

# MAT 303 Project One Summary Report

Daniel Gorelkin

Daniel.Gorelkin@SNHU.edu

Southern New Hampshire University

## 1. Introduction

*The explored dataset will be read from the “housing\_v2.csv” file. This dataset contains 2692 records of randomly examined houses that sold in the Seatle region and their 22 variables such as footage, number of bathrooms, and more. The statistical analyses in this report will generate several models based on the dataset variables in order to predict the most accurate and reasonable house’s selling price to help the real estate company set better prices when listing a home for a client. This will ensure that listings can be sold within a reasonable amount of time and at a fair price. This project will perform the First Order Regression analysis with Quantitative and Qualitative Variables, which will help to predict the quantitative value of a property and allow us to understand how the variables influence the price. In addition, the Second Order Multiple Regression Model will be generated to examine non-linear relationships between our variables if such will be found. Lastly, the Nested Models F-Test will be conducted to compare and determine if the more complex model with more variables provides a statistically significant improvement in the fitment compared to the simpler model with fewer variables.*

## 2. Data Preparation*.*

*As mentioned above, the dataset contains 2692 records of different assets and their properties as variables. Each asset has 22 variables, where 19 are* ***quantitative,*** *such as price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, condition, grade, sqft\_basement, yr\_built, yr\_renovated, zipcode, sqft\_living15, sqft\_lot15, age, appliance\_age, crime, and the local school\_rating. In addition, our dataset contains three* ***qualitative*** *variables, e.g., backyard, view, and renovated.*

*For the purpose of this analysis, we will compare the effect of the size of the living area in sqft (sqft\_living), the size of the upper level in sqft (sqft\_above), the number of bathrooms (bathrooms), the age of the home (age), and lastly the view from the property as (view). The examined variables against the sale price of the home (price) will allow us to build the prediction model equation for the expected home sale price.*

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

### Correlation Analysis

*A graph of a blue line

AI-generated content may be incorrect.A graph of red dots

AI-generated content may be incorrect.*

*A close-up of a text

AI-generated content may be incorrect.*

*Based on these analyses, we can observe a relationship between price and living area with a value of r≈0.689, which indicates that the relationship exists, but this is not a perfect correlation, and other factors also influence price. The relationship between Price and Age, on the other hand, has a value of r≈-0.075, which indicates that age has a very weak effect on price, and age alone does not determine the home’s price.*

### Reporting Results

* *The general form and the prediction equation of the multiple regression model:*

*The general form:*

***Price=+(x1)+(x2)+(x3)+(x4)+(x5)+(x6)***

*The prediction equation:*

***Price^=+​(sqft\_living)+​(sqft\_upper)+​(age)+​(bathrooms)+(view1)+(view2)***

*Where to represent the living area, upper-level area,* ***​*** *the age of the home, the number of bathrooms and as predictor variables represent the view with the base case of backs out to a road when the dummy variables for* ***view1*** *and* ***view2 = 0.***

* *The prediction model equation using outputs obtained from the multiple regression model:*

*A screenshot of a computer

AI-generated content may be incorrect.*

*Hence, the updated prediction equation for the home price is:*

***Price^=7709 + 129.3​(sqft\_living) + 19.51​(sqft\_upper) + 1451(age) + 43970(bathrooms) + 167500(view1) + 249000(view2)***

*Where a home that backs out to a road (view1=0, and view2=0), a lake (view1=0, and view2=1), or a home that backs out to trees (view1=1, and view2=0).*

* *From our model and the value of R2 (R-squared)=0.6029, we can learn that our model explains 60.29% of the variation in price, and the other 40% could be explained by other predictor variables or due to chance. Yet, the model suggests moderate to strong prediction ability, and it explains a significant amount of price variation. Likewise, Our slightly lower Adjusted R-Squared (adjusted R-squared) value for the model of 0.602 indicates that adding predictors did not overwhelm the explanatory power of the model, and most predictors contribute to the model meaningfully.*
* *The beta estimates for the living area and lake view give us a clue that lake view significantly ramps up the price for a home by $249000, where for each additional unit of the property’s square fit, the price will increase by $129.3​ and for the ability of any additional family member to take a shower at the same time as any other family member it will cost the buyer additional $43970. To conclude, water is a pricy resource.*
* *The residuals and fitted values plots.*

