# Housing Price Prediction Model using modeldata package - a subset of ames housing data

Saurav Mukherjee

2023-02-16

## Introduction

I am building home price prediction model. I am using Ames Housing dataset to explore the attributes which have been identified somehow influencing the housing cost.

Initially I wanted to use the 'Ames Housing Data" - a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. Howvever, I looked at a dataset which is a subset of this dataset and which is available within modeldata package created by https://modeldata.tidymodels.org/. I did some research and looked at the model - Hedonic Pricing Method to predict the house price. The Hedonic Pricing Method talks about internal characteristics as well as the external factors affecting the price of a good. Based on the idea of hedonic price modeling I am looking the is that neighborhood-specific and unit-specific characteristics help determine house prices.

### Data - Ames Housing Data

A data set from De Cock (2011) has 82 fields were recorded for 2,930 properties in Ames IA. I used a version from the package modeldata dataset name as ames which is copies from the original AmesHousing package but does not include a few quality columns that appear to be outcomes rather than predictors.

### Load required Libraries

### Load ames dataset

### Setup environments

## Exploratory Data Analysis

### Explore Ame Dataset - Dimension, Columns and Datatypes

### Explore Sales Price Distribution

Table 1: Ames Housing Dataset dimension

| x |
|---:|
| 2930 |
| 74 |

```
## [1] "Ames Housing Dataset Columns"
```

```
##  [1] "MS_SubClass"        "MS_Zoning"          "Lot_Frontage"
##  [4] "Lot_Area"           "Street"             "Alley"
##  [7] "Lot_Shape"          "Land_Contour"       "Utilities"
## [10] "Lot_Config"         "Land_Slope"         "Neighborhood"
## [13] "Condition_1"        "Condition_2"        "Bldg_Type"
## [16] "House_Style"        "Overall_Cond"       "Year_Built"
## [19] "Year_Remod_Add"     "Roof_Style"         "Roof_Matl"
## [22] "Exterior_1st"       "Exterior_2nd"       "Mas_Vnr_Type"
## [25] "Mas_Vnr_Area"       "Exter_Cond"         "Foundation"
## [28] "Bsmt_Cond"          "Bsmt_Exposure"      "BsmtFin_Type_1"
## [31] "BsmtFin_SF_1"       "BsmtFin_Type_2"     "BsmtFin_SF_2"
## [34] "Bsmt_Unf_SF"        "Total_Bsmt_SF"      "Heating"
## [37] "Heating_QC"         "Central_Air"        "Electrical"
## [40] "First_Flr_SF"       "Second_Flr_SF"      "Gr_Liv_Area"
## [43] "Bsmt_Full_Bath"     "Bsmt_Half_Bath"     "Full_Bath"
## [46] "Half_Bath"          "Bedroom_AbvGr"      "Kitchen_AbvGr"
## [49] "TotRms_AbvGrd"      "Functional"         "Fireplaces"
## [52] "Garage_Type"        "Garage_Finish"      "Garage_Cars"
## [55] "Garage_Area"        "Garage_Cond"        "Paved_Drive"
## [58] "Wood_Deck_SF"       "Open_Porch_SF"      "Enclosed_Porch"
## [61] "Three_season_porch" "Screen_Porch"       "Pool_Area"
## [64] "Pool_QC"            "Fence"              "Misc_Feature"
## [67] "Misc_Val"           "Mo_Sold"            "Year_Sold"
## [70] "Sale_Type"          "Sale_Condition"     "Sale_Price"
## [73] "Longitude"          "Latitude"
```

