# MAT 303 Project One Summary Report

Kathryn Delaney-Seamans

Kathryn.delaney-seamans@snhu.edu

Southern New Hampshire University

## 1. Introduction

In Project One, I will be exploring the housing\_v2 data set, which contains information on several houses in regards to variables like price of the home, age, living area, and more. For the first model, we will focus on price in relation to living area square footage and price to start, then looking at variables like number of bathrooms and view for a regression model. The second model will focus on comparing price with factors like average school rating and crime rate per 100,000 people. The results of my analysis will be used to better understand the relationship between price of a home and various factors that may influence it. This data may be useful for consumers and real estate agents alike to predict home prices based on the variables included in the study. Multiple regression analysis will be performed to look at the relationship between price and the predictor variables (living area, age of home, number of bathrooms, etc.). We will look at how the variables collectively influence the price of a home and examine qualitative and quantitative variables to build a comprehensive model.

## 2. Data Preparation

There are several variables in this data set that will be examined more closely. The price of the home will serve as the dependent variable. The independent variables will be square footage, number of bathrooms, average school rating, and crime rate per 100,000 people, as well as their interactions. There are 23 columns and 2692 rows.

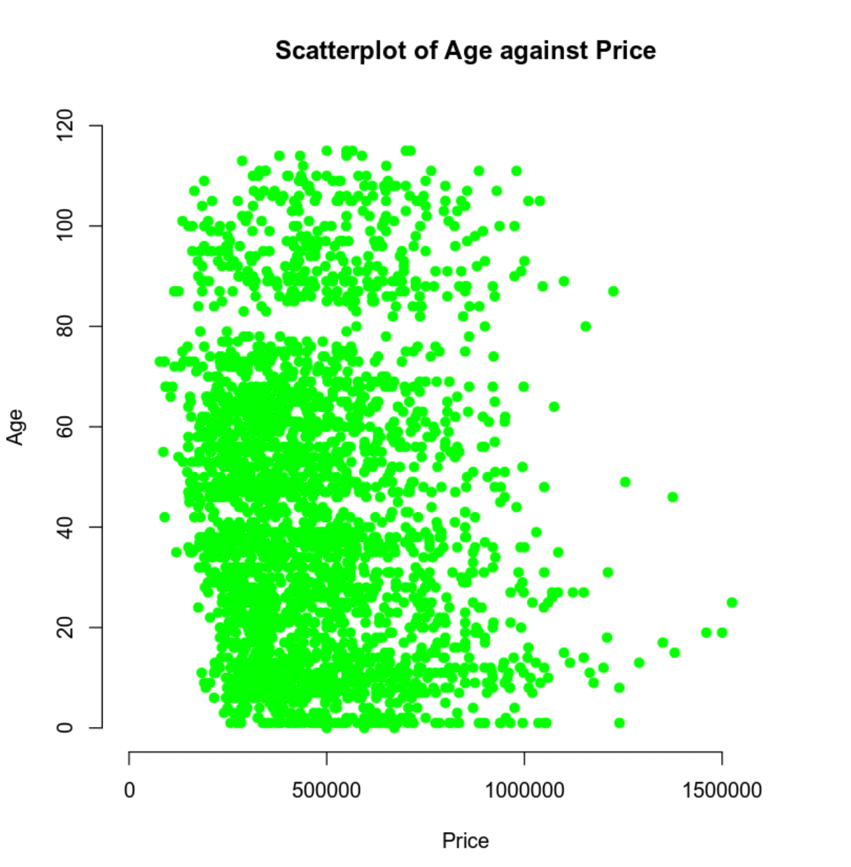
## 3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

