1. Introduction

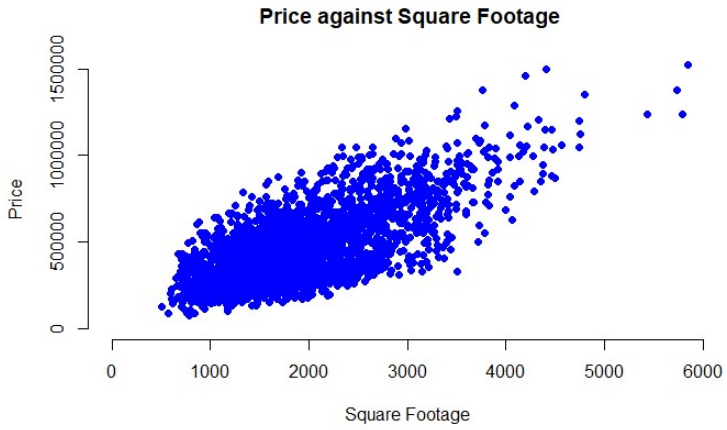
This project is based on a historical real estate dataset. It features attributes commonly associated with house buying and selling. Key variables include house price, square footage of the home and the lot, and additional qualitative factors such as the view from the home, the local crime and school ratings. By using this dataset, real estate companies can better price their homes for a quicker sale. Through the application of multiple and quadratic regression models both qualitative and quantitative variables will be analyzed in order to offer deeper insights into the property market and aiding faster transactions.

2. Data Preparation

This dataset contains 2692 rows and 23 columns. Important variables that will be considered for this project include price, bedrooms, square feet of living space, square feet above ground, view, crime rate and school rating.

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

Correlation Analysis



The above scatterplot suggests a positive linear relationship between the price and square footage of a home. I ran a correlation test between the two variables and the result was 0.6894838 which suggests a positive and moderately strong linear relationship. Based on this, it could be said that as the square footage of a property increases, the price is also likely to increase and vice versa. The value is closer to 1 but not quite there, which indicates that while the relationship clearly exists, it's not a perfectly direct one.

The above scatterplot shows a weak negative linear relationship between the price of a home and its age. This is further supported by a correlation test output of -0.07460764. This indicates that as the age of a home increases, its price tends to decrease slightly, and vice versa. With the correlation being closer to 0 than -1 it can be concluded that this is a weak correlation and home age doesn't significantly impact its price based on this particular dataset.

Reporting Results

General Form: Price = $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \varepsilon$

Prediction Equation: Price = $\beta_0 + \beta_1$ * sqft_living + β_2 * sqft_above + β_3 * age + β_4 * bathrooms + β_5 * view + ε

Model 1 multiple regression model:

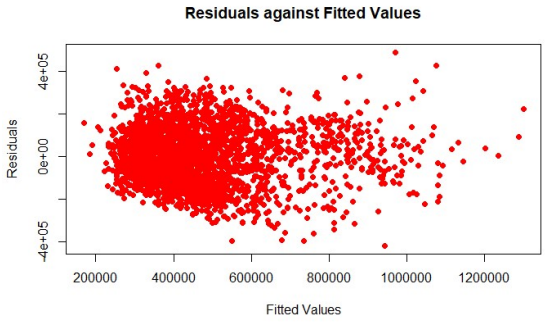
```
Call:
lm(formula = price ~ sqft_living + sqft_above + age + bathrooms +
    view, data = housing_v2)

Residuals:
    Min      1Q  Median      3Q     Max
-419299 -101792   -5606   93896  489323

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.709e+03  1.411e+04   0.546  0.58495
sqft_living 1.293e+02  8.123e+00  15.916  < 2e-16 ***
sqft_above  1.951e+01  7.458e+00   2.616  0.00894 **
age         1.451e+03  1.199e+02  12.098  < 2e-16 ***
bathrooms   4.397e+04  6.126e+03   7.178 9.13e-13 ***
view1       1.675e+05  1.071e+04  15.640  < 2e-16 ***
view2       2.490e+05  1.201e+04  20.739  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133600 on 2685 degrees of freedom
Multiple R-squared:  0.6029, Adjusted R-squared:  0.602
F-statistic: 679.3 on 6 and 2685 DF,  p-value: < 2.2e-16
```

