***Introduction***

The United States housing bubble is one of the sectors that was severely affected by the 2007-2009 recession in the US. The extremely rise in subprime mortgage delinquencies and foreclosures, and the resulting decline of securities leads to a financial crisis over the United States economy (North Carolina Department of Statistics 2016). Over half of the US states was affected by the declination of housing price and reached its largest drop in 2008(cite). 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). Housing price, hence, is an important sector in predicting and presenting the economic growth of a country, especially the United States. Ultimately, having a model to determine which is the group of significant factors 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 categorical variables in home value assessments from a data science view, where numbers will mainly explain and mark any significant correlations. The Ames Housing dataset 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 dataset provides an opportunity to explore and conduct multiples 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 dataset, there are many advanced regression and machine learning models that have been built on this dataset with an RMSE up to 0.08021. However, in my project, I focus on multiples 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 the Ames Housing dataset, conducted by Dean De Cook, was utilized in analysis. The dataset is originally recorded by the Ames City Assessor’s Office for the city’s assessment process purpose and was cleaned and decoded by Dean. It describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of categorical variables (23 nominal, 23 ordinals, 14 discrete, and 20 continuous) involved in home value assessments. Due to the large amount of observations compared to a multiple linear regression model, the dataset was divided into training and testing dataset with 50/50 ratio. With the variance in the number of variables, this dataset 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 dataset.

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). 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 according to this paper.

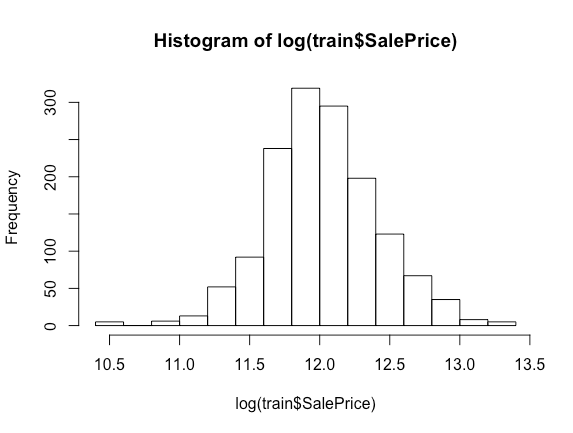
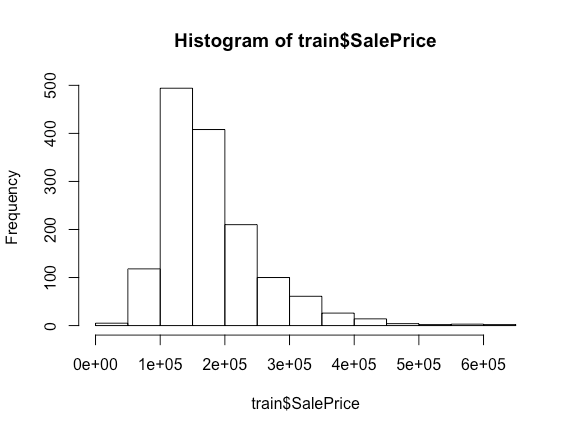
The primary focus of this project was to predict the housing sale price in Armes (IW) and to determine the best group of explanatory variables by implementing multiple linear regression model. According to the data description provided (Cock 2011), many entries are labeled as “NA” to indicate absence or null value for the majority of the predictor variables. There are some exception 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 restaurant, school districts, mobile service and hospitals are top 5 criteria for homeowner choosing a new house. I also intuitively removed factors that are not related to homeowner’s preferences. In terms of outliners, I deleted observations that have Great Living Area greater than 4000 square feet, according to Dean’s note.

After removing factors based on the location and home assessment criteria, I have 14 variables in my model, which include neighborhoods (Neighborhoods), the types of dwelling involved in the sale (MSSubClass), lot area (LotSize), overall condition (OverCondition), basement type 1 finished area in square feet (BsmtFinSF1), basement unfinished area in square feet(BsmtUnFinSF), 1st floor square feet (`1stFlrSF`), 2nd floor square feet (`2ndFlrSF`), full bath, garage area, basement exposure (referring to walkout or garden level walls), fireplace, and year remodel. I changed the type of MSSubClass and neighborhoods from numerical and categorical factor to indicator variables. For fields like fireplace, full bath, basement exposure, and house remodeled, I recoded them as binary values.

Two multiples linear regression models were created, with one model exploring the relationship between all home assessment variables without neighborhood factor and the housing sale price, and the other model depicting the association between all home assessment variables with house’s location to the housing sale price, while holding all variables to constant. 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 $625000 and the minimum sale price is $34900. The average of housing sale price in the dataset is 180 151$. x% of the observations has sale price more than 500 000$. According to the histogram below, the variance of Sale Price is skewed with a longer tail to the right. Of the 80 variables in the data set, nineteen (19) had missing values and eighteen (18) had missing values above five (5) percent.



