MAT 303 Project Summary Report

Bailey Davidson
Bailey.davidson@snhu.edu
Southern New Hampshire University

1. Introduction

The data set to be analyzed is economic data that gives information on features that are included in a house. All of these variables are important in determining the price of the home and helps people plan out buying and selling homes as well as the government determining laws and taxes.

To analyze this data, we are going to do a First Order Regression mode, then a Second Order Regression Model, and then compare the statistics of a reduced model and a complete model to determine which one would be best to use when making future predictions on housing prices. First, we will look at the relationship between the square feet of main living area, square feet of upper living area, age, bathrooms, if the house has a view of trees versus the view of a lake, and how all of these features affect the price of a house. Then, we are going to create a second regression model looking at how the local school rating and crime score, their interaction term, and how each one of these affects the price of a house.

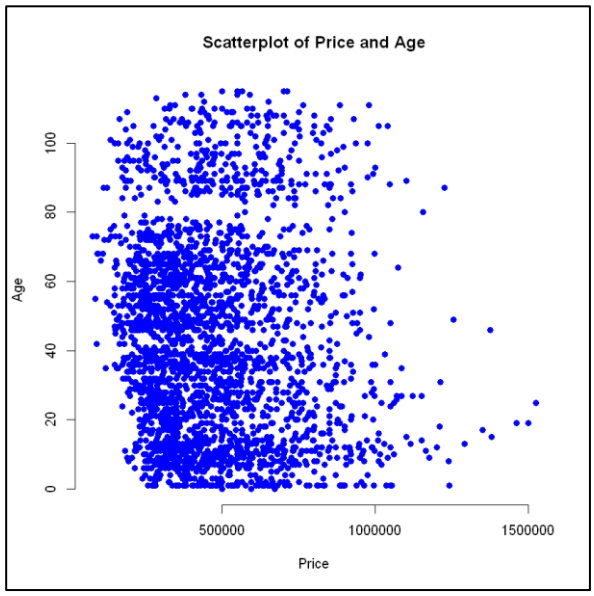
Lastly, we are going to create a first order regression model also using the school rating and crime score. Mainly to compare the output statistics with those of the Second Order Regression model to find out if those extra terms the second order model has makes the performance of the model better.

2. Data Preparation

There are 2692 rows and 23 columns in this data set. The columns are all features that are included in a house. Amongst these features, we chose to take a closer look at the price, number of bedrooms, number of bathrooms, different calculations of square feet, different views, age of home, local school rating, and local crime score.

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

Correlation Analysis



The scatterplot of price vs square foot shows that there is a positive linear relationship between these two. Also, the majority of house prices are under $1 million, and the area is under approximately 3500 square feet. On the other hand, the scatterplot of price and age doesn't show a clear relationship and, again, we see the majority of house prices are under $1 million. However, the age of these houses vary greatly.

The correlation coefficient between *Price vs. the living area* (square feet) is 0.6895, which justifies the scatterplot which shows a relatively positive trend, but with a skew towards the bottom left. This is the same with the correlation coefficient of *Price vs. age* which is -0.0746. This says there is barely a relationship between the two variables, as seen in the scatterplot above.

Reporting Results

The general form of the regression model for price as a response variable with living area, upper-level area, age of the home, number of bathrooms, and view as predictor variables:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6$$

Where y is the predicted value of *price* and x_1 is *sqft_living*, x_2 is *sqft_above,* x_3 is *age,* x_4 is *bathrooms,* x_5 is *view1*, and x_6 is *view2.* $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6$ are estimates of $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, \beta_4, \beta_5, \beta_6$ respectively.

