Categorical Analysis of Housing Prices

Eduardo Gonzalez, Milica Miskovic, Travis Barton

May 14, 2018

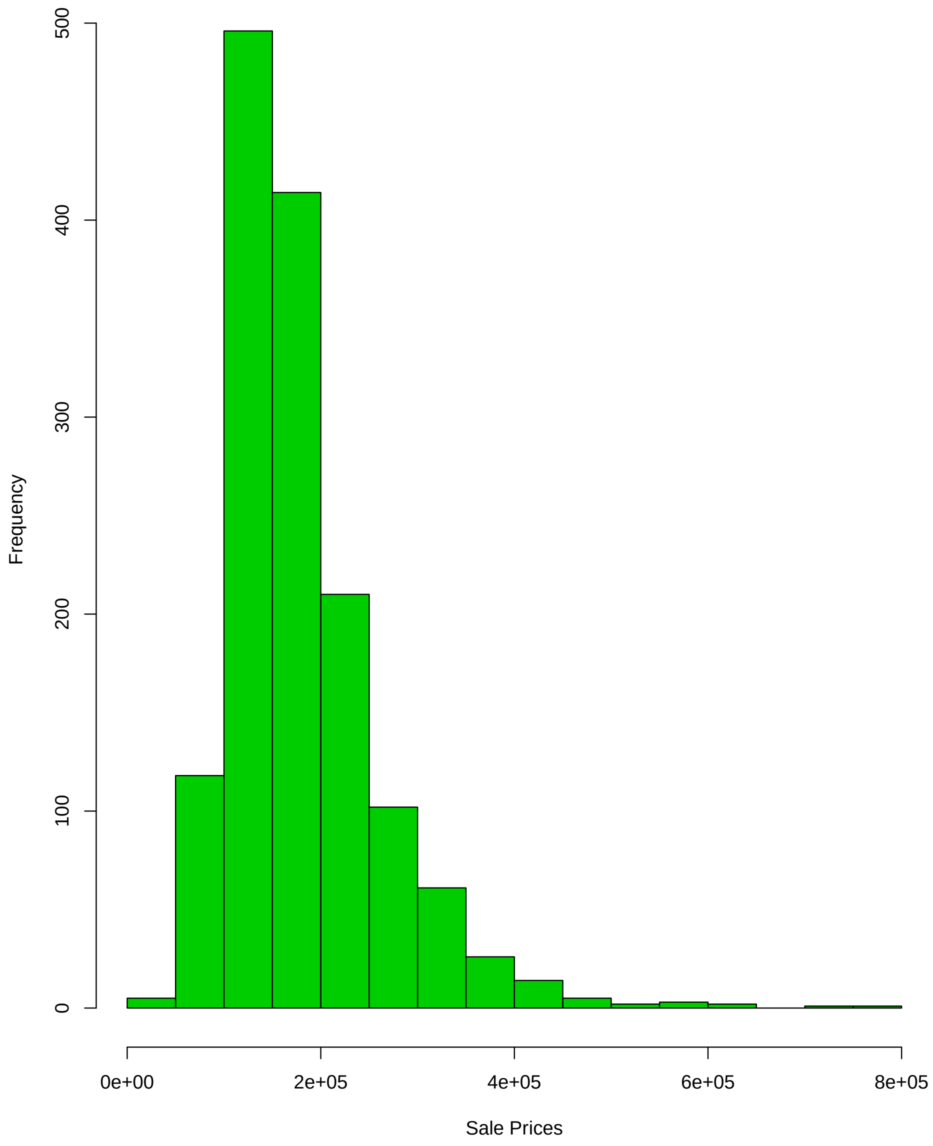
1 Purpose

The purpose of this project is to explore the best model for explaining house sale prices using the categorical data from a Kaggle data science competition. The dataset contains information on houses for sale in Ames, Iowa and has 79 variables. One of the variables were 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. Since the focus of this project will be based on design and analysis of experiments, we will only be concerned with qualitative variables. To perform our analysis we used the R packages: Readr, PCAmixdata, and Python, and JMP statistical software.