*A diagram of red dots

AI-generated content may be incorrect.A graph of a normal q-q plot

AI-generated content may be incorrect.*

*The plots of fitted values and the normality suggest that residuals are randomly distributed, and heteroscedasticity is not a major concern as there is no obvious shape of a funnel visible, though some outliers are present. The Q-Q plot, on the other hand, shows normally distributed residuals as the data points lie closely and follow the blue line. Yet, some outliers can be observed at both tails of the plot. Therefore, our model seems to be reasonable, with residuals behaving relatively randomly.*

### Evaluating the Significance of the Model

Evaluating the overall significance level for the regression model at a significance level of 5%:

*The null hypothesis H0:* ***=​=​=​==****=0 (no relationship exists between any of the predictor variables and price)*

*The alternative hypothesis Ha: ≠0 for i = 1, 2, 3, 4, 5, 6 (relationship exists with at least one variable)*

*The overall F-statistic is F=679.3, with a corresponding P-value of 2.2e-16. The very large F-statistic value suggests that our model is explaining a significant amount of the variance in the data compared to the unexplained one. In addition, our very small P-value indicates strong evidence against the null hypothesis, and under the significance level of 5%, there is sufficient evidence exists to reject the null hypothesis and conclude that a statistically significant linear relationship exists between any or some of the predictive variables and our model explaining well the variability in the response variable.*

*Evaluate the significance of the living area at a 5% level of significance*

*The null hypothesis H0: =0 (no relationship exists between sqft\_living and price)*

*The alternative hypothesis Ha: ≠0 (relationship exists between sqft\_living and price)*

*The t-statistic for the sqft\_living estimate is 15.916, with a corresponding P-value of 2e-16.*

*Since our obtained P-value is < 0.05, sufficient evidence exists to reject the null hypothesis and conclude that a statistically significant linear relationship exists between sqft\_living and price.*

*Evaluate the significance of all other variables at a 5% significance level*

*In the same manner, we can observe that the corresponding t-statistics values for:  
The size of the upper level (sqft\_above) = 2.616, with P-value = 0.0089 < 0.05*

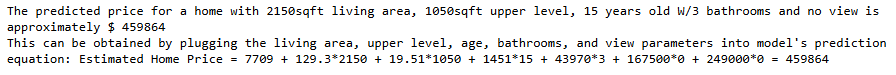
*The Age of the home (age) = 12.098, with P-value = 0.0000 < 0.05*

*The Number of bathrooms (bathrooms) = 7.178, with P-value = 0.0000 < 0.05*

*The view to the lake or the trees (view1) = 15.640 / (view2), with P-values of 0.0000 respectively*

*Therefore, all of the individual variables are statistically significant as their P-values are < 5%. Hence, all predicting variables are reliable and contribute to the model’s accuracy.*

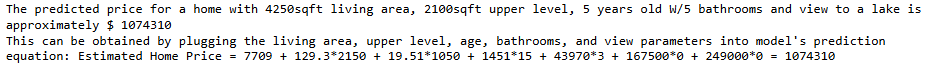
### Making Predictions Using Model



*A screenshot of a white background

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.*

**

*A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.*

*This prediction interval provides a wide range where we expect the price of an arbitrary home with the given characteristics to fall 90% of the time. Hence, it suggests wider price estimate fluctuations for similar properties, whereas the confidence interval is narrower and provides a narrower range where we expect the mean price of all similar homes to fall 90% of the time.*

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

### Correlation Analysis

*A graph with blue dots

AI-generated content may be incorrect.A graph with blue dots

AI-generated content may be incorrect.*

*As we can see from the price vs. average school rating scatterplot, despite the strong positive relationship of r=0.786 existing between the price and the school\_rating variables, a non-linear relationship can be seen as the rate at which the price increases with school rating is not constant, and the data points clearly follow a curved, bending upwards pattern as the school rating increases.*

