# MAT 303 Project One Summary Report

Tammy Hartline

Week 4

tammy.hartline@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

*In this project, I am exploring a dataset that contains information about various housing properties. The dataset includes several key variables such as square footage, age of the home, number of bathrooms, view, average school rating in the area, and crime rate per 100,000 people. My primary objective in this statistical analysis is to develop predictive models for house prices based on these variables. I aim to gain insights into what factors influence house prices and to what extent.*

*The results of my analysis will be used to make informed decisions related to real estate, particularly in understanding the impact of various property attributes and neighborhood characteristics on pricing. By gaining a deeper understanding of these relationships, potential buyers, sellers, and real estate professionals can make more accurate pricing predictions and investment decisions.*

*Throughout this project, I will employ various statistical analyses, including correlation analysis, multiple regression modeling, and hypothesis testing. These analyses will help me identify the key predictors of house prices, evaluate the significance of my models, and assess whether additional quadratic terms are necessary to improve predictions. Ultimately, my goal is to provide valuable insights into the factors that drive housing prices, contributing to more informed decision-making in the real estate market.*

## 2. Data Preparation

*In this housing dataset analysis, two regression models were developed to predict housing prices using a set of important independent variables (predictors) and a dependent variable (Price), representing the selling/listing price of homes. The dataset, containing 2692 rows and 23 columns, provided a robust population for analysis.*

*For Model #1, a first-order regression model was constructed with quantitative independent predictor variables, which included "sqft\_living," "sqft\_above," "age," and "bathrooms." The results of this model revealed the significant impact of these variables on housing prices, explaining approximately 51.78% of the price variability. Correlation analysis showed that "sqft\_living" had the highest positive correlation (0.689) with price, while "age" exhibited a negative correlation (-0.074).*

*In contrast, Model #2 was developed as a second-order regression model, incorporating "school\_rating" and "crime" as predictor variables. This model successfully explained approximately 80.88% of the variability in housing prices. Diagnostic plots for Model #2 indicated that the model adhered to the assumptions of linearity, normality of residuals, and homoscedasticity.*

*These two models, each with its set of predictor variables, provide valuable insights into the factors influencing housing prices in the dataset. Model #1 emphasizes the importance of physical characteristics such as size and age of the property, while Model #2 highlights the role of neighborhood attributes, specifically school ratings and crime rates, in predicting housing prices. These findings contribute to a comprehensive understanding of the factors impacting housing prices, aiding in more accurate predictions and informed decision-making in the real estate market. Lastly, I performed an F-test to compare Model #2 with a reduced model (r\_model2) without the second-order terms. The result indicated that Model #2 is significantly better in explaining the variability in housing prices.*

*Overall, these regression models provide valuable insights into the relationships between the predictor variables and housing prices, which can assist in setting better prices for listings.*

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

### Correlation Analysis

*In Model #1, a first-order regression model was constructed using quantitative independent predictor variables, including "sqft\_living," "sqft\_above," "age," and "bathrooms." This model was analyzed to understand the influence of these variables on housing prices, and the results demonstrated their significant or insignificant impact, explaining approximately 51% of the variability in housing prices.*

*Correlation analysis was conducted to assess the relationships between these predictor variables and the dependent variable, Price.*

*The correlation coefficient between Price and Living Area (sqft\_living) was 0.689. This indicates a strong positive correlation, implying that there is a substantial tendency for housing prices to increase as the square footage of the living area increases.*

*A graph of a number of dots

Description automatically generated with medium confidence*

*The correlation coefficient between Price and Above Ground Area (sqft\_above) was 0.566. This also indicates a strong positive correlation, suggesting that there is a significant positive relationship between the above-ground square footage of a home and its price.*

*A graph of a graph

Description automatically generated with medium confidence*

*The correlation coefficient between Price and Age of the Home (age) was approximately -0.075. This signifies a weak negative correlation, suggesting that older homes may, on average, have slightly lower prices. However, the impact of the age variable on housing prices is not as substantial as that of living area or above-ground area.*

