Final Project

Hunter Garrison, Kevin Smith, Reese Madsen, Giulio Martini

2024-04-26

Section 1: Introduction

In this project, we are working with a dataset comprising of the observed prices of houses and corresponding attributes of each house. This dataset is from the database website Kaggle and is comprised of 545 observations of houses in India with 13 variables recorded for each specific house. Below are the attributes of each house recorded in the dataset:

Attribute	Description	Data Type
price	Price of the houses in rupees. (this is our response variable)	Integer
area	Area of a house in square feet	Integer
bedrooms	Number of bedrooms in house	Integer
bathrooms	Number of bathrooms in house	Integer
stories	Number of stories in house	Integer
mainroad	Whether or not house is connected to main road	Boolean
guestroom	Whether or not house has a guest room	Boolean
basement	Whether or not house has a basement	Boolean
hotwaterheating	Whether or not house has a hot-water heater	Boolean
airconditioning	Whether or not house has air conditioning	Boolean
parking	Number of parking spots at house	Integer
prefarea	Whether or not the house is in a preferred area	Boolean
furnishingstatus	Furnishing status of the house (furnished, semi-furnished, unfurnished)	String

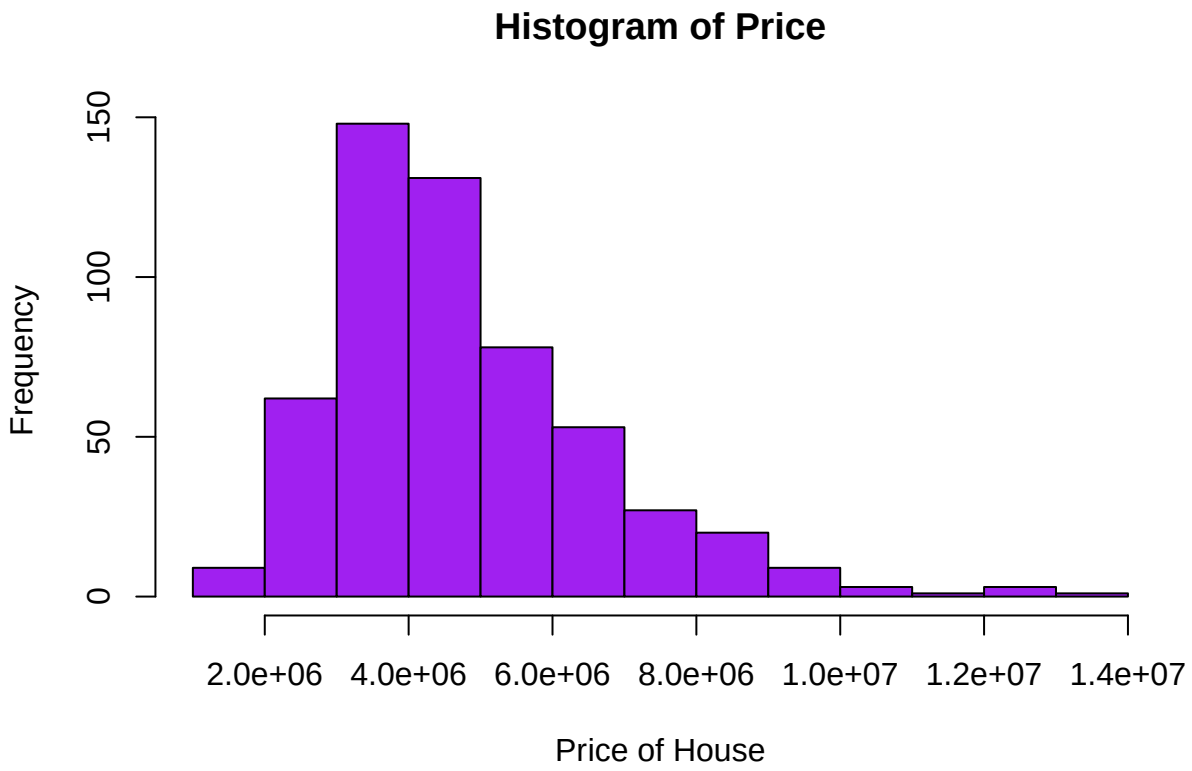
The majority of our predictors in this dataset are of type boolean, with the remainders being integers and one categorical predictor comprising of strings. Below is the first 5 lines of our dataset:

