More Than a White Picket Fence:

How Home Features Influence Sale Price

Yvonne Renard, Alex Teboul, Sara Elkasevic

DePaul University

Professor Besser

DSC 324/424

November 22nd, 2019

Executive Summary

The process of buying and selling homes can be daunting. Buyers come into the process with a laundry list of desired home attributes and a target price in mind. But what features are they willing to sacrifice and how do the characteristics of a home ultimately influence its sale price? And from the seller's perspective, what is the maximum sale price you can list your property for and still attract buyers?

The common real-estate mantra "*location, location, location"* suggests that what really matters in property values is where a property is geographically located. A wealth of research exists to support this claim. For example, a 2016 report using data from over 160,000 properties listed on Zillow found an average difference in home prices between Urban and Rural areas of over $100,000 across the entire United States (Fuller, 2016). Further breakdown on location leads to more stark differences in property values. The findings can be considered fairly obvious; it's no surprise to estimate a house on the island of Manhattan or downtown Chicago would be more expensive than one in rural America on average. In our analysis, we attempt to go beyond the location mantra to better understand what other factors influence residential property values and determine the extent to which we can predict prices - based on features and not location.

In order to better understand the relationship between residential property sale prices and property characteristics, we performed an exploratory data analysis on a housing dataset from Kaggle. Specifically, our group chose the *Ames Housing dataset*, which consists of 2930 observations and 79 explanatory variables describing the different aspects of residential homes in Ames, Iowa and the price they sold for. This dataset was compiled by Dean De Cock at Truman University for educational purposes and serves as an extension to the popular Boston Housing Prices dataset. Of these variables, 38 are numeric values describing characteristics of area and amount, including sale price. Some of the home feature variables include lot area, number of bedrooms, and number of bathrooms. An additional seventeen ordinal variables rate different aspects of the house, like basement condition and heating quality ranked on a scale from poor to excellent. Before we further analyzed the data, we took multiple preprocessing steps, including imputing missing values by their median and mapping descriptive ordinal variables to a numeric 1-5 scale.

To understand the underlying structure of our dataset, we performed a principal component analysis and a common factor analysis. From these analyses, we determined factors that involve overall living area and square footage explained a significant proportion of the variance in the dataset. Essentially, most of the differences between homes can be explained by home square footage and the amount of features the home has. This suggests to us that a property's sale price could also be strongly influenced by these features. From the different descriptive statistics run on our dataset and associated charts, we confirmed that there is an influence from these features on a property's sale price. We also confirmed through our principal component and factor analyses that square feet and counting variables had positive relationships

with each other. In English, we find that a house with more bedrooms also has more bathrooms and more living space on average. Larger homes also tend to have higher quality features in them, such as finished basements and remodeled kitchens.

Given that our goal was to not only understand the factors influencing home prices, but also to estimate the marginal effects a home's features on its sale price, we created models using random forest classification and linear regression methods. From the random forest classification, we were able to predict whether a home would be sold at a low, low middle, high middle, or high sale price based on home features with about 80% accuracy and high sensitivity and specificity. This means that for homes in Ames, Iowa, we can predict the sale price range about 8 times out of 10 based on a subset of home features. We are also able to determine that the top five most important features for determining sale prices were the overall quality of a property, 1st floor square footage, the year the property was built, lot area, and year remodeled from the random forest classification method. It stands to reason that a high-quality, new or recently remodeled home with a large 1st floor square footage and a big lot would sell at a high price point. To predict marginal effects on sale price more specifically, we used a backward elimination linear regression that resulted in similarly successful estimates. One key metric thats supports our linear regression model is the adjusted-R^2 value of 85.4%. This suggests that about 85% of the changes in sale price in Ames, Iowa could be explained by the 35 home features we included in the model.

Although our results suggest strong explanatory power, our methods still have some limitations. Our sample only includes data on 2,930 homes, which is a small number in the grand scheme of the real-estate market. Our data is also localized to Ames, Iowa, which is a city of about 66,000 people. We are not able to extrapolate our findings to all markets across the US or to other places in the world because we do not have sufficient data. We would argue that the methods we use in this paper could be replicated in other cities or neighborhoods within cities with a fair amount of success. Another glaring limitation is that our data consists of house sales from 2006-2010, which is the same period as the Great Recession. This significantly impacted home prices, so our models would need to be adjusted to current prices before being used to predict home values today. In summary, our main limitations include our sample size, the localization of our data, and the time period the data comes from.

Despite these limitations, our analysis uncovers the important features that influence the sale prices of homes in Ames, Iowa these features could also be important in most real-estate markets. These important features include a property's square footage, the year the home was built, if and when the home was remodeled, and the overall quality. While a homeowner looking to sell may not be able to control for square footage and size, they could prioritize remodeling if they are looking to sell at a higher price. The practice of home-flipping, in which investors purchase a home, remodel it, and sell it at a higher price, could greatly benefit from such an analysis. Our findings suggest that a basement, kitchen, and/or bathroom remodel are most critical to achieving a higher sale price.

More Than a White Picket Fence:
How Home Features Influence Sale Price

Yvonne Renard, Alex Teboul, Sara Elkasevic

DePaul University - DSC 324/424: Advanced Data Analysis

Abstract. This paper explores the relationship between home features and sale prices using a combination of exploratory and predictive techniques. More specifically, we seek to answer the question of what home features influence the sale prices of homes in Ames, Iowa. Home feature effects have long been leveraged in real-estate markets through intuition or research conducted by those seeking greater profits or better deals. To extend our understanding of this critical field, we performed a principal component analysis, common factor analysis, random forest classification, and linear regression in R. Our principal component analysis and common factor analysis uncovered key underlying structures regarding quality, quantity, and room types as the most significant features to explain our data. Through the use of random forest classification and linear regression, we determined that features that add value to a home are commonly associated with larger square footage and the perceived newness of a home, such as when it was built or remodeled. While these features were expected to be important, it is significant to note that our models were also highly predictive of sale price. Although our findings are generally consistent, it should be acknowledged that predicting sale price is generally difficult, given that our data comes from a single city in the Midwest and occurs during the last recession. Nonetheless, the important features we have extracted from the analysis can help those seeking a greater understanding of the market and how to approach home pricing.