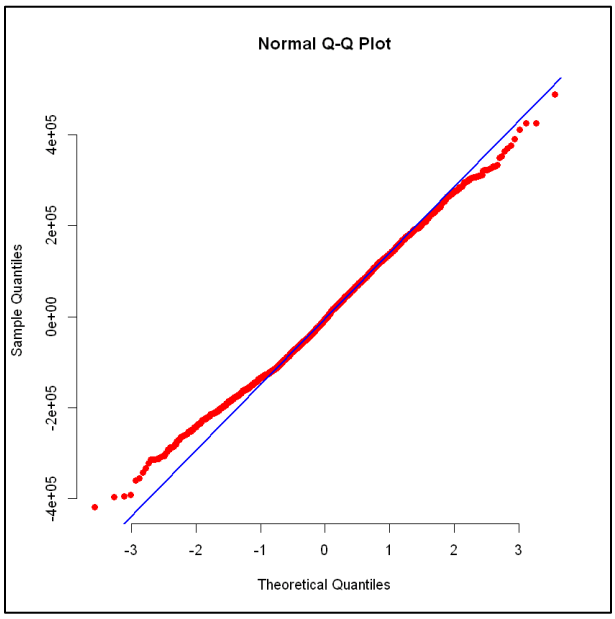
```
Call:
lm(formula = price ~ sqft_living + sqft_above + age + bathrooms +
    view, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-419299 -101792   -5606   93896  489323

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.709e+03  1.411e+04   0.546  0.58495
sqft_living 1.293e+02  8.123e+00  15.916  < 2e-16 ***
sqft_above  1.951e+01  7.458e+00   2.616  0.00894 **
age         1.451e+03  1.199e+02  12.098  < 2e-16 ***
bathrooms   4.397e+04  6.126e+03   7.178 9.13e-13 ***
view1       1.675e+05  1.071e+04  15.640  < 2e-16 ***
view2       2.490e+05  1.201e+04  20.739  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133600 on 2685 degrees of freedom
Multiple R-squared:  0.6029,    Adjusted R-squared:  0.602
F-statistic: 679.3 on 6 and 2685 DF,  p-value: < 2.2e-16
```

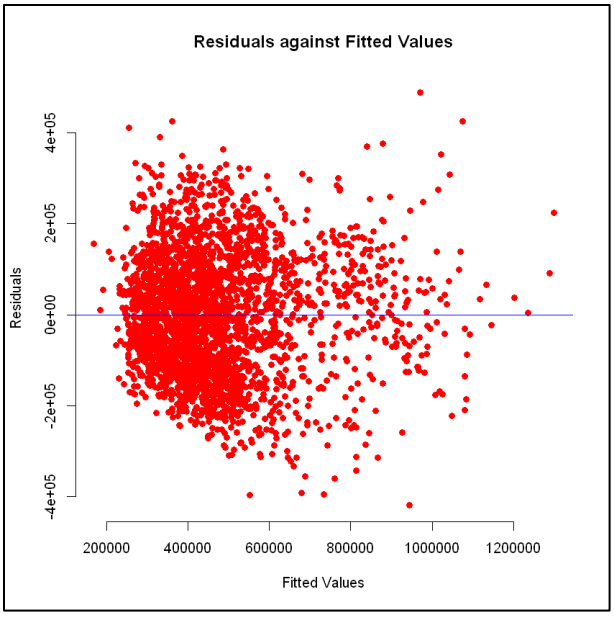
Prediction model equation using outputs obtained from your R script:

$$price = 7709 + 129.3x_1 + 19.51x_2 + 1451x_3 + 43970x_4 + 167500(view1) + 24900(view2)$$

The R-squared value is 0.6029 and the adjusted R-squared is 0.602. This tells us the proportion of the variance in price that is explained by the regression model. So, we are simply looking at the percentage that is predictable by the model (regression).

The beta results for living area in square-feet is 129.3, which means that when an additional square foot is added to the house the value increases by $129.30. There is a similar measure if the house has a lake view, which is represented by the number 2 in the data set. If the house has a lake view the price is $249,000 more than the houses with a view of a road or trees.





The "Residuals against Fitted Values" is biased and heteroscedastic because the plotted values do not equal zero across any thin vertical strip and the spread of values are not the same in any vertical strip. (Jost, 2017)

The "Normal Q-Q Plot" has "strong positive linear correlation", if not on the upper end of a "moderate positive linear correlation". This is because the points between –1 and 2 "Theoretical Quartiles" are directly on the qqline. However, the points on either end vary from the qqline and are not "strong". (Newcastle University. (n.d.))

Evaluating the Significance of Model

Null Hypothesis: Housing features have no effect on house prices.

Alternative Hypothesis: Housing features have a significant effect on house prices.

P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the model is significant.

Conclusion:
The p-value of this model is less than 2.2e-16, which is much less than 0.05. Therefore, we can reject the null hypothesis and conclude that housing features have a significant effect on their price.

