Exploring Value Impacting Factors on Real Estate Prices

York University, CSDA 1000 Introduction to Big Data, June 2019

Vaibhav Rastogi(ID#306309)

Table of Contents

ABSTRACT:

The purpose of this paper is to determine whether a predictive model can be used to determine the value of real estate. Specifically, it will attempt to assess which quantitative factors that most influence the value of real estate in a given housing market. Using the Kaggle housing data for Ames, Iowa, we will examine the various attributes of properties and, removing qualitative elements, we hope to generate a functioning model for determining the value impact of a given property's features (in the baseline context of its market). Furthermore, beyond exploring the correlation between the property's value and its quantitative attributes, the model will seek to assess the weight of each attribute; while we expect there to be significant correlation between attributes and price, the greater focus of this project will be understanding the weight of each attribute as value-impacting-factors and will require further assessment.
Finally after calculating the correlations, features were selected and a linear regression model was created. With each iteration the model was tested for the statistical significance of its output and validated using the error and fit metrics

INTRODUCTION:

Due to the significant impact qualitative factors have on perceived value of real estate (location, "Trendiness" of a given area, "heat" of a real estate market, proximity to points of interest and amenities, personal preferences of buyer, etc.), it becomes invaluable to have access to tools that can reasonably assess the objective value of the impact quantitative attributes of the property will have, and the weight of a given feature on that impact.

In having such tools at one's disposal, one can better appraise the overall value of a property; by accounting for a "baseline" value rooted in qualitative features, one can apply the value-impacting-factor formula to predict and determine a property's relative worth.

The benefit and utility of such a tool is fourfold:

1) In and of itself, it will provide builders, realtors, and prospective buyers with a means of better assessing the value of given properties, allowing for more informed and data-based decision-making.
2) By inversion, it will provide those same parties with the ability to assess the qualitative value of properties and areas (IE: given the value of the quantitative features, they can determine the qualitative impact factor of a given property/group of properties).
3) In broader application, it will allow for the forecasting of market trends (Ex: the increasing or decreasing 'trendiness' of a neighbourhood, the heat & health of a market, etc.) by

flagging transaction patterns that fall outside the parameters of the established factor impact evaluations. This will be invaluable to market analysts, urban planners, and financial institutions.

4) In scalability, this tool can provide a foundation for the assessment of any business or market that relies on meta-bivariate evaluation (IE existence & quantity of particular features and value of qualitative factors & preferences).

DATA

The dataset used for this analysis came courtesy of the training set for the Kaggle competition *House Prices: Advanced Regression Techniques*. It provides final sale prices for 1,460 properties in Ames, Iowa, as well as 79 attributes for the properties themselves, both qualitative and quantitative. Attributes included age of property, the neighbourhoods they were located in, floor area, lot area, year renovated, the category of house-type, the category of porch on the property, etc.

Given that the data was provided for an official Kaggle competition and its utilization of actual realtor data, it is reasonable to expect the data to be both rigorous & sufficient for the purposes of our analysis, as well as accurate for the purposes of scaling and extrapolation. The 79 attributes provide ample opportunities for feature engineering. The data allows for separation based on qualitative and quantitative attributes, allowing for modeling that accounts for the impact of either as a dependent variable.

Before modeling could be undertaken, it was imperative to analyse the attributes, separate the qualitative elements, transform quantifiable data into appropriate workable formats, create custom attributes that better quantify the significance of ordinal and nominal data (IE year renovated: years since renovation, buildings with any porch type combined into single category, etc.), and isolate those that had the greatest direct correlation with final prices.

METHODOLOGY:

1. Data preparation

First, our team scanned the data set and identified the dependent and independent variables. Second, we attempted to clean the dataset of any observations that are outliers. We created box plots for every variable to check how many outliers were observed and to determine if these outliers were statistically significant.
Finally, the team discussed how to determine the strength of correlation between the dependent and independent variables.

The team prepared the data set to make it suitable for the model. This included a lengthy process of converting string data into binary data so it can be interpreted by the analysis tool and we can generate the statistics of the data set.
In order to be entirely objective about cleaning the data in the dataset, the number of unique values in each column were counted.
With Price being the dependent variable all other values in the table other than Model were considered possible independent variables. The columns containing categorical variables are strings are separated into unique binary variables in Creating Value Attributes. The same was done for variables containing numerical values but which were ordinal in nature

The Scatter plot & box plot graphs were utilized to reach the following:

- § The 1st Quartile (Q1), 3rd Quartile (Q3) and Inter Quartile (IQ) were calculated;
- § Using a lower boundary of Q1 – 3*IQ and an upper boundary of Q3 + 3*IQ the outliers were identified (SEMATECH, 2019);
- § If the outlier was obviously due to incorrectly measured or entered data, the row was removed from the dataset;
- § A regression model was created with the independent variable and the dependent variable; and,
- § If the outlier did not affect the results of the regression graph, then the outlier rows were not removed from the dataset – if the outlier did affect the results of the regression graph then the rows with the outlier values were removed from the dataset (Grace-Martin, 2019).

The following section describes each variable in the “Train” dataset and if we did anything in terms of imputing the data. The variables of importance to us have been highlighted and any imputation were carried forward to the “Test” data set as well:

We had 79 variables in addition to 1 target variable. In total we had 1460 records in our train data set and 1518 in our test data set. These are the steps we took for each

Variable Type	**Number of variables**	**Reasoning / Action**
Removed	11	Removed because their variability was too low (more than 98% in one value) or had no connection with the data we were using
Target Variable	1	SalePrice
Numerical Variables	30	Kept as is
Categorical Variables	35	Converted to 177 Dummy Variables
Time / Year variables	3	Converted to a single age variable

2. Feature Engineering

We calculated the correlation with respect to Sales Price for each of the 204 variables

These variables had strong correlation with the sales price. The threshold for strong in our case was greater than 0.5 or less than -0.5

Variable Name	Correlation with Sales Price
OverallQual	0.790981601
GrLivArea	0.708624478
GarageCars	0.640409197
GarageArea	0.623431439
TotalBsmtSF	0.613580552
1stFlrSF	0.605852185
F_EXTERQ_TA	-0.589043523
FullBath	0.560663763
IF_BSMTQ_Ex	0.553104847
TotRmsAbvGrd	0.533723156
IF_KITQ_TA	-0.519297854
AGE_IN_YEARS	-0.509078738
IF_KITQ_OTHER	0.504093676

AGE_IN_YEARS , IF_KITCHENQUALITY_TYPICAL , TotalRoomsAboveGround , IF_BasementQuality_Excellent , No. of FullBathrooms , IF_EXTERnalQuality_typical , 1stFlr SqFeet , TotalBasementSqFeet , GarageArea , GarageCars , abovegrade Living Area , OverallQual, (overall material and finish of the house) IF_Kitchen_quality_excellent

Age, If kitchen quality is typical and if external quality is typical are NEGATIVELY correlated with the price, meaning these values have a downward effect on the price as they go up.

