# MAT 303 Project One Summary Report

Andrea Dale

Andrew.dale@snhu.edu

Southern New Hampshire University

## Introduction

The purpose of this report is to look at different models and how we at Eoreza Housing can utilize these models to better predict housing prices so that we can maximize profit while keeping the houses moving quickly. Throughout the models we will be doing significance tests to make sure that the predictors we are using actually have an impact on housing price. We will also be paying attention to correlation coefficients known as r-squared which help show how strongly a given mode’s predictors have a relationship to the response variable. It is important to note as we go through this report though that a strong correlation does not always mean a good prediction model. As it is important to check the significance tests, residuals against fitted values and Q-Q plot as well.

## Data Preparation

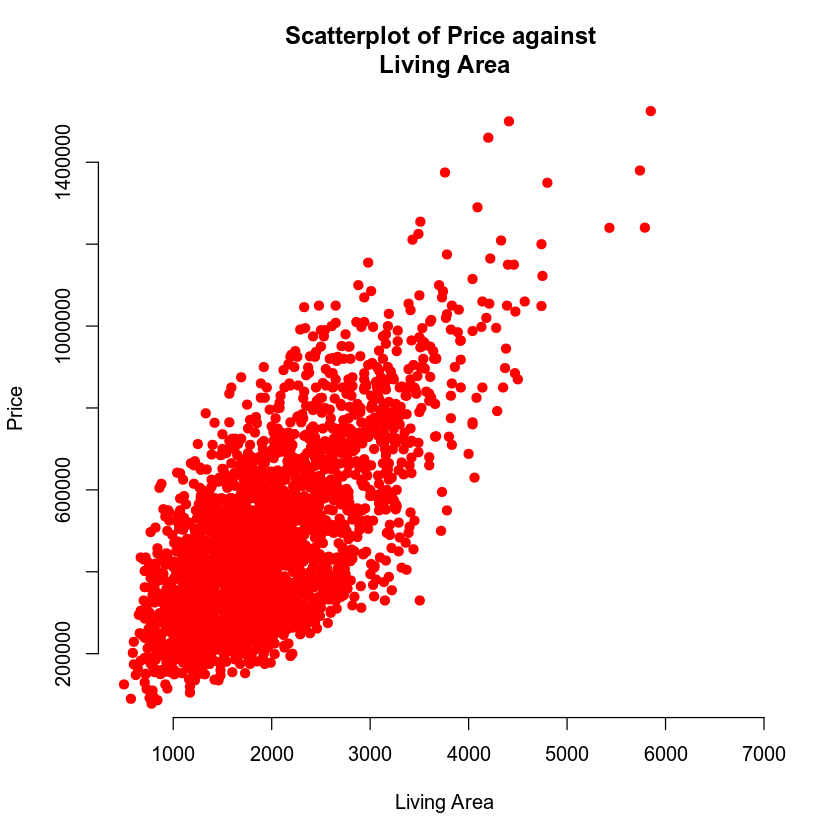
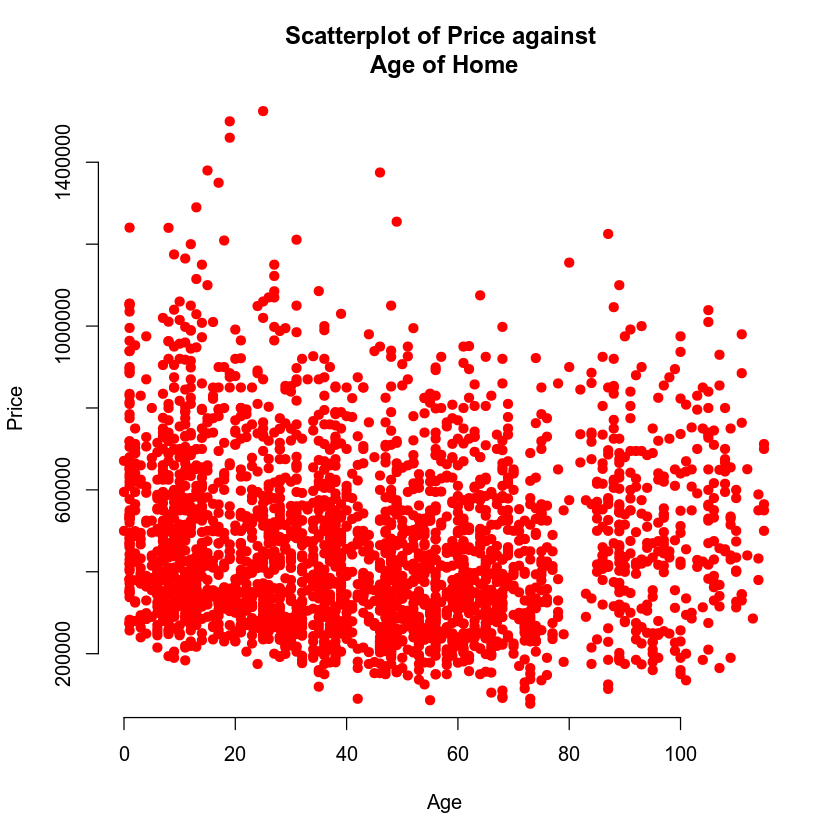
| **Variable** | **What does it represent?** |
| --- | --- |
| price | Sale price of the home |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms |
| sqft\_living | Size of the living area in sqft |
| sqft\_above | Size of the upper level in sqft |
| sqft\_lot | Size of the lot in sqft |
| age | Age of the home |
| grade | Measure of craftsmanship and the quality of materials used to build the home |
| appliance\_age | Average age of all appliances in the home |
| crime | Crime rate per 100,000 people |
| backyard | Home has a backyard (backyard=1) or not (backyard=0) |
| school\_rating | Average rating of schools in the area |
| view | Home backs out to a lake (view=2), backs out to trees (view=1), or backs out to a road (view=0) |