price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement
13300000	7420	4	2	3	yes	no	no
12250000	8960	4	4	4	yes	no	no
12250000	9960	3	2	2	yes	no	yes
12215000	7500	4	2	2	yes	no	yes
11410000	7420	4	1	2	yes	yes	yes

hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
no	yes	2	yes	furnished
no	yes	3	no	furnished
no	no	2	yes	semi-furnished
no	yes	3	yes	furnished
no	yes	2	no	furnished

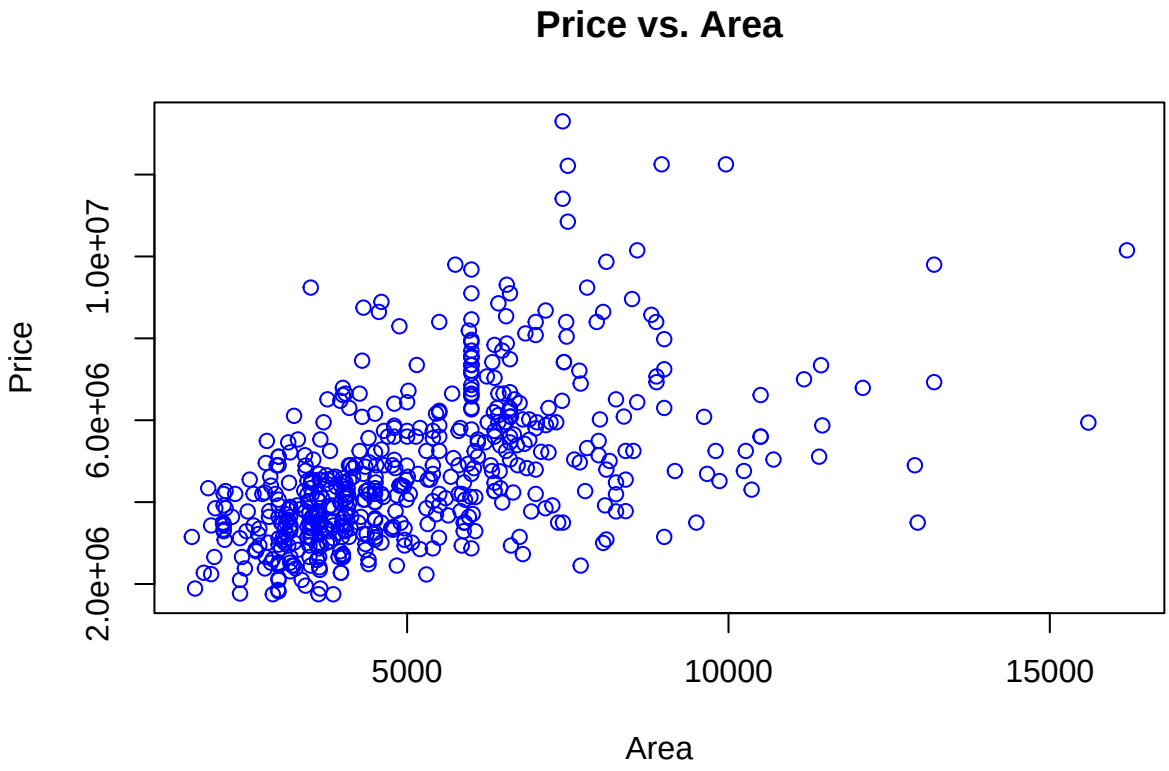
We are attempting to create a linear regression model that best explains the variance in the price of the houses using the predictors in our dataset. In this report, we will explore our data by finding collinearity, outliers, influential points, and by checking error assumption violations. We will also discover which variables are the most important by using variable selection, and test OLS as well as GLS and WLS regressions. After doing this, we will combine all the knowledge we have gained about our data set in order to find the model in which we believe best explains the variance in the price of the houses.

