# MAT 303 Project One Summary Report

Ashley De Venuto

ashley.devenuto@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this Project, the dataset being used is based on historical real estate that holds different variables common for buying or selling a house. There are multiple variables within this dataset. Some of those variables include price, sqft\_living, age and much more. The main variables being used within this project are price, sqft\_living, sqft\_above, age, bathrooms, school\_rating, view, and crime. The results from the analysis will be used by real estate companies to determine pricing for homes. Having the right price will help sell houses at a faster rate. To find this information, an analysis must be conducted. The two models being used are the first order regression model with quantitative and qualitative variables, a complete second order regression model with quantitative variables, and the nested models F-Test.

## 2. Data Preparation

In this dataset, there are 13 main variables. The variables that the project will focus on are price, sqft\_living, sqft\_above, age, bathrooms, school\_rating, view, and crime. In total, there is 23 columns represented by 23 variables/attributes. There are 2692 rows represented by the records in the dataset.

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

### Correlation Analysis

This project started with creating two different scatterplots. The first scatterplot was for price vs the living area (sqft\_living) and the second scatterplot is for price vs the age of the home (age). Looking at the first scatterplot, there is a positive correlation between price and living area. This means that when the living area increases the price also increases. When looking at the scatterplot for price vs age, there is no real relationship between the two variables. There is a slight negative correlation between price and age of home.

A red dot diagram with numbers

Description automatically generated A graph of red dots

Description automatically generated

### Reporting Results

The general form for the analysis where the variables are price, living area, upper level area, age of the home, number of bathrooms and view is shown below:

The prediction equation looks like this:

{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mover><mi>y</mi><mo>&#x2227;</mo></mover><mo>=</mo><mover><msub><mi>&#x3B2;</mi><mn>0</mn></msub><mo>&#x2227;</mo></mover><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>1</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>2</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>3</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>3</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>4</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>4</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>5</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>5</mn></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"}

The multiple regression model will look like this:

{"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\" style=\"font-family:stix;font-size:16px;\"><mi>E</mi><mfenced><mi>y</mi></mfenced><mo>=</mo><mn>7</mn><mo>.</mo><mn>709</mn><mi>e</mi><mo>+</mo><mn>03</mn><mo>&#xA0;</mo><mo>+</mo><mo>&#xA0;</mo><mn>1</mn><mo>.</mo><mn>293</mn><mi>e</mi><mo>+</mo><mn>02</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>+</mo><mn>1</mn><mo>.</mo><mn>951</mn><mi>e</mi><mo>+</mo><mn>01</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mn>1</mn><mo>.</mo><mn>451</mn><mi>e</mi><mo>+</mo><mn>03</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>3</mn></msub><mo>+</mo><mn>4</mn><mo>.</mo><mn>397</mn><mi>e</mi><mo>+</mo><mn>04</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>4</mn></msub><mo>+</mo><mn>2</mn><mo>.</mo><mn>490</mn><mi>e</mi><mo>+</mo><mn>05</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>5</mn></msub></math>","origin":"MathType for Microsoft Add-in"}{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mi>E</mi><mfenced><mi>y</mi></mfenced><mo>=</mo><mn>7</mn><mo>.</mo><mn>709</mn><mi>e</mi><mo>+</mo><mn>03</mn><mo>&#xA0;</mo><mo>+</mo><mo>&#xA0;</mo><mn>1</mn><mo>.</mo><mn>293</mn><mi>e</mi><mo>+</mo><mn>02</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>+</mo><mn>1</mn><mo>.</mo><mn>951</mn><mi>e</mi><mo>+</mo><mn>01</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mn>1</mn><mo>.</mo><mn>451</mn><mi>e</mi><mo>+</mo><mn>03</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>3</mn></msub><mo>+</mo><mn>4</mn><mo>.</mo><mn>397</mn><mi>e</mi><mo>+</mo><mn>04</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>4</mn></msub><mo>+</mo><mn>1</mn><mo>.</mo><mn>675</mn><mi>e</mi><mo>+</mo><mn>05</mn><mo>&#xA0;</mo><msub><mi>x</mi><mn>5</mn></msub></mstyle></math>","origin":"MathType for Microsoft Add-in"}

I wrote down two different ones representing the two different views within the model using the coefficients.

The value of is 0.6029 and the value of the is 0.602. This shows that the predictor variables of living area, upper level area, age of the home, number of bathrooms, and view accounts for around 60% of the variation in price of homes. It can be said that price goes up by 2.490e+05 if the home has a lake view and the price will go up by 1.293e+02 if the home has a nice sized living area. When wanting to find out the assumption of homoscedasticity, look at both the scatterplot based on residuals vs fitted values and the Q-Q plot. When looking at the scatterplot there is no clear pattern. I believe that this model has a homoscedasticity. The Q-Q plot shows that the residuals are normally distributed since none of the points deviate from the line. Here is what the scatterplot and Q-Q plot look like:

A red dot diagram with numbers and a white background

Description automatically generated A graph with red and blue lines

Description automatically generated

### Evaluating Significance of Model

To find out if the model is significant at a 5% level of significance, a overall F-test has to be carried out. The null hypothesis looks like this:

The alternative hypothesis looks like this:

The p-value of the model is 2.2e-16. This is less than the 0.05 or 5% level of significance. This means that the null hypothesis is rejected, and the alternative hypothesis is accepted. The results show that the model is significant at 5% level of significance. From the results, it can be concluded that at least one of the predictor variables has a significant relationship with the response variable (price of home).

