Final Project

ParticipationGrade

2022-12-02

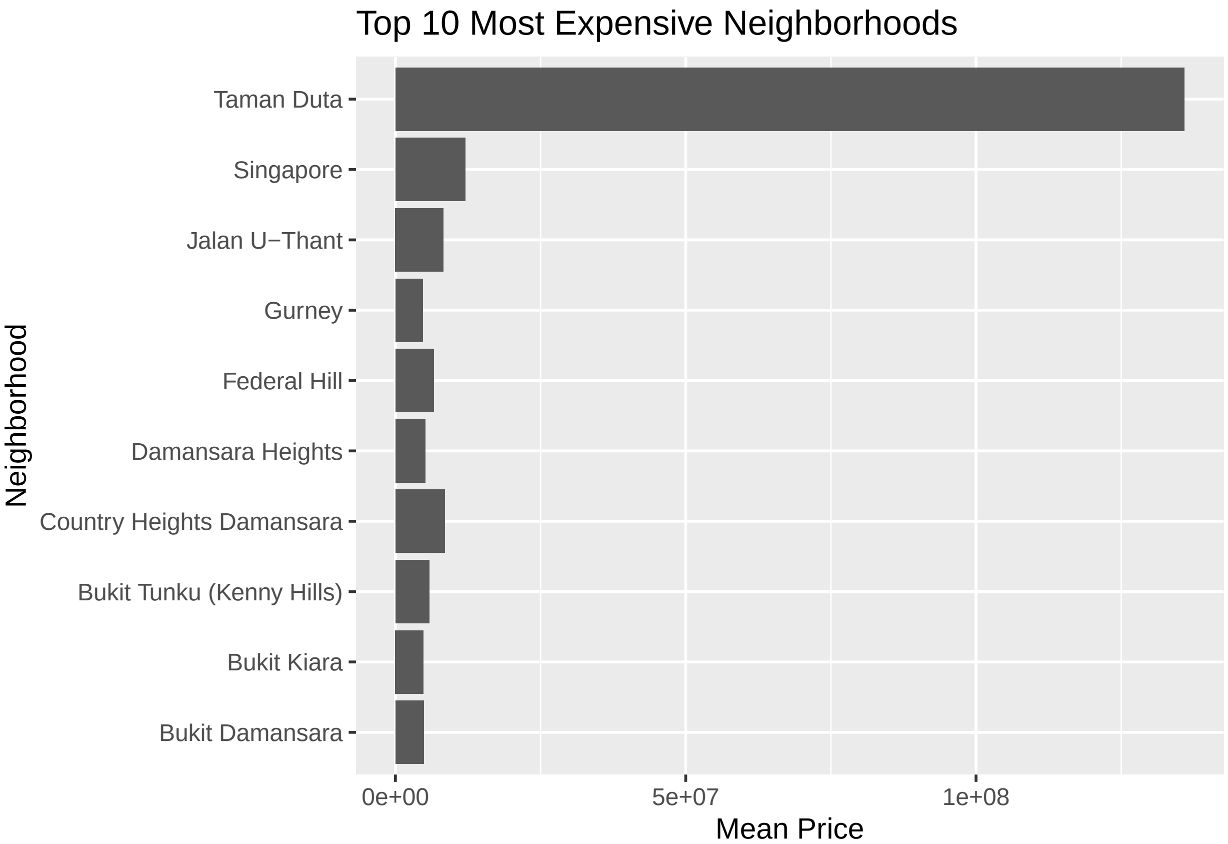
Revisions Summary: - Added new best subset model using a dummy variable for most and least expensive neighborhoods - Added new EDA plot to further explore relationship between price and continuous variables - Changed KNN regression k-fold CV to only use original train data in training - Added explanation of best subset regression choice - Removed train RMSE plots and code for best subset regression - Explained why we chose the four predictor model over the two predictor model for BSR

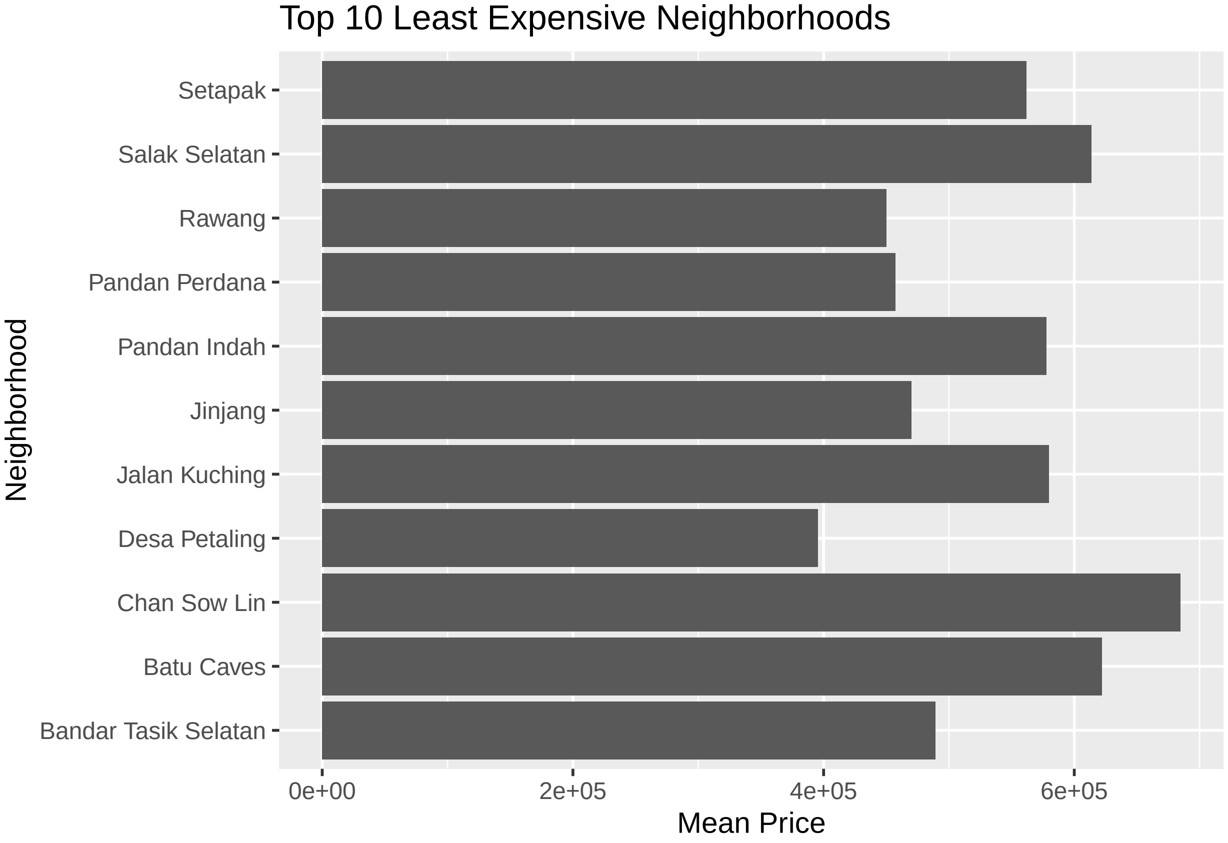
Introduction:

