Ames House Data Analysis

MA584

Ekaterina Gorbunova
Kelly Huang

05/02/2019

Table of Contents

Page

Abstract

The housing and real estate industry has always been, and remains a very popular and important business area, as people constantly buy, sell, and rent houses. Buying a house is a very important decision in anyone's life and should be looked at from different perspectives. When completing a purchase, one should be sure about the conditions of the house, its quality, and its price. The last one is especially significant in people's decisions and thus led to this study based on the dataset of 1200 houses in Ames. The goal of this research was to understand better what drives changes in housing pricing.

Introduction

Our research on house sale price in Ames included several steps. First of all, we analyzed the data given, to indicate any outliers, to spot any particular correlation, and check assumptions for the linear regression model. Secondly, we plotted the best fit line to look at the significance of each numerical variable related to size, as size factors appear to be dominating in affecting the sale price. Thirdly, we looked at how significant is each categorical feature after adjusting for the size effect. Then, we saw if there was any statistically significant difference across neighborhoods, as it would help us answer our main research question: which neighborhoods are overpriced or underpriced. We also conducted a clustering analysis to see what are the main factors for houses groups. In the end, we concluded which particular neighborhoods had higher or lower prices than expected.

Research Question

Given all background, we wanted to investigate what features of the house affect its sales price and try to spot particular neighborhoods in Ames with underpriced and overpriced houses.

Question 1

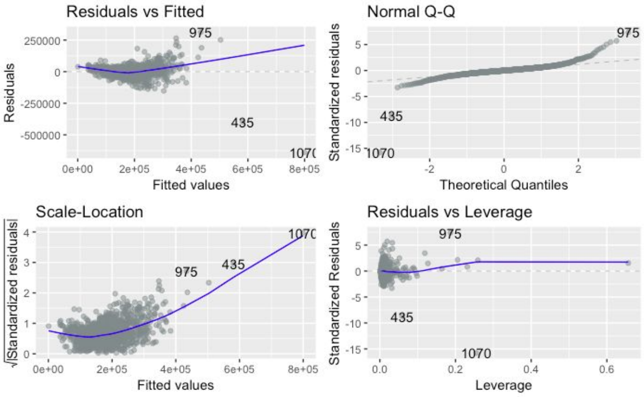
1. *Check of assumptions*

Before fitting a linear regression model, we had to check if all assumptions hold. These include linear relationship, multivariate normality, and no or little multicollinearity. We made some plots to visualize our assumptions. As we can see from the graph, some assumptions are violated.

The first plot is for the linearity assumption, i.e. whether there is a linear relationship between the dependent and independent variables. As seen, the residual plot shows a fitted pattern, as our data is not generally symmetrically distributed around y=0 (purple line). This indicates that there is a difference in the gathered data and predicted values, which doesn't satisfy a linearity condition.

The second graph is a normal Q-Q (Quantile-Quantile), which indicates to what extent our data is normally distributed. Since the plotted "experimental" line doesn't identically mimic a straight fitted line, we cannot assume normality.

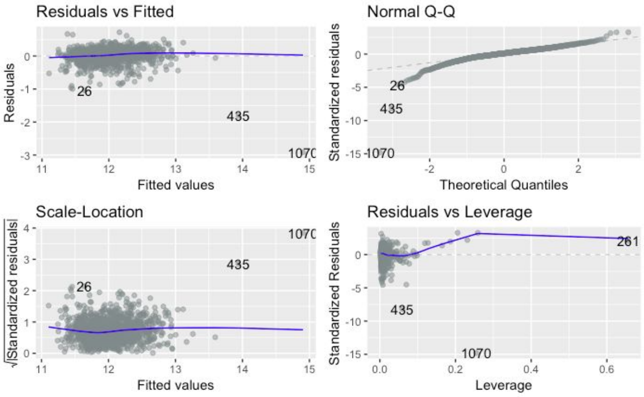
The third plot portrays a Scale-Location relationship, which provides information on whether residuals are spread equally along with the ranges of predictors. The horizontal blue line should have an equal spread of data points, which would mean that all of the independent variables have the same variance (deviation of a variable from its mean).



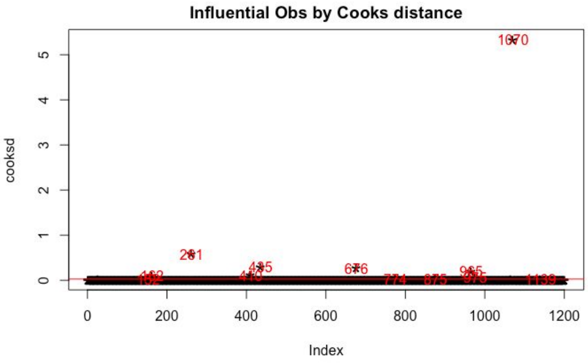
The last plot we created was fitting Residuals vs Leverage (a measure of how far away independent values are from each other). We then proceeded to spot any outliers that would affect the normality of the data. As we can see, there are some extreme points like "975", "435" and "1070", which we decided to omit.