Correlation Matrix

Now we will check for correlations within these variables to identify if in any of the cases is the correlation very high (>0.7 or < -0.7). By doing this we reduce redundancies and improve the statistical significance of our results

Full Correlation Matrix

	SalePrice	IF_BSMTQ_Ex	TotalBsmtSF	X1stFlrSF	FullBath	IF_KITQ_TA
SalePrice	1	0.5531	0.61358	0.60585	0.56066	-0.5193
IF_BSMTQ_Ex	0.5531	1	0.40317	0.37739	0.2374	-0.26292
TotalBsmtSF	0.61358	0.40317	1	0.81953	0.32372	-0.31119
X1stFlrSF	0.60585	0.37739	0.81953	1	0.38064	-0.27357
FullBath	0.56066	0.2374	0.32372	0.38064	1	-0.4187
IF_KITQ_TA	-0.5193	-0.26292	-0.31119	-0.27357	-0.4187	1
TotRmsAbvGrd	0.53372	0.24058	0.28557	0.40952	0.55478	-0.21043
GarageCars	0.64041	0.35642	0.43458	0.43932	0.46967	-0.39945
GarageArea	0.62343	0.37022	0.48667	0.48978	0.40566	-0.37093
AGE_IN_YEARS	-0.50908	-0.28632	-0.2921	-0.24131	-0.44041	0.57706
IF_EXTERQ_TA	-0.58904	-0.33323	-0.3904	-0.31284	-0.46634	0.6716

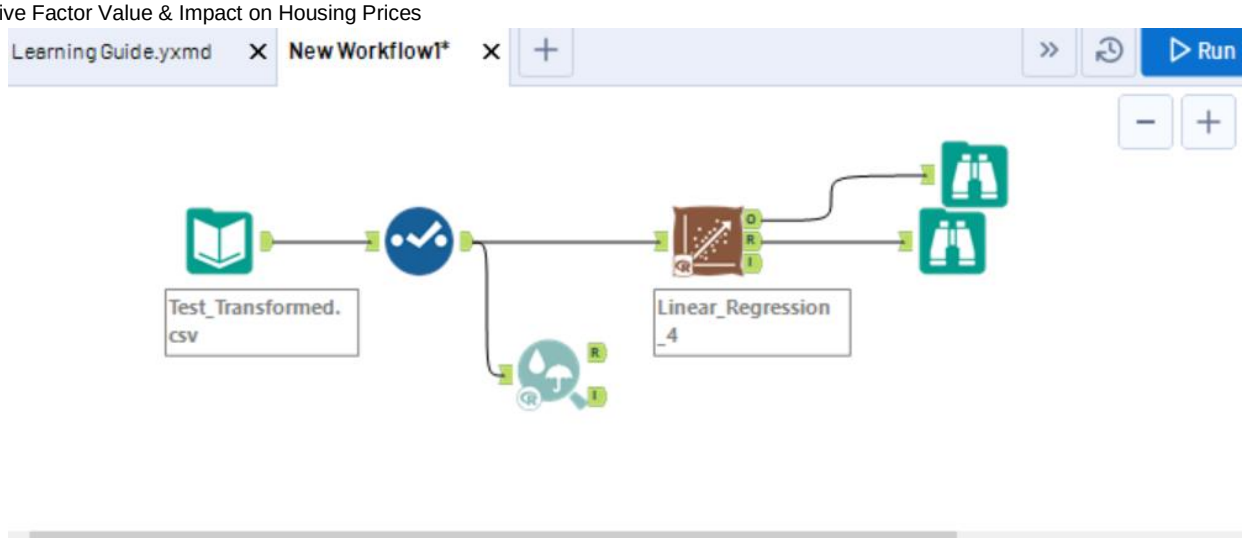
	TotRmsAbvGrd	GarageCars	GarageArea	AGE_IN_YEARS	F_EXTERQ_TA	
SalePrice	0.53372	0.64041	0.62343	-0.50908	-0.58904	
IF_BSMTQ_Ex	0.24058	0.35642	0.37022	-0.28632	-0.33323	

TotalBsmtSF	0.28557	0.43458	0.48667	-0.2921	-0.3904	
X1stFlrSF	0.40952	0.43932	0.48978	-0.24131	-0.31284	
FullBath	0.55478	0.46967	0.40566	-0.44041	-0.46634	
IF_KITQ_TA	-0.21043	-0.39945	-0.37093	0.57706	0.6716	
TotRmsAbvGrd	1	0.36229	0.33782	-0.194	-0.24248	
GarageCars	0.36229	1	0.88248	-0.42324	-0.48183	
GarageArea	0.33782	0.88248	1	-0.37345	-0.44526	
AGE_IN_YEARS	-0.194	-0.42324	-0.37345	1	0.56562	
F_EXTERQ_TA	-0.24248	-0.48183	-0.44526	0.56562	1	

We have identified two cases

1. Correlation between Total Basement Square footage and 1st floor square footage is very high (0.82) . This is intuitively also correct. Hence in this case we will only use Total Basement SQft for our model because the correlation between that and saleprice is marginally higher than the correlation between 1st floor square footage and price
2. Correlation between Garage Area and Garage Cars is also very high (0.88). This is intuitively also correct. Hence in this case we will only use Garage Cars in our model because correlation between that and sale price is marginally higher than the correlation between garage area and sale price
3. General Living area and Total rooms above ground were highly correlated. We used the former in our model because of the limited

Data Modeling



Workflow

1st Iteration of the model

We ran the first iteration of the linear regression model with the following variables – AGE_IN_YEARS , IF_KITQ_TA , IF_KITQ_Other, IF_BSMTQ_Ex , FullBath , F_EXTERQ_TA ,TotalBsmtSF, GarageCars , GrLivArea , OverallQual

Report

Report for Linear Model Linear_Regression_6

Basic Summary

Call:

lm(formula = SalePrice ~ OverallQual + IF_BSMTQ_Ex + TotalBsmtSF + GrLivArea + FullBath + IF_KITQ_TA + IF_KITQ_OTHER + GarageCars + AGE_IN_YEARS + F_EXTERQ_TA, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-473816	-16392	-379	14856	258738

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-22774.81	7578.294	-3.00527	0.0027	**
OverallQual	14724.33	1175.636	12.52457	< 2.2e-16	***
IF_BSMTQ_Ex	38494.82	4243.404	9.07168	< 2.2e-16	***
TotalBsmtSF	24.14	2.663	9.06497	< 2.2e-16	***
GrLivArea	46.06	2.582	17.84086	< 2.2e-16	***
FullBath	-100.35	2403.188	-0.04176	0.9667	
IF_KITQ_TA	-5952.27	2691.101	-2.21183	0.02713	*
IF_KITQ_OTHER	33135.70	4510.751	7.34594	3.39e-13	***
GarageCars	15284.30	1640.841	9.31492	< 2.2e-16	***
AGE_IN_YEARS	-264.40	61.142	-4.32441	2e-05	***
F_EXTERQ_TA	-7549.62	2969.206	-2.54264	0.01111	*

In this iteration we notice that one variables is not statistically significant in the model because its p values is much above 0.05. We were taking a 95% confidence interval

In the next iteration we will remove the variables –number of full bathrooms and run the model again

