# MAT 303 Project One Summary Report

Cadden Buist

cadden.buist@snhu.edu

Southern New Hampshire University

## 1. Introduction

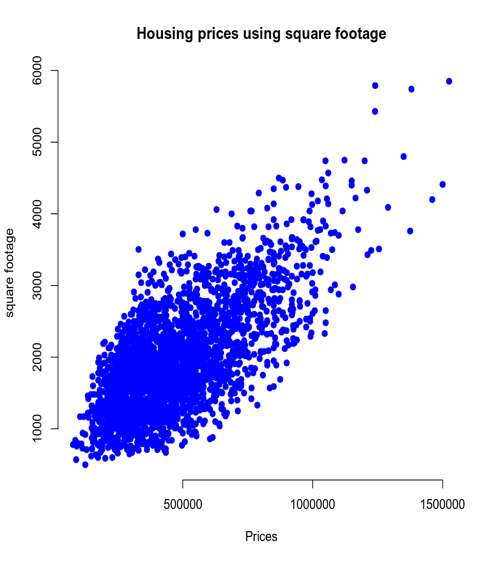
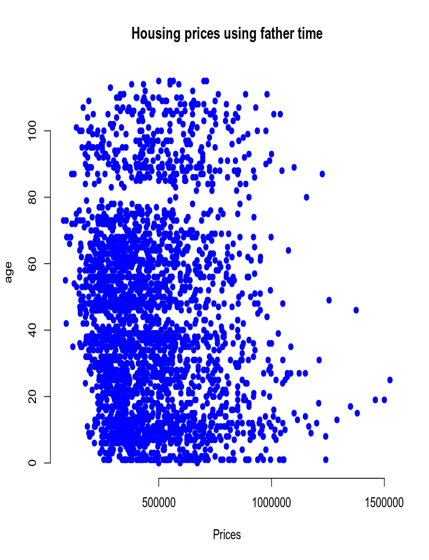
I will be exploring historical housing data set that lists housing attributes of housing such as square footage, quantity of rooms and bathrooms, etc. These results will help us to determine fair prices on our houses and get them sold within a more reasonable amount of time. They analyses that will be ran in this project will be first order models, second order models, and nested f-test. These factors and analyses that I will be running in this should help me find prices on these houses that are more reasonable for the customers but still give us a profit.

## 2. Data Preparation

Important variables in this data set is going to be price, bedrooms, bathrooms, sqft\_living, sqft\_above, sqft\_lot, age, grade, appliance\_age, crime, backyard, school\_rating, and view. There are 2692 rows and 6 columns in this dataset.

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

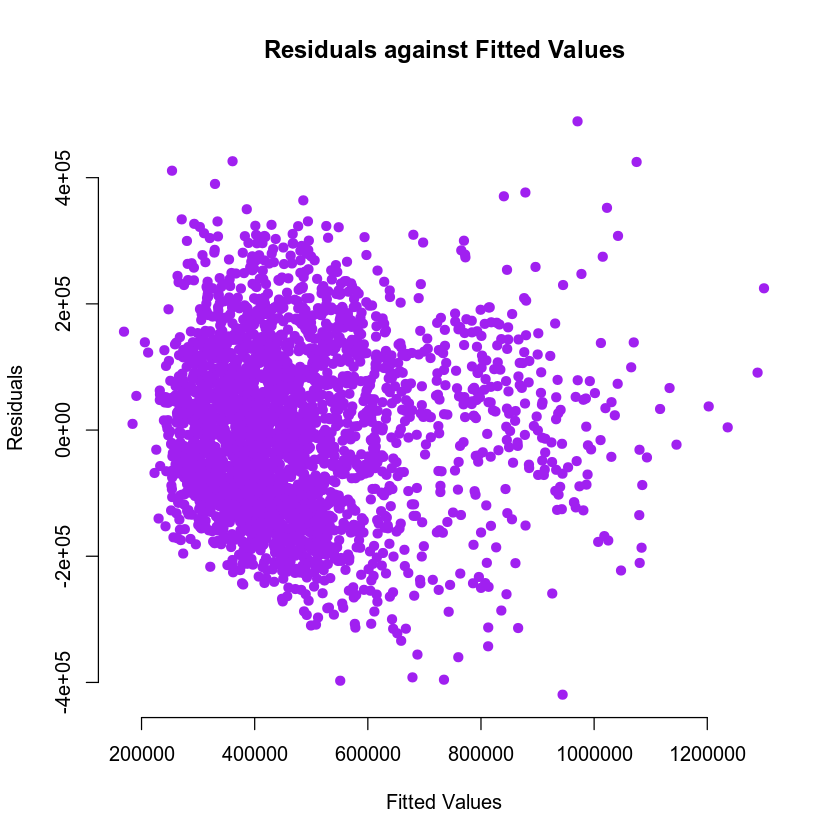
### Correlation Analysis

**

In the prices and square footage scatterplot it is easy to see that there is a positive moderate correlation while our age and price scatterplot shows no trend at all telling us that the age is not really a factor when pricing a house. From looking at the prices and square footage scatterplot we can visually see that as square footage increases so does the price of the house. On the other hand when using age it looks like even as the age of the house increases the value of the price remains the same. This tells me that square footage has a lot of pull on the prices while the age has yet to determine how much the house costs.

