IE Group 4 Final Report: Home Price

Daniel Huang, Nicole Bronchuk, Ashley Hoang, Jacqueline Ryan

May 11, 2021

1 Introduction

Buying a home is a hallmark of the American Dream, so determining a fair home price is crucial to both selling and buying homes. While in the past realtors have relied on local knowledge and common sense to price homes, utilizing mathematical models such as linear regression allows us to determine house prices using more analytical methods. Many housing price studies have been conducted over the years, and these studies have very similar goals to ours — to create some model to estimate housing prices based on certain factors. It is important to consider the numerous factors that play a role in pricing, because every house is not one in the same as mentioned in [Pal84]. Understanding and having a working knowledge of housing demand is also integral to the housing market and possible related services. Some studies have also looked into specific geographical factors and their effect on house pricing. A case study conducted in Eugene, Oregon, the same area in which we will be focused on, measured the impact proximity of bus stations had on single-family home pricing [PCMC17]. In our report, after examining the data collected from 76 single-family homes in Eugene, Oregon, our goal is to create a working model that helps determine a reasonable home price in this location based on the many characteristics of the home given to us. From our own initial beliefs and a similar study done in LA, we hypothesize that some of the key factors will include but may not be limited to square feet, number of bedrooms, number of bathrooms, the year built, and schools nearby [Yua19].

2 Data Exploration

2.1 Data Overview

The data file contains information on 76 single-family homes in Eugene, Oregon during 2005. There are 17 specific variables observed and recorded for each home, 1 higher order variable, and 1 interaction term. These variables are ID, $Price$ (our response variable), $Size$, Lot, $Bath$, Bed, $BathBed$ (interaction term), $Year$, Age, Age^2 (higher order variable), $Garage$, $Status$, $Active$, $Elem$, $Edison$, $Harris$, $Adams$, $Crest$, and $Parker$.

At the time the data was collected, the data submitter was preparing to place his house on the market, and it was important for him to be able to come up with a reasonable asking price. He

collected and put all this information together to help himself make that decision. Each single-family home was assigned an ID number from 1-76 as reference. This is represented by the variable ID.

2.2 Quantitative Variables

Nine of the variables are quantitative. These variables are $Price$, $Size$, $Bath$, Bed, $BathBed$, $Year$, Age, Age^2, and $Garage$. The sale price of each home is given in thousands of dollars and the floor size is listed for each home in thousands of square feet. The number of bathrooms and bedrooms are recorded with half-bathrooms being represented as 0.1 and the number of bedrooms ranging from 2 to 6. The variable $BathBed$ is the interaction term between the two variables $Bath$ and Bed. The variable $Year$ is straightforward being the year that the home was built, and Age is calculated by dividing the difference that results from subtracting 1970 from the year the home was built by 10. Thus, Age is measured in decades. The variable Age^2 takes the variable Age and squares it in order to obtain a positive number so that whether the age is positive or negative is disregarded. Lastly, the garage size is quantified by the number of cars it can fit, ranging from 0 to 3.

2.3 Categorical Variables

Three of the variables are categorical. These include Lot, $Status$, and $Elem$. The lot size of each single-family home is categorized by being assigned a number from 1 to 11. The status of each home is classified as act, pen, or sld, meaning active listing, pending sale, or sold, respectively. Lastly, the nearest elementary school for each home is recorded. Among these elementary schools are Edgewood, Edison, Harris, Adams, Crest, and Parker.

2.4 Elementary School and Active

The last six variables are indicator variables, which take on the value 0 or 1 and help to more clearly identify homes with certain characteristics. Each home has the indicator variable $Active$ (that corresponds to $Status$), with 1 representing a home active on the market and 0 representing a pending or sold home. Additionally, each home has five indicator variables corresponding to the closest elementary school. Each of the elementary schools except for Edgewood has an indicator variable corresponding to whether or not it is the closest school to that home. If all these indicators are 0, then the house is closest to Edgewood.