2nd Iteration of the model

Report

Report for Linear Model Linear_Regression_6

Basic Summary

Call:

lm(formula = SalePrice ~ OverallQual + IF_BSMTQ_Ex + TotalBsmtSF + GrLivArea + IF_KITQ_TA + IF_KITQ_OTHER + GarageCars + AGE_IN_YEARS + F_EXTERQ_TA, data = the.data)

Residuals:

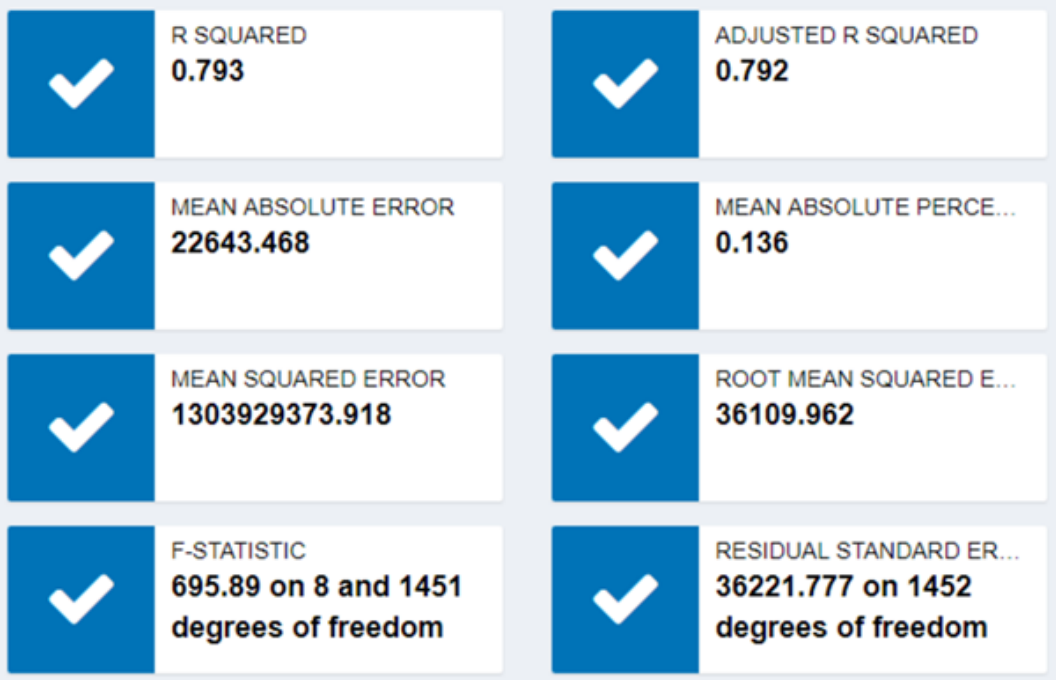
Min	1Q	Median	3Q	Max
-473671	-16420	-360	14862	258743

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-22843.87	7393.090	-3.090	0.00204	**
OverallQual	14720.10	1170.855	12.572	< 2.2e-16	***
IF_BSMTQ_Ex	38498.30	4241.130	9.077	< 2.2e-16	***
TotalBsmtSF	24.14	2.657	9.087	< 2.2e-16	***
GrLivArea	46.01	2.283	20.153	< 2.2e-16	***
IF_KITQ_TA	-5948.29	2688.483	-2.213	0.02709	*
IF_KITQ_OTHER	33156.41	4481.861	7.398	2.33e-13	***
GarageCars	15276.41	1629.381	9.376	< 2.2e-16	***
AGE_IN_YEARS	-263.96	60.198	-4.385	1e-05	***
F_EXTERQ_TA	-7538.43	2956.087	-2.550	0.01087	*

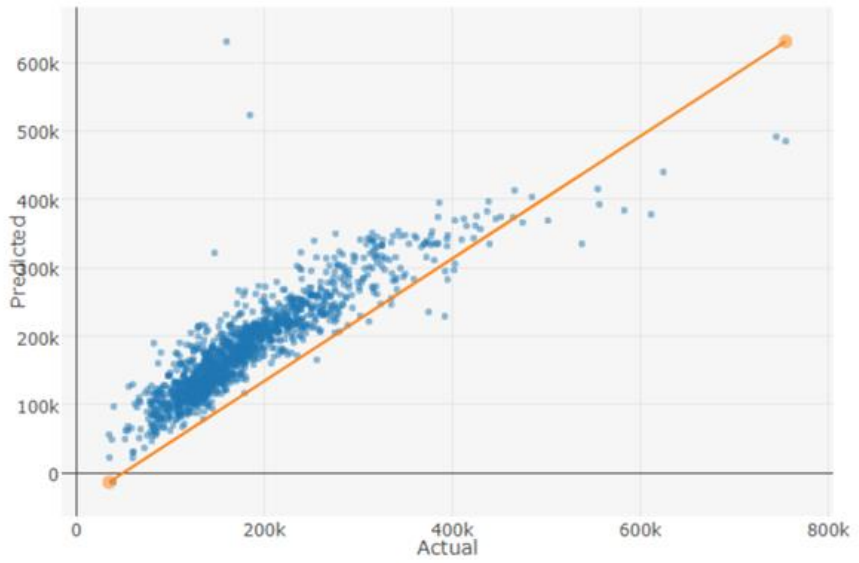
In this iteration all variables are coming out as statistically significant. This is our final model

Model Validation

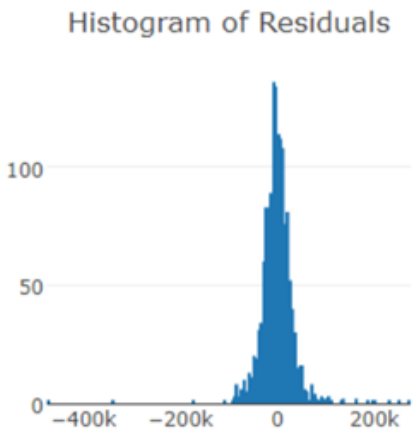


An adjusted Rsq of 79.3% is reasonably high

Justification for linear Model –



In the graph given above a linear shape is the best fir the the data points. Moreover look at the residual error plot below. It forms a bell curve around zero value



Statistic	Value
Minimum	-470650
1st Quartile	-16544
Median	-891
Mean	0
3rd Quartile	15442
Maximum	270139

PREDICTIONS & CONCLUSION

This is our final linear regression equation

SALE PRICE =
-30375.05 + 15794.94x(Overall Quality) + 50450.17x(IF_BASEMENTQUALITY_EXCELLENT) + 26.05x(Total_Basement_SqFeet) + 46.76x(General_Living_Area) – 7270.37x(IF_KitchenQuality_Typical) + 14793.1x(GarageCars) -254.68x(Age_in_Years) – 6897.75x(IF_ExternalQuality_Typical)

This consists of 8 variables out of a total of 204 actual and dummy variables. The Accuracy with this predicts the value of the target variable is 79.3%

Adjusted R Squared Value - {0.8}
Key Predictor Variables -
Age, Quality of Basement, External Structure, Kitchen, #Garagecars, Area of Basement, General Living and Overall Quality
Surprising misses -
Neighbourhood, Number of rooms, Number of stories, Lotsize
Future Step:
Build a clustering model on the housing data