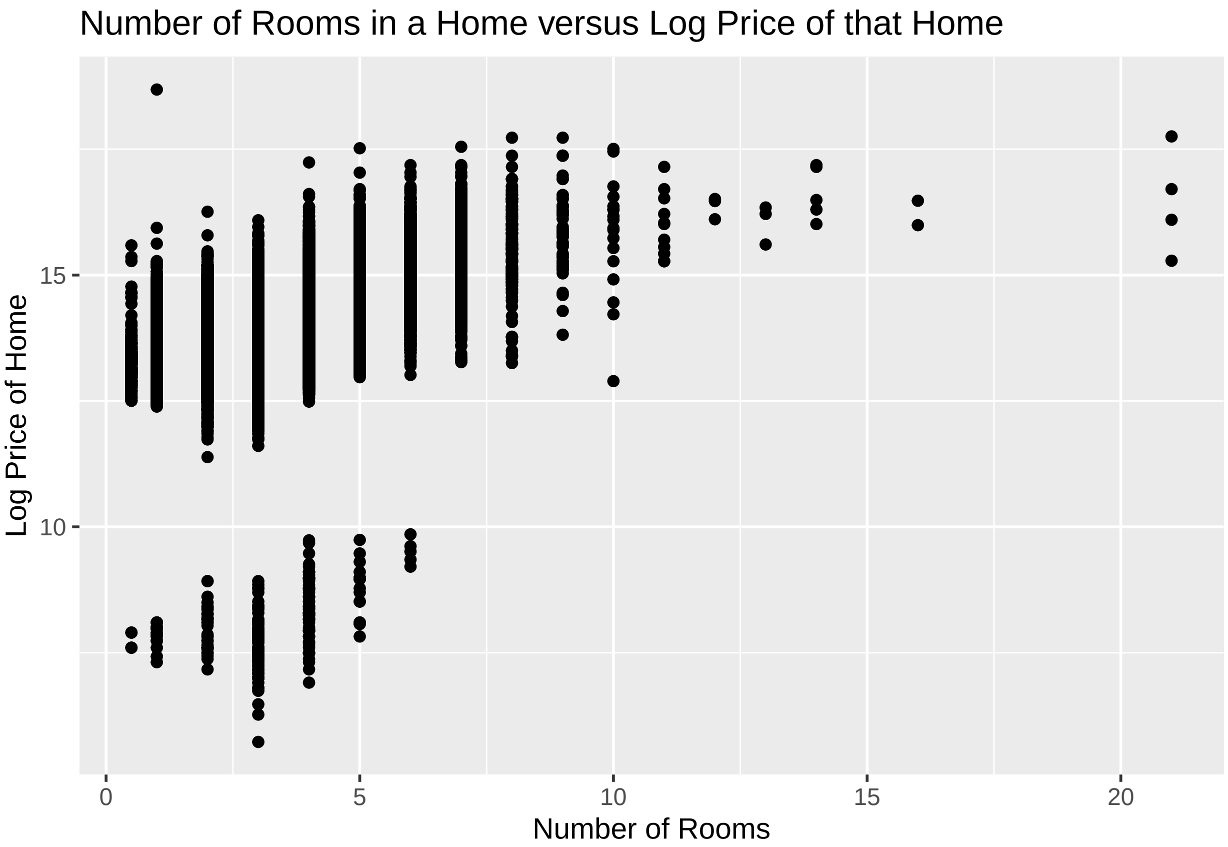
The main focus of our research was to determine if we could predict the price of houses in Kuala Lumpur given data from a Malaysian house listing website. The dataset was taken from Kaggle, and was scraped directly from a real estate website on a single (unspecified) day in 2019. The variables explored in this dataset include the house location, furnishing status, size, number of rooms, car parks, and bathrooms, the size of the house, as well as the kind of property. For our analysis, we wanted to attempt to use two distinct methods of regression in order to compare which predictors were most important in accurately predicting the price of a house, and to analyze the differences in prediction accuracy between them.