### Correlation Analysis

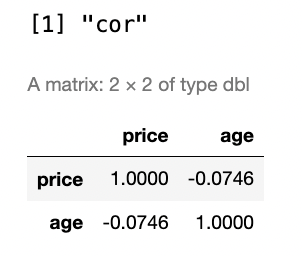
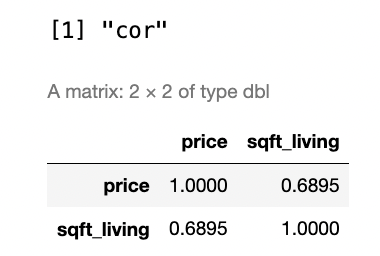
Price vs. Living Area Scatterplot

**

Price vs. Age Scatterplot:



There appears to be a positive trend for the scatterplot of living area against price. As the living area increases, the price of a home also increases. This would indicate that larger homes would tend to have higher prices. The trend in the scatterplot between age of a home and price is a bit more difficult to interpret. I see a wide distribution of prices across varying ages of homes. There is a cluster of data around the 500,000 price range, under 80 years old, which tells me that many of the homes in this data set fall in that trend. There does not seem to be a clear relationship between age and price.



The correlation coefficients for price vs. living area and price vs. age of the home are 0.6895 and -0.0746 respectively. The correlation coefficient for price vs. living area indicates a moderately positive correlation. For price vs. age of the home, there is a weak negative correlation between the two. The price may tend to decrease slightly as the age of the home increases. These correlation coefficients match the observations made from the scatterplots

### Reporting Results

The general form of this regression model is:

In this model represents living area, represents upper level living area, represents age of the home, represents number of bathrooms, and represents view.

The prediction equation of the multiple regression model is:

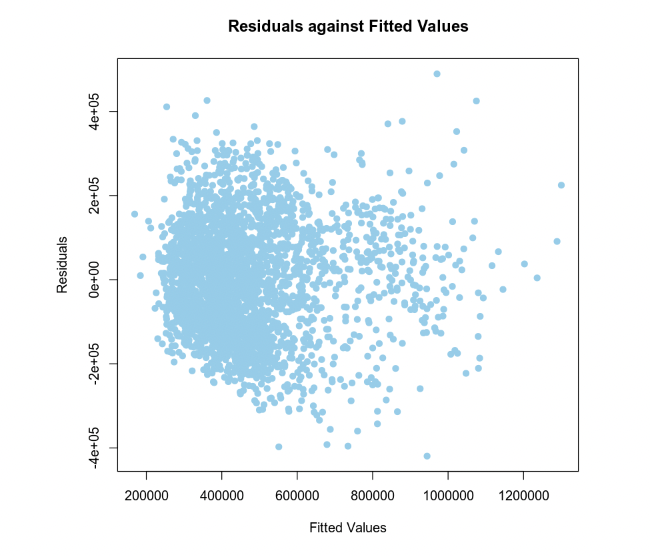
The prediction equation of the multiple regression model using outputs from the R script is:

The outputs from the multiple regression model indicated the difference in the variable view, where represents a view of trees and represents a lake view.

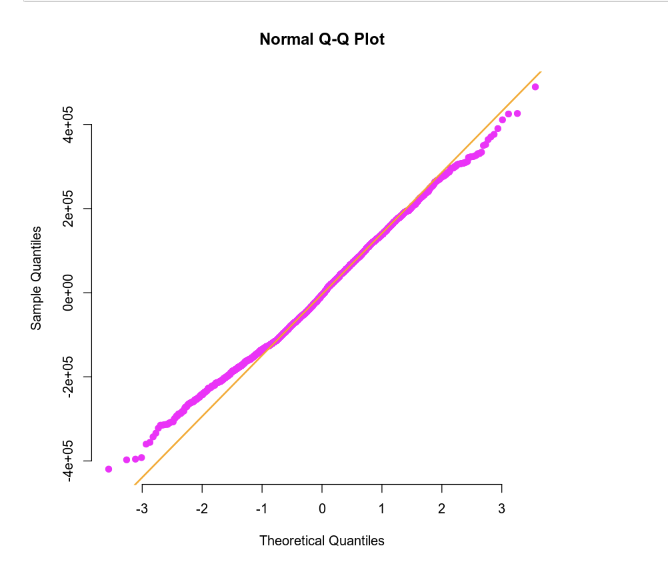
The value of is 0.6029 and the value of is 0.602. The value indicates that approximately 60.29% of the variance in price can be explained by the independent variables in the model. The value is the adjusted value to account for the number of predictors in the model and the sample size. These values suggest a reasonably good fit, however, it also may indicate that there are other factors that affect the price of homes not included in this model, such as the interaction between the variables included.

