Melbourne Housing Prediction Final Report

Leonardo Guzman, Tashin Azad, Kathryn Chen, Lawrence Wu

Spring 2023

Exploration of the data

We are given the task of predicting the **Price** of housing in Melbourne. Using the Melbourne Housing data set, our goal is to select relevant explanatory variables to create a model to predict the price of housing. As such, we present the Melbourne data set.

Relevant quantitative variables:

1. Rooms, which lists the number of rooms in the house
2. Bathroom, which lists the number of bathrooms in the house
3. Distance, which lists the distance of the property from the Central Business District
4. Price, which lists the price of the property in dollars
5. Car, which lists the number of car spots on the property
6. Landsize, which lists the size of the land the property is on
7. Building area, which lists the area of the building

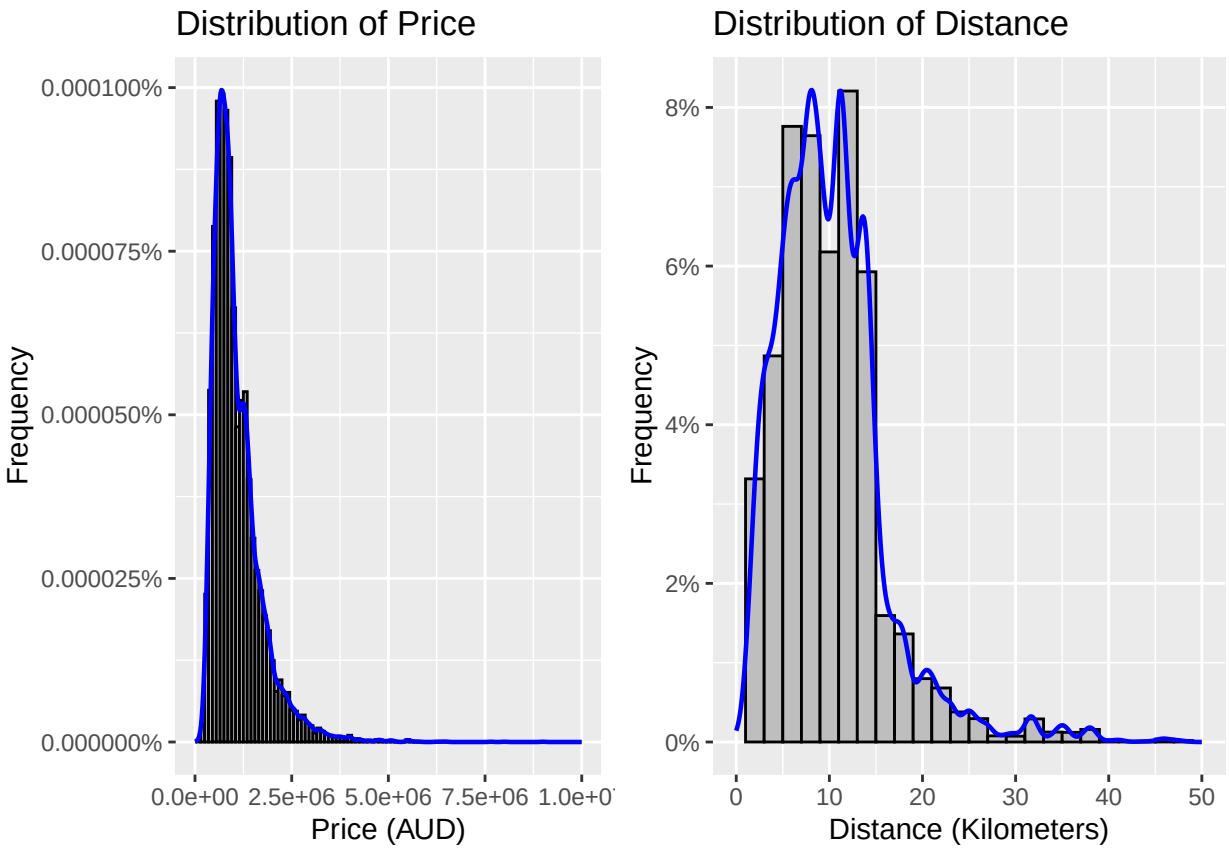
Relevant categorical variables:

1. Regionname, which lists the general region the property is located in
2. Type, which lists the property type

Observational unit: one house or apartment in Melbourne

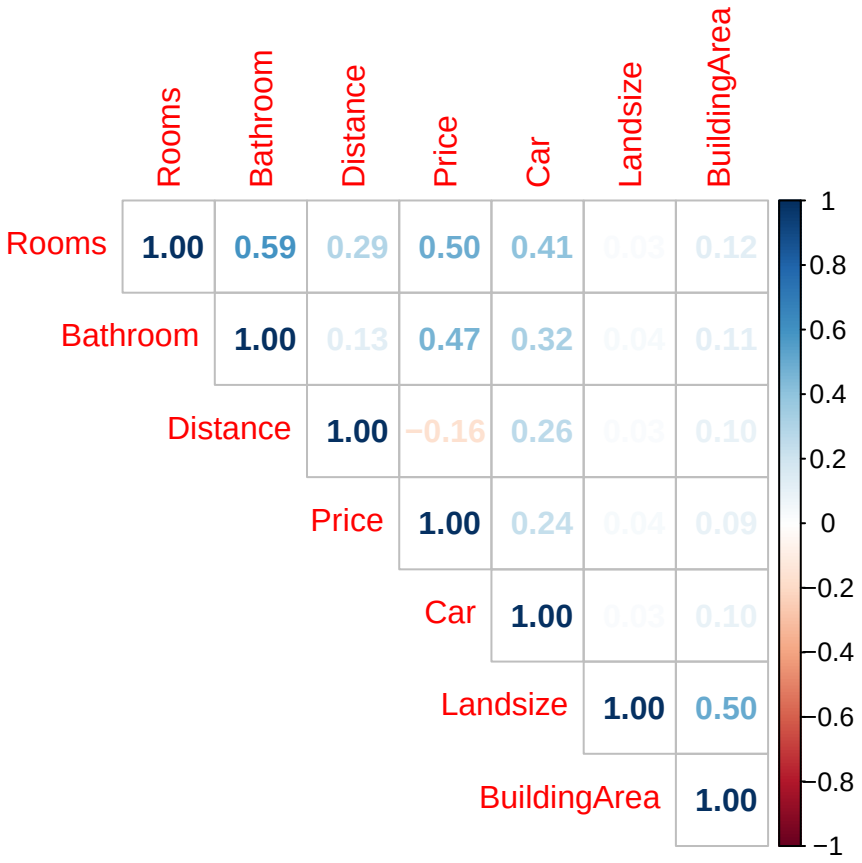
Columns: 21
Rows: 13,580

The following visualizations demonstrated below provide an important picture of our data.



The price histogram shows a distribution that is positively skewed or skewed right for the price in dollars of the houses in Melbourne. There is a long tail on the right, meaning few houses are on the further, pricier end. Majority of the prices are clustered around the lower end of the price range.

The distance histogram indicates that most the sampled housing units are concentrated about 10 Kilometers from the Central Business District. However, observe that there is quite a long tail as well.



This is a correlation plot of the correlations between many of the quantitative variables. The blue indicates a positive correlation while the red indicates a negative correlation and the opacity indicates the strength of the correlation (darker is stronger, correlation coefficients are also shown).

Here is a quick overview of our data.