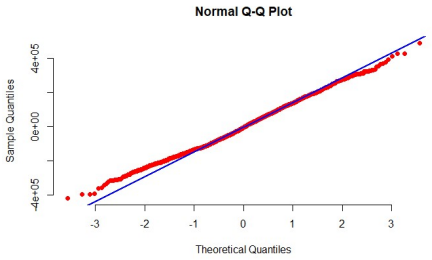
The R-squared value for this model is 0.6029, meaning that approximately 60.29% of the variability in the home price is explained by the variables in the model (living area, above area, age, number of bathrooms, and view). The adjusted R-squared value is 0.602, which takes into account the number of

predictors in the model. Since the adjusted R-squared is very close to the R-squared, it indicates that all the predictors are relevant to the model.

Interpretation of beta estimates for the living area and lake view:
The beta estimate for the living area (sqft_living) is approximately 129.3, indicating that for every additional square foot of living area, the home price should be expected to increase by about $129.3, given all other variables remain constant.

This model included 2 different categories for the 'view' variable shown as 'view1' and 'view2.' The beta estimate for 'view1' is approximately 167500 which suggests that homes with a 'view1' rating show an average price increase of $167,500 when compared to homes that do not. The beta estimate for 'view2' is approximately 249,000, which suggests that homes that a 'view2' rating have an average increase of about $249,000 when compared to homes without it, and under the assumption that all variables remain constant.





The scatterplot of residuals against fitted values does not display any clear pattern. Therefore, the assumption of homoscedasticity would be met. The plots on the normal Q-Q plot do not deviate from the line. The residuals falling along the reference line suggests a normal distribution which works in support of the validity of my linear model.

Evaluating Significance of Model

An overall F-test was conducted to examine whether at least one of the predictors is significantly related to the response variable.

The null hypothesis (H_0) states that all regression coefficients are equal to zero indicating no relationship between predictors and the response variable. The alternate hypothesis (H_1) states that at least one regression coefficient is not equal to zero indicating that there is a relationship between at least one predictor and the response variable.

The smallest p-value in the ANOVA table is less than 2.2e-16 which is practically zero and far less than 0.05. Therefore, the null hypothesis is rejected, and it can be concluded that at least one predictor is significantly related to the response variable at the 5% level of significance.

The corresponding p-values for all predictors, (sqft_living, sqft_above, age, bathrooms, view) well less than 0.05, indicating that each predictor significantly contributes to the model at the 5% significance level. In this case the null hypothesis should be rejected for all predictors.

Making Predictions Using Model

The equation for 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 road is as follows:

Price = 7709.048 + 129.2846 * 2150 + 19.51206 * 1050 + 1450.617 * 15 + 43970.12 * 3 + 167491.5 * 1 + ε

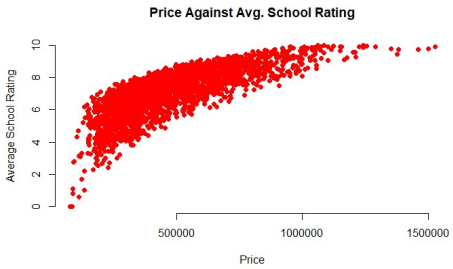
With the confidence interval being an estimate of the range in which the true population mean is expeted to fall based on the sample data we have a confidence interval that ranges from approximately $446,088 to $473,569. This means that, with 90% confidence the average price of a home with the features in the equation will fall within this range. The prediction interval is from approximately $239,563 to $680,093. This indicates that the average price of a similar home would fall within this range in the future.