2 Data Description and Edits

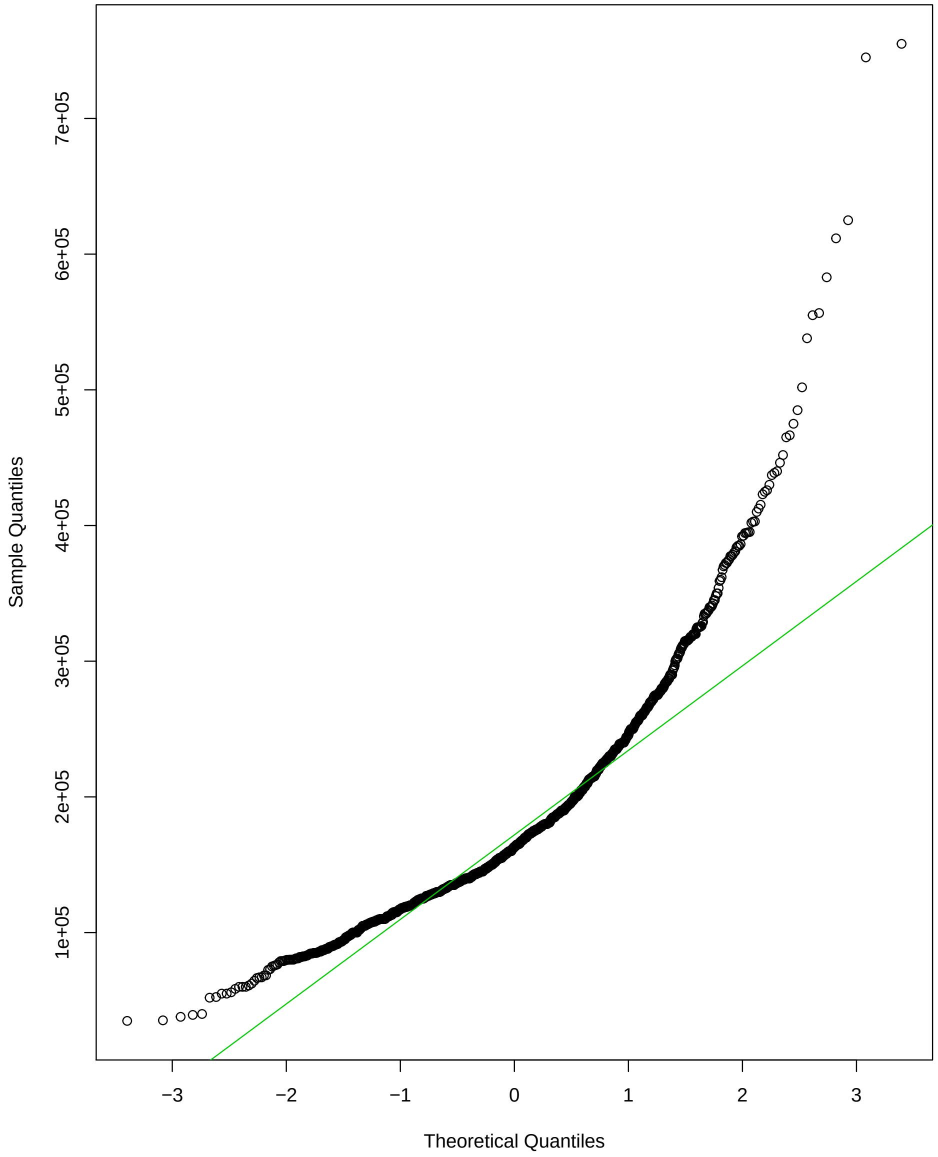
The 36 qualitative variables (descriptions on appendix) consist of a variety of factors describing the house's interior, exterior and surrounding neighborhood. While they cover wide variety of useful topics, many variables had enormous imbalances within their different levels. For example, MSSubClass, a variable examining the type of house that was involved in the sale, had 11 different levels, however, levels 1, 5 and 6 held nearly all of the data. We remedied this by transforming the variable into a 4 level factor, with levels 1, 5 and 6 becoming levels 1,2 and 3, with a 4th level labeled "Other". This allowed us to have a much more balanced variable, with the proportion of data being more evenly distributed over the different levels. The cost of this fix, is that we lose some of our interpretability. Now, if levels 1, 2 or 3 are not significant, we will not know which of the other MSSub-Classes are the main contributor, just that it is not levels 1, 2 or 3. Other variables were so dominant in one of their levels that adding it to our model would not aid in our goal of finding the main factors of house sales. The variable Street, for example, had only two levels: "Pave", meaning that the house was along a paved street, and "Grvl", meaning that the house was along a gravel road. Of these two variables, "Pave" consisted of 1,454 of the 1,460 (or 99.5%) of the observational units. If street was deemed significant, we would not be able to trust it, as we might not have had enough gravel facing houses to realize their importance. The final Modification came for the response variable. Looking at the plots of the sales values, one can see a definite pattern.

Fig 1



Indeed, if we examine the distribution of the response, we will see that it does not follow a normal distribution.