The above table shows all of the variables within our data set. The important ones that will be used in our models are price(the price of a house) and will be our response variable we are trying to predict. For predictor values we will be looking at living area(sqft\_living), age of the home(age), upper level(sqft\_above), view(view), average school rating of area(school\_rating), crime rate per 100,000 people(crime). The total rows and columns for our data set is 2692 and 23 respectively.

## 3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

### Correlation Analysis

The first thing we will want to look at as far as correlation analysis is the scatterplots. Which we will do one of price vs the living area and price vs age. Which are the two following scatterplots:



The first graph has a clear correlation that as living area goes up. Price goes up. Making it a good predictor of price and worth looking at in multiple models. The second graph however is flat across at every price point meaning that age has little or perhaps no effect on the price someone is willing to pay for their home. That is the first flag for age and will be something important for us to keep an eye on throughout additional models and tests.

**Table 1. R squared and adjusted R squared**

|  |  |  |
| --- | --- | --- |
| Model with Predictor | R2 | Adjusted R2 |
| Living Area | 0.4754 | 0.4752 |
| Age | 0.0056 | 0.0052 |

The above table shows the correlation coefficients of our two models using just living area and age as predictors for housing price. What we can tell from each model is that one age is not scientifically a good predictor of price. More surprising though is living area is not as well, at least by itself. That is what we can get from the living area’s correlation coefficient when we can clearly see there is a relation between the living area and price in its scatterplot. We can verify this by looking at the p-value of living area to see if it is significant or not. A positive correlation means that as living area increases so does the price.

### Reporting Results

For this section the management of Eoreza Housing want a regression model using price as the response variable. The first step of creating a model to test is to set up its general form and prediction equations at the base of foundational level. The general form is:

E(y)=β0 + β1 x1 + β2 x2 + β3 x3 + β4 x4 +β5 x5 + β6 x6

The prediction equation for this model is going is:

y^=β0^ + β1^ x1 + β2^ x2 + β3^ x3+ β4^ x4 + β5^ x5+ β6^ x6

Where B5 and B6 are dummy variables for the qualitative variable View And Bi where I > 0 and I < 5 is the index of the predictor variable.