Keywords: principal component analysis, common factor analysis, random forest classification, linear regression, home prices, real-estate

1. Introduction

The housing market has always been of interest to various sectors, ranging from mortgage lenders to home-flippers. Many people seek to investigate what attributes of a home can influence the sale price of a house. Our research focuses on data from 2006-2010 in Ames, Iowa with 79 characteristics of 2,930 residential home sales with a median home price of $174,300.. Of these 79 characteristics, 38 are numeric, 17 are ordinal, and 24 are categorical. This research investigates the relationship between a home's sale price and its features, such as square footage, overall conditions, and other characteristics that add value to a home. We sought to answer what variables influence the sale price of homes in Ames, Iowa and their magnitudes to make conclusions on which features would make a home more valuable. Such information would be useful for actors within the housing market to better sell and prepare houses, such as homeowners, investors, real estate agents, and insurance companies.

The housing market is something that can affect everybody, from people's rents to their investment allocations. Having the ability to estimate house prices can give us a benchmark of

housing affordability and housing quality. The direct relationship between affordable housing and a local economy's health has been measured in multiple studies, where rising house prices indicate a healthy or expanding economy. Because the housing market is a major macroeconomic indicator, having the ability to forecast housing prices could aid policy makers, investors, and real estate parties to set accurate and realistic expectations in the long-run.

2. Literature Review

The most common method of analysis performed on house pricing is a linear multivariate regression to predict a home's sale price. Although linear regressions are most common, some studies also include nonlinear methods, such as neural networks. In *Performance of Multiple Linear Regression and Non-Linear Neural Networks and Fuzzy Logic Techniques in Modeling House Prices* by Siti Amari and Gurudeo Anand Tularam, an analysis on homes in Bathurst, Australia investigates the various factors that drive up housing prices, despite any changes in the financial market. The researchers included several independent variables, such as land value, land area in meters, and distances to points of interest, to perform the multivariate regression. They used a step-wise approach and ran their model twice, with and without the inclusion of variables related to macroeconomic variables. After adding these social indicators the model revealed that location in regard to the city area and land value were significant. The other method, neural networks, showed to perform better than linear regression in regards to prediction accuracy. This method used the Levenberg-Marquardt technique in which the networks were trained until the mean-square error was a low value. The researchers concluded that although neural networks do have better performance, multivariate regression can be improved to show the same performance.

In the research article *Forecasting House Prices in OECD Economies* by N.Kundan Kishor and Hardik A. Marfatia, the authors attempt to forecast the movement of house prices within the unique economies of the OECD, an intergovernmental economic organization. They cite inflation, interest rates, money supply, and level of economic activity as some of the strongest indicators in housing markets. In order to better predict future home prices, the authors' analysis uses global and local factors to predict house prices across the different OECD countries. The combination forecasts can incorporate information from several predictors and estimate more accurate forecasts in changing environments to allow for changes in macroeconomic trends over time. Their data is from OECD countries on housing prices from 1975 to 2013, with a seasonally adjusted housing price index from the Federal Reserve of Dallas. With several domestic and global factors, including personal consumption expenditures and interest rates, the movements in macroeconomic trends and their effects on housing prices can predict most influence on pricing forecasts. As important as overall macroeconomic variables are, the authors find that the most significant factor in a house's price is the house's own previous selling prices.

In the journal *Property Renovations and Their Impact on House Price Index Construction,* the authors A. N. Bogin and Hardik M. Doerner state that most literature on

housing prices and housing indices potentially contain bias from an omitted renovation variable. While picking apart assumptions regularly made by other housing price predictions, the authors highlight that single-city analyses are more likely to contain systematic renovation bias. While controlling for renovation may seem like a simple task, the extent and overall quality of renovation is seldom recorded in house sales. In an attempt to capture the marginal effects of improvement types and extent of renovation, the authors combine historical data from independent real estate multiple listing services (MLS) and mortgage transactional data from the Federal Housing Finance Agency (FHFA). By introducing the renovation indicator into a model, any potential omitted variable bias that is commonly present in housing price research should diminish. The authors note that renovation bias is more likely to occur in localized studies, and use linear regressions on inner-city variations to find the average magnitudes of a home's renovation effects. As the magnitude of research's radius from a city center expands, the marginal effect of renovations vanishes. Because of this, the authors conclude that the magnitude and persistence of the renovation bias predictably varies.

The overall theme of the literature mentioned above concludes that the housing system in the United States is a complex system with multiple levels of global and local factors. Global factors, such as money supply, unemployment rates, and income levels, play significant roles in driving market dynamics and forecasting housing prices across the world. Local factors, which are the structural features of a home, play a more direct role on housing prices. Because our data set focuses solely on homes sold in Ames, Iowa, we are interested in estimating the role and magnitudes of different local effects on a property's market price. The rest of the paper is organized as follows. Section 3 includes the various methodologies used in our analysis. Section 4 covers the discussion and results of each method, and our conclusions are in section 5.