Fig 2



With the modification of a log transform, the scatter plots and QQs become normal, with constant variance and a linear relationship.

Fig 2

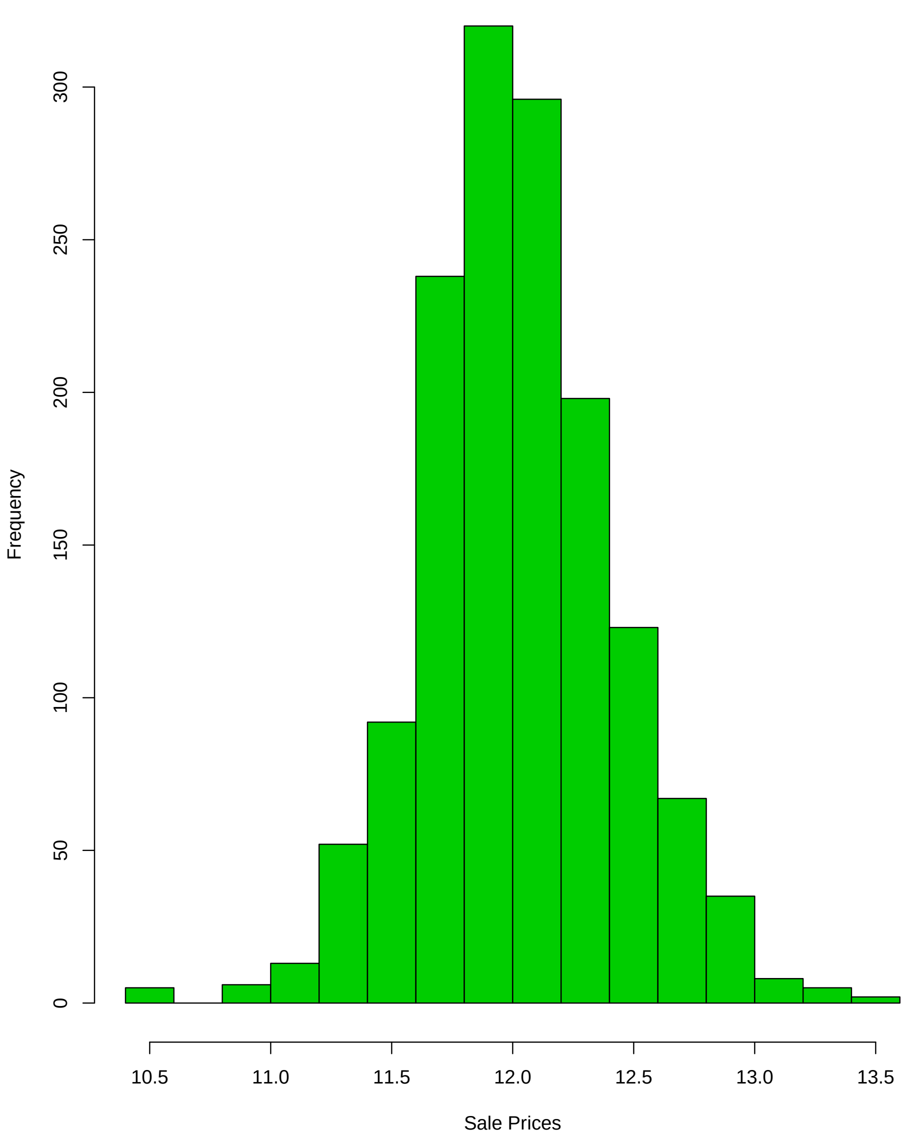
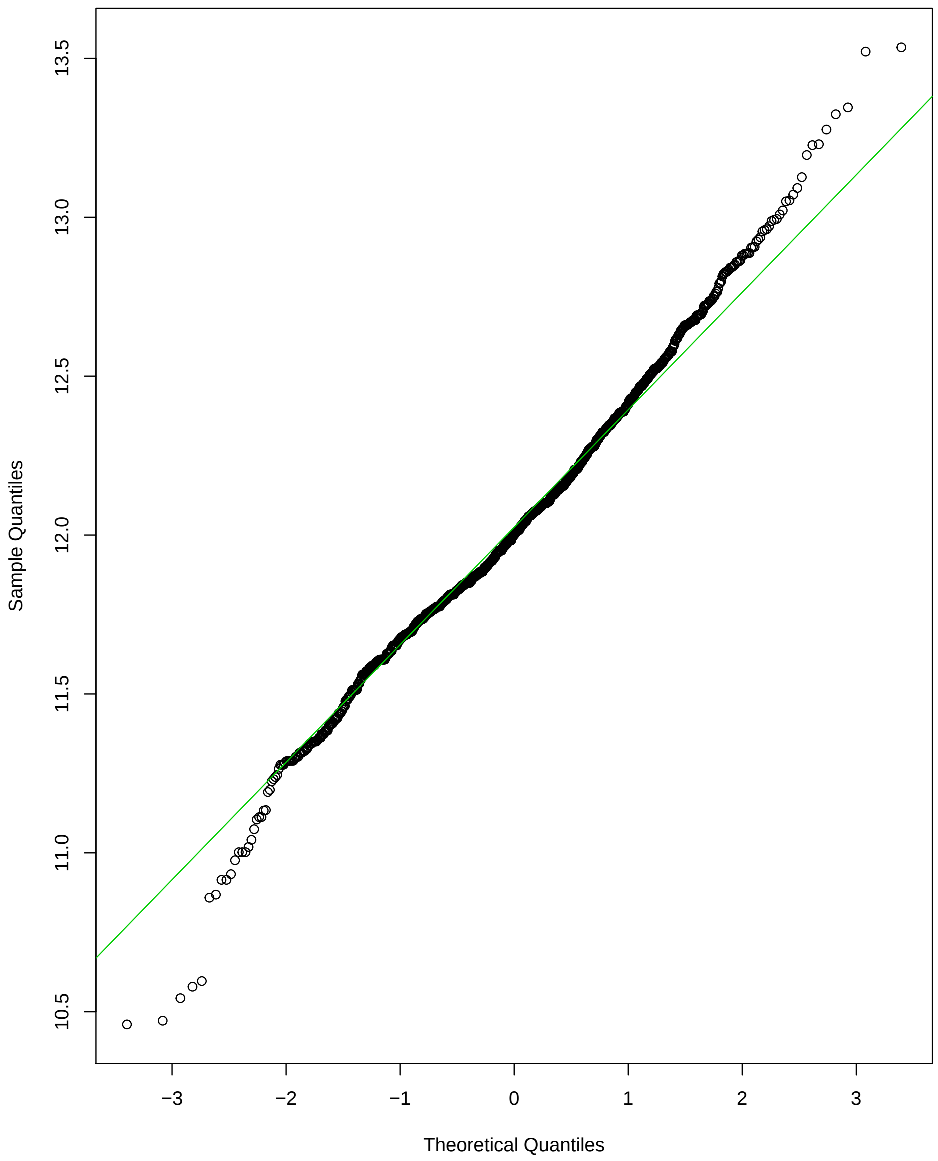