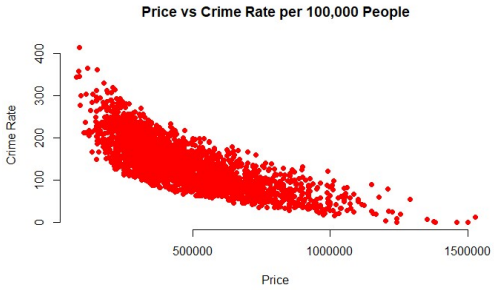
When taking into account a home that has 4250 sqft living area, 2100 sqft upper-level living area, is 5 years old, has 5 bathrooms, and backs out to a lake the equation produces a predicted price of:
Price = 1086988.465 + ε

According to the prediction interval, the predicted price for a home with 4250 sqft living area, 2100 sqft upper-level living area, 5 years old, 5 bathrooms, and a lake view is $1,074,285. This offers 90% confidence that is a new observation with these same characteristics was added, its price would fall somewhere between $852,522.60 and $1,296,048. The confidence interval makes reference to the reliability of the estimate of the mean response. Here we can see 90% confidence that the true average price of homes with these features is between $1,045,117 and $1,103,454. The prediction interval, which is intended to cover a single new observation, is wider due to the extra uncertainty associated with predicting a single observation as opposed to the mean of several observations.

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

Correlation Analysis





The scatterplot comparing average school rating with home prices demonstrates a positive correlation. This indicates that home prices tend to rise in areas with higher school ratings. This can be interpreted as a reflection of the perceived value that homeowners place on high-quality education. It may suggest that people are willing to invest more in a home if it affords their children access to better schools.

Then we have the scatterplot contrasting home prices with crime rates presents a negative correlation. In essence, as crime rates increase, the prices of homes tend to decrease. This underscores the importance homeowners place on safety and security. Areas with higher crime rates are typically less desirable, which can depress home values.

Reporting Results

The general form and 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:

$Price = \beta_0 + \beta_1*(school\_rating) + \beta_2*(crime) + \beta_3*(school\_rating\text{^}2) + \beta_4*(crime\text{^}2) + \beta_5*(school\_rating*crime) + \varepsilon$

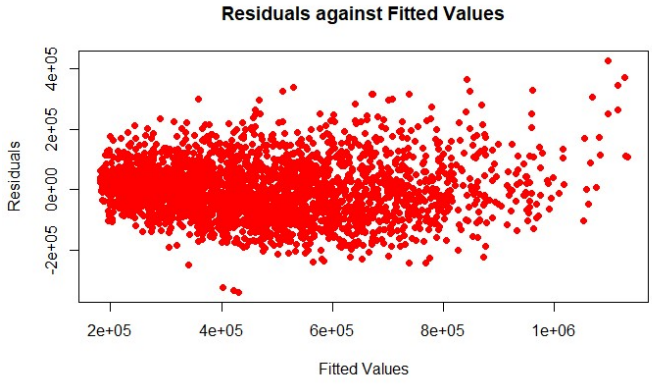
```
Call:
lm(formula = price ~ school_rating + crime + school_rating:crime +
    I(school_rating^2) + I(crime^2), data = housing_v2)

Residuals:
    Min      1Q  Median      3Q     Max
-340729  -61055   -6288   56875  427915

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.339e+05  1.032e+05   7.113 1.45e-12 ***
school_rating       -7.375e+04  2.083e+04  -3.541 0.000406 ***
crime               -3.155e+03  5.235e+02  -6.027 1.90e-09 ***
I(school_rating^2)   1.165e+04  1.109e+03  10.497  < 2e-16 ***
I(crime^2)           6.377e+00  7.265e-01   8.777  < 2e-16 ***
school_rating:crime -5.227e+01  4.853e+01  -1.077 0.281513
---
```