*A graph of a graph with numbers and dots

Description automatically generated with medium confidence*

*The correlation coefficient between Price and Number of Bathrooms (bathrooms) was 0.486. This indicates a moderate positive correlation, implying that homes with more bathrooms tend to command higher prices.*

*A graph of different sizes and colors

Description automatically generated*

*In summary, the correlation analysis confirms the strong positive correlations between Price and Living Area (sqft\_living), Size of Upper Level (sqft\_above), and Number of Bathrooms (bathrooms). These findings suggest that larger living spaces, more above-ground square footage, homes with water views like lake or ocean front properties, and a greater number of bathrooms tend to be associated with higher housing prices. Additionally, the age of the home (age) exhibits a weak negative correlation with prices, indicating that older homes may, on average, have slightly lower prices. The scatterplot summaries highlight the varying degrees of linear trends observed in the relationships between the independent variables and housing prices. While living area and above-ground area displayed strong positive trends, indicating their significant impact on prices, the age of the home exhibited a subtle negative trend, suggesting a weaker influence on prices. Lastly, the number of bathrooms showed a moderate positive trend, indicating its relevance in determining housing prices. In conclusion, the correlation analysis appears to indicate that a positive correlation implies an increase in the predictor variable is associated with an increase in housing prices, while a negative correlation suggests the opposite.*

### Reporting Results

*General Form:*

 E(Y) = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2 + 𝛽3𝑥3 + 𝛽4𝑥4 + 𝛽5𝑥5

*Where price is the predicted housing price, is the intercept (constant) term, is predictor sqft\_living, is predictor sqft\_above, is predictor age, is predictor bathrooms, is predictor view, and lastly represents the error term.*

*Prediction Equation:*

*where is predicted price = 7709.048 + 129.2846 \* sqft\_living + 19.51206 \* sqft\_above + 1450.617 \* age + 43970.12 \* bathrooms + 167491.5 \* view1 + 259000.00 \* view2*

*The R² value was 0.6029, and the Adjusted R² value was 0.602. These values can now be used to interpret the approximate variability in housing prices as 60%, explained by the combination of predictor variables used in this model. In this scenario, due to the five predictor variables, both squared values are approximately 60%, suggesting that model has a good fit.*

*Interpretation of the beta estimates for the intercept (, with a p-value of 0.58495, we can infer that this is not statistically significant. This estimate suggests when all predicator variables are zero, then the predicted price is not significantly different from zero (which may not be meaningful in practice for some variables). The residual standard error is 133,600 on 2685 DF. The Interpretation of the estimate was approximately 129.28, indicating for every additional square foot increase in the living area, then the predicted housing price increases by our approximate value of $129.28. For the approximate estimate was 19.51, indicating similar results to sqft\_living, for every additional square foot increase in the upper-level area (holding all variables constant), the predicted housing price increases by approximately $19.51. This indicates for each additional year of age on the home, and again holding all other variables constant, the predicted housing price increases by approximately $1,450.62. estimate was approximately 43,970.12, indicating for every additional bathroom, (holding all other variables constant), the predicted housing price increases by the approximate estimate value of $43,970.12. Lastly, (view2), had an approximate estimate of 249000.0, while view1 had an approximate estimate of 167,491.50. These approximations indicate that homes with waterfront views, have a higher valued average by $249,000.00 and homes with forest views have a higher valued average by $167,491.50.*

*Residuals are a crucial aspect of regression analysis, and it's important to examine them to assess the validity of model assumptions. In the Residuals vs. Fitted Values plot, I observed that the residuals appear randomly scattered around the horizontal line at y = 0. This suggests that there is no strong pattern or curvature in the residuals, indicating that the linearity assumption is reasonable.*