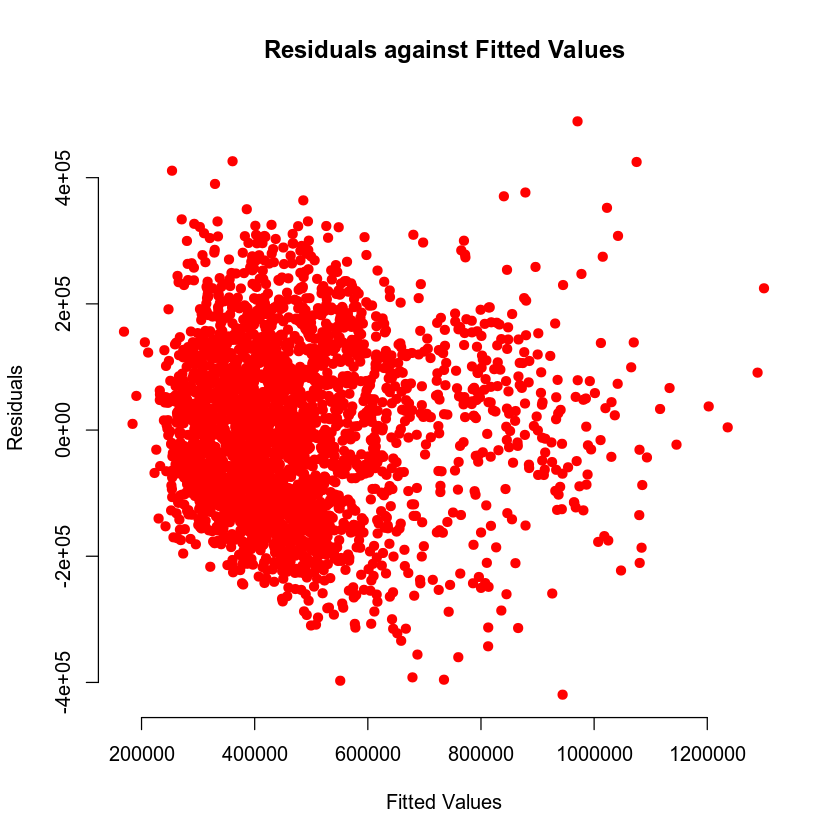
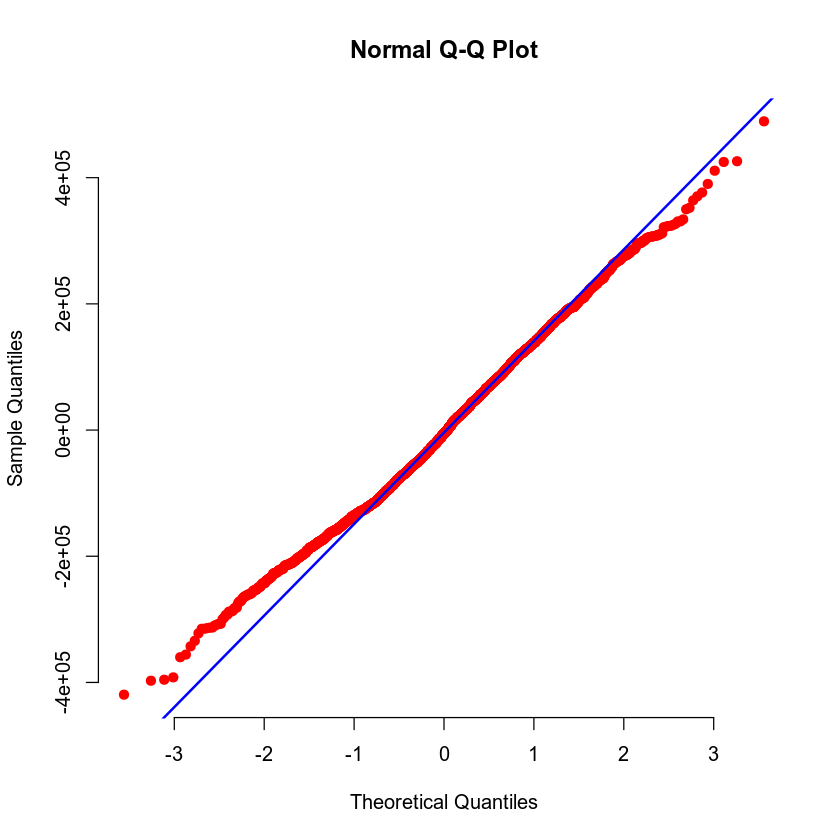
Once we have the model built we will want to look at the r-squared values to get a feel of the correlation between our model and housing price.