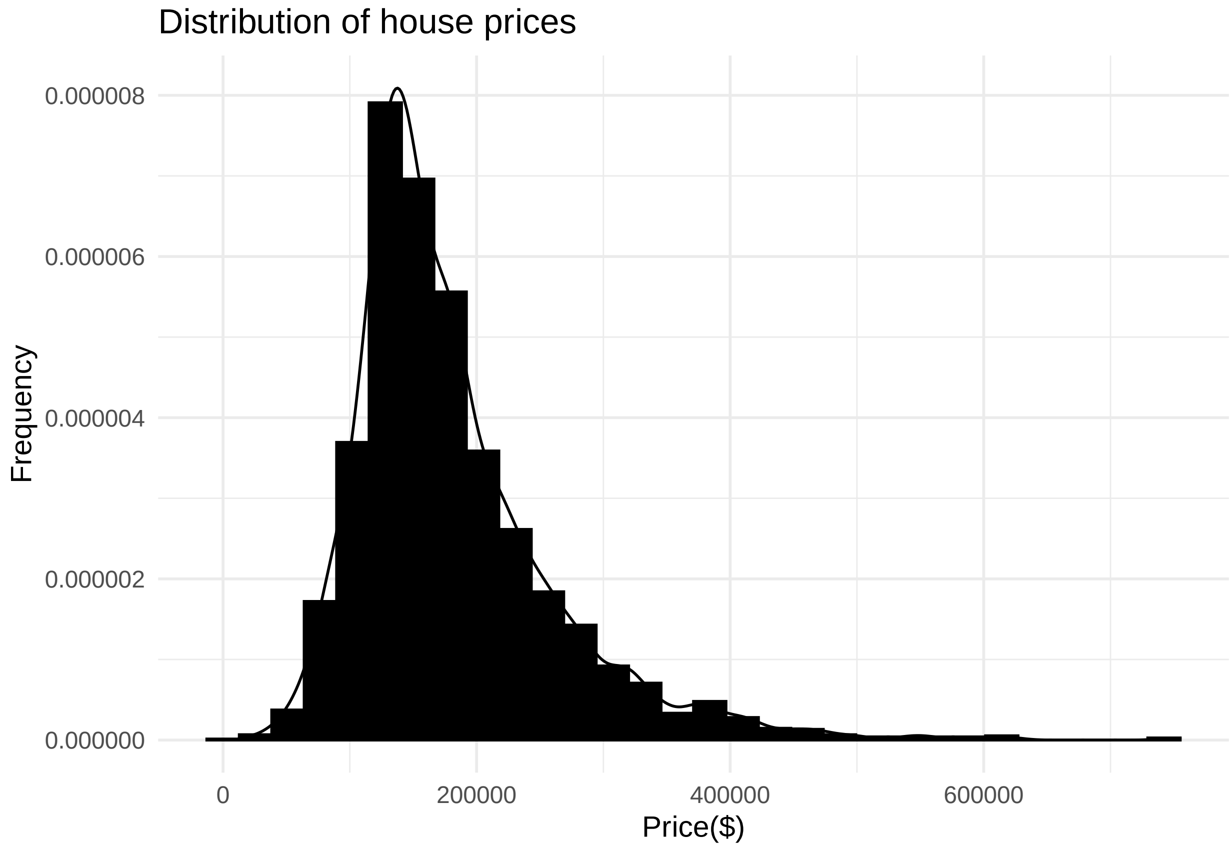
```
## tibble [2,930 x 74] (S3: tbl_df/tbl/data.frame)
##  $ MS_SubClass       : Factor w/ 16 levels "One_Story_1946_and_Newer_All_Styles",..: 1 1 1 1 6 6 12
##  $ MS_Zoning         : Factor w/ 7 levels "Floating_Village_Residential",..: 3 2 3 3 3 3 3 3 3 3 ...
##  $ Lot_Frontage      : num [1:2930] 141 80 81 93 74 78 41 43 39 60 ...
##  $ Lot_Area          : int [1:2930] 31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
##  $ Street            : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley             : Factor w/ 3 levels "Gravel","No_Alley_Access",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Lot_Shape         : Factor w/ 4 levels "Regular","Slightly_Irregular",..: 2 1 2 1 2 2 1 2 2 1 ...
##  $ Land_Contour      : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 2 4 4 ...
##  $ Utilities         : Factor w/ 3 levels "AllPub","NoSeWa",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Lot_Config        : Factor w/ 5 levels "Corner","CulDSac",..: 1 5 1 1 5 5 5 5 5 5 ...
##  $ Land_Slope        : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood      : Factor w/ 29 levels "North_Ames","College_Creek",..: 1 1 1 1 7 7 17 17 17 7 .
##  $ Condition_1       : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 3 3 3 ...
##  $ Condition_2       : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Bldg_Type         : Factor w/ 5 levels "OneFam","TwoFmCon",..: 1 1 1 1 1 1 5 5 5 1 ...
##  $ House_Style       : Factor w/ 8 levels "One_and_Half_Fin",..: 3 3 3 3 8 8 3 3 3 8 ...
##  $ Overall_Cond      : Factor w/ 10 levels "Very_Poor","Poor",..: 5 6 6 5 5 6 5 5 5 5 ...
```

```
##  $ Year_Built        : int [1:2930] 1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Year_Remod_Add    : int [1:2930] 1960 1961 1958 1968 1998 1998 2001 1992 1996 1999 ...
##  $ Roof_Style        : Factor w/ 6 levels "Flat","Gable",..: 4 2 4 4 2 2 2 2 2 2 ...
##  $ Roof_Matl         : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior_1st      : Factor w/ 16 levels "AsbShng","AsphShn",..: 4 14 15 4 14 14 6 7 6 14 ...
##  $ Exterior_2nd      : Factor w/ 17 levels "AsbShng","AsphShn",..: 11 15 16 4 15 15 6 7 6 15 ...
##  $ Mas_Vnr_Type      : Factor w/ 5 levels "BrkCmn","BrkFace",..: 5 4 2 4 4 2 4 4 4 4 ...
##  $ Mas_Vnr_Area      : num [1:2930] 112 0 108 0 0 20 0 0 0 0 ...
##  $ Exter_Cond        : Factor w/ 5 levels "Excellent","Fair",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation        : Factor w/ 6 levels "BrkTil","CBlock",..: 2 2 2 2 3 3 3 3 3 3 ...
##  $ Bsmt_Cond         : Factor w/ 6 levels "Excellent","Fair",..: 3 6 6 6 6 6 6 6 6 6 ...
##  $ Bsmt_Exposure     : Factor w/ 5 levels "Av","Gd","Mn",..: 2 4 4 4 4 4 3 4 4 4 ...
##  $ BsmtFin_Type_1    : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 2 6 1 1 3 3 3 1 3 7 ...
##  $ BsmtFin_SF_1      : num [1:2930] 2 6 1 1 3 3 3 1 3 7 ...
##  $ BsmtFin_Type_2    : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 7 4 7 7 7 7 7 7 7 7 ...
##  $ BsmtFin_SF_2      : num [1:2930] 0 144 0 0 0 0 0 0 0 0 ...
##  $ Bsmt_Unf_SF       : num [1:2930] 441 270 406 1045 137 ...
##  $ Total_Bsmt_SF     : num [1:2930] 1080 882 1329 2110 928 ...