3. Methods

3.1. Data Exploration and Preprocessing

Our data captures 2,930 houses sold in Ames Iowa from the years 2006 - 2010. These 2,930 observations of residential homes are include 79 variables. Of these 79 variables, 38 are numeric, 17 are ordinal, and 24 are categorical. The majority of the ordinal variables have scales from poor to excellent. In preprocessing the data, we converted all such scales to 1-5 numerical scales. At the end of this process, 54 numeric variables remained, which were then included in our PCA, CFA, random forest classification, and linear regression.An additional step was taken to clean the dataset by replacing NA with 0 for specific scale variables where NA did not indicate a missing value, but instead a lack of the particular home feature. An example of this is basement quality, where an NA value meant that the home did not have a basement. Because the basement quality variable is a scored variable, we interpret this to mean that not having a basement is worse than having even a poor quality basement.

In addition to fitting those ordinal variables to numeric scales, we addressed specific missing values. There were 490 missing values in lot frontage, which were replaced with the

median value of lot frontage due to the right skewness of the variable. Stability and reliability of our components and data were tested prior to our analysis with KMO (0.83), Bartlett's Test of Sphericity ($p<2.22 \times 10^{-16}$), and Cronbach's Alpha (0.17).(Figures 6-8) KMO and Bartlett's confirmed stability and reliability. The low Cronbach's Alpha can be explained by the dependencies in our data between related variables. Specifically, there are variables that measure the same aspect like quality or square footage for variables in finished and unfinished states. An example of this is the unfinished basement square footage variable and the related finished square footage variable. We present some important descriptive statistics regarding home features from our dataset in Figure 1.

	min.	max.	median	mean
Sale Price ($)	12,789	755,000	160,000	180,796
Lot Area (sf)	1,300	215,245	9,436	10,148
Bedrooms	0	8	3	-
Full Bathrooms	0	4	2	-
Fireplaces	0	4	1	-
Overall Quality	1	10	6	-
Garage Size in Cars	0	5	2	-

Figure 1: Ames Dataset Descriptive Statistics Sample.

3.2. Principal Component Analysis

Principal component analysis (PCA) is a method used in order to reduce dimensionality. Essentially through finding the correlation amongst a group of variables and placing them in a component, uncorrelated variables are achieved. From such analysis, variance in the dataset and how well each component explains the variance can be concluded. We used R to conduct our principal component analysis with a varimax rotation. Three components were used in the final analysis after examining the scree plot as shown in Figure 9. There were three components with an eigenvalue of 1.0 or above, and two components of the knee in the plot is observed. The initial analysis ranged from two to five components, and where three was decided to be optimal due to its variance values and how the variables were distributed into the components with no cross-loadings. The PCA was run at least three times with a cut off value of 0.4 in order to remove variables that were not being loaded into any components or to remove variables that showed cross loading.

3.3. Common Factor Analysis

Following PCA, a common factor analysis (CFA) was performed in order to further assess underlying factors among housing features. We ran CFA on our subset data of 54 numeric and ordinal variables using a varimax rotation for greater interpretability. Analysis was run on 2-10 factors with a cutoff of 0.4 for our loadings. The optimal number of factors was determined to be three by selecting the number of factors that presented interpretable results, without cross-loadings, or variables failing to load. Cumulative variance explained by each factor was also used to determine whether the factor analysis was sufficient and results were then compared with PCA.

3.4. Random Forest Classification

A random forest classifier (RF) computes the majority classification among a set of decision trees. Our aim with this method is to determine the extent to which RF can predict the sale price range of a home based on its features. Our model used default settings in R consisting of 500 decision trees with 7 variables tried at each split and split on impurity. In order to perform RF on the dataset, we first grouped home sale prices into 4 equally sized groups: Low (<$129,499), Low Middle ($129,500 - $160,000), High Middle ($160,001 - $213,500), and High (>$213,501). A train/test split of 70% train to 30% testing was used on our 54 numeric variables subset. Initially our model performed quite well, but to increase our confidence in the technique, we ran the random forest classifier 10 times on different random splits of the dataset and averaged our accuracy, sensitivity, and specificity. This was an important step for this method, as our first test was significantly more successful than subsequent runs. We believe this should aid in reproducibility for this technique.

3.5. Multivariate Linear Regression

As most of the literature supports, we also use a backward elimination linear multivariate regression to remove any redundant or insignificant variables and then to measure the marginal house feature effects on overall price. We chose to utilize the backward elimination method to build our model because it begins with all of the variables then removes any variables that are not statistically significant. Because we are using a linear method, we are assuming four things. The first assumption is that house sale prices are linearly related with the independent variables in our model. We also assume no severe multicollinearity exists, there are no influential outliers in the data set, and the errors are homoscedastic and not autocorrelated.

Before we conducted a linear regression analysis, we analyzed a correlation matrix between all of the numeric variables to test for early signs of multicollinearity, as seen in Figure 10. The most positively correlated variables are between the number of fireplaces and fireplace quality, exterior quality and overall property quality, and pool area and pool quality variables. The correlations are all over 0.7, so it will be important to remember the high associations between these variables in our final model. The most negatively correlated variables are the basement square feet variables with associations ranging from -0.4 to -0.5. In order to run a VIF test, we ran a full regression that included every numerically coded variable in the data set. The

results from the full regression can be observed in Figure 11 and the results from the VIF can be seen in Figure 12. The highest VIF score is 6.5775, which is less than 10, so we can comfortably move forward with our linear analysis without being concerned with multicollinearity.

4. Discussion and Results

4.1. Principal Component Analysis

Principal component analysis on the numerical variables of the dataset revealed three significant components. The scree plot in Figure 9 shows that three components would suffice, however both two and four components also were tested. From the loadings shown in Figure 13, the variables distribute well into three components with at least three variables in each component. There is a cumulative variance of 0.62, which indicates that the model explains the data set well. In the first principal component can be called House Modernity, with variables such as the garage year built, house year built and year remodeled have the most significance on the component with values of 0.86, 0.83, 0.81. The second component is most weighted by the finished square footage of basement type 1 with a value of 0.84. Therefore, we call the second component Basement Properties. The third component is most influenced by the lot frontage calculated in linear feet and lot area calculated in square feet, so we classified it as Size. From this analysis, variables related to the size and condition of a house are significant in the grouping of components and thus would be significant in prediction models such as regression.