2.5 Relationships and Residual Analysis

After conducting single variable regression between $Price$ and each other predictor and obtaining the corresponding residual graphs, we found a few interesting results. We discovered that $Bath$ and $Price$ were directly related and Bed and $Price$ were inversely related when we plotted $Bath$ and Bed individually against $Price$. In other words, $Price$ increased with number of bathrooms

but decreased with number of bedrooms. It seemed like there could be some relation between $Bath$ and Bed since their interaction term $BathBed$ was given. From the residual graph of $BathBed$ shown below, the linear model seems appropriate and there is constant variance.

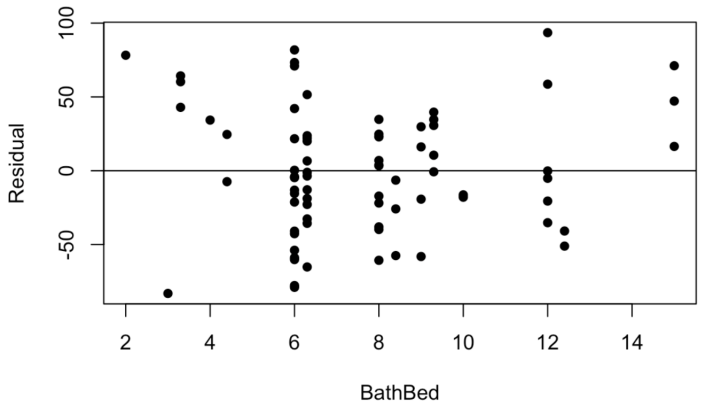
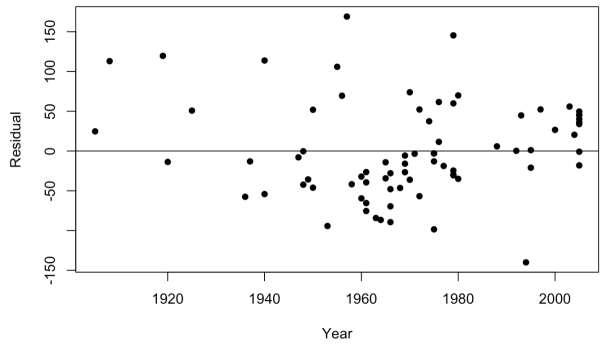
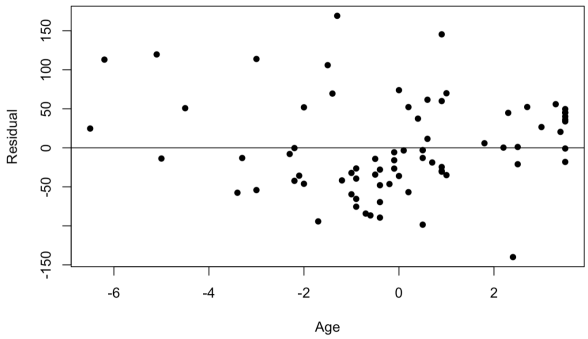


Figure 1: $BathBed$ Residual Graph.

The predictor variables $Year$ and Age we found to simply be transformations of each other. Therefore, their graphs were identical with the only difference being the numbers on the x-axis as seen below, so there was only need to focus on one and Age seemed to be a better quantifier.



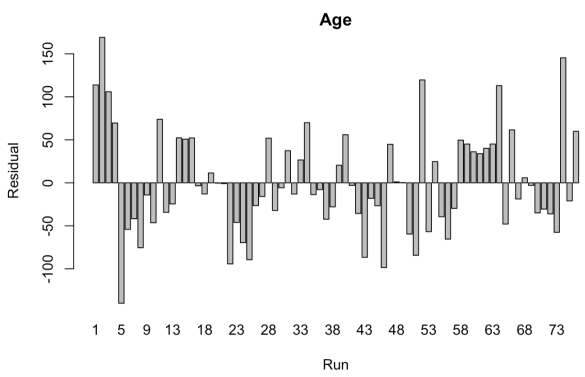
(a) $Year$ Residual Graph.