Fig 4



While examining our data, we realize that many of our variables are not capable of handling interaction terms. Since we do not have balance (see Variable Selection for details) we cannot introduce interaction terms without creating more bias. We know from the interaction plots that there may indeed be interactions between our variables, but we do not know if our F tests for those interactions will be valid. So instead of attempting inference on variables that are known to be biased, we have chosen to forego any interaction variables.

If given more time and resources, we could apply some practical fixes including: deriving the true expected values of the interaction mean squares, collecting more data and then sampling down until we have balance, bootstrapping our current data (this would require a more powerful computer than we had access to), coercing the data to be orthogonal via MCA (this was not allowed for the scope of this project), and many more.

3 Variable Selection

3.1 Reduction by binning

In order to remove redundant variables, we decided to use a combination matching algorithm of our own creation (appendix 5.1). The algorithm starts with making 'bins' that can hold all possible combinations of predictor variables (This was done after initial variable reduction.) We separated out columns that will not be predictors (X and SalePrice), and ran the algorithm on the remaining predictors. The 'bins' can take values from $[0, 0, ..., 0]$ to $[v_{1_{max}}, v_{2_{max}}, ..., v_{N_{max}}]$. Thus, every possible combination of variable levels can be recorded. If a variable didn't change the various bin numbers when removed/added, then we assumed that levels of that variable occur in the same place as levels of some other variables, therefore they are redundant in the information that they give. We would then leave that variable out of data. During this 'in and out' process we were only able to remove one variable, CentralAir.