Table 1: Data summary

Name	melbourne_housing
Number of rows	13580
Number of columns	21
Column type frequency:	
character	8
numeric	13
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Suburb	0	1	3	18	0	314	0
Address	0	1	8	27	0	13378	0
Type	0	1	1	1	0	3	0
Method	0	1	1	2	0	5	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SellerG	0	1	1	23	0	268	0
Date	0	1	9	10	0	58	0
CouncilArea	0	1	0	17	1369	34	0
Regionname	0	1	16	26	0	8	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Rooms	0	1.00	2.94	0.96	1.00	2.00	3.0	3.00	10.00	
Price	0	1.00	1075684.08	639310.72	85000.00	650000.00	903000.0	1330000.00	9000000.00	
Distance	0	1.00	10.14	5.87	0.00	6.10	9.2	13.00	48.10	
Postcode	0	1.00	3105.30	90.68	3000.00	3044.00	3084.0	3148.00	3977.00	
Bedroom2	0	1.00	2.91	0.97	0.00	2.00	3.0	3.00	20.00	
Bathroom	0	1.00	1.53	0.69	0.00	1.00	1.0	2.00	8.00	
Car	62	1.00	1.61	0.96	0.00	1.00	2.0	2.00	10.00	
Landsize	0	1.00	558.42	3990.67	0.00	177.00	440.0	651.00	433014.00	
BuildingArea	6450	0.53	151.97	541.01	0.00	93.00	126.0	174.00	44515.00	
YearBuilt	5375	0.60	1964.68	37.27	1196.00	1940.00	1970.0	1999.00	2018.00	
Lattitude	0	1.00	-37.81	0.08	-38.18	-37.86	-37.8	-37.76	-37.41	
Longtitude	0	1.00	145.00	0.10	144.43	144.93	145.0	145.06	145.53	
Propertycount	0	1.00	7454.42	4378.58	249.00	4380.00	6555.0	10331.00	21650.00	

As shown in the skim function above, we have a significant amount of missing or empty values from the columns BuildingArea, and YearBuilt. We should point out that the standard deviation for price is very high. We reasonably expect that there is variation of price dependent on geographic location and proximity to the central business district (this data is captured by the distance statistic). Our findings show that there is a negative correlation between distance from CBD and the price of housing.

A couple interesting findings and comments on the dataset: When comparing the land size of the property with the types of property, the type "h" has the highest average land size as well as the highest average price. Type "t" is in the middle for average price, yet it interestingly has the lowest average land size. Finally, type "u" had the lowest price yet was in the middle for average land size.

Overall there were a few things of interest, first is that the vast majority of the missing observations seem to be concentrated with BuildingArea and YearBuilt. Another is that all of the distribution of the variables we analyzed seem to be skewed to the right which gives us some indication of the direction of the outliers in that they lie further to the right. Although there are a lot of missing observations, the fact that they are limited to only two of the variables is promising in that we got a representative sample of the population.

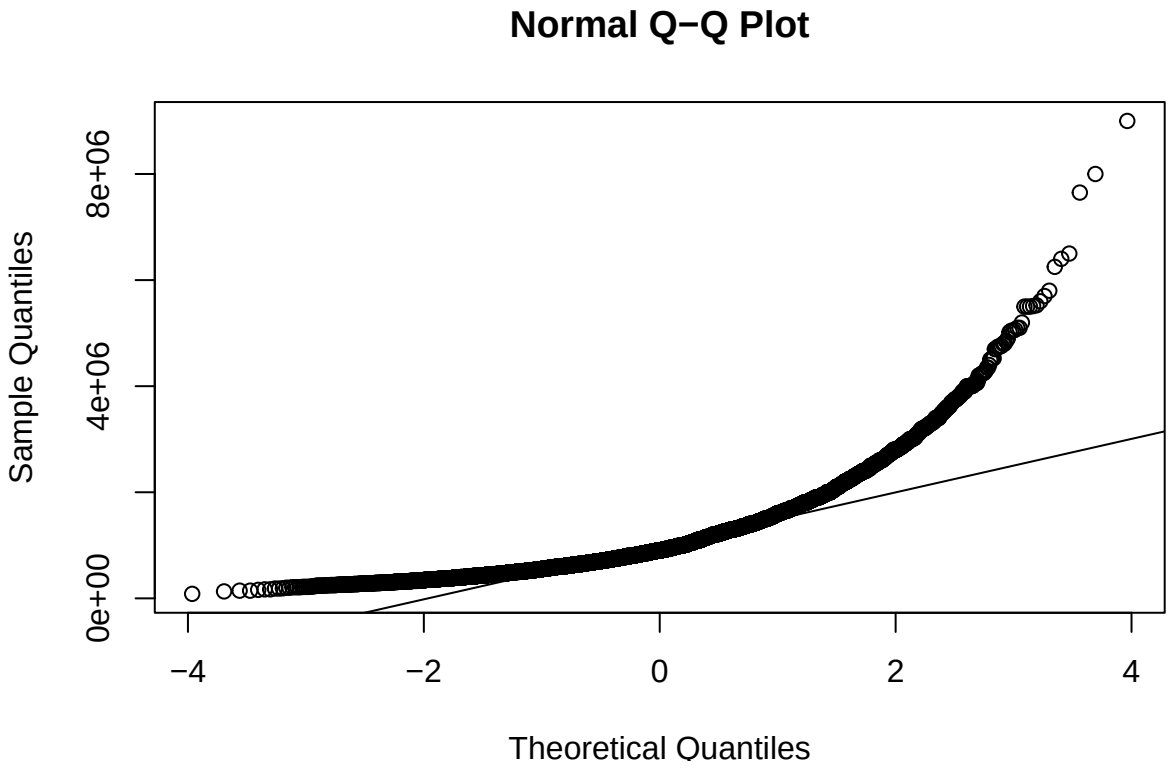
First Model: Simple Linear Regression

Having done our basic analysis of the variables in the previous section, we now turn out attention to checking the necesary assumptions in order to create a technically sound model. Recall that we are attempting to model price and a function of the other variables in the datset. A reminder of the dataset citation:

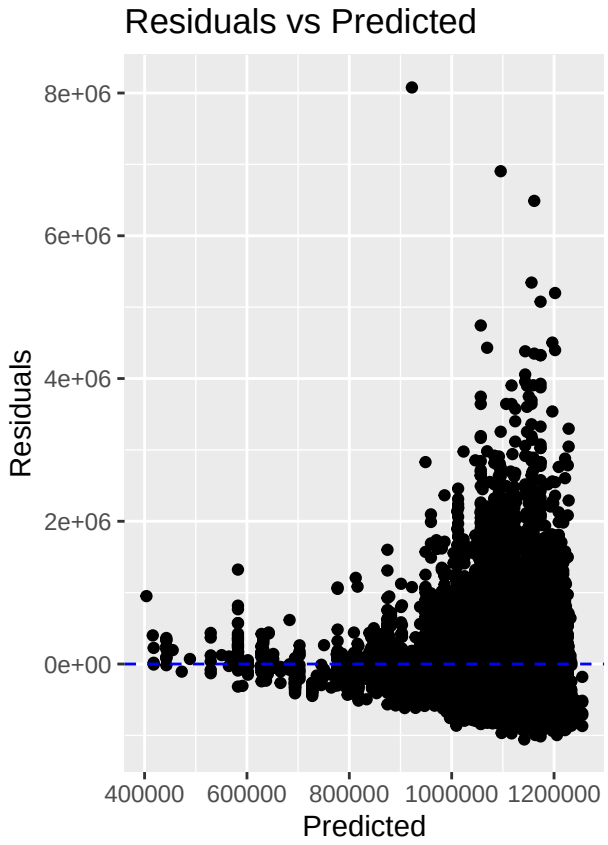
Becker, D. (2017, September). Melbourne Housing Snapshot, Version 5. Retrieved May 11, 2023 from https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot.

