# Why Is This House So Expensive?

DSC 324/424

Augustine Chiu, David Cornes, Jacob Wursteisen, Rikesh Patel, Dean Zillner

## Abstract

The Ames home sales dataset includes data on over 2,900 home sales in Ames Iowa from 2006-2010. To an outsider of the real estate market in this area, this seems to be a high transaction rate as the population for this area was only about 66,000 during this timeframe, with college students making up roughly half of that. Our goal is to develop a model that can accurately model what a home should be priced for, given input descriptive details/variables about the home and property. This can be useful for both home buyers and sellers for whatever the reason. During our research we used multiple regression analysis, correspondence analysis (CA), principal component analysis (PCA), k-means cluster analysis and canonical correlation analysis (CCA). Our results seemed to be aligned indicating a few select variables were significant factors in home pricing. CCA and regression analysis has the most robust results. Unsurprisingly, the homes' overall condition and quality were a leading factors, considering nothing else unchanged. Also homes with 1 or more full bathrooms in their basement were significant, which is interesting as many of the homes surveyed did not have basements. We were able to predict coefficients' significance using regression for over 78% of the errors in our regression model. In other words, we will be able to provide a prediction for a home's value, given our model's input data is provided, and provide an output which is 78% accurate. We suggest future modeling using a catered scope to more specific sub-markets within the dataset.

## Introduction

Real estate provides a data playground for analysis. From firms that compute algorithms for flipping homes, real estate investors looking for passive income, and families looking to buy or sell a family home, buyers and sellers must meet at a price that satisfies both parties. Using home sales data from the surrounding area is a great way to factor relative pricing of a home.

The Ames, Iowa dataset that we found includes 82 data points for 2,930 home sales from 2006-2010. This includes extensive home details ranging from neighborhood, lot square footage and descriptive characteristics of the home/property. Our goal is to take this data and parse it into a real life opportunity to properly price homes in the area. We consider our dataset to be especially extensive as the population of Ames in 2019 was around 66,000, with half of the population being students at Iowa State University. Throughout our analysis we will explore 5 different statistical approaches to building models to explain the pricing in the area.

## Literature Review

One paper sought to predict the price of real estate in Kuala Lumpur based on data of 352 past sales (Sarip et al. 2016). Fuzzy Least Squares Regression (FLSR) and Artificial Neural Networks (ANN) were applied in this research. Mean Absolute Error was used to evaluate both models rather than adjusted r-squared. The researchers found that applying FLSR, a more complex methodology compared to multiple regression analysis, was highly effective in predicting real estate prices compared to ANN. The most significant takeaway is that FLSR is also computationally efficient versus ANN, and therefore an ideal method to predict real estate prices based on characteristics of historical sales.

Another research paper sought to determine real estate prices in China (Li et al. 2011). The researchers utilized PCA to reduce complexity in the dataset and classify the factors that influence the Chinese real estate market. Specifically, they sought to reduce redundancy in the dimensions in their dataset. The results yielded 7 principal components that explained 89.7% of variance in the data. Their analysis found that there were 7 categories that influence real estate pricing in Chica ranging from political and social factors to characteristics of the house. Overall, this research showed that PCA can be “effective and precise” when applied to real estate price data.

The third research paper we looked at applied cluster analysis to real estate sales in Turkey in order to determine which areas of the country saw the highest return on investment over time (Hepşen et al. 2011). The researchers examined 71 metropolitan residential markets in Turkey and applied hierarchical clustering, which yielded three clusters. The first cluster, composed of 29 cities, had real estate properties with the lowest ROI, while the third cluster was composed of 8 cities with the highest ROI. The most significant takeaway here is cluster analysis is effective when applied to real estate sales data. Furthermore, clustering can provide useful results for potential buyers of residential properties in a particular market, such as Ames, Iowa.

## Methods

The techniques we used in our analysis include multiple regression, canonical correlation analysis, correspondence analysis, cluster analysis, and principal component analysis. It was appropriate to use PCA because of the amount of explanatory variables we had. Correspondence analysis was used because there were numerous categorical variables in the dataset that were not going to be used by the other methods. Correspondence analysis allows us to explore the relationships between these variables.To be thorough we ran linear, lasso and ridge regressions in attempt to scope the best fit we could for this dataset.

Professional assessors in Ames conducted an overall quality score of each house in the dataset. The score was a number between one and ten. One of course resulting in the lowest quality and ten resulting in the highest quality. This led us to a visualization with k-means cluster analysis on sale price and the overall quality score. The analysis was a random sample of fifty observations.

Finally, our group decided that Canonical Correlation Analysis (CCA) would be beneficial in understanding how different sets of predictor variables affected each other. CCA is commonly used when the researcher can make two or more distinct groups of variables using the variables in the dataset, and then wishes to see how much correlation is present between the two variables. For our dataset, we split the variables into two different groups; 1. size variables (Y set), and 2. home quality variables (X set). Home quality variables included an overall quality score of the home as well as numeric variables concerning number of rooms, bathrooms, porches, etc. the sale price of each home was also included in this set. The purpose of this split was to see how the size of a home and it's rooms are associated with the quality of the house and it's price. In other terms, this split should show evidence of the association "bigger is better".

## Discussion and Results

K-means score in cluster one; the overall quality scores and sale price for the houses are negative. Also, as for cluster two it is further from zero than cluster one. Interpreting this image from the k-means cluster plot (Figure 2), we see the house in data point seventeen or eight received a good deal with a high quality score by the assessor and a low sale price (from buyer's perspective); while perhaps house in data point 48 or data point 16 paid too much. This visualization will not be explored further to draw a working model due to weak variance explained. However, if investigated further, added data in this set having access to geography contents could determine where a specific neighborhood in Ames contains higher price points.