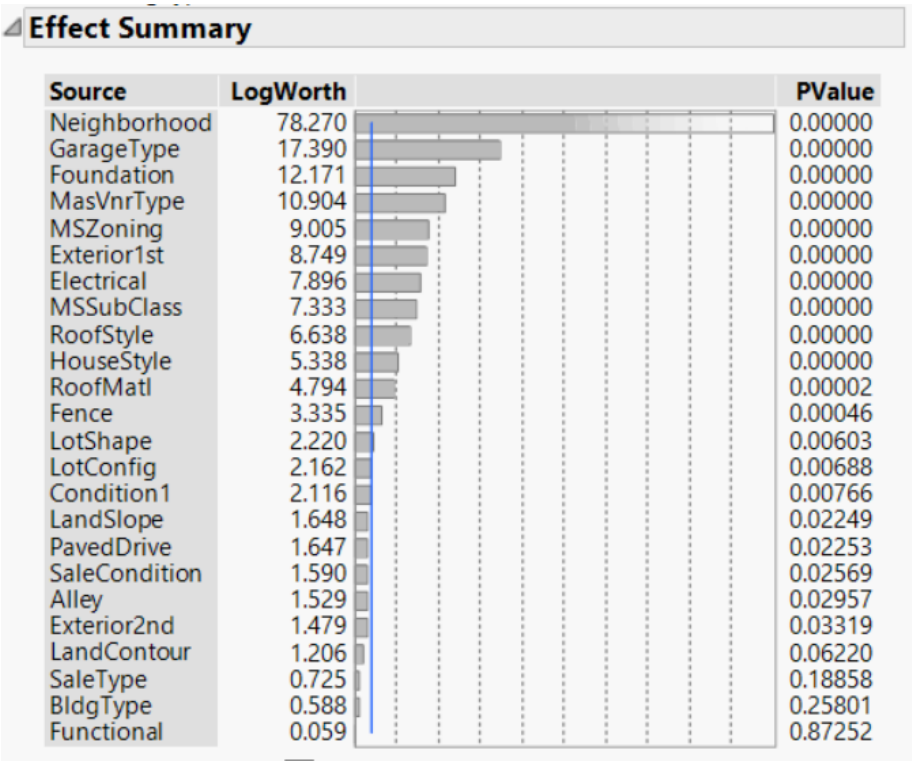
After the combination algorithm, we attempted to extract the biggest balanced table from our dataset. From prior theory, we knew that Neighborhood was the most important variable for our analysis. Because of this, it was guaranteed to be included and unchanged in our table. Unfortunately, the vast majority of 'bins' were empty and only a few had the vast majority of our data. Even after repeating this process with systematic variable selection, we ended up with a set of 'bins' which were as empty as in our previous attempts. We believe that the reason that this method did not produce any results lies in the fact that for most of our variables, they have some number of levels with disproportionally many observations in tandem with other levels of other variables. Also, since our observational units are houses it is understandable that certain features naturally go together. It was hard to achieve even one observation per combination of variable levels, although we have around 1400 observations and 23 variables. Since systematic variable selection process didn't produce a balanced table with reasonable number of non-empty variables we can not fully trust our significant tests. This is because of the expected values of the mean squares. Normally, with a balanced data set, our expected mean squares are well defined. Without balance, we do not know what they are. When we perform our F-tests, our ratios of $\frac{MS_{factor}}{MS_{error}}$ will not have the correct numerator. Instead, the numerator will have extra covariance terms added onto it that we are not accounting for. As a result, our F values will be farther from 1 than we would expect, and our p-values will be lower than they should be. In an attempt to fight this, we will use a more conservative .01 for our p-value. We recognize that this is not a perfect fix and that there will still be an unknown amount of bias involved. To fix this problem completely, we would have to use the appropriate F ratio which would involve calculating the covariances between our factors, and deriving the appropriate expected values for our mean squares.

3.2 ANOVA Variable selection

In order to establish the most optimal model, we will run our data through a two tier selection process and use k-fold validation to determine which of the two preforms the best and therefor earns our recommendation.

3.2.1 Model 1

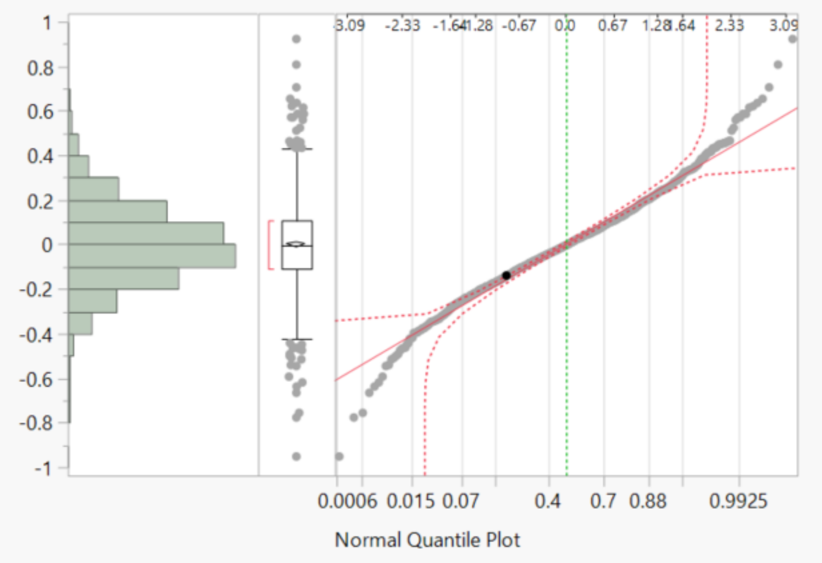
Our first model was fit with everything inside of it and then reduced with our .01 p-value cut off. The resulting JMP output shows that even with our conservative p-value estimate, we were able to reduce 9 of our 24 variables.