Data Exploration

CHARTS & VISUALIZATIONs:

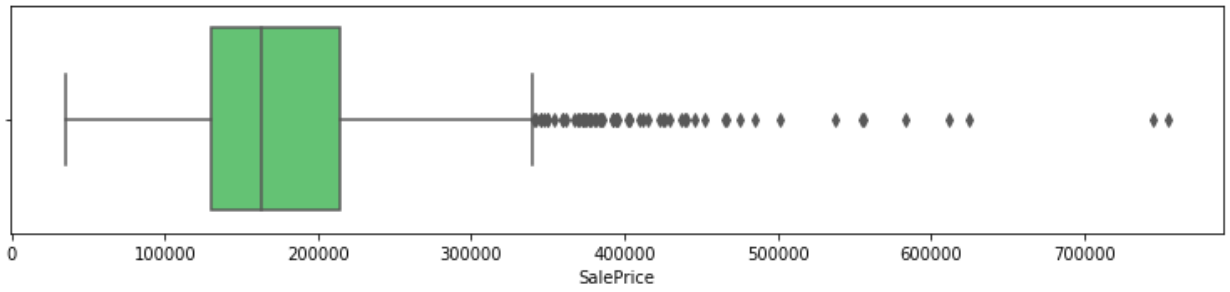


Figure 1. Sale price statistics:

range:	$34,900 - $755,000
mean:	$180,921.19
std:	$79,442.50
2nd quartile:	$129,975
median:	$163,000
3rd quartile:	$214000