Principal Component Analysis (PCA) was administered to carry out feature reduction. As our dataset contains 39 explanatory variables, it was deemed important to explore feature reduction in order to reduce complexity in the data. First, we looked at the Cumulative Proportion of Variance explained by the Principal Components to see if we could use PCs in model building. Our results show us that 23 PCs can be used to explain 90% of variance in our data. Furthermore, the cumulative proportion of variance levels off past 30 principal components (Figure 7). Variables contributing most to the first Principal Component- which explained 22% of variance in our data were: Sale Price, Total Above Ground Living Area, Total Basement

Square Footage, Overall Quality, 1st Floor Square Footage, Number of Full Baths, Year Built, and Year Garage Built (Figure 8). In other words, it can be said that the price, square footage, age, and quality of the residential property contributed most to the first Principal Component.

For correspondence analysis, the first pairing we observed was Neighborhood and House.Style.  Looking at the plot (Figure 3), there are some conclusions that we can draw. If arrows were to be drawn from the origin to all the points, we would see that the angles between the 2Story point and the Blueste and Gilbert points are small. Because of this, we can say that the 2 story house style corresponds strongly with the Blueste and Gilbert neighborhoods. If someone was looking for a 2 story house, then they should probably start by looking in the Blueste and Gilbert neighborhoods.

The next correspondence I looked at was between the variables House.Style and Roof.Style.  As can be seen from Figure 4 below, the gable roof style corresponds strongly with the 1.5Unf house style. Meanwhile, the gambrel and mansard roof styles don't really correspond with any house style. Other than the gable roof style, there doesn't seem to be much strong correspondence between other house and roof styles.

The last correspondence that was analyzed was House.Style vs Garage.Type. From Figure 5 below, we can see that there are strong correspondences. If someone was looking for a detached garage type, then there are a few house styles that they should start with. The 1.5Fin, 2.5Unf, and 1.5Unf house styles all correspond strongly with the detached garage. We can also see that the built in garage corresponds strongly with 2 story houses.

We used linear, lasso and ridge regressions in our research in an attempt to capture as many errors as we could. The linear regression provided the best results and accounted for ~78% of the error in our predictions. The most significant coefficient suggested the condition of the home/property was the most significant individual factor of our results. This resulted in a positive $8,236 move in home price for each 1 point increase in the overall condition variable. The second most significant variable was having a full basement bathroom, resulting in a positive $7,580 increase for each bathroom added noted in Regression Coefficients (Figure 6).

After performing canonical correlation analysis (CCA) the correlations or associations between the two groups of variables were clear. First, the communalities in this analysis were rather promising. The smallest communality came from the "Overall Quality" variable with a value of .7221, and the greatest came from "Number of full baths in the basement" with a value of .9962. Communalities represent the fraction of total variance explained for each variable, this split appears to be able to explain each variable's variance very well. Next, the aggregate redundancy coefficients generated from this analysis can provide some insight. From this test, the Y set is able to explain 27% of variance in the X set, and the X set is able to explain 32.7% of variance within the Y set. This is important, because there is a moderate amount of information that can be explained about the opposite group for each of the groups.

Additionally, the Cancor values for the potential variates are very strong values which show a very strong association between size and home quality. The Cancors for the first three significant variates are 0.9135 0.7902 0.7045. These values show that there are strong, positive correlations between size and home quality. After running both a Bartlett's Chi squared test, and a Wilk's Lambda test, there was evidence of 12 different significant variates, although the Cancor values drop off after the third one. Included in the appendix there are helio plots (Figure 10) and the table of loadings for each variable (Figure11). The most influential variables by loading are "Basement Square Feet" and "First Floor Square Feet" for the Y set, and "Sale Price" in the X set.

## Conclusion

One topic that cannot be ignored is that we omitted over half of the variables from our dataset. Without consideration of other research methods, we would encourage future researchers to encompass a wider range of variables than were accessible for this project. Additionally, our best results show there is a strong positive correlation between the size of the house and the overall quality and price of the house. The main driving factors for the size set are the basement square footage, and the first floor area. This is beneficial to users of this model, because anyone who is trying to sell their house can see that a potential addition of a new room would benefit their house's value more than a renovation of their kitchen or bathroom would.

## References

Sarip, A., Hafez, M., & Daud, M. (2016). Application Of Fuzzy Regression Model For Real Estate Price Prediction. *Malaysian Journal Of Computer Science, 29*(1), 15-27. doi:10.22452/mjcs.vol29no1.2

Shi, H., Li, Wanqing, L. (2011). Applying Unascertained Theory, Principal Component Analysis and ACO-based Artificial Neural Networks for Real Estate Price Determination. *Journal of Software, 6*(9), 1672-1679. doi:10.4304/jsw.6.9.1672-1679

Hepşen, Ali & Vatansever, Metin. (2011). Using Hierarchical Clustering Algorithms for Turkish Residential Market. *International Journal of Economics and Finance*. 4. 10.5539/ijef.v4n1p138.

Ames, Iowa. Wikipedia. Retrieved August 18, 2020, from https://en.wikipedia.org/wiki/Ames,_Iowa