EDA:

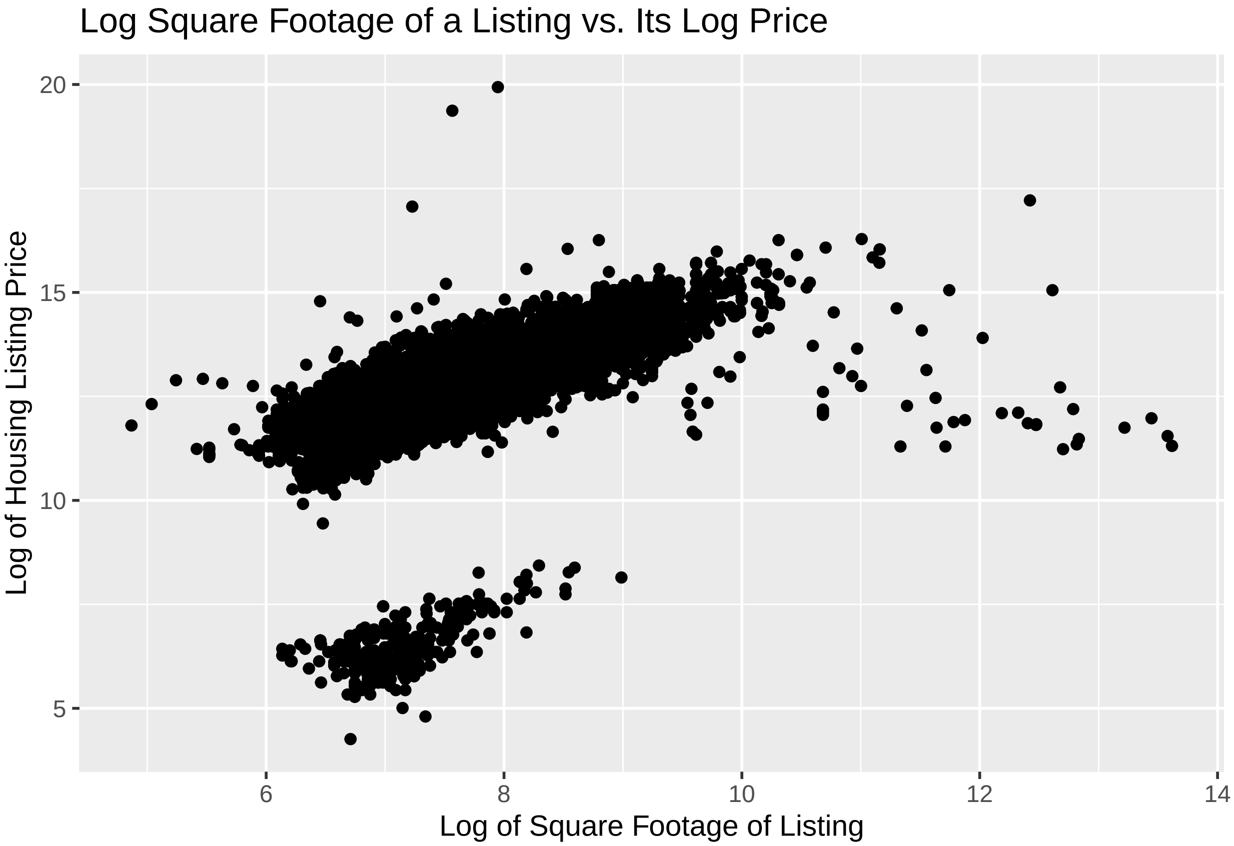
The dataset contains seven potential predictors which are as follows: Location, Rooms, Bathrooms, Car.Parks, Property.Type, Size, and Furnishing, and one response variable of interest for our purposes: Price. Location specifies the neighborhood and city of the listed house (all Kuala Lumpur), the Price variable gives the price in Malaysian ringgits, the Rooms variable gives the number of rooms in the house (sometimes given as an expression and not a whole number), the Bathroom variable gives the number of bathrooms, Car.Parks refers to the number of parking spots, Property.Type refers to the kind of property the house is listed as (condominium, etc.), Size is given mostly in sq. ft., although also occasionally in acres and hectares, and Furnishing gives the furnishing status at the time of listing (fully furnished, partly furnished, etc.) To explore our dataset preliminarily we examined the relationship between various features of the house listing and its price, including number of rooms, location, and square footage.


Top 10 Most Expensive Neighborhoods
Neighborhood
Taman Duta
Singapore
Jalan U−Thant
Gurney
Federal Hill
Damansara Heights
Country Heights Damansara
Bukit Tunku (Kenny Hills)
Bukit Kiara
Bukit Damansara
0e+00
5e+07
1e+08
Mean Price


Top 10 Least Expensive Neighborhoods
Neighborhood
Setapak
Salak Selatan
Rawang
Pandan Perdana
Pandan Indah
Jinjang
Jalan Kuching
Desa Petaling
Chan Sow Lin
Batu Caves
Bandar Tasik Selatan
0e+00
2e+05
4e+05
6e+05
Mean Price



Number of Rooms in a Home versus Log Price of that Home



Log Square Footage of a Listing vs. Its Log Price

These graphs indicate that regressing on number of rooms alone may not be a very powerful predictor of price, as the same number of rooms, e.g. 3, can have an incredibly low log price (close to 0), or a much higher log price (above 15). Location on the other hand, seems to have a strong effect on house price. Given our location plot we can also see that there may be potential outliers in this data set. The Neighborhood of Taman Duta has multiple, multi-billion dollar listings that skew it to be far more expensive than all the other neighborhoods. We used the log of the home price as the response value of these plots in order to better standardize our data and see trends more clearly.

The Log Square Footage vs. Log Price and the Room Number vs. Log Price visualizations both show slight relationships between the two predicting variables and the response variable of price, yet they also exhibit gaps which could be better explained by using these predicting variables in conjunction with one another and other predicting variables such as number of bathrooms and number of parking spaces. We took the log value of the variables (SqFootage and USPrice) which exhibited extremely wide ranges due to some outliers. These outliers included large compounds and entire apartment buildings, which had enormous prices and enormous square footage, and luxury million-dollar apartments which were comprised of smaller square footage and large price values, in order to get a clearer picture of the relationship between these variables and the price of a listing.