We will now perform some exploratory data analysis to show the distribution of our response variable as well as some key relationships in this dataset. Below is a histogram of our response variable, price, as well as a table showing some key statistical information about the response:




```
##       Min      Q1  Median    Mean      Q3      Max NA_Count
## 1 1750000 3430000 4340000 4766729 5740000 13300000        0
```

From our histogram and table, we see that our data is right skewed with a small amount of houses showing prices much farther than the mean. These houses may get removed as outliers later on in our report. Below, we also show a scatter plot of price vs. area:



We can see a visible linear correlation between the two variables, suggesting that price will be a large factor in the models we will test.

All in all, this dataset is very interesting to us because of how it can teach us the relationships between a house's price and attributes about the house and can help us better understand the housing market when we need to go out and purchase a house for ourselves.

Section 2: Regression Analysis

Collinearity

In our study of the dataset we must ask ourselves whether or not the data is plagued by the issue of collinearity. Do the predictors present a linear relationship, not just with the response, but within each other as well? Collinearity can hurt a model's performance; inflated R^2 scores and diminished p-values are but a few of the problems that arise alongside predictor collinearity... and so, the dataset must be tested for it.

Let us begin by creating a ***correlation heat-map*** for each predictor. This measures the correlation between predictors, and provides us with a visual method of estimating it. Each categorical column is *label encoded* - their class value is turned to a constant integer - and is added to the correlation matrix.

As we can see by the correlation heat-map, correlations between each predictor are realtively small. The strongest correlation is between *stories* and *bedrooms*, with a value of 0.41, swiftly followed by *bedrooms* and *bathrooms* with a coefficient of 0.37, and by *parking* and *area*, with a value of 0.29. These coefficients are decently small, however we will proceed with a more thorough investigation to make sure collinearity is not one of this dataset's problems.

The next step is viewing the dataset's condition numbers, based on a linear model's (that is using the dataset) values.

Eigenvalue	Condition Index
11.4591	1.0000
0.5846	4.4275
0.2140	7.3182
0.1537	8.6333
0.1229	9.6550
0.1016	10.6223
0.0886	11.3721
0.0778	12.1358
0.0622	13.5738
0.0569	14.1939
0.0403	16.8673
0.0309	19.2715
0.0075	38.9969

The condition number is 38.9969, and usually a value above 30 marks the dataset for collinearity. This value of condition number suggests that collinearity should be a problem in this dataset. To further confirm this

hypothesis, we shall check each predictor's VIF value.

Variable	VIF
area	1.325208
bedrooms	1.367503
bathrooms	1.286559
stories	1.478029
mainroad	1.172661
guestroom	1.212687
basement	1.320749
hotwaterheating	1.039293
airconditioning	1.207262
parking	1.211959
prefarea	1.148598
furnishingstatus	1.095641

Only predictors with VIFs above 5 are considered problematic, however we can see that, in this dataset, the highest VIF value is 1.478029. We could stop, however we wish to make the investigation very thorough.

We can then check if the variables are orthogonal. An orthogonal variable is not plagued by multicollinearity, and has a R^2_k of less than 0.3. Below is a table of these values for each predictor:

Variable	R_k^2
area	0.24540168
bedrooms	0.26874003
bathrooms	0.2273304
stories	0.32342322
mainroad	0.14723849
guestroom	0.17538517
basement	0.24285414
hotwaterheating	0.03780781
airconditioning	0.17167947
parking	0.17488936
prefarea	0.12937340
furnishingstatus	0.08729247

We can see that most predictors are orthogonal, with the exception of *stories* which has a value of 0.3234.