The coefficient for living area is 129.3. This indicates that for every additional square foot of living space, the price of the home will increase by $129.30, assuming that other variables are held constant. We can interpret this as that larger homes will have an increased selling price. The coefficient for lake view is 249000. This indicates that having a lake view significantly increases the home’s price. If two identical homes were compared, with one having a lake view and the other not having a lake view, the one with the lake view is predicted to be worth $249,000 more. This tells us that homes with a lake view will have a substantially increased price.

Residuals against Fitted Values



Normal Q-Q Plot



Based on the plots above, the residuals against fitted values plot shows that the values are randomly scattered around the horizontal axis with no discernable patterns or trends. Therefore, the assumption of homoscedasticity is likely met by looking at the plot.

In terms of the normality of the residuals, the plotted points lie relatively close to the diagonal line, so we could discern that the residuals are normally distributed.

### Evaluating Significance of Model

The p-value associated with the F-test is less than . This is far below the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that the model is statistically significant.

After examining the T-tests, the information gathered is listed below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| Sqft\_living |  | t = 15.916 | p < 2 | Reject the null |
| Sqft\_upper |  | t = 2.616 | p = 0.00894 | Reject the null |
| age |  | t = 12.098 | p < 2 | Reject the null |
| bathrooms |  | t = 7.178 | p = 9.13 | Reject the null |
| View 1 (trees) |  | t = 15.640 | p < 2 | Reject the null |
| View 2 (lake) |  | t = 20.739 | p < 2 | Reject the null |

All values exhibit a high level of significance and are below the 5% threshold.

### Making Predictions Using Model

The predicted price for a home that 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 road is $459,828.20. In terms of the 90% prediction interval for the home described above, the interval is [239563, 680093.4]. This signifies we can be 90% confident that the price will fall within this interval. The 90% confidence interval is [446087.9, 473568.5]. This indicates that we can be 90% confident that the true mean price will fall within the range above. The prediction interval is wider than the confidence interval because it accounts for both the uncertainty in estimating the mean response and the variability of individual responses around that mean.

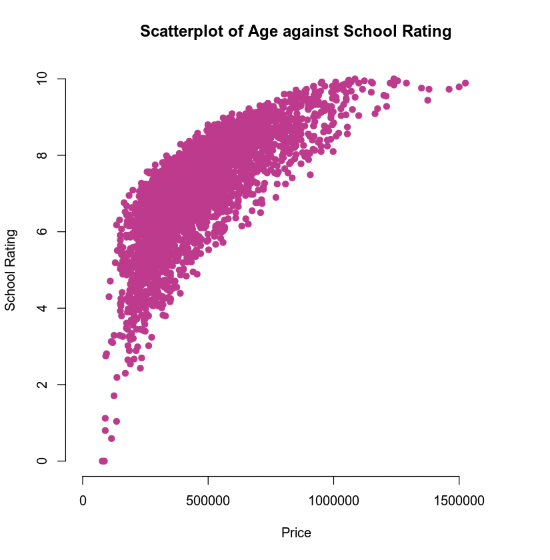
The predicted price for a home that has 4250 square feet of living area, 2100 square feet of upper level living area, is 5 years old, has 5 bathrooms, and backs out to a lake is $1,074,285. In terms of the 90% prediction interval for the home described above, the interval is [852522.6, 1296048]. This signifies we can be 90% confident that the price will fall within this interval. The 90% confidence interval is [1045117, 1103454]. This indicates that we can be 90% confident that the true mean price will fall within the range above.

