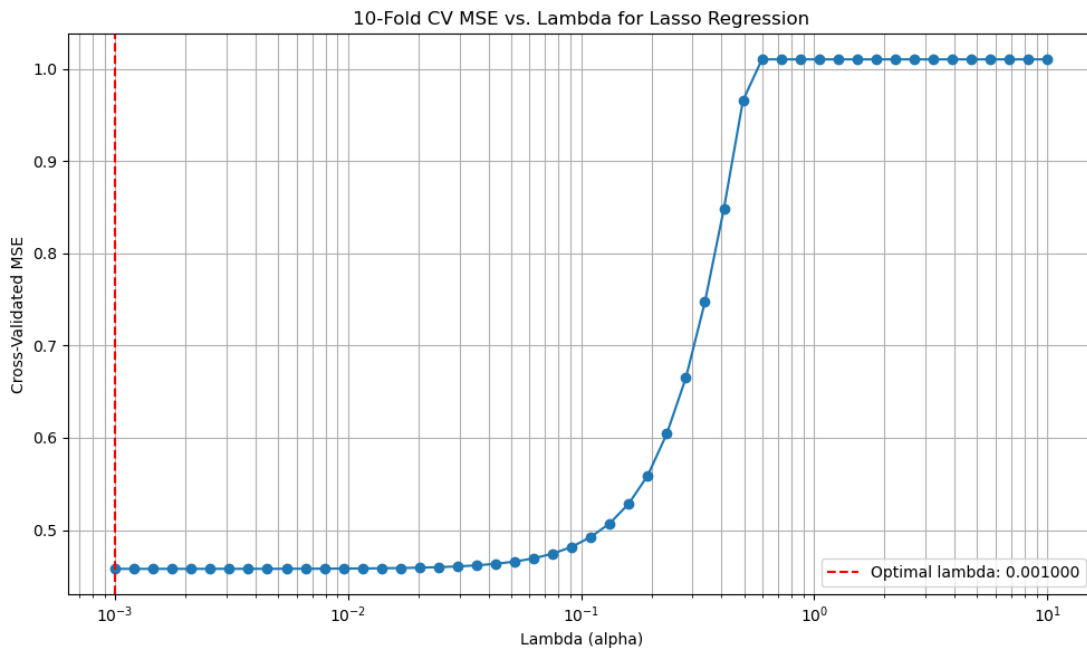
Figure 3. MSE plotted against lambda regularization parameter.

3 Part 3: Plot of Regression Coefficients

Figure 4 plots the coefficients of each feature change against the regularization parameter lambda. As lambda increases the coefficients shrink towards zero, indicating that those features are being regularized. The optimal values of lambda are shown on the left with the block dotted line at around 0.001. Comparing the coefficients of each feature the features with higher coefficients have more relative importance in predicting house prices. We can see that area has the highest coefficient of all the features shown on the graph.
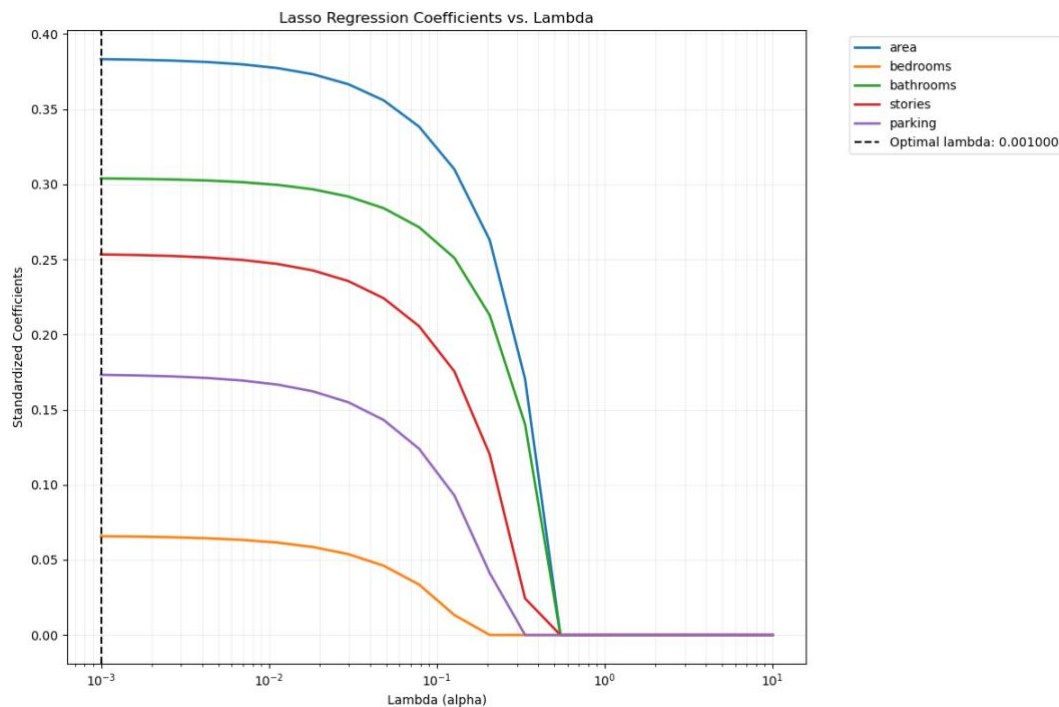
Figure 4. MSE plotted against lambda regularization parameter.

4 Resources used to achieve this goal

Canvas: Homework template

Python Libraries: NumPy, pandas, matplotlib, seaborn, scikit learn.

5 References

"Sklearn.linear_model.Lasso." *Scikit-Learn*, scikit-

learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.