If, perchance, there was an issue of collinearity, we could check the individual condition numbers. They would have to be above 30. Each value above 30 would signify a singular instance of a problematic linear dependence between predictors. Luckily, there is none here, as you can see from the values below:

Eigenvalue	Condition Index	intercept	area	bedrooms	bathrooms	stories
11.459	1.000	0.000	0.001	0.000	0.001	0.001
0.585	4.428	0.000	0.003	0.000	0.000	0.001
0.214	7.318	0.000	0.001	0.004	0.036	0.340
0.154	8.633	0.001	0.070	0.000	0.008	0.008
0.123	9.655	0.000	0.467	0.005	0.108	0.001
0.102	10.622	0.000	0.177	0.004	0.472	0.093
0.089	11.372	0.000	0.011	0.003	0.006	0.083
0.078	12.136	0.000	0.096	0.000	0.180	0.123

Eigenvalue	Condition Index	intercept	area	bedrooms	bathrooms	stories
0.062	13.574	0.006	0.001	0.046	0.087	0.017
0.057	14.194	0.004	0.121	0.014	0.030	0.138
0.040	16.867	0.000	0.000	0.751	0.068	0.147
0.031	19.271	0.003	0.053	0.067	0.002	0.041
0.008	38.997	0.986	0.001	0.106	0.001	0.008

mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0.000	0.001	0.001	0.000	0.001	0.002	0.001	0.001
0.000	0.001	0.001	0.000	0.000	0.775	0.001	0.007
0.001	0.012	0.058	0.001	0.013	0.005	0.014	0.034
0.000	0.030	0.046	0.007	0.011	0.093	0.032	0.423
0.004	0.033	0.128	0.001	0.018	0.052	0.029	0.014
0.001	0.020	0.007	0.003	0.156	0.040	0.058	0.004
0.002	0.029	0.000	0.003	0.475	0.005	0.419	0.013
0.002	0.271	0.001	0.009	0.233	0.000	0.265	0.006
0.027	0.363	0.047	0.157	0.007	0.014	0.097	0.205
0.028	0.184	0.586	0.108	0.005	0.000	0.021	0.105
0.091	0.040	0.107	0.011	0.002	0.000	0.012	0.000
0.519	0.001	0.010	0.404	0.020	0.001	0.041	0.003
0.325	0.015	0.009	0.296	0.060	0.014	0.010	0.184

With all the information provided, we then conclude, with confidence, that multicollinearity is not a problem of this dataset.

Variable Selection

We now move on to selecting variables to use in our final model. We will perfrom this using forward, backward, and step-wise selection, using both AIC and BIC, and reporting the LOO-CV RMSE and adjusted R^2 for each model to determine which model is the best.

Model	LOO-CV RMSE	Adjusted R^2
Backwards AIC	1087331	0.6740117
Backwards BIC	1090310	0.6706188
Forwards AIC	1087331	0.6740117
Forwards BIC	1090310	0.6706188
Step-wise AIC	1087331	0.6740117
Step-wise BIC	1090310	0.6706188

After finding all of the values for LOO-CV RMSE and adjusted R^2 for each model, we see that the best models are the backwards, forwards, and step-wise models that used AIC as their metric. All three of these models are the full model with no variables removed. Therefore, we will stick with the model containing all 12 of the original predictors and move on to model diagnostics.

Model Diagnostics

In order to remove problematic data points from the data set, we must perform a series of tests to identify them. To begin, we will examine our error assumptions to see if any have been violated using graphical methods and hypothesis tests.

Constant Variance Assumption To verify that our model follows the constant variance assumption, we will be using the Breush-Pagan test at a significance level of $\alpha = .05$. Our null and alternative hypotheses are H_0: Homoscedastic errors and H_1: Heteroscedastic errors.