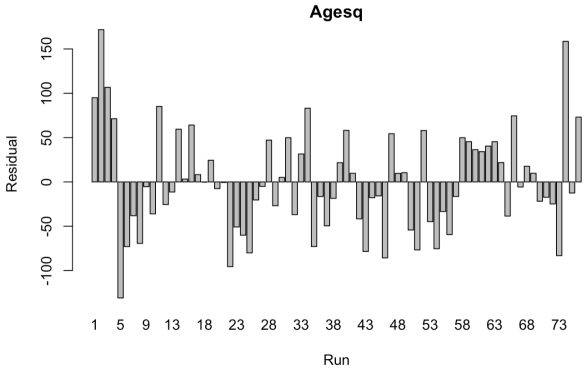
(b) Age Residual Graph.

With Age standardized using the year 1970, the distribution of ages appeared to be close to symmetrical with about three outliers. The linear model with Age seemed to be appropriate and the variance relatively constant. The error terms for Age, however, looked almost "cyclical" meaning they might not be independent. When we looked at the transformed variable Age^2, on the other hand, it looked like it had a more appropriate linear model and more constant variance.

As seen below, when the error terms of Age and Age^2 were plotted against i, the error terms of Age^2 displayed more of a horizontal band (not considering the first few and last few i's) and they were normally distributed by the looks of the normal Q-Q plot, which we will discuss further later. In the graphs of the residuals plotted against X_i, it is interesting to note that Age^2 condensed all the points of the variable Age. The distribution of points for Age^2 is skewed to the left meaning



(a) *Age* Residual vs. Run Graph.



(b) Age^2 Residual vs. Run Graph.

more of the homes that data was collected from were built around 1970.

The distribution for *Garage* was very skewed to the right. *Garage* demonstrated that its linear model was appropriate and error terms had constant variance, however, error terms were not independent and not normally distributed. *Size* and *Lot* also had a slight positive correlation with *Price*, which makes sense given that bigger houses and larger plot sizes tend to be more expensive. For the most part, the linear models of both seemed appropriate, variances were constant, and error terms were independent. The error terms were not normally distributed, though, for both *Size* and especially *Lot*.

When comparing the *Price* of houses that are currently *Active* on the market with the *Price* of those that have already been sold, there was a clear difference in the range. The prices of the houses that had already been sold took on a much larger range of values than those that were currently on the market; however, the median price of an *Active* home was slightly higher than that of an already sold home.

When looking at the elementary schools, the homes within the *Edison* and *Harris* elementary school districts assumed significantly higher prices when compared with the rest of the data. Homes within the *Parker* elementary school district had a range of prices on the lower end of the spectrum, minus one outlier. There were only six homes in the data set located in the *Crest* elementary school district, and only three in the *Adams* district, which did not really give us sufficient information to speak on how these two districts might contribute to *Price*.

2.6 Correlations Between Quantitative Variables

Since *Year* and *Age* were transformations of each other, their correlation, not surprisingly, was 1. Besides this pair, there were not any that had a degree of correlation raising concern. *Garage* was found to have a positive correlation with *Age* and a slight negative correlation with Age^2. Similarly, *Size* had a higher correlation with *Age* than Age^2, although it was very little for both. *Bath* and *Bed* also had little correlation along with *Size* and *Garage*.

3 Preliminary Models

Once we had done initial data exploration as well as outlier and residual analysis, we used multiple criteria such as AICp, BIC, and adjusted R-squared to determine a preliminary model. We excluded Age^2 and $BathBed$ when conducting our initial analysis, since those were modified and interaction variables, as well as ID and the categorical variables that had indicator counterparts. Once we decided to use Age, $Year$ was also unnecessary. Ultimately, we selected from the following variables: $Size$, $Bath$, Bed, Age, $Garage$, Lot, and the five indicators for elementary schools, $Edison$, $Harris$, $Parker$, $Adams$, and $Crest$. We used forward stepwise regression for all three criteria, but interestingly enough we got the same preliminary model for all three:

$$\begin{aligned} E(Y) &= 62.336 + 18.674(Garage) + 82.059(Edison) + 59.962(Harris) \\ &\quad +10.877(Lot) + 29.305(Active) + 59.487(Size) \end{aligned} \tag{1}$$

None of the three preliminary models actually included the variables Age, $Bath$, or Bed, which was interesting since we believed (based on prior knowledge) that these factors would have some sort of impact on the price of a home. In order to know more, however, we would have to turn to more in-depth analysis on each of those individual variables. The adjusted R-squared value of this preliminary model was 0.4502, which is a good starting point, but not ideal. Analysis of the residuals is shown in Figure 2.