## Appendix

Figure 1: Sale Price Histogram

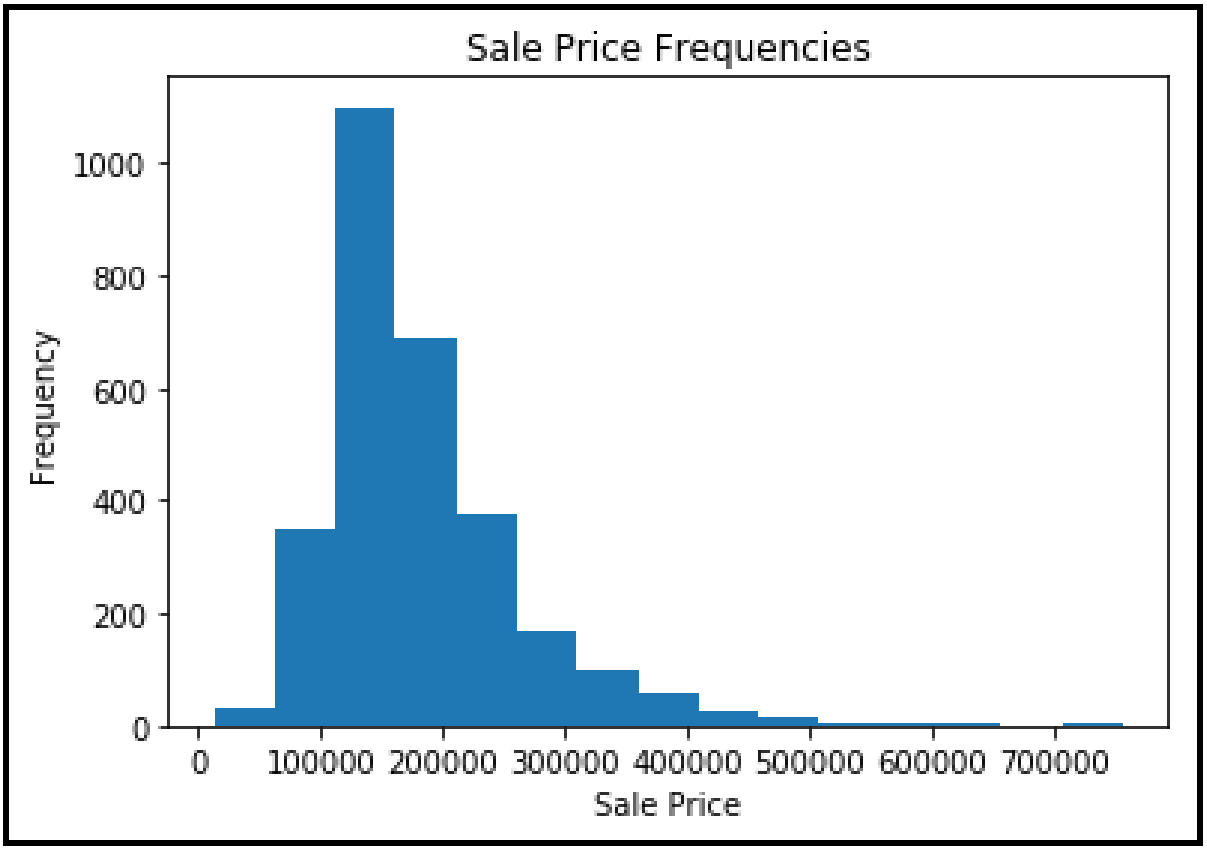


Figure 2: K-means Cluster Plot

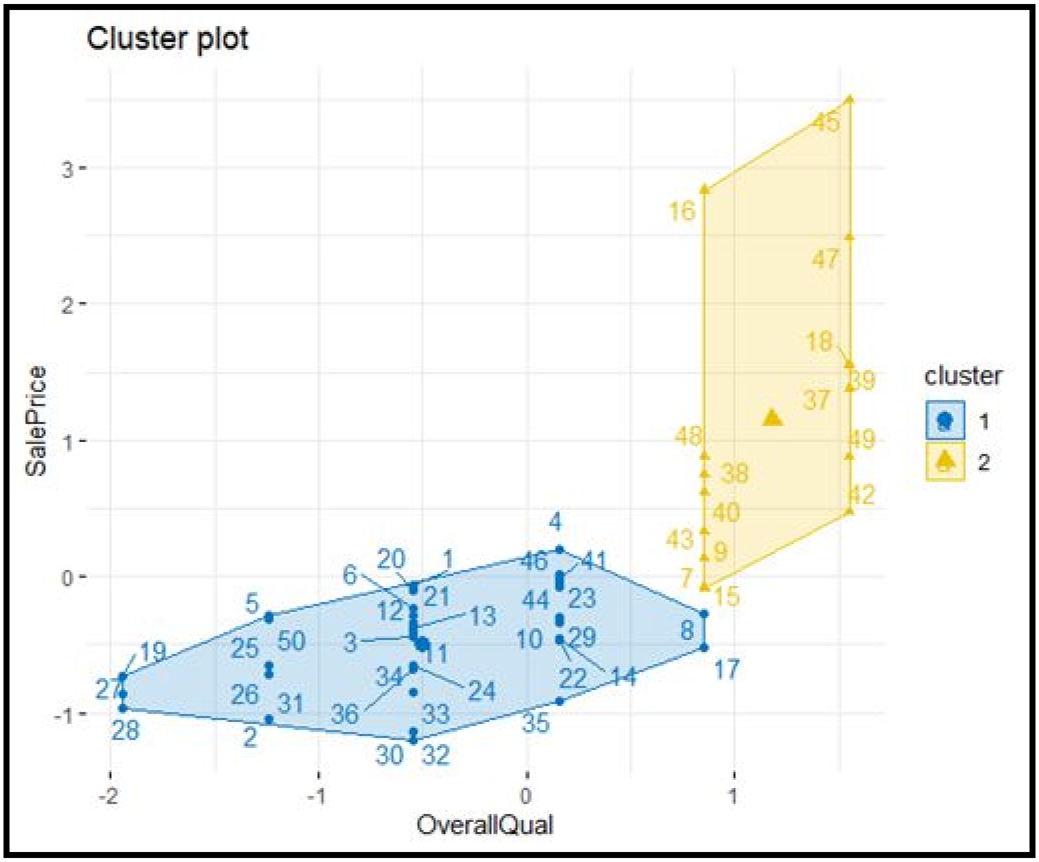


