Final Project: Predictive models for housing prices in Ames, Iowa

Huong Dang, Japneet Kohli, Zulay Melendez Valle, Shalini Nambiar

George Mason University

# Abstract

Accurate predictions of property prices are useful to homeowners, potential buyers and sellers, real estate agents and financial institutions. This paper describes and evaluates three machine learning models, namely decision tree, random forest, and linear regression , that can predict the prices of residential homes in Ames, Iowa. While there are various methods of evaluating machine learning models, we used only the Root Mean Squared Error (RMSE) metric. Given the nature of the data, we came to the conclusion that a random forest is the best model for making predictions. The most important predictor of price appeared to be the overall quality of the material and finish of the house, across the models followed by footage area of various aspects such as each floor, basement and total footage area.

# The Dataset

**Source:** The dataset was obtained from Kaggle, but was originally compiled by Dean De Cock and published in the Journal of Statistics Education in 2011. It was presented as an alternative to the more popular Boston Housing dataset used by students and professors in the data analytics field (De Cock, D., 2011). Since Kaggle provided this data as part of a competition, it was divided into ‘test’ set and ‘train’ set with the test set not having sales price information. As a result, we have only used 1460 data points from the original dataset.

The dataset has highly detailed information about the houses sold in Ames, from 2006 to 2010. It contains 80 variables (excluding the ID), out of which 34 variables are numerical and the rest are categorical, including ordinal data (for details please see appendix). The variables describe various aspects of the house, ranging from building properties (year built, number of bedrooms, etc.), and space (total ground living area, basement height, basement area, etc.) to location (neighborhood and zoning classification).

Similar to the goals of the competition, our objective are to explore the dataset, try to find a good model for predicting housing prices in Ames, Iowa and identify the best predictive variables.

# Data Exploration and Preprocessing

One of the first steps of data exploration was to look for missing information. While the dataset had numerous variables with NA values, it was discovered that NA stood for “Not Applicable” rather than “Not Available”. Therefore, these were not treated as missing values and were coded as an additional level for categorical variables. For numerical variables they were coded as zero. However, 4 variables – the type of alley access, quality of the pool, quality of the fence and miscellaneous features had 1369,1453, 1179 and 1406 NA values respectively. In a dataset 1460 values, these variables would not add value and were removed from the analysis.

Using a scatter plot (Fig.1) we identified two houses that appeared to be outliers. They had a very large living area (>4,000) and a disproportionately lower price (<200,000). These data points were removed from the dataset.

Data about the year the property was built, the year the garage was built, and the year the remodeling was done, was transformed into age data by subtracting these years from the year the property was sold. All of the year related data was then removed from the dataset.

The ordinal variables related to quality and condition of the houses were converted to numerical form. All numerical data were then scaled to avoid variables with larger values from overshadowing the ones with smaller values, while being used by the models. Principal Component Analysis was then performed on the numerical variables for dimension reduction.

# Predictive Models

**Decision Tree**

We began with a decision tree as it is simple and intuitive and can deal with both categorical and numerical data without any transformations. The first model, using all the variables, identified neighborhood as an important variable. However, this variable has 25 levels which is not ideal. Additional analysis performed on this variable (explained later in the report) failed to provide us a method to group the data. Therefore, we decided to drop this variable, for all of the models going forward. It should be noted that dropping the neighborhood variable impacted the performance of the model by increasing the RMSE.

To prune the tree, we set the complexity parameter at 0, so that it can be plotted to identify the optimal number of levels (Fig.2) and it was observed that having 11 splits had the lowest relative error. However, the best trade-off between the size of the tree and the error seemed to be at the value of complexity parameter as 0.18, where the relative error is close to one standard deviation with the least number of splits for the tree, which in this case was 6.

This model identified the overall quality of the material and finish of the house as the most important variable and was the first split reducing nearly 50% of the relative error. The other variables that the model used were: the square footage of the above ground living area, square footage of the 1st and 2nd floor, square footage of the basement and a variable MSSubClass. Although the MSSubClass is a categorical variable with 16 levels, each category contains information pertaining to the number of storeys, the age and the finished/unfinished status of the building. Therefore we decided to keep this variable as it is packed with information.

**Random Forest**

To improve on the tree model, we used a random forest and since our tree model used 6 variables, we used that as a starting reference point for the mtry (number of variables) tuning parameter of the random forest. As expected, the RMSE improved drastically. We then proceeded to use various different values for the number of variables and the number of trees. It appeared that increasing the number of variables had a much bigger impact in improving the RMSE than increasing the number of trees. The marginal improvement in RMSE also started to fall as the number of variables increased.

The important variables remained largely the same. Overall quality and above ground living area were flipped in importance, with the living area now being most important. Garage area joined the ranks of important variables in this model.

# Linear Regression (Using PCA)

We tried a regression model using only the numerical variables in the form of principal components. Since the scree plot (Fig.3) did not have a clear elbow it was decided to use 23 principal components (less than half the number of numerical variables) since they would account for more than 80% of the variation in the data. We used cross validation to ensure that the model is stable and not too biased, and backward selection to obtain the best mix of variables. Unsurprisingly, this model did not perform as well as the random forest, but was better than the decision tree.

In hopes of improving the model, we created a random forest using just the categorical variables and then identified the top 10 most important variables from that model based on increase in MSE. These categorical variables were then added to the regression model. Surprisingly the model performance decreased and the RMSE went up when we did this.

**Understanding the Principal Components**

To get an understanding of what variables are being used in the model we looked at the loadings on the principal component. The first component has the highest loading for overall quality followed by size of the garage in terms of car capacity. The other important variables are the quality of the material used in the exterior of the house, kitchen and basement, square footage of the garage, area above ground and basement. Most of these variables are the ones that the tree model uses to make its predictions. So essentially this component is separating the big designer mansions from the rest of the properties.

The second component has high negative loadings for multiple variables relating to the basement of the property, its type, condition, exposure and the number of bathrooms it has. At the same time, it also has high positive loadings for variables related to the rooms and the area above ground. So, the properties in this component with a high value would be the ones that don’t have basements or not very big ones.