Methodology:

Cleaning Data - First, we converted all of the housing prices from string representations of the price in RM currency to numerical values representing their USD prices. Then, we cleaned the awkward and sometimes typo-laden entries of the column representing the number of rooms in a house listing. Some of the room number entries were listed as expressions that had to be evaluated, such as “5+1” rooms, so to remedy this, we converted strings to expressions and evaluated them using the eval(parse()) function to convert them into numerical values that we could use in a regression model. Then we cleaned the extremely inconsistent and often erroneous variable representing the square footage of a listing. We had to manually convert around 120 entries which were listed in acres rather than sq. ft. Some entries didn’t even have a numerical value and were rather referencing an entire district or reference code, and so we assigned these incorrect entries as NA values and removed them from the dataset. Finally, we manually removed observations that were not caught previously by our many cleaning techniques, which seemed to have square footages that were implausibly small or large. The resulting dataset resides within the variable cleandata. These values are now within the cleandata columns “USPrices”, representing the converted housing price; “RoomNum”, representing the evaluated number of rooms of a property; and “SqFootage”, representing the evaluated square footage of a listed property.

Train & Test Split - We split our clean data into training and testing data sets for our model generation. The training dataset contained 70% of the cleaned observations, or 24523 randomly selected rows of the cleaned data. The testing dataset consisted of all the rows not included in the training dataset. The data used in the K-fold cross validation for the KNN regression was taken from a 70-30 split of the already split train data.

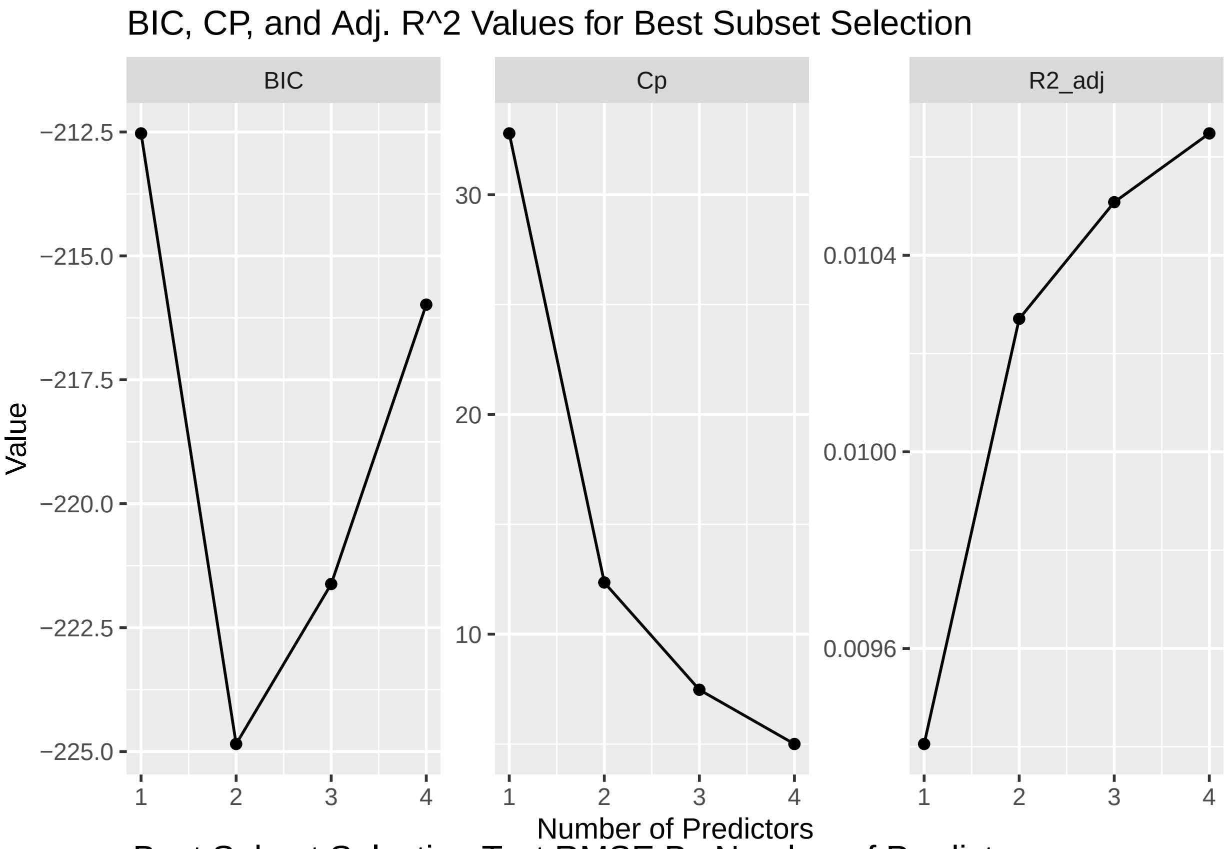
Best Subset Regression - For our first method of regression, we chose to use best subset regression. Best subset regression identifies and returns the best model of each predictor size, in our case one predictor up to four predictors. We selected this method as we thought it would be useful in answering our initial research question about which predictors were important for predicting total price. Initially, we intended to use the categorical variables (Location, Property.Type, SizeType, and Furnishing) as well as continuous variables in our model creation, which would have lead to potential model sizes of one predictor up to ten predictors. We successfully translated these variables into dummy variables using the dummy_cols function from the R package fastDummies, however our computers were not powerful enough to run that many resulting variables in a reasonable time frame, and so categorical variables were excluded from further analysis. For best subset regression, this left four potential predictors; Bathrooms, Car.Parks, SqFootage, and RoomNum. We used the regsubsets function from the R package leaps in order to construct the models, regressing the training data for USPrice on the four predictors. We then calculated the test root mean sure error, BIC, Cp, and adjusted R^2 values for each model size given the outputs of the predict.regsubset function found in Lab 04 Selection and the summary statistic of the model. All were then visualized in plots using the R package ggplot2.