3.2. Common Factor Analysis

The common factor analysis produced three significant factors which cumulatively explain 58.3% of the variance. The full CFA results can be seen in Figure 2:

	Factor1	Factor2	Factor3
overallqual	0.803		
yearbuilt	0.750		
yearremodadd	0.725		
fullbath	0.596		
garagecars	0.674		
garagearea	0.639		
exterqual	0.824		
bsmtqual	0.673		
heatingqc	0.586		
kitchenqual	0.763		
bsmtfinsf1		0.813	
bsmtunfsf		-0.695	
bsmtfullbath		0.692	
bsmtfintype1		0.756	
fireplaces			0.936
fireplacequ			0.868
x1stflrsf	0.459		
	Factor1	Factor2	Factor3
SS loadings	5.518	2.346	2.047
Proportion Var	0.325	0.138	0.120
Cumulative Var	0.325	0.463	0.583

Figure 2: Common Factor Analysis showing 3 factors and variance explained.

The first factor is called Quality, Quantity, and Age because of variables it includes: overall property quality, year built, and number of full bathrooms. All of the variables in this factor have positive relationships with on another. For example, newer, higher overall quality

homes are more likely to have higher quality kitchens, basements, and larger garages and in some cases more square footage. Factor 2 is called Basement because it includes variables related to basement features, such as finished basement square feet and number of full bathrooms in the basement. We note that unfinished basements square footage has a negative relationship with a finished basement's square footage. The final factor had only 2 variables, both related to fireplaces, which gave it the name Fireplace. While it only has 2 variables, it is interesting that having a fireplace at all explains about as much variance in the data as having a basement. These results are comparable to the results of the PCA, which also concludes that 3 components were sufficient to explain our data. It is important to note that multiple factor numbers and variable subsets were run in our CFA, and of all the models, this 3 factor one was the most interpretable.

3.3. Random Forest Classification

Our random forest classification results in a model that predicts sale price range with roughly 80.1% test set accuracy on average. The model was run 10 times on different training and testing subsets of the data. Initially, the model had 92.7% accuracy, but subsequent runs within the 77-80% accuracy range revealed that this was likely due to a lucky train/test split for our model. Sensitivity and specificity are also noted for the four sale price ranges and are an average. We usually achieve high specificity on the high and low price ranges, and moderate specificity on the middle ranges. Sensitivity is similarly higher for the low and high sale price ranges than for the middle ranges. This suggests that the model would rarely confuse a high and low priced home based on features, but has some difficulty distinguishing between closely related homes selling at similar price points on both sides of the ranges. Figure 3 demonstrates an average run of our Random Forest Classifier. The random forest classifier also indicates the most important features for effecting the sale price range classification. Our most important features include overall quality, 1st floor square footage, year built, lot area, and year remodeled. Feature importance to the RF model is summarized in Figure 4 and Figure 14.

```
housing_prediction high low middle high middle low
       high         188   0          15          1
       low            1 172           1         33
       middle high   27   1         164         26
       middle low     3  42          52        153

Overall Statistics

               Accuracy : 0.7702
                 95% CI : (0.7409, 0.7976)
    No Information Rate : 0.2639
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6937

 Mcnemar's Test P-Value : 0.01894

Statistics by Class:

                     Class: high Class: low Class: middle high Class: middle low
Sensitivity               0.8584     0.8000             0.7069            0.7183
Specificity               0.9758     0.9473             0.9165            0.8544
```

Figure 3: Random Forest Testing Set Sample run showing Confusion Matrix, 77% accuracy, Moderate Sensitivity, and High Specificity.

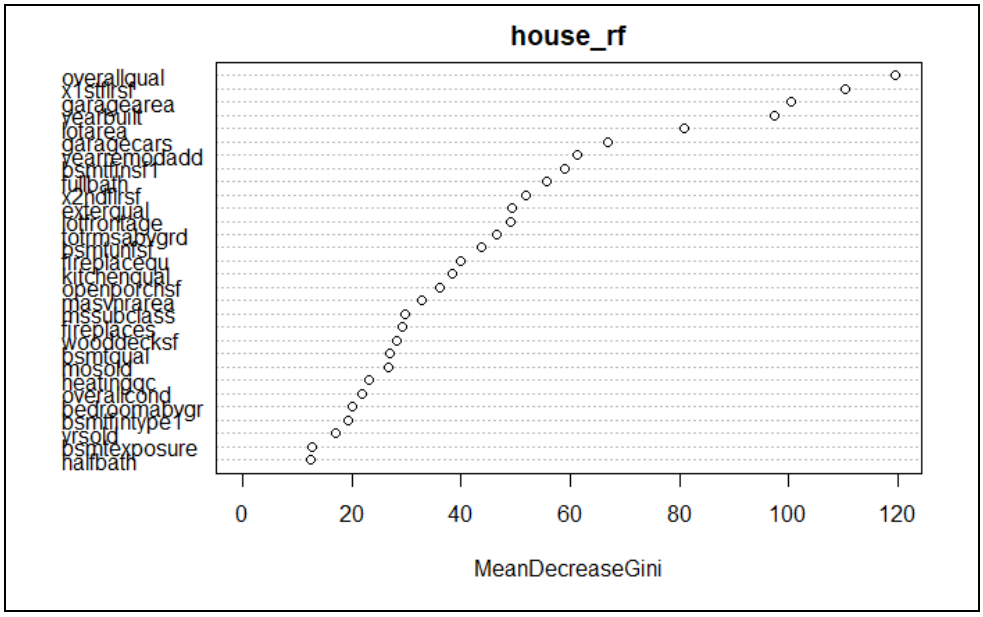

Figure 4: Feature importance using Mean Decrease in Gini Index for the Random Forest model. Overall Quality, First Floor Square Footage, Garage Area, Year Built, Lot Area lot area are among the most important and Half Baths, Basement Exposure, and Year sold are among the least important to this RF model.