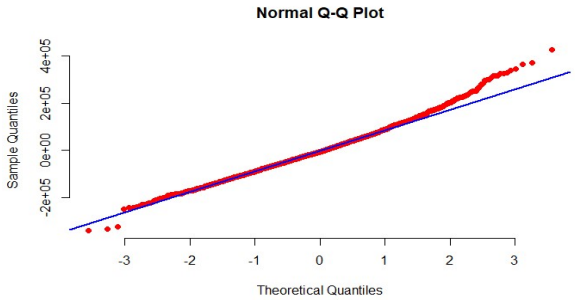
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92690 on 2686 degrees of freedom
Multiple R-squared:  0.8088, Adjusted R-squared:  0.8084
F-statistic:  2272 on 5 and 2686 DF,  p-value: < 2.2e-16
```

Replacing the coefficients with the values obtained from the R script the equation is:

Price = 733900 - 73750*(school_rating) - 3155*(crime) + 11650*(school_rating^2) + 6.377*(crime^2) - 52.27*(school_rating*crime) + ε

The R-squared value is 0.8088 which indicates that approximately 80.88% of the variation in the house prices can be explained by the average school rating in the area and the crime rate per 100,000 people. This relatively high value suggests that the model explains a large proportion of the variability in house prices. The adjusted R-squared value is 0.8084 which is very close to the R-squared value. This suggests that all of the variables in the model are contributing meaningfully to the prediction of the house prices, and that there isn't an issue with adding predictors that don't actually prove the model present.





From the scatterplot of the residuals shown above, there is no distinguishable pattern noted, and therefore homoscedasticity is assumed. The plots do not deviate significantly from the line in the Q-Q plot so it can be concluded that the plot holds residuals with normal distribution.

Evaluating Significance of Model

Based on the overall F-test, the null hypothesis (H_0), all predictor variables (school_rating, crime, I(school_rating^2), I(crime^2), school_rating:crime) have no effect on the home price and all the regression coefficients are zero.
The alternative hypothesis (H_1), states that at least one predictor variable has a non-zero coefficient. Given that the p-value of the F-statistic is less than 2.2e-16, the null hypothesis can be rejected which implies that at least one of the predictor variables significantly contributes to the model. Therefore, the model is significant at the 5% level.

Individual Beta Tests:

school_rating:
Null Hypothesis (H_0): The coefficient of school_rating is zero.
Alternative Hypothesis (H_1): The coefficient of school_rating is not zero.
Given the p-value is less than 2e-16, we reject H_0. School_rating is significant at a 5% level of significance.

crime:
Null Hypothesis (H_0): The coefficient of crime is zero.
Alternative Hypothesis (H_1): The coefficient of crime is not zero.
Given the p-value is less than 2e-16, we reject H_0. Crime is significant at a 5% level of significance.

I(school_rating^2):
Null Hypothesis (H_0): The coefficient of I(school_rating^2) is zero.
Alternative Hypothesis (H_1): The coefficient of I(school_rating^2) is not zero.
Given the p-value is less than 2e-16, we reject H_0. I(school_rating^2) is significant at a 5% level of significance.

I(crime^2):
Null Hypothesis (H_0): The coefficient of I(crime^2) is zero.
Alternative Hypothesis (H_1): The coefficient of I(crime^2) is not zero.
Given the p-value is less than 2e-16, we reject H_0. I(crime^2) is significant at a 5% level of significance.

school_rating:crime (interaction term):
Null Hypothesis (H_0): The coefficient of the interaction term is zero.
Alternative Hypothesis (H_1): The coefficient of the interaction term is not zero.
Given the p-value is 0.2815, which is greater than 0.05, we fail to reject H0. Hence, the interaction term is not significant at a 5% level of significance.

This information shows that all variables are significant at a 5% level of significance except for the interaction term (school_rating:crime).

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 1000,000 individuals is approximately $874,497. The 90% prediction interval is between $721,606 and $1,027,388. The means that we can be 90% certain that the true price of a home in this area would fall within this range, considering room for error. The 90% confidence interval for this prediction is between $863,681.4 and $885,312.7. This means that we can be 90% confident that the average price for a home in this area lies within this range. The

confidence interval is narrower than the prediction interval as it does not account for the error variance, only the uncertainty about the population parameter.
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 approximately $199,706. The 90% prediction interval for this prediction is between $46,991.65 and $352,421.7. This means that we can be 90% sure that the true price for a home in this area would fall within this range, considering any variability in the predictions. The 90% confidence interval for this prediction is between $191,753.5 and $207,659.9. This implies that we can be 90% confident that the average price for a home in this area lies within this range.