*A graph of a number of dots

Description automatically generated with medium confidence*

*The Normal Q-Q Plot shows that the residuals follow a fairly straight line, which is an indication that they approximately follow a normal distribution. This suggests that the assumption of normality of residuals is not severely violated.*

*A graph of a normal q-q plot

Description automatically generated*

*Overall, based on the examination of these plots, the assumptions of linearity and normality of residuals seem to be reasonable for this multiple regression model, which adds to the model’s validity.*

*In summary, the multiple regression model with living area, upper-level area, age, number of bathrooms, and view as predictor variables provides a reasonably good fit to the data, explaining a significant portion of the variability in housing prices. The beta estimates suggest the direction and strength of the relationships between these predictors and prices. The diagnostic plots indicate that the model's assumptions are met to a reasonable extent.*

### Evaluating Significance of Model

*The significance of the multiple regression model was assessed using an overall F-test at a 5% level of significance. In this test, the null hypothesis (H0) states that all the regression coefficients, except the intercept, are equal to zero, implying that none of the predictor variables have a significant impact on the dependent variable (housing price). The alternative hypothesis (Ha), on the other hand, asserts that at least one of the coefficients is not equal to zero, indicating that there is a significant relationship between the predictor variables and housing price. The F-statistic for this model was 808.18, and the associated p-value was found to be less than 2.2e-16, which is much smaller than the 5% significance level. Consequently, we reject the null hypothesis (H0) in favor of the alternative hypothesis (Ha), concluding that the overall model is statistically significant.*

*To identify which individual terms (beta coefficients) are significant at a 5% level of significance, beta tests were conducted for each predictor variable. In these tests, the null hypothesis (H0) posits that a specific beta coefficient is equal to zero, signifying that the corresponding predictor variable has no significant effect on housing price. The alternative hypothesis (Ha) suggests that the beta coefficient is not equal to zero, indicating a significant relationship between the predictor variable and housing price. Among the predictor variables, sqft\_living, age, bathrooms, view1, and view2 were found to be significant at a 5% level of significance based on their p-values, which were all less than 0.05. These results imply that changes in these variables have a statistically significant impact on the predicted housing prices. However, sqft\_above was not considered statistically significant in this analysis since its p-value exceeded the 5% threshold. Therefore, we can conclude that not all predictor variables are equally significant in explaining housing prices, and only the mentioned significant variables should be considered in practical applications of the model.*

*A screenshot of a computer

Description automatically generated*

### Making Predictions Using Model

*For the first home, which has 2150 square feet of living area, 1050 square feet of upper-level living area, is 15 years old, has 3 bathrooms, and backs out to the road, the predicted price is approximately $329,349.3. The 90% prediction interval for this home's price ranges from $108,264.1 to $550,434.5. This means that we are 90% confident that the actual price of the home will fall within this interval. The prediction interval is wider than the confidence interval because it not only accounts for the variability in the model but also considers the additional uncertainty associated with predicting a specific individual's home price.*

*For the second home, which boasts 4250 square feet of living area, 2100 square feet of upper-level living area, is 5 years old, has 5 bathrooms, and backs out to a lake, the predicted price is approximately $830,347.9. The 90% prediction interval for this home's price ranges from $608,274.6 to $1,052,421. Similar to the first home, this interval represents our 90% confidence that the actual price will fall within this range.*

*In both cases, the prediction intervals provide a broader range of potential prices due to the inherent uncertainty in predicting specific home prices, while the confidence intervals typically refer to the variability of the model's predictions for the population as a whole. In conclusion of Model1, it can be used to assist setting better prices for listings and provide insights into the factors that influence housing prices.*