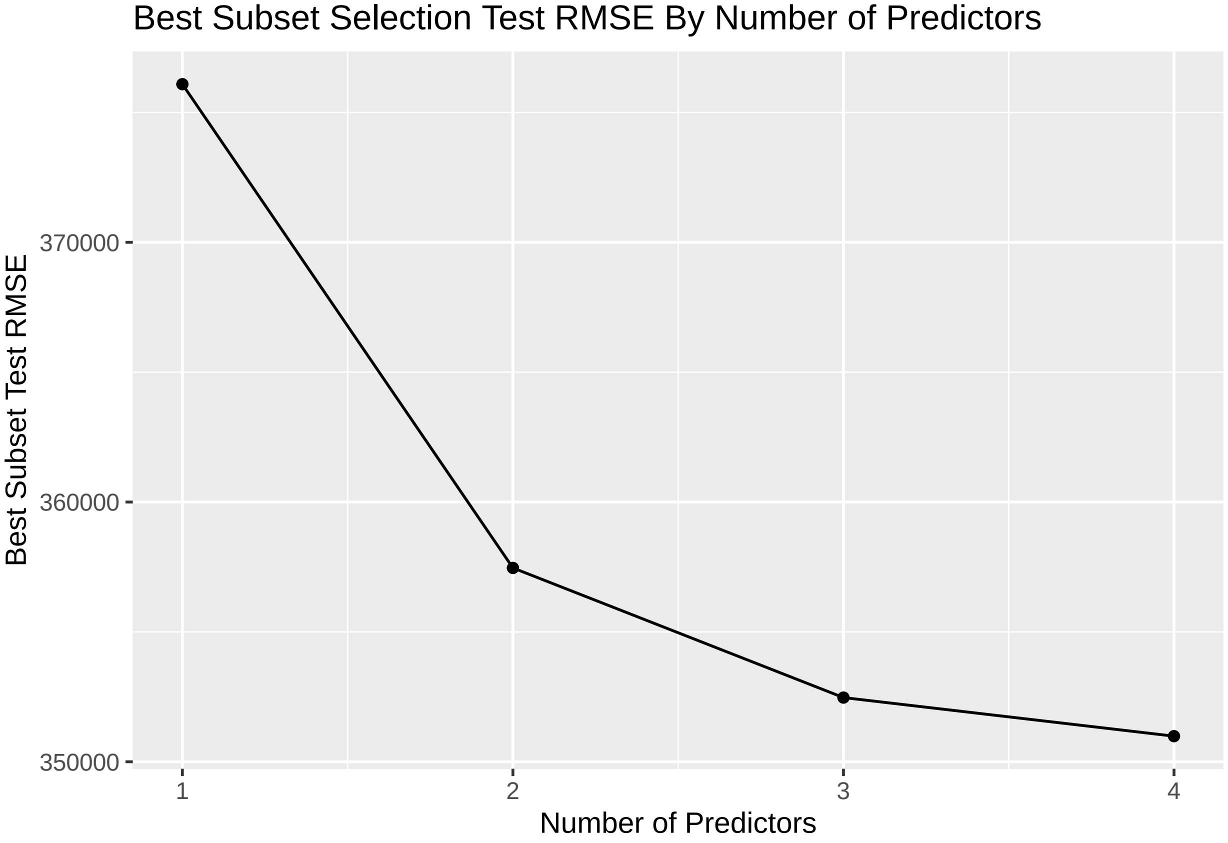
As part of our revision process we also included another round of best subset regression, with the intention of being able to use at least some of the categorical variables. For example, in our EDA, we found that Taman Duta was the most expensive neighborhood by a large margin. Therefore, when we run best subset regression on only a subset of the categorical variables, we figured that whether or not the listing was in Taman Duta would be a valuable predictor. Computationally we could include more than just the dummy variable for Taman Duta, but not the dummy variables for all neighborhoods. Since we are including the most expensive neighborhood as a potential, and had power for another potential predictor, we also decided to include least expensive neighborhood as a dummy variable, which in this case is Chan Sow Lin. The final list of the 5 predictors used in this model were Bathrooms, SqFootage, RoomNum, Taman Duta (yes/no), and Chan Sow Lin (yes/no), which again is a rather small amount of variables to use with best subset prediction, but in order to be consistent with the already performed model, we still used best subset prediction with the understanding that it would most likely still choose the model with all predictors. The dummy variable used to indicate the neighborhood was either a 1 (in Taman Duta or Chan Sow Lin) or a 0 (not in TD or CSL).

K Nearest Neighbors Regression - We sought to compare the differences in accuracy between this method and the Best Subset Regression of 4 predictors. We used four predictors in our KNN model, as was determined most accurate by the Best Subset method. Using all four predictors, RoomNum, SqFootage, Car.Parks, and Bathrooms, we sampled 40 values of K between 1 and 500 to test which K value would yield the lowest RMSE when used in a KNN regression model. Our testing gave us the optimal neighbor set size which gives us the lowest RMSE within the training data of the model. Using this K value, we generated the most accurate KNN regression model using our four predictors to predict the USPrices variable of our test data.

