NOVEMBER 15, 2018

Examining potential factors in predicting Ames housing sale price

MATH-299 FINAL PROJECT

TAM NGUYEN

Introduction

The United States housing bubble is one of the sectors that was severely affected by the 2007-2009 recession in the US. The extreme rise in subprime mortgage delinquencies and foreclosures and the resulting decline of securities lead to a financial crisis for the United States economy (North Carolina Department of Statistics 2016). Over half of the US states were affected by the declination of housing prices which reached its largest drop in 2008 (Mantell). The unemployment rate then reached its peak of 10.2 percent in October 2009, after the recession had ended. According to Mark Zandi, chief economist of Moody's Economy.com, housing prices declines of 10-15 percent are enough to get significant credit problem and to eliminate the homeowner's equity (Bernasek). Ultimately, having a model to determine which is the group of significant factors such as the number of rooms, the lot size area, and the house's location in predicting housing price is a big contribution to avoid future recession and a potential financial crisis.

The purpose of this project is to build a regression model to predict the housing sale prices in Ames, Iowa. I will analyze the relationship between the housing market and a large number of variables in home value assessments from a data science view, where numbers will mainly explain and mark any significant correlations. The Ames Housing data set conducted by Dean De Cock for data science education from 2006 to 2010 was utilized in the analysis. With an easily understood predictor variable of home sale prices, this data set provides an opportunity to explore and conduct multiple regression models from simple models to the more complex ones. According to the leaderboard of the competition House Prices: Advanced Regression Techniques from Kaggle using the same data set, there are many advanced regression and machine learning models that have been built on this data set with an RMSE (square root of mean square error) up to 0.08021. However, in my project, I focus on using multiple linear regression model to predict the final housing sale price. I hypothesize that factors related to housing location and housing assessment are significant in my future model.

Materials & Methods

Data from the 2006-2010 Ames Housing data set, conducted by Dean De Cook, was utilized in the analysis. The data set is originally recorded by the Ames City Assessor's Office for the city's assessment process purpose and was cleaned and decoded by De Cock. It describes the sale of individual residential property in Ames, Iowa from 2006 to 2010 (De Cock). The data set contains 2930 observations and a large number of variables (23 nominal, 23 ordinals, 14 discrete, and 20 continuous) involved in home value assessments. Due to a large number of observations compared to a multiple linear regression model, the data set was divided into training and testing data set with 50/50 ratio (Kaggle.com). With the variance in the number of variables, this data set is good for education purpose and also can explain some of the home buyer habits. However, it cannot be a representative of the United States housing market due to the lack of randomness in the process of collecting data set.

In general, the 20 continuous variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square footage found on most common home listings, other more specific variables are quantified in the data set. Area measurements on the basement, main living area, and even porches are broken down into individual categories based on quality and type. The 14 discrete variables typically quantify the number of items occurring within the house. Most are specifically focused on the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Additionally, the garage capacity and construction/remodeling dates are also recorded. There is a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being “Street” (gravel or paved) and the largest being “Neighborhood” (physical locations within the Ames city limits) (De Cock). The nominal variables typically identify various types of dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property. For more details on the unit and level of each factor, I attached the text file with more description as an appendix to this paper.