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 68.416, df = 13, p-value = 1.569e-09
```

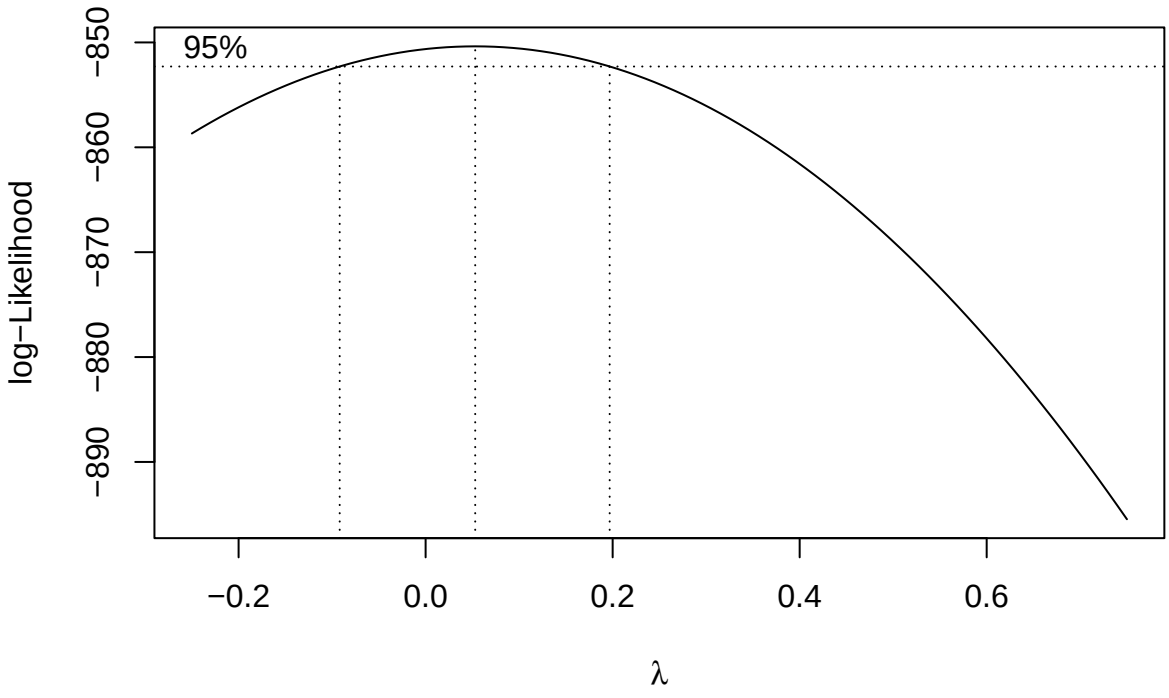
The value of our test-statistic is 68.416 and our p-value is $1.569 \times e^{-9}$. We reject the null hypothesis at the $\alpha = .05$ significance level and conclude that constant variance assumption is violated.

Normality Assumption In order to check the normality assumption, we will be performing the Shapiro-Wilk test at the $\alpha = .05$ significance level. Our null and alternative hypotheses are H_0: The errors are normally distributed and H_1: The errors are not normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.95399, p-value = 5.31e-12
```

The value of our test-statistic is .95399 and our p-value is $5.31 \times e^{-12}$. We reject the null hypothesis at the $\alpha = .05$ significance level and conclude that errors are not normally distributed.

In order to try to correct for this, we will be using the Box-Cox method.



```
## [1] 0.0530303
```

According to the plot, and the print out, we can tell that $\hat{\lambda} = .05303$. So we will be performing an OLS regression with price$^{.05303}$ as the response and performing the Shapiro-Wilk test at the $\alpha = .05$ significance level.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_bc)
## W = 0.99525, p-value = 0.0939
```

The value of our test-statistic is 0.9953 and our p-value is 0.0939. We accept the null hypothesis at the $\alpha = .05$ significance level and conclude that the errors are now normally distributed.

We will also check the RMSE and the percent variation in price explained by both models.

Model	RMSE	% Variation
Standard Model	1054129	68
Transformed Model	1041366	69

Since our errors are now normally distributed, we will use the transformed model instead of the standard model.

Linearity Assumption It is clear that there is some linear relationship between the response and the numerical predictors (area, bathrooms, bedrooms, stories, parking).

Highly Influential Points