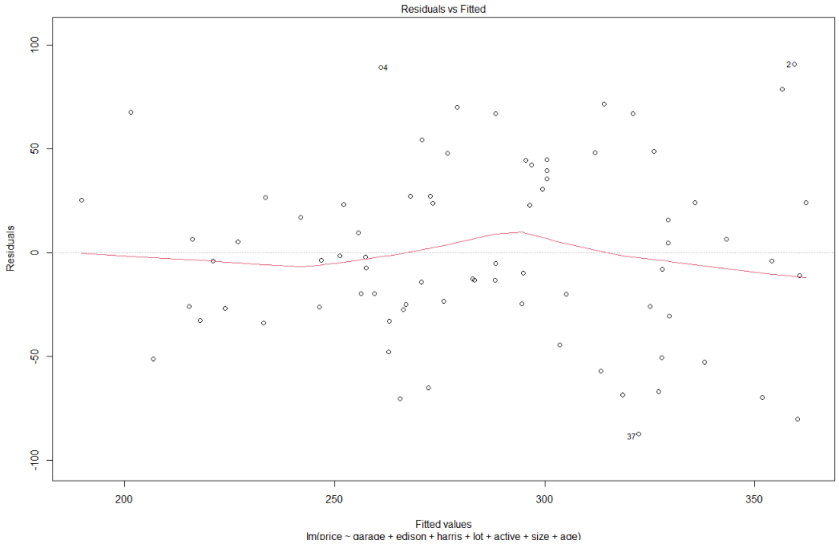


Figure 2: Residual analysis for the preliminary model in (1)

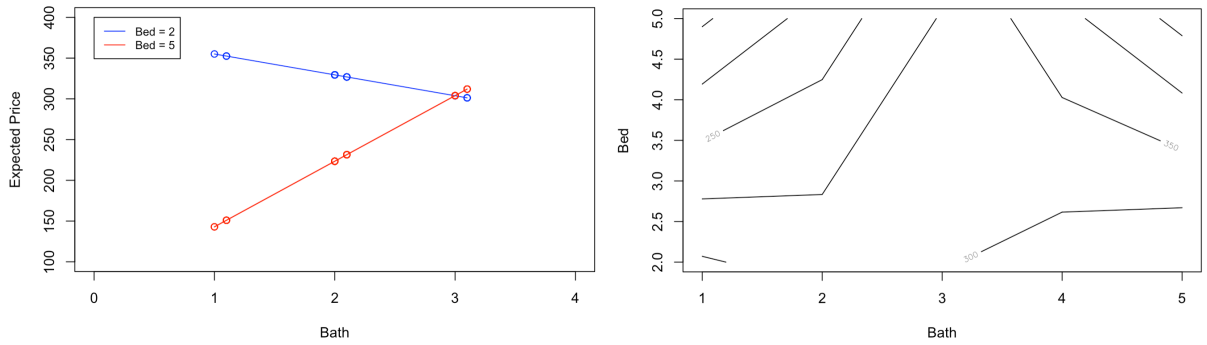
From the plot, we can see that the there are no clear patterns in the residuals, but the low adjusted R-squared value leads us to believe we can do better.

4 Procedures to Choose Variables

Once we had our preliminary model, we decided to explore the remaining variables in order to determine whether or not to include them in our model.

4.1 Interaction Terms

When we fit the additive model, adding in the two variables *Bath* and *Bed* to our preliminary model, the results were consistent with our previous findings. The regression coefficient of *Bath* was positive and the regression coefficient of *Bed* was negative.



(a) Conditional Effects Plot.

(b) Contour Curves.

It is clear from the conditional effects plot of the response function against *Bath* when *Bed* is equal to 2 and 5 (on the left) and the set of plotted contour curves for the response surface (on the right) that there is an interference interaction between *Bath* and *Bed*. To test the interaction between *Bath* and *Bed*, we first checked the correlation. Since the correlation, calculated as 0.147, was low, there was no need to do any centering.