# Model Comparison and Conclusion

Based on the RMSE metric, the random forest provides us the best result. Table 1 in the appendix shows the RMSE for various models, using different subsets of data and different tuning parameters. The average price of the houses in the dataset is $180,933. The lowest RMSE we obtained was $25,201, 14% of the average price. This number is rather high and indicates that there is a lot more improvement to be made to the model.

The important variables were mostly similar across the models with overall material finish and quality of the property and the square footage of the house above ground being the most important ones across the models.

# Challenges and Additional Efforts

One of the biggest challenges was the mere size of the data. A large number of categorical variables, with large number of levels within each, increased the size of the data manifold. The re-coding of NA values was another challenge. Despite re-coding it as a separate level, it was discovered that random forests treat NA values as missing values even if it is coded as a separate level. Multiple iterations were required to get the data into a format that would be treated correctly by all the models.

In order to understand and group the neighborhood variable, we performed hierarchical clustering and identified 5 clusters based on the dendrogram (Fig.4). We then tried to see if certain neighborhoods belonged primarily to one cluster, as this would justify clubbing neighborhoods based on the cluster they belong to. In other words, say cluster 1 contained houses with multiple floors, superior material, bigger garages and 5 neighborhoods fell primarily in this cluster. We could club the 5 neighborhoods into 1, with the justification that they are the posh neighborhoods. Unfortunately, out of the 25 neighborhoods, 13 had a less than 50% membership to any one particular cluster (Table 2). For example in the Old Town neighborhood 31% of the values belong to cluster 1, 28% to cluster 2, 29% to cluster 3 and 12% to cluster 4. In this case it is hard to say which cluster this neighborhood belongs to. The north Ames neighborhood on the other hand had 65% of the values belong to cluster 2, Sawyer had 66% and Timber 63%. So we could possibly club them together.

# 

# Limitations and Next Steps

**Diagnostics on the regression**

On running the diagnostics on the linear regression model, it was observed that some of the assumptions of linear regression were violated. Fig.5 shows that since the red line does not fall at zero especially on the higher end of prices the assumption about the mean of the error terms being 0 and the standard deviation being constant is violated. Fig.6 shows that the assumption about the normality of the error terms is also violated at the tail ends of the data. From Fig.7 we can see that the variance in the error term is also not constant as the red line is slightly curved and not a flat line. Fig.8 is related to the presence of variables with high leverage and the presence of the Cook’s distance dotted line indicates that there are indeed some variables with high leverage. The data points marked by 1182, 691 and 1046 seem like the main culprits for most of these issues and should be investigated as next steps.

**Possible Next Steps**

This project can only be viewed as a starting level for analyzing the data and there is much more to be done. Since the data exploration was limited, it would be useful to look into this in more detail, as this can reveal certain transformations that can be made, which in turn would improve the models. This can be in the form of combining some of the levels in the categorical variables, such as the MSSubClass that was identified as a significant predictor. A new variable can be created that captures all the data for a particular aspect of the property, such as the basement. So instead of having 10 variables that describe the basement area quality, height, condition etc, everything gets condensed into one variable. Additional analysis of the variables can also improve the data in terms of skewness and outliers, possibly leading to an improvement in the models.

A cluster analysis could be done to subset the data and build a model for each subset. For example, the clusters could be based on the nature of neighborhoods. Housing prices in affluent neighborhoods may depend on different factors than those in comparatively poorer neighborhoods. Moreover, a logistic regression model to predict neighborhood membership may help understand the characteristics of each neighborhood in the dataset. This in turn can help combine the levels into meaningful categories.

If unrestricted by computational power the creation of a loop function to cycle through various values of mtry (number of variables) and ntree (number of trees) tuning variables and recording the RMSE at each combination would be an improvement on the manual method used here and can help find the best combination.