The primary focus of this project was to predict the housing sale price in Ames (IA) and to determine the best group of explanatory variables by implementing multiple linear regression model. According to the data description provided (De Cock), many entries are labeled as “NA” to indicate absence or null value for the majority of the predictor variables. There are some exceptions to this explanation in the case of linear feet of street connected to property, recorded as LotFrontage, where there is no explicit mention of NA; Masonry veneer area in square feet, recorded as MasVnrType, where an explicit None level is defined; Electrical system, recorded as Electrical, where no explicit mention of NA is provided. It can be explained that for those without definition, we can understand that there is no appearance of that factor in the house. For example, if there is no (N/A) masonry veneer type, there is no masonry veneer in that house. This indicates that the majority of NAs in the data set are meaningful and may be replaced by reasonable choices. Focusing on the simplicity of a good model, I deleted high missing factors (more than 80% of the total amount of observations), low variances and multicollinearity factors by Max Kuhn's method. Additionally, according to a survey of 2000 Americans from RootMetrics, location, amenities like shops, parks and restaurants, school districts, mobile service, and hospitals are top five criteria for homeowner choosing a new house. I also intuitively removed factors that are not directly related to homeowner's preferences such as month sold or year garage built. De Cock notes that there are five abnormal observations, and three of them are true outliers (partial sales that likely don't represent actual market values) and two of them are simply unusual sales (very large house priced relatively appropriately). Hence, I deleted those observations that have great living area greater than 4000 square feet as he recommended.

After removing factors based on the location and home assessment criteria, I have 11 variables in my analysis with 1456 observations, which include neighborhoods (Neighborhoods), lot area (LotSize), overall condition (OverCondition), basement finished area in square feet (BsmtFinSF1), basement unfinished area in square feet (BsmtUnFinSF), 1st floor square feet (`1stFlrSF`), 2nd floor square feet (`2ndFlrSF`), number of full bath (FullBath), garage area (GarageArea), basement exposure (referring to whether a house has walkout or garden level walls - BsmtExposure_new), and fireplace (Fireplace). I changed the type of neighborhoods from categorical factor to indicator variable. For fields like the fireplace and basement exposure, I recoded them as binary values. One represents a house that has at least one fireplace, and zero represents a house that doesn't have a fireplace. Similarity, if a basement of a house has walking or garden level walls, I recoded them as one; otherwise, I recoded it as zero.

Two multiple linear regression models were created, with one model exploring the relationship between all home assessment variables without the neighborhood factor and the housing sale price, and the other model depicting the association between all home assessment variables and adding the house's location to predict the housing sale price. The Variance Inflating Factor (VIF) was used to examine the relationship between explanatory variables in each model. ANOVA table was utilized to test the effectiveness of the model.

Results

Of the 1456 housing sale price observations over the four-year period, the maximum sale price is $625,000 and the minimum sale price is $34,900. The average of housing sale price in the data set is $180,151. 1.78% of the observations have sale price more than $500,000. According to the histogram below, the variance of Sale Price is skewed with a long tail to the right. In my analysis, I tried to predict the natural log of housing sale price, which gives me a more normal distribution. Table 1 shows the summary statistics for all variables that are included in the model, except for categorical Neighborhood dues to its complicatedness.