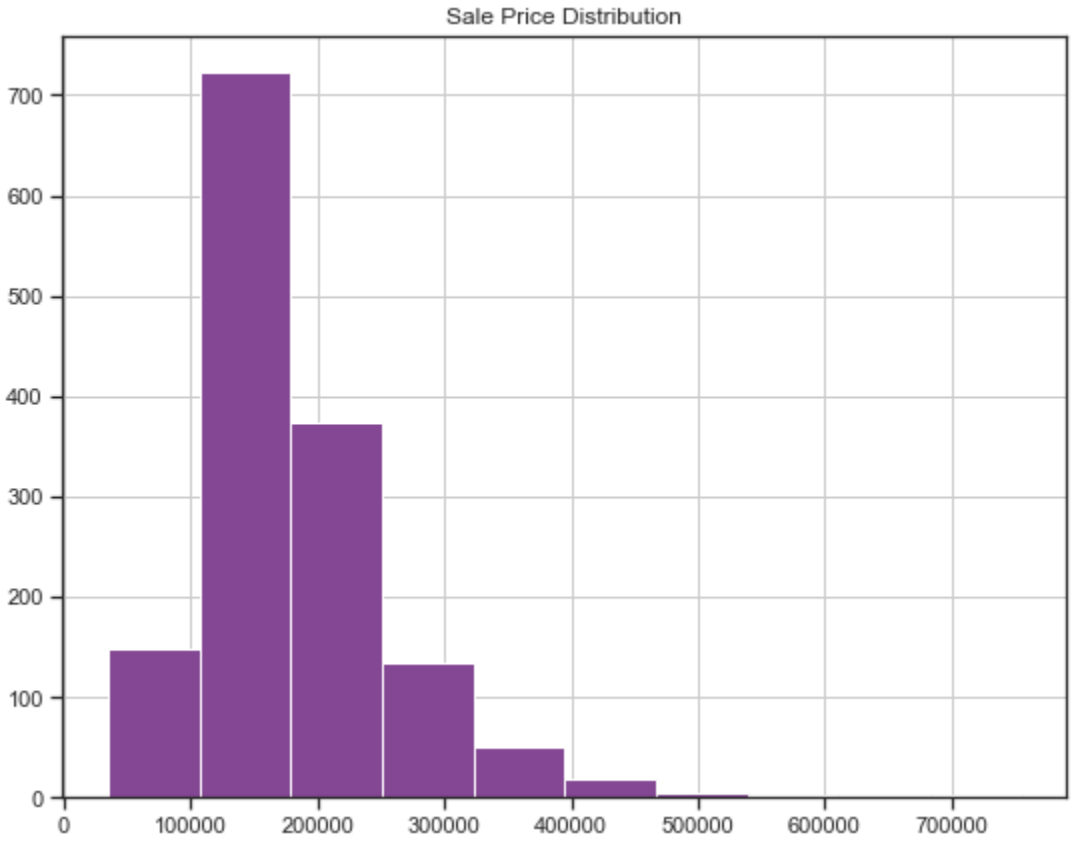


Figure 2. Sale Price distribution

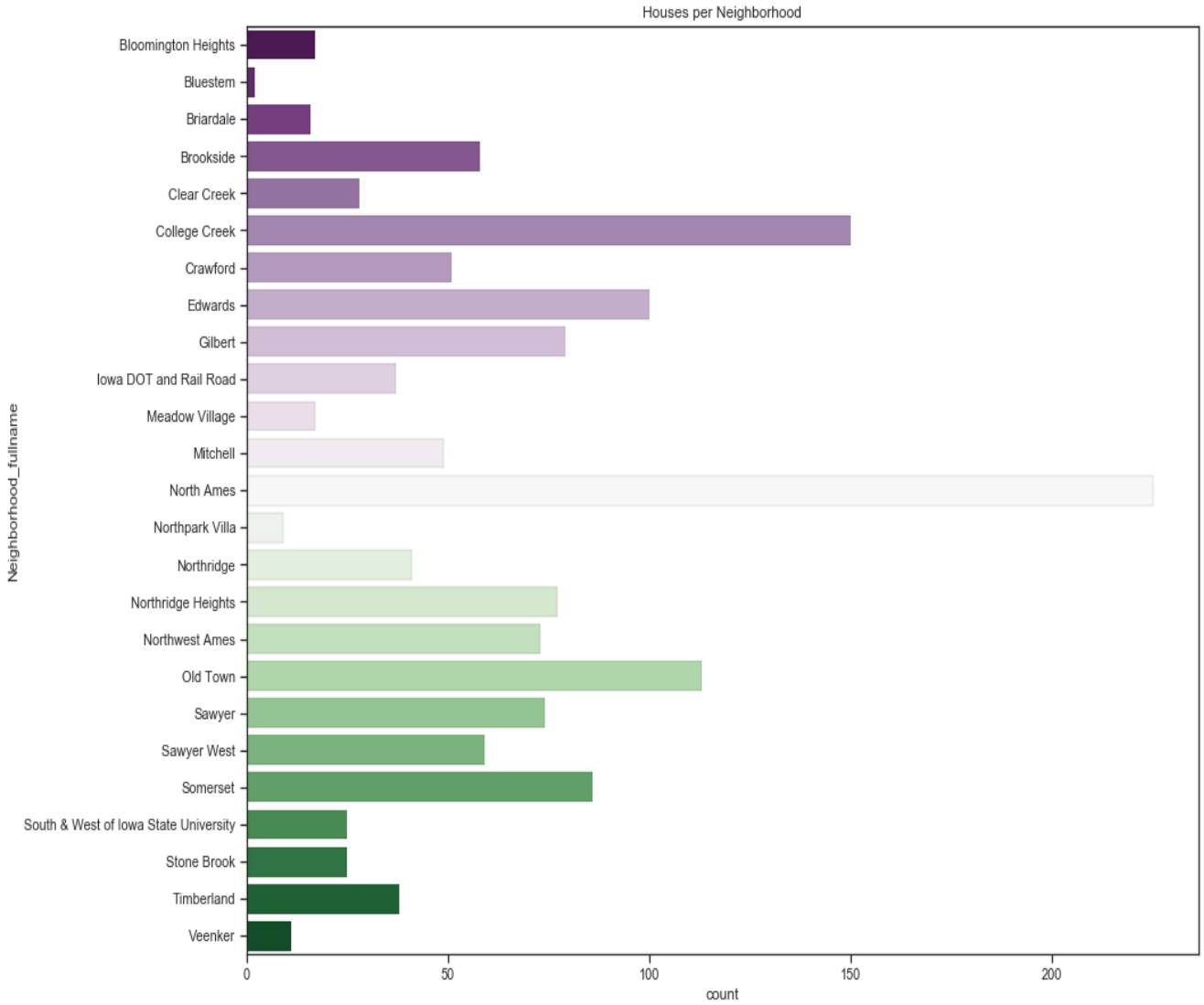


Figure 3

This graph demonstrates that North Ames has, by far, the largest count of houses per neighborhood. Bluestem holds the least amount of houses followed by Northpark villa and Veenker.

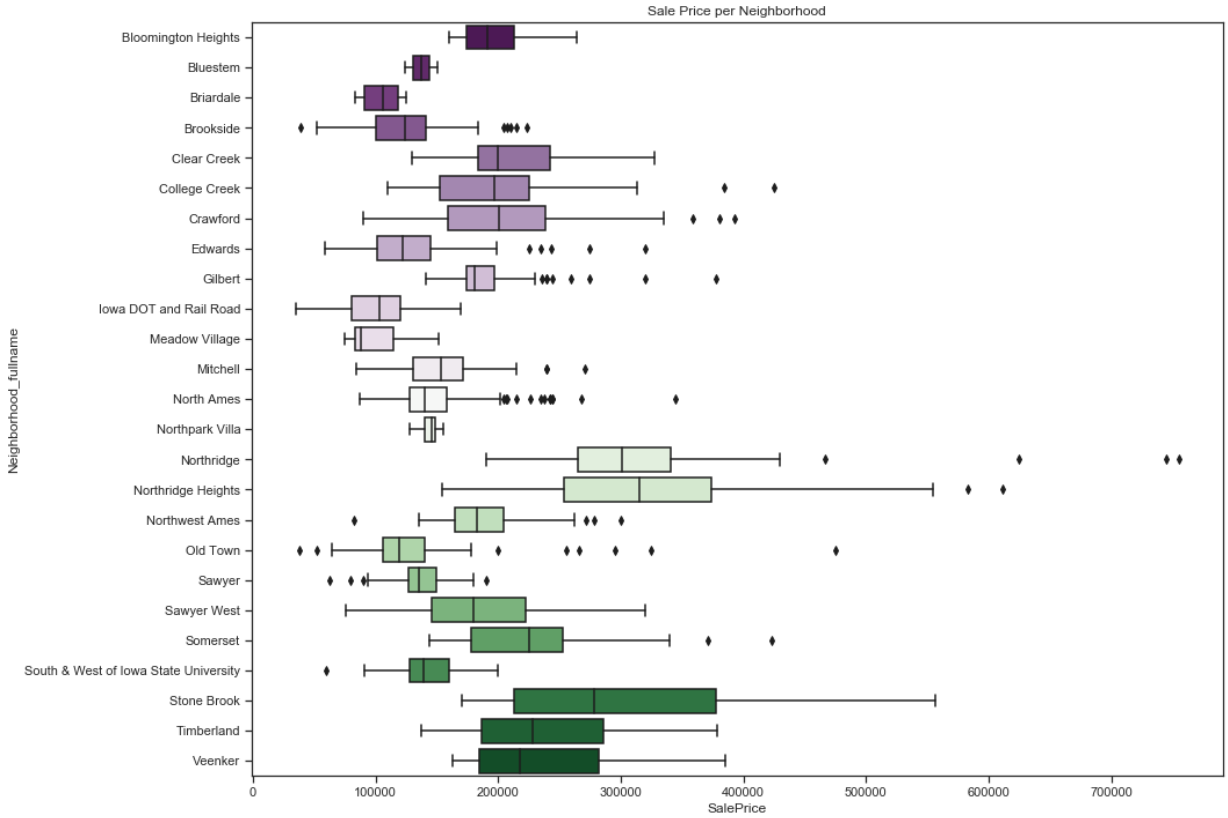


Figure 4

Sale prices of houses in the North Ames region remain fairly low, similar cost to those in the Northpark Villa region. This demonstrates that the number of houses in a neighbourhood may not be a reliable factor in determining the fluctuating sale prices.

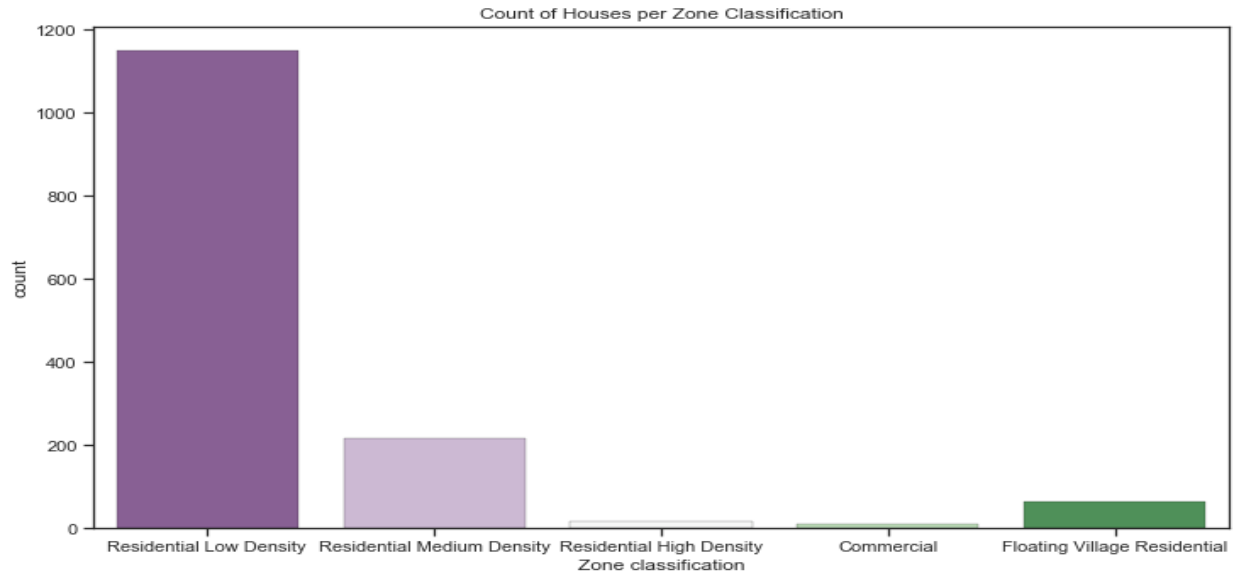


Figure 5

This graph demonstrates that their are substantially more houses in Residential low density zones. Commercial zones and Residential high density zones have the lowest count of houses. This demonstrates that the zone classification is a factor influencing the count of houses.

