# MAT 303 Project One Summary Report

Yaqub Mohamud

yaqub.mohamud@snhu.edu

Southern New Hampshire University

## 1. Introduction

The housing\_v2 data set is the one that is being examined for statistical analysis. This data set's information will be used to examine changes in how various home attributes—like living space, upper-level space, age, number of bathrooms, and view—affect a property's selling price. The results of this research will be used to help the real estate firm establish more accurate prices for the properties listed by their clients. First and second order regression models containing both quantitative and qualitative variables, as well as F-test for model comparison, will be among the analytical techniques used in this project.

## 2. Data Preparation

Thirteen variables will be included in the dataset, but for my analysis, I will be concentrating on the following: price, living area square footage (sqft\_living), age of the house, number of bathrooms, upper-level area square footage (sqft\_upper), view from the house, local crime rating, and school rating. This dataset has 2692 rows of data in total, with 23 columns.

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

### Correlation Analysis

A diagram of a living area

Description automatically generated

A graph of red dots

Description automatically generated

There is a positive trend between living space and price. If the living area increases, the price of the home increases. There is no discernible trend between age of home and price.

The association coefficients for age and living area are -0.0746 and 0.6895, respectively.

The correlation coefficients for living area and home age have the following strengths:

* Living area has a moderately positive association, as indicated by its moderate positive correlation coefficient, which ranges from 0.40 to 0.80.
* Because the correlation coefficient is negative and ranges from 0.0 to 0.40, there is a weak negative association with age.

Overall, compared to the age of the home, living area has a stronger impact on home prices.

### Reporting Results

The general form equation for the multiple regression model using price as the response variable and living area, upper level area, age of the home, number of bathrooms and view as predictor variables is:

Y=β\_0+β\_1 X\_1+β\_2 X\_2+β\_3 X\_3+β\_4 X\_4+β\_5 X\_5

E(y)=β0+β1⋅sqft\_living+β2⋅sqft\_above+β3⋅age+β4⋅bathrooms+β5⋅view

y ̂=7709+129.3⋅sqft\_living+19.51⋅sqft\_above+1451⋅age+43970⋅bathrooms+167500⋅view1+249000⋅view2

The value R2 (R-squared) is 0.6029 and Adjusted R-Squared (adjusted R-squared) is 0.602. This means that approximately 60.29% of the variance in price can be explained by the model. Despite being somewhat lower, the adjusted R-squared value—which takes into consideration the number of predictors in the model—still shows a decent match.

Furthermore, Living Area (sqft\_living): Keeping all other factors equal, the coefficient of 129.3 indicates that the price of the house rises by around $129.30 for every square foot of living area.

The Lake View (view2): Keeping all other factors equal, the coefficient of 249000 shows that properties with a lake view (view2) are around $249,000 more expensive than those with a road view (view0).

A red dot diagram with numbers and a white background

Description automatically generated

A graph with a red line

Description automatically generated

The residual vs. fitted values plot, which lacks a discernible pattern, would serve as the foundation for the homoscedasticity assumption. The points primarily lie along the reference line in the Normal Q-Q Plot, indicating that the residuals are roughly normally distributed. This suggests that the normality assumption is logically satisfied. The residuals' normal distribution and random distribution around zero indicate that the model fits the data well.

### Evaluating Significance of Model

At the 5% level of significance, the model's significance was assessed using the overall F-test. The F-test's null hypothesis states that the model is not significant as all regression coefficients are equal to zero. The alternative hypothesis holds that a meaningful model is implied if at least one coefficient is non-zero. Given that the p-value is less than 0.05 (2.2e-16), we reject the null hypothesis and call the model significant. The p-value for each variable is much below the significance level of 0.05. The alternative hypothesis can be accepted in place of the null hypothesis. Every predictor variable separately has a statistically significant link with the response variable, based on the data.