Figure 3: House Style vs Neighborhood

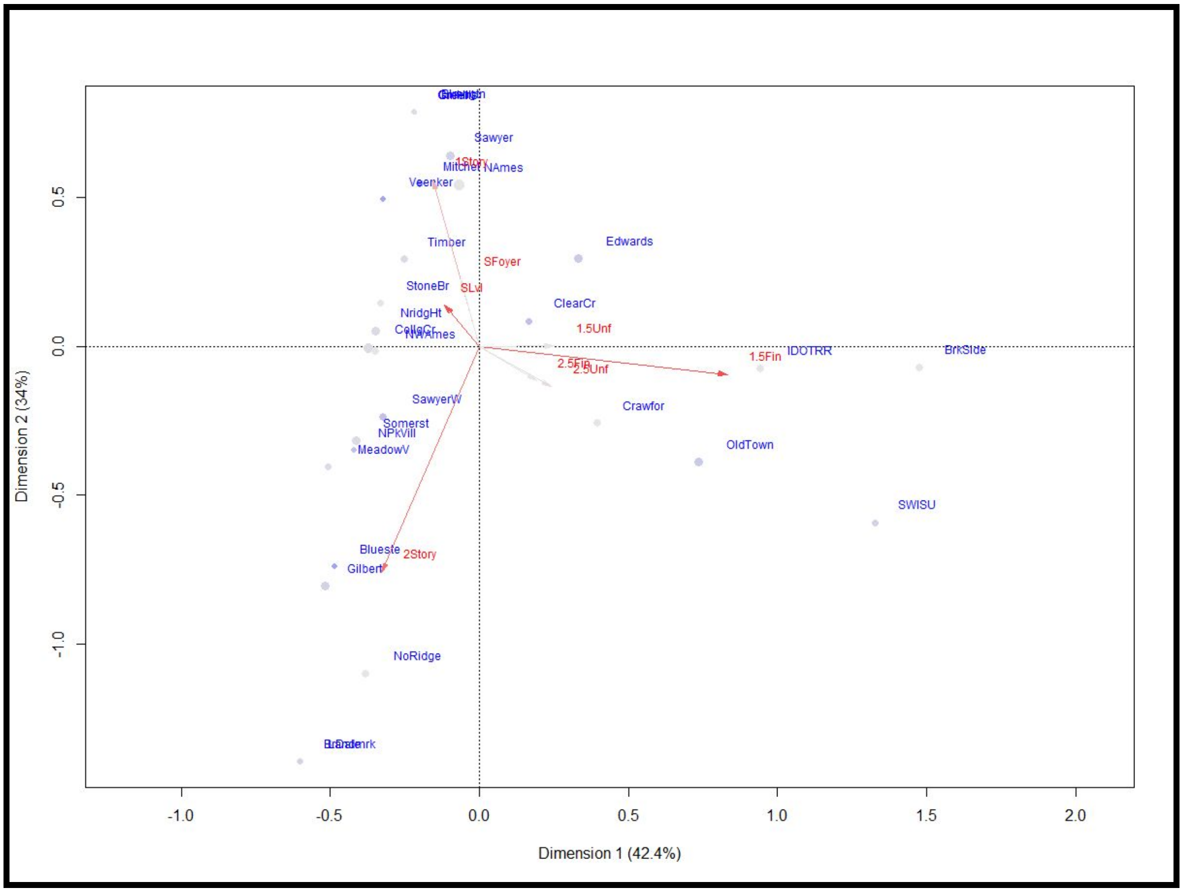


Figure 4: House Style vs Roof Style