```
## [1] 35
```

To check for highly influential points, we will be checking the cooks distances of the data. Using this distance, we find that there are 35 highly influential points. We will check to see if removing these points from the model will in any way correct our models assumption violations. First we will check the constant variance assumption at the $\alpha = .05$ significance level.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_fix
## BP = 31.917, df = 13, p-value = 0.002471
```

The value of our test-statistic is 31.917 and our p-value is .00247. We reject the null hypothesis at the $\alpha = .05$ significance level and conclude that constant variance assumption is violated.

Next we will check our constant variance assumption at the $\alpha = .05$ significance level.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_fix)
## W = 0.99865, p-value = 0.9681
```

The value of our test-statistic is .9987 and our p-value is .9681. We accept the null hypothesis at the $\alpha = .05$ significance level and conclude that errors are still normally distributed.

We will also check the R^2 of both models,

```
## [1] 0.6818018
```

```
## [1] 0.7651252
```

Model	R^2
Standard Model	.682
Model with Influential Points Removed	.765

It is clear that the Model with Influential Points removed is the preferred model as it explains almost 8% more of the observed variability in price with the predictors.

Outliers To check for outliers we will be using the studentized residuals at the $\alpha = .05$ significance level.

```
## [1] 0
```

We can see that there are zero outliers, so we move on. Our model now has normally distributed errors, however the equal variance assumption is violated. Therefore, we will look to use a regression method other than ordinary least squares. Since our errors are not correlated (heteroscedastic) we will use weighted least squares.

Weighted Least Squares (WLS)

We now begin performing the weighted least squares regression. We create a model using the absolute values of the residual and then calculate weights as $\frac{1}{fittedvalues^2}$. We then create our model with these weights.

After creating our model, we will perform a t-test at the 5% significance for each parameter for the OLS model and the WLS model.

Below are the estimates and p-values for the OLS regression:

Coefficient	Estimate	P-value
Intercept	2.14	0.000
area	5.91e-06	2.69e-28
bedrooms	0.00307	0.0361
bathrooms	0.0203	1.74e-19
stories	0.0111	6.06e-17
mainroadyes	0.0111	0.000110
guestroomyes	0.00833	0.00206
basementyes	0.00896	0.000086
hotwaterheatingyes	0.0182	0.000538
airconditioningyes	0.0215	5.06e-21
parking	0.00519	0.0000168
prefareayes	0.0158	4.39e-11
furnishingstatussemi-furnished	0.000412	0.862
furnishingstatusunfurnished	-0.0135	1.53e-07

Coefficient	Estimate	P-value

Below are the estimates and p-values for the WLS regression:

Coefficient	Estimate	P-value
Intercept	2.03e+05	0.315
area	236	4.09e-29
bedrooms	5.94e+04	0.312
bathrooms	9.73e+05	1.90e-31
stories	4.46e+05	1.83e-17
mainroadyes	4.14e+05	0.0000875
guestroomyes	3.14e+05	0.00219
basementyes	2.02e+05	0.0186
hotwaterheatingyes	8.42e+05	5.62e-07
airconditioningyes	8.47e+05	1.69e-22
parking	2.31e+05	1.50e-06
prefareayes	6.59e+05	1.33e-13
furnishingstatussemi-furnished	-15.5e+03	0.860
furnishingstatusunfurnished	-3.71e+05	0.000411

For OLS every variable is significant except furnishing status semi-furnished. For WLS every variable is significant except bedrooms and furnishing status semi-furnished.

```
summary(model_fix)$r.squared
```

```
## [1] 0.7651252
```

```
summary(model_wls)$r.squared
```

```
## [1] 0.7780183
```

Model	R^2
OLS	0.765
WLS	0.778

After finding the R^2 score of each regression, we see that WLS has a better score.