Individual beta tests:
P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the variable is significant.

Null Hypothesis (H_0) for Each Term: This specific housing feature has no effect on house prices.

Alternative Hypothesis (H_1) for Each Term: This specific housing feature has a significant effect on house prices.

Confidence Intervals:

```
[1] "confint"
```

A matrix: 7 × 2 of type dbl

	5 %	95 %
(Intercept)	-15513.0058	30931.1022
sqft_living	115.9193	142.6499
sqft_above	7.2402	31.7839
age	1253.3171	1647.9160
bathrooms	33890.4136	54049.8359
view1	149870.7002	185112.3814
view2	229279.0983	268796.5052

Confidence intervals from R output:

Intercept

1. The confidence interval includes zero, thus we fail to reject the null hypothesis.
2. The t-value is 0.546, close to the mean, and the p-value is 0.58495, higher than 0.05 (5%).

Living area

3. The confidence interval does not include zero, thus we reject the null hypothesis.
4. The t-value is 15.916, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Upper-level area

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 2.616, not close to the mean, and the p-value is 0.00894, less than 0.05 (5%).

Age

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 12.098, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Bathrooms

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 7.178, close to the mean, and the p-value is 9.13e-13, less than 0.05 (5%).

View1

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 15.640, close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

View2

1. The confidence interval does not include zero, thus we reject the null hypothesis.

2. The t-value is 20.739, close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Conclusion:
The variables that are significant at 5% are the living area (sqft), upper-level area, age, bathrooms, view1, and view2, because they all have a p-value less than 0.05 and their t-values are not close to the mean, specifically more than 2. In addition to not including zeros in their confidence intervals. This means that we reject the null hypothesis and can say that for every change in each of these features, the price of a house is significantly affected.

Making Predictions Using Model

The predicted price for a home that has 2150 sqft living area, 1050 sqft upper-level living area, is 15 years old, has 3 bathrooms, and backs out to road is $459,828.20. At a 95% confidence level, the prediction intervals are [239563, 680093.4] and the confidence intervals are [446087.9, 473568.5]. The prediction intervals tell us that we can be 95% confident that a house with these features will have a price between $239,563.00 and $680,093.40. The confidence intervals tell us that we are 95% confident that the average housing price will fall between $446,087.90 and $473,568.50 for a house with these features.

The predicted price for a home that has *4250* sqft living area, *2100* sqft upper-level living area, is 5 years old, has 5 bathrooms, and backs out to lake is $1,074,285.00. At a 95% confidence level, the prediction intervals are [852522.6, 1296048] and the confidence intervals are [1045117, 1103454]. The prediction intervals tell us that we can be 95% confident that a house with these features will have a price between $852,522.60 and $1,296,048.00. The confidence intervals tell us that we are 95% confident that the average housing price will fall between $1,045,117.00 and $1,103,454.00 for a house with these features.

Prediction intervals are wider than confidence intervals, such as in our example, because they include an additional source of uncertainty, the random variation of individual observations, in this case houses. Rather than just including the uncertainty in estimating the mean, which is the confidence intervals. (Forthofer et al., 2007)

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

Correlation Analysis

A second order model is appropriate when using these variables because there isn't a perfect linear relationship between price and either school rating or crime. It looks like when price increases the school rating has a downward concave and is also increasing. When price increases crime decreases and creates an upward concave.