2. *Log transformation*

In order to achieve normality in our results, we chose to use the natural logarithm of our dependent variable "SalePrice". Using the natural log for our dependent variable only linearizes the relationship curve and results in a better fitted model, as we can see in the graphs below.



3. Managing Outliers and Redundant Values

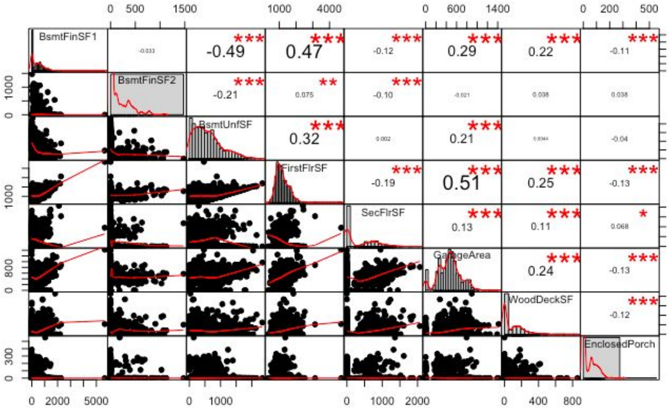


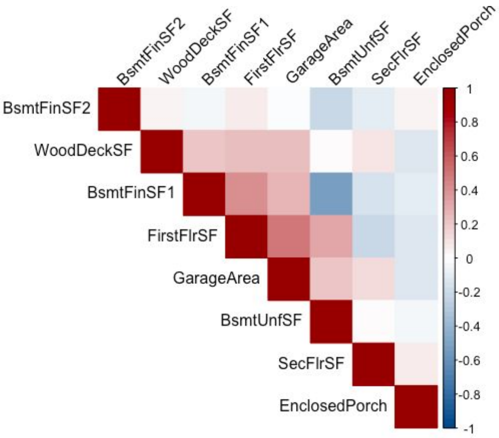
As we have noticed in the residuals vs leverage plot, there are some significant outliers. To check those more specifically, we constructed a Cooks distance plot to spot influential outliers. We decided to omit only the most significant ones, and thus created a new dataset excluding rows with indexes “1070”, “261”, “435”.

We also had to remove some of the variables from our model that were redundant and created NA’s. Such features were TotalBsmtSF, which was the sum of BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, and GrLivArea, a combination of 1stFlrSF, 2ndFlrSF, LowQualFinSF.

4. Fitting Regression Model

Finally, we were able to fit a linear regression model on continuous numerical variables with “sale price” being our response variable. We decided to use the backward selection as our variable selection method. At first, we included all variable in the model (without redundant ones) and concluded on the significance of each feature, based on p-values. Then we eliminated each feature one by one until we got a model with only significant variables (Appendix 1). We decided to check for the correlation between those variables as well. Included below is a correlation matrix of the significant variables and a more visualized plot. We decided that the correlation between variables was not affecting the model so much. We left the variables as is, eliminating only Pool Area as it didn’t make sense to correlate it with other variables.





5. *Analysis of the Model*

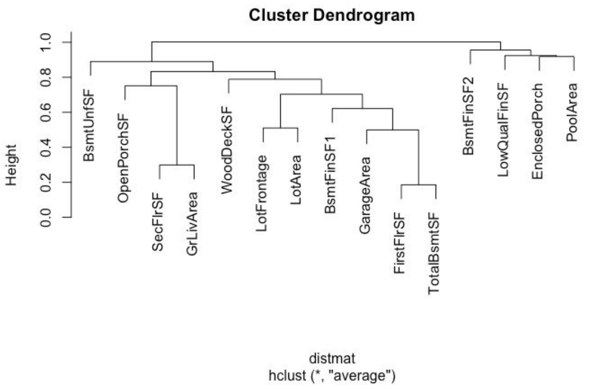
In the end, we got a model that helped us conclude which size variables are the most important in affecting the sale price. Garage Area appeared to have the biggest correlation coefficient, thus affecting the price of houses the most. Interestingly enough, an enclosed porch and pool area had negative coefficients, indicating that the price would drop with that area being bigger. We concluded that the maintenance of an enclosed porch in Iowa is more expensive and, thus, not beneficial; this is why the price would drop. Otherwise, the most important size features with a positive effect on sales appeared to be: BsmtFinSF1, BsmtUnfSF, FirstFlrSF, SecFlrSF, Garage Area and WoodDeckSF.

