Analysis Of Housing Data

Travis Barton, Eddie Gonzalez

4/30/18

1 Purpose

The purpose of this project is to explore the best model for explaining house sale prices in a kaggle data science competition. The dataset contains information on houses for sale in Ames, Iowa and has 79 variables. One of the variables was deemed too sparse for use, so after removing the predictor variable (Sales Price), we are left with 36 qualitative and 41 quantitative variables to work with. To perform our analysis we used the R packages: FactoMineR, Readr, PCAmixdata, and GGPlot2.

2 Data

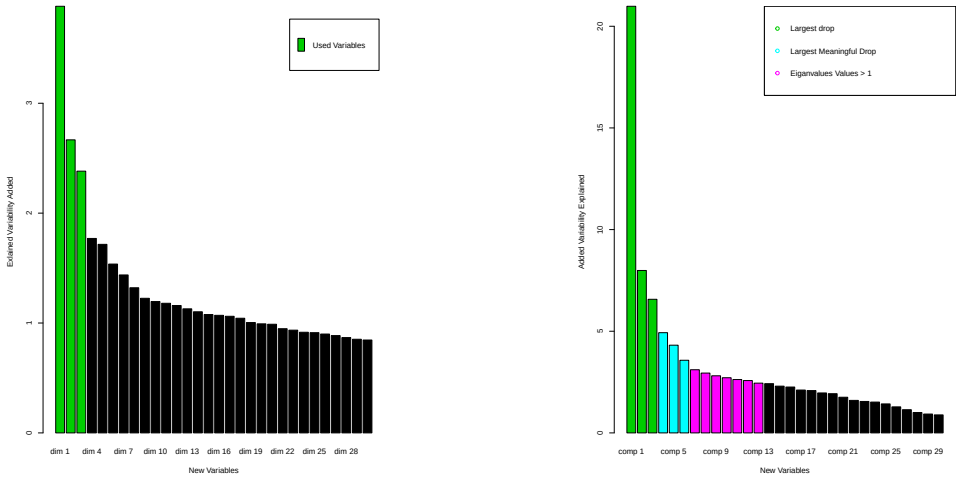
The 36 qualitative variables (tables on appendix) consist of a variety of factors describing the house's interior, exterior and surrounding neighborhood. Since there is a mixture of characters and numbers to describe the different factor levels, all qualitative variables were changed to letters/characters and all quantitative variables were changed to numbers/scales. An example of this lies in the variable MSSubClass (identifying the type of building involved in the sale), which was originally listed as a factor of numbers (20 = 1-Story 1946 style house or newer, 30 = 1-Story 1945 house or older, ...), and was changed to a list of letters instead with 20 = "A", 30 = "B", etc. This variable conversion became a problem because during our analysis we chose to perform Principal Component Analysis (PCA) on the quantitative variables, and Multiple Components Analysis (MCA) on the qualitative variables in order to reduce the number of predictors in our various models. During this endeavor, we used a function called "splitmix" from the package PCAmixdata that splits a data set into it's qualitative and quantitative variables. In its input, it required that qualitative variables only have characters and quantitative variables only have numbers.

The same problem of variable-data type conversion came up during our examination of the quantitative variables. When examining many of the ordinal variables like ExterQual (which evaluates the quality of the material on the outside of the house on a scale from

excellent to poor represented by "Ex" and "Po") we noticed that, while they do have a ranking, they also are listed as character inputs.

3 Component Analysis

Once the data was separated into its quantitative and qualitative components, we ran our PCA and MCA. PCA is a method of dimensionality reduction that takes advantage of the correlation between the given variables to create new variables that are linear combinations of the old ones. This has several perks, including the fact that the new variables are guaranteed to be linearly independent of each other and that, often, a fraction of the new variables can explain almost all of the variability in that data. MCA is the same general principle except for qualitative variables.