The prediction interval is wider than the confidence interval in both cases because it accounts for both the uncertainty in estimating the mean response and the variability of individual responses around that mean. A prediction interval is designed to say that a single response will fall within that interval, while a confidence interval reflects where we expect the mean response to fall.

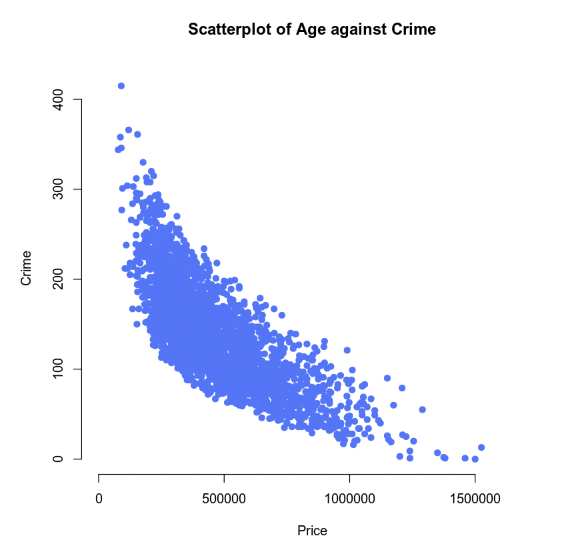
## 4. Model #2 - Complete Second Order Regression Model with Quantitative Variables

### Correlation Analysis

Price vs. School Rating Scatterplot



Price vs. Crime Scatterplot



There appears to be a strong positive trend for the scatterplot of school rating against price. As the school rating increases, the price of a home also increases. I would argue that there is a non-linear relationship between these two variables because there is a sharp incline in home price with the initial rise of school rating, then, the rate of increase does slow down as home price continues to rise. This may indicate that price does not change at a constant rate and a second order model may be more appropriate in this case. In terms of crime against price, there appears to be a strong negative trend. Again, this seems like a non-linear relationship because there is a sharp decline with the initial rise of home price, then, the rate of decrease does slow down as home price continues to rise. This may indicate that crime does decrease initially as home prices increase, but that the rate of change stagnates as in the more expensive homes. I believe second order regression models would be appropriate both these variables.

### Reporting Results

The general form of the second order regression model is:

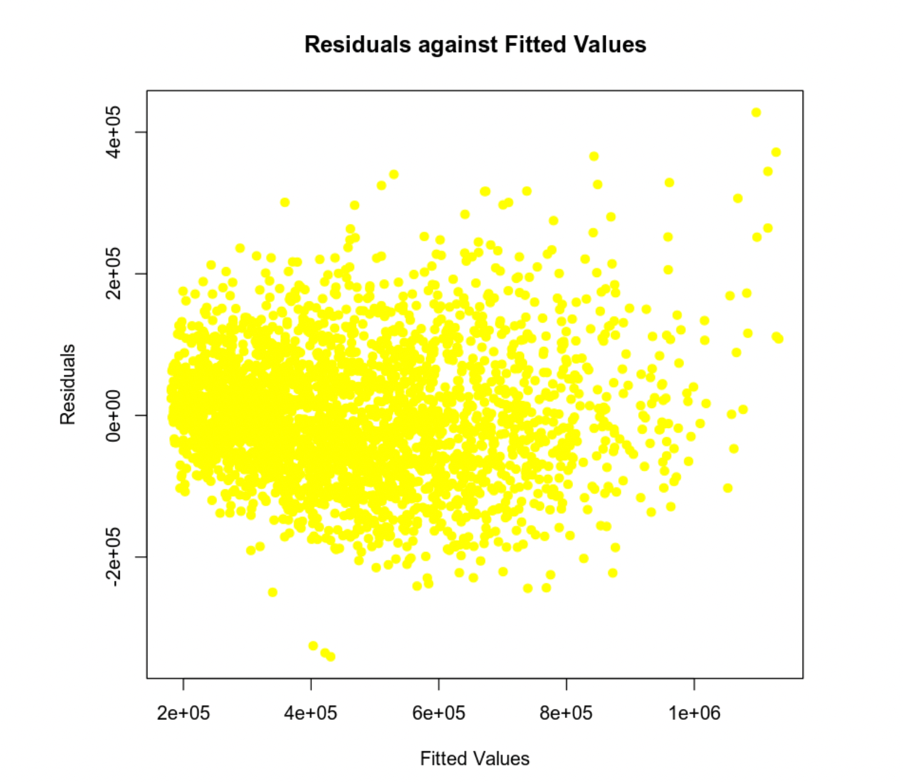
In this model, x1 represents school rating and x2 represents crime rate.