Results:

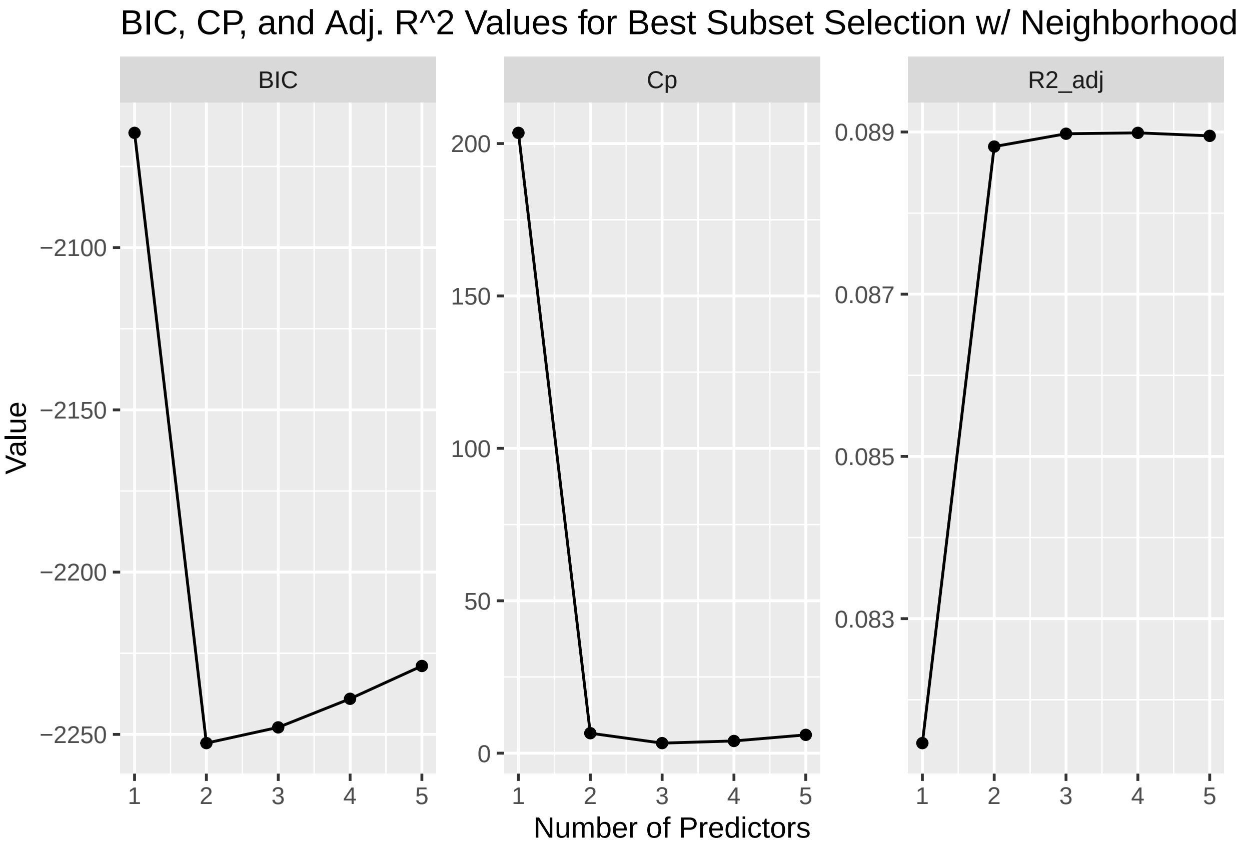
Best Subset Results - The R^2 adjusted for best subset selection was maximized at four predictors, while the Cp and BIC values were minimized at 4 and 2 respectively. The test RMSE was minimized in a model with all four predictors. We chose 4 as the optimal model size. Given the already small amount of predictors we were working with we weren't concerned about the possibility of overfitting resulting from a larger number of predictors, and the CV test error, as well as the Cp value, was minimized at 4. The test RMSE at four predictors was $350,985.22. The resulting model was: USPrice = -2.090762e+05 + 1.962128e+05 x Bathrooms + 1.127425e+05 x Car.Parks + 4.953569e+00 x SqFootage - 6.058989e+04 x RoomNum