Question 2

1. *Correlation and Clustering*

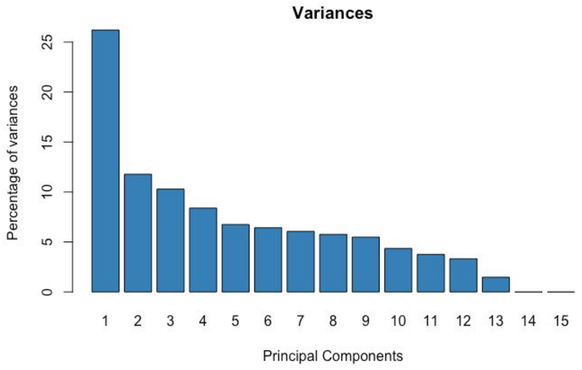
We then further analyzed the relationship between these continuous variables that described the size of the house. As described above, we computed the correlation between these variables and found that most were not very strongly correlated. Some had significant positive correlations, which agrees with the logic that more area allows for a larger home with more spacious rooms. Only the finished basement area and the unfinished basement area had a significant negative correlation, which is also reasonable.

Next, we used cluster analysis to group these variables. The resulting dendrogram implies that these continuous features form two main groups. Finished basement (2) area, low-quality finished area, enclosed porch pool area, and pool area are clustered together and are very similar to each other. All the other variables more similar to each other than to these four variables and thus form another group. However, the variables in this group are not all equally similar to each other. For example, the first floor area and total basement area are much more similar to each other than to the other variables. Lot frontage and lot area appear just as similar to each other but are not as close to the area of the first floor or basement.

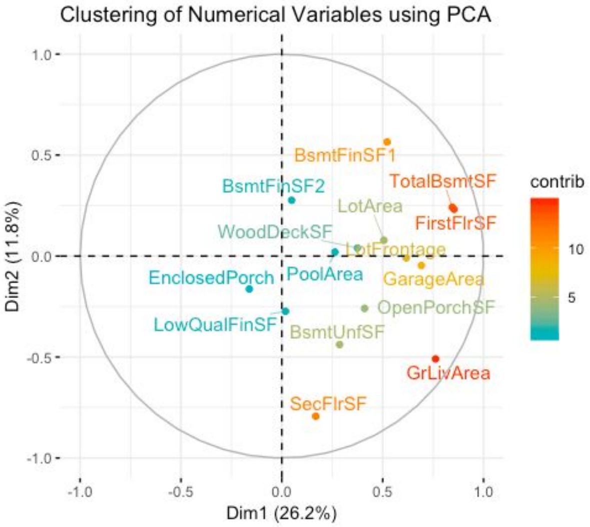


2. *PCA Analysis*

Finally, we performed a principal component analysis to determine if fewer dimensions could be used to represent these continuous features. Our goal was to use the appropriate number of principal components which captures about 80-90% of the variance. In our analysis, we found that that the first nine principal components explained 87% of the variance.



Additionally, we plotted the features using the first two components which resulted in the grouping of the features that agreed with the clustering analysis. From the biplot, we can observe which variables are related to each other, looking at their positioning, and how much they contribute to the model, looking at their color. The features that had a very low contribution formed the cluster that included the four variables that were very similar to each other. The other variables had a higher contribution and those that were found to be more similar were positioned more closely together.



Question 3

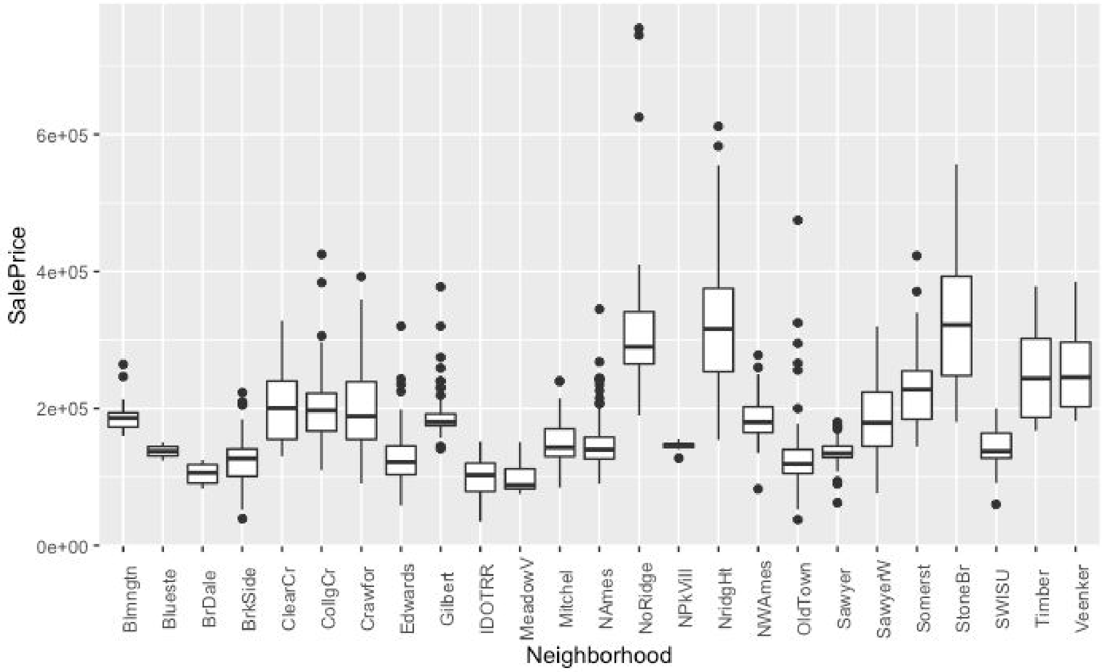
The model we built only considered the effect of the house size on the sale price, but this only explained about 70% of the variance. This led us to believe that the condition and quality of the house also had an effect on the sale price. We used ANOVA Type III SS, which describes how each variable contributes to the model after all others variables are in the model already. We determined the importance of these features by adding each one to the model individually and finding its significance after adjusting for size. We found that almost all the categorical features were significant. This makes sense, as the quality and condition of the living spaces and amenities in a home should be important to homeowners.

