# MAT 303 Project One Summary Report

Brandon Gibbs

brandon.gibbs@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data set being explored is a comprehensive collection of historical data on real estate transactions. This data includes various attributes of houses, such as square footage, number of bathrooms, age of the home, upper-level living area, school rating, crime rate, and views. These attributes will be analyzed to understand their relationship with the houses' selling prices.

The results of this analysis will be used by the real estate company to set more accurate and competitive listing prices for homes. By understanding the impact of different attributes on home prices, the company can ensure that their listings are priced to reflect market conditions, which will help in selling homes within a reasonable amount of time. This information will also aid in advising clients on pricing strategies and potential improvements that could increase their home's value.

The project will involve several types of statistical analyses, including correlation analysis to identify and measure the strength and direction of the relationships between different house attributes and their selling prices. First-order regression models will be used to predict house prices based on key quantitative and qualitative variables. Second-order regression models, which include interaction terms and squared terms, will capture non-linear relationships and interactions between predictors. Additionally, an F-test for model comparison will be conducted to determine if more complex models provide significantly better predictions than simpler ones. ANOVA will be used to analyze the variance and test the significance of different predictors in the model. Finally, confidence and prediction intervals will be provided to quantify the uncertainty in predictions.

By using these analyses, the project aims to build strong models that can accurately predict house prices based on critical variable factors, ultimately aiding in better decision-making for pricing strategies.

## 2. Data Preparation

There are several important variables are used to predict home prices. These variables include both quantitative and qualitative factors that significantly impact the model's accuracy and reliability. Here are the key variables this project will fucus on:

* **Living Area (sqft\_living),** quantitative variable represents the total living area of the home in square feet.
* **Upper Level Area (sqft\_above),** quantitative variable indicates the living area above 2nd floor level.
* **Age of the Home (age)**, quantitative variable measures the age of the home in years. Older homes may have different pricing dynamics compared to newer ones, making this an important factor.
* **Number of Bathrooms (bathrooms)**, quantitative variable represents the total number of bathrooms in the home.
* **View (view)**, qualitative variable indicates the type of view the home has, such as a lake view, road view, or no view.
* **School Rating (school\_rating)**, quantitative variable represents the average school rating in the area.
* **Crime Rate (crime)**, quantitative variable measures the crime rate per 100,000 people in the area.

The dataset being used for this project contains 23 columns and 2692 rows.

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

### Correlation Analysis

A green and white diagram

Description automatically generated

A graph of purple dots

Description automatically generated

A table with numbers and text

Description automatically generated

The scatterplots reveal distinct trends between the variables. The scatterplot of price versus living area shows a clear positive trend, indicating that homes with larger living areas tend to have higher prices. This is supported by the correlation coefficient of 0.6895, which signifies a strong positive relationship. On the other hand, the scatterplot of price versus the age of the home shows a slight negative trend, suggesting that older homes tend to have lower prices. However, the correlation coefficient of -0.0746 indicates that this relationship is weak. Overall, living area has a stronger impact on home prices compared to the age of the home.

### Reporting Results

The general for equation for this regression:

(R-squared) of 0.6029 and Adjusted R-Squared (adjusted R-squared) of 0.602 These values indicate that approximately 60.29% of the variability in home prices can be explained by the model. The adjusted R-squared value, which accounts for the number of predictors in the model, is slightly lower but still indicates a good fit.

Additionally, Living Area (sqft\_living): The coefficient of 129.3 suggests that for each additional square foot of living area, the price of the home increases by approximately $129.30, holding all other variables constant. The Lake View (view2): The coefficient of 249000 indicates that homes with a lake view (view2) are priced approximately $249,000 higher than homes with a road view (view0), holding all other variables constant.

A green and white diagram

Description automatically generated

A graph with purple and green lines

Description automatically generated

The Residuals vs. Fitted Values plot residuals appear to be scattered around zero, with no clear pattern. This indicates that the model's errors are randomly distributed, suggesting that the assumptions of linearity and homoscedasticity are met. While The Normal Q-Q Plot the points mostly lie along the reference line, suggesting that the residuals are approximately normally distributed. This indicates that the normality assumption is reasonably met. The residuals are randomly distributed around zero, and they follow a normal distribution suggests that the model is a good fit for the data.