### Making Predictions Using Model

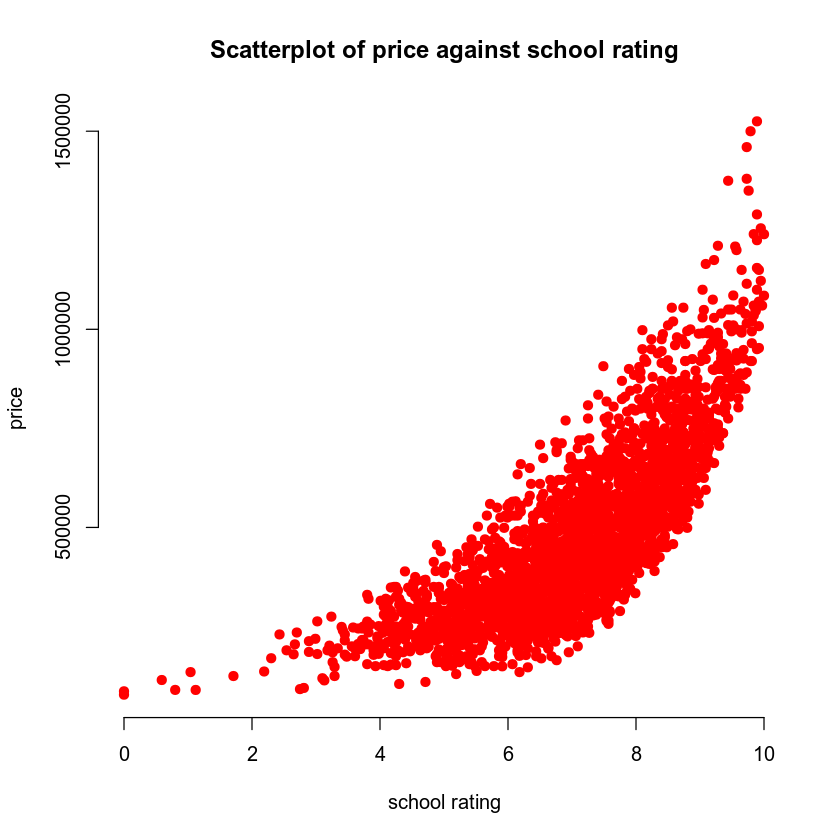
A house that is 15 years old, has three bathrooms, backs out to the road, is 2150 square feet in size, and has an upper-level living space of 1050 square feet is expected to cost $459,828. The range of predictions is [$239,563, $680,093]. This interval leads us to the conclusion that we have 90% confidence that the home's price will fall within this range. There is a [$446,087, $473,568] confidence interval. With 90% confidence, we may infer from this interval that the average home price will fall within this range. This increases the accuracy of calculating the average property price.

A house with 4250 square feet of living space, 2100 square feet of upper-level living space, five years of age, five bathrooms, and a lake view is expected to cost $1,074,285. The range of predictions is [$852,522, $1,296,048]. With 90% certainty, we can say that the home's price will be in this range. There is a [$1,045,117, $1,103,454] confidence interval. With 90% certainty, we can say that this home's mean price will be in this range.

Because the prediction interval takes into consideration both the variability in individual property prices and the uncertainty in the model's projections, it is larger than the confidence interval.

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

### Correlation Analysis

****

**A red dotted graph with white text

Description automatically generated**

A non-linear trending curve was seen in both the scatterplots showing price in connection to the average local school rating and price in relation to the crime rate per 100,000 people. In particular, as the price falls and the crime rate rises, the scatterplot showing the relationship between price and crime rate moves downward. On the other hand, when prices rise and school ratings rise, the scatterplot showing the relationship between price and rating rises higher. For these relationships, it would be reasonable to use a second order model.

### Reporting Results

Using the local average school rating and the crime rate per 100,000 residents as predictors, the general form of a complete second order model for pricing would be:

E(y)=β0+β1⋅school\_rating+β2⋅crime+β3⋅(school\_rating⋅crime)+β4⋅( school\_rating^2)+β5⋅( crime^2)

y ̂=733900+(-73750)⋅school\_rating+(-3155)⋅crime+(-52.27)⋅(school\_rating⋅crime)+11650⋅( school\_rating^2)+6.377⋅( crime^2)

The value R2 (R-squared) is 0.8088 and Adjusted R-Squared (adjusted R-squared) is 0.8084. This means that approximately 80.88% of the variance in price can be explained by the model. Despite being somewhat lower, the adjusted R-squared value still shows a good match. This demonstrates that a significant amount of the variation in home prices can be explained by the model.