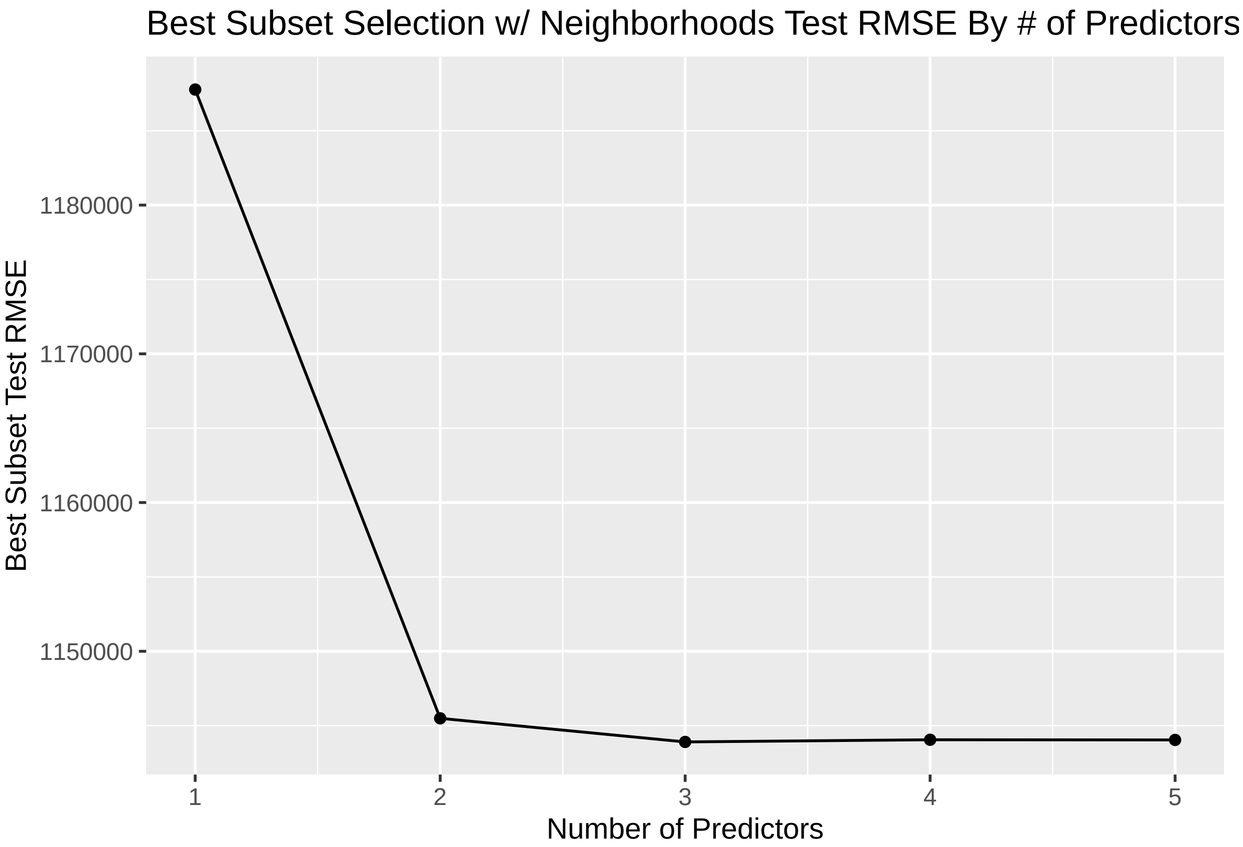


```
##   (Intercept)     Bathrooms     Car.Parks     SqFootage      RoomNum
## -2.090762e+05  1.962128e+05  1.127425e+05  4.953569e+00 -6.058989e+04
```

The best subset model that included the two neighborhood predictors (Taman Duta & Chan Sow Lin) performed best, contrary to our initial thinking, at only 4 predictors, not all 5. The Rˆ2 was maximized at 4 predictors, while the Cp and BIC values were minimized at 3 and 2 respectively. The test RMSE was minimzed at 4 predictors. These results obviously made it harder to select the best model size, as the various diagnostic parameters somewhat disagree. In the end, we chose the model with four predictors, since it minimizes both the test RMSE and the adjusted Rˆ2 value, and similarly to the previous model we were not overly concerned with the possibility of overfitting given we only had 5 potential predictors. The final test RMSE of the model with four predictors was $1,143,903.41. The final model given by this method was: USPrice = -1.526937e+05 + 2.118226e+05 x Bathrooms + 4.164160e+00 x SqFootage - 3.040803e+04 x RoomNum + 4.724859e+07 x Taman Duta (1 = yes, 0 = no).