##  $ Heating           : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Heating_QC        : Factor w/ 5 levels "Excellent","Fair",..: 2 5 5 1 3 1 1 1 1 3 ...
##  $ Central_Air       : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical        : Factor w/ 6 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ First_Flr_SF      : int [1:2930] 1656 896 1329 2110 928 926 1338 1280 1616 1028 ...
##  $ Second_Flr_SF     : int [1:2930] 0 0 0 0 701 678 0 0 0 776 ...
##  $ Gr_Liv_Area       : int [1:2930] 1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
##  $ Bsmt_Full_Bath    : num [1:2930] 1 0 0 1 0 0 1 0 1 0 ...
##  $ Bsmt_Half_Bath    : num [1:2930] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Full_Bath         : int [1:2930] 1 1 1 2 2 2 2 2 2 2 ...
##  $ Half_Bath         : int [1:2930] 0 0 1 1 1 1 0 0 0 1 ...
##  $ Bedroom_AbvGr     : int [1:2930] 3 2 3 3 3 3 2 2 2 3 ...
##  $ Kitchen_AbvGr     : int [1:2930] 1 1 1 1 1 1 1 1 1 1 ...
##  $ TotRms_AbvGrd     : int [1:2930] 7 5 6 8 6 7 6 5 5 7 ...
##  $ Functional        : Factor w/ 8 levels "Maj1","Maj2",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ Fireplaces        : int [1:2930] 2 0 0 2 1 1 0 0 1 1 ...
##  $ Garage_Type       : Factor w/ 7 levels "Attchd","Basment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Garage_Finish     : Factor w/ 4 levels "Fin","No_Garage",..: 1 4 4 1 1 1 1 3 3 1 ...
##  $ Garage_Cars       : num [1:2930] 2 1 1 2 2 2 2 2 2 2 ...
##  $ Garage_Area       : num [1:2930] 528 730 312 522 482 470 582 506 608 442 ...
##  $ Garage_Cond       : Factor w/ 6 levels "Excellent","Fair",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ Paved_Drive       : Factor w/ 3 levels "Dirt_Gravel",..: 2 3 3 3 3 3 3 3 3 3 ...
##  $ Wood_Deck_SF      : int [1:2930] 210 140 393 0 212 360 0 0 237 140 ...
##  $ Open_Porch_SF     : int [1:2930] 62 0 36 0 34 36 0 82 152 60 ...
##  $ Enclosed_Porch    : int [1:2930] 0 0 0 0 0 0 170 0 0 0 ...
##  $ Three_season_porch: int [1:2930] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Screen_Porch      : int [1:2930] 0 120 0 0 0 0 0 144 0 0 ...
##  $ Pool_Area         : int [1:2930] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Pool_QC           : Factor w/ 5 levels "Excellent","Fair",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Fence             : Factor w/ 5 levels "Good_Privacy",..: 5 3 5 5 3 5 5 5 5 5 ...
##  $ Misc_Feature      : Factor w/ 6 levels "Elev","Gar2",..: 3 3 2 3 3 3 3 3 3 3 ...
##  $ Misc_Val          : int [1:2930] 0 0 12500 0 0 0 0 0 0 0 ...
##  $ Mo_Sold           : int [1:2930] 5 6 6 4 3 6 4 1 3 6 ...
##  $ Year_Sold         : int [1:2930] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Sale_Type         : Factor w/ 10 levels "COD","Con","ConLD",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ Sale_Condition    : Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 5 5 5 5 5 5 5 ...
```

```
##  $ Sale_Price         : int [1:2930] 215000 105000 172000 244000 189900 195500 213500 191500 236500 18
##  $ Longitude          : num [1:2930] -93.6 -93.6 -93.6 -93.6 -93.6 ...
##  $ Latitude           : num [1:2930] 42.1 42.1 42.1 42.1 42.1 ...
```