Effect Summary

Source	LogWorth	PValue
Neighborhood	78.270	0.00000
GarageType	17.390	0.00000
Foundation	12.171	0.00000
MasVnrType	10.904	0.00000
MSZoning	9.005	0.00000
Exterior1st	8.749	0.00000
Electrical	7.896	0.00000
MSSubClass	7.333	0.00000
RoofStyle	6.638	0.00000
HouseStyle	5.338	0.00000
RoofMatl	4.794	0.00002
Fence	3.335	0.00046
LotShape	2.220	0.00603
LotConfig	2.162	0.00688
Condition1	2.116	0.00766
LandSlope	1.648	0.02249
PavedDrive	1.647	0.02253
SaleCondition	1.590	0.02569
Alley	1.529	0.02957
Exterior2nd	1.479	0.03319
LandContour	1.206	0.06220
SaleType	0.725	0.18858
BldgType	0.588	0.25801
Functional	0.059	0.87252

This model also displayed normal residuals, linear relationships and equal variance.



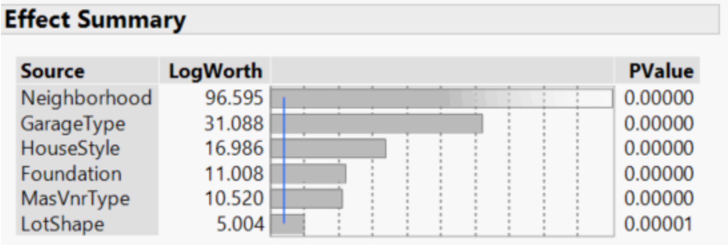
The resulting model is as such:

$$log(salesprice) \sim Condition1 + LotConfiguration + LotShape + Fence + RoofMaterial + HouseStyle + MSSubClass$$

$$+Electrical + Exterior1st + MSZoning + MasonryVeneer + Foundation + GarageType + Neighborhood$$

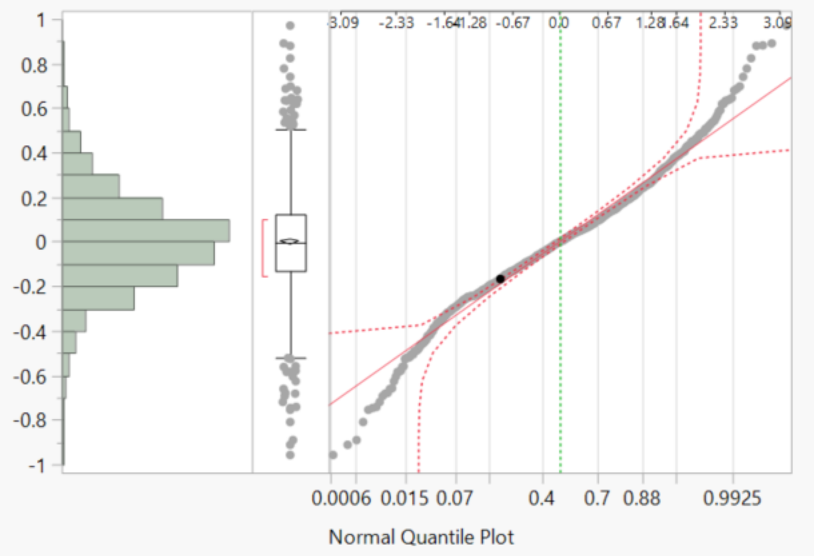
3.2.2 Model 2

After our first round of eliminations, we preformed a second selection process. Any variable that had over 70% of its data or more in one level was excluded from our model. The thought behind this action was that heavily biased data might lead our model away from the true values of sale price. This resulted in the following model:



Effect Summary

Source	LogWorth		PValue
Neighborhood	96.595		0.00000
GarageType	31.088		0.00000
HouseStyle	16.986		0.00000
Foundation	11.008		0.00000
MasVnrType	10.520		0.00000
LotShape	5.004		0.00001



$$log(\text{sales price}) \sim \text{Neighborhood} + \text{Garage Type} + \text{House Style} + \text{Foundation} + \text{Masonry Veneer Type} + \text{Lot Shape}$$