Model 1: Predicting housing sale price without neighborhood indicator (12 variables)

The first multiples 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 overall condition of the house in scale of 10, Basement finished and unfinished area in square feet, and garage area. As indicators, I included whether or not that house has either full bath or half bath, whether that house has a fireplace, and whether that house has a walking garden (recoded as BsmtExposure\_new). Coefficients, p-values and confidence interval for all variables included in the model 1 are displayed in the table ().

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Coefficient** | **p-value** | **2.5 %** | **97.5 %** | **Significant rate** |
| **(Intercept)** | 10.73 | < 2e-16 | 1066.5930% | 1079.8060% | \*\*\* |
| **MSSubClass30** | -0.2279 | < 2e-16 | -27.1935% | -18.3809% | \*\*\* |
| **MSSubClass40** | -0.08792 | 0.296426 | -25.3027% | 7.7194% |  |
| **MSSubClass45** | -0.08901 | 0.071617 | -18.5861% | 0.7837% | . |
| **MSSubClass50** | -0.1233 | 9.97E-09 | -16.5279% | -8.1360% | \*\*\* |
| **MSSubClass60** | 0.111 | 2.94E-05 | 5.9046% | 16.2919% | \*\*\* |
| **MSSubClass70** | -0.09602 | 0.002004 | -15.6868% | -3.5169% | \*\* |
| **MSSubClass75** | -0.1532 | 0.001718 | -24.8912% | -5.7534% | \*\* |
| **MSSubClass80** | -0.00993 | 0.687102 | -5.8282% | 3.8422% |  |
| **MSSubClass85** | -0.04167 | 0.291104 | -11.9057% | 3.5726% |  |
| **MSSubClass90** | -0.1793 | 7.95E-12 | -23.0289% | -12.8300% | \*\*\* |
| **MSSubClass120** | 0.1175 | 5.54E-09 | 7.8182% | 15.6743% | \*\*\* |
| **MSSubClass160** | 0.01597 | 0.588407 | -4.1904% | 7.3841% |  |
| **MSSubClass180** | -0.1321 | 0.017133 | -24.0707% | -2.3527% | \* |
| **MSSubClass190** | -0.2512 | 5.57E-13 | -31.8900% | -18.3492% | \*\*\* |
| **LotArea** | 0.000001734 | 0.000338 | 0.000079% | 0.000268% | \*\*\* |
| **OverallCond** | 0.04829 | < 2e-16 | 3.9947% | 5.6626% | \*\*\* |
| **BsmtFinSF1** | 0.0002536 | < 2e-16 | 0.0216% | 0.0291% | \*\*\* |
| **BsmtUnfSF** | 0.0002235 | < 2e-16 | 0.0190% | 0.0257% | \*\*\* |
| **`1stFlrSF`** | 0.0003604 | < 2e-16 | 0.0318% | 0.0403% | \*\*\* |
| **`2ndFlrSF`** | 0.0003269 | < 2e-16 | 0.0280% | 0.0374% | \*\*\* |
| **FullBath** | 0.04579 | 0.000168 | 2.1981% | 6.9593% | \*\*\* |
| **GarageArea** | 0.0003657 | < 2e-16 | 0.0314% | 0.0418% | \*\*\* |
| **BsmtExposure\_new** | 0.07375 | 9.38E-12 | 5.2701% | 9.4803% | \*\*\* |
| **Fireplace** | 0.06389 | 2.52E-09 | 4.2995% | 8.4779% | \*\*\* |

According to the residual plot, there is a long-left tail in the normal QQ plot and the normality is questionable. Otherwise, the conditions for the inference were met, the variance of the residuals was constant, and the model overall looks normal. This model explained approximately 82.25% of the variability in the data, with an adjusted R2 of 0.8225, F-stat equals 281.8 on 24 and 1431 DF, and the ρ-value (< 2.2e-16) is approximately 0, indicated that the model was significant.

In this model, all factors have a p-value of approximately 0, show the significant relationships between the natural log of sale price and all the variable. The group of least significant variables compared to the most significant ones are whether the house has full bath or not, whether the basement has walkout or garden level walls or not, and whether the house has fireplace or not. Otherwise, the group of factors represent the area of the house such as garage area, basement finished area, first and second floor area is significant to response variable. Additionally, the overall condition factor which rates the overall condition of the house is also significant.