Table: Ames Housing Dataset

|| || || ||

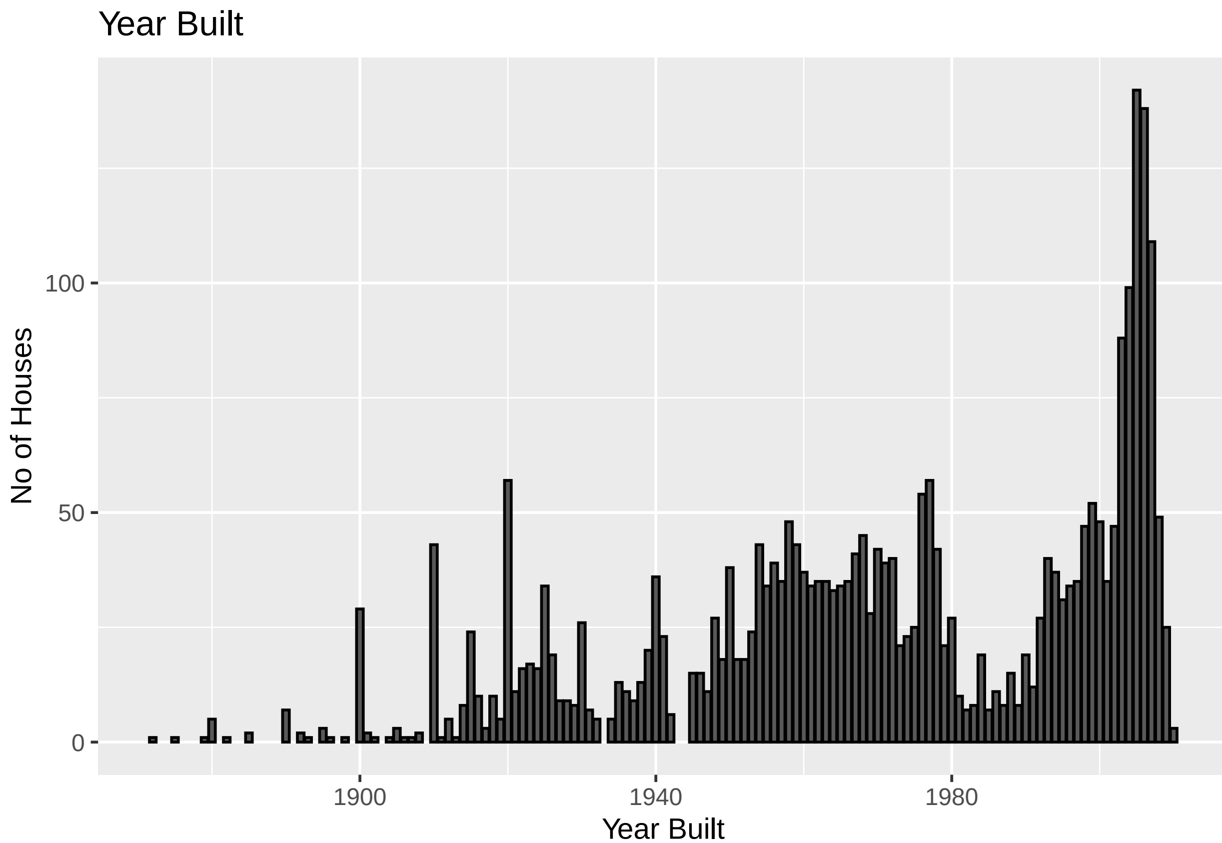



## Sale Price Observation The Sale Price is right-skewed

```
##
## Sale Price skewness : 1.742607
```

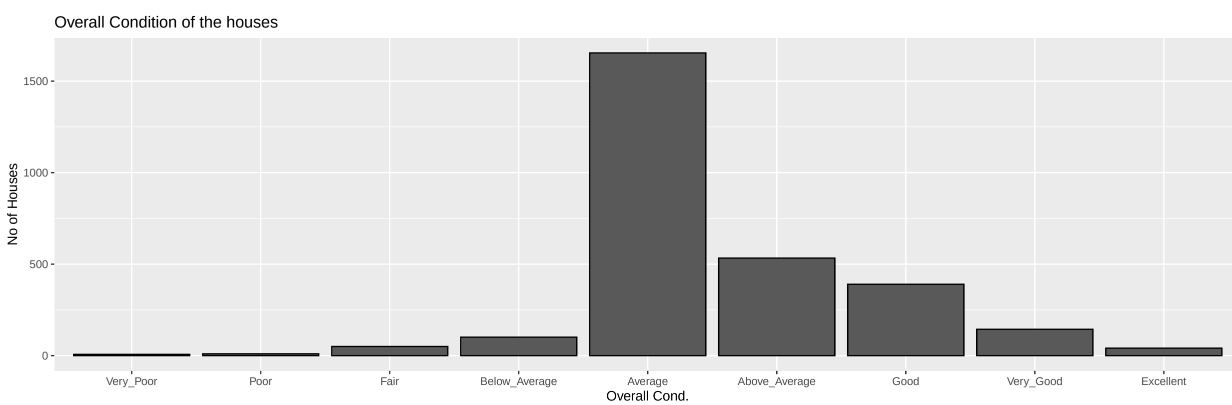
```
##
## Sale Price kurtosis : 8.108122
```