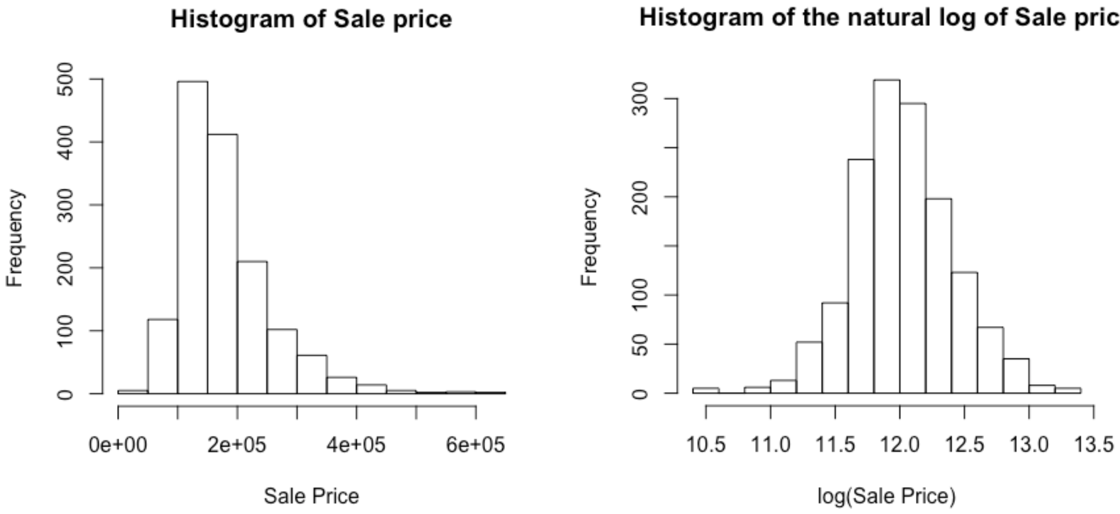


Figure 1: Histogram of housing sale price and its natural log

Table 1: Summary statistics variables with min, max, median, mean, standard deviation, and number of missing value

	min	median	max	mean	sd	missing	n
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
Lot size area (sq/ft)	1300	9468.5	215245	10448.8	9860.763449	0	1456
Overall Condition (scale 1-10)	1	5	9	5.57624	1.1139656	0	1456
Basement finished area (sq/ft)	0	381	2188	436.991	430.2550517	0	1456
Basement unfinished area (sq/ft)	0	477.5	2336	566.99	442.1971819	0	1456
1st floor area (sq/ft)	334	1086	3228	1157.11	369.3073305	0	1456
2nd floor area (sq/ft)	0	0	1818	343.533	431.5289149	0	1456
# of Full Bath	0	2	3	1.56181	0.5476829	0	1456
Garage Area (sq/ft)	0	478.5	1390	471.569	211.986421	0	1456
Basement exposure (Binary)	0	0	1	0.31937	0.466392	0	1456
Fire place (Binary)	0	1	1	0.5261	0.4994899	0	1456

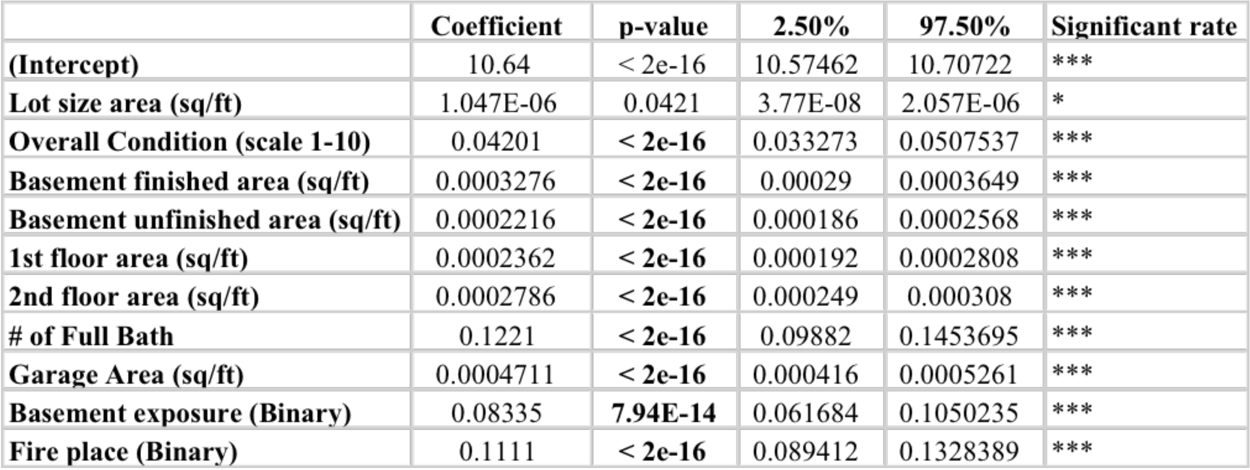
Model 1: Predicting housing sale price without neighborhood indicator (10 variables)

The first multiple linear regression model predicts the natural log of housing sale price based on housing assessment variables such as lot size area in square feet, the rate of the overall condition of the house in the scale of 10, the number of full bath, basement finished and the unfinished area in square feet, and garage area in square feet. As indicators, I included whether that house has a fireplace and whether that house's basement has walking or garden level walls. Coefficients, p-values and confidence intervals for all variables included in model 1 are displayed in table 2.