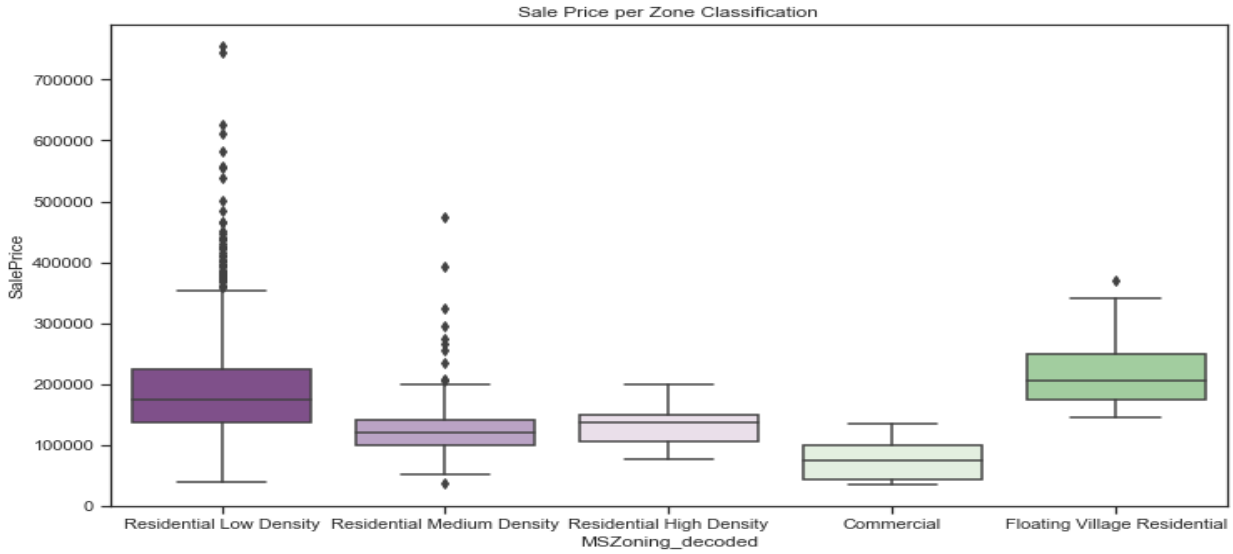


Figure 6

The range in sale price of Residential low density zones appears to be larger than the other zones. Floating village residential although containing a lower count of houses, has the highest median sale price. Commercial zones have the lowest sale prices. This demonstrates that the zone of a house plays a factor in the sale price of the houses.

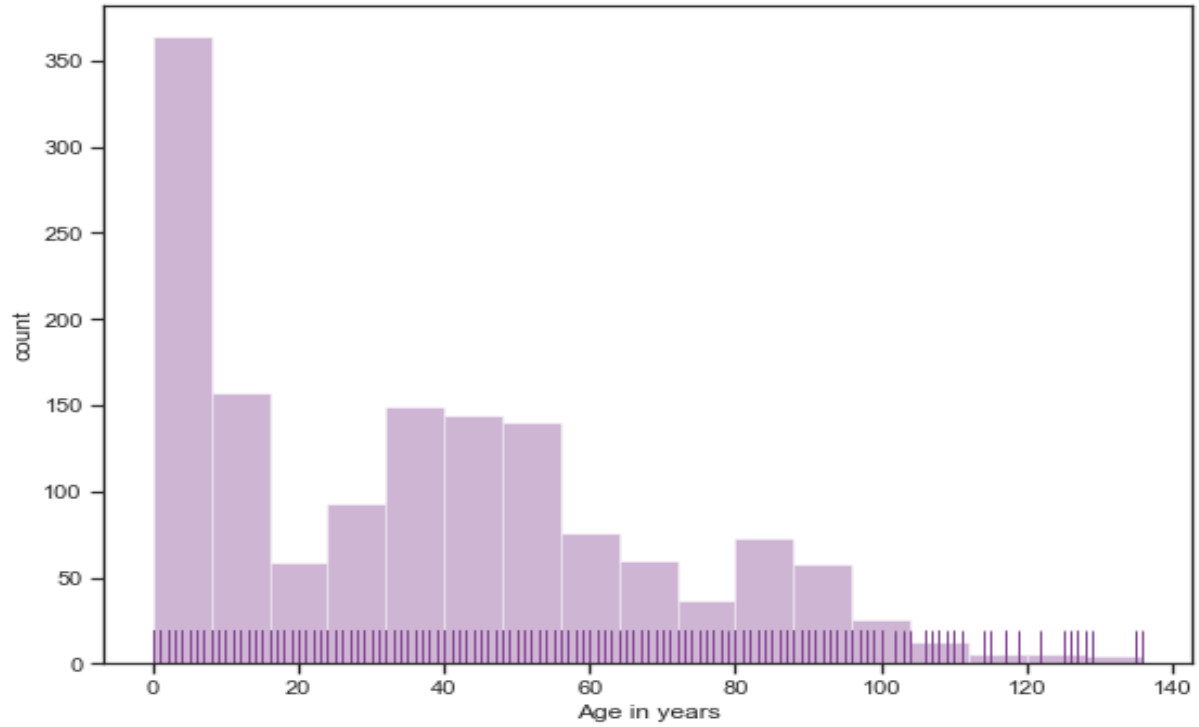


Figure 7

This figure demonstrates that there are substantially more new houses aging only 0-10 years. Houses aging more than 100 years have a much lower count. We can conclude that in recent years there has been a large influx in houses built and a large portion of older houses being demolished.

REFERENCES

Grace-Martin - 2019
Sematech - 2019 version
“House Prices: Advanced Regression Techniques.” *Kaggle*, 2019, www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

Detailed Data preparation steps

The following section describes each variable in the “Train” dataset and if we did anything in terms of imputing the data. The variables of importance to us have been highlighted and any imputation were carried forward to the “Test” data set as well:

1. ID – Unique Identifier
2. MS Subclass 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

This is the frequency distribution

20	536
60	299
50	144
120	87
30	69
160	63
70	60
80	58
90	52
190	30
85	20
75	16
45	12
180	10
40	4

Because the frequency distribution is too less for subclass 190 , 180, 85, 75, 45, 40 - we will club all these under “others category” . We created dummy variables for each of the categories and add them into our dataset

3. MSZONING

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

In the Train Data set this is the frequency distribution

RL	1151
RM	218
FV	65
RH	16
C (all)	10

We will create dummy variables for RL, RM and FV categories and club all others into OTHER_MSZONE.