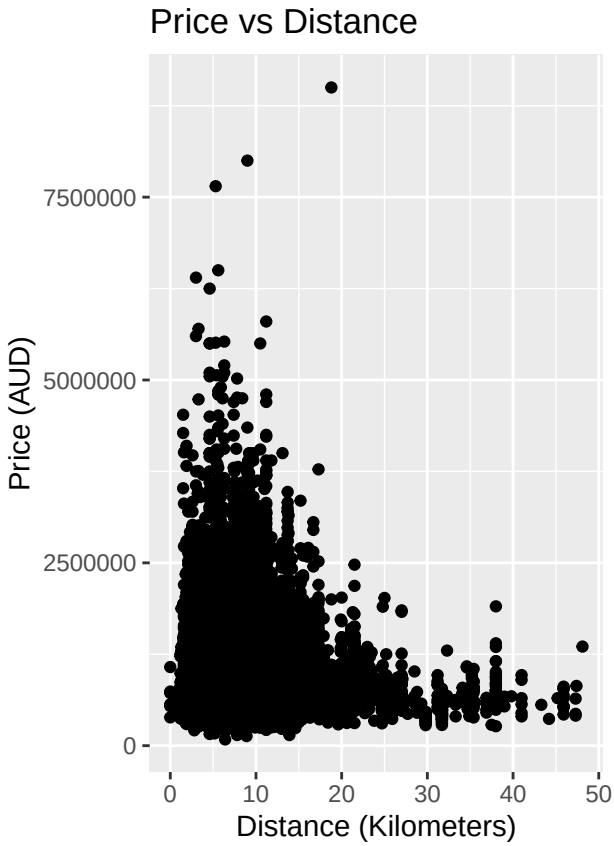
Working with the Melbourne Housing data set, we are specifically interested in the relationship between price of housing and distance to the Central Financial District. We intuitively expect that price is negatively correlated with the distance, as proximity to the CBD is a desirable trait for housing. We will consider a linear model for our data and apply a logarithmic transformation to the price. To test our speculation, we consider the following hypothesis $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$.

First, we check the necessary assumptions on a non-transformed linear model.



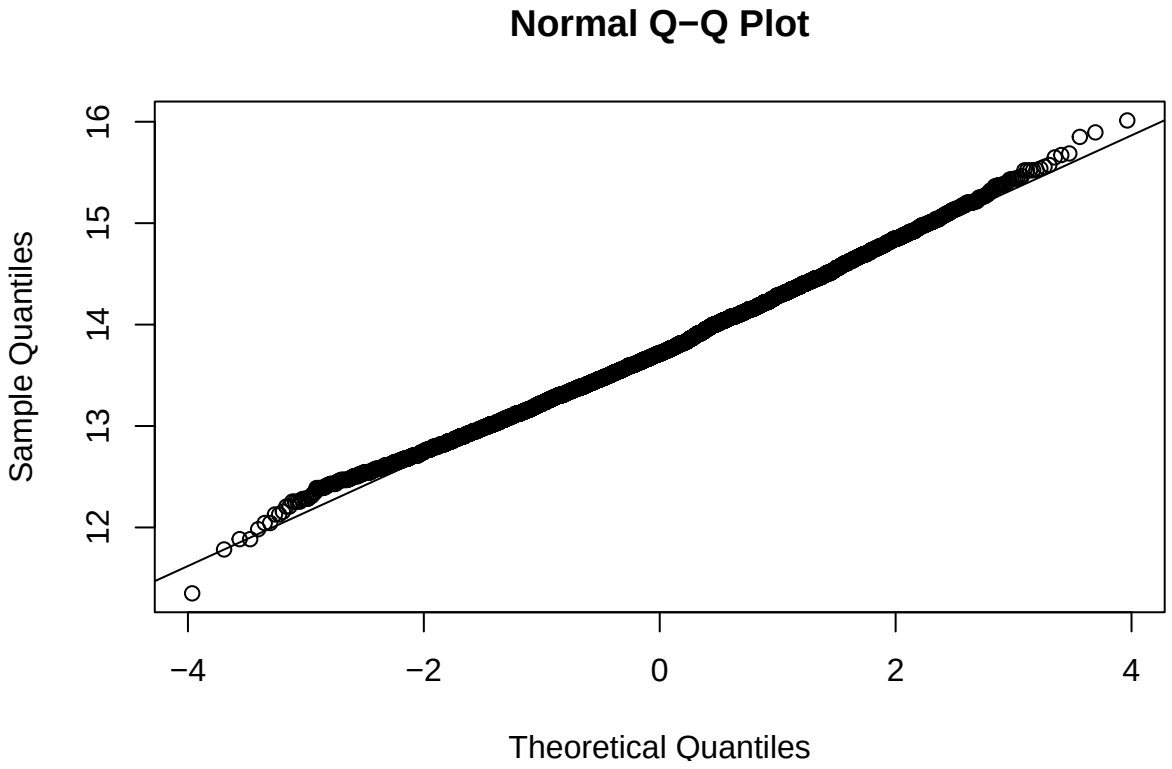
As demonstrated above, the qq plot indicates the simple linear model is not very normal.



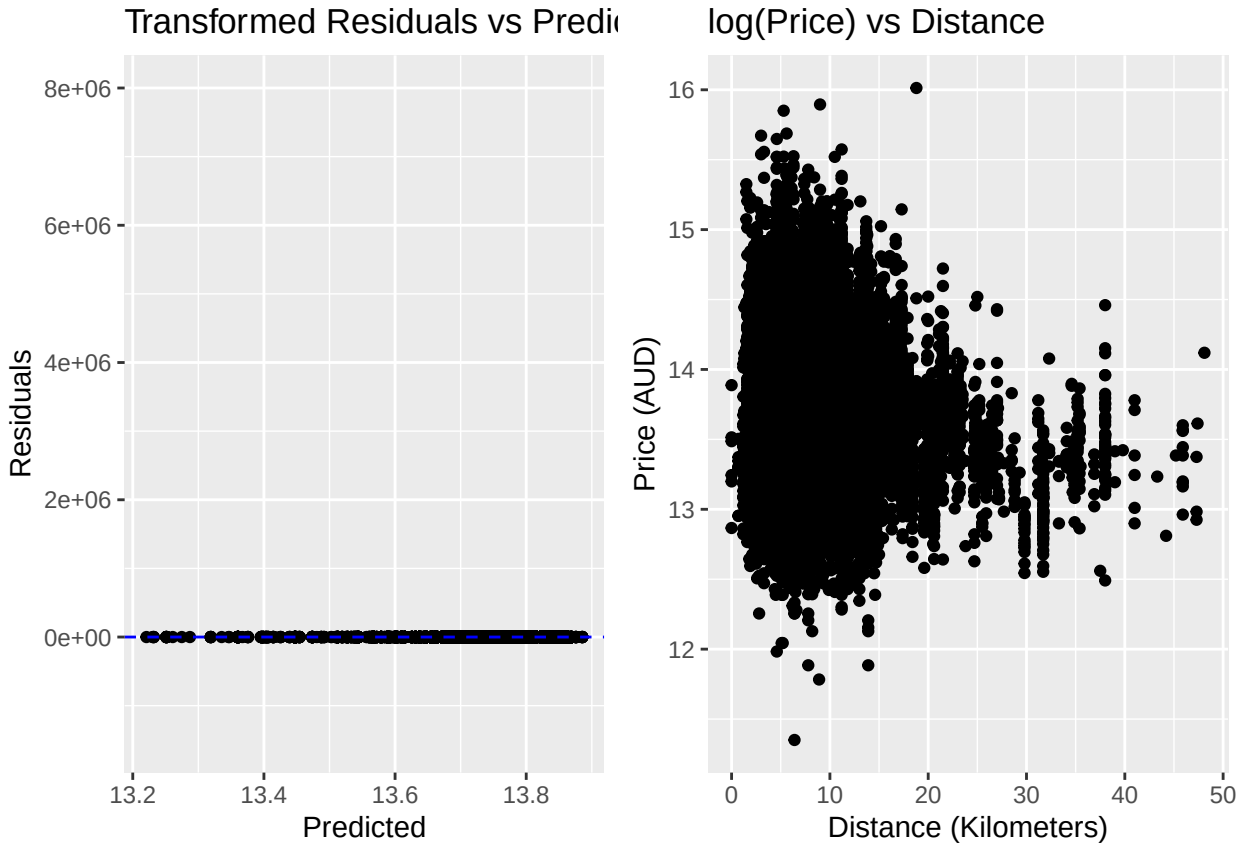


The residual plot for the model indicates both a lack of linearity and constant variance. This is further confirmed by plotting the distance and the house prices themselves. For this reason, we concluded that we should perform some kind of transformation on the variables.