# Appendix

Fig.1: Correlation between GrLivArea and SalePrice

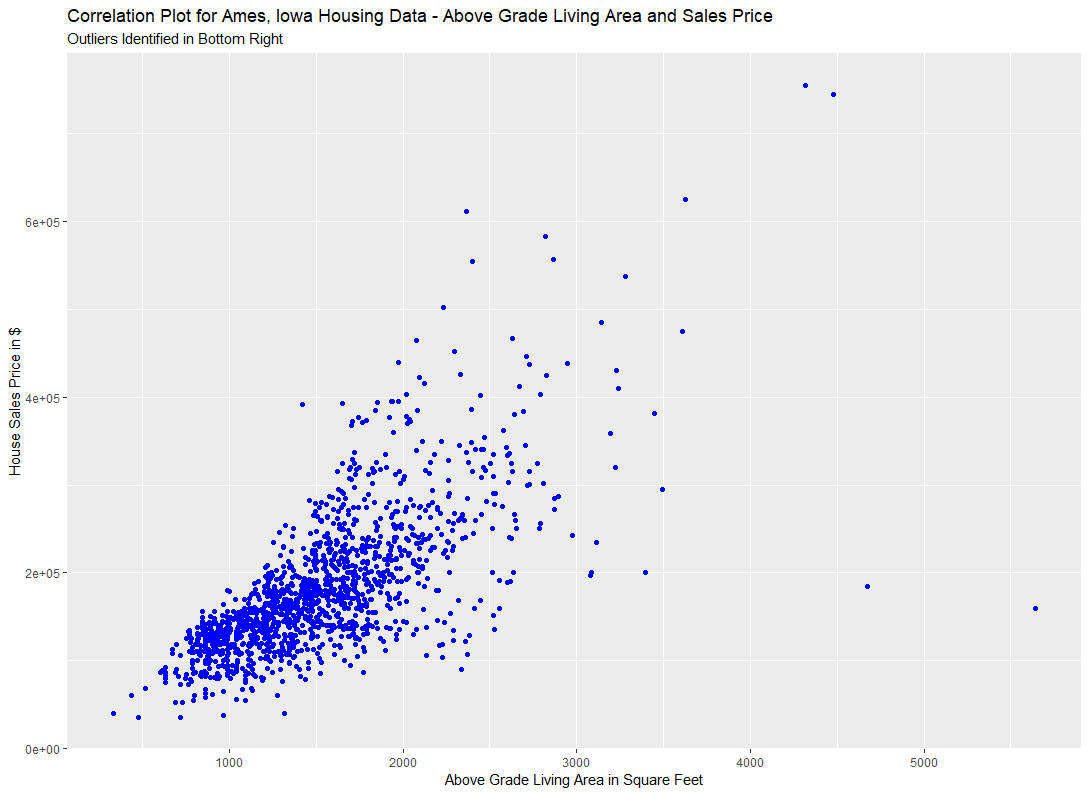


Fig. 2: Identifying the Optimal Size of the Decision Tree

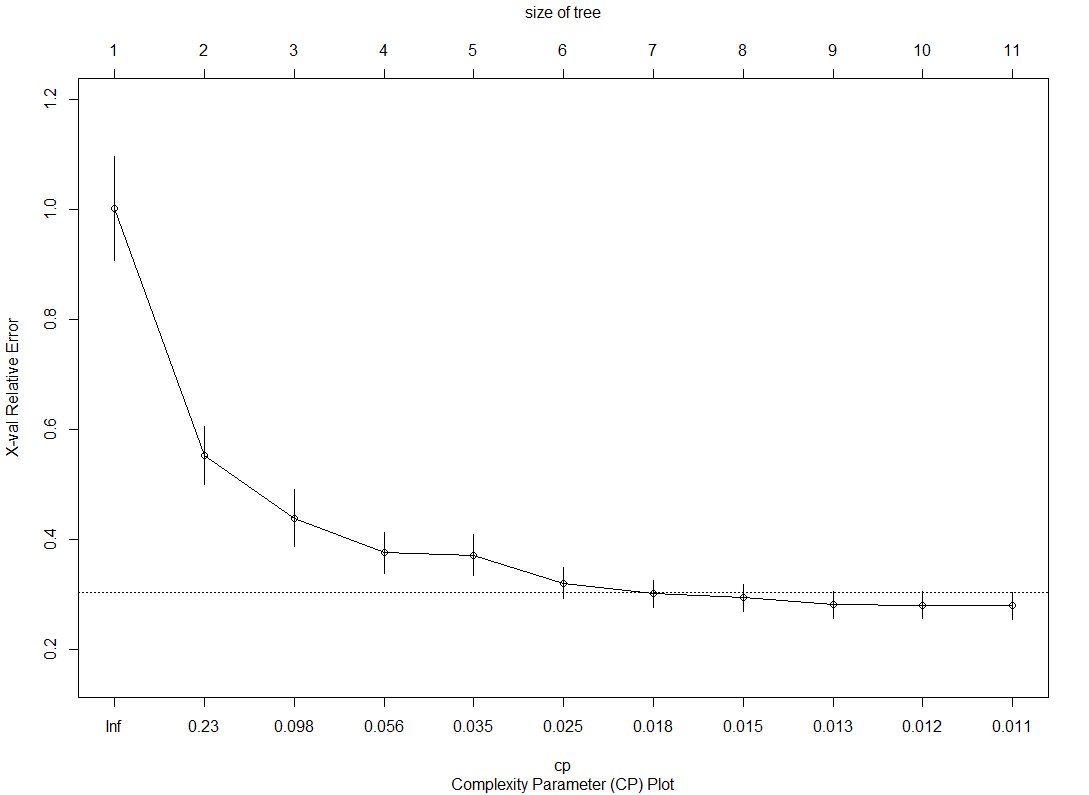