*A number and numbers on a white background

Description automatically generated*

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

### Correlation Analysis

*For the Price vs. Average School Rating scatterplot, where price is the dependent variable and school\_rating is the independent variable, there appears to be a general positive trend, indicating higher priced homes do tend to have higher average school\_rating in their local area. That being said, the relationship does not appear to be perfectly linear. Instead, it exhibits some curvature, suggesting that a simple linear model may not fully capture the complexity of the relationship between a home’s price and the local average school rating. Based on the second order model scatterplot, it appears the quadratic model might be appropriate for modeling this relationship, as it can capture the curvature seen in the data, which allows a more accurate representation of the relationship.*

*A graph of a school rating

Description automatically generated*

*For the Price vs. Crime Rate per 100,000 people, where price is our dependent variable and crime rate is our independent variable, the scatterplot appears to show a general negative trend, suggesting that, in general, homes within areas where the crime rate is low, tend to have higher values. Similar to the school rating scatterplot, the relationship here does not appear to be strictly linear, as there is some curvature, indicating that a simple linear model may not capture the full complexity of the relationship between housing prices and crime rates. Therefore, it seems that the second order quadratic model would be more appropriate for modeling the relationship, as it allow the capture of the curvatures.*

*A graph of a graph

Description automatically generated with medium confidence*

*For both correlation analysis’, the scatterplots indicate that the relationships between price and crime rate and price and school rating, are not purely linear. Instead each plot exhibits some curvature, suggesting that a second order model, that will allow quadratic terms, will likely be more suitable to capture the underlying patterns in the data more accurately. It is worth noting however, that further analysis and modeling should be conducted to confirm the appropriateness of the quadratic terms and to build more robust predictive models.*

### Reporting Results

*General Form:*

*P = B0 + B1X1 + B2X2 + B3(X1^2) + B4(X2^2) + B(X1) (X2) + E*

*Where:*

*P represents price, the dependent variable, (All B’s mean beta), and the X variables are our independent variables (coefficients, predictors), X1 represents the average school rating, X2 represents the crime rate per 100,000 people, B0 represents the intercept, B1 the slope parameter for school rating, B2 the slope parameter for crime rate, B3 the coefficient for the square of school rating, B4 the coefficient for the square of the crime rate, B5 the coefficient for the interaction term between school rating and crime rate, and lastly, E represents the error term.*

*Using the outputs from the R-script, the specific prediction model equation for this model is:*

*Price = 733,910.73 – 73,748.17 (school\_rating) – 3,154.77 (crime) + 11,646.80 (school\_rating^2) + 6.38 (crime^2) – 52.27 (school\_rating) (crime) + E*

*Interpreting R^2 and adjusted R^2 for this model, allowed me to determine the R-squared value(R^2) for this model is approximately 0.8088, which indicates about 80.88% of the variability in the price can be explained by the predictors (school rating and crime rate), which translates to the model accounts for a substantial portion of the variance in price. The adjusted R-squared is also approximately 0.8084 and adjusts for the number of predictors in this model. Because both R^2 and R^2 adjusted are so close in value, it suggests the inclusion of the quadratic and interaction terms in the model does not significantly improve the model’s explanatory power beyond the linear terms.*

*The residuals vs Fitted plot shows no clear pattern or curvature, which is a positive sign. This suggests that the linearity assumption holds, and the residuals are fairly evenly spread around zero. However, there are some outliers with larger residuals, signaling further investigation is needed.*

*A graph with a red line

Description automatically generated*

*The normal Q-Q plot is relatively linear, indicating the residuals are approximately normally distributed. However, there is some deviation from linearity in the tails suggesting that the normality assumption may not be perfectly met, especially in the extreme values of residuals. This also signals further investigation or transformation of day may be needed to address this.*

*A graph of a normal graph

Description automatically generated with medium confidence*

*In summary, the second-order regression model appears to explain a significant portion of the variance in price, but there are outliers and potential deviations from normality in the residuals that should be considered in the interpretation and further analysis of the model.*