Having tested several different transformations of our data, we ultimately settle with the logarithmic transformation. Consider the following qqplot.



In comparison to the non-transformed plot, the qqplot of the logarithmic transformation appears to be much more normal. With this promising result at hand, we plot our transformed linear model.



Besides the relatively normal qq plot, the residuals for this model were much more evenly distributed on both sides of the x-axis, which indicates linearity. While the residuals still did not have constant variance, this variation of the model seemed like a much better fit for the data. This is especially shown when the residual plot is scaled to the same y-axis values as the previous residual plot, which shows that compared to the first simple linear model, the residuals of this model are much smaller.

We further assess the fit of the model.

```
## # A tibble: 6 x 2
##   Distance  pred
##      <dbl> <dbl>
## 1      0    13.9
## 2      0.7  13.9
## 3      1.2  13.9
## 4      1.3  13.9
## 5      1.5  13.9
## 6      1.6  13.9
```

```
##   Distance     pred   ci.fit   ci.lwr   ci.upr
## 1      0.0 13.88546 13.88546 13.86798 13.90293
## 2      0.7 13.87578 13.87578 13.85920 13.89236
## 3      1.2 13.86888 13.86888 13.85292 13.88483
## 4      1.3 13.86749 13.86749 13.85167 13.88332
## 5      1.5 13.86473 13.86473 13.84915 13.88031
## 6      1.6 13.86335 13.86335 13.84789 13.87881
```

The first table shows the transformed model's predicted log(Price) values given specific distance values. The second table also shows the lower and upper bound of the 95% prediction intervals for each predicted value. As the log(Price) values are around 13.9 when distance is low, this is relatively expected.

```
##             R^2 Transformed R^2
##      0.02641335      0.02369682
```

As shown by the summary results, the R^2 of both models are very low. The simple linear models are likely not good predictors of price on their own.

Using our transformed simple linear model, we calculate the 95% confidence interval for β_1.

```
##                2.5 %      97.5 %
## Distance -0.01530965 -0.01232576
```

```
## (Intercept)    Distance
## 13.90293293 -0.01232576
```

```
## (Intercept)    Distance
## 13.86798019 -0.01530965
```

```
## [1] -0.0138177
```

The 95% confidence interval estimation for β_1: (-0.01677631 -0.01253827)

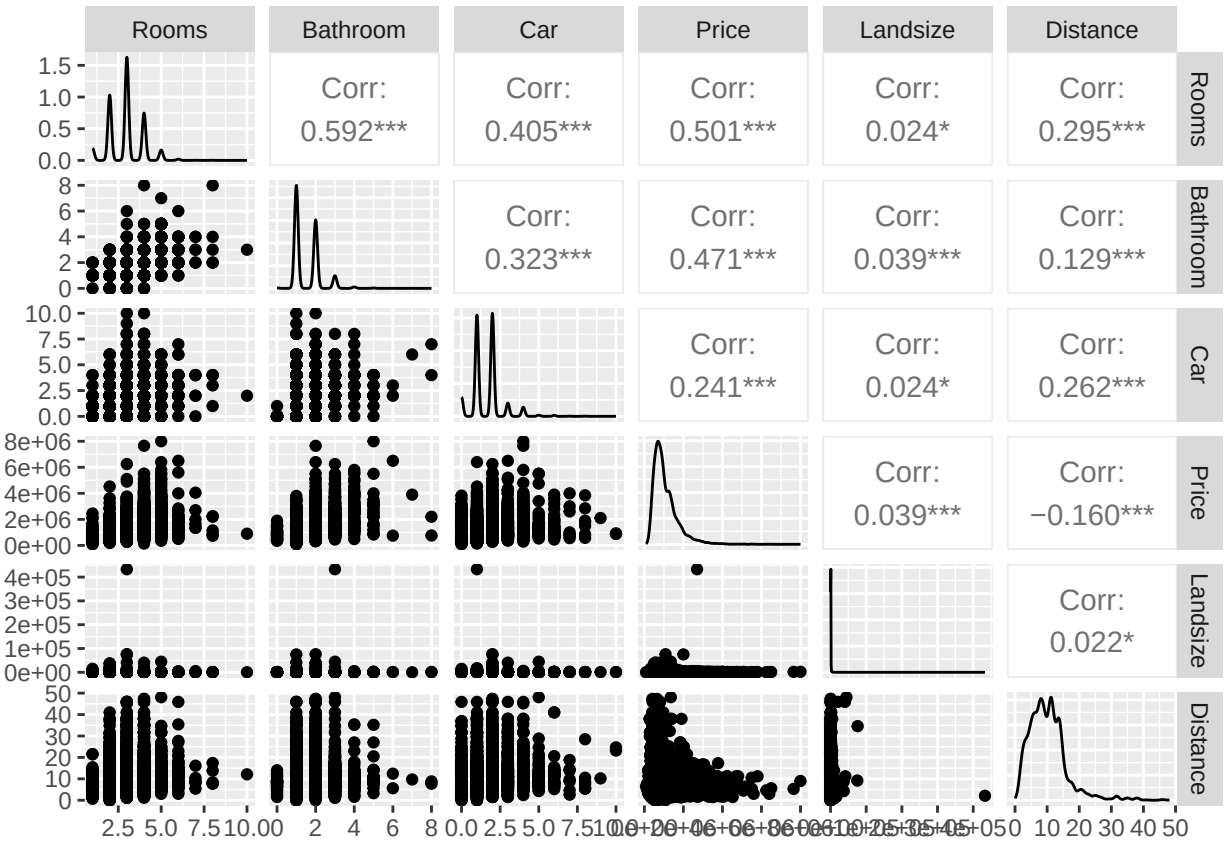
Interpretation of CI: We are 95% confident that the coefficient of the distance variable is between -0.01677631 and -0.01253827. Furthermore, since the confidence interval does not contain zero, we have significant evidence to reject the null hypothesis ($\beta \neq 0$).

While we did find that there is significant evidence for $\beta_1 \neq 0$, further analysis indicates distance's limited ability to predict price. For this reason, we need to construct a better, larger model.

The Full Model: Multiple Linear Regression

Our goal is to perform statistical analysis on the Melbourne Housing Dataset and map the relationships between the various variables. Specifically, we are interested in modeling price as a response of the other variables in order be able to predict housing prices in Melbourne for any housing with just a few variables. Once again, a reminder of the dataset citation:

Becker, D. (2017, September). Melbourne Housing Snapshot, Version 5. Retrieved May 11, 2023 from https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot.



Here we are creating a pairs plot to check the correlation of the quantitative variables and from it we can see that price is significantly correlated with Rooms, Bathroom, Car, Price, Landsize and Distance. This is a promising sign that these variables can be used as predictors for Price.

Knowing this, consider the following candidate models.