## Houses and Year Built



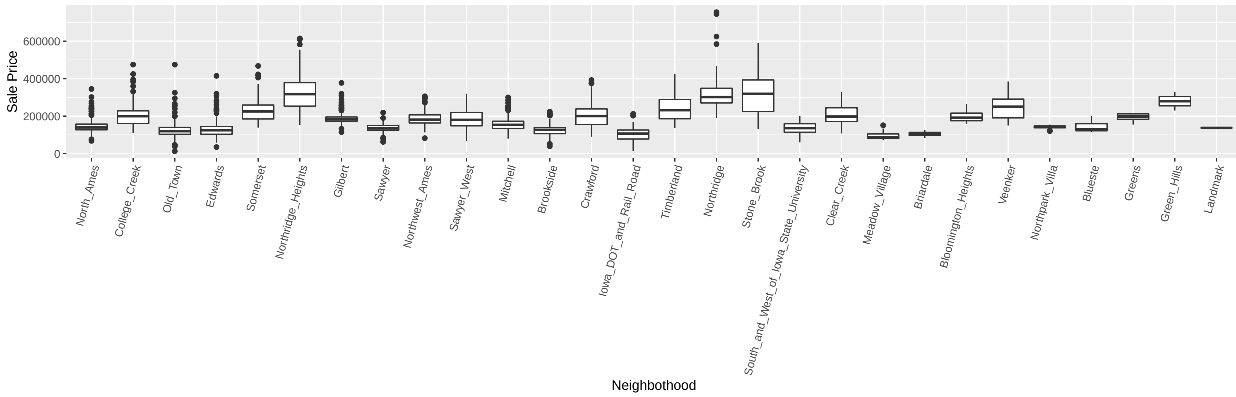

It looks that we have more houses were built at the beginning of 2000.

## Condition of the houses




House condition - most of the houses are of average condition

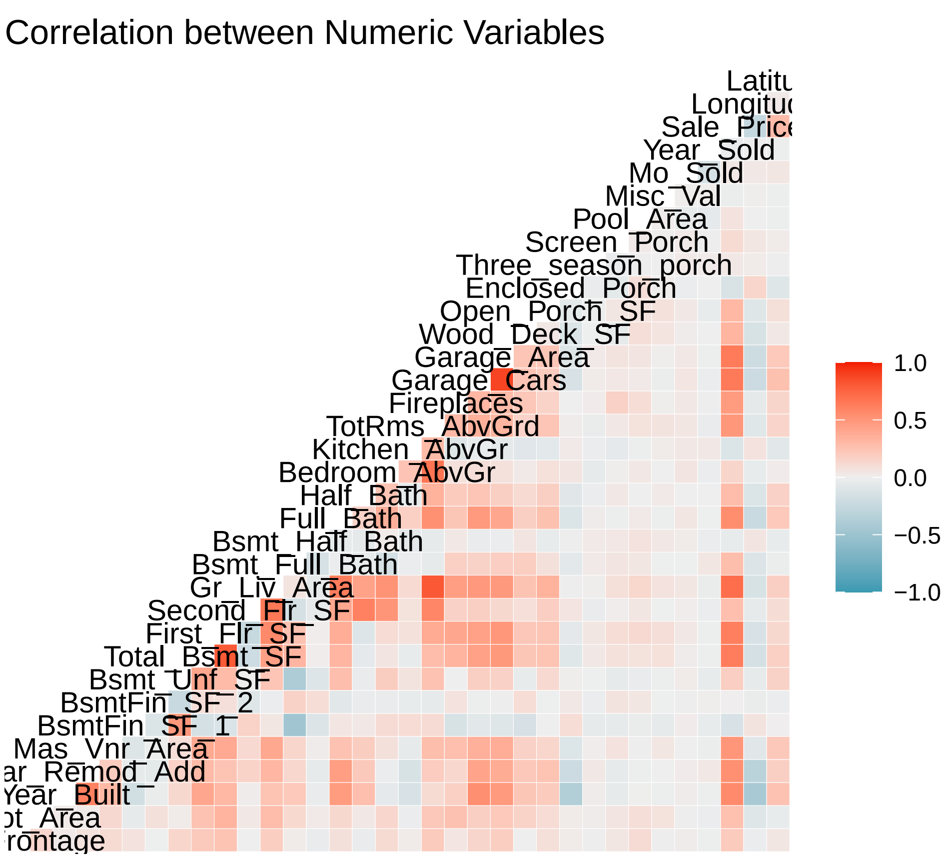
### Neighborhood and House Price




House Price varies with the neighborhood with few outliers by neighborhood. Also, the median house price by neighborhood is roughly between 200,000 and 400,000. It seems Neighborhood would have some impact on housing price.

## Correlation between Sale Price and other variables

### Correlation between numeric variables



There are some high correlations between variables mostly positive but with some negative. I did further analysis and added pairwise correlation between other numeric variables and sales price.