According to the normal QQ plot, there is a little skew in the right-tail. Otherwise, the conditions for the inference were met, and the variance of the residuals was constant. This model explained approximately 78.68% of the variability in the data, with an adjusted R^2 of 0.7868, F-stat equals 538 on 10 and 1445 DF, and the ρ-value (< 2.2e-16) is approximately 0, indicating that the model was significant.

In this model, all factors except lot size area have a p-value of approximately 0, showing the significant relationships between the natural log of the sale price and all the variables. The least significant variable in the model is the lot size area in square feet with p-value equaling 0.04. Otherwise, the group of factors representing the area of the house such as garage area, basement finished area, and first and second-floor area is significant to the response variable. The group of binary factor is also statistically significant. Additionally, the overall condition factor which rates the overall condition of the house is also significant. Factors in the model don't show a multicollinearity relationship to each other.

*Table 2: Coefficient, CIs, and p-value summary of the first regression model. *** Coefficients were moderately different from 0 at α = 0.05.CI for the confidence interval.*



	Coefficient	**p-value**	**2.50%**	**97.50%**	**Significant rate**
(Intercept)	10.64	< 2e-16	10.57462	10.70722	***
Lot size area (sq/ft)	1.047E-06	0.0421	3.77E-08	2.057E-06	*
Overall Condition (scale 1-10)	0.04201	**< 2e-16**	0.033273	0.0507537	***
Basement finished area (sq/ft)	0.0003276	**< 2e-16**	0.00029	0.0003649	***
Basement unfinished area (sq/ft)	0.0002216	**< 2e-16**	0.000186	0.0002568	***
1st floor area (sq/ft)	0.0002362	**< 2e-16**	0.000192	0.0002808	***
2nd floor area (sq/ft)	0.0002786	**< 2e-16**	0.000249	0.000308	***
# of Full Bath	0.1221	**< 2e-16**	0.09882	0.1453695	***
Garage Area (sq/ft)	0.0004711	**< 2e-16**	0.000416	0.0005261	***
Basement exposure (Binary)	0.08335	**7.94E-14**	0.061684	0.1050235	***
Fire place (Binary)	0.1111	**< 2e-16**	0.089412	0.1328389	***

While holding all other variables constant, model 1 95% confidence predicts that for every one square feet increase in the lot size area, the natural log of sale price will increase in the range from 0.00000004 to 0.000002. With every one point higher in the rate of the overall condition will affect the natural log of sale price in the range from 0.0334 to 0.0507. With every one square feet increase in first and second floor area, the model is 95% confident that the natural log of the sale price will respectively increase from 0.0002 to 0.0003. Increasing every one unit of full bathroom will raise the natural log of Sale Price from 0.0988 to 0.1453. With indicator variables, whether or not having a fireplace will increase the natural log of housing price from 0.089 to 0.132; whether or not having a walkout or garden level walls will raise the natural log of housing price from 0.061 to 0.105.

Based on the coefficient in table 2, all factors in the model have positive relationships with the natural log of housing sale price. Increasing one unit of full bath affecting the housing price more than increasing one unit of fireplace or having walkout or garden level walls in the house's basement.

Model 2: Predicting housing sale price with neighborhood indicator (11 variables)