3.4. Multivariate Linear Regression

Because we know that the full regression is not likely to be the best fitting model for predicting house prices, we then ran a backward elimination regression. The resulting model is shown in Figure 5.

$$\widehat{Sale\ Price} = Type\ of\ Home + Lot\ Area + Overall\ Quality + Overall\ Condition + Year\ Built + Masonry\ Type + Sq\ Ft\ of\ Finished\ Basement + Rating\ of\ Basement + Sq\ Ft\ of\ Unfinished\ Basement + Full\ Bath\ in\ Basement + Full\ Bathrooms + Bedrooms\ Above\ Ground + Kitchen\ Above\ Ground + Total\ Rooms\ Above\ Ground + Garage\ Cars + Garage\ Area + Wood\ Deck\ Sq\ Ft + Open\ Porch\ Sq\ Ft + Screen\ Porch\ Sq\ Ft + Pool\ Area + Year\ Sold + Exterior\ Quality + Basement\ Quality + Basement\ Condition + Basement\ Exposure + Rating\ of\ Second\ Basement + Heating\ Quality + Kitchen\ Quality + Home\ Functionality + Fireplace\ Quality + Garage\ Quality + Pool\ Quality + \varepsilon$$

Figure 5: Linear Regression Model with significant features included.

The results from our final regression can be seen in Figure 15. With an increase in the adjusted R-squared from our full regression, we can see that our model's overall fit has improved. The most significant features that raise a house price include the lot area, overall condition of the property, square feet, year built, number of full bathrooms in the basement, and kitchen quality. The most significant property features that have a negative effect on a house's price are garage quality, basement condition, bedrooms above ground, and pool area. While some of these variables may appear to be surprising to have negative effects, it is important to remember that these are the marginal effects of above ground bedrooms, and basement condition, while we also control for the square feet of every floor in a home.

3.5. Limitations

There are limitations in analyzing static data, especially such as forecasting home prices during the Great Recession. Since the data is not being tracked over a period of time, it is difficult to accurately predict the price of a home. There may have been a single year in which the prices of homes were below average and it could not have anything to do with the features of the home. Therefore, it is important to take this into consideration when examining such data. We can draw conclusions on what influences the price of a home but unfortunately it is difficult to predict home price accurately. Another limitation is that the dataset only included features in regard to the home itself. There were no other macroeconomic features such as percent of population that are family households, or features related to distances to the city center or parks. Including information like this could have better deepen the understanding of what influences home prices, and what give cities greater insight into what types of homes to establish where. We also do not have every property's history of renovation, which has shown to be influential from the literature. Future work could focus on addressing these limitations by combining data from multiple locations and exploring other aspects of the homes or neighborhoods.

4. Conclusion

After conducting analysis on the Ames housing data using principal component analysis, common factor analysis, random forest classification, and linear regression we can conclude that although it is difficult to accurately predict house price given the time period in which the data was collected, it is possible to determine what features most influence the price of a home. House price is highly influenced both by the size of its features and its newness in regard to its year built or year remodeled. If a home is newer and more spacious, then the price of the house will be higher on average. More specifically, influential features on a home's price include square footage, the year the home was built, if and when it was remodeled, and overall quality. While a homeowner looking to sell may not be able to control for square footage and size, they could prioritize remodeling if they are looking to sell at a higher price. The practice of home-flipping, in which investors will purchase a home, remodel it, and sell it at a higher price is evidence of the success of this approach. The profits made in this business rely on selling a home for a higher price than it was bought for due to its remodeling. Our findings suggest a basement, kitchen, and/or bathroom remodel are most critical to achieving a higher sale price. Further research can include more features related location such as to distances to the city center or more macroeconomic data such as inflation and interest rates.

5. 1 References

1. Amari S., & Tularam G.A. (2012) Performance of Multiple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices. *Journal of Mathematics and Statistic*s *2012*, 8 (4), 419-434.
2. Bogin, A.N., & Doerner, W.M. (*2019*) Property Renovations and Their Impact on House Price Index Construction. *Journal of Real Estate Research*: 2019, 41(2), 249-283.
3. Kishor, K.N., & Marfatia, H.A. (2016) Forecasting House Prices in OECD Economies. Journal of Forecasting, 37(2), 170-190.
4. Fuller, C. (2016) Rockin' the Suburbs: Home Values and Rents in Urban, Suburban and Rural Areas. *Zillow Internal Research*.

6. Appendix

```
        Bartlett's Test of Sphericity

Call: bart_spher(x = housing_data_nums)

     X2 = 68414.163
     df = 1128
p-value < 2.22e-16
```

Figure 6: Bartlett's Test

```
Reliability analysis
Call: alpha(x = housing_data_nums, check.keys = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N    ase  mean  sd median_r
      0.09      0.84    0.91     0.099 5.3 0.0048 58614 173    0.053

 lower alpha upper     95% confidence boundaries
0.08 0.09 0.1
```