(a) (b)

Figure 1: PCA vs MCA variance contributions

After performing MCA, we made the decision to cut off the number of new variables at three, as after three the amount of explained variability that we gained by adding another did not outweigh the cost of making our model less parsimonious. This decision can be seen as a sudden drop in the explained variability (Fig 1a). When it came to PCA, the number of variables that we should keep was not clear. The graph of added variability drops off in two places, once after the third variable, and again after the sixth (Fig 1b). What is more, there is a general rule of thumb for PCA that one should never keep a variable with an eigenvalue lower than 1 (as it is generally assumed that the original data could have explained as much or more variability than an eigan valued variable with value

1) and our eigenvalues have values above 1 as far down as the 13th variable. As a result, we performed many analysis with all three variants and compared their performances.

4 Analysis

In order to establish optimal model adequacy, we tested our models on two goals. One where we attempted to predict the exact sale price of the home with linear regression techniques, and the other where we attempted to predict whether a home would sell above or below the average price of homes using logistic regression. During the analysis, we discovered that Sale Price did not follow a normal distribution. Since Sale Price does not follow a normal distribution, we modified Sale Price with a log transformation. After the transformation, the new Sale Price followed a normal distribution, and thus our assumptions for linear regression were met. The models, R squared/accuracy values, RMSE values and Ranks are tabled below.

Basic Linear Model	R^2	RMSE	Rank
$Y = PCA_1 + PCA_2 + PCA_3 + MCA_1 + MCA_3$	.7903	36892.7	1
$Y = PCA_1 + PCA_2 + PCA_3 + ... + PCA_6 + MCA_1 + MCA_2 + MCA_3$	.802	36113.3	2
$Y = PCA_1 + PCA_2 + ... + PCA_{13} + MCA_1 + MCA_2 + MCA_3$	.8116	36154.7	3

Log Transformed Model	R^2	RMSE	Rank
$log(Y) = PCA_1 + PCA_2 + PCA_3 + MCA_1$	.8339	.1632	1
$log(Y) = PCA_1 + PCA_2 + PCA_3 + ... + PCA_6 + MCA_1 + MCA_2$	.8473	.1562	3
$log(Y) = PCA_1 + ... + PCA_6 + PCA_8 + ... + PCA_{10} + PCA_{12} + PCA_{13} + MCA_1 + MCA_2$	.8534	.1521	2

Logistic Model	Accuracy	AIC	Rank
$Y_{binary} = PCA_1 + MCA_1 + MCA_3$	.9233	608	3
$Y_{binary} = PCA_1 + PCA_3 + PCA_5 + MCA_3$	.9322	578.37	2
$Y_{binary} = PCA_1 + PCA_3 + PCA_4 + PCA_5 + PCA_8 + PCA_{13}$	.9369	531.04	1

5 Conclusion

Based on the results of the Basic Linear Regression model, we chose the first model as the best model over the other three. The first model has far fewer variables than the other linear regression models and only has a slightly lower R^2 value. For the Log Transformed models, we again choose the simplest model because the amount of increase in R^2 that more variables provided did not out way the cost of adding complexity. Since Sale Price does not follow a normal distribution, using a basic linear regression model would not be the best choice to use for analyzing and predicting the housing prices of Ames, Iowa. The basic linear regression models also tend to overestimate the Sale Price of the houses as sale

price increases. Our final choice of model for predicting housing prices was the simplet model of the log transformation options. For the Logistic Regression model, we chose the model number three because this model had the highest accuracy and the lowest AIC value.

6 Appendix

MSSubClass: 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

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

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

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

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

MasVnrArea: Masonry veneer area in square feet

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 (<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

BsmtFinSF1: Type 1 finished square feet

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

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

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

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: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

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

Fireplaces: Number of fireplaces

```
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
       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

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

       Fin        Finished
       RFn        Rough Finished
       Unf        Unfinished
       NA        No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

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

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

       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

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

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 cond
       Family        Sale between family members
       Partial        Home was not completed when last assessed (associated with New Homes)
```