3.3 Model Performances

With our two models selected, we ran k-fold validation with k equaling 5 and 10. We then ranked our model's performance based on their Adjusted R Squared and RMSE values. The results are below:

Model Performance, K = 5				Model Performance, K = 10			
Model	R^2	RMSE	Rank	Model	R^2	RMSE	Rank
1	.6821	.2258	1	1	.6845	.2253	1
2	.6589	.2339	2	2	.6654	.2317	2

4 Conclusion

When attempting to predict the housing prices in the town of Ames, Iowa, we were able to create two models using a data set of 36 qualitative house attributes with 1,460 observations. While we did not have adequate balance, we attempted to run an ANOVA analysis anyway. We compensated for the inflated MS values with an extra conservative p-value of .01 instead of .05. Once we had our models, we ran two k-fold validation tests, one with $K = 5$ and another with $K = 10$. The resulting best model for predicting house sale prices is model 1. This can be expanded and improved with the addition of the forgone continuous variables.

5 Appendix

5.1 Data Description

```
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 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

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

5.2 Cleaning Algorithm

```
cdata<-read.table("Clean_data.csv", header=TRUE, sep=",")


cdata$BsmtQual<- NULL #ordinal
cdata$BsmtFinType1<- NULL #ordinal
cdata$BsmtFinType2<- NULL #ordinal
cdata$PoolQC<- NULL #ordinal
cdata$GarageCond<- NULL #ordinal
cdata$GarageFinish<-NULL#ordinal

levels[length(levels) + 1] <- "None"
cdata<-na.omit(cdata)
cdata$GarageType <- factor(cdata$GarageType, levels = levels)

cdata$GarageType[is.na(cdata$GarageType)] <- "None"
rm(levels)

table(cdata$MSSubClass)
levels(cdata$MSSubClass)[levels(cdata$MSSubClass)=="C"] <- "Other"
levels(cdata$MSSubClass)[levels(cdata$MSSubClass)=="H"] <- "Other"
levels(cdata$MSSubClass)[levels(cdata$MSSubClass)=="O"] <- "Other"

table(cdata$MSZoning)
levels(cdata$MSZoning)[levels(cdata$MSZoning)=="C (all)"] <- "Other"
levels(cdata$MSSubClass)[levels(cdata$MSSubClass)=="RH"] <- "Other"


table(cdata$Street) #there is only 6 obs of GRVL the rest are PAVE
cdata$Street<-NULL

table(cdata$Utilities) # only 1 is NOSEWA
cdata$Utilities<- NULL

table(cdata$LotConfig)
levels(cdata$LotConfig)[levels(cdata$LotConfig)=="FR2"] <- "FR2and3"
levels(cdata$LotConfig)[levels(cdata$LotConfig)=="FR3"] <- "FR2and3"


table(cdata$HouseStyle)
levels(cdata$HouseStyle)[levels(cdata$HouseStyle)=="2.5Fin"] <- "2.5FU"
levels(cdata$HouseStyle)[levels(cdata$HouseStyle)=="2.5Unf"] <- "2.5FU"

table(cdata$RoofStyle)
levels(cdata$RoofStyle)[levels(cdata$RoofStyle)=="Mansard"] <- "Other"
levels(cdata$RoofStyle)[levels(cdata$RoofStyle)=="Shed"] <- "Other"
levels(cdata$RoofStyle)[levels(cdata$RoofStyle)=="Gambrel"] <- "Other"
levels(cdata$RoofStyle)[levels(cdata$RoofStyle)=="Flat"] <- "Other"

table(cdata$RoofMatl)
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="ClyTile"] <- "Other"
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="Membran"] <- "Other"
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="Metal"] <- "Other"
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="Roll"] <- "Other"
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="WdShake"] <- "Other"
levels(cdata$RoofMatl)[levels(cdata$RoofMatl)=="WdShngl"] <- "Other"
```