Figure 7: Chronbach's Alpha

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = housing_data_nums)
Overall MSA =  0.8
MSA for each item =
   mssubclass   lotfrontage       lotarea   overallqual   overallcond     yearbuilt
         0.62          0.87          0.86          0.95          0.60          0.82
 yearremodadd    masvnrarea    bsmtfinsf1    bsmtfinsf2    bsmtunfsf     x1stflrsf
         0.90          0.96          0.65          0.43          0.58          0.71
    x2ndflrsf  lowqualfinsf  bsmtfullbath  bsmthalfbath      fullbath      halfbath
         0.61          0.53          0.82          0.46          0.85          0.71
 bedroomabvgr  kitchenabvgr  totrmsabvgrd    fireplaces    garagecars    garagearea
         0.78          0.61          0.80          0.75          0.86          0.86
  wooddecksf   openporchsf enclosedporch    x3ssnporch   screenporch      poolarea
         0.91          0.92          0.82          0.59          0.70          0.54
       mosold        yrsold         alley      lotshape     exterqual     extercond
         0.54          0.48          0.76          0.94          0.95          0.68
    bsmtqual      bsmtcond  bsmtexposure  bsmtfintype1  bsmtfintype2     heatingqc
         0.91          0.52          0.87          0.92          0.66          0.95
  kitchenqual    functional   fireplacequ     garagequ        poolqc         fence
         0.95          0.61          0.79          0.87          0.53          0.93
```

Figure 8: Kaiser-Meyer-Olkin factor adequacy

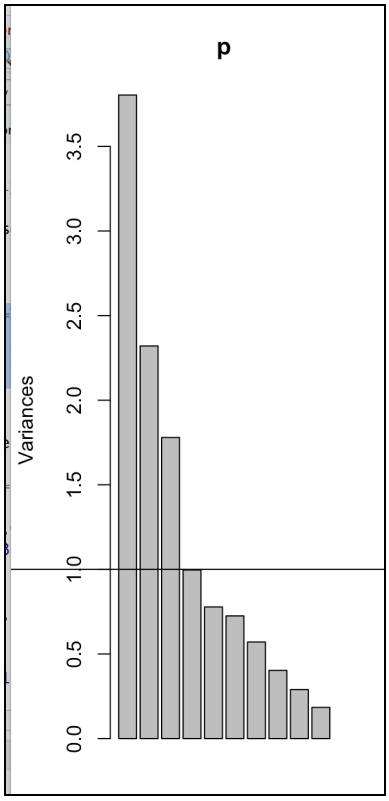

Figure 9: Scree plot for PCA Analysis

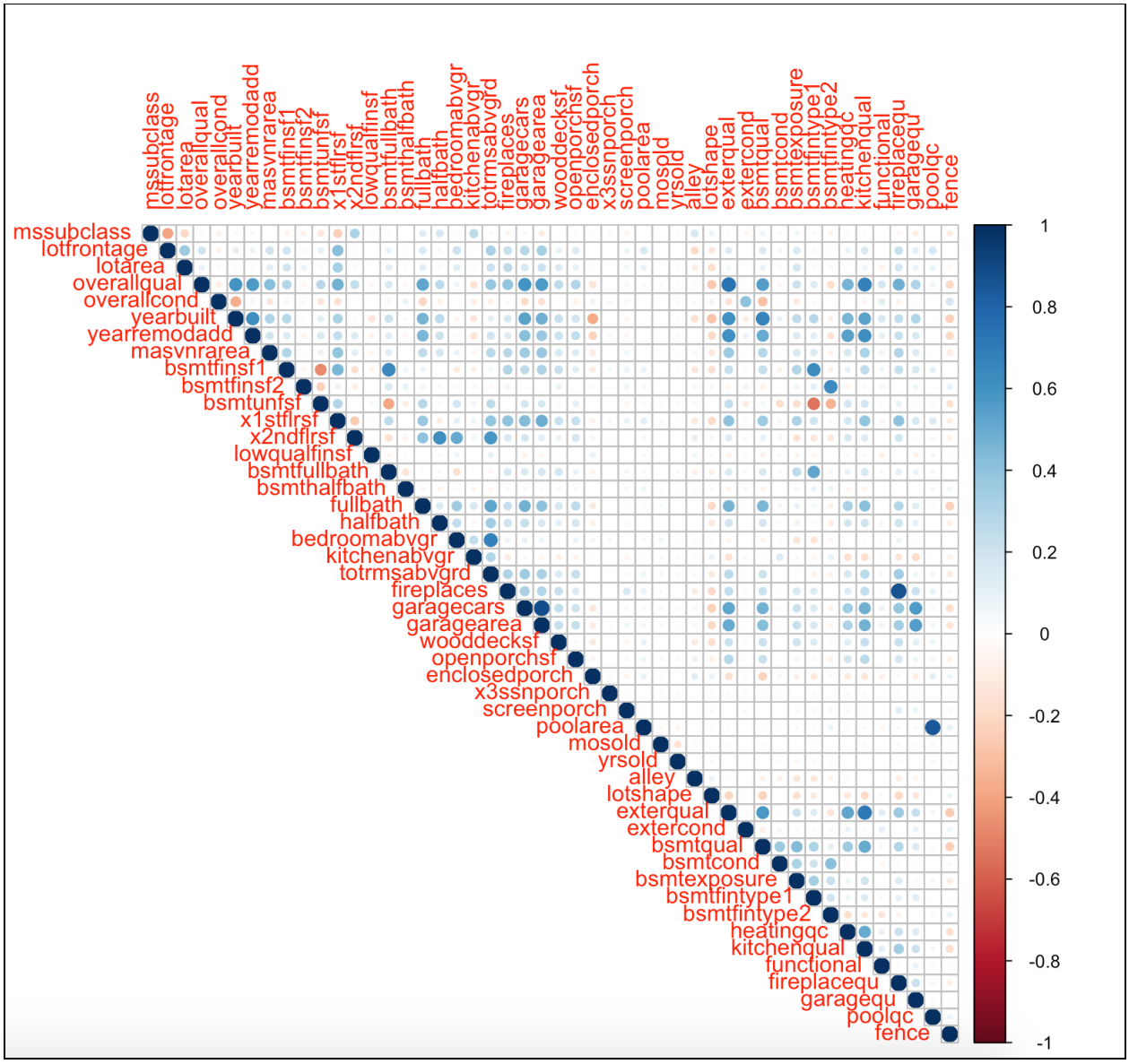

Figure 10: Correlation Matrix