The plot of the residuals of the additive model against the interaction term *BathBed* can be referred to in subsection 2.5. Since the pattern seemed to warrant including the interaction term, we refitted the model including *BathBed*. We then conducted an F-test testing the regression coefficient of the interaction term in this fitted model.

At a 0.05 level of significance, there was sufficient evidence to reject the null hypothesis and conclude that the regression coefficient of the interaction term was not equal to 0 further supporting it would be helpful to include the interaction term in the model. In the process of finalizing our model, we kept in mind that we could only include this term if we decided to include both *Bath* and *Bed*. Our preliminary model did not select the two variables *Bath* or *Bed*. It was interesting to us, though, that the dataset had included the interaction term for us.

4.2 Higher Order Terms

As discussed earlier, in our preliminary exploration of the variables *Age* and Age^2 when residual analysis was done, it was hypothesized that Age^2 may be a variable worth considering that may

better fit the data than the variable Age itself. As previously mentioned, looking at the normal Q-Q plots shown below, we can see that the data points form a straighter line in the plot of Age^2 compared to the plot of Age. This is why we believed Age^2, a transformation of the variable Age, to be an important variable to consider and possibly include when fitting our final linear regression model.

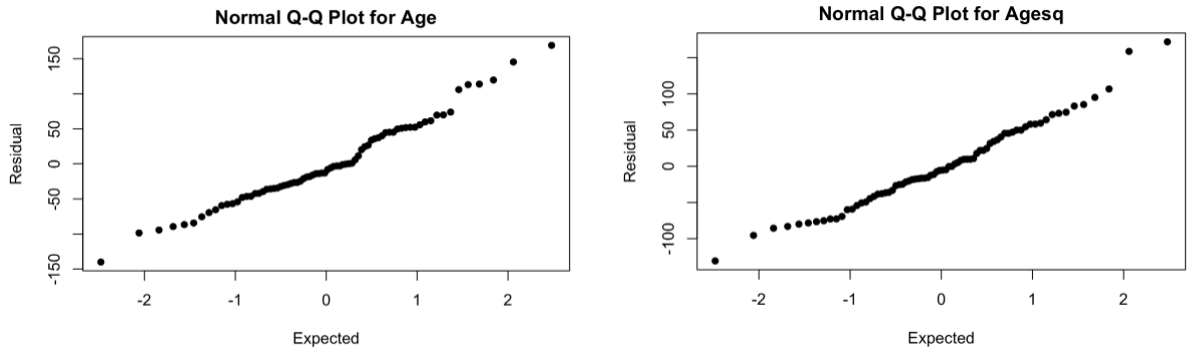


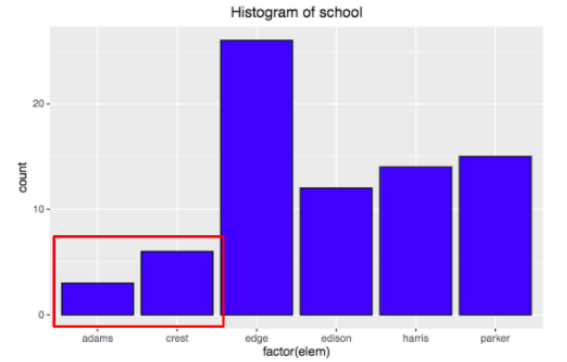
Figure 3: Normal Q-Q plots for Age and Age^2

Since the variable Age acts as a centering variable for the variable $Year$, and the variable Age^2 is its quadratic term, we knew that if we decided to include Age^2 we would have to include the variable Age also. To determine if including Age^2 in our model would be beneficial, we fit the quadratic model with Age and Age^2 and conducted an F-test testing the regression coefficient of the quadratic term Age^2.

At a 0.05 level of significance, there was sufficient evidence to reject the null hypothesis concluding that the regression coefficient of Age^2 was not equal to 0, therefore, we could accept the quadratic model. The variable Age was not selected in our preliminary model, but we have reason to believe that the variable Age^2 makes the variable Age more effective.