*The same could be observed from the Price vs. Crime plot. There is a strong negative relationship of r=-768, though it does not indicate the type of the relationship. In addition, we can observe that the rate at which the price decreases is not linear, and the data points clearly follow a curved, bending downward pattern as the crime rating increases. Therefore, we can expect the relation for these variables on price to be of a second degree.*

### Reporting Results

*The general form equation of a complete second-order model for price using the average school rating in the area and crime rate per 100,000 people as predictors is:*

***E(y) =*** ***+ (x1) + (x2) + (x1)(x2) + (1) + (x²2)***

*The estimated prediction equation is:*

***Price^ =*** ***+ (s.rating) + (crime) + (s.rating)(crime) + (s.rating****²)* ***+ (crime****²****)***

*Where* ***s.rating*** *represents the**average school rating, and* ***crime*** *represents the crime rate in the area and , … represent are estimates of  , …  respectively.*

*The complete second-order model for price using the average school rating and crime rate:*

*A screenshot of a computer

AI-generated content may be incorrect.*

*Hence, the updated estimated prediction equation is:****Price^ =*** *733,900**- 73,750****(s.rating)*** *- 3,155****(crime)*** *- 52.27****(s.rating)(crime)*** *+ 11,650****(s.rating****²)**+ 6.377****(crime****²****)***

*Where* ***s.rating*** *represents the**average school rating, and* ***crime*** *represents the crime rate in the area.*

*From this model and the value of R2=0.8088, we can learn that our model explains 80.88% of the variation in price, whereas the other 19.11% could be explained by other predictor variables or due to chance. Yet, the model suggests a strong correlation, and it can explain a significant amount of price variation. Likewise, Our slightly lower, nearly perfect Adjusted R-Squared value for the model of 0.8084 indicates that the predictors did not overwhelm the explanatory power of the model, and all predictors contribute to the model meaningfully.*

*A red dot diagram with white text

AI-generated content may be incorrect.A graph of a normal q-q plot

AI-generated content may be incorrect.*

*From the plot of the residuals against fitted values, we can see that there are no signs of fanning or funneling patterns that would directly indicate heteroscedasticity, and our residuals are scattered randomly. However, we can observe that our points are randomly spread around the zero, which supports the linearity assumption, and because the variance increases with fitted values, this violates the assumption of homoscedasticity. Our normal Q-Q plot shows that our residuals are mostly normally distributed though some extreme values deviate at both tails. The above confirms that our model should fit well with the non-linear relationship between the variables.*

### Evaluating the Significance of the Model

Evaluating the overall significance level for the regression model at a significance level of 5%:

*The null hypothesis H0:* ***=​=​=​=​=****0 (no relationship exists between any of the predictor variables and price), meaning the predictors do not explain variation in price.*

*The alternative hypothesis Ha: ≠0 for i = 1, 2, 3, 4, 5 At least one regression coefficient is not equal to zero, e.g., at least one predictor significantly explains the price.*

*The overall F-statistic is F=2272, with a corresponding P-value of 2.2e-16. The very large F-statistic value with a very small P-value indicates strong evidence against the null hypothesis, and under the significance level of 5%, there is sufficient evidence exists to reject the null hypothesis and conclude that there is strong evidence that at least one of the predictors such as school\_rating, crime, their interaction, or quadratic terms significantly explains variation in house price. Hence, our model is statistically significant.*

*Evaluate the significance of each term at a 5% level of significance*

* *The null hypothesis H0: =0 for i = 1, 2, 3, 4, 5 (no relationship exists between school\_rating, crime, school\_rating², crime², school\_rating:crime and price respectively).*
* *The alternative hypothesis Ha: ≠0 for i = 1, 2, 3, 4, 5 (relationship exists between school\_rating, crime, school\_rating², crime², school\_rating:crime and price respectively).*

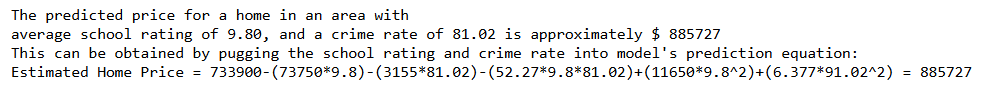
*The t-statistic for each term estimate, with corresponding p-values summarized in the table:*

|  |  |  |  |
| --- | --- | --- | --- |
| ***Test under 5% significance*** | ***t-value*** | ***P-value*** | ***conclusion*** |
| ***school\_rating*** | *-3.541* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***crime*** | *-6.027* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***school\_rating²*** | *10.497* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***crime²*** | *8.777* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***school\_rating:crime*** | *-1.077* | *0.281* | ***P-value*** *> 0.05 (Not significant)* |