**Table 2. R squared and adjusted R squared**

|  |  |
| --- | --- |
| R2 | Adjusted R2 |
| 0.6029 | 0.602 |

Given that the closer r-squared is to 1 or -1 the better the correlation is between the predictors and the response variable. This model might not be the best choice to use when it comes to predicting housing price but there is more statistics for us to look at first. Such as the residuals against fitted values below and Q-Q plot.

Before we get to those though I was asked to interpret the beta values of the living area and lake view. 1.293e+02 is the beta value for sqft\_living which means for every sqft of living space in this model housing price goes up 129.30 whereas a lake view is 2.490e+05 meaning a lake view will increase price by 249,000.



From these plots we can see that data is normally distributed and that the residuals are evenly spaced around 0. While the correlation does not seem to be the strongest the validity of the model is good.

### Evaluating Significance of Model

The fist step of testing the significance level of a model and its predictors is to set the null hypothesis and alternative hypothesis. To note for our text it is the slope of any given variable would be 0 where the alternative is at least one is not 0. The individual tests however look at each variable and its relationship to price and the model. The hypothesis for our test is:

H0: βi = 0

Ha: βi =/= 0

Where i is equal to the index of the predictor variable.

For the individual terms we get hypothesis of:

H0: β1 = 0 or β2 = 0 or β3 = 0 or β4 = 0 or β5 = 0 or β6 = 0

Ha: β1 =/= 0 or β2 =/= 0 or β3 =/= 0 or β4 =/= 0 or β5 =/= 0 or β6 =/= 0

Where or is the separation of individual tests.

For the f-test we have a p-value of 2.2e-16 meaning the test is very close to 0 and that the overall model is significant against a test value of 5% or .05. Our individual values for each predictor are in the table below:

**Table 3. Individual P-Values of predictor Variables**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Living Area(X1) | Upper Level (X2) | Age(X3) | Bathrooms (X4) | View1(X5) | View2(X6) |
| 2e-16 | 0.0089 | 2e-16 | 9.13e-13 | 2e-16 | 2e-16 |

What we know from the table is that all predictors in the model are statistically significant. Which means while the other stats are not where we fully want them. Every variable in this model has some kind of effect on price. Which means it might take multiple tested models to find the right one to use as our prediction model to have higher quality predictions.

### Making Predictions Using Model

When doing predictions using a model the first thing to do is set the prediction equation but with the beta values from our model. So our equation becomes:

y^=7,709 + 129.30 x1 + 19.51 x2 + 1,451x3+ 43,970 x4 + 167,500 x5+ 249,000x6

I was given two sets of data to use as tests to find the predicted house price using this model. The first set of data was 2150 sqft of living area, 1050 sqft in an upper level, 15 years old with 3 bathrooms and backs out onto a road. This house would have a predicted price of $459,828.20. With this data it gives us prediction and confidence intervals of:

**Table.4 Confidence and Prediction Interval of Model3 prediction test 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Fit | Lower | Upper |
| Prediction | 459828.2 | 239563 | 680093.4 |
| Confidence | 459828.2 | 446087.9 | 473568.5 |

What this means from the table above is with this data and model we can say confidently that any single housing price will be between 239,563 and 680,093.40 and the price means of like data sets will be between 446,087.90 and 473,568.50.