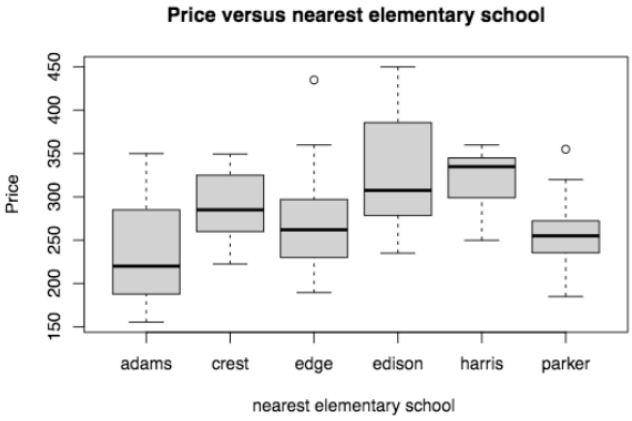
4.3 Elementary School Selection

$Price$ tended to be higher when the house was close to $Edison$ and $Harris$, lower when it was close to $Adams$, and broadly similar in a moderate range when it was near the other three schools as shown below in the box plot. $Parker$ and $Adams$ school districts had negative coefficients with home price, which we did not want to include in our model without significantly worsening its fit. Therefore, we concluded to move forward with choosing only $Edison$ and $Harris$ elementary schools indicators in our final model as was chosen by our preliminary model.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   269.76      10.80  24.969  < 2e-16 ***
edison         57.34      19.23   2.983  0.00393 **
harris         49.35      18.26   2.702  0.00864 **
parker        -12.31      17.86  -0.689  0.49295
adams         -27.92      33.59  -0.831  0.40862
crest          18.06      24.95   0.724  0.47160
```

(a) Histogram of Elementary School

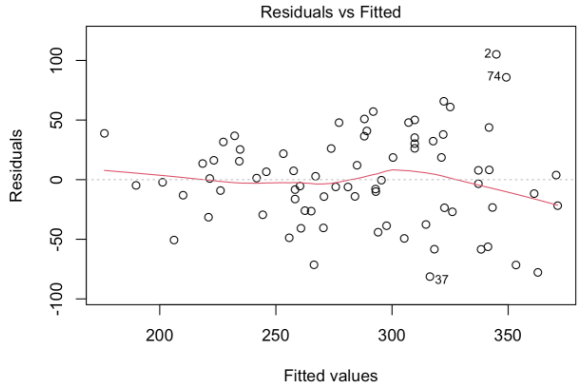


(b) Boxplot of Elementary School

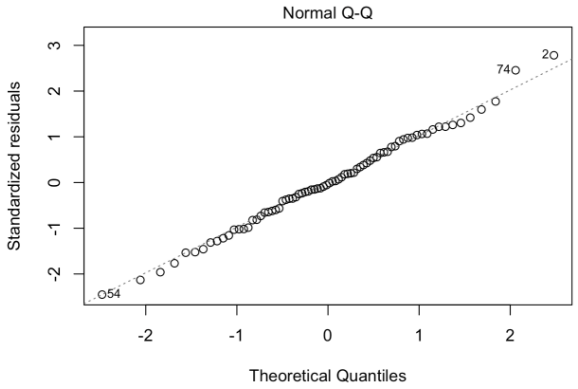
5 Final Model

Once we completed all of our analysis, our final model ended up including eleven variables, a jump from the six we included in our preliminary model. The variables incorporated in our final model were:

$$\begin{aligned} E(Y) &= 332.48 + 13.12(Garage) + 67.06(Edison) + 47.27(Harris) + 9.92(Lot) + \\ &\quad 27.42(Active) + 56.72(Size) + 3.30(Age) + 1.64(Age^2) - 98.15(Bath) \\ &\quad -78.91(Bed) + 30.39(BathBed) \end{aligned} \tag{2}$$



(a) Residual analysis for the final model in (2)



(b) Normal Q-Q plot for the final model in (2)

As you can see from our residual analysis plots, there is once again no clear pattern established by the residuals when plotted against the fitted values, and the Q-Q plot looks good as well. Our

R-squared and adjusted R-squared values increased when going from our preliminary model to our final model. Our R-squared started out as 0.49 and ended as 0.58, and our adjusted R-squared started out as 0.45 and ended as 0.51. This shows us that including the five extra variables that we ended up incorporating into our model helped us reduce the unexplained variation in price by almost ten percent, and almost seven percent adjusted.

6 Conclusions

Overall, our model was not able to predict the variation in home price as well as we would have hoped. With an adjusted R-squared of 0.51, we were only able to capture about 51% of the variation using our model. However, given the nature of the data and the market we were trying to model, we believe these results to be quite solid.