Reporting Results

General Form:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$$

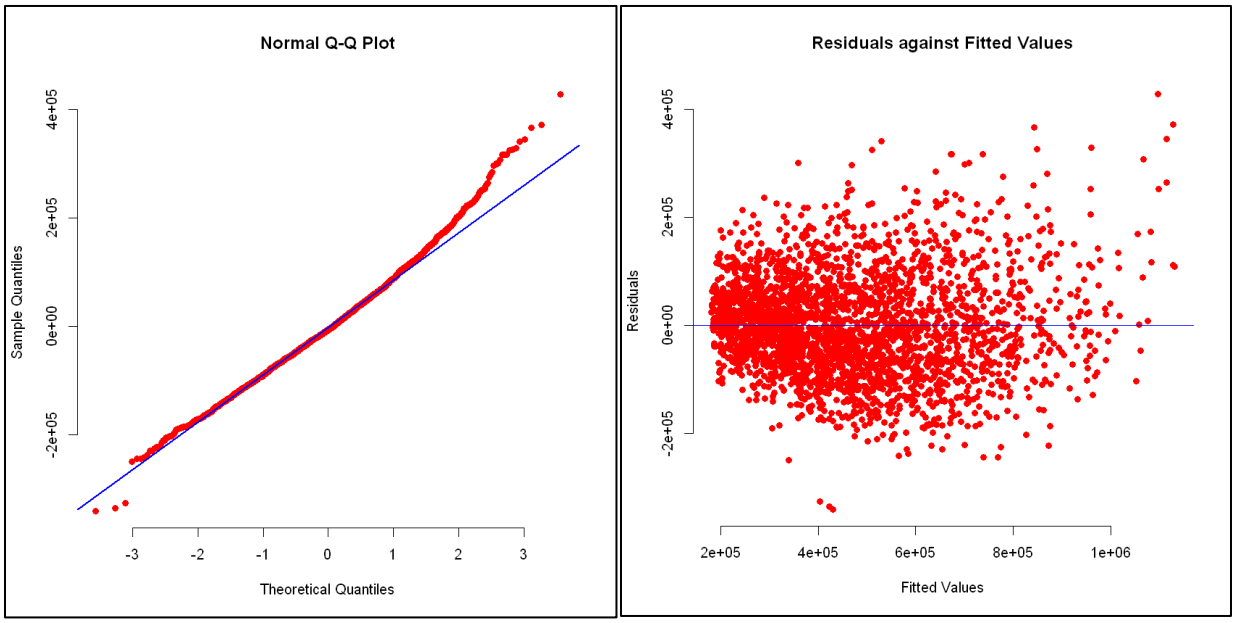
Where y is the predicted value of *price* and x_1 is *school rating*, and x_2 is *crime*.
$\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ are estimates of $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, \beta_4, \beta_5$ respectively.

```
Call:
lm(formula = price ~ school_rating + crime + school_rating:crime +
    I(school_rating^2) + I(crime^2), data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-340729  -61055   -6288   56875  427915

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.339e+05  1.032e+05   7.113 1.45e-12 ***
school_rating       -7.375e+04  2.083e+04  -3.541 0.000406 ***
crime               -3.155e+03  5.235e+02  -6.027 1.90e-09 ***
I(school_rating^2)   1.165e+04  1.109e+03  10.497  < 2e-16 ***
I(crime^2)           6.377e+00  7.265e-01   8.777  < 2e-16 ***
school_rating:crime -5.227e+01  4.853e+01  -1.077 0.281513
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92690 on 2686 degrees of freedom
Multiple R-squared:  0.8088,    Adjusted R-squared:  0.8084
F-statistic:  2272 on 5 and 2686 DF,  p-value: < 2.2e-16
```

Prediction equation with the outputs from my R code the new prediction model is:

$$\hat{y} = 733900 - 73750x_1 - 3155x_2 - 52.27x_1x_2 + 11650x_1^2 + 6.377x_2^2$$

The R-squared value is 0.8088 and the adjusted R-squared is 0.8084. This tells us the proportion of the variance in MPG that is explained by the MLR model. So, we are simply looking at the percentage that is predictable by the model (regression).



The “Normal Q-Q Plot” has “strong positive linear correlation”, if not on the upper end of a “moderate positive linear correlation”. This is because the points between –3 and 1.5 “Theoretical Quartiles” are

directly on the qqline. However, the points on right end vary from the qqline and are as not “strong”. (Newcastle University. (n.d.))

The “Residuals against Fitted Values” is biased and heteroscedastic because the plotted values do not equal zero across any thin vertical strip and the spread of values are not the same in any vertical strip. Especially in the top-right vs bottom-right and the middle-left vs middle-right of the graph. (Jost, 2017)

Evaluating the Significance of Model

Null Hypothesis: Housing features have no effect on house prices.

Alternative Hypothesis: Housing features have a significant effect on house prices.

P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the model is significant.

Conclusion:
The p-value of this model is less than 2.2e-16, which is much less than 0.05. Therefore, we can reject the null hypothesis and conclude that housing features have a significant effect on their price.