Fig.3: Scree Plot for Principal Component Analysis

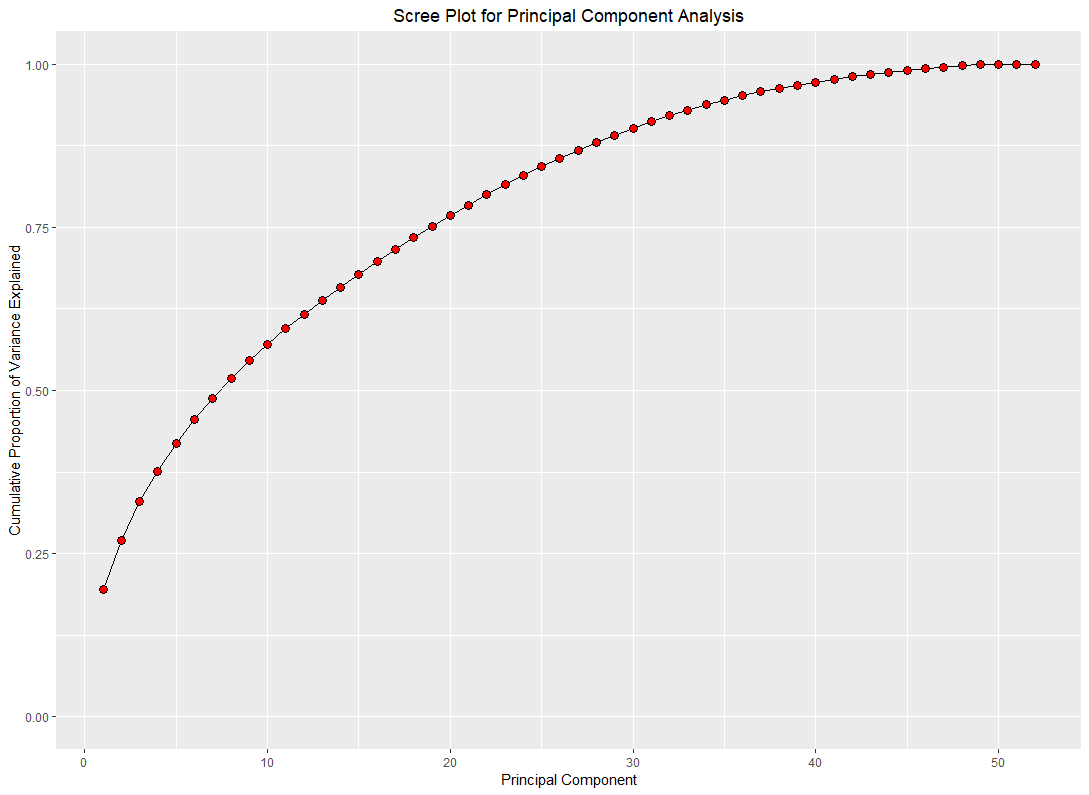


Fig.4: Identifying Clusters in the Data

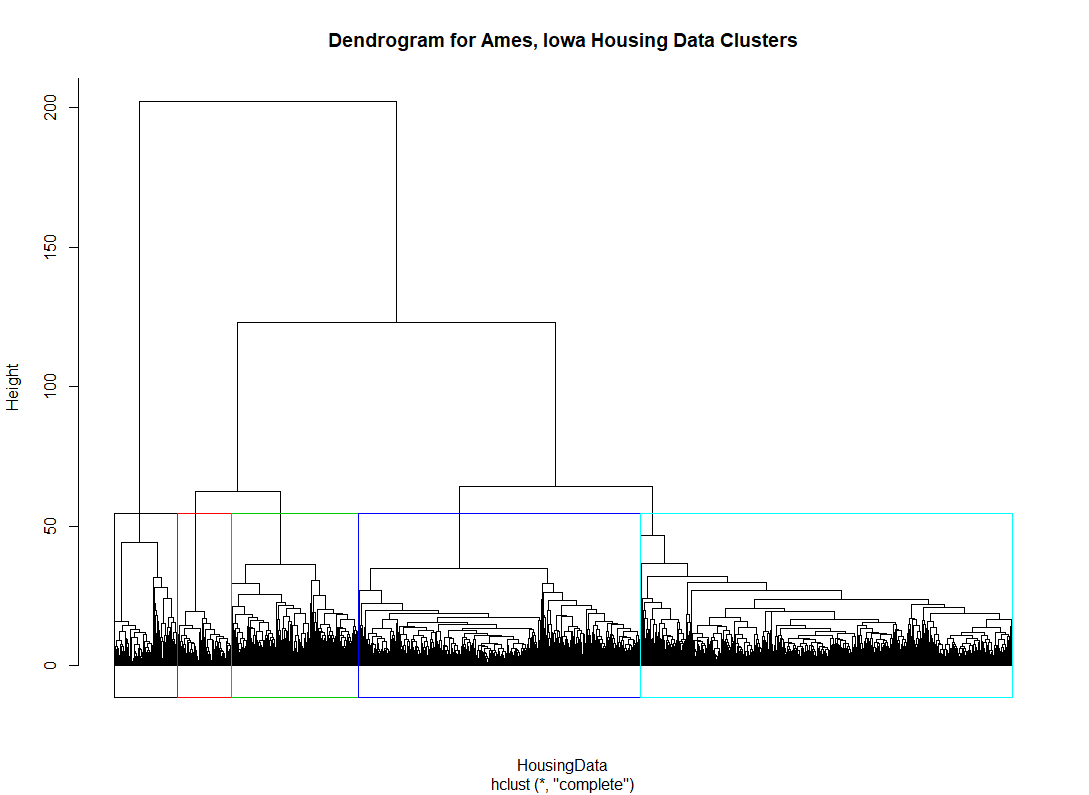


Fig.5: Testing the Mean and Standard Deviation Assumptions of Error Terms