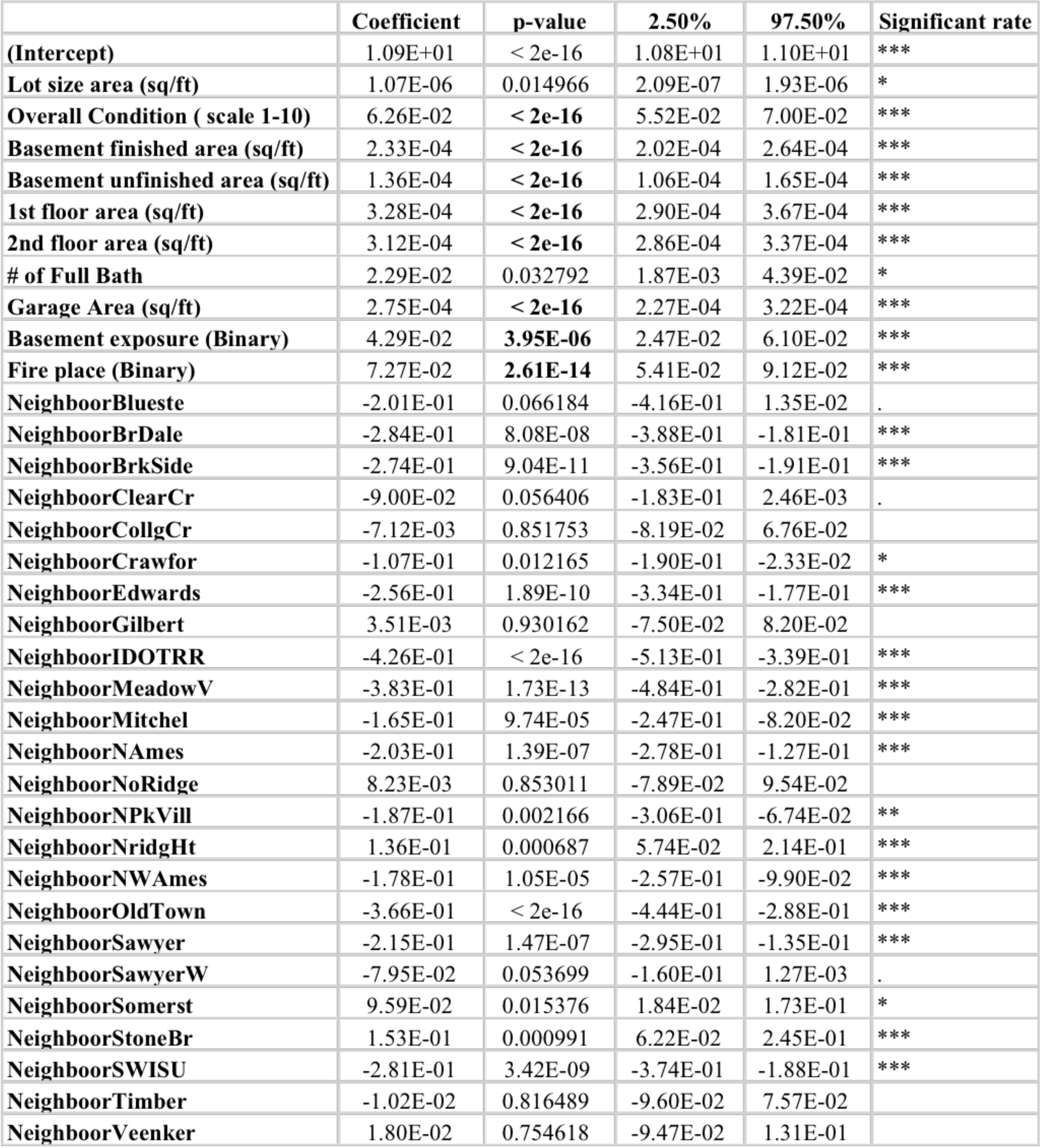
The second model predicts the natural log of housing price based on all housing assessment factors like the first model and housing location factor which is recorded as a neighborhood. Coefficients, p-values and confidence intervals for all variables included in model 2 are displayed in table 3.

According to the normal Q-Q plot, there is also a little skew in the right tail. Otherwise, the conditions for the inference were met, and the variance of the residuals was constant. The second model explained approximately 86.57% of the variability in the data, with an adjusted R^2 of 0.8657, F-stat equals 276.8 on 32 and 1421 DF, and the ρ-value (< 2.2e-16) is approximately 0, indicating that the model was significant.