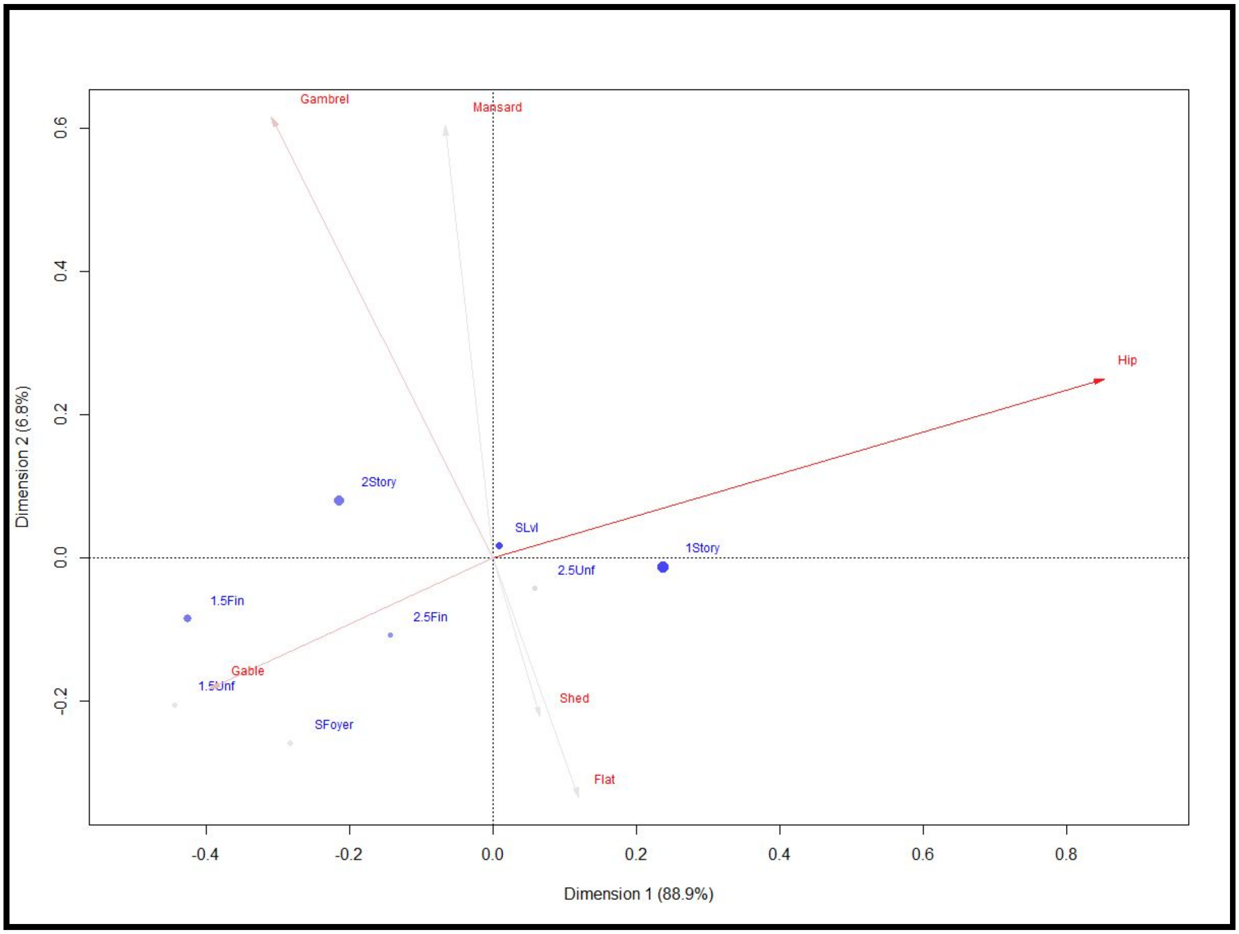


Figure 5: House Style vs Garage Type

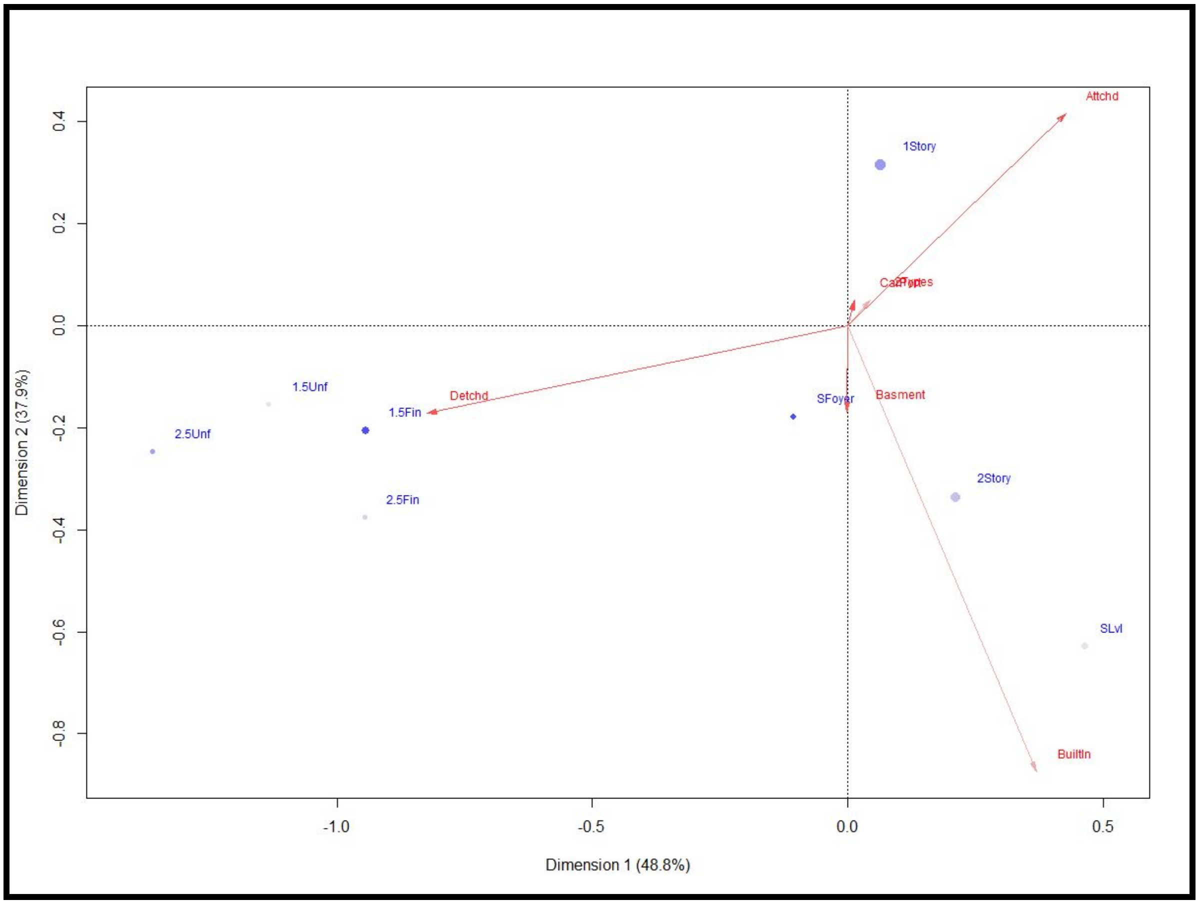
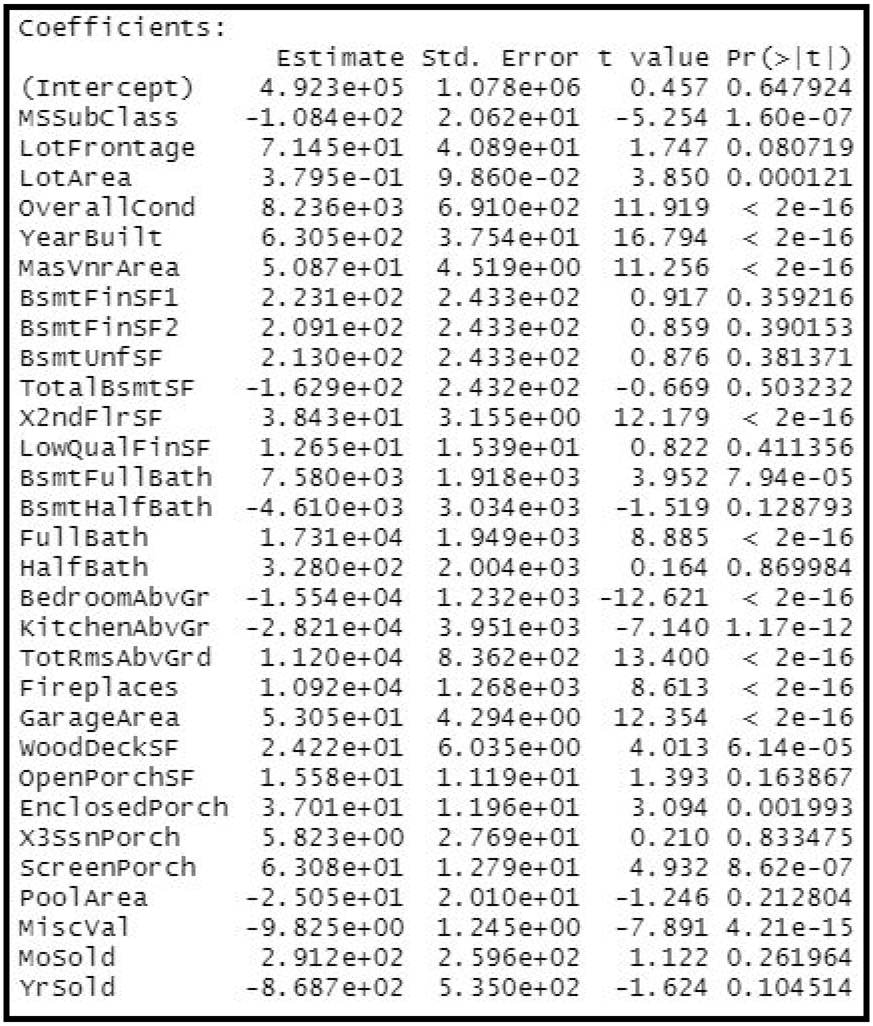