The only ones that were not significant were roof style, roof material, type of dwelling (MSSubClass), condition2, and miscellaneous features. It was interesting for both variables that described the roof were not significant. It implies that the roof is not a very important feature visually and functionally to homeowners, especially compared to size features. The type of home was also insignificant, which makes sense because the type of home would correlate to the size. After adjusting for size, the type of home would not be as important. Condition2 and miscellaneous features were also insignificant, which is probably explained by the small number of homes that actually have additional features that were not described by the other variables.

Question 4

In order to test if there is a significant difference between across means of house size and sale price across neighborhoods, we used MANOVA to get the F-value for each feature and for the model overall. Our model included numerical continuous variables grouped by neighborhoods. In the end, our model showed very significant p-value ($2.2 \times e^{-16}$) with Wilk's lambda = 0.11, so we were able to conclude that the means of house size and sale price are different. Pool Area is only variable with no statistically significant difference across neighborhoods.

We also created a box plot diagram to visualize sale price variable across neighborhoods. As we see on the right, the boxes vary a lot, as supported by our conclusion in MANOVA.

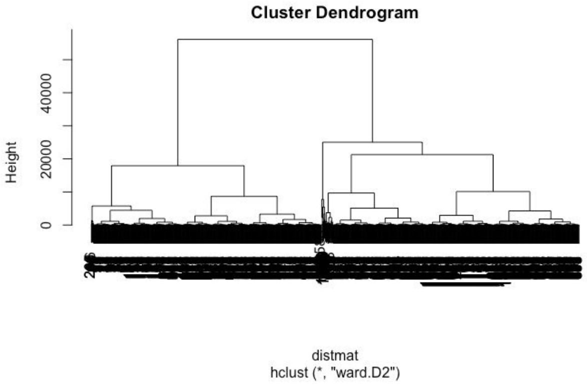


It is interesting to note that when we previously created a model with Neighborhood in addition to all the significant size variables, Neighborhood was still significant after adjusting for size effect. This seems to disagree with our findings in the MANOVA test, which suggests that the sizes differ across the neighborhoods. However, MANOVA only tests for whether all the means are equal or not. So rejecting the null hypothesis in this MANOVA test but still finding that Neighborhood is significant even after adjusting for size suggests that, while there are differences in size across some of the neighborhoods, the sales prices across the neighborhoods are still different enough for the neighborhood to still be an important feature in explaining the price.

Question 5

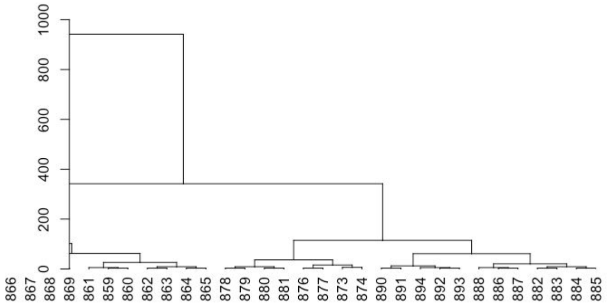
1. *Full data clustering*

Next we performed a cluster analysis on the individual houses to see how similar they are to each other and if we can group them into a number of categories. The following dendrogram was created to cluster the houses based on features, except for neighborhood, type of dwelling (MSSubClass), and zone classification (MSZoning). With so many observations though, it is very difficult to see which houses are clustered together, which houses are outliers, and how many clusters there are. However, this dendrogram does indicate that many of the houses are very dissimilar from each other, as the heights are very large.