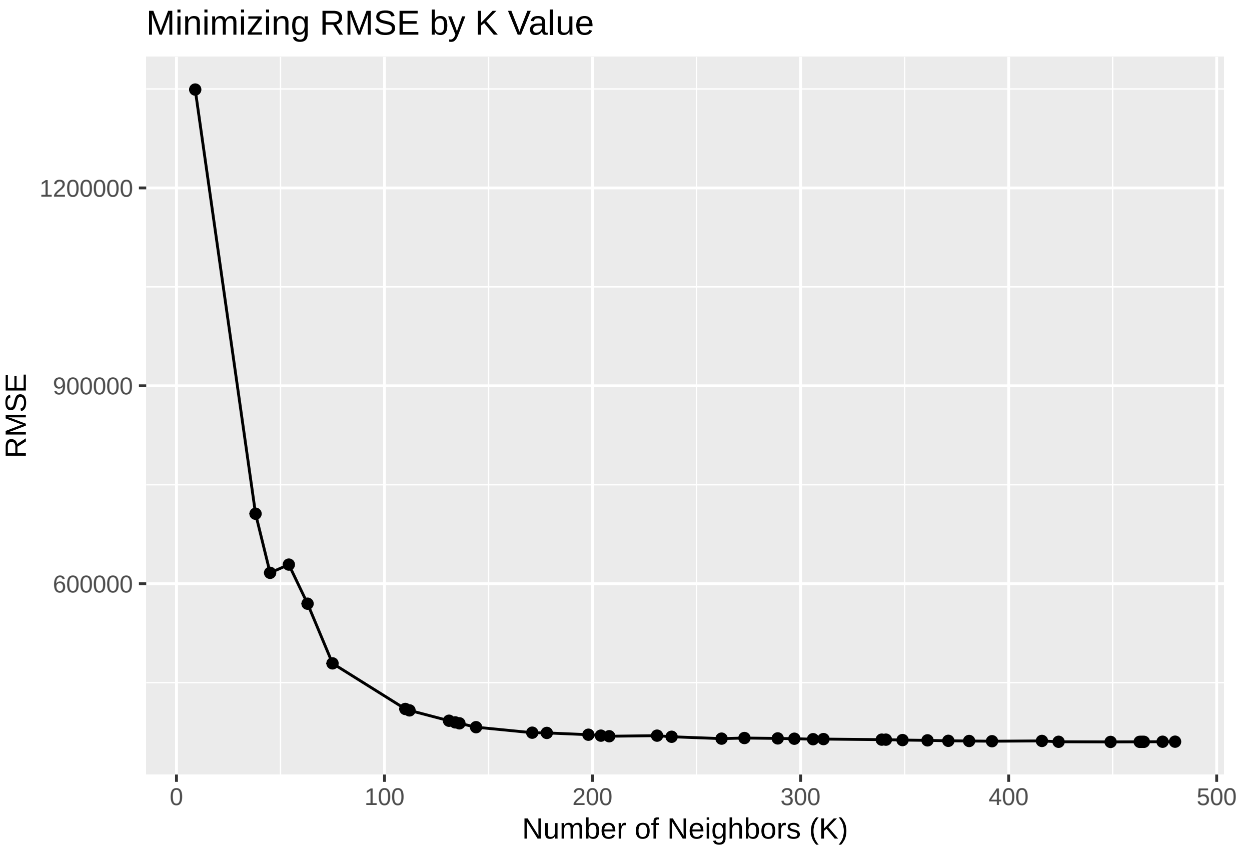


BIC, CP, and Adj. R^2 Values for Best Subset Selection w/ Neighborhood



```
##   (Intercept)     Bathrooms     SqFootage       RoomNum    taman_duta
## -1.526937e+05  2.118226e+05  4.164160e+00 -3.040803e+04  4.724859e+07
```

KNN Regression Results - Using all four predictors, as recommended by our Best Subset Selection model, we found that the optimal K value for a KNN regression model would be 449 neighbors. This value of K gave us the lowest RMSE by a large margin. We then created a KNN regression model that bases its decisions upon the 449 nearest neighbors which predicted the prices of the house listings in our test data with a RMSE of $342,937.40, which was slightly lower than the RMSE of our Best Subset Selection model.



```
## [1] "Optimal K value for KNN regression:  449"
```

```
## [1] 342937.4
```

```
## [1] "RMSE when K is 449:  342937.37906048"
```

Discussion:

Using Subset Selection and K-Nearest-Neighbors, we were able to predict the price of a house listing in Kuala Lumpur given its number of rooms, bathrooms, parking spaces, and its square footage within a margin of error around $360,000. While this may seem like a large margin, it is helpful to note that Kuala Lumpur, Malaysia has one of the most disparitous housing markets. It is the most expensive real estate market out of any city in Malaysia, while still encompassing districts that are essentially slums. The average house price in 2020 was $185,648 in the city, yet this data set also includes listings for entire multi-building complexes that can easily reach up to billions of dollars in price.

Subset Selection confirmed our hypothesis that all of the four predictors we were testing would be useful in predicting the price of a house. This method's RMSE was $350,985.22. Using many KNN regression models and comparing their RMSE, we identified that using a K value of 449 yielded the most accurate KNN regression model. This optimal model was slightly more accurate than the Subset Selection model, with a RMSE of $342,937.40. Due to the very similar values of these error margins, we suspect that this is near the accuracy limit of most regression models using only these four predictors.

To critique our methodology, only using the quantitative variables of our data in our models probably does not give us the most accurate price predictions. However, because of the size of our data set, we were limited to only using these quantitative variables due to the hundreds of different categorical data labels produced when trying to incorporate dummy variables. Also, it seems to us that listings for entire apartment buildings and complexes should not be included in a dataset of housing prices, and likely negatively affect the accuracy of our overall price predictions. Removing these observations entirely from the data set may improve overall confidence in our predictions. Another aspect of our methodology that was effected by our discovery that we did not have the power to run the quantitative variables was our choice to use best subset regression as a potential model. This model made a lot more sense when we were going to be choosing from potentially hundreds of variables (the original variables as well as the many, many dummy variables), but became less reasonable once we were limited to just four predictors anyways. Had we had more time to go back knowing we would only be able to use four predictors, we may have found more interesting conclusions using other methods such as a decision tree, which would have allowed us to analyze how consequential each of the four predictors may be.

In the future, we could improve this investigation by utilizing the neighborhood variable as a predictor in our different models, as the neighborhood that a house is in is usually a very large factor in the house's pricing, which we confirmed via our EDA. Unfortunately, we were both limited by computational power and patience within this project, and were unable to configure the neighborhood variable in such a way that our computers could create a model using it in under 4 hours. It would not only be interesting to use such a variable in our predictive models, but it may also be interesting to see how accurate a tree model would be, seeing as our RMSE of the price is around $350,000, so there is clearly room for improvement. Perhaps we could regress all of the variables but using smaller data sets that are defined by neighborhood, resulting in far smaller datasets but possibly a more accurate model, if the neighborhoods are fairly homogeneous in price.

Revision Additions: Best subset selection may seem like an unnecessary choice for running models with only four and six variables respectively, and we generally certainly agree. However, we began our analysis and code with the thinking that we would be using dozens of dummy variables representing the many categorical variables in our dataset. When that proved impossible to run, we already had our best subset code written out, and decided that for the sake of time it would not be possible to restart the entire section, and that the model would most likely just return a model with all predictors anyways.

The results of our best subset model including binary dummy variables for the most and least expensive neighborhoods, Taman Duta and Chan Sow Lin, interestingly did not improve our model accuracy. Contrary to our initial assumptions, the model actually performed worse than the model that did not include any categorical variables. It is not entirely surprising that the least expensive neighborhood was not included as a predictor in the final four predictor model, given the fact that it was not the least expensive neighborhood by any large margin and so most likely did not indicate any particularly important general information for listings in the neighborhood. At first, it was surprising that including the Taman Duta variable did not increase the model accuracy, given that the neighborhood is so much more expensive than any other neighborhood in the dataset, but upon further reflection it actually seems very logical. The most likely explanation is that the Taman Duta neighborhood has a few incredibly expensive outliers that pull the average house value in the neighborhood much higher. Since the coefficient of the Taman Duta neighborhood predictor in our final model is such a large positive number, this most likely means that the model predicts any house in that neighborhood to have a very high listing price, when in reality only a few houses in the neighborhood have such high values, and the average is most likely much more normal. Given more time and computational power, it would be interesting to explore the inclusion of more neighborhood dummy variables, and whether or not they would be useful, as well as the property type.