All variables except the lot size area in the second model are significant with p-value approximately 0. However, it is worth noting that some neighborhoods appear to be uncorrelated with any change in the sale price. All other variables used in the model appear to have significant relationships with the natural log of sale price. Including the neighborhood factor also decreases the significance of lot size area in the model. Confidence intervals of factors in the second model were interpreted similarity with those in the first model. Factors in the model don't show a multicollinearity relationship to each other, except for neighborhood which has VIF equaling to 6.549 (see the appendix). The relationship between whether or not a house has basement exposure to the natural log of housing sale price is less significant in the second model than their relationship in the first model.

Two models both have little skew tails on the normal probability plot. However, it doesn’t affect the overall performance of the model. To fix the problem of the normality of two models, there are some models with interaction terms, some models with additional housing assessment factors such as type of dwelling or whether or not that house is remodeled, and some models with factor transformations made to examine the relationship between housing sale price as a response variable and housing assessment related variables and housing neighborhood as explanatory variables. However, the adjusted R-square was decreased and made the overall model too complicated.

*Table 3: Coefficient, CIs, and p-value summary of the second regression model. *** Coefficients were moderately different from 0 at 〈 = 0.05. CI for the confidence interval.*



	Coefficient	p-value	2.50%	97.50%	Significant rate
(Intercept)	1.09E+01	< 2e-16	1.08E+01	1.10E+01	***
Lot size area (sq/ft)	1.07E-06	0.014966	2.09E-07	1.93E-06	*
Overall Condition (scale 1-10)	6.26E-02	**< 2e-16**	5.52E-02	7.00E-02	***
Basement finished area (sq/ft)	2.33E-04	**< 2e-16**	2.02E-04	2.64E-04	***
Basement unfinished area (sq/ft)	1.36E-04	**< 2e-16**	1.06E-04	1.65E-04	***
1st floor area (sq/ft)	3.28E-04	**< 2e-16**	2.90E-04	3.67E-04	***
2nd floor area (sq/ft)	3.12E-04	**< 2e-16**	2.86E-04	3.37E-04	***
# of Full Bath	2.29E-02	0.032792	1.87E-03	4.39E-02	*
Garage Area (sq/ft)	2.75E-04	**< 2e-16**	2.27E-04	3.22E-04	***
Basement exposure (Binary)	4.29E-02	**3.95E-06**	2.47E-02	6.10E-02	***
Fire place (Binary)	7.27E-02	**2.61E-14**	5.41E-02	9.12E-02	***
NeighboorBlueste	-2.01E-01	0.066184	-4.16E-01	1.35E-02	.
NeighboorBrDale	-2.84E-01	8.08E-08	-3.88E-01	-1.81E-01	***
NeighboorBrkSide	-2.74E-01	9.04E-11	-3.56E-01	-1.91E-01	***
NeighboorClearCr	-9.00E-02	0.056406	-1.83E-01	2.46E-03	.
NeighboorCollgCr	-7.12E-03	0.851753	-8.19E-02	6.76E-02	
NeighboorCrawfor	-1.07E-01	0.012165	-1.90E-01	-2.33E-02	*
NeighboorEdwards	-2.56E-01	1.89E-10	-3.34E-01	-1.77E-01	***
NeighboorGilbert	3.51E-03	0.930162	-7.50E-02	8.20E-02	
NeighboorIDOTRR	-4.26E-01	< 2e-16	-5.13E-01	-3.39E-01	***
NeighboorMeadowV	-3.83E-01	1.73E-13	-4.84E-01	-2.82E-01	***
NeighboorMitchel	-1.65E-01	9.74E-05	-2.47E-01	-8.20E-02	***
NeighboorNAmes	-2.03E-01	1.39E-07	-2.78E-01	-1.27E-01	***
NeighboorNoRidge	8.23E-03	0.853011	-7.89E-02	9.54E-02	
NeighboorNPkVill	-1.87E-01	0.002166	-3.06E-01	-6.74E-02	**
NeighboorNridgHt	1.36E-01	0.000687	5.74E-02	2.14E-01	***
NeighboorNWAmes	-1.78E-01	1.05E-05	-2.57E-01	-9.90E-02	***
NeighboorOldTown	-3.66E-01	< 2e-16	-4.44E-01	-2.88E-01	***
NeighboorSawyer	-2.15E-01	1.47E-07	-2.95E-01	-1.35E-01	***
NeighboorSawyerW	-7.95E-02	0.053699	-1.60E-01	1.27E-03	.
NeighboorSomerst	9.59E-02	0.015376	1.84E-02	1.73E-01	*
NeighboorStoneBr	1.53E-01	0.000991	6.22E-02	2.45E-01	***
NeighboorSWISU	-2.81E-01	3.42E-09	-3.74E-01	-1.88E-01	***
NeighboorTimber	-1.02E-02	0.816489	-9.60E-02	7.57E-02	
NeighboorVeenker	1.80E-02	0.754618	-9.47E-02	1.31E-01	