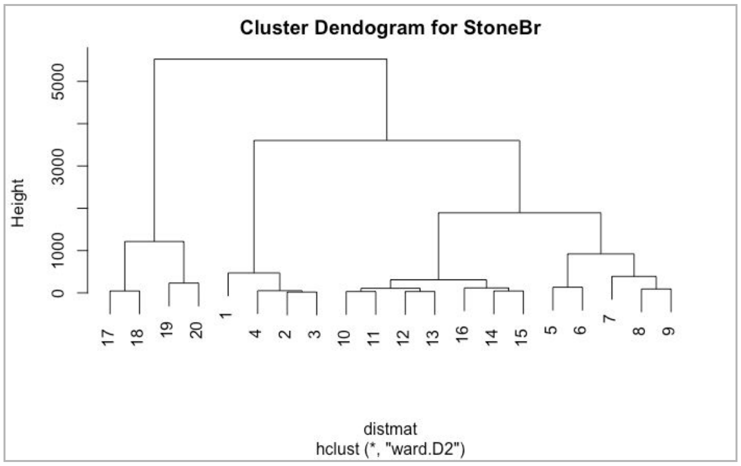
2. *Zoomed in branch*

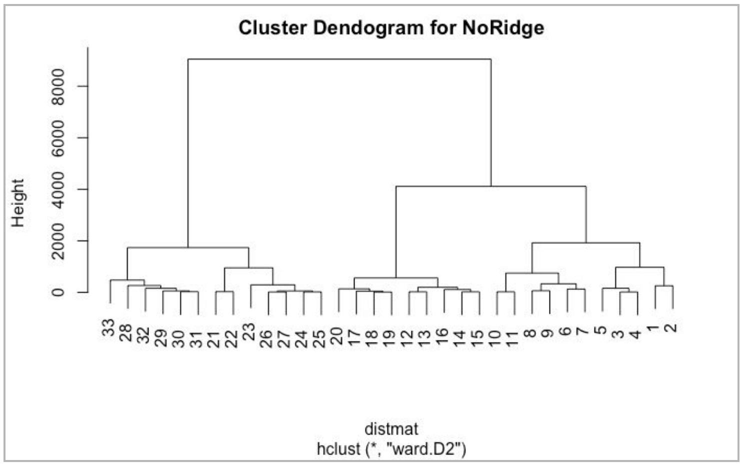
We were able to zoom in on parts of this dendrogram to get a closer look at the houses that were grouped together. For example, this second graph is a look at the right most branches of the first dendrogram. It appears that there are many houses that very similar to each other. It was found that these houses here share many of the same characteristics, such as many of the variables related to the lot, land, roof, garage, basement, building type, and house style. It is also important to note that these houses also shared the MSZoning and MSClass. However, these houses did come from a few different neighborhoods.

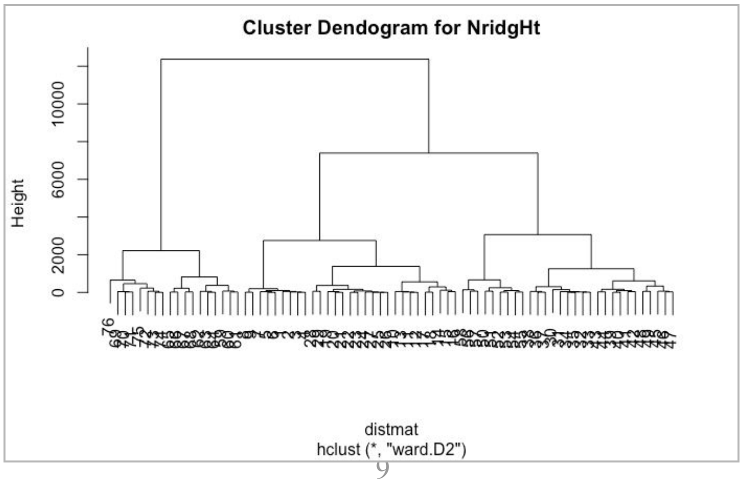


3. *Clustering by Neighborhoods*

With so many observations, it is very difficult to visualize all the clusters. Thus it would be much easier to try to cluster a smaller group of houses. We decided to separate the houses by neighborhood and then try to cluster them. Below are the dendrograms for three of the neighborhoods (Northridge, Stone Brook, and Northridge Heights). These dendrograms show that within the neighborhoods, there are houses that are very similar to each other. These three neighborhoods in particular have about four clusters. However, the groups are still very dissimilar from each other.