Model 1

```
##
## Call:
## lm(formula = log(Price) ~ Rooms + Bathroom + Car + Landsize +
##     Distance + Landsize:Distance, data = train)
##
## Coefficients:
##       (Intercept)              Rooms           Bathroom                Car
##         1.294e+01          2.982e-01          1.188e-01          4.198e-02
##          Landsize           Distance  Landsize:Distance
##         6.861e-07         -3.186e-02          4.540e-07
```

Model 2

```
##
## Call:
## lm(formula = log(Price) ~ Rooms + Bathroom + Distance + Car +
##     Landsize + I(Rooms^2) + I(Bathroom^2) + I(Distance^2) + I(Car^2) +
##     I(Landsize^2) + Landsize:Distance, data = train)
##
## Coefficients:
##       (Intercept)              Rooms           Bathroom           Distance
##         1.240e+01          7.634e-01          1.001e-01         -5.014e-02
```

```
##               Car           Landsize         I(Rooms^2)      I(Bathroom^2)
##         3.236e-02          7.403e-06         -7.419e-02          9.736e-03
##     I(Distance^2)           I(Car^2)      I(Landsize^2)  Distance:Landsize
##         5.366e-04          1.857e-03         -1.539e-11          1.298e-07
```

```
Model 3
```

```
##
## Call:
## lm(formula = log(Price) ~ I(Rooms^2) + I(Bathroom^2) + I(Distance^2) +
##     I(Car^2) + I(Landsize^2), data = train)
##
## Coefficients:
##   (Intercept)     I(Rooms^2)  I(Bathroom^2)  I(Distance^2)       I(Car^2)
##     1.336e+01      3.771e-02      3.198e-02     -7.105e-04      8.182e-03
## I(Landsize^2)
##     4.659e-12
```

Here we have our three candidate models and the first one is a simple linear model with one interaction term present. We chose to have this interaction term present because it had the strongest effect on R^2 out of multiple interaction terms we tested. The second model has added quadratic term to each of the quantitative variables because after some plotting we found that this transformation helped with linearity. The final model is a simplified version of model 2 in which only the transformed quantitative variables are present.

The first step in our model selection process was to optimize our candidate models and to do this we used `stepAIC()` to perform backwards selection. We chose to do backwards selection because it allows us to consider the effects of all variables simultaneously, which is important in the case of collinearity because backwards selection may be forced to keep variables with strong interactions in the model.

```
## Model 1 R^2 Model 2 R^2 Model 3 R^2
##   0.4497117   0.4931888   0.3558125
```

```
## Model 1 BIC Model 2 BIC Model 3 BIC
##   10379.987    9508.696   12064.816
```

```
## Model 1 AIC Model 2 AIC Model 3 AIC
##   10321.680    9435.812   12021.085
```

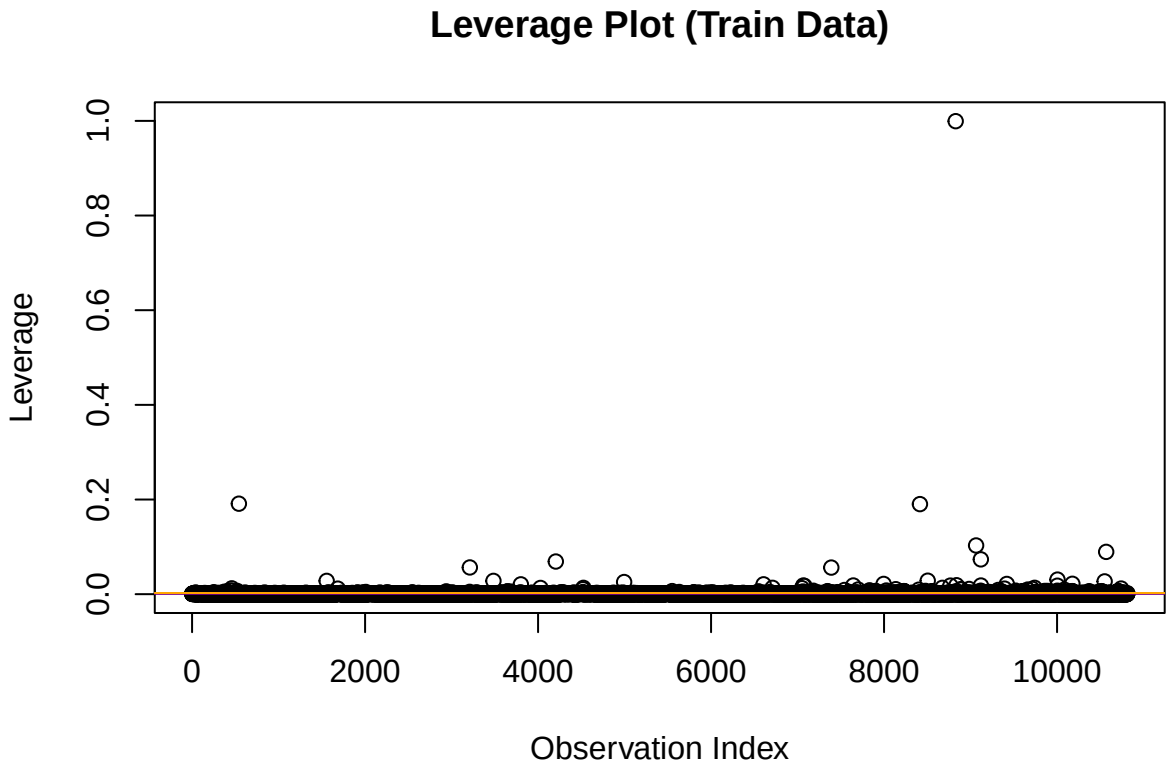
Here we can see after comparing the R-squared, AIC and BIC values that in terms of all three statistics, our second candidate model is preferred.

```
##                    Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)    1.237669e+01 2.542497e-02 486.792777  0.000000e+00
## Rooms          7.535927e-01 1.620677e-02  46.498640  0.000000e+00
## Bathroom       1.384503e-01 6.627818e-03  20.889279  4.994558e-95
## Distance      -5.056305e-02 1.726907e-03 -29.279543 2.044056e-181
## Car            4.131689e-02 4.225514e-03   9.777954  1.735404e-22
## Landsize       9.977773e-06 2.653961e-06   3.759578  1.710973e-04
## I(Rooms^2)    -7.274986e-02 2.494095e-03 -29.168839 4.112662e-180
## I(Distance^2)  5.501529e-04 5.143709e-05  10.695645  1.448581e-26
## I(Landsize^2) -2.065841e-11 6.512809e-12  -3.171966  1.518344e-03
```