```
table(cdata$Exterior1st)
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="AsphShn"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="BrkComm"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="CBlock"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="ImStucc"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="Stone"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="STorWO"] <- "Other"
levels(cdata$Exterior1st)[levels(cdata$Exterior1st)=="AsbShng"] <- "Other"

table(cdata$Exterior2nd)
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="AsphShn"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="BrkComm"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="CBlock"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="ImStucc"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="Stone"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="STorWO"] <- "Other"
levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="Cmn"] <- "Other"

table(cdata$Foundation)
levels(cdata$Foundation)[levels(cdata$Foundation)=="ST"] <- "Other"
levels(cdata$Foundation)[levels(cdata$Foundation)=="Wood"] <- "Other"
levels(cdata$Foundation)[levels(cdata$Foundation)=="Slab"] <- "Other"
levels(cdata$Foundation)[levels(cdata$Foundation)=="Stone"] <- "Other"

table(cdata$BsmtExposure) #ordinal
cdata$BsmtExposure<- NULL

table(cdata$Heating)
levels(cdata$Heating)[levels(cdata$Heating)!="GasA"] <- "Other"
levels(cdata$Heating)

table(cdata$CentralAir)

table(cdata$Electrical)

levels(cdata$Electrical)[levels(cdata$Electrical)=="FuseF"] <- "FuseFP"
levels(cdata$Electrical)[levels(cdata$Electrical)=="FuseP"] <- "FuseFP"
levels(cdata$Electrical)[levels(cdata$Electrical)=="Mix"] <- "FuseFP"
levels(cdata$Electrical)

table(cdata$GarageType)
levels(cdata$GarageType)[levels(cdata$GarageType)=="2Types"] <- "Other"
levels(cdata$GarageType)[levels(cdata$GarageType)=="Basment"] <- "Other"
levels(cdata$GarageType)[levels(cdata$GarageType)=="CarPort"] <- "Other"
levels(cdata$GarageType)

table(cdata$SaleType)
levels(cdata$SaleType)[levels(cdata$SaleType)=="Con"] <- "Other"
levels(cdata$SaleType)[levels(cdata$SaleType)=="ConLD"] <- "Other"
levels(cdata$SaleType)[levels(cdata$SaleType)=="ConLI"] <- "Other"
levels(cdata$SaleType)[levels(cdata$SaleType)=="ConLw"] <- "Other"
levels(cdata$SaleType)[levels(cdata$SaleType)=="CWD"] <- "Other"
levels(cdata$SaleType)[levels(cdata$SaleType)=="Oth"] <- "Other"

table(cdata$SaleCondition)
levels(cdata$SaleCondition)[levels(cdata$SaleCondition)=="AdjLand"] <- "AdjAlloca"
levels(cdata$SaleCondition)[levels(cdata$SaleCondition)=="Alloca"] <- "AdjAlloca"
```

```
table(cdata$Condition1)
levels(cdata$Condition1)[levels(cdata$Condition1)=="PosA"] <- "Other"
levels(cdata$Condition1)[levels(cdata$Condition1)=="PosN"] <- "Other"
levels(cdata$Condition1)[levels(cdata$Condition1)=="RRAe"] <- "Other"
levels(cdata$Condition1)[levels(cdata$Condition1)=="RRNe"] <- "Other"
levels(cdata$Condition1)[levels(cdata$Condition1)=="RRNn"] <- "Other"

table(cdata$Functional)
levels(cdata$Functional)[levels(cdata$Functional)=="Maj1"] <- "Other"
levels(cdata$Functional)[levels(cdata$Functional)=="Maj2"] <- "Other"
levels(cdata$Functional)[levels(cdata$Functional)=="Mod"] <- "Other"
levels(cdata$Functional)[levels(cdata$Functional)=="Sev"] <- "Other"

cdata$Condition2<-NULL

levels(cdata$Exterior2nd)[levels(cdata$Exterior2nd)=="Brk Cmn"] <- "Other"
table(cdata$Exterior2nd)

cdata$Heating<-NULL
cdata$CentralAir<-NULL #since it is correlated with something
cdata<-na.omit(cdata)
write.csv(cdata, file = "data.csv")
```

5.3 Binning Algorithm

```
def bin_data(data):
    """ we will try to see how many different combinations are in existence"""
    properties_combinations = []
    occurrences = []
    dupes = []
    current_bin = 1
    # data[l] is the l-th observation
    # data[l][0] is the observation ID (it equals l)
    # data[l][1] is the vector containing all the variables
    # data[l][2] contains selling price
    # finally, data[l][3] contains bin ID
    for l in range(len(data)):
        if not (data[l][1] in properties_combinations):
            properties_combinations.append(data[l][1])
            occurrences.append(1)
        else:
            for temp in range(len(properties_combinations)):
                if data[temp][1] == properties_combinations[temp]:
                    occurrences[temp] += 1
                    dupes.append(temp)
    for property_combination in properties_combinations:
        for row in data:
            if row[1] == property_combination:
                row[-1] = current_bin
        current_bin += 1

    return data, len(properties_combinations), dupes
```