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 0.000079% to 0.000268%. With every one point higher in the rate of overall condition will affect the natural log of sale price in range from 3.99% to 5.67%. With every one square feet increases in first and second floor area, the model 95% confident that the natural log of the Sale Price will increase from 0.03% to 0.04%. With indicator variables, whether or not having a fireplace will increase the natural log of housing price from 4.3% to 8.5%; whether or not having a walkout or garden level walls will raise the natural log of housing price from 5.3% to 9.5%; and whether or not having bath will increase the natural log of housing price from 2.2% to 6.95%.

Model 2: Predicting housing sale price with neighborhood indicator

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 recorded as neighborhood. Coefficients, p-values and confidence interval for all variables included in the model 2 are displayed in the table ().

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Coefficients** | **p-value** | **2.50%** | **97.50%** | **Significant rate** |
| **Intercept** | 10.95 | < 2e-16 | 1084.98% | 1104.81% | \*\*\* |
| **`1stFlrSF`** | 0.0003636 | < 2e-16 | 0.03% | 0.04% | \*\*\* |
| **`2ndFlrSF`** | 0.0002979 | < 2e-16 | 0.03% | 0.03% | \*\*\* |
| **BsmtExposure\_new** | 0.03527 | 0.000193 | 1.84% | 5.53% | \*\*\* |
| **BsmtFinSF1** | 0.0001981 | < 2e-16 | 0.02% | 0.02% | \*\*\* |
| **BsmtUnfSF** | 0.0001387 | < 2e-16 | 0.01% | 0.02% | \*\*\* |
| **Fireplace** | 0.05407 | 1.18E-08 | 3.77% | 7.46% | \*\*\* |
| **FullBath** | 0.04118 | 6.58E-05 | 2.01% | 6.04% | \*\*\* |
| **GarageArea** | 0.0002327 | < 2e-16 | 0.02% | 0.03% | \*\*\* |
| **LotArea** | 0.000001463 | 0.000717 | 0.000000548 | 0.000002241 | \*\* |
| **MSSubClass120** | 0.009407 | 0.638329 | -3.09% | 4.71% |  |
| **MSSubClass160** | -0.05122 | 0.100648 | -11.21% | 1.00% |  |
| **MSSubClass180** | 0.0177 | 0.751019 | -8.97% | 12.92% |  |
| **MSSubClass190** | -0.1581 | 3.92E-07 | -21.91% | -9.76% | \*\*\* |
| **MSSubClass30** | -0.1546 | 2.18E-11 | -19.64% | -10.65% | \*\*\* |
| **MSSubClass40** | -0.02944 | 0.678235 | -17.13% | 10.74% |  |
| **MSSubClass45** | -0.04646 | 0.277867 | -12.93% | 3.89% |  |
| **MSSubClass50** | -0.05274 | 0.010249 | -9.30% | -1.27% | \*\* |
| **MSSubClass60** | 0.06213 | 0.006687 | 1.38% | 10.35% | \* |
| **MSSubClass70** | -0.04542 | 0.117393 | -10.24% | 1.08% |  |
| **MSSubClass75** | -0.01564 | 0.718305 | -10.35% | 6.61% |  |
| **MSSubClass80** | 0.01092 | 0.60068 | -2.96% | 5.21% |  |
| **MSSubClass85** | -0.005088 | 0.878157 | -7.32% | 5.72% |  |
| **MSSubClass90** | -0.1243 | 2.53E-08 | -16.52% | -7.82% | \*\*\* |
| **NeighboorBlueste** | -0.133 | 0.220873 | -35.34% | 7.32% |  |
| **NeighboorBrDale** | -0.2443 | 2.39E-05 | -34.78% | -12.35% | \*\*\* |
| **NeighboorBrkSide** | -0.1965 | 1.82E-05 | -28.77% | -10.86% | \*\*\* |
| **NeighboorClearCr** | -0.09642 | 0.043789 | -18.92% | -0.17% | \* |
| **NeighboorCollgCr** | -0.03132 | 0.436054 | -10.92% | 4.84% |  |
| **NeighboorCrawfor** | -0.0399 | 0.376883 | -13.67% | 3.96% |  |
| **NeighboorEdwards** | -0.2231 | 1.82E-07 | -30.64% | -13.96% | \*\*\* |
| **NeighboorGilbert** | -0.01947 | 0.646191 | -10.39% | 6.24% |  |
| **NeighboorIDOTRR** | -0.3444 | 8.21E-13 | -43.87% | -25.18% | \*\*\* |
| **NeighboorMeadowV** | -0.3807 | 3.94E-11 | -49.66% | -27.22% | \*\*\* |
| **NeighboorMitchel** | -0.1561 | 0.000386 | -24.23% | -7.00% | \*\*\* |
| **NeighboorNAmes** | -0.2118 | 2.42E-07 | -28.97% | -12.98% | \*\*\* |
| **NeighboorNoRidge** | -0.02905 | 0.528488 | -10.72% | 7.16% |  |
| **NeighboorNPkVill** | -0.1319 | 0.028517 | -25.47% | -1.85% | \* |
| **NeighboorNridgHt** | 0.1201 | 0.003035 | 5.28% | 21.01% | \*\* |
| **NeighboorNWAmes** | -0.1917 | 7.24E-06 | -27.15% | -10.48% | \*\*\* |
| **NeighboorOldTown** | -0.2838 | 1.04E-10 | -37.13% | -20.05% | \*\*\* |
| **NeighboorSawyer** | -0.2168 | 5.61E-07 | -30.13% | -13.22% | \*\*\* |
| **NeighboorSawyerW** | -0.0944 | 0.026999 | -17.51% | -0.79% | \* |
| **NeighboorSomerst** | 0.1011 | 0.01484 | 2.10% | 18.27% | \* |
| **NeighboorStoneBr** | 0.1403 | 0.002104 | 5.41% | 23.28% | \*\* |
| **NeighboorSWISU** | -0.2057 | 4.93E-05 | -30.64% | -10.82% | \*\*\* |
| **NeighboorTimber** | -0.01878 | 0.678678 | -10.43% | 7.34% |  |
| **NeighboorVeenker** | 0.005734 | 0.918935 | -10.73% | 11.40% |  |
| **OverallCond** | 0.05997 | < 2e-16 | 5.33% | 6.83% | \*\*\* |
| **YearRebuilt** | -0.01502 | 0.098863 | -3.47% | 0.08% | . |