5. Nested Models F-Test

Reporting Results

The general form for a first-order model that includes an interaction term between two predictor variables would be written as:

$Price = \beta_0 + \beta_1 * school\_rating + \beta_2 * crime + \beta_3 * school\_rating * crime + \varepsilon$

Evaluating Significance of Model

The null hypothesis (H_0: $\beta_1 = \beta_2 \ldots = \beta_n = 0$): would be that all the predictor variables (crime rate, school rating, and their interaction) in the model are not statistically significant, and all the regression coefficients are equal to zero.

The alternative hypothesis (H_a: at least one β_1 does not = 0 for I = 1, 2, ….,n) would be that at least one predictor variable is significant, meaning at least one of the regression coefficients is not equal to zero.

The p-values associated with the crime rate, school rating, and the interaction term between the two, are identical at a value of 2.2e-16. This minute p-value can be rounded to zero, which is below our designated threshold of 5%, leading us to reject the null hypothesis for each of these variables. This suggests that each of these variables - crime rate, school rating, and their combined influence - holds statistical significance in this model.

Model Comparison

In regression analysis, when comparing two models, the "reduced model" is a simpler model with fewer predictor variables, while the "complete model" includes all potential predictor variables, as well as potential interactions and higher order terms of the variables.

The reduced model and the complete model were compared to see if the additional variable in the complete model significantly improves the fit to the data. This comparison is done using a nested model F-test.
The general form of the reduced model is:
$Price = \beta_0 + \beta_1 * school\_rating + \beta_2 * crime + \varepsilon$

The general form of the complete model is:

$Price = \beta_0 + \beta_1 * school\_rating + \beta_2 * crime + \beta_3 * school\_rating * crime + \varepsilon$

Null Hypothesis (H_0): The reduced model is sufficient to explain the data. This implies that the interaction term (school_rating:crime) does not significantly contribute to the prediction of the price.

Alternative Hypothesis (H_a): The complete model is better at explaining the data, implying that the interaction term does significantly contribute to the prediction of the price.

The test result shows a p-value of less than 2.2e-16, which is less than the 5% significance level.

The null hypothesis should then be rejected, concluding that the complete model with the interaction term provides a significantly better fit for the data compared to the reduced model. This indicates that including the interaction between school_rating and crime significantly improves the prediction of the home prices.

6. Conclusion

In this analysis, I leveraged several statistical models including a first-order regression model, a comprehensive second-order model, a reduced model, and a nested F-test to compare the reduced and complete models. These tools helped to ascertain whether the simplified model would suffice or if the more detailed second-order model was necessary.

The first-order regression model revealed a significant statistical relationship between the price of a home and a number of variables including its living area, above-ground square footage, age, the number of bathrooms, and the view from the property.

In the complete second-order regression model, the home's price showed a significant correlation with the average school rating in the area and the crime rate.

While the reduced model successfully highlighted the statistical significance between home value and its associated factors, it would not be my preference for use. The nested F-test indicated that the inclusion of additional predictor variables was necessary for an accurate model, which makes the complete second-order model a more fitting choice.

This comprehensive model incorporated crime rate, school rating, the crime rate and school ratings, and the interaction term for school and crime rating as predictors. It yielded the highest R-squared value, indicating its superior capacity to explain the variance among variables. This model also illustrated the connection between these variables and the home's value, our response variable.

These insights can be valuable for real estate professionals when determining a property's optimal pricing, considering the influence of its various attributes. The significant correlation found between these variables and home values suggests that these models can serve as tools for pricing homes accurately.