*Since our obtained P-values for school\_rating, crime, school\_rating², and crime² are < 0.05, sufficient evidence exists to reject the null hypothesis and conclude that a statistically significant relationship exists and these predictors significantly affect house price.*

*However, because the Interaction term >0.05, not sufficient evidence exists to reject the null hypothesis and conclude that school\_rating:crime does not significantly affect house price and could potentially be removed from the model.*

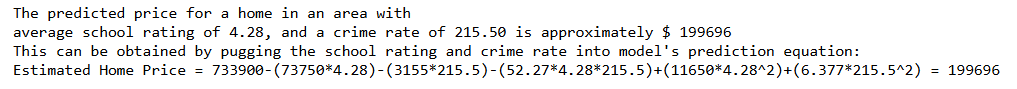
### Making Predictions Using Model

****

**A screenshot of a computer

AI-generated content may be incorrect.A close-up of a number

AI-generated content may be incorrect.**

****

**A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

*This prediction interval provides a wider range where we expect the price of an arbitrary home with the given characteristics to fall 90% of the time. Hence, it suggests wide price estimate fluctuations for similar properties, whereas the confidence interval is narrower and provides a shallow range where we can expect to find the mean price of all similar homes to fall 90% of the time.*

## 5. Nested Models F-Test

### Reporting Results

*The first-order model for home price using average school rating and crime rate, including average school rating and crime rate as interaction terms predictors general form is:*

***E(y) =*** ***+ (x1) + (x2) + (x1)(x2)*** where *represents the slope for school rating, represents the slope of the crime rate and represents the slope of the interaction terms between the two.*

*Where the estimated prediction equation is:****Price^ =*** ***+ (school\_rating) + (crime) + (school\_rating)(crime)***

*The first-order regression model for price using average school rating in the area and crime rate per 100,000 people as predictors, including the interaction terms:*

*A screenshot of a computer

AI-generated content may be incorrect.*

*The prediction model equation using outputs from the R script is:*

***Price^ =*** *-410233.37* ***+*** *155559.97****(school\_rating) +*** *2230.07****(crime)*** *-564.85****(school\_rating)(crime)***

### Evaluating the Significance of the Model

Evaluating the overall significance level for the model at a significance level of 5%:

*The null hypothesis H0:* ***=​==****0 (no relationship exists between any of the predictor variables and price), meaning the predictors do not explain variation in price.*

*The alternative hypothesis Ha: ≠0 for i = 1, 2, 3. At least one regression coefficient is not equal to zero, e.g., at least one predictor significantly explains the price.*

*The overall F-statistic is F=3573, with a corresponding P-value of 2.2e-16. The very large F-statistic value with a very small P-value indicates strong evidence against the null hypothesis, and under the significance level of 5%, there is sufficient evidence exists to reject the null hypothesis and conclude that there is strong evidence that at least one of the predictors such as school\_rating, crime, or their interaction, significantly explains variation in house price. Hence, our model is statistically significant.*

*Evaluate the significance of each term at a 5% level of significance*

* *The null hypothesis H0: =0 for i = 1, 2, 3 (no relationship exists between school\_rating(), crime(), school\_rating:crime() and price, respectively).*
* *The alternative hypothesis Ha: ≠0 for i = 1, 2, 3 (relationship exists between school\_rating(), crime(), school\_rating:crime() and price, respectively).*

*The t-statistic for each term estimate, with corresponding p-values summarized in the table:*

|  |  |  |  |
| --- | --- | --- | --- |
| ***Test under 5% significance*** | ***t-value*** | ***P-value*** | ***conclusion*** |
| ***school\_rating*** | *49.65* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***crime*** | *17.20* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |
| ***school\_rating:crime*** | *-31.63* | *0.000…* | ***P-value*** *< 0.05 (Significant)* |