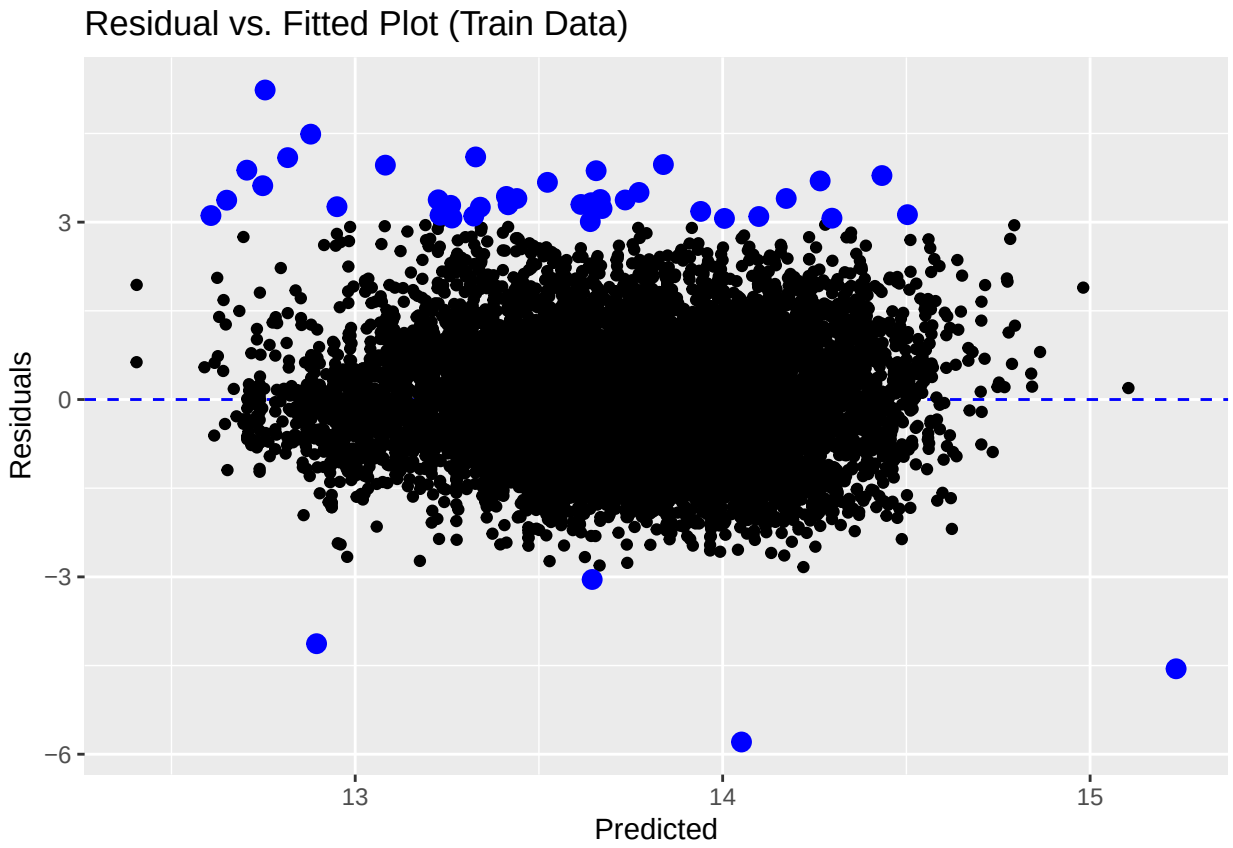
Using the summary function it is easy to see that all of our coefficients are statistically significant at the 0.05 level.

```
##    R^2       R^2 adjusted
##    0.4930865    0.4915828
```

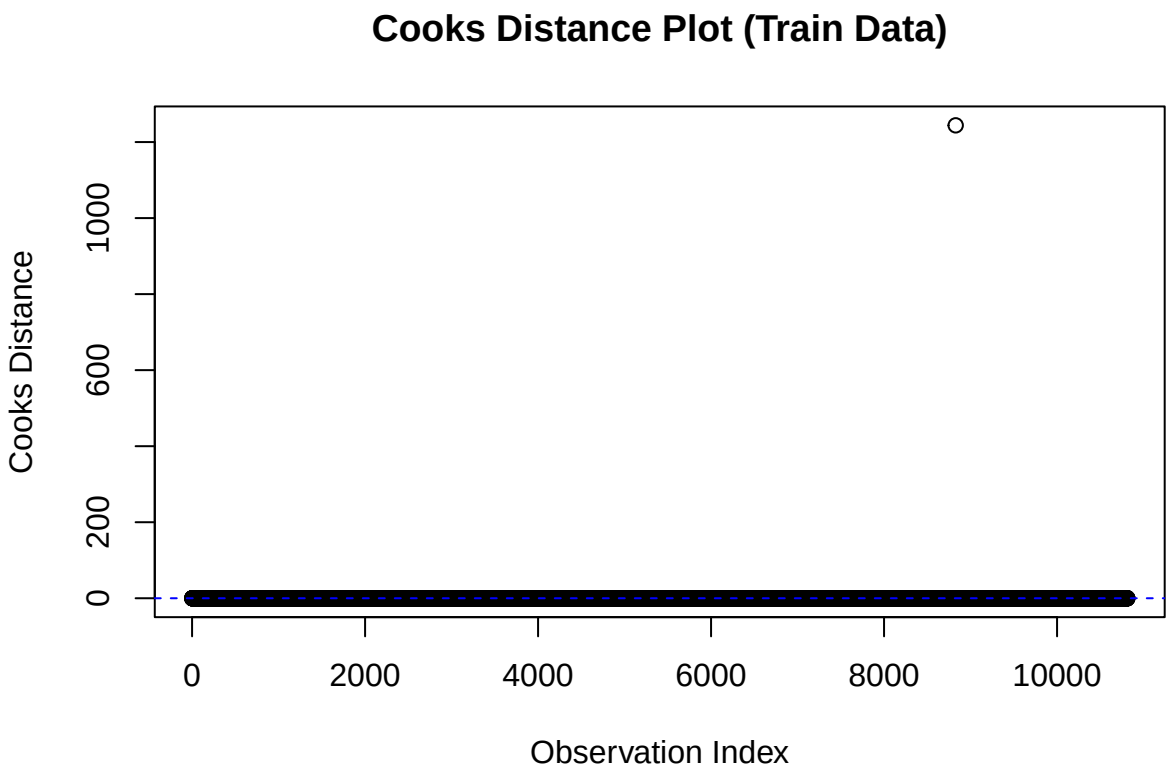
Here we have the variability explained by our model and it is important to keep in mind that a high R-squared does not always mean the model is a good one, it could be that the model is overfitting which is why it is important to use a train/test split to build the model. Our model has a decent R-squared value, there's oppurtunity to improve the model by introducing more complex transformations or a greater number of predictor variables.



This is a plot of the high leverage points with the uppermost cutoff being `3 p / n`. It is very clear to see that there exist a high number of leverage points but it is important to note that this does not necessarily mean that they have a great effect on the model, simply that they are "extreme" in the predictor space.



Here we have a plot of the outliers with the ones in blue having a standardized residual with an absolute value greater than 3, which is often considered to be the standard criteria for identifying outliers. From the plot we can see that there are relatively few of them when compared to the total number of observation in the dataset.



This is a Cooks Distance plot which identifies influential observation with the cutoff being any value that has a cooks distance greater than `4 / n` is considered an influential observation. Influential observation are the observation that have a large effect on the model. Hence, we are omitting the influential observations from the training data and the refitting the model with the new data.

```
## Refitted Model Original Model
##      0.4944250      0.4931888
```

As expected, by omitting the influential observations we yield a slightly higher R^2.

Using our refitted model, we construct a confidence interval for the mean predicted response.

```
##    Actual       Fit       Lwr       Upr
## 1075684.1  974639.9  965834.1  983526.1
```

Although the actual sample mean does not lie within constructed confidence interval, our prediction interval captures the actual value. However, this is likely due to the fact that our interval is huge.

```
##        Actual       Fit      Lwr     Upr
## 3425   515000  852990.3 409014.2 1778893
## 5249   810000 1032259.8 495109.0 2152173
## 13067 1138000 1074301.2 515256.4 2239900
## 5200   790000 1330808.8 638304.0 2774621
## 2020  1370000  992017.2 475801.8 2068294
```