Figure 6: Regression Coefficients



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.923e+05  1.078e+06   0.457 0.647924
MSSubClass    -1.084e+02  2.062e+01  -5.254 1.60e-07
LotFrontage    7.145e+01  4.089e+01   1.747 0.080719
LotArea        3.795e-01  9.860e-02   3.850 0.000121
OverallCond    8.236e+03  6.910e+02  11.919  < 2e-16
YearBuilt      6.305e+02  3.754e+01  16.794  < 2e-16
MasVnrArea     5.087e+01  4.519e+00  11.256  < 2e-16
BsmtFinSF1     2.231e+02  2.433e+02   0.917 0.359216
BsmtFinSF2     2.091e+02  2.433e+02   0.859 0.390153
BsmtUnfSF      2.130e+02  2.433e+02   0.876 0.381371
TotalBsmtSF   -1.629e+02  2.432e+02  -0.669 0.503232
X2ndFlrSF      3.843e+01  3.155e+00  12.179  < 2e-16
LowQualFinSF   1.265e+01  1.539e+01   0.822 0.411356
BsmtFullBath   7.580e+03  1.918e+03   3.952 7.94e-05
BsmtHalfBath  -4.610e+03  3.034e+03  -1.519 0.128793
FullBath       1.731e+04  1.949e+03   8.885  < 2e-16
HalfBath       3.280e+02  2.004e+03   0.164 0.869984