## Correlation of Sales Price with other numeric variables

Table 2: Ames Housing Dataset - correlated numeric variables with the Sale Price

| | x |
|---|---|
| Lot_Frontage | 0.2018745 |
| Lot_Area | 0.2665492 |
| Year_Built | 0.5584261 |
| Year_Remod_Add | 0.5329738 |
| Mas_Vnr_Area | 0.5021960 |
| BsmtFin_SF_1 | -0.1349055 |
| BsmtFin_SF_2 | 0.0060176 |
| Bsmt_Unf_SF | 0.1833076 |
| Total_Bsmt_SF | 0.6325288 |
| First_Flr_SF | 0.6216761 |
| Second_Flr_SF | 0.2693734 |
| Gr_Liv_Area | 0.7067799 |
| Bsmt_Full_Bath | 0.2758227 |
| Bsmt_Half_Bath | -0.0358166 |
| Full_Bath | 0.5456039 |
| Half_Bath | 0.2850560 |
| Bedroom_AbvGr | 0.1439134 |
| Kitchen_AbvGr | -0.1198137 |
| TotRms_AbvGrd | 0.4954744 |
| Fireplaces | 0.4745581 |
| Garage_Cars | 0.6475616 |
| Garage_Area | 0.6401383 |
| Wood_Deck_SF | 0.3271432 |
| Open_Porch_SF | 0.3129505 |
| Enclosed_Porch | -0.1287874 |
| Three_season_porch | 0.0322246 |
| Screen_Porch | 0.1121512 |
| Pool_Area | 0.0684032 |
| Misc_Val | -0.0156915 |
| Mo_Sold | 0.0352588 |
| Year_Sold | -0.0305691 |
| Sale_Price | 1.0000000 |
| Longitude | -0.2513973 |
| Latitude | 0.2908914 |

Thus, I identified variables which has higher correlations (correlation > 0.5 and < - 0.2)

I also looked at some non-numeric variables and their relatins with the Sale Price

### Correlation of Sales Price with non-numeric variables

Table 3: Ames Housing Dataset - correlated non-numeric variables with the Sale Price

| | x |
|---|---|
| MS_SubClass | -0.0347748 |
| MS_Zoning | -0.3064225 |
| Street | 0.0595193 |
| Alley | 0.1088436 |
| Lot_Shape | 0.3026647 |
| Land_Contour | -0.0693388 |
| Utilities | -0.0310365 |
| Lot_Config | -0.0587875 |
| Land_Slope | 0.0685534 |
| Neighborhood | 0.1575002 |
| Condition_1 | 0.1590773 |
| Condition_2 | 0.1048063 |
| Bldg_Type | -0.0952280 |
| House_Style | 0.2310546 |
| Overall_Cond | -0.1635790 |
| Roof_Style | 0.2546450 |
| Roof_Matl | 0.0720760 |
| Exterior_1st | 0.0550217 |
| Exterior_2nd | 0.0535448 |
| Mas_Vnr_Type | -0.0763142 |
| Exter_Cond | 0.1206939 |
| Foundation | 0.4579558 |
| Bsmt_Cond | 0.1095363 |
| Bsmt_Exposure | -0.3519094 |
| BsmtFin_Type_1 | -0.0975925 |
| BsmtFin_Type_2 | 0.1074020 |
| Heating | -0.0728977 |
| Heating_QC | -0.4426972 |
| Central_Air | 0.2645064 |
| Electrical | 0.2378218 |
| Functional | 0.1192451 |
| Garage_Type | -0.4061833 |
| Garage_Finish | -0.4494826 |
| Garage_Cond | 0.2750657 |
| Paved_Drive | 0.2749134 |
| Pool_QC | -0.0919699 |
| Fence | 0.1745827 |
| Misc_Feature | -0.0574683 |
| Sale_Type | -0.1845079 |
| Sale_Condition | 0.3330831 |

### Looking at the non-numeric variable, I identified few variables which are highly correlated -

MS_Zoning, Lot_Shape, Foundation, Sale_Condition , Garage_Finish, House_Style, Heating_QC,