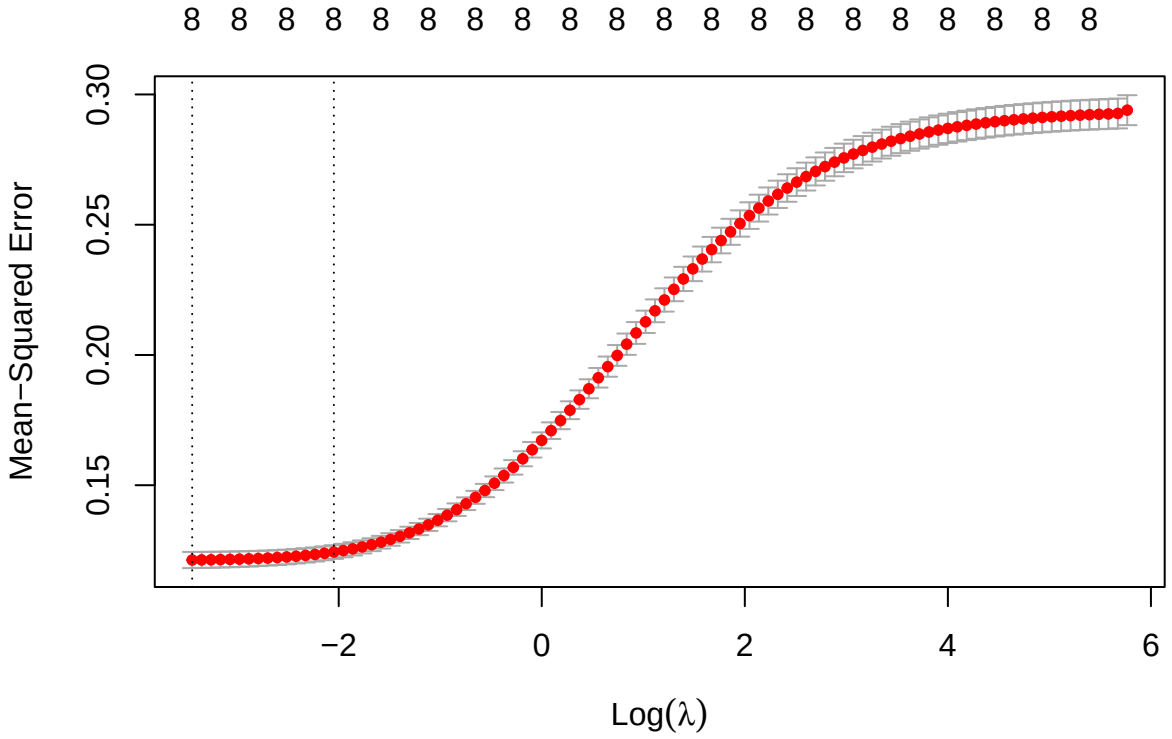
Our model has a fairly intuitive interpretation, all of the variables that were kept in our final model are typically characteristics that are desirable when house searching. Distance (which is the distance to Melbourne's Business District) has negative correlation which is intuitive to understand as property prices rise the closer they are to urban hubs. The rest have positive correlation which is also simple to explain as the more rooms, bathrooms, car spots and the bigger the landsize the bigger the house which comes with an added cost.

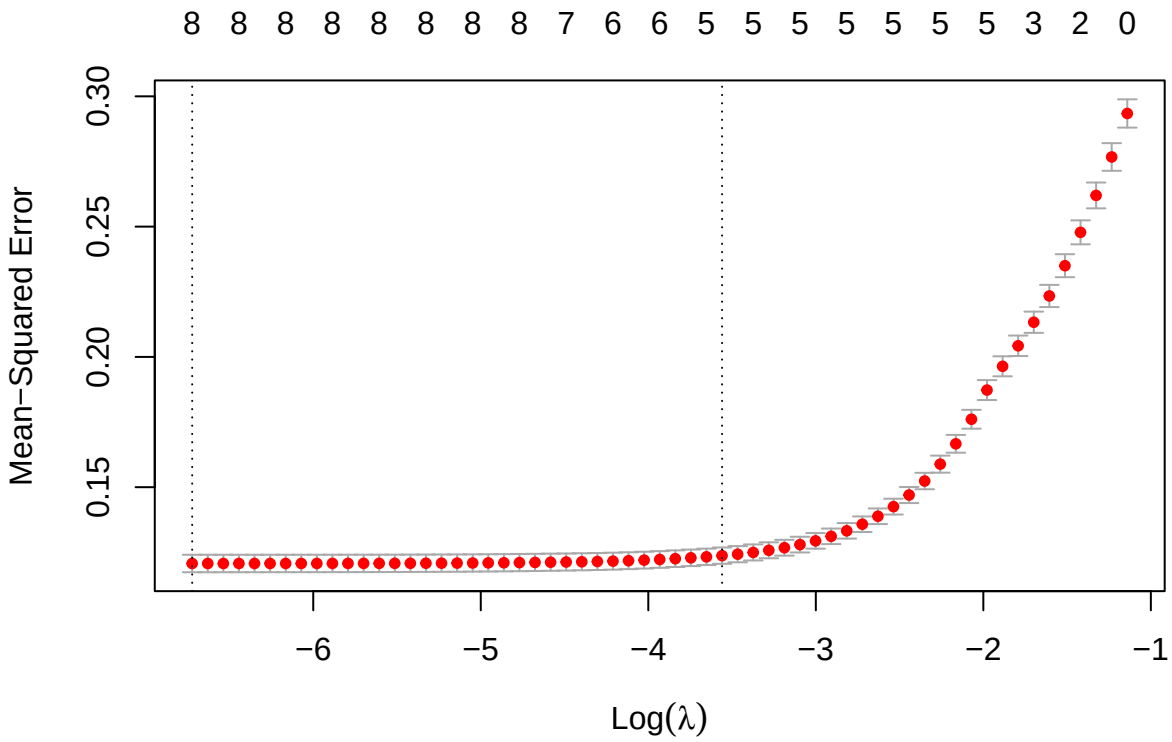
In summary, the quadratic multiple linear model was the strongest model from the three candidate models we proposed. We were able to further improve the model by trimming influential points. When constructing

the prediction interval for the model, we observed that the prediction interval was usually very large. This is expected as the housing market is very complex. ### Beyond the Linear Model

Having now chosen a model that fits the log of the price to our chosen significant explanatory variables, we now use the shrinkage methods lasso and ridge regression to optimize the model's parameters. We do this to reduce a large number of predictors while preserving useful information. We first load the melbourne data once again. A reminder of the dataset citation:

Becker, D. (2017, September). Melbourne Housing Snapshot, Version 5. Retrieved May 11, 2023 from https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot.





Note that the ridge MSE increases similar to a sigmoid activation as lambda increases, while lasso MSE increases exponentially. This may mean that for this model, as lambda increases and more predictor coefficients are set to zero, the lack of predictors increases MSE.

```
## Best Ridge Lambda Best Lasso Lambda
##        0.03195859        0.00120322
```

The following values above are the best lambda values for each respective model.

Coefficients of best ridge model.

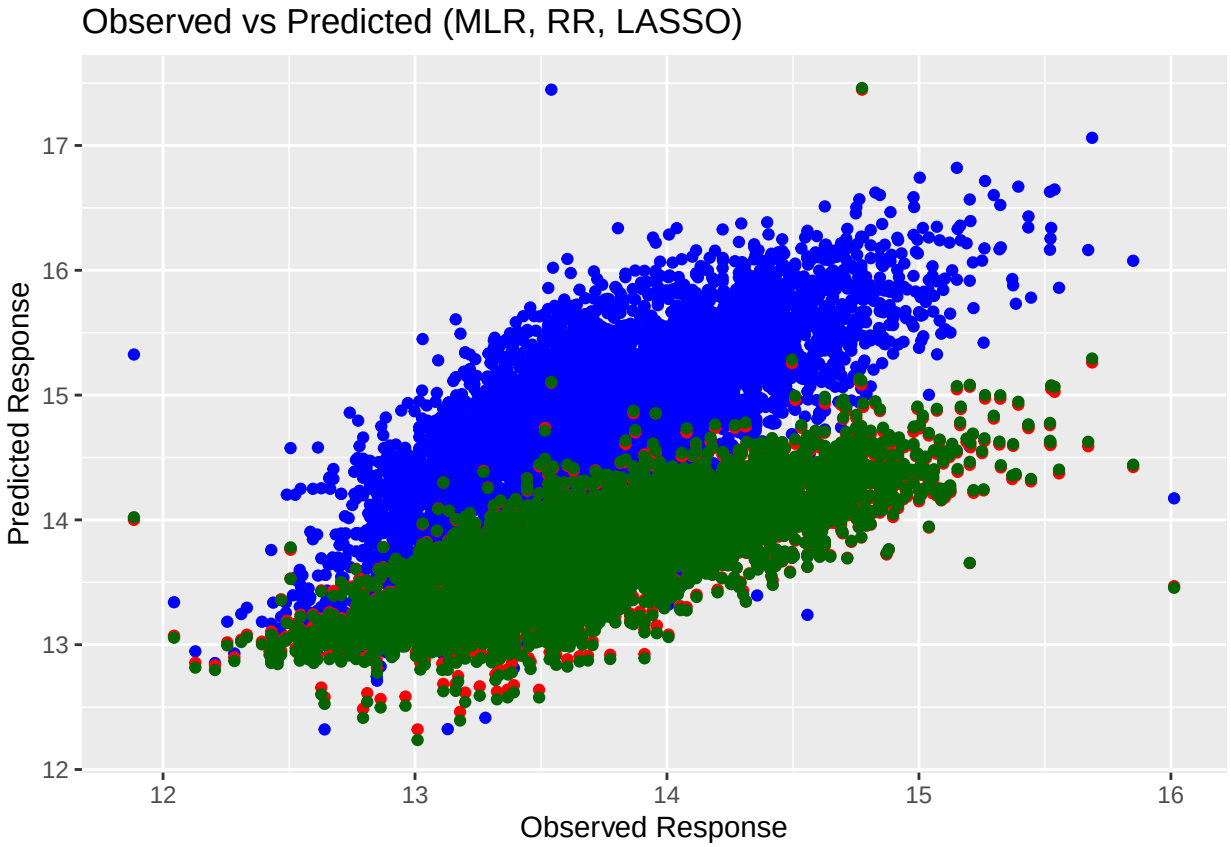
```
##  (Intercept)        Rooms     Bathroom     Distance          Car         Type
##  13.73723179   0.14063904   0.08856513  -0.19082280   0.02069681  -0.17865502
## BuildingArea   Regionname     Landsize
##   0.09690724   0.01085016   0.00951464
```