For the second set of data we have a house of 4250 living area with a 2100 upper level both in sqft. Then it is 5 years old with 5 bathroom and backs out to a lake. This will give us a predicted price of $1,074,285.39. This data with model3 leading us to prediction and confidence intervals of:

**Table.5 Confidence and Prediction Interval of Model3 prediction test 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Fit | Lower | Upper |
| Prediction | 1074285.39 | 852522.6 | 1296048 |
| Confidence | 1074285.39 | 1045117 | 1103454 |

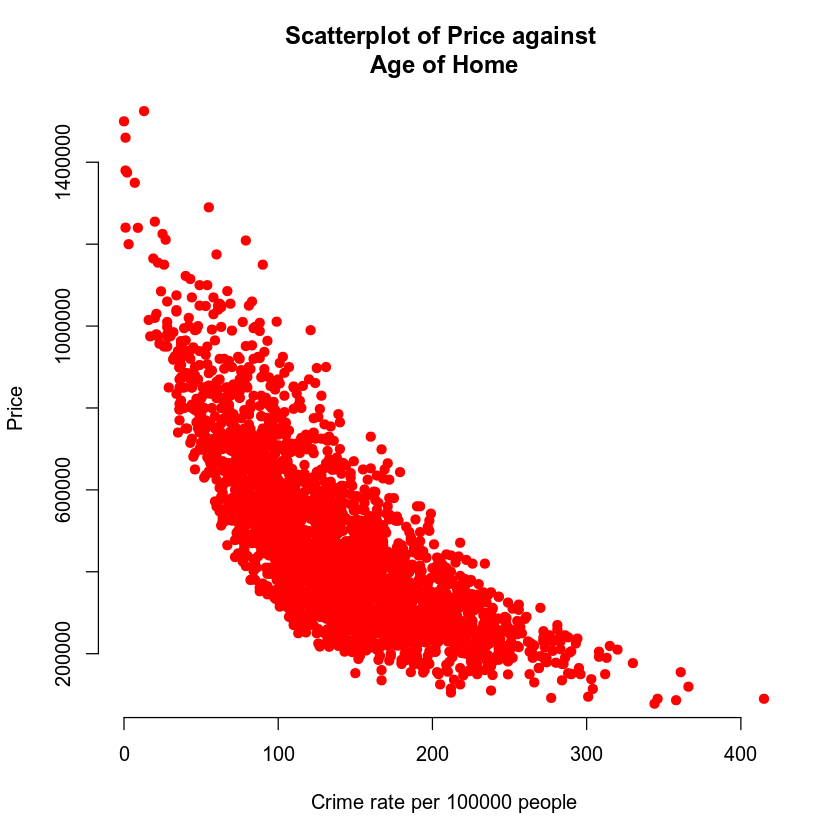
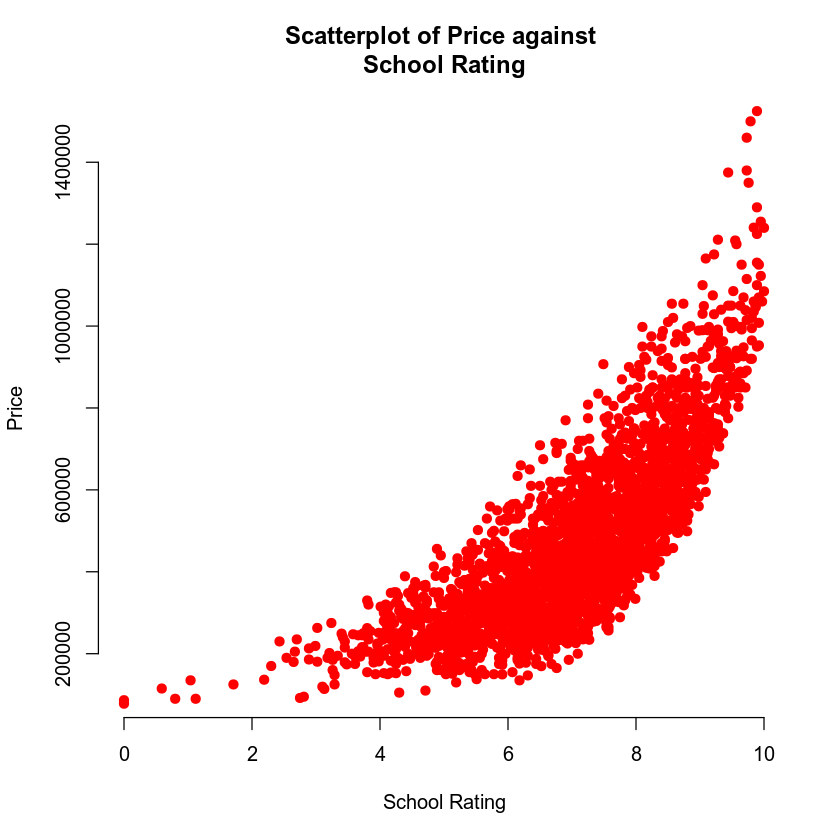
Like with prediction test 1 the prediction interval says of like data we could predict that a price would be between 852522.6 with 90% confidence. As only 10% of values would be outside of that. The confidence interval states though the mean of same sized data sets of like values being between 1045117 and 1103454.