To figure out which variables have a significant relationship with the response variable (price of home) at a 5%level of significance, a individual beta test has to be conducted for each of the predictor variables. The null hypothesis looks like this:

The alternative hypothesis looks like this:

The p-value for living area (sqft\_living) is 2e-16, which is less than 0.05 or 5%. The p-value for upper-level living area (sqft\_above) is 0.00894, which is less than 0.05 or 5%. The p-value for the age of home (age) is 2e-16, which is less than 0.05 or 5%. The p-value for the number of bathrooms (bathrooms) is 9.13e-13, which is less than 0.05 or 5%. The p-value for each of the view variables is 2e-16, which is less than 0.05 or 5%. Based on these results, the null hypothesis is rejected, and the alternative hypothesis is accepted. Also based on the results, it can be determined that all the predicted variables have a significant relationship with the response variable (price of home) at a 5% level of significance.

### 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 the road is 459828.2. The prediction interval falls between 239563 and 680093.4. This means that given any restrictions or errors, there is a 90% chance that a data point will fall within that range. The confidence interval falls between 446087.9 and 473568.5. 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 a lake is 1073285. The prediction interval falls between 852522.6 and 1296048. This means that there is a 90% chance that a data point will fall within that range. The confidence interval is between 1045117 and 1103454. There is uncertainty involved with the parameters included with prediction and confidence intervals. Error is also involved with the prediction interval as well. That is why there is a difference between the confidence and prediction intervals. The prediction interval is wider than the confidence interval due to the prediction interval being an estimate.

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

### Correlation Analysis

For model 2, a second order regression is used. Two scatterplots were created to show the relationship between price vs average school rating in the area and price vs the crime rate per 100,000 people. Looking at the scatterplot price vs average school rating, there is a positive relationship. This means that when school rating rises so does the price of a home. Looking at the scatterplot price vs crime rate, there is a negative relationship. This means that as the crime rate goes up the price of a home goes down. The scatterplots for each are down below:

A graph of a scatter plot

Description automatically generated A graph of a crime rate

Description automatically generated

### Reporting Results

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

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

{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mover><mi>y</mi><mo>&#x2227;</mo></mover><mo>=</mo><mover><msub><mi>&#x3B2;</mi><mn>0</mn></msub><mo>&#x2227;</mo></mover><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>1</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>2</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>3</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>4</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msubsup><mi>x</mi><mn>1</mn><mn>2</mn></msubsup><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>5</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msubsup><mi>x</mi><mn>2</mn><mn>2</mn></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"}

The complete second order model for price using average school rating in the area and crime rate per 100,000 people as predictors is:

The value of is 0.8088 and the value of is 0.8084. The results of the value of this model indicates that there is an 80% variation between the predictor variables (school ratings and crime) and the response variable (price of homes). Looking at the scatterplot that shows fitted values vs. residuals, it shows no clear pattern. It is believed that this model has homoscedasticity. Looking at the Q-Q plot, the points don’t deviate far from the line. This means the results of the Q-Q plot shows that the residuals are normally distributed.

The plots are shown down below:

A red dot diagram with white text

Description automatically generated A graph with red line and blue line

Description automatically generated

### Evaluating Significance of Model

To find if the model is significant at a 5% level of significance, an F-test needs to be conducted. The null hypothesis is:

The alternative hypothesis is:

The p-value for the F-test is 2.2e-16. This is less than 0.05 or 5%. The null hypothesis is rejected, and the alternative hypothesis is accepted. This means that the model is significant at a 5% level of significance. This also means at one of the predictor variables (school rating and crime rate) has a significant relationship with the response variable (price). To find out which predictor variable(s) is significant with the response variable at a 5% level of significance, individual beta tests must be conducted for each predictor variable. The null hypothesis looks like this:

The alternative hypothesis looks like this:

The p-value for school rating is 0.00406 which is less than 0.05 or 5%. Crime has a p-value of 1.90e-09 which is less than 0.05 or 5%. I(school\_rating^2) has a p-value of 2e-16 which is less than 0.05 or 5%. The p-value of I(crime^2) is 2e-16 which is less than 0.05 or 5%. The school\_rating:crime variable has a p-value of 0.2815 (rounded to 4 decimal places). The results show that all the predictor variables except for interaction variable school\_rating:crime has a statical significance at a 5% level of significance.