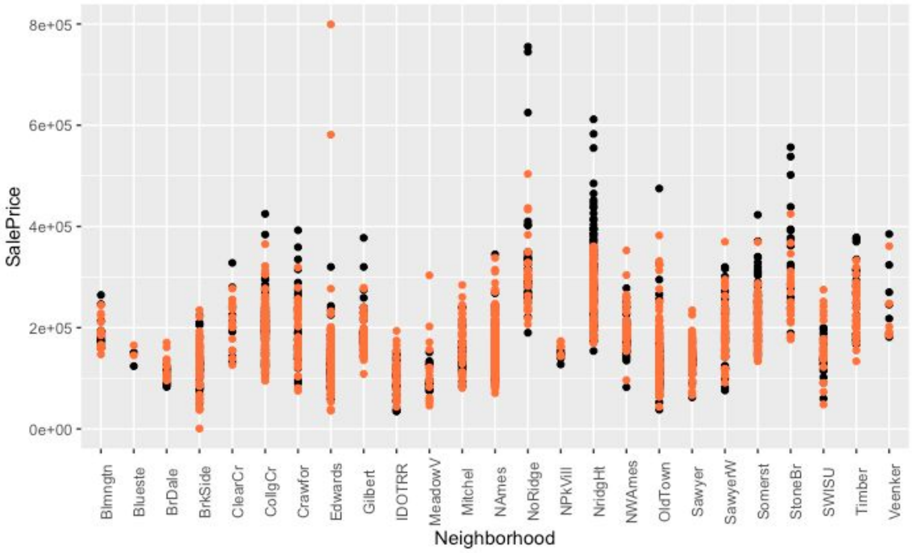
Conclusion

1. *Discussion*

After analyzing our variables and constructed plots for visualization, we were able to make some conclusions. We found out that the size of the home and living areas are the most dominating effect in the dataset and plays a bigger role in affecting price than quality and other conditions. An interesting insight was that the pool didn't affect the price positively and only created a disadvantage. Overall, we were also able to conclude that prices and various size features are different across neighborhoods. Finally, we wanted to see if some of those neighborhoods were overpriced or underpriced.

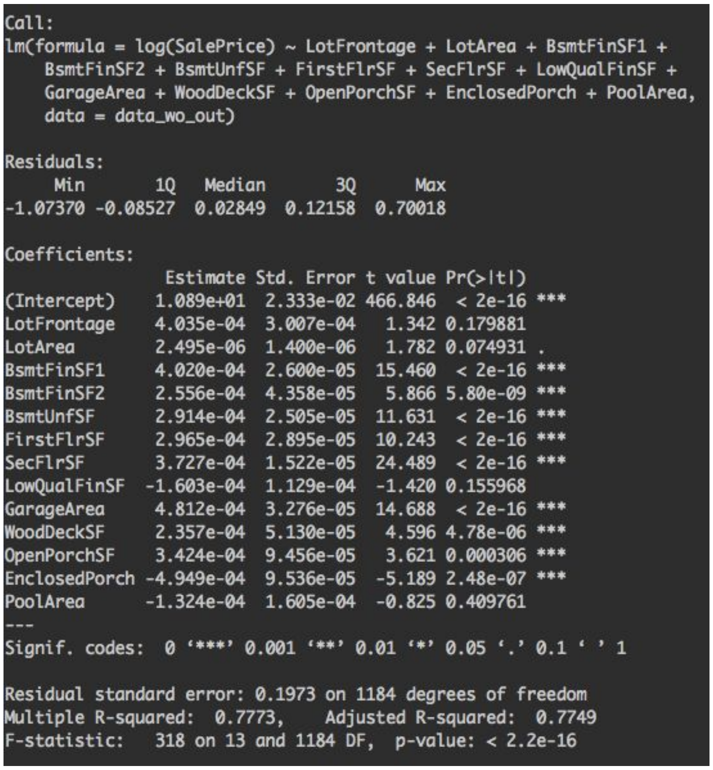
2. *Question 6*

In order to analyze neighborhoods that were overpriced or underpriced in our dataset, we decided to fit a linear regression model with both size variables and categorical ones to get the most accurate predictions. Our R-value was close to 80.48%, so we knew our model was good enough. We then constructed a plot with actual values represented by black dots and predicted values in orange. As we saw, dots overlap almost everywhere, indicating that the listed sales prices for most homes were similar to our model's predictions. However, there were some neighborhoods with many homes that were more expensive than our predictions. In Northridge Heights, there are very many homes that our model project a much lower price for. Northridge and Stone Brook also had a few homes that are considerably more expensive than predicted. None of the neighborhoods appeared to be severely underpriced. Most neighborhoods only had a few homes that our model predicted to valued higher. Edwards stands out because there are two homes with prices that are far less than predicted. Meadow Village also appears to have two homes that were valued higher by our model than actually listed, though the difference is much less drastic than in Edwards.