## Feature Engineering and additional visualizations

**Created a variable total_area = First_Flr_SF + Second_Flr_SF + Total_Bsmt_SF**

**Created a variable total_Bathroom = Full_Bath + Bsmt_Full_Bath + 0.5* Half_Bath+ 0.5 * Bsmt_Half_Bath**

**Created a variable ornarall_Condition_n a numeric representation of overall_Condition**

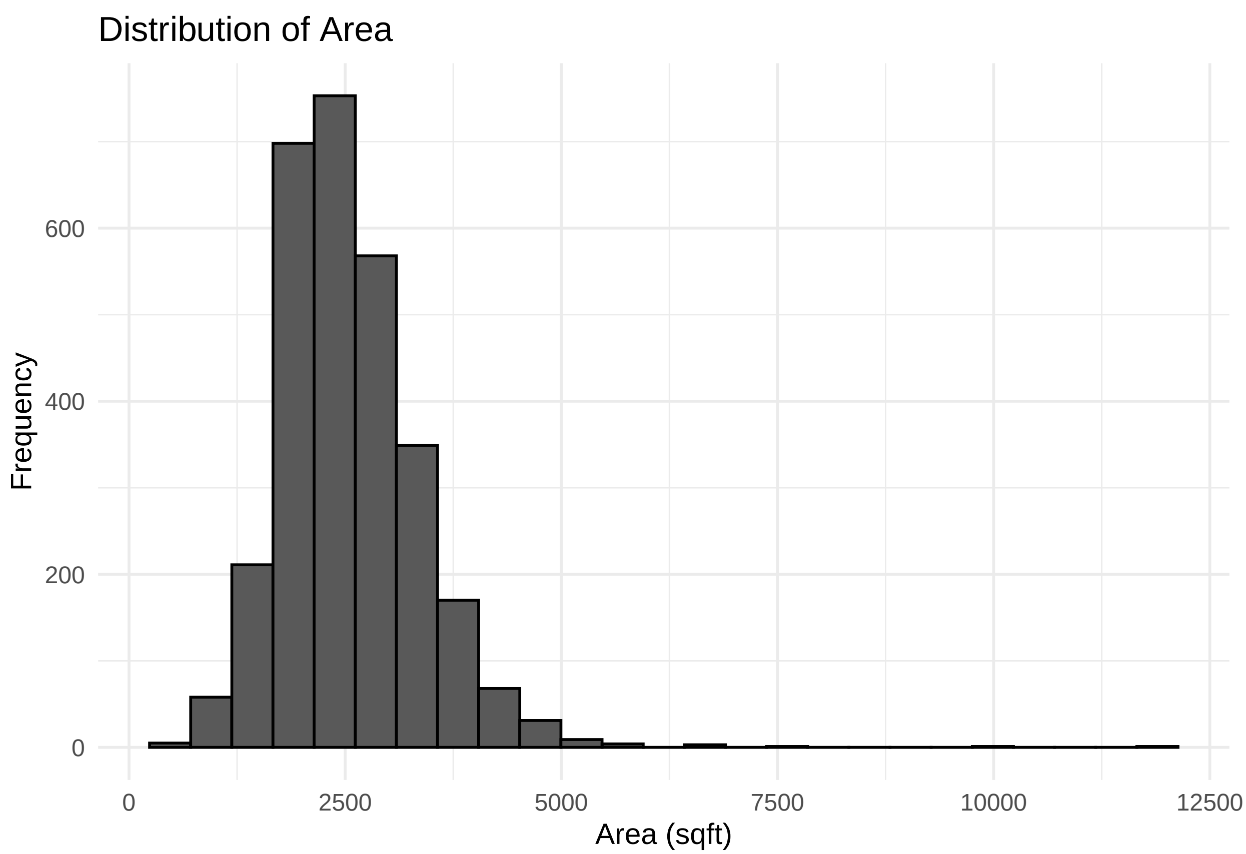
**Created a variable house_Age = year_Sold - year_Build**

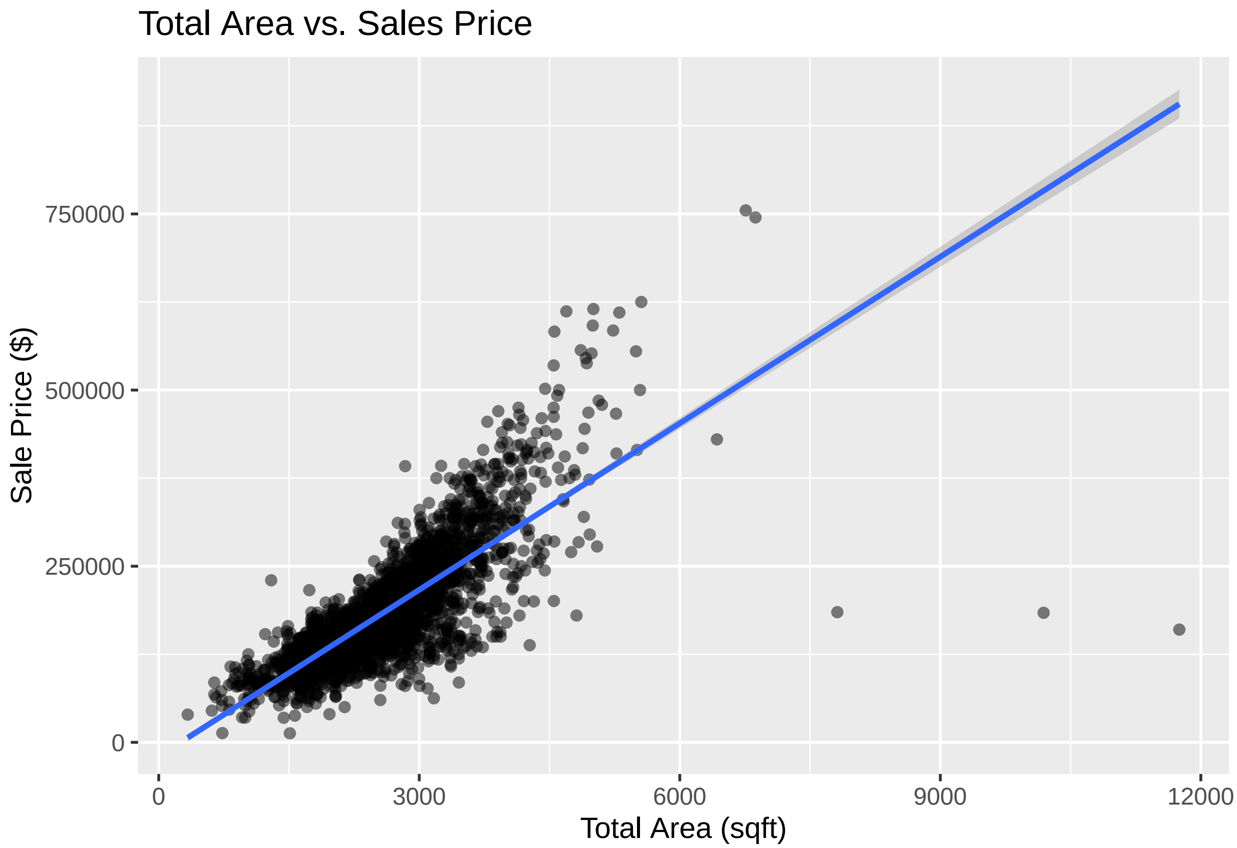
```
##
## Corelation between Total Area and Sale Price : 0.7931272
```