One interesting result of our model was the negative coefficients that we ended up getting for the variables *Bath* and *Bed*. These negative coefficients imply a negative impact on price for every additional bed/bathroom, which would seem counter-intuitive. One possibility our group came up with was that the number of beds/bathrooms would decrease as you got closer to a more densely populated area such as a city, and it would increase as you got to more rural locations. Rural locations tend to have lower prices than cities, thus explaining the negative relationship between price and bed/bathroom.

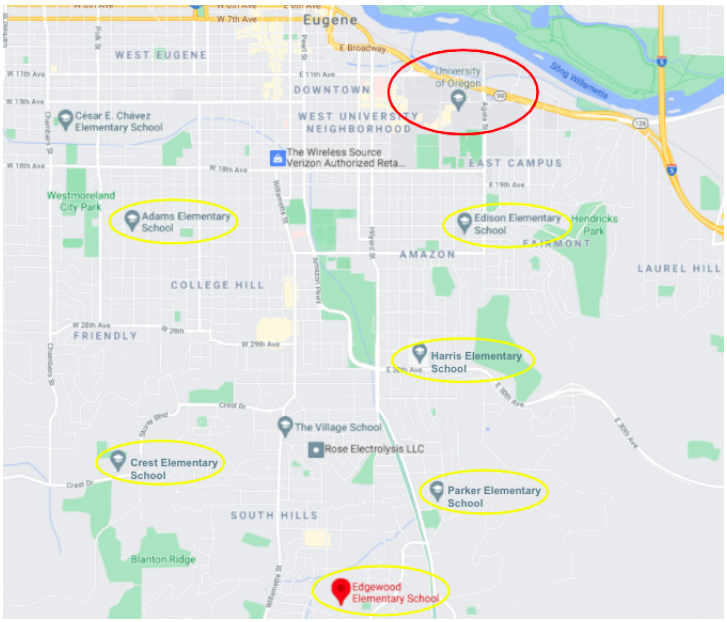


Figure 4: Map of the local area in Eugene, Oregon (captured in Google Map)

Another interesting discussion point is the inclusion of various elementary schools. We ultimately only decided to keep two of the elementary schools in our model and not the others. While this could definitely be attributed to the number of homes for each one, there may be other geo-

graphic factors at play here. From our map above, it seems like both Edison and Harris are close to city parks, and close to the local college university as well. These factors were not included in the data, and thus may have had some effect on price that we were not able to capture within our model. Overall, we would have liked to have more data and see more studies in order to corroborate the results of our model.

7 Future Steps

Given this set of predictor variables, we found that the best model we could come up with did not do an amazing job of capturing the full variation in home price. However, due to time constraints, we were only able to do full outlier analysis towards the end, leading to a few outliers in X, but none in Y. This analysis could definitely be used to improve our model had we had more time. Some future steps could include investigating further the effect of the closest elementary school, as well as other potential geographic factors not included in the model currently. There were also many assumptions made for the data, such as the weighting of half bathrooms as only 0.1 of a full bathroom. This assumption could be tested further in order to determine if it is truly accurate or not. Additionally, just having more data from not only Oregon, but also other cities and states would help corroborate this model and give more insight into factors that affect home price.

References

[Pal84] Raymond B. Palmquist. Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics*, 66(3):394, Aug 1984.

[PCMC17] Victoria A Perk, Martin Catalá, Maximillian Mantius, and Katrina Corcoran. 2017.

[Yua19] Lishun Yuan. A regression model of single house price in la constructing a predicted model for house prices, 2019.