Confidence Intervals:

```
[1] "confint"
```

A matrix: 6 × 2 of type dbl

	5 %	95 %
(Intercept)	564127.8529	903693.6158
school_rating	-108021.8320	-39474.5138
crime	-4016.0800	-2293.4529
I(school_rating^2)	9821.2268	13472.3788
I(crime^2)	5.1816	7.5724
school_rating:crime	-132.1146	27.5761

Individual beta tests:

P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the variable is significant.

Null Hypothesis (H_0) for Each Term: This specific housing feature has no effect on house prices.

Alternative Hypothesis (H_1) for Each Term: This specific housing feature has a significant effect on house prices.

Confidence intervals from R output:

Intercept

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 7.113, not close to the mean, and the p-value is 1.45e-12, less than 0.05 (5%).

School Rating

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is -3.541, not close to the mean, and the p-value is 0.000406, less than 0.05 (5%).

Crime

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is -6.027, not close to the mean, and the p-value is 1.90e-09, less than 0.05 (5%).

School Rating-Squared

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 10.497, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Crime-Squared

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is 8.777, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

School Rating:Crime

1. The confidence interval includes zero, thus we fail to reject the null hypothesis.
2. The t-value is -1.077, close to the mean, and the p-value is 0.281513, higher than 0.05 (5%).

Conclusion:
All of the variables are significant at 5% besides the School Rating:Crime interaction term, which is close to the mean and has a high p-value. The variables that are significant also do not include zeros in their confidence intervals. Therefore, we can reject our null hypothesis and conclude that all of these housing features in this model, besides the interaction term between School Rating and Crime, have a significant effect on the price of a house. In other words, when a housing feature changes it can significantly change the price of that house.

Making Predictions Using Model

The predicted price for a home that in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is $874,497.00. At a 95% confidence level, the prediction intervals are [721606.2, 1027388] and the confidence intervals are [863681.4, 885312.7]. The prediction intervals tell us that we can be 95% confident that a house with these features will have a price between $721,606.20 and $1,027,388.00. The confidence intervals tell us that we are 95% confident that the average housing price will fall between $863,681.40 and $885,312.70 for a house with these features.

The predicted price for a home in an area with an average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is $199,706.70. At a 95% confidence level, the prediction intervals are [46991.65, 352421.7] and the confidence intervals are [191753.5, 207659.9]. The prediction intervals tell us that we can be 95% confident that a house with these features will have a price between $46,991.65 and $352421.70. The confidence intervals tell us that we are 95% confident that the average housing price will fall between $191753.50 and $207659.90 for a house with these features.

5. Nested Models F-Test

Reporting Results

General Form:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

Where y is the predicted value of *price* and x_1 is *school_rating*, x_2 is *crime,* and $x_1 x_2$ is *school_rating: crime.* $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \hat{\beta}_3$ are estimates of $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3$ respectively.

```
Call:
lm(formula = price ~ school_rating + crime + crime:school_rating,
    data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-336984  -63754   -4397   58894  440377

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -410233.37   25261.25  -16.24   <2e-16 ***
school_rating        155559.97    3133.06   49.65   <2e-16 ***
crime                  2230.07     129.70   17.20   <2e-16 ***
school_rating:crime    -564.85      17.86  -31.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94870 on 2688 degrees of freedom
Multiple R-squared:  0.7995,    Adjusted R-squared:  0.7993
F-statistic:  3573 on 3 and 2688 DF,  p-value: < 2.2e-16
```

Prediction model equation using outputs obtained from your R script:

$$price = -410233.37 + 155559.97 x_1 + 2230.07 x_2 - 564.85 x_1 x_2$$

Evaluating the Significance of Model

Null Hypothesis: Housing features have no effect on house prices.

Alternative Hypothesis: Housing features have a significant effect on house prices.

P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the model is significant.

Conclusion:
The p-value of this model is less than 2.2e-16, which is much less than 0.05. Therefore, we can reject the null hypothesis and conclude that housing features have a significant effect on their price.

Individual beta tests:

P-Value: If the p-value is less than 0.05 (5%) we reject the null hypothesis and conclude the variable is significant.

Null Hypothesis (H_0) for Each Term: This specific housing feature has no effect on house prices.

Alternative Hypothesis (H_1) for Each Term: This specific housing feature has a significant effect on house prices.