BedroomAbvGr  -1.554e+04  1.232e+03 -12.621  < 2e-16
KitchenAbvGr  -2.821e+04  3.951e+03  -7.140 1.17e-12
TotRmsAbvGrd   1.120e+04  8.362e+02  13.400  < 2e-16
Fireplaces     1.092e+04  1.268e+03   8.613  < 2e-16
GarageArea     5.305e+01  4.294e+00  12.354  < 2e-16
WoodDeckSF     2.422e+01  6.035e+00   4.013 6.14e-05
OpenPorchSF    1.558e+01  1.119e+01   1.393 0.163867
EnclosedPorch  3.701e+01  1.196e+01   3.094 0.001993
X3SsnPorch     5.823e+00  2.769e+01   0.210 0.833475
ScreenPorch    6.308e+01  1.279e+01   4.932 8.62e-07
PoolArea      -2.505e+01  2.010e+01  -1.246 0.212804
MiscVal       -9.825e+00  1.245e+00  -7.891 4.21e-15
MoSold         2.912e+02  2.596e+02   1.122 0.261964
YrSold        -8.687e+02  5.350e+02  -1.624 0.104514
```

Figure 7: Cumulative Proportion of Variance

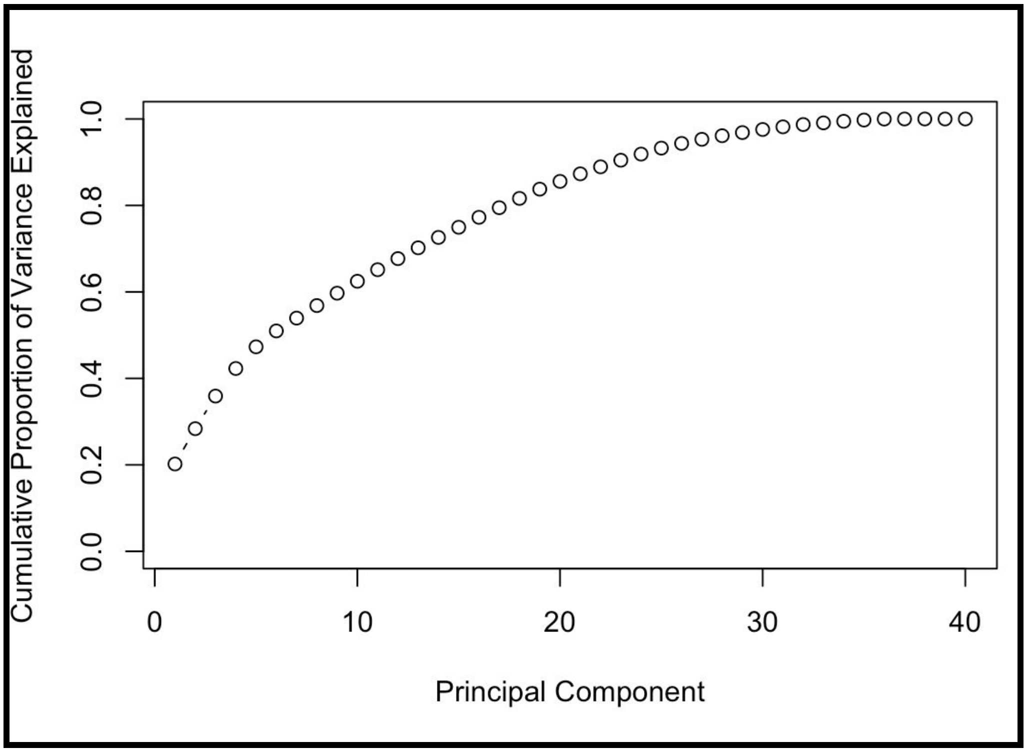


Figure 8:

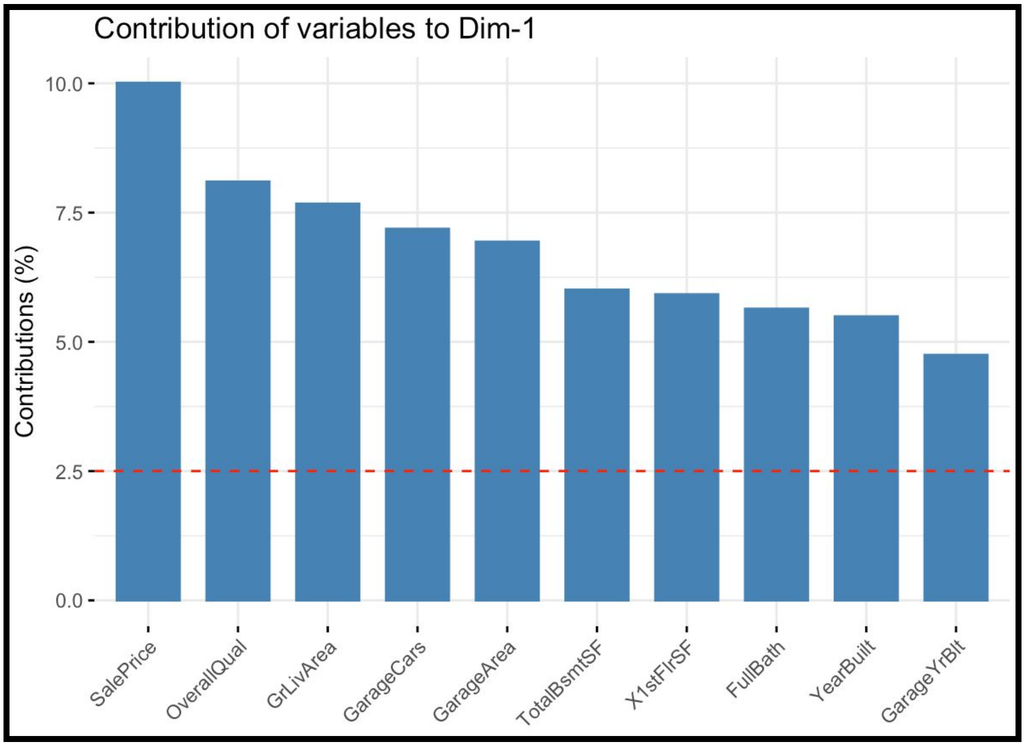
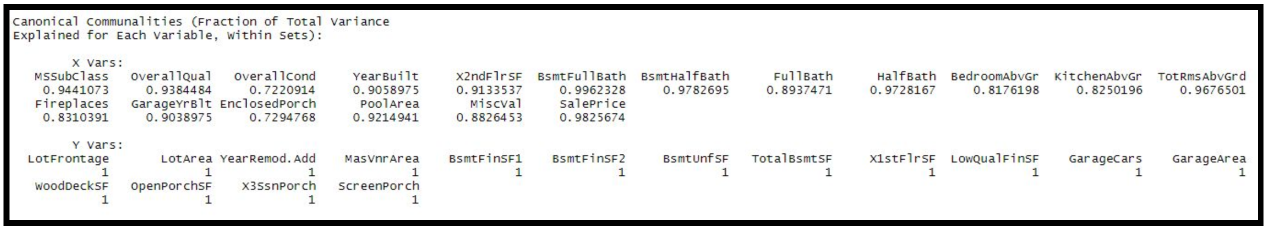


Figure 9: CCA Communalities Table