According to the residual plot, there is also a long-left tail in the normal QQ plot and the normality is questionable. Otherwise, the conditions for the inference were met, the variance of the residuals was constant, and the model overall looks normal. The second model explained approximately 87.64% of the variability in the data, with an adjusted R2 of 0.8764, F-stat equals 270.4 on 49 and 1406 DP, and the ρ-value (< 2.2e-16) is approximately 0, indicated that the model was significant. With an addition of the neighborhood variable, this model explained almost 6% more than the variability in the dataset compared to the model 1.

All variables in this model are significant with small p-value (~0), which shows a strong relationship between explanatory factor and the response variable. 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 sale price at the 1% level. Including the neighborhood factor also decreases the significant of lot size area in the model. Overall, while holding all other variable constant, the confidence interval of all variables in this model has pretty similar range and effect to the natural log of housing sale price.

To fix the questionable about normality of two models, there are some models with interaction terms and other transformations were made to examine the relationship between housing sale price as 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.

***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 MSE equals to 0.019. Model 2 increases the adjusted R-squared by almost 6% of the variability in the dataset, which is nearly 87 observations. However, without the location (neighborhood) explanatory, the first model still can explain up to 82.24% of the variability in the data, with a MSE equals to 0.028. House assessment related variables are significant in predicting housing sale price. Although many advanced models has been built on this dataset to predict housing price, my findings indicate that a simply multiple regression models can nicely predict housing sale price with 13 over 80 explanatory variables given in the dataset.

Lot size area of the house in square feet is significant. 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 feet 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 has a strong relationship with housing sale price, as predicted.

Housing neighbor indicator, which shows the housing physical locations within Ames city limits, has a strong relationship with housing sale price. Ames (IW) is a college town, best known as the home of Iowa State University (ISU), with leading in Agriculture, Design, Engineering, and Veterinary Medicine colleges (cite). The student population of ISU is more than 36,000 students, which makes up approximately one half of the city’s population in 2017. In 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, beside selecting house that close to college, homeowner is also consider the crime rates and other amenities such as airport or golf course in the town. Those reasons can explain for 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 close to the memorial park.

Through the findings of my analysis, homebuyers do care about home amenities such as fireplace, garage and walkout garden beside location and size of the house. The rate of housing overall condition is included other amenities, for example, electrical, heating system, kitchen condition. That’s why I did not include many indicators in my dataset.

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

Comparing to other advanced models, the accuracy of the second model in our analysis is significant. With a highest RMSE equal to 0.127, is moderately lower than RMSE of the highest model in the leaderboard of Kaggle. Only by decoding variables and doing log transformation for housing sale price, our model is statistically significant.

Given the shortcomings of this analysis, there are many possible future directions. To test the accuracy of those models and homeowner preferences, we should try to predict housing sale price in other places. This dataset is not representative of the housing market in the United States, however, it is still able to summarize homebuyer preferences. Additionally, with other town that have the same construction like Ames, my model can be a start for exploring 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. Additionally, using transformation for control variables is a room to try. More advanced models can also be used, since this dataset has a large amount of observations with many control variables. However, simplicity should be considered as the priority.