Confidence Intervals:

```
[1] "confint"
```

A matrix: 4 × 2 of type dbl

	5 %	95 %
(Intercept)	-451798.765	-368667.983
school_rating	150404.774	160715.161
crime	2016.670	2443.479
school_rating:crime	-594.228	-535.464

Confidence intervals from R output:

Intercept

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is -16.24, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

School-Rating

1. The confidence interval does not include zero, thus we reject the null hypothesis.
1. The t-value is 49.65, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Crime

2. The confidence interval does not include zero, thus we reject the null hypothesis.
3. The t-value is 17.20, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

School-Rating:Crime

1. The confidence interval does not include zero, thus we reject the null hypothesis.
2. The t-value is -31.63, not close to the mean, and the p-value is less than 2e-16, less than 0.05 (5%).

Conclusion:

All variables are significant at 5% because they all have a p-value less than 0.05 and their t-values are not close to the mean, specifically more than 2. In addition to not including zeros in their confidence intervals. This means that we reject the null hypothesis and can say that for every change in each of these features, the price of a house is significantly affected.

Model Comparison

The F-Test for nested models is used to determine if there's a statistically significant difference in the fit of two models, with one being "reduced" and another being "complete". The "complete" or "full" model has the variables that the "reduced" model has plus additional ones. The F-test answers the question: "Do the additional variables make the model significantly different?".

Reduced model:

General Form:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$$

Where y is the predicted value of *price* and x_1 is *school_rating*, x_2 is *crime,* and $x_1 x_2$ is *school_rating: crime.* $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \hat{\beta}_3$ are estimates of $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3$ respectively.

Complete model:

General Form:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \ \beta_4 x_1^2 + \beta_5 x_2^2$$

Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \ \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$$

Where y is the predicted value of *price* and x_1 is *school rating*, and x_2 is *crime*.
$\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ are estimates of $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, \beta_4, \beta_5$ respectively.

F-Test Results:

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2686	2.307469e+13	NA	NA	NA	NA
2	2688	2.419501e+13	-2	-1.120319e+12	65.20513	2.22716e-28

Null Hypothesis: The reduced model fits the data as well as the complete model.
Alternative Hypothesis: The complete model fits the data significantly better than the reduced model.

P-Value: We want the p-value less than 0.05 to be able to reject the null hypothesis, the F-test gives us 2.22716e-28.

Conclusion: The p-value is less than 0.05, therefore we reject the null hypothesis. The F-statistics is 65.20513, which is the difference between the two models by calculating the difference in RSS and degrees of freedom difference and is quite large. As a result, the complete model fits the data significantly better than the reduced model.

This means that the additional terms, school_rating-squared and crime-squared are needed to make more accurate predictions of housing prices based on school ratings and crime scores.

6. Conclusion

I would choose the complete second order regression model with quantitative variables. I would choose the complete over the reduced because it is significantly better based on the nested F-tests completed. I would choose the second order regression model over the first order because the predictors, in this case the features, do not have a perfect linear relationship to the outcome, the price. The adjusted R for the first order model is 0.6029, which means that only 60% of the values are explained by the model. The second order model has a percentage of almost 81%. Even though the variables are different, each variable in both models have individual significance on a house's price. Also, the first order model has more significant variables than the second order, so it should have a higher adjusted R-squared. However, the second order model is just a better fit and I recommend to use this one for better predictions of house prices.

A practical importance of this analysis would be for home buyers and sellers to predict price based on these features. Sellers would be able to list their house for an appropriate amount and buyers would be able to make appropriate offers to the sellers. Also, government agencies can use this information to help plan on movements towards affordable housing. Another example could be for insurance agencies to use the crime score and price of house to help set premiums. These are only a handful of examples, amongst countless, that this analysis can be used for.

7. Citations

Forthofer, R. N., Lee, E. S., & Hernandez, M. (2007). Interval estimation. In *Biostatistics* (pp. 169–212). doi:10.1016/b978-0-12-369492-8.50012-1

Jost, S. (2017). *Linear Regression*. DePaul University College of Computing and Digital Media: IT223. https://condor.depaul.edu/sjost/it223/documents/regress.htm

Newcastle University. (n.d.). Strength of Correlation. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html