### Making Predictions Using Model

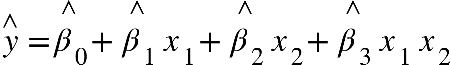
The predicted price for a home in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is 874497. The prediction interval of 90% is between 721696.2 and 1027388. This means that there is a 90% certainty that at least one data point will fall within this range. The confidence interval of 90% is between 863681.4 and 885312.7. This means that homes that have these parameters will fall within this range of numbers. 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 199706.7. The prediction interval at 90% is between 46911.65 and 352421.7. This means that there is 90% certainty that a data point will fall within this range. The confidence interval at 90% is 191753.5 and 207659.9. This means that homes that have these parameters will fall within this range of numbers.

## 5. Nested Models F-Test

### Reporting Results

The general form of the first order model for price using average school rating in the area and crime rate per 100,000 people is:

The prediction equation of the first order model for price using average school rating in the area and crime rate per 100,000 people is:



The first order regression model for price using average school rating in the area and crime rate per 100,000 people is:

### Evaluating Significance of Model

The overall F-test involves the null hypothesis and alternative hypothesis. The null hypothesis is:

The alternative hypothesis is:

The p-value for as part of the overall F-test is 2.2e-16 which is less than 0.05 or 5%. This means that the null hypothesis is rejected, and the alternative hypothesis is accepted. This also means that this model is significant at a 5% level of significance. The results also show that at least one of the predictor variables (school rating and crime) has a significant relationship with the response variable (price of home). Individual beta tests must be conducted to find which variable is significant to the price of a home. The null hypothesis looks like this:

The alternative hypothesis looks like:

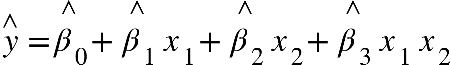
The school\_rating, crime, and the interaction between school\_rating:crime all have a p-value of 2e-16 which is less than 0.05 or 5%. The null hypothesis is rejected, and the alternative hypothesis is accepted. This means that all the predictor variables have a significant relation with the response variable (price of home).

### Model Comparison

When comparing this model with the second order model for pricing using average school rating in the area and crime rate per 100,000 people, a nested model F-test needs to be performed. This will help determine if the quadratic terms can help predict the prices of a home. The reduced model takes some of the terms into consideration while the complete model uses all the terms into consideration. Conducting the nested model F-test helps determine which variables can be ignored during testing or if the variables must be included. The F-test also helps decide if the full second order model must used instead of the reduced one.

The general form of the reduced model is:

The prediction equation for the reduced model is:



The general form for the complete model is:

The prediction equation for the complete model is:

{"mathml":"<math style=\"font-family:stix;font-size:16px;\" xmlns=\"http://www.w3.org/1998/Math/MathML\"><mstyle mathsize=\"16px\"><mover><mi>y</mi><mo>&#x2227;</mo></mover><mo>=</mo><mover><msub><mi>&#x3B2;</mi><mn>0</mn></msub><mo>&#x2227;</mo></mover><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>1</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>2</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>3</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msub><mi>x</mi><mn>1</mn></msub><mo>&#xA0;</mo><msub><mi>x</mi><mn>2</mn></msub><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>4</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msubsup><mi>x</mi><mn>1</mn><mn>2</mn></msubsup><mo>+</mo><mover><msub><mi>&#x3B2;</mi><mn>5</mn></msub><mo>&#x2227;</mo></mover><mo>&#xA0;</mo><msubsup><mi>x</mi><mn>2</mn><mn>2</mn></msubsup></mstyle></math>","origin":"MathType for Microsoft Add-in"}

The nested model F-test is ran to test a 5%level of significance to evaluate if the quadratic (squared terms) are needed. The null hypothesis is:

The alternative hypothesis is:

The nested model has a p-value of 2.22716e-28. The results determined that the null hypothesis should be rejected since the p-value is less than 0.05 or 5% level of significance. The results also show that the quadratic terms are significant and should be included in the model. The reduced model is insufficient which means that the complete second order model needs to be used.

## 6. Conclusion

While working on this analysis, a first order regression model, a complete second order regression model, a reduced model, and a nested F-test were utilized. The first order regression model was used to determine if living area, upper-level area, age, bathrooms, and view had a significant relationship with how a home was priced. The results showed that all the variables had a significant statical relationship. The complete second order regression model was used to determine if there was a significant relationship between the price of a home, the average school rating and the crime rate. The results showed that each of the variables besides the interaction term of school rating and crime had a significant statistical relationship. The reduced model also showed a significant relationship between the price of a home and all the predictor variables. I would not use the reduced model since the F-test indicated that quadratic (squared terms) should be added to the model. I would choose to use the complete second order model that had school rating, crime rate, school rating squared, crime rate squared, and the interaction term of school rating and crime rate. The variance is better explained with the second regression order model due to having an 80% variance compared to the 60% variance with the first order regression model. The analysis can be used by real estate companies to help determine the correct pricing for a home based on these variables. The real estate companies can also use this model to test for different variables as well. The results show that the variables tested in this analysis have a significant relationship in determining the pricing for homes. The visual results also show that certain variables have a positive correlation, and others have negative correlations with pricing. All the results from this model will help real estate companies with creating more sales by pricing homes that clients want at the right price.