4. Lot Frontage - Linear feet of street connected to property

The values range from 21 to 313. Median is 68. 87 values are beyond the interquartile range but they are genuine entries

5. LotArea: Lot size in square feet

The Values range from 1300 to 215242. 69 values fall as outliers but we intend to keep them because they are genuine

6. Street: Type of road access to property

Grvl	Gravel
Pave	Paved

Only 6 values correspond to Gravel out of 1460 in the train data set. This is not suitable for analysis and we will be ignoring this column for that purpose

Alley: Type of alley access to property

Grvl	Gravel
Pave	Paved
NA	No alley access

This is the frequency distribution

Grvl	50
NA	1369
Pave	41

We will be making categorical variables for each

Lot Shape: General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

This is the distribution of the data

Reg	925
IR1	484
IR2	41
IR3	10

Categorical variables will be made for Reg and IR1. IR2 and IR3 will be clubbed as other

Land Contour: 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

This is the distribution

Lvl	1311
Bnk	63
HLS	50
Low	36

We Will make categorical variables for each value

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

This is the distribution

AllPub	1457
NA	2
NoSeWa	1

Since all almost entries have all public utilities it does not make sense to include this in our analysis or make categorical variables here

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

The frequency distribution for this variable is

Inside	1052
Corner	263
CulDSac	94
FR2	47
FR3	4

We will make dummy variables for Inside, Corner and CulDsac. We will club the remaining categories.

LandSlope: Slope of property

Gtl	Gentle slope
Mod	Moderate Slope
Sev	Severe Slope

This is the distribution

Gtl	1382
Mod	65
Sev	13

We will create dummy variables for each of the above categories

Neighbourhood : 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

This is the frequency distribution for the neighbourhood variables

NAmes	225
CollgCr	150
OldTown	113
Edwards	100
Somerst	86
Gilbert	79
NridgHt	77
Sawyer	74
NWAmes	73

SawyerW	59
BrkSide	58
Crawfor	51
Mitchel	49
NoRidge	41
Timber	38
IDOTRR	37
ClearCr	28
SWISU	25
StoneBr	25
MeadowV	17
Blmngtn	17
BrDale	16
Veenker	11
NPkVill	9
Blueste	2

We will create independent dummy variables for all categories with more than 50 values. For the remaining we will club those categories under the others heading

Condition 1 : 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

This is the frequency distribution

Norm	1260
Feedr	81
Artery	48
RRAn	26
PosN	19
RRAe	11
PosA	8
RRNn	5
RRNe	2

We will make Dummy variables for Norm, Feedr, Artery and club the remaining under others

Condition 2 : 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

The frequency Distribution of this variable is

Norm	1445
Feedr	6
RRNn	2
PosN	2
Artery	2
RRAn	1
RRAe	1
PosA	1

Since almost all the values are “normal” it is not suitable for our analysis as the data is too skewed and we will not make dummy variables

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

The frequency distribution for this is -

1Fam	1220
2fmCon	31
Duplex	52
Twnhs	43

TwnhsE	114

We will create dummy variables for each category

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

The frequency distribution for this variable is –

1Story	726
2Story	445
1.5Fin	154
SLvl	65
SFoyer	37
1.5Unf	14
2.5Unf	11
2.5Fin	8

We will make dummy variables for 1story,2story,1.5Fin,SLvl and club the remaining into others

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

We will use these variables as numerical variables

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

We will use these variables as numerical variables

YearBuilt: Original construction date

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

YrSold: Year Sold (YYYY)

Since the YearRemodAdd will be the same as Yearbuilt for houses that did not have a remodelling done we will convert these variables into 2 new variables

A) Age of the house from the more recent of year built or year remod add
B) Whether remodelling took place or not will be depicted with a binary variable

7. Roof Style : Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

The frequency distribution of this variable is –

Gable	1141
Hip	286
Flat	13
Gambrel	11
Mansard	7

Shed	2

We will create dummy variables for Gabby, Hip and club the remaining into other roof styles

8. Roof Matl : 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

The frequency distribution is given below

CompShg	1434
Tar&Grv	11
WdShngl	6
WdShake	5
Roll	1
Membran	1
ClyTile	1
Metal	1

Since more than 98% of the values are in one category, it is not suitable to be used for analysis and we will not make dummy variables for this

9. 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

10. 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

Most of the times Exterior 1 and Exterior 2 are the same value

AsbShng	AsbShng	17
	Plywood	2
	Stucco	1
AsphShn	AsphShn	1
BrkComm	Brk Cmn	2
BrkFace	AsbShng	1
	BrkFace	24
	HdBoard	3
	Plywood	6
	Stone	2
	Stucco	1
	Wd Sdng	12
	Wd Shng	1
CBlock	CBlock	1
CemntBd	CmentBd	59
	Wd Sdng	1
	Wd Shng	1

HdBoard	AsphShn	1
	HdBoard	193
	ImStucc	2
	MetalSd	1
	Plywood	23
	Wd Sdng	1
	Wd Shng	1
ImStucc	ImStucc	1
MetalSd	AsphShn	1
	HdBoard	3
	MetalSd	212
	Stucco	1
	Wd Sdng	2
	Wd Shng	1
Plywood	Brk Cmn	5
	HdBoard	2
	ImStucc	3
	Plywood	96
	Wd Sdng	2
Stone	HdBoard	1
	Stone	1
Stucco	CmentBd	1
	Stone	1
	Stucco	20
	Wd Shng	3
VinylSd	AsbShng	1
	HdBoard	1
	ImStucc	1
	Other	1
	Plywood	2
	Stucco	1
	VinylSd	502
	Wd Sdng	1
	Wd Shng	5
Wd Sdng	AsbShng	1
	BrkFace	1
	HdBoard	2
	ImStucc	3
	MetalSd	1
	Plywood	8
	Stone	1
	Stucco	1
	VinylSd	2
	Wd Sdng	177
	Wd Shng	9
WdShing	HdBoard	2

Plywood	5
Stucco	1
Wd Sdng	1
Wd Shng	17

Hence we will make dummy variables for only Exterior 1

11. MasVnrType : Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
NoneNone	
Stone	Stone

The frequency distribution of this variable is

None	864
BrkFace	445
Stone	128
BrkCmn	15
NA	8

This table also contains some NA Values which will be clubbed with "others"
We will make dummy variables for None, BrkFAce,Stone and Others

12. MasVnrArea: Masonry veneer area in square feet

    This ranges from 0 to 1600 will be taken as a numeric variable

13. Centralair – binary variable indicating whether the house has central airconditioning or not. 95 values indicate no centralair and 1365 values indicate centralair

14. Exterqual - Evaluates the quality of the material on the exterior

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

The frequency distribution for this is –

TA	906

Gd	488
Ex	52
Fa	14

We will make categorical variables for TA , GD and Ex. We will club the other outcomes

15. Extercond : Evaluates the quality of the material on the exterior

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

TA	1282
Gd	146
Fa	28
Ex	3
Po	1

We will make categorical variables for TA, Gd and club the remaining into others

`30. Foundation – Type of foundation

BrkTil	Brick & Tile
CBlock	Cinder Block
PConc	Poured Contrete
Slab	Slab
Stone	Stone
Wood	Wood

Frequency distribution is given as follows

PConc	647
CBlock	634
BrkTil	146
Slab	24
Stone	6
Wood	3

We will make dummy variables for Pconc, CBlock, BrkTil and club the others into one dummy variable