Discussion

As hypothesized, location and house assessment related variables are significant in predicting the housing sale price. The model constructed with neighborhood involvement as a representative for location factor was moderately strong, with an MSE equaling 0.021. Model 2 increases the adjusted R-squared by almost 8% of the variability in the data set. However, without the location (neighborhood) explanatory variable, the first model still can explain up to 78.68% of the variability in the data, with an MSE equals 0.033. House assessment related variables are significant in predicting housing sale price. Although many advanced models have been built on this data set to predict housing price, my findings indicate that simple multiple regression models can effectively predict housing sale price with 11 over 80 explanatory variables given in the data set.

Lot size area of the house in square feet is less significant than other variables in the model. According to the model, with every one square feet change in the lot size area, the natural log of housing sale price will increase by $0.000001463. Changing only one square foot will not significantly affect the housing price, since this town is small and is not a tourist attraction town to affect the housing price by square feet. In general, lot size area still has a moderately strong relationship with housing sale price.

Housing neighbor indicator, which shows the housing physical locations within Ames city limits, has a strong relationship with housing sale price. Ames (IA) is a college town, best known as the home of Iowa State University (ISU). The student population of ISU is more than 36,000 (Iowa State University), which makes up approximately one half of the city's population in 2017, according to the US Census. With the size of 24.27 square miles, residents in Ames can easily drive across the town in approximately 12 minutes with a good highway system. Hence, besides selecting a house that is close to the college, a homeowner also considers the crime rates and other amenities such as shopping mall or golf course in the town. Those reasons can explain the negative and positive relationship between the neighborhood and the housing price. Some neighborhoods appear to be uncorrelated with any change in the sale price such as Veenker which is close to the memorial park. We can also notice that Stone Brooke is the most significant place that is correlated to the housing price in Ames. This is also explainable since Stone Brooke is close to Ames Shopping mall, according to StoneBrooke.org.

The VIF of neighborhood which is greater than 5 is a concern in the second model. It means that neighborhood factor has multicollinearity relationship with other variables in the model. This seems understandable, since a good neighborhood will include a good housing condition and assessments such as public transportation and restaurants. The relationship between overall condition of the house and the neighborhood might be a good explanation for the high VIF. I still prefer the second model, since residents in Ames can look at the model and know which is a good neighborhood. Also, the VIF of neighborhood is 6.5 which is not significantly higher than 5.

Through the findings of my analysis, homebuyers do care about home amenities such as a fireplace, garage and walkout or garden level walls in the basement beside location and size of the house. Especially, having more full bath will increase the housing sale price. The rate of housing overall condition includes other amenities, for example, electrical, heating system, and kitchen condition. That's why I did not include many indicators in my data set.