### Reporting Results

The general form equation for price, living area, upper level, age of the home, number of the bathrooms, and view would be E(y)=B0+B1X1+B2X2+B3X3+B4X4+B5X5 with E(y) being price, B0 is y-intercept, B1 is living area, B2 is upper level area, B3 is the age of the home, B4 is number of bathrooms, and B5 is view with X 1-5 being the amount for each term. The prediction formula is y=B0+B1X1+B2X2+B3X3+B4X4+B5X5 with the variables being the same terms as the general equation but y is the predicted value and the variables are estimates of the general equation with B1=1.293e3, B2=1.951e1, B3=1.451e3, B4=4.397e4, and B5=1.675e5 or 2.490e5. The R-squared value on the multiple regression model is 0.6029 and the adjusted r-square is 0.602 which tells us that our values fit this model at a moderate level but it isn’t a strong fit like what we would love to see. The beta estimates for living area and lake view are 1.293e^02 and 1.675e^05 which tells us that as price moves up so does living area and lake view by their respected beta estimates. As we can see below our residual against fitted value plot and our Q-Q plot have a strong correlation and a strong linear regression correlation. These plots are very accurate to where we expect them to be.



### A red line with green lines Description automatically generated

### Evaluating Significance of Model

This model is not significant at a 5 percent level as significance as we can see with the p-value being 2.2e-16 which tells us that the p-value is less than the level of significance value. This means we can reject the null hypothesis for this model. There is only one term that is significant at the level of significance at 5 percent and that is our y-intercept which doesn’t affect our model and we are able to continue using the set values we used for this model.

### Making Predictions Using Model

The first predicted price for the home has a predicted interval fit of 459,828.2 with an lower of 239,563 and an upper of 680,093.4 while the confidence interval has a fit identical to our predicted interval and a lower interval that is 446,087.9 and a upper interval that is 473,568.5. For the second home price we have a predicted interval fit of 1,074,285 a lower of 852,522.6 and an upper of 1,296,048 with the confidence interval fit identical to the prediction intervals like the first house with a lower interval that is 1,045,117 and an upper interval that is 1,103,454. The reason the predicted intervals are wider than the confidence intervals is because the prediction intervals account for the uncertainty of the populations mean and it random variations individual values.

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

### Correlation Analysis

*A graph of scatter plot

Description automatically generated*

**A purple dotted graph with white text

Description automatically generated**

We can see from the first scatterplot that as the school rating gets better the price for the homes goes up and eventually starts to straighten itself out. In the second plot we can see that as the house price goes up the crime drops a lot. After this analysis I believe a second order regression model is appropriate with price affecting both of these values.

### Reporting Results

The general form equation for price, school rating, and crime would be E(y)=B0+B1X1+B2X2 with E(y) being price, B0 is y-intercept, B1 school rating, B2 is crime with X 1-5 being the amount for each term, and B3= the school\_rating to crime value. The prediction equation is y=B0+B1X1+B2X2+B3X3 with y as price B0 the intercept is -410233.37, B1=155559.97, and B2=2230.07 and B3=-564.85. R-squared is 0.7995 or 75.95% and adjusted r-squared is 0.7993 or 79.93% which tells us that our values are a strong fit together. The two plots below have a moderate to low validity so we can assume that they match kind of close to what we expect our price outcome to be.

A diagram of a graph

Description automatically generated

*A graph with red line and green line

Description automatically generated*

### Evaluating Significance of Model

Yes, this model is significant at a 5% level of significance with our p-value=2.2e-16 which will be less than 0.05 or a 5% level of significance. The 3 values we have each have a p-value of 2e-16 letting us know that the null hypothesis can be rejected for the model and all of our values.

### Making Predictions Using Model

The fit for our predicted intervals for the first house is 767,820.9 with a lower of 584,820.6 and an upper price of 950,821.2. The confidence intervals for the first house has a fit of 767,820.9 a lower price of 760,030.6 and an upper price of 775,611.2. On our second house there is a prediction interval fit of 147,849.3 and a lower price of -35,142.11 with an upper price of 330,840.7. The confidence intervals on this house has a fit price of 147,849.3 a lower price of 140,270.6 and an upper price of 330,840.7. As we can see from these intervals school rating and crime have a huge impact on the price of a house.

## 5. Nested Models F-Test

### Reporting Results

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

The general and prediction equation are going to look similar with the general equation Ey=B0+B1X1+B2X2 and the prediction equation y=B0+B1X1+B2X2 with Ey and y being the price of the house both B0’s being the y-intercept, both B1’s being the school rating, and both B2’s being the crime per 100,000 residents with the X’s being our variables for our values. Our prediction model is y=190932.08+72408.15X1-1638.00.

### Evaluating Significance of Model

This model is significant at a 5% significance level which is greater than the p-value 2.2e-16 and so is the individual values that are all 2e-16 making all p-values less than 0.05 the 5% significance value. This means we can reject the null hypothesis for all of the values.

### Model Comparison

A reduced model is the model with fewer variables because it doesn’t have the full data that the other model does so in our case the complete model is our second regression model and our reduced model is our first regression model. The general form equation for the reduced model is Ey=B0+B1X1+B2X2 with the prediction equation just like it y=B0+B1X1+B2X2 and for this particular complete model it is going to be the exact same but with a B3 so it will be Ey=B0+B1X1+B2X2+B3X3. The Null hypothesis is H0=Bj=Bi+1=0 with the alternative hypothesis using the same formula but equals anything but 0. Our p-value is 5.863269e-187 which is much lower than 0.05 giving us the conclusion that we can reject the null hypothesis.

## 6. Conclusion

I would choose the first order regression model from our first test because it had way more factors taken into account then the other models did. I did have a lower r and alternative r-square then the other models but it took more into consideration for a house which is more important when it comes to the housing market. The practical importance of the analyses we did was to reinforce our skills with first and second order regression models and to use the nested models f-test in a realistic situation.