The prediction equation is as follows:

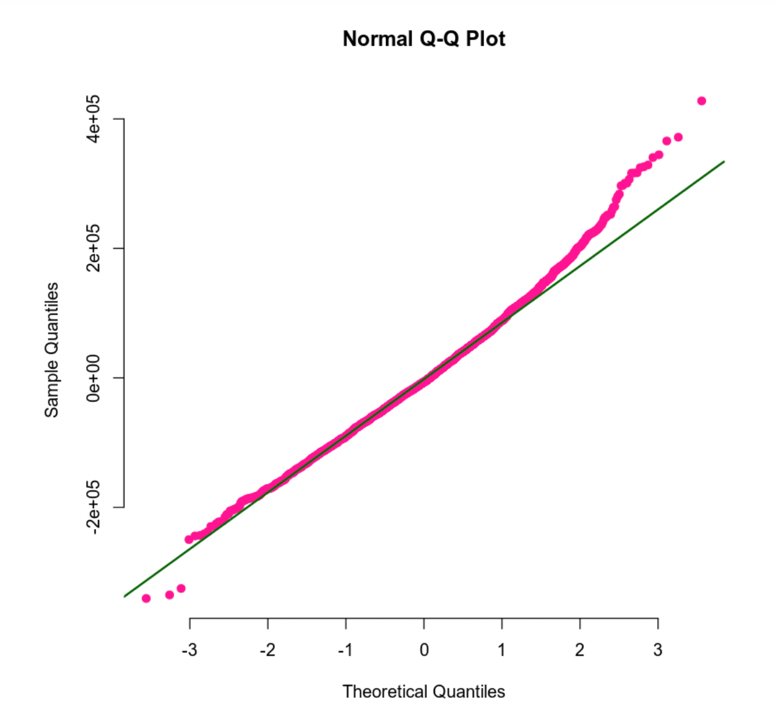
The second order regression model is:

The values of and are 0.8088 and 0.8084 respectively. The value is the ratio of variance in the dependent variable (price) that is explained by the independent variables, school rating, school rating2, crime, crime2, and the interaction between school rating and crime. It indicates that approximately 80.81% of the variance in wage price can be explained by the predictor variables. The high value suggests a strong fit of the model to the data. is very slightly lower than the value. This again suggests that the model is a good fit for the data gathered.

Residuals against Fitted Values



Normal Q-Q Plot



Based on the plots above, the residuals against fitted values plot shows that the values are randomly scattered around the horizontal axis with no discernable patterns or trends. Therefore, the assumption of homoscedasticity is likely met by looking at the plot.

In terms of the normality of the residuals, the plotted points lie relatively close to the diagonal line, except above 2 and below -3, where the data has a slight deviation. However, we can still assume the residuals are normally distributed.

### Evaluating Significance of Model

The p-value associated with the F-test is less than . This is far below the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that the model is statistically significant.