The prediction interval is wider because it is accounting for any possible value of the housing price. Where the confidence interval is accounting for the mean of multiple datasets. There will be values that can be higher or lower individually however, the mean is easier to predict given like data and data set sizes.

## 4. Model #2 - Complete Second Order Regression Model with Quantitative Variables

### Correlation Analysis

The first step of using our new predictors is to create a scatterplot of each to see how the data set might correlate to housing price. Those graphs are:



From the two graphs above we can clearly see that both have a relationship with price. For crime rate it is a curved graph that as crime rate gets higher housing price decreases. Where with school rating we have a curved graph that as school rating increases so does the price of a house. It is important to note that since each of these are curved graphs there is a squared value for them to go with their normal value as well.

### Reporting Results

First step of setting up model4 is setting the general form and prediction equations for our second order model. Those equations starting with general form are:

E(y)=β0+ β1 x+ β2 x2+ β3 x12+ β4 x22

For our prediction equation we have:

ŷ =β0^+ β1^ x + β2^ x2+ β3^ x12+ β4^ x22

Where ŷ  is the predicted value of house price and β0^,β1^, β2^, β3^, β4^ are estimates of β0 ,β1 ,β2, β3, β4 respectively.

The next step is to write our prediction equation using the beta estimates from our model which gives us:

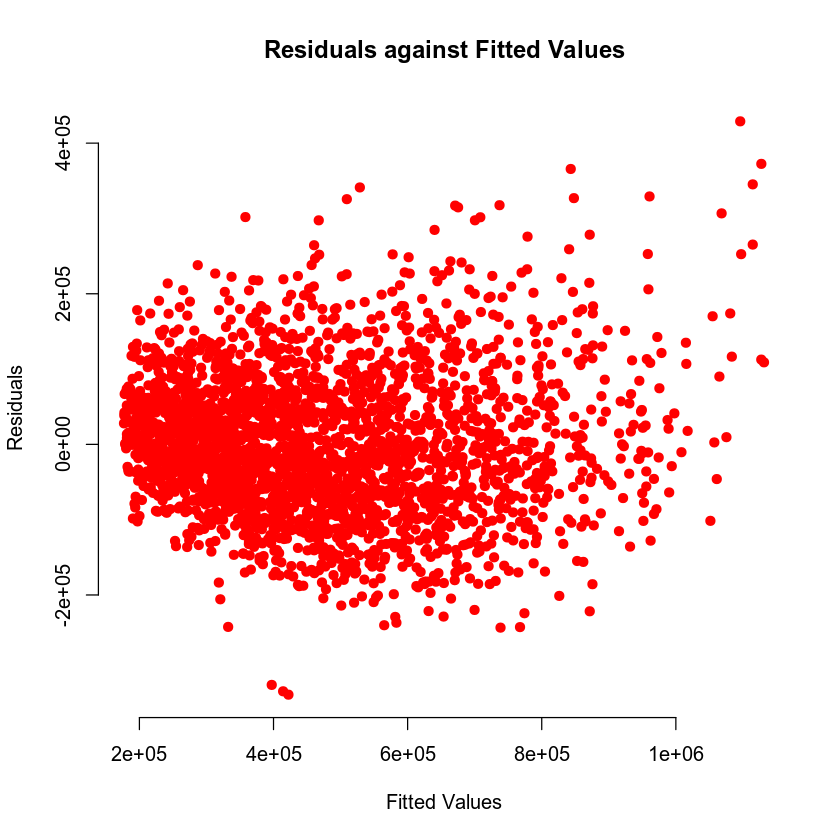
ŷ =841000- 94140x -3696 x2+ 12600 x12+ 6.995 x22

**Table 6. R squared and adjusted R squared**

|  |  |
| --- | --- |
| R2 | Adjusted R2 |
| 0.8087 | 0.8084 |

Unlike model3 these r-squared values show a much stronger correlation between our predictors and the response variable of housing price. By having a stronger correlation, it is possible the model will be better at predicting housing price as is shown by the r-squared being closer to 1.

To verify if this a good model though or next step is to check the residuals against fitted values and Q-Q plots for abnormalities.



Looking at our graphs above while there does not appear to be any issues with the residuals being they flutter about the same above and below about 0. The Q-Q plots tail at the end gives reasons to doubt normality of the model and data and other comparisons of predictors and models might be needed.

### Evaluating Significance of Model

For our significance test of model4 the first task will be setting the null and alternative hypothesis which are:

For the F-test or our overall test of the model’s significance:

H0: βi = 0

Ha: βi =/= 0

Where i is equal to the index of the predictor variable.

For the individual terms we get hypothesis of:

H0: β1 = 0 or β2 = 0 or β3 = 0 or β4 = 0

Ha: β1 =/= 0 or β2 =/= 0 or β3 =/= 0 or β4 =/= 0

Where or is separating individual tests.

Our F-test p-value for our model is 2.2e-16 meaning that our overall model is scientifically significant and a good candidate to be a predictor of housing price. But we have to look at the individual results as well.

**Table 7. Individual P-Values of predictor Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| School rating(X1) | Crime Rate (X2) | School rating^2(X1^2) | Crime^2(X2^2) |
| 2e-16 | 2e-16 | 2e-16 | 2e-16 |

The good news is that all of our variables are significant however because they all have the same p-value it is a bit inconclusive as to the model being the best fit given the tail before. One thing to look into is the possible interaction between these two predictors.

### Making Predictions Using Model

Since we did the prediction equation update in a previous section, we can move into the prediction tests using model4. First test is with a school rating of 9.80 and a crime rate of 81.02. This data gives us a housing price of $875,478.21 and prediction and confidence intervals of:

**Table.8 Confidence and Prediction Interval of Model4 prediction test 1**

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Fit | Lower | Upper |
| Prediction | 875478.21 | 722590.2 | 1028366 |
| Confidence | 875478.21 | 864766.6 | 886189.8 |

From the table above we prediction interval with a lower and upper bound of 722,590.20 and 1,028,366. This means based on this model we can say any predicted house price will 90% of the time be in that range. Where the confidence interval is saying the mean price in our model of like data sets will be between 864766.60 and 886189.8.