### Evaluating the Significance of Model

To determine if the model is significant at a 5% level of significance, we perform the overall F-test. The null hypothesis () states that the model with no predictors is as good as the model with predictors (i.e., all regression coefficients are zero), while the alternative hypothesis () states that the model with predictors is better than the model with no predictors (i.e., at least one regression coefficient is not zero). With a P-value of < 2.2e-16, we reject the null hypothesis, indicating that the model is significant at a 5% level of significance. For individual beta tests, the null hypothesis () for each predictor states that the coefficient is zero, while the alternative hypothesis () states that the coefficient is not zero. The P-values for the intercept, living area (), upper level area (), age of the home (), number of bathrooms (), view1 (), and view2 () are 0.58495, < 2e-16, 0.00894, < 2e-16, 9.13e-13, < 2e-16, and < 2e-16, respectively. Since the P-value for the intercept is greater than 0.05, we fail to reject the null hypothesis, indicating that the intercept is not significant at a 5% level of significance. However, the P-values for living area, upper level area, age of the home, number of bathrooms, view1, and view2 are all less than 0.05, leading us to reject the null hypothesis for these predictors. This indicates reveals that living area, upper level area, age of the home, number of bathrooms, view1, and view2 are significant at a 5% level of significance. Overall model is significant at a 5% level of significance, and all individual predictors except the intercept are significant at a 5% level of significance indicating that living area, upper level area, age of the home, number of bathrooms, and view are important predictors of home prices.

### Making Predictions Using Model

Home 1

2150 sqft living area, 1050 sqft upper level living area, 15 years old, 3 bathrooms, backs out to road

Predicted Price: $459,828.20

90% Prediction Interval: $239,563 to $680,093.40

90% Confidence Interval: $446,087.90 to $473,568.50

The prediction interval provides a range within which we expect the price of a similar home to fall 90% of the time. It is wider because it accounts for both the uncertainty in the model and the variability in individual home prices. While the confidence interval provides a range within which we expect the average price of homes with these characteristics to fall 90% of the time. It is narrower because it only accounts for the uncertainty in the model's estimate of the average price.

Home 2: 4250 sqft living area, 2100 sqft upper level living area, 5 years old, 5 bathrooms, backs out to a lake

Predicted Price: $1,074,285

90% Prediction Interval: $852,522.60 to $1,296,048

90% Confidence Interval: $1,045,117 to $1,103,454

The prediction interval for this home is wider, ranging from $852,522.60 to $1,296,048, reflecting the variability in individual home prices.Thus confidence interval is narrower, ranging from $1,045,117 to $1,103,454, reflecting the uncertainty in the model's estimate of the average price.

The prediction interval is wider than the confidence interval because it accounts for both the uncertainty in the model's predictions and the variability in individual home prices. The confidence interval only accounts for the uncertainty in the model's estimate of the average price, making it narrower.

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

### Correlation Analysis

A graph of purple dots

Description automatically generated

A green graph with numbers and a line

Description automatically generated with medium confidence

Based on the scatterplots and the correlation coefficients, a second-order model could be appropriate for these variables. The scatterplot of price versus living area shows a strong positive trend with some curvature, indicating that a quadratic term might improve the model. Similarly, the scatterplot of price versus age of the home shows a weak negative trend with some curvature, suggesting that a second-order model could better capture the relationship between these variables and home prices.

### Reporting Results

The R-squared value () of 0.8088 indicates that approximately 80.88% of the variability in home prices can be explained by the model, suggesting a strong relationship between the predictors (school rating and crime rate) and the response variable (price). The adjusted R-squared value () of 0.8084, which accounts for the number of predictors in the model, is slightly lower but still indicates a strong fit. This confirms that the model explains a substantial portion of the variability in home prices.

A chart with purple dots

Description automatically generated

A graph of a normal q-q plot

Description automatically generated

### Evaluating the Significance of Model

To determine if the model is significant at a 5% level of significance, we perform the overall F-test. The null hypothesis () states that the model with no predictors is as good as the model with predictors (i.e., all regression coefficients are zero), while the alternative hypothesis () states that the model with predictors is better than the model with no predictors (i.e., at least one regression coefficient is not zero). With a P-value of < 2.2e-16, we reject the null hypothesis, indicating that the model is significant at a 5% level of significance. For individual beta tests, the null hypothesis () for each predictor states that the coefficient is zero, while the alternative hypothesis () states that the coefficient is not zero. The P-values for the intercept (), school rating (), crime rate (), school rating squared (), crime rate squared (), and the interaction term (school rating \* crime rate) () are 1.45e-12, 0.000406, 1.90e-09, < 2e-16, < 2e-16, and 0.281513, respectively. Since the P-value for the interaction term is greater than 0.05, we fail to reject the null hypothesis, indicating that the interaction term is not significant at a 5% level of significance. However, the P-values for the intercept, school rating, crime rate, school rating squared, and crime rate squared are all less than 0.05, leading us to reject the null hypothesis for these predictors. This indicates that the intercept, school rating, crime rate, school rating squared, and crime rate squared are significant at a 5% level of significance. The overall model is significant at a 5% level of significance, and all individual predictors except the interaction term are significant at a 5% level of significance. This indicates that the intercept, school rating, crime rate, school rating squared, and crime rate squared are important predictors of home prices.