```
Residuals:
    Min      1Q  Median      3Q     Max
-525119  -15011   -1241   12994  248940

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.498e+06  8.901e+05   1.683 0.092443 .
mssubclass    -1.683e+02  1.751e+01  -9.609  < 2e-16 ***
lotfrontage   -3.864e+01  3.424e+01  -1.129 0.259094
lotarea        3.624e-01  8.350e-02   4.340 1.47e-05 ***
overallqual    1.142e+04  7.914e+02  14.431  < 2e-16 ***
overallcond    4.812e+03  6.974e+02   6.900 6.38e-12 ***
yearbuilt      1.091e+02  4.142e+01   2.634 0.008487 **
yearremodadd  -9.070e+00  4.466e+01  -0.203 0.839072
masvnrarea     2.836e+01  3.792e+00   7.478 9.93e-14 ***
bsmtfinsf1     2.576e+01  3.232e+00   7.968 2.29e-15 ***
bsmtfinsf2     3.211e+01  5.293e+00   6.066 1.48e-09 ***
bsmtunfsf      1.462e+01  3.105e+00   4.710 2.60e-06 ***
x1stflrsf      4.235e+01  3.777e+00  11.211  < 2e-16 ***
x2ndflrsf      4.732e+01  3.199e+00  14.792  < 2e-16 ***
lowqualfinsf   2.201e+01  1.274e+01   1.728 0.084048 .
bsmtfullbath   6.059e+03  1.600e+03   3.786 0.000156 ***
bsmthalfbath  -2.878e+03  2.501e+03  -1.151 0.249952
fullbath       3.675e+03  1.728e+03   2.127 0.033510 *
halfbath       1.095e+03  1.677e+03   0.653 0.513765
bedroomabvgr  -3.419e+03  1.080e+03  -3.166 0.001562 **
kitchenabvgr  -1.306e+04  3.435e+03  -3.802 0.000147 ***
totrmsabvgrd   2.060e+03  7.659e+02   2.690 0.007186 **
fireplaces     2.104e+03  1.880e+03   1.119 0.263341
garagecars     6.148e+03  1.857e+03   3.310 0.000944 ***
garagearea     2.084e+01  6.279e+00   3.318 0.000917 ***
wooddecksf     8.257e+00  5.029e+00   1.642 0.100741
openporchsf   -2.333e+01  9.341e+00  -2.498 0.012562 *
enclosedporch  8.454e+00  9.879e+00   0.856 0.392182
x3ssnporch     3.062e+00  2.276e+01   0.135 0.892992
screenporch    5.355e+01  1.054e+01   5.081 3.99e-07 ***
poolarea      -1.281e+02  3.043e+01  -4.207 2.66e-05 ***
mosold        -5.119e+01  2.133e+02  -0.240 0.810327
yrsold        -9.173e+02  4.417e+02  -2.077 0.037901 *
alley         -1.120e+03  1.648e+03  -0.680 0.496685
lotshape      -4.935e+02  4.352e+02  -1.134 0.256842
exterqual      1.227e+04  1.756e+03   6.990 3.40e-12 ***
extercond     -1.654e+03  1.705e+03  -0.970 0.332011
bsmtqual       1.136e+04  1.393e+03   8.151 5.32e-16 ***
bsmtcond      -4.169e+03  1.800e+03  -2.316 0.020620 *
bsmtexposure   5.103e+03  6.597e+02   7.736 1.41e-14 ***
bsmtfintype1   3.042e+02  4.124e+02   0.738 0.460793
bsmtfintype2  -1.821e+03  6.669e+02  -2.730 0.006374 **
heatingqc      1.642e+03  7.636e+02   2.151 0.031594 *
kitchenqual    9.677e+03  1.380e+03   7.014 2.88e-12 ***
functional     4.245e+03  9.345e+02   4.543 5.79e-06 ***
fireplacequ    1.547e+03  6.813e+02   2.271 0.023214 *
garagequ      -4.929e+03  1.064e+03  -4.631 3.80e-06 ***
poolqc         2.319e+04  5.544e+03   4.183 2.96e-05 ***
fence          3.805e+02  3.870e+02   0.983 0.325616

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

Residual standard error: 30620 on 2881 degrees of freedom
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.853
F-statistic: 355.2 on 48 and 2881 DF,  p-value: < 2.2e-16
```

Figure 11: Full regression results

```
> vif(m_back)
  mssubclass      lotarea  overallqual  overallcond    yearbuilt   masvnrarea   bsmtfinsf1   bsmtfinsf2
    1.507704     1.267891     3.859669     1.451136     3.394876     1.411412     6.269837     2.453822
   bsmtunfsf    x1stflrsf    x2ndflrsf lowqualfinsf bsmtfullbath     fullbath bedroomabvgr kitchenabvgr
    5.492501     6.577461     4.480004     1.082011     1.939169     2.401716     2.442124     1.668573
totrmsabvgrd   garagecars   garagearea   wooddecksf  openporchsf  screenporch     poolarea    exterqual
    4.461581     6.160388     5.606137     1.228920     1.212963     1.064184     3.622630     3.158533
    bsmtqual     bsmtcond bsmtexposure bsmtfintype2    heatingqc  kitchenqual   functional  fireplacequ
    3.160169     1.904018     1.590520     2.837706     1.603348     2.500687     1.183285     1.583116
   garagequ       poolqc       yr2006
    1.763241     3.688002     1.020786
```

Figure 12: VIF Test