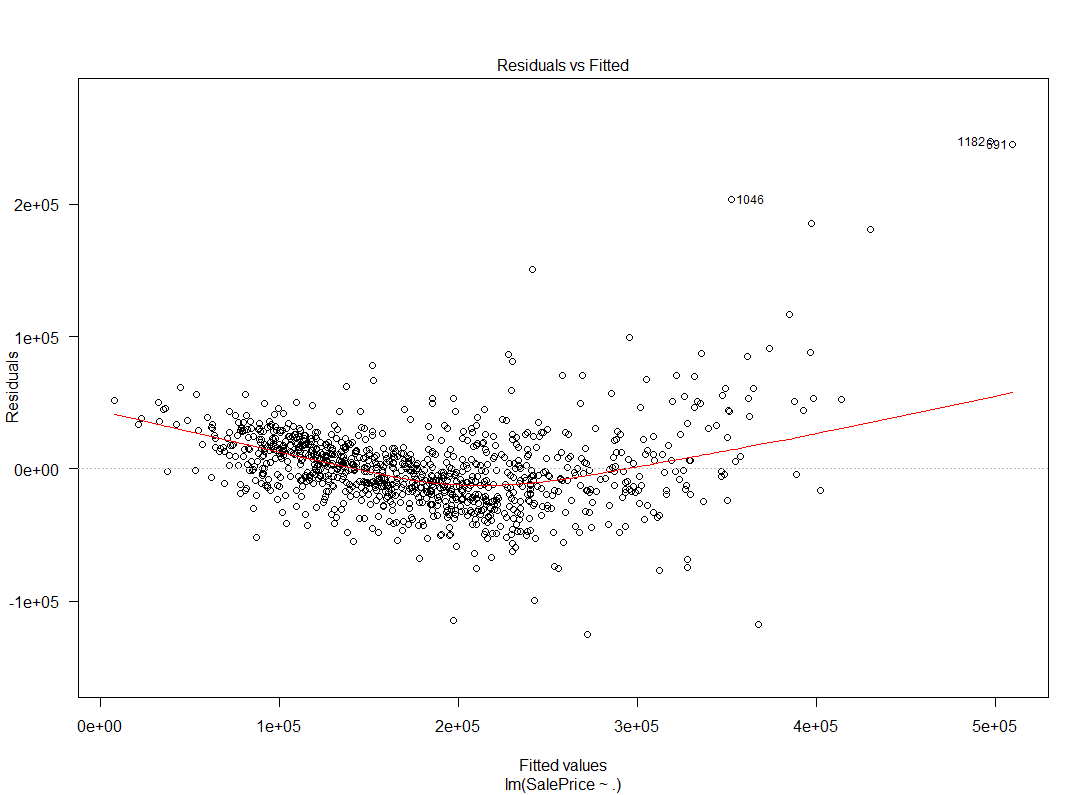


Fig.6: Testing the Normality of the Error Terms

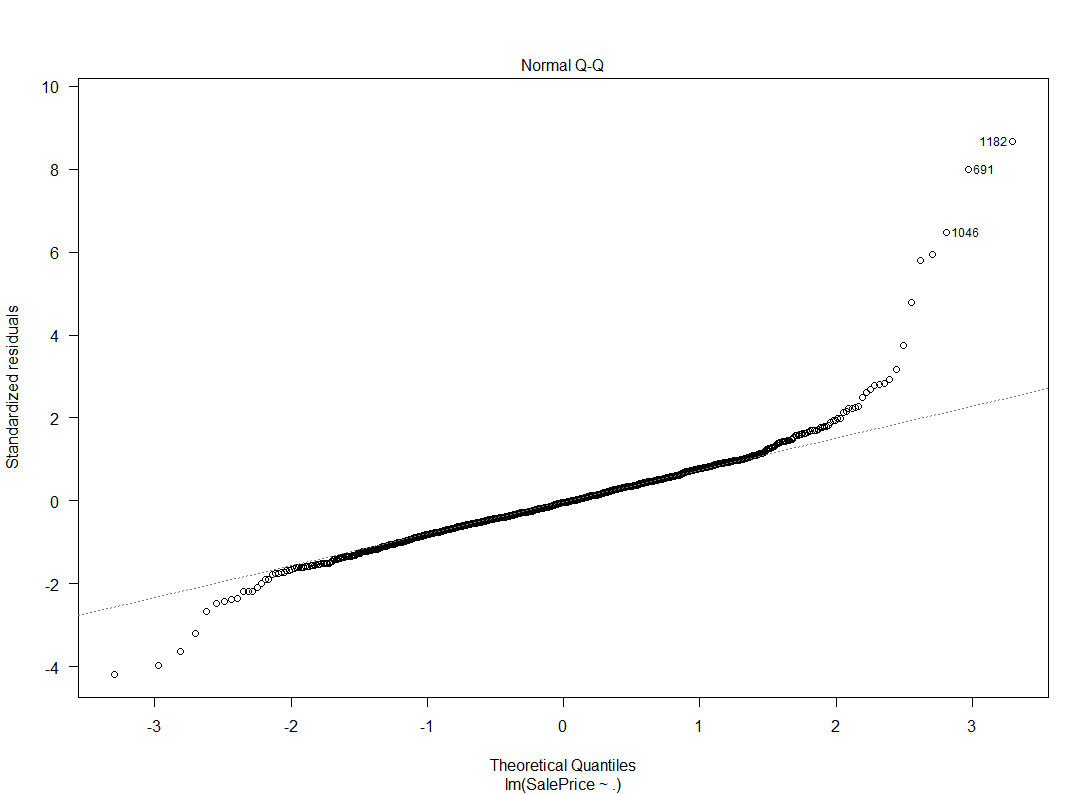


Fig.7: Testing for Heteroscedasticity