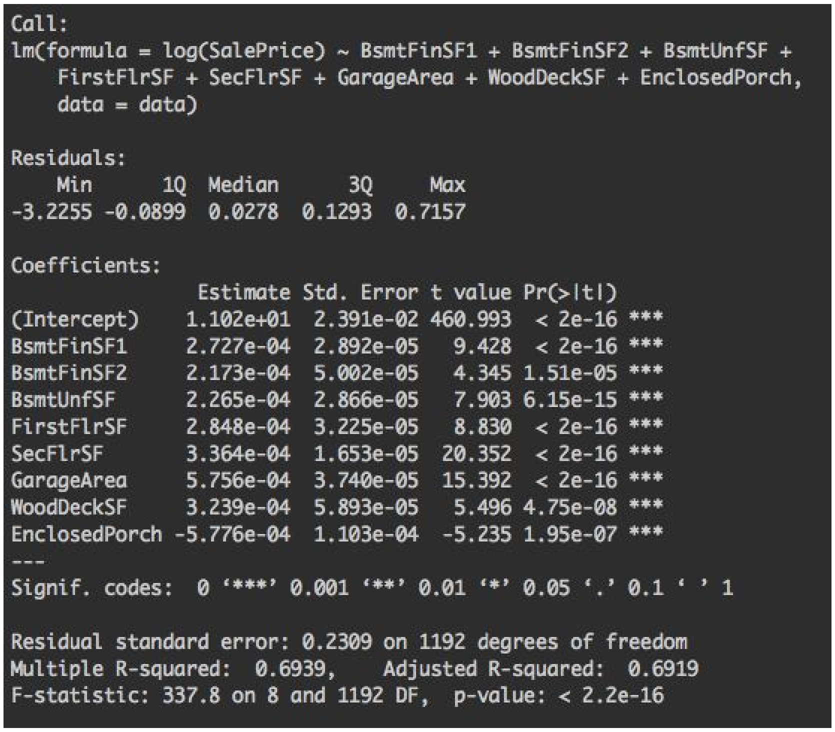
Appendix 1

Model with continuous variables before selection



```
Call:
lm(formula = log(SalePrice) ~ LotFrontage + LotArea + BsmtFinSF1 +
    BsmtFinSF2 + BsmtUnfSF + FirstFlrSF + SecFlrSF + LowQualFinSF +
    GarageArea + WoodDeckSF + OpenPorchSF + EnclosedPorch + PoolArea,
    data = data_wo_out)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07370 -0.08527  0.02849  0.12158  0.70018 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.089e+01  2.333e-02 466.846  < 2e-16 ***
LotFrontage    4.035e-04  3.007e-04   1.342 0.179881    
LotArea        2.495e-06  1.400e-06   1.782 0.074931 .  
BsmtFinSF1     4.020e-04  2.600e-05  15.460  < 2e-16 ***
BsmtFinSF2     2.556e-04  4.358e-05   5.866 5.80e-09 ***
BsmtUnfSF      2.914e-04  2.505e-05  11.631  < 2e-16 ***
FirstFlrSF     2.965e-04  2.895e-05  10.243  < 2e-16 ***
SecFlrSF       3.727e-04  1.522e-05  24.489  < 2e-16 ***
LowQualFinSF  -1.603e-04  1.129e-04  -1.420 0.155968    
GarageArea     4.812e-04  3.276e-05  14.688  < 2e-16 ***
WoodDeckSF     2.357e-04  5.130e-05   4.596 4.78e-06 ***
OpenPorchSF    3.424e-04  9.456e-05   3.621 0.000306 ***
EnclosedPorch -4.949e-04  9.536e-05  -5.189 2.48e-07 ***
PoolArea      -1.324e-04  1.605e-04  -0.825 0.409761    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1973 on 1184 degrees of freedom
Multiple R-squared:  0.7773,	Adjusted R-squared:  0.7749 
F-statistic:   318 on 13 and 1184 DF,  p-value: < 2.2e-16
```