*Since our obtained P-values for school\_rating, crime, and school\_rating:crime are < 0.05, sufficient evidence exists to reject the null hypothesis for each one of the terms and conclude that a statistically significant relationship exists for each and each one of these predictors significantly affects house price.*

### Model Comparison

*When comparing two models, such as a reduced and a complete model, we compare two models where one model is a simplified (reduced) model of the more complex (complete) model that contains all terms of the simple model in addition to having at least one additional predictor to try and improve the reduced model. By doing so, we can try to add complexity to the simplified model that could significantly improve its fit to the data. In other words, it helps us compare and evaluate the significance of the added variable/s and determine if the increased complexity of a model is justified.*

*The general form and prediction equation of the reduced model:*

*General form:* ***E(y) =*** ***+ (x1) + (x2) + (x1)(x2)***

*Prediction equation:* ***Price^ =*** ***+ (school\_rating) + (crime) + (school\_rating)(crime)***

*The general form* *and prediction equation of the complete model:*

*General form:* ***E(y) =*** ***+ (x1) + (x2) + (x1)(x2) + (1) + (x²2)***

*Prediction equation:* ***Price^ =*** ***+(s.rating) +(crime) +(s.rating)(crime) +(s.rating****²)****+ (crime****²****)***

*The nested model F-test at a 5% level of significance:*

***A close up of a number

AI-generated content may be incorrect.***

*Evaluate the significance of adding quadratic (squared) terms at a 5% level of significance*

*The null hypothesis H0:* ***=​=****0 (The coefficient for the quadratic terms is 0. The reduced model is sufficient), meaning the predictors do not add value to the model.*

*The alternative hypothesis Ha: ≠0 for i = 4, 5. At least one quadratic coefficient is not equal to zero, hence the complete model should be used.*

*The overall F-statistic is F=65.20, with a corresponding P-value of 2.2e-28. The very small P-value < 0.05 which indicates strong evidence against the null hypothesis, and under the significance level of 5%, there is sufficient evidence exists to reject the null hypothesis and conclude that quadratic terms should be used in predicting the home price, and the complete model should be chosen as it provides a better fit.*

## 6. Conclusion

**Models comparison table**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Test*** | ***Model #1 - First Order Regression W/Quantitative and Qualitative Variables*** | ***Model #2 - Complete Second Order Regression W/Quantitative Variables*** | ***Model #3 - First Order Regression W/ one interaction term*** |
| ***Residual standard error*** | *133600* | *92690* | *94870* |
| ***R-squared*** | *60.29%* | *80.88%* | *79.95%* |
| ***Adjusted R-squared*** | *60.2%* | *80.84%* | *79.93%* |
| ***F-statistic*** | *679.3* | *2272* | *3573* |
| ***P-value*** | *2.2e-16* | *2.2e-16* | *2.2e-16* |
| ***Significant predictors under 5%*** | *5/5* | *4/5* | *3/3* |

*By comparing the results of the three models examined, we can see that all three models are significant based on their very low P-value and can predict the house price based on the predictor variables. From the tests, we can see that model #1 can explain only 60% of the price variance compared to models #2 and #3, which explain approximately 81% and 80% of the variance. In addition, we can see from the adjusted R-squared value that there is not much space for improvement to improve the models for models #2 and 3. Model #2, however, provides the lowest residual standard error, which means its prediction results are expected to be in a narrower error window and eventually indicate a better model fit. Therefore, model #2 is the best choice for predicting house prices because it has the lowest residual error, the highest R-squared, and a strong F-statistic, and it does a good job of predicting house prices accurately while effectively capturing the relationship between variables and keeping prediction errors low. With that said, the practical importance of the analyses is that a model with a better fit can provide more accuracy in house price prediction and pinpoint which factors or interactions most influence house prices. By choosing the best model, investors and realtor companies can make more weightened pricing decisions and set the home price fairly.*

## 7. References

*ZyBooks. (n.d.). https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/3/section/1*

*ZyBooks. (n.d.-b). https://learn.zybooks.com/zybook/MAT-303-12932.202516-1/chapter/4/section/1*