### Evaluating Significance of Model

*Null hypothesis (HO), for crime coefficient is not significant as it is equal to zero. Alternative hypothesis (Ha), for crime coefficient is significant as it is not equal to zero. The p-value for the coefficient of crime is essentially zero (p <2.2e-16), which is much less than the 5% significance level of 0.05. This leads to a conclusion that the coefficient for crime is highly significant at 5% level, indicating that it has a significant effect on the price.*

*Null Hypothesis (HO), for school rating coefficient is not significant as it is equal to zero. Alternative hypothesis (Ha), for school rating coefficient is significant as it is not equal to zero. The p-value for the coefficient of school rating is 0.000406, which is less than the 5% significance level of 0.05. This leads to a conclusion that the coefficient for school rating is significant at a 5% level, indicating that it has a significant effect on the price.*

*In summary, both crime and school rating are individually significant at a 5% level of significance in explaining the variation in the price.*

### Making Predictions Using Model

*In prediction scenario 1, the predicted price $874,515.60 is the estimated price for a home in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 people. The prediction interval for this scenario is in the range of $874,497 to $1,027,388. The 90% confidence interval indicates that we are 90% confident that the population mean of home prices for this scenario falls within the mentioned range, indicating that there is a 90% probability that the true price of a home with these variables would be within this interval range.*

*In prediction scenario 2, the predicted price $199,844.50 is the estimated 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 people. The prediction interval for this scenario is in the range of $199,706.70 to $352,421.70. The 90% confidence interval indicates that we are 90% confident that the population mean of home prices for this scenario falls within the mentioned range, indicating that there is a 90% probability that the true price of a home with these variables would be within this interval range.*

*Therefore, these intervals provided valuable estimates of the range in which we can expect the true price of a home to fall in each scenario, along with our level of confidence in those estimates. The prediction intervals are wider because they account for both the variability of the data and the uncertainty in predicting individual prices. The confidence levels on the other hand, are narrower as they estimate the average price for the given scenario.*

## 5. Nested Models F-Test

### Reporting Results

*In this analysis, I constructed a first-order regression model for predicting housing prices, incorporating average school rating in the area and crime rate per 100,000 people as predictors. Notably, I included an interaction term between these predictors to capture potential combined effects. The resulting prediction equation takes the form*

***Price = -410,233.37 + 155,599.97 \* school\_rating + 2,230.07 \* crime – 564.85 \* (school\_rating \* crime).***

*This equation enabled the ability to estimate housing prices while considering the influence of average school rating, crime rate and their interaction. The model appears highly significant, as indicated by a substantial R-squared value of 0.7995 and a statistically significant F-statistic. These results suggest that the predictors, along with their interaction, collectively explain a significant portion of the variation in housing prices.*

### Evaluating Significance of Model

*In this model, where price is our dependent variable and our predictors are average school rating and crime, and their interaction term, it does appear to be significant at a 5% level of significance.*

*To determine this, I conducted an overall F-test.*

*The null hypothesis (HO), all coefficients in the model, including the intercept and the coefficients for school rating, crime, and their interaction, are equal to zero, meaning none of the predictors have any significant effect on housing prices.*

*The alternative hypothesis (Ha), at least one of the coefficients in the model is not equal to zero, indicating that at least one predictor has a significant impact on housing prices.*

*The calculated p-value for the overall F-test is much than the 0.05 (5%) level of significance, likely approaching zero. Given this extremely low p-value, the null hypothesis is rejected, concluding the model is significant at a 5% level of significance. This translates to at least one of the predictors, school\_rating, crime, or their interaction, significantly influences housing prices.*

*To identify which individual terms are significant at a 5% level, I examined the p-values associated with each coefficient in the model. In this scenario, all three coefficients have p-values much less than 0.05. Consequently, we can conclude that all these terms are statistically significant at the 5% significance level, indicating that school rating, crime, and their interaction have a significant impact on housing prices.*

*Therefore, the overall F-test demonstrated the model’s significance and individual beta tests confirmed the significance of all predictors, emphasizing their importance in explaining variations in housing prices.*