```
Loadings:
             RC1    RC2    RC3
overallqual   0.764
yearbuilt     0.831
yearremodadd  0.812
fullbath      0.645
garageyrblt   0.858
bsmtqual      0.720
heatingqc     0.671
kitchenqual   0.744
bsmtfinsf1           0.840
bsmtfullbath         0.824
bsmtfintype1         0.837
lotfrontage                 0.753
lotarea                     0.710
x1stflrsf                   0.705
masvnrarea                  0.429

                 RC1   RC2   RC3
SS loadings    4.850 2.345 2.080
Proportion Var 0.323 0.156 0.139
Cumulative Var 0.323 0.480 0.618
```

Figure 13: Loadings for three principal components where RC1 = “House Modernity”, RC2 = ”Basement Properties”, and RC3 = “Size”.

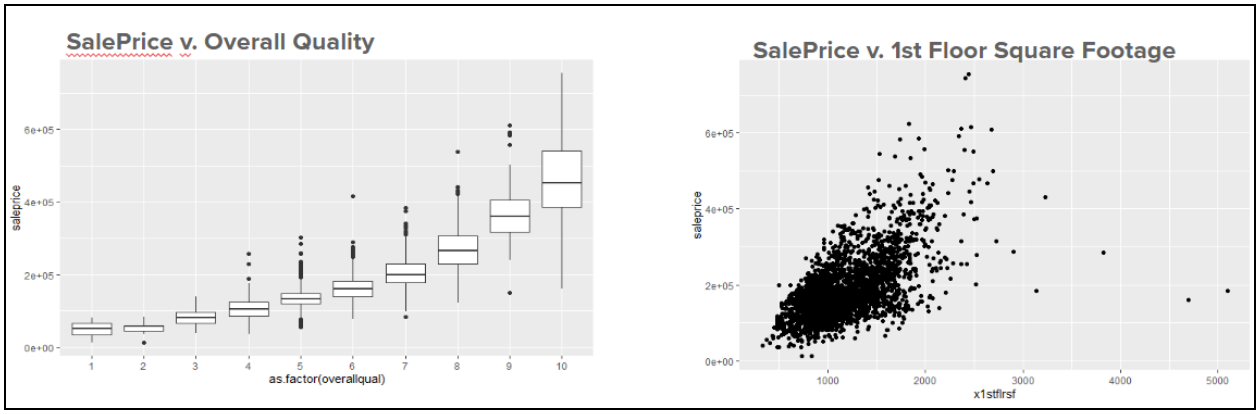

Figure 14: Top 2 Important Variables for the Random Forest Classifier against SalePrice

```
Residuals:
    Min      1Q  Median      3Q     Max
-531948  -14943    -986   12750  249347

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.374e+06  8.743e+05   1.572 0.116173
mssubclass  -1.657e+02  1.628e+01 -10.181  < 2e-16 ***
lotarea      3.614e-01  8.081e-02   4.472 8.05e-06 ***
overallqual  1.157e+04  7.873e+02  14.693  < 2e-16 ***
overallcond  4.593e+03  6.118e+02   7.508 7.95e-14 ***
yearbuilt    1.173e+02  3.445e+01   3.404 0.000674 ***
masvnrarea   2.806e+01  3.761e+00   7.462 1.12e-13 ***
bsmtfinsf1   2.609e+01  3.108e+00   8.395  < 2e-16 ***
bsmtfinsf2   3.181e+01  5.238e+00   6.072 1.42e-09 ***
bsmtunfsf    1.408e+01  3.015e+00   4.670 3.15e-06 ***
x1stflrsf    4.299e+01  3.701e+00  11.615  < 2e-16 ***
x2ndflrsf    4.876e+01  2.794e+00  17.449  < 2e-16 ***
lowqualfinsf 2.288e+01  1.270e+01   1.802 0.071727 .
bsmtfullbath 6.624e+03  1.502e+03   4.411 1.07e-05 ***
fullbath     3.132e+03  1.585e+03   1.976 0.048222 *
bedroomabvgr -3.522e+03  1.067e+03  -3.299 0.000981 ***
kitchenabvgr -1.316e+04  3.414e+03  -3.855 0.000118 ***
totrmsabvgrd 1.959e+03  7.590e+02   2.581 0.009894 **
garagecars   6.305e+03  1.844e+03   3.419 0.000638 ***
garagearea   1.918e+01  6.222e+00   3.083 0.002067 **
wooddecksf   8.758e+00  4.960e+00   1.766 0.077570 .
openporchsf -2.445e+01  9.232e+00  -2.648 0.008142 **
screenporch  5.355e+01  1.040e+01   5.149 2.80e-07 ***
poolarea    -1.250e+02  3.022e+01  -4.137 3.62e-05 ***
yrsold      -8.741e+02  4.341e+02  -2.014 0.044139 *
exterqual    1.186e+04  1.732e+03   6.848 9.12e-12 ***
bsmtqual     1.129e+04  1.378e+03   8.191 3.84e-16 ***
bsmtcond    -4.141e+03  1.772e+03  -2.337 0.019517 *
bsmtexposure 5.070e+03  6.455e+02   7.854 5.64e-15 ***
bsmtfintype2 -1.744e+03  6.625e+02  -2.632 0.008522 **
heatingqc    1.510e+03  7.474e+02   2.020 0.043471 *
kitchenqual  9.503e+03  1.348e+03   7.047 2.27e-12 ***
functional   4.129e+03  9.277e+02   4.451 8.88e-06 ***
fireplacequ  2.208e+03  3.942e+02   5.602 2.31e-08 ***
garagequ    -4.769e+03  1.051e+03  -4.539 5.89e-06 ***
poolqc       2.251e+04  5.513e+03   4.084 4.55e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30610 on 2894 degrees of freedom
Multiple R-squared:  0.855,     Adjusted R-squared:  0.8532
F-statistic: 487.5 on 35 and 2894 DF,  p-value: < 2.2e-16
```

Figure 15: Backward Elimination Regression