I also examined the relationship between whether or not remodeling the house affected the housing sale price by recoding that factor as a binary variable. However, it is not significant with a high p-value (~0.099). This seems explainable, since homeowners are less likely to care about house reconstruction history and more likely to focus on the present condition of the house.

Comparing to other advanced models, the accuracy of the second model in our analysis is significant. Our model has a RMSE result of 0.145 which is moderately higher than the RMSE of the highest model in the leaderboard of Kaggle which is 0.08021. Only by decoding explanatory variables and doing log transformation for housing sale price does our model becomes statistically significant. Also, our model has a significant constant variance in the residuals plot. The only concern is the normality of both models since the normal QQ plot has a little right skew.

Given the shortcomings of this analysis, there are many possible future directions. As mentioned at the beginning of the analysis, predicting housing sale price is extremely important since it could be a preliminary for any financial crisis. The value of housing is the main contributor to real estate, one of the integral roles in the economics of the United States. According to Amadeo of the Balance.com, real estate construction contributed $1.34 trillion to the nation's economic output, which takes up to 7% percent of the U.S gross domestic product in 2017 (GDP). Predicting housing price then is crucial to the economy of the United States. To test the accuracy of those models and homeowner preferences, we should try to predict housing sale price in other places. This data set is not representative of the housing market in the United States; however, it is still able to summarize homebuyer preferences. Additionally, with another town that has the same construction like Ames, my model can be a start for exploring the relationship between housing sale price and housing assessment factors. Other control variables should be included in the future model to raise the accuracy of the predicted housing price. More advanced models can also be used, since this data set has a large number of observations with many control variables. However, simplicity should be considered as the priority.

Bibliography

Amadeo, Kimberly. “Why Buying a Home Helps Build the Nation.” The Balance. Accessed November 21, 2018. https://www.thebalance.com/how-does-real-estate-affect-the-u-s-economy-3306018.

Bernasek, Anna. “When Does a Housing Slump Become a Bust?” *The New York Times*, June 17, 2007, sec. Business Day. https://www.nytimes.com/2007/06/17/business/yourmoney/17view.html.

De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” *Journal of Statistics Education* 19, no. 3 (November 2011). https://doi.org/10.1080/10691898.2011.11889627.

Home Guru. “Information About Factors That Determine Property Prices - HomeGuru.” Accessed October 18, 2018. http://www.homeguru.com.au/house-prices/.

Iowa State University. “Fall Enrollment Is Second Highest on Record - Inside Iowa State for Faculty and Staff - Iowa State University.” Accessed November 19, 2018. https://www.inside.iastate.edu/article/2017/09/07/enroll.

Kaggle.com. “House Prices: Advanced Regression Techniques.” Accessed November 21, 2018. https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Kuhn, Max, and Kjell Johnson. “An Introduction to Feature Selection.” In *Applied Predictive Modeling*, edited by Max Kuhn and Kjell Johnson, 487–519. New York, NY: Springer New York, 2013. https://doi.org/10.1007/978-1-4614-6849-3_19.

Macdonald, Alan F. “What Does Above Grade Mean? | Real Estate Definition.” Accessed November 15, 2018. https://www.gimme-shelter.com/above-grade-50066/.

Mantell, Ruth. “Home Prices off Record 18% in Past Year, Case-Shiller Says.” MarketWatch. Accessed November 15, 2018. https://www.marketwatch.com/story/home-prices-off-record-18-in-past-year-case-shiller-says.

StoneBrooke.org. “- Home.” Accessed December 7, 2018. http://www.stonebrooke.org/.

Time.com. “Cell Reception Is More Important To Home Buyers Than Schools.” Money. Accessed November 17, 2018. http://time.com/money/3904761/buy-home-good-cell-mobile-reception/.

U.S Census Beureau. “U.S. Census Bureau QuickFacts: Ames City, Iowa.” Accessed November 19, 2018. https://www.census.gov/quickfacts/fact/table/amescityiowa/PST045217.