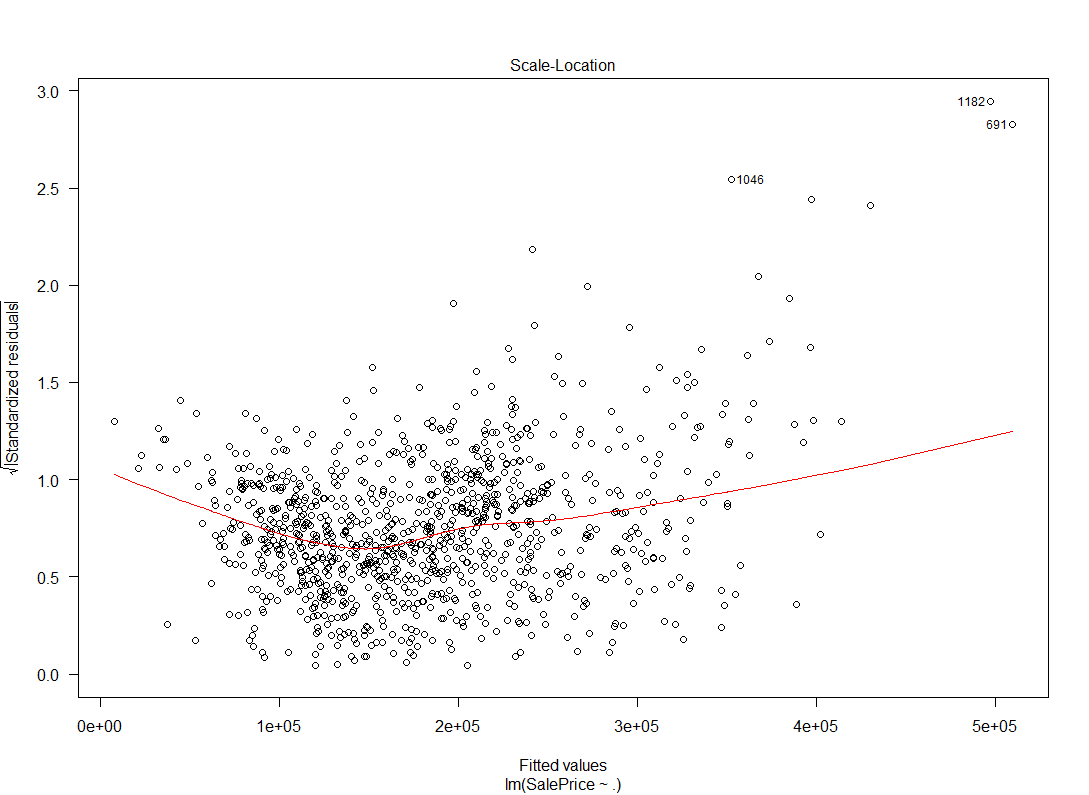


Fig. 8: Testing for Data Points with High Leverage

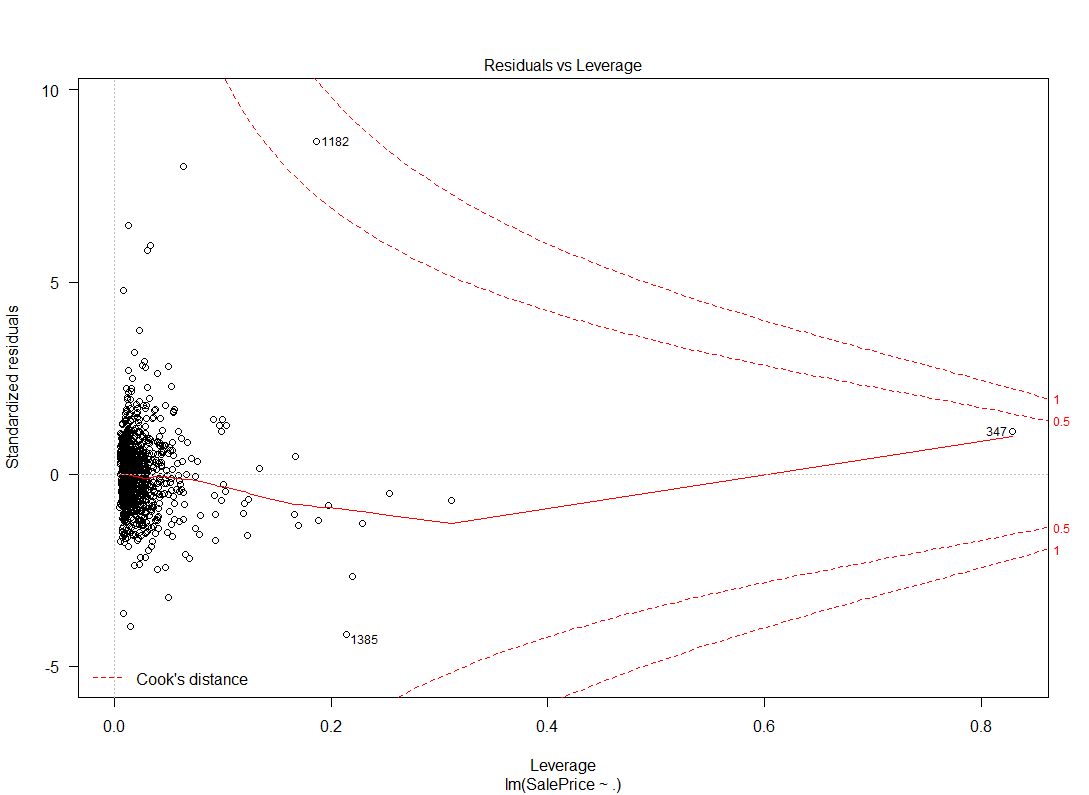


Fig. 9: Decision Tree Using All Variables

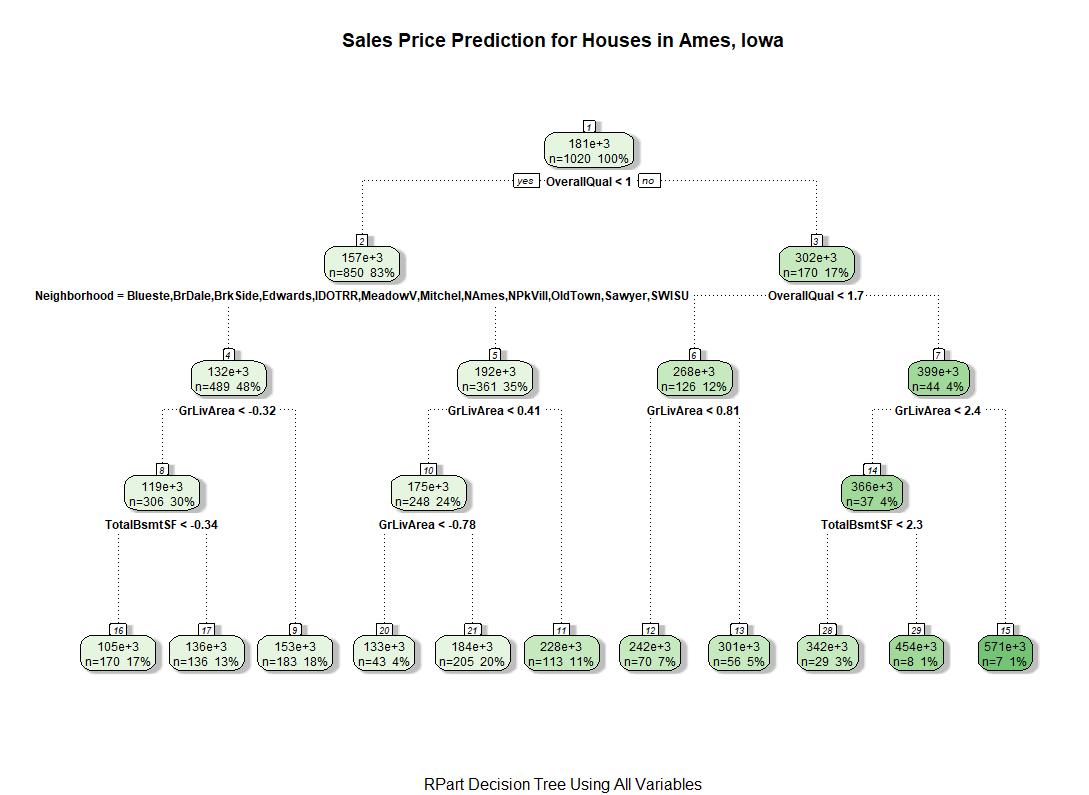