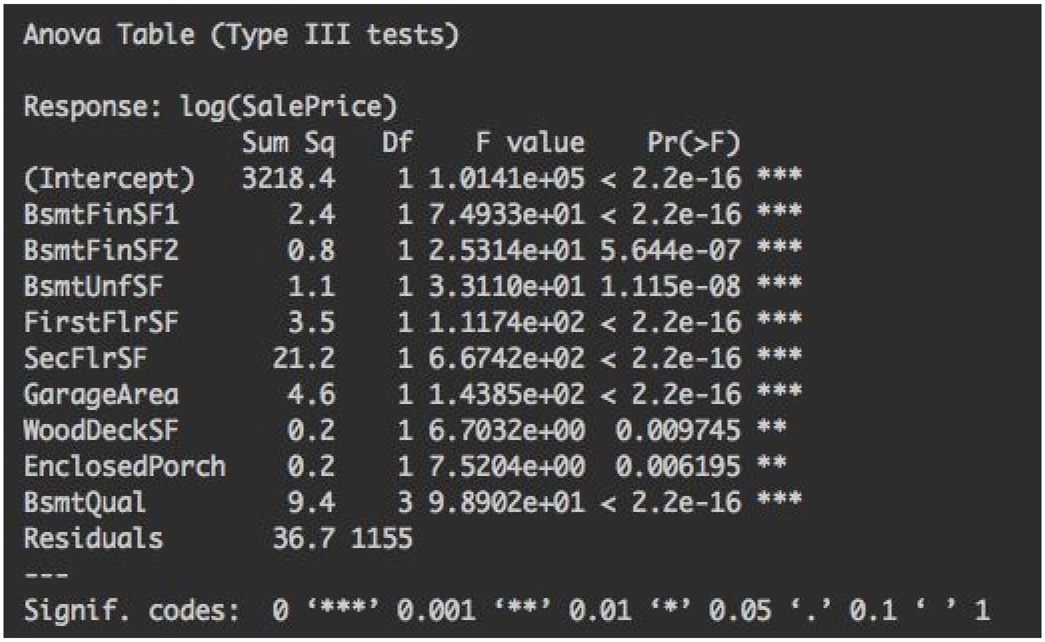
Model after backwise selection



```
Call:
lm(formula = log(SalePrice) ~ BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
    FirstFlrSF + SecFlrSF + GarageArea + WoodDeckSF + EnclosedPorch,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2255 -0.0899  0.0278  0.1293  0.7157 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.102e+01  2.391e-02 460.993  < 2e-16 ***
BsmtFinSF1     2.727e-04  2.892e-05   9.428  < 2e-16 ***
BsmtFinSF2     2.173e-04  5.002e-05   4.345 1.51e-05 ***
BsmtUnfSF      2.265e-04  2.866e-05   7.903 6.15e-15 ***
FirstFlrSF     2.848e-04  3.225e-05   8.830  < 2e-16 ***
SecFlrSF       3.364e-04  1.653e-05  20.352  < 2e-16 ***
GarageArea     5.756e-04  3.740e-05  15.392  < 2e-16 ***
WoodDeckSF     3.239e-04  5.893e-05   5.496 4.75e-08 ***
EnclosedPorch -5.776e-04  1.103e-04  -5.235 1.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2309 on 1192 degrees of freedom
Multiple R-squared:  0.6939,	Adjusted R-squared:  0.6919 
F-statistic: 337.8 on 8 and 1192 DF,  p-value: < 2.2e-16
```

Appendix 2

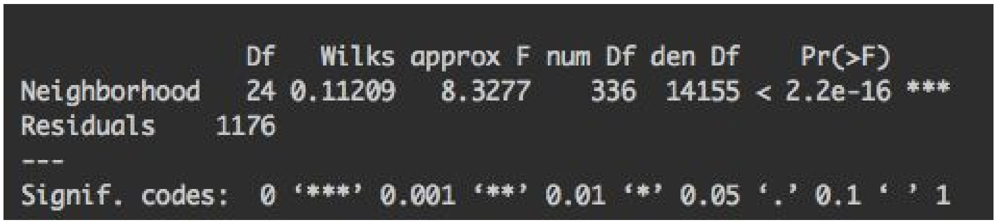
Example of significance test after adjusting for size effect with "BsmtQual" categorical feature, using Type III SS



```
Anova Table (Type III tests)

Response: log(SalePrice)
              Sum Sq   Df    F value    Pr(>F)    
(Intercept)   3218.4    1 1.0141e+05 < 2.2e-16 ***
BsmtFinSF1       2.4    1 7.4933e+01 < 2.2e-16 ***
BsmtFinSF2       0.8    1 2.5314e+01 5.644e-07 ***
BsmtUnfSF        1.1    1 3.3110e+01 1.115e-08 ***
FirstFlrSF       3.5    1 1.1174e+02 < 2.2e-16 ***
SecFlrSF        21.2    1 6.6742e+02 < 2.2e-16 ***
GarageArea       4.6    1 1.4385e+02 < 2.2e-16 ***
WoodDeckSF       0.2    1 6.7032e+00  0.009745 ** 
EnclosedPorch    0.2    1 7.5204e+00  0.006195 ** 
BsmtQual         9.4    3 9.8902e+01 < 2.2e-16 ***
Residuals       36.7 1155                         
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 3

Overall MANOVA test for the model



```
               Df   Wilks approx F num Df den Df    Pr(>F)    
Neighborhood   24 0.11209   8.3277    336  14155 < 2.2e-16 ***
Residuals    1176                                             
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```