After examining the T-tests, the information gathered is listed below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| School rating |  | t = -3.541 | p =0.000406 | Reject the null |
| crime |  | t = -6.207 | p = 1.9 | Reject the null |
| School rating2 |  | t = 10.497 | p < 2 | Reject the null |
| Crime2 |  | t = 8.777 | p < 2 | Reject the null |
| School rating:crime |  | t = -1.077 | p = 0.281513 | Fail to reject the null |

All values exhibit a high level of significance and are below the 5% threshold, except for the interaction between school rating and crime. The p value in that case is above the 0.05 threshold at 0.281513.

### Making Predictions Using Model

The predicted price for a home that has an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is $874,497. In terms of the 90% prediction interval for the home described above, the interval is [721606.2, 1027388]. This signifies we can be 90% confident that the price will fall within this interval. The 90% confidence interval is [863681.4, 885312.7]. This indicates that we can be 90% confident that the true mean price will fall within the range above.

The predicted price for a home that has an average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is $199,706.70. In terms of the 90% prediction interval for the home described above, the interval is [46991.65, 352421.7]. This signifies we can be 90% confident that the price will fall within this interval. The 90% confidence interval is [191753.5, 207659.9]. This indicates that we can be 90% confident that the true mean price will fall within the range above.

## 5. Nested Models F-Test

### Reporting Results

The general form of this regression model is:

In this model represents average school rating and represents crime rate per 100,000 people.

The prediction equation of the multiple regression model is:

The first order regression model is:

### Evaluating Significance of Model

The p-value associated with the F-test is less than . This is far below the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that the model is statistically significant.

After examining the T-tests, the information gathered is listed below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| School rating |  | t = 49.65 | p < 2.2 | Reject the null |
| crime |  | t = 17.20 | p < 2.2 | Reject the null |
| School rating:crime |  | t = -31.63 | p < 2.2 | Reject the null |

All values exhibit a high level of significance and are below the 5% threshold.

### Model Comparison

When comparing the two models, a reduced model is a model that contains a subset of the predictors and a complete model includes more variables such as additional predictors or terms. For example, in the complete model, we will be using quadratic terms.

The general form of the reduced model is:

In this model represents average school rating and represents crime rate per 100,000 people.

The prediction equation of the reduced model is:

The general form of the complete model is:

The prediction equation of the complete model is:

To determine if the quadratic terms are needed, we can run the nested model F-test.

The null hypothesis would be that the quadratic terms, are not needed (equal to 0) and the reduced model is sufficient. The alternative hypothesis is that at least one of the above terms is needed (and the complete model is necessary. The table below describes the results from the F test.

|  |  |  |  |
| --- | --- | --- | --- |
| Hypothesis | F-test | P value | Conclusion |
|  | 65.20513 |  | Reject the null hypothesis |

Since the p-value is less than 0.05, we can reject the null hypothesis and conclude that the quadratic terms should be used in the model. The complete model provides a better fit for the data than the reduced model.

## 6. Conclusion

There were three different models used: the first model with a significant number of predictors without any interaction or quadratic terms, the second model with less predictors but an interaction term and two quadratic terms, as well as the third model, with the same predictors as model two but without the quadratic terms.

Based on the analysis of the complete model (model 2), this is the best choice for predicting house prices. This model provides a better fit of the data compared to the other models. The nested model F-test showed that the quadratic terms improve the model’s fit, proven by the very small p-value 2.22716. This model showed that the coefficients for school rating, crime, and their quadratic terms were all statistically significant. The only term proven to not be statistically significant was the interaction between school rating and crime rate. This model had the highest value at 0.8088 and value at 0.8084, which again shows a good fit of the data.

This analysis has several important implications. This model can be used by home buyers, real estate professionals, and investors to help predict home prices. Politicians and policymakers can utilize the information in this analysis to make informed decisions regarding education and crime prevention to boost home value. Additionally, those involved in real estate investment and insurance could use this model to develop pricing based on the information gathered from these predictions.