```
Canonical Communalities (Fraction of Total Variance
Explained for Each Variable, Within Sets):

      X Vars:
   MSSubClass   OverallQual   OverallCond     YearBuilt     X2ndFlrSF  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
    0.9441073     0.9384484     0.7220914     0.9058975     0.9133537     0.9962328     0.9782695     0.8937471     0.9728167     0.8176198     0.8250196     0.9676501
   Fireplaces   GarageYrBlt EnclosedPorch      PoolArea       MiscVal     SalePrice
    0.8310391     0.9038975     0.7294768     0.9214941     0.8826453     0.9825674

      Y Vars:
  LotFrontage       LotArea YearRemod.Add    MasVnrArea    BsmtFinSF1    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF     X1stFlrSF  LowQualFinSF    GarageCars    GarageArea
            1             1             1             1             1             1             1             1             1             1             1             1
   WoodDeckSF   OpenPorchSF    X3SsnPorch   ScreenPorch
            1             1             1             1
```

Figure 10: Helio Plots-First 2 Significant Variates

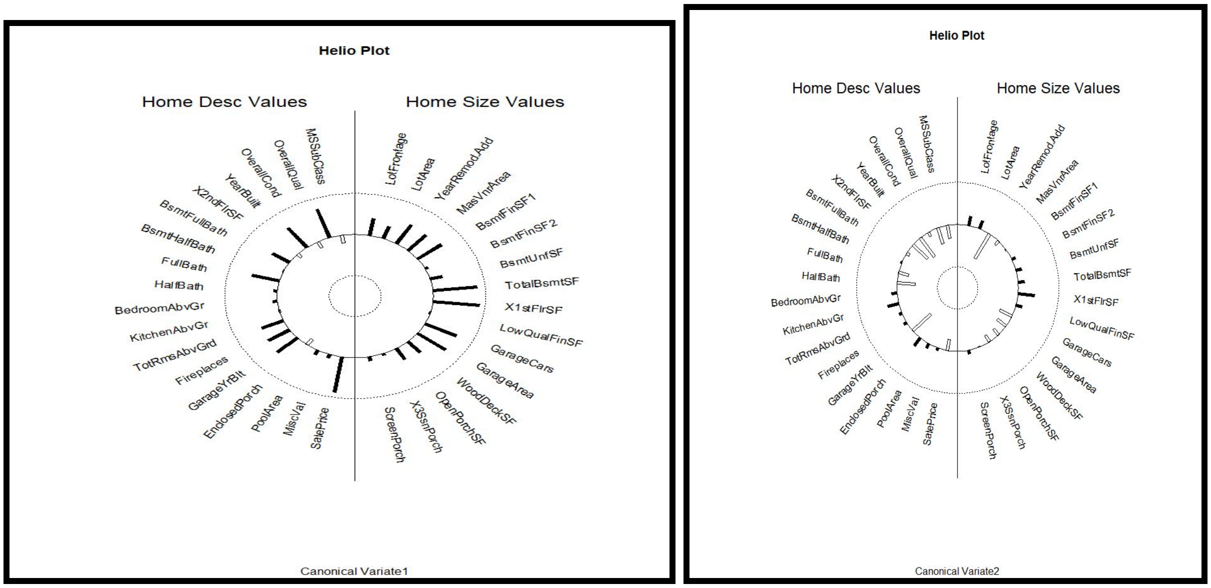
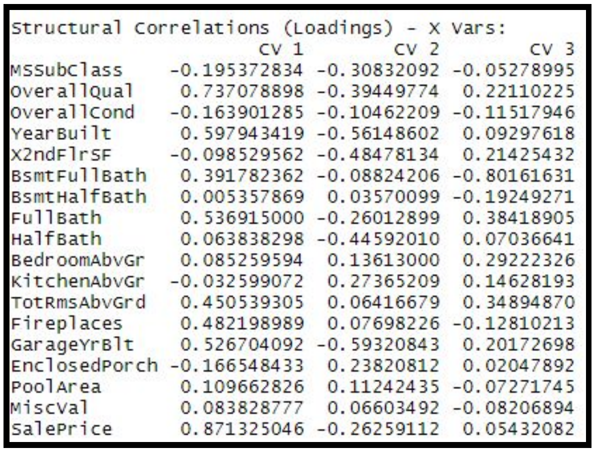
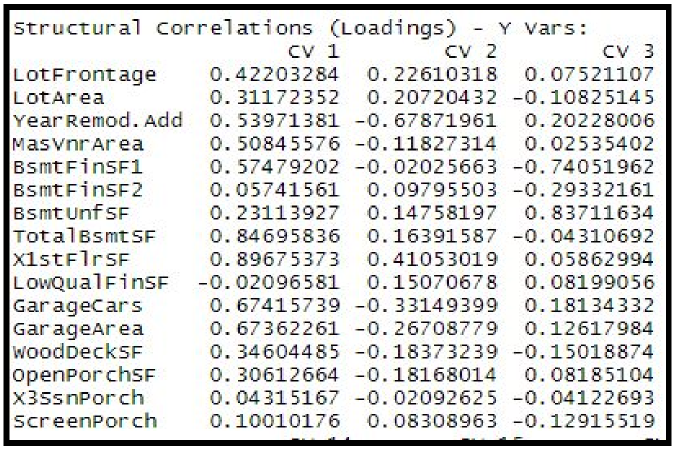


Figure 11: Quality and Size Loadings (Coefficients)



```
Structural Correlations (Loadings) - X Vars:
                     CV 1        CV 2        CV 3
MSSubClass    -0.195372834 -0.30832092 -0.05278995
OverallQual    0.737078898 -0.39449774  0.22110225
OverallCond   -0.163901285 -0.10462209 -0.11517946
YearBuilt      0.597943419 -0.56148602  0.09297618
X2ndFlrSF     -0.098529562 -0.48478134  0.21425432
BsmtFullBath   0.391782362 -0.08824206 -0.80161631
BsmtHalfBath   0.005357869  0.03570099 -0.19249271
FullBath       0.536915000 -0.26012899  0.38418905
HalfBath       0.063838298 -0.44592010  0.07036641
BedroomAbvGr   0.085259594  0.13613000  0.29222326
KitchenAbvGr  -0.032599072  0.27365209  0.14628193
TotRmsAbvGrd   0.450539305  0.06416679  0.34894870
Fireplaces     0.482198989  0.07698226 -0.12810213
GarageYrBlt    0.526704092 -0.59320843  0.20172698
EnclosedPorch -0.166548433  0.23820812  0.02047892
PoolArea       0.109662826  0.11242435 -0.07271745
MiscVal        0.083828777  0.06603492 -0.08206894
SalePrice      0.871325046 -0.26259112  0.05432082
```



```
Structural Correlations (Loadings) - Y Vars:
                     CV 1        CV 2        CV 3
LotFrontage     0.42203284  0.22610318  0.07521107
LotArea         0.31172352  0.20720432 -0.10825145
YearRemod.Add   0.53971381 -0.67871961  0.20228006
MasVnrArea      0.50845576 -0.11827314  0.02535402
BsmtFinSF1      0.57479202 -0.02025663 -0.74051962
BsmtFinSF2      0.05741561  0.09795503 -0.29332161
BsmtUnfSF       0.23113927  0.14758197  0.83711634
TotalBsmtSF     0.84695836  0.16391587 -0.04310692
X1stFlrSF       0.89675373  0.41053019  0.05862994
LowQualFinSF   -0.02096581  0.15070678  0.08199056
GarageCars      0.67415739 -0.33149399  0.18134332
GarageArea      0.67362261 -0.26708779  0.12617984
WoodDeckSF      0.34604485 -0.18373239 -0.15018874
OpenPorchSF     0.30612664 -0.18168014  0.08185104
X3SsnPorch      0.04315167 -0.02092625 -0.04122693
ScreenPorch     0.10010176  0.08308963 -0.12915519
```