### Model Comparison

*To compare the two models, Model1 and Model2, I assessed whether including quadratic (squared) terms contributes significantly to predicting home prices. Model1, having a simpler baseline, includes average school rating and crime rate per 100,000 people as predictors. Its general form can be expressed as*

***Y = Beta0 + Beta1(X1) + Beta2(X2^2) + E***

*and the prediction equation is*

***Y(top-hat) = Beta0 + Beta1(X1) + Beta2(X2)***

*Adding the quadratic terms for the predictors results in a more complex equation of*

***Y = Beta0 + Beta1(X1) + Beta2(X2) + Beta3(X1^2) + Beta4(X2^2) + Beta5(X1) (X2) + E***

*with the corresponding prediction equation*

***Y(top-hat) = Beta0 + Beta1(X1) + Beta2(X2) + Beta3(X1^2) + Beta4(X2^2) + Beta5(X1) (X2)***

*Where firstly, the intercept (Beta0) represents the baseline price of homes when all predictor variables are zero. In other words, it signifies the starting point for home prices in the absence of any influence from the predictors.*

*Next, Beta1 (X1) corresponds to the effect of the average school rating in the area on home prices. It quantifies how a one-unit change in the average school rating impacts home prices, assuming all other predictors remain constant.*

*Beta2 (X2^2) introduces a quadratic term for the crime rate per 100,000 individuals. This term explores whether there's a non-linear relationship between the crime rate and home prices. It captures potential curvature in the relationship.*

*In the more complex equation, we see the addition of Beta3 (X1^2) and Beta4 (X2^2), which represent the squared terms of the average school rating and crime rate, respectively. These terms allow us to assess if there are non-linear associations between these variables and home prices.*

*Furthermore, Beta5 (X1 \* X2) accounts for the interaction effect between the average school rating and the crime rate on home prices. It helps us understand whether the combined influence of these two predictors on home prices deviates from what we would expect based on their individual effects.*

*Finally, the error term (E) encompasses all other unmeasured factors and random variations that affect home prices but are not explicitly considered in the model. It represents the unexplained variability in home prices.*

*In essence, this model attempts to capture the complex interplay between average school ratings, crime rates, and potentially their non-linear and interactive effects, providing a more nuanced understanding of how these factors relate to home prices.*

*To evaluate the significance of these quadratic terms, I conducted a nested model F-test at a 5% level of significance. The null hypothesis posits that adding quadratic terms (model2), does not significantly improve the model fit compared to the first regression model (model1), while the alternative hypothesis (Ha), suggests that model2 does offer significant improvement.*

## 6. Conclusion

*In evaluating both Model #1 and Model #2, it becomes evident that Model #2, the second-order regression model, is the preferred choice for predicting house prices. This preference arises from several critical factors. First, Model #2 outperforms Model #1 in terms of explanatory power, as it accounts for approximately 80.88% of the variability in housing prices, compared to Model #1's 60.29%. This higher R-squared value indicates that Model #2 provides a more comprehensive understanding of the factors influencing housing prices.*

*Secondly, Model #2 incorporates quadratic terms and an interaction term, allowing it to capture potential non-linear relationships and interactions between predictors more effectively. This feature is particularly valuable in the real estate market, where relationships between variables are often complex and may not adhere strictly to linear patterns.*

*Lastly, the nested models F-test confirmed that Model #2 significantly improves the model fit compared to Model #1, reinforcing its superiority in explaining price variability.*

*In practical terms, choosing Model #2 empowers real estate professionals, buyers, and sellers with a more accurate and nuanced tool for predicting housing prices. By considering both linear and non-linear relationships, as well as interactions between variables, Model #2 offers a more reliable basis for setting listing prices, making investment decisions, and understanding the intricate dynamics of the housing market. Therefore, Model #2 emerges as the preferred choice for predicting house prices due to its superior predictive capability and ability to capture complex relationships between variables.*