For our second prediction test the data is a school rating of 4.28 and a crime rate of 215.50 which gives us a predicted house price of $197,345.37 with prediction and confidence intervals of:

**Table.9 Confidence and Prediction Interval of Model4 prediction test 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | Fit | Lower | Upper |
| Prediction | 197345.37 | 44668.41 | 350023.3 |
| Confidence | 197345.37 | 190257 | 204433.8 |

As was the case with the first test that for like data sets and predictors we can say 90% of the time that any predicted housing price will be between 44,668.41 and 350,023.30. With mean prices being between 190,257 and 204433.8.

## 5. Nested Models F-Test

### Reporting Results

Before we make model5 it is first important to set up the general form and prediction equation of the model. For this model those are:

General Form:  
E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2

Prediction Equation:

ŷ =β0^+ β1^ x1 + β2^ x2+ β3^ x1 x2

where ŷ  is the predicted value of the housing price, x1 is school rating, x2 is crime rate.β0^,β1^, β2^, β3^  are estimates of β0 ,β1 ,β2 , β3  respectively

With these made we can run our model through R and get the beta estimates. This gives us an updated prediction equation of:

ŷ =-410233.37+ 155559.97 x1 + 2230.07x2-564.85 x1 x2

### Evaluating Significance of Model

First step of any significance test is to set the null and alternative hypothesis which for model5 are as follows:

For the F-test or our overall test of the model’s significance:

H0: βi = 0

Ha: βi =/= 0

Where i is equal to the index of the predictor variable.

For the individual terms we get hypothesis of:

H0: β1 = 0 or β2 = 0 or β3 = 0

Ha: β1 =/= 0 or β2 =/= 0 or β3 =/= 0

Where or is separating individual tests.

The f-test p-value for model5 is 2.2e-16 meaning that our overall model is scientifically significant and a good candidate to be a predictor of housing price. The next step being to look at the individual significance of each predictor which is in the table below:

**Table 10. Individual P-Values of predictor Variables**

|  |  |  |
| --- | --- | --- |
| School rating(X1) | Crime Rate (X2) | School rating:Crime(X1X2) |
| 2e-16 | 2e-16 | 2e-16 |

From the table above we can clearly see that all three predictors are statistically significant and valid predictors of housing price. Meaning that this model or this model combined with another could be the most desired for giving us a good view of what a home’s price should be.

General Form:  
E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2

Prediction Equation:

ŷ =β0^+ β1^ x1 + β2^ x2+ β3^ x1 x2

where ŷ  is the predicted value of the housing price, x1 is school rating, x2 is crime rate.β0^,β1^, β2^, β3^  are estimates of β0 ,β1 ,β2 , β3  respectively

And for model4:

### Model Comparison

For this section management of Eoreza Housing would like like a nested f-test done in order to compare model5 to model4. How this works is you take a model that doesn’t have as many predictors such is the case with model5 and compare it against model4 which has more predictors using a chosen significance level. Which in our case is 5%. To restate the general form and prediction equations of each model for model5 we have:

E(y)=β0+ β1 x+ β2 x2+ β3 x12+ β4 x22

For our prediction equation we have:

ŷ =β0^+ β1^ x + β2^ x2+ β3^ x12+ β4^ x22

Where ŷ  is the predicted value of house price and β0^,β1^, β2^, β3^, β4^ are estimates of β0 ,β1 ,β2, β3, β4 respectively.

For this test we are figuring out if the quadratic model is better than the one with only an interactive term so our null hypothesis and alternative hypothesis fit this test are:

H0: βj = βk = 0

Ha: βj =/= 0 or βk =/= 0

Where βj and βk are the bets estimates for school\_rating^2 and crime ^2.

Doing the comparison test for these models gives us a p-value of 2.770393e-29. This means that without a reasonable doubt that the quadratic predictors are needed to get an accurate prediction of housing prices.

## 6. Conclusion

To close out this report I will say I am not confident in using any of the models asked for by the decision makers of Eoreza Housing to predict housing price. With that said given the data tested and the available data I have included in the attached document a model6 with like tests done on it as to the other models that I feel would be a more ideal model to use as a predictor of housing price.