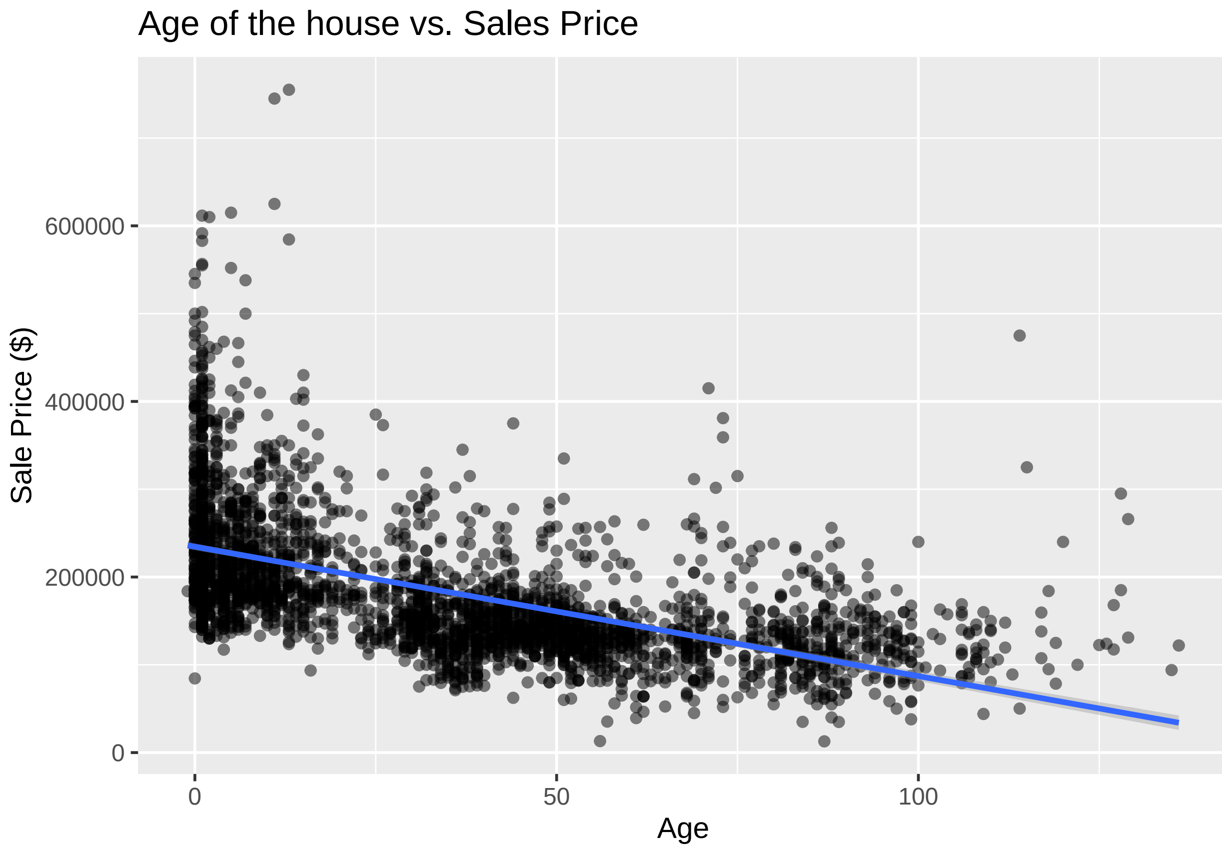
```
##
## Corelation between Total Bathroom and Sale Price : 0.636175
```