Coefficients of best lasso model.

```
##  (Intercept)        Rooms     Bathroom     Distance          Car         Type
## 13.737231787  0.147160206  0.088364225 -0.204504828  0.019109609 -0.187939633
## BuildingArea   Regionname     Landsize
##  0.096695573  0.009592913  0.009171623
```

```
## R^2 Lasso R^2 Ridge   R^2 MLR
## 0.5966001 0.5954708 0.5330962
```

The R^2 value of the LASSO, Ridge and MLR prediction are noted and compared. As observed, the R^2 value of Lasso is slightly higher than ridge and significantly higher than the linear model.



Here we have MLR in blue, LASSO in green, and Ridge Regression in red. Ridge Regression (RR) and LASSO add a penalty term to the loss function, which constrains the coefficients to be relatively small. Therefore, the predicted value range for RR and LASSO tends to be smaller compared to MLR.

The Melbourne Housing dataset has many variables, and even with everything we have done to far to capture the relationships between them, it is difficult to fully do so. Principal component analysis (PCA) is another way to capture variable information: it identifies principal components, linear combinations of the original predictors that explain the most amount of variation in the dataset in the fewest number of components.

We performed PCA on every quantitative predictor in the dataset in order to identify which predictors would display the largest amount of variation, then compared those predictors to the ones we used in our model. Below are the variances explained by each of 12 principal components:

```
##  [1] 0.25857747 0.15239333 0.12688124 0.08475448 0.08147569 0.06821104
##  [7] 0.05787945 0.05000137 0.03955929 0.03649329 0.02426800 0.01950536
```

Below is the full PCA.

```
##                     PC1          PC2         PC3         PC4          PC5
## Price         -0.39257159  0.189896847 -0.42239194  0.04154562 -0.025430799
## Rooms         -0.46223251  0.190811985  0.07752977 -0.01173652  0.045732978
## Bathroom      -0.43408249  0.149396636  0.08067574  0.04821017  0.148203771
## Distance      -0.20559328 -0.391190151  0.43559961 -0.09898889 -0.002761026
## Car           -0.32001458  0.081132537  0.26793807  0.06195989 -0.041463114
## Landsize      -0.10072925  0.005521388  0.12560905  0.10665133 -0.959175086
## Postcode      -0.19247170 -0.556184421 -0.06322524  0.04961030 -0.010410455
## BuildingArea  -0.43330875  0.183138237  0.01113172  0.03913072  0.089284088
## YearBuilt     -0.01246113 -0.148811368  0.59007806  0.02833449  0.190234181
## Lattitude      0.14020572  0.412103163  0.35648432  0.20001869 -0.007922527
```

```
## Longitude     -0.20086793 -0.453120679 -0.22581006  0.09857991  0.004369152
## Propertycount  0.06335657 -0.073450642 -0.03695152  0.95722461  0.095931365
##                       PC6         PC7          PC8         PC9         PC10
## Price          0.07961846 -0.03570625 -0.005454077  0.20118926 -0.10844624
## Rooms         -0.20565162  0.17499009  0.010359719 -0.16434189  0.39744922
## Bathroom       0.35127420  0.14943280  0.062305808  0.33507399  0.42095487
## Distance      -0.42491651  0.18141112  0.114242888 -0.26289331  0.17457853
## Car           -0.27832776 -0.81388884 -0.066573655  0.18549815 -0.13844916
## Landsize       0.18851835  0.06641597  0.028074371 -0.02947818  0.01610828
## Postcode      -0.14815535  0.24178392  0.205203620  0.59188793 -0.31107495
## BuildingArea   0.05367671  0.22357706  0.061372201 -0.43425112 -0.66692254
## YearBuilt      0.63876838 -0.06342456 -0.037853685  0.01597224 -0.14147771
## Lattitude     -0.29679203  0.36134216 -0.510473156  0.35832482 -0.16647038
## Longitude      0.11115017 -0.04556817 -0.796792626 -0.17428248  0.07849718
## Propertycount -0.03345274 -0.03039397  0.186922313 -0.13385473  0.08787364
##                       PC11        PC12
## Price          0.56338553  0.50175121
## Rooms          0.37292419 -0.58704663
## Bathroom      -0.53347888  0.18647479
## Distance       0.05455851  0.52541941
## Car           -0.11311001 -0.05453526
## Landsize      -0.02072512 -0.01891439
## Postcode       0.01622274 -0.27778844
## BuildingArea  -0.27697912 -0.04863697
## YearBuilt      0.39663513 -0.03801038
## Lattitude      0.03913126  0.08053070
## Longitude     -0.08030086 -0.03623164
## Propertycount  0.02109564  0.02601995
```

Looking at the three largest principal components, it seems that Price, Rooms, Bathroom, Car, BuildingArea, Latitude, and Longitude had fairly consistently high values, meaning the principal component analysis determined these variables displayed a fairly high amount of variation. Combined with the fact that the no one principal component explained a significantly large amount of variance, this indicates that the variables are very interconnected. Due to the complex nature of the data, any attempt to use just a few variables to predict price would be difficult.

The above results also indicate that with the exception of Landsize, the predictors we used in our model were contributing significantly to the model. As for the other promising predictors, we did not use them in the model due to the high number of missing values those columns contained.

Conclusion

As mentioned before, due to the complex nature of the data, it is incredibly difficult to predict the price of housing. For instance, variables such as rooms, bathrooms, buildingarea (intuitively) are all necessarily interconnected. While we were able to construct models that are mostly able to capture new values relatively close to the price, the prediction interval itself is incredibly large and uncertain. The fact is that the housing market is intrinsically extremely complex and our data set alone is insufficient in capturing the true nature of the housing economy.

Starting off with the simple linear model, we did expect to see some prediction power. However, upon seeing our residual plot and the R^2 value (about 0.02), we quickly realize just how complex the housing market really is. Using backwards selection, we created the Lasso and Ridge Models which are significantly better than our simple linear model, but still falls short in predicting housing prices.

Ultimately, because our data set was so limited in comparison to the full complexity of the housing market, we are unable to create a model that meaningfully predict the price of housing. With a much larger and

powerful data set, we would be able to generate a much more powerful and accurate model.