Fig. 10: Pruned Decision Tree With 6 Splits

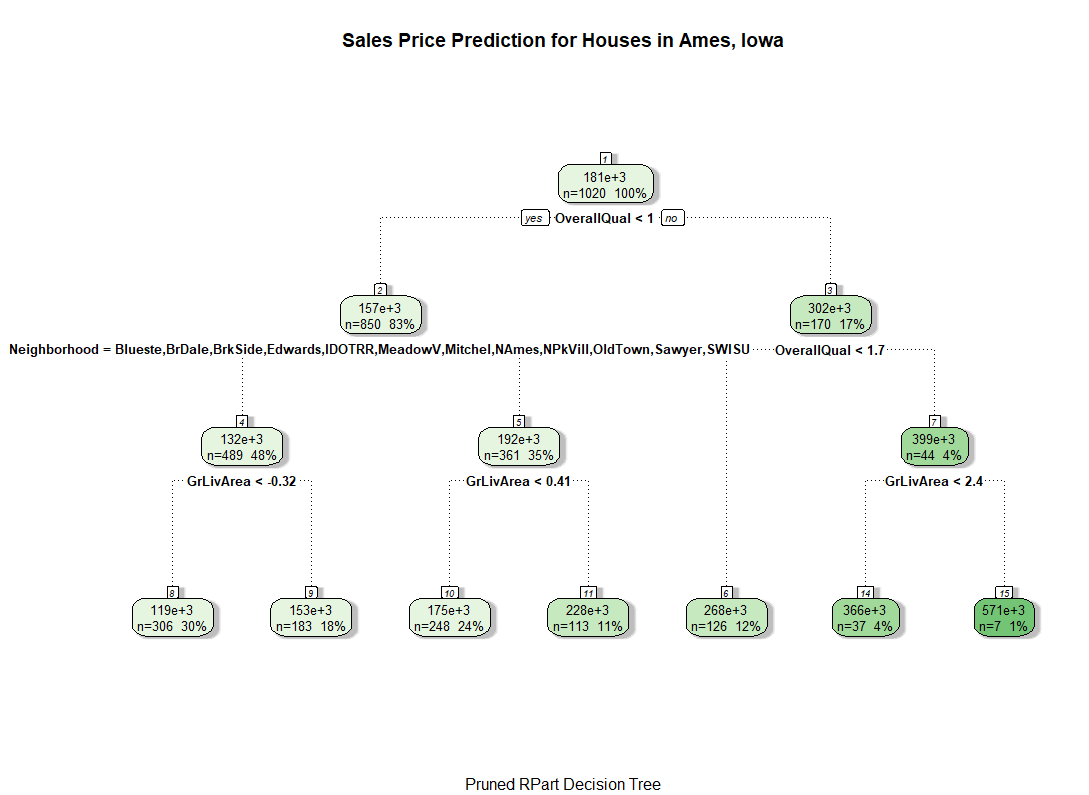


Fig. 11: Pruned Decision Tree Without Variable Neighborhood

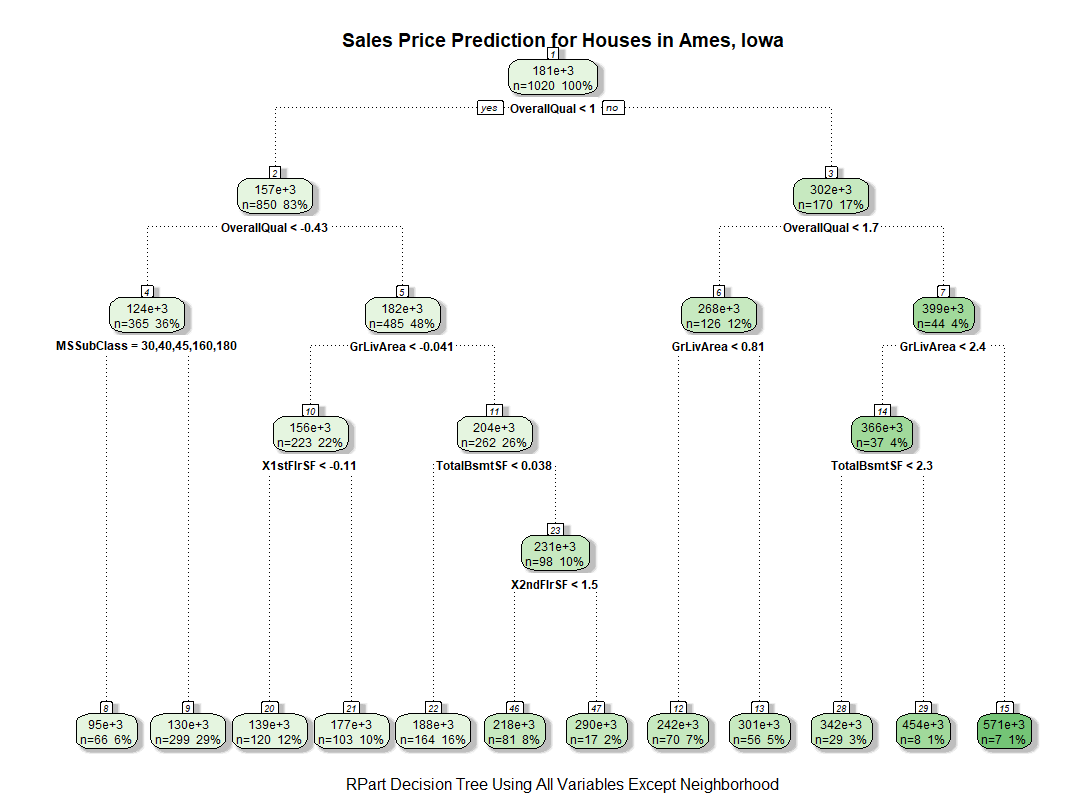
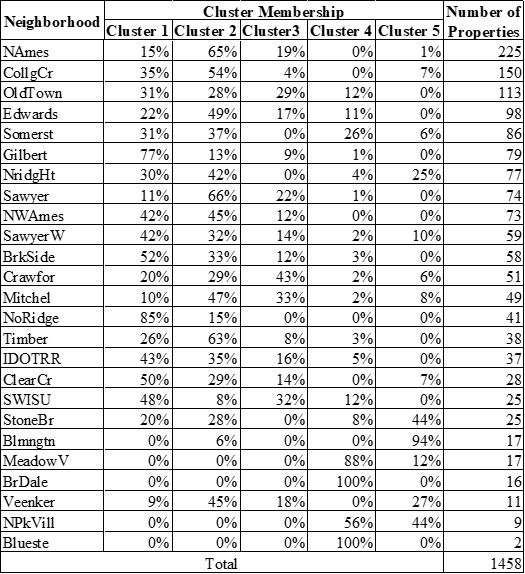