16. 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

The frequency distribution is

TA	649
Gd	618
Ex	121
NA	37
Fa	35

We will make dummy variables for the top 3 categories and club the others

17. 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

The frequency distribution of this variable is

TA	1311
Gd	65
Fa	45
NA	37
Po	2

We will make dummy variables for top 3 categories and club the others

18. 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

The frequency distribution of this variable is

No	953
Av	221
Gd	134
Mn	114
NA	38

We will create dummy variables for each of the above categories

19. 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

The frequency distribution of this variable is as follows

Unf	430
GLQ	418
ALQ	220
BLQ	148
Rec	133
LwQ	74
NA	37

We will make dummy variables for all categories

20. BsmtFinSF1: Type 1 finished square feet

We will treat this as a numerical variable. The values range from 0 to 5644. The outliers do not skew the calculations and are relevant for them

21. 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

The frequency distribution for this variable is

```
We will create dummy variables for the top 4 categories and club the
remaining ones

  22.  BsmtFinSF2: Type 2 finished square feet
  23.  BsmtUnfSF: Unfinished square feet of basement area
  24.  TotalBsmtSF: Total square feet of basement area

     Above three varables will be considered numeric variables. Outliers do
     not skew model calculations and are genuine values.

  25.  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
```

Since more almost 98% of the values are for GasA , this variable is not suitable for use in the model and analysis and we won't make dummy variables for this

HeatingQC: Heating quality and condition

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

The frequency distribution for this variable is

Ex	741
TA	428
Gd	241
Fa	49
Po	1

We will make dummy variables for top 3 categories and club the remaining into others

26. 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

The frequency distribution of this variable is

SBrkr	1334
FuseA	94
FuseF	27
FuseP	3
NA	1
Mix	1

We will make dummy variables for top 2 categories and club the remaining into others

27. 1stFlrSF: First Floor square feet

This is considered as a numerical variable. The values range from 334 to 5642. Outliers do not skew calculations and contain genuine values

28. 2ndFlrSF: Second floor square feet

This is considered as a numerical variable. The values range from 0 to 2165. Outliers do not skew calculations and contain genuine values

29. LowQualFinSF: Low quality finished square feet (all floors)

For 1434 entries this value is 0. Hence it will not be used for analysis

30. GrLivArea: Above grade (ground) living area square feet

This is considered as a numerical variable. The values range from 334 to 5642. Outliers do not skew calculations and contain genuine values

31. BsmtFullBath: Basement full bathrooms

This is considered as a numerical variable. Values range from 0-3, no outliers

32. BsmtHalfBath: Basement half bathrooms

This is considered as a numerical variable. Values range from 0-2, no outliers

33. FullBath: Full bathrooms above grade

This is considered as a numerical variable. Values range from 0-3, no outliers

34. HalfBath: Half baths above grade

This is considered as a numerical variable. Values range from 0-2, no outliers

35. Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

This is considered as a numerical variable. Values range from 0-8, no outliers

36. Kitchen: Kitchens above grade

This is considered as a numerical variable. Values range from 0-3, outliers do not skew the calculations

37. KitchenQual: Kitchen quality

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

The frequency distribution for this is

TA	735
Gd	586
Fa	39
Ex	100

We will make dummy variables for the top 3 categories and club the remaining 2

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

This will be taken as a numerical variable. It ranges from 0-8 and the outliers do not skew the model but are genuine values

39. 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
SalSalvage only

The frequency distribution of this variable is

Typ	1360
Min2	34
Min1	31
Mod	15
Maj1	14
Maj2	5
Sev	1

The top 3 categories will have dummy variables created for them and the remaining will be clubbed into other
Fireplaces: Number of fireplaces
This is taken as a numerical variable. It ranges from 0-3. Outlier are not there

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

The frequency distribution for the variable is –

Attchd	870
Detchd	387
BuiltIn	88
NA	81
Basment	19
CarPort	9
2Types	6

We will make dummy variables for top 4 categories and club the remaining ones

GarageYrBlt: Year garage was built
We will convert this variable to a numerical variable indicating age of garage when sold. Values containing NA will be treated as 0 and the value of 2207 is corrected to 2007.

GarageFinish: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

The frequency distribution of the variable is

Fin	352
NA	81
RFn	422
Unf	605

We will make separate dummy variables for all categories

GarageCars: Size of garage in car capacity

This is a numeric variable ranging from 0-5. No outliers

GarageArea: Size of garage in square feet
This is a numerical variable ranging from 0 to 1600. Outliers are genuine values and do not skew the model results

GarageQual: Garage quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

The frequency distribution for this variable is

TA	1311
NA	81
Fa	48
Gd	14
Ex	3
Po	3

We will make dummy variables for the top 3 categories and club the remaining into others

GarageCond: Garage condition

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

The frequency distribution of this variable

TA	1326
NA	81
Fa	35
Gd	9
Po	7
Ex	2

We will make dummy variables for the top 3 categories and club others

PavedDrive: Paved driveway

Y Paved

P Partial Pavement
N Dirt/Gravel

The frequency distribution of the variable is –

Y	1340
N	90
P	30

We will make dummy variables for each of the categories

WoodDeckSF: Wood deck area in square feet
This is a numerical variable and ranges from 0 to 857. Outliers are genuine values and do not skew the model.

OpenPorchSF: Open porch area in square feet
This is a numerical variable and ranges from 0 to 547. Outliers are not there

EnclosedPorch: Enclosed porch area in square feet
This is a numerical variable and ranges from 0 to 552. Outlier values do not skew calculation

3SsnPorch: Three season porch area in square feet
For 1436 out of 1460 entries this value is zero. This is not suitable for our modelling analysis

ScreenPorch: Screen porch area in square feet
This is a numerical variable which ranges from 0 to 480. Outliers are genuine values which we need in the modelling analysis

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool

1453 out of 1460 entries have no pool, hence both #70 and #71 variables are not suitable for use in our analysis

Fence: Fence quality

GdPrv	Good Privacy
MnPrv	Minimum Privacy
GdWo	Good Wood
MnWw	Minimum Wood/Wire
NA	No Fence

The frequency distribution of this variable is –

NA	1179
MnPrv	157
GdPrv	59
GdWo	54
MnWw	11

We will make dummy variables for the top 2 categories and club the remaining ones

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

NA	1406
Shed	49
Gar2	2
Othr	2
TenC	1

We will make dummy variables for top 2 categories and club the remaining ones

MiscVal: $Value of miscellaneous feature
This is a numerical variable. Outliers are genuine values and needed for the modelling analysis

MoSold: Month Sold (MM)

The frequency distribution for this variable is

6	253
7	234
5	204
4	141
8	122
3	106
10	89
11	79
9	63
12	59
1	58
2	52

We will make dummy variables for each of the months

74. 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
OthOther	

The frequency distribution of this variable is

WD	1267
New	122
COD	43
ConLD	9
ConLI	5
ConLw	5
CWD	4
Oth	3
Con	2

We will make dummy variables for top 3 categories and club the remaining under “OTHER TYPES”

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)

The frequency distribution for this variable is –

Normal	1198
Partial	125
Abnorml	101
Family	20
Alloca	12
AdjLand	4

We will create dummy variables for the first three categories and club the others