A red dot diagram with white text

Description automatically generated

A graph with a red line

Description automatically generated

The plots above show that the residuals show no obvious trend against the fitted values, and the usual q-q plot's points mostly follow the line, with only a few marginally deviating from or exceeding it. The assumption of homoscedasticity is mostly supported by the data from the plots above.

### Evaluating Significance of Model

When using the overall F-test, the alternative hypothesis suggests that at least one of the predictor variables is related to the response variables, while the null hypothesis asserts that there is no significant relationship between the response variables and predictor variables. The alternative hypothesis is accepted and the null hypothesis is rejected since the model is considered significant at the 5% level of significance, with a P-value of 2.2e-16. At the 5% level of significance, it is determined that the factors of age, view, and living room square footage are significant. The alternative hypothesis is rejected and the null hypothesis is accepted when the P-value is less than 5%.

### Making Predictions Using Model

A home in an area with an average crime rate of 81.02 and a school rating of 9.80 is expected to cost $874,497. The range of predictions is [$721,606, $1,027,388]. This interval leads us to the conclusion that we have 90% confidence that the home's price will fall within this range. The interval of confidence is [$863,681, $885,321]. With 90% confidence, we may infer from this interval that the average home price will fall within this range. This increases the accuracy of calculating the average property price.

A home in an area with a crime rate of 215.50 and an average school rating of 4.28 is expected to cost $199,706. The range of predictions is [$46,991, $352,421]. This interval leads us to the conclusion that we have 90% confidence that the home's price will fall within this range. There is a [$199,706, $207,659] confidence interval. With 90% confidence, we may infer from this interval that the average home price will fall within this range.

## 5. Nested Models F-Test

### Reporting Results

The first-order regression model's general form is:  
  
β0+β1x1+β2x2+β3x1x2 = E(y).  
  
Using the results from the R script, the prediction model equation is:   
  
y ̂=-410233.37+155559.97⋅school\\_rating+2230.07⋅crime-564.85⋅(school\\_rating ⋅crime)

### Evaluating Significance of Model

The p-value for this F-test is 2.2e-16, which is rounded to zero. We conclude that the model is significant and reject the null hypothesis as the p-value is less than the 5% significance level. In terms of the correlation between the response variable for house price and other factors, we can say that there is statistical significance for at least one of the predictor variables. Both crime and the relationship between crime rate and school rating had p-values of 2e-16. Since all of the p-values are less than the significance level of 0.05, we exclude the null hypothesis and conclude that every variable is statistically significant.

### Model Comparison

To find out if the extra terms in the more complicated model significantly increase the model's fit, the F-statistic is calculated using the ANOVA procedure for comparing nested models. This is the general Nested F-test formula:

F=((RSSreduced-RSScomplete)/(dfreduced-dfcomplete))/(RSScomplete/dfcomplete)

The full second-order model (Model 2) yields noticeably better predictions than the reduced first-order model, according to the nested models F-test. The p-value of 2.22716e-28 and the F-statistic of 65.20513 show that the second-order model's added complexity is warranted. This implies that the model's capacity to forecast property values based on crime rate and school rating is enhanced with the addition of quadratic and interaction factors.

## 6. Conclusion

In conclusion, the statistical analyses conducted for this project show that the crime rate and school rating are important factors that affect housing costs, and their relationships are not linear. The second-order regression model outperforms the first-order model on the data because it includes both linear and quadratic components for these variables. The quadratic terms significantly improve the model, according to the significance and F-tests for nested models. According to these findings, homes in areas with better school ratings and lower crime rates are worth more. Real estate companies may use this information to better predict and set property prices based on local conditions.