Table 1: Comparing the model performance

|  |  |
| --- | --- |
| RMSE Value | Model Details |
| $38,134 | Tree, with 10 splits using all the data |
| $42,524 | Tree, with 6 splits using all the data |
| $40,721 | Tree, with 11 splits excluding the neighborhood variable |
| $27,086 | Forest, mtry=6, ntree=500 |
| $26,150 | Forest, mtry=9, ntree=1000 |
| $26,288 | Forest, mtry=10, ntree=1000 |
| $25,894 | Forest, mtry=12, ntree=500 |
| $26,030 | Forest, mtry=12, ntree=2000 |
| $25,571 | Forest, mtry=15, ntree=500 |
| $25,451 | Forest, mtry=15, ntree=2000 |
| $25,201 | Forest, mtry=16, ntree=3000 |
| $32,986 | Regression, 23 principal components, cross validated backward selection |
| $54,299 | Forest, mtry=10, ntree=500, only categorical variables |
| $34,654 | Regression, 20 principal components, 10 categorical variables |

Table 2: Membership of neighborhoods to each cluster



Data Dictionary

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| MSSubClass | The building class |
| MSZoning | The general zoning classification |
| LotFrontage | Linear feet of street connected to property |
| LotArea | Lot size in square feet |
| Street | Type of road access |
| Alley | Type of alley access |
| LotShape | General shape of property |
| LandContour | Flatness of the property |
| Utilities | Type of utilities available |
| LotConfig | Lot configuration |
| LandSlope | Slope of property |
| Neighborhood | Physical locations within Ames city limits |
| Condition1 | Proximity to main road or railroad |
| Condition2 | Proximity to main road or railroad (if a second is present) |
| BldgType | Type of dwelling |
| HouseStyle | Style of dwelling |
| OverallQual | Overall material and finish quality |
| OverallCond | Overall condition rating |
| YearBuilt | Original construction date |
| YearRemodAdd | Remodel date |
| RoofStyle | Type of roof |
| RoofMatl | Roof material |
| Exterior1st | Exterior covering on house |
| Exterior2nd | Exterior covering on house (if more than one material) |
| MasVnrType | Masonry veneer type |
| MasVnrArea | Masonry veneer area in square feet |
| ExterQual | Exterior material quality |
| ExterCond | Present condition of the material on the exterior |
| Foundation | Type of foundation |
| BsmtQual | Height of the basement |
| BsmtCond | General condition of the basement |
| BsmtExposure | Walkout or garden level basement walls |
| BsmtFinType1 | Quality of basement finished area |
| BsmtFinSF1 | Type 1 finished square feet |
| BsmtFinType2 | Quality of second finished area (if present) |
| BsmtFinSF2 | Type 2 finished square feet |
| BsmtUnfSF | Unfinished square feet of basement area |
| TotalBsmtSF | Total square feet of basement area |
| Heating | Type of heating |
| HeatingQC | Heating quality and condition |
| CentralAir | Central air conditioning |
| Electrical | Electrical system |
| 1stFlrSF | First Floor square feet |
| 2ndFlrSF | Second floor square feet |
| LowQualFinSF | Low quality finished square feet (all floors) |
| GrLivArea | Above grade (ground) living area square feet |
| BsmtFullBath | Basement full bathrooms |
| BsmtHalfBath | Basement half bathrooms |
| FullBath | Full bathrooms above grade |
| HalfBath | Half baths above grade |
| Bedroom | Number of bedrooms above basement level |
| Kitchen | Number of kitchens |
| KitchenQual | Kitchen quality |
| TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |
| Functional | Home functionality rating |
| Fireplaces | Number of fireplaces |
| FireplaceQu | Fireplace quality |
| GarageType | Garage location |
| GarageYrBlt | Year garage was built |
| GarageFinish | Interior finish of the garage |
| GarageCars | Size of garage in car capacity |
| GarageArea | Size of garage in square feet |
| GarageQual | Garage quality |
| GarageCond | Garage condition |
| PavedDrive | Paved driveway |
| WoodDeckSF | Wood deck area in square feet |
| OpenPorchSF | Open porch area in square feet |
| EnclosedPorch | Enclosed porch area in square feet |
| 3SsnPorch | Three season porch area in square feet |
| ScreenPorch | Screen porch area in square feet |
| PoolArea | Pool area in square feet |
| PoolQC | Pool quality |
| Fence | Fence quality |
| MiscFeature | Miscellaneous feature not covered in other categories |
| MiscVal | $Value of miscellaneous feature |
| MoSold | Month Sold |
| YrSold | Year Sold |
| SaleType | Type of sale |
| SaleCondition | Condition of sale |

1. Details of the levels in categorical data

1MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

Street: Type of road access to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

RoofStyle: Type of roof

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (&lt;70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

Heating: Type of heating

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes

References

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester

regression project. Journal of Statistics Education, 19(3).

Fox, J. & Weisberg, S. (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL:<http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

Grothendieck, G. (2017). sqldf: Manipulate R Data Frames Using SQL. R package version 0.4-11. [https://CRAN.R-project.org/package=sqldf](https://cran.r-project.org/package=sqldf)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). ISLR: Data for an Introduction to

Statistical Learning with Applications in R. R package version 1.2.

[https://CRAN.R-project.org/package=ISLR](https://cran.r-project.org/package=ISLR)

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. RNews 2(3),

18--22.

Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A.,

Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A.,

Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2019). caret: Classification and Regression

Training. R package version 6.0-84. [https://CRAN.R-project.org/package=caret](https://cran.r-project.org/package=caret)

Makowski, D. (2018). The Psycho Package: An Efficient and Publishing-Oriented Workflow for Psychological Science. Journal of Open Source Software, 3(22), 470. Available from

<https://github.com/neuropsychology/psycho.R>

Milborrow, S. (2019). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R

package version 3.0.7. [https://CRAN.R-project.org/package=rpart.plot](https://cran.r-project.org/package=rpart.plot)

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.

[https://CRAN.R-project.org/package=RColorBrewer](https://cran.r-project.org/package=RColorBrewer)

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition.

Springer, New York. ISBN 0-387-95457-0

Wickham., H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham., H (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. [https://CRAN.R-project.org/package=tidyverse](https://cran.r-project.org/package=tidyverse)

Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for

Knowledge Discovery, Use R!, Springer.

Therneau, T. & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. R

package version 4.1-13. [https://CRAN.R-project.org/package=rpart](https://cran.r-project.org/package=rpart)