### Making Predictions Using Model

Home with School Rating of 9.80 and Crime Rate of 81.02 per 100,000 Individuals Predicted

Price: $874,497

90% Prediction Interval: $721,606 to $1,027,388

90% Confidence Interval: $863,681 to $885,313

The prediction interval provides a range within which we expect the price of a similar home to fall 90% of the time. It is wider because it accounts for both the uncertainty in the model and the variability in individual home prices.

The confidence interval provides a range within which we expect the average price of homes with these characteristics to fall 90% of the time. It is narrower because it only accounts for the uncertainty in the model's estimate of the average price.

Home with School Rating of 4.28 and Crime Rate of 215.50 per 100,000 Individuals

Predicted Price: $199,707

90% Prediction Interval: $46,992 to $352,422

90% Confidence Interval: $191,753 to $207,660

The prediction interval for this home is wider, ranging from $46,992 to $352,422, reflecting the variability in individual home prices.

The confidence interval is narrower, ranging from $191,753 to $207,660, reflecting the uncertainty in the model's estimate of the average price.

## 5. Nested Models F-Test

### Reporting Results

The general form of the first-order regression model is:

Using the outputs obtained from your R script, the prediction model equation can be written as:

Evaluating Significance of the Model

The overall F-test is used to determine if the model is significant at a 5% level of significance. The null hypothesis (H₀) states that all the regression coefficients are equal to zero, meaning that none of the predictor variables have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that at least one of the regression coefficients is not equal to zero, meaning that at least one predictor variable has a significant effect on the response variable (price). The p-value for the overall F-test is less than 2.2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis and conclude that the model is significant at a 5% level of significance.

For the individual beta tests, the null hypothesis (H₀) for each term states that the corresponding regression coefficient is equal to zero, meaning that the predictor variable does not have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that the corresponding regression coefficient is not equal to zero, meaning that the predictor variable has a significant effect on the response variable (price). Based on the model summary, the p-values for the intercept, school\_rating, crime, and the interaction term (school\_rating:crime) are all less than 2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis for each term and conclude that all terms are significant at a 5% level of significance.

* Write the general form and the prediction equation of a first order model for price using average school rating in the area and crime rate per 100,000 people as predictors. Include the interaction term between average school rating and crime rate. Use (where i equals 1, 2, ... ) to represent the slope parameters for all predictor variables.
* Create the first order regression model for price using average school rating in the area and crime rate per 100,000 people as predictors. Include the interaction term between average school rating and crime rate. Write the prediction model equation using outputs obtained from your R script.  
  Note: Use average school rating and crime rate as quantitative variables in this model. Use the equation editor to write the prediction model equation with the outputs

 Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.

### Evaluating Significance of Model

The overall F-test is used to determine if the model is significant at a 5% level of significance. The null hypothesis (H₀) states that all the regression coefficients are equal to zero, meaning that none of the predictor variables have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that at least one of the regression coefficients is not equal to zero, meaning that at least one predictor variable has a significant effect on the response variable (price). The p-value for the overall F-test is less than 2.2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis and conclude that the model is significant at a 5% level of significance.

For the individual beta tests, the null hypothesis (H₀) for each term states that the corresponding regression coefficient is equal to zero, meaning that the predictor variable does not have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that the corresponding regression coefficient is not equal to zero, meaning that the predictor variable has a significant effect on the response variable (price). Based on the model summary, the p-values for the intercept, school\_rating, crime, and the interaction term (school\_rating:crime) are all less than 2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis for each term and conclude that all terms are significant at a 5% level of significance.

### Model Comparison

The ANOVA formula for comparing nested models involves calculating the F-statistic to determine if the additional terms in the more complex model significantly improve the fit of the model. Here's the general formula for Nested F-test

The nested models F-test shows that the complete second-order model (Model 2) provides significantly better predictions than the reduced first-order model. The F-statistic of 65.20513 and the p-value of 2.22716e-28 indicate that the additional complexity of the second-order model is justified. This suggests that including the quadratic terms and interaction terms improves the model's ability to predict home prices based on school rating and crime rate.

Conclusion

The complete second-order model has a higher R-squared (0.8088) and adjusted R-squared (0.8084) compared to the reduced first-order model's R-squared (0.7995) and adjusted R-squared (0.7993). This indicates that the complete model explains more variability in home prices. The nested models F-test shows that the complete second-order model provides significantly better predictions than the reduced first-order model, with an F-statistic of 65.20513 and a p-value of 2.22716e-28. This justifies the additional complexity of the second-order model, as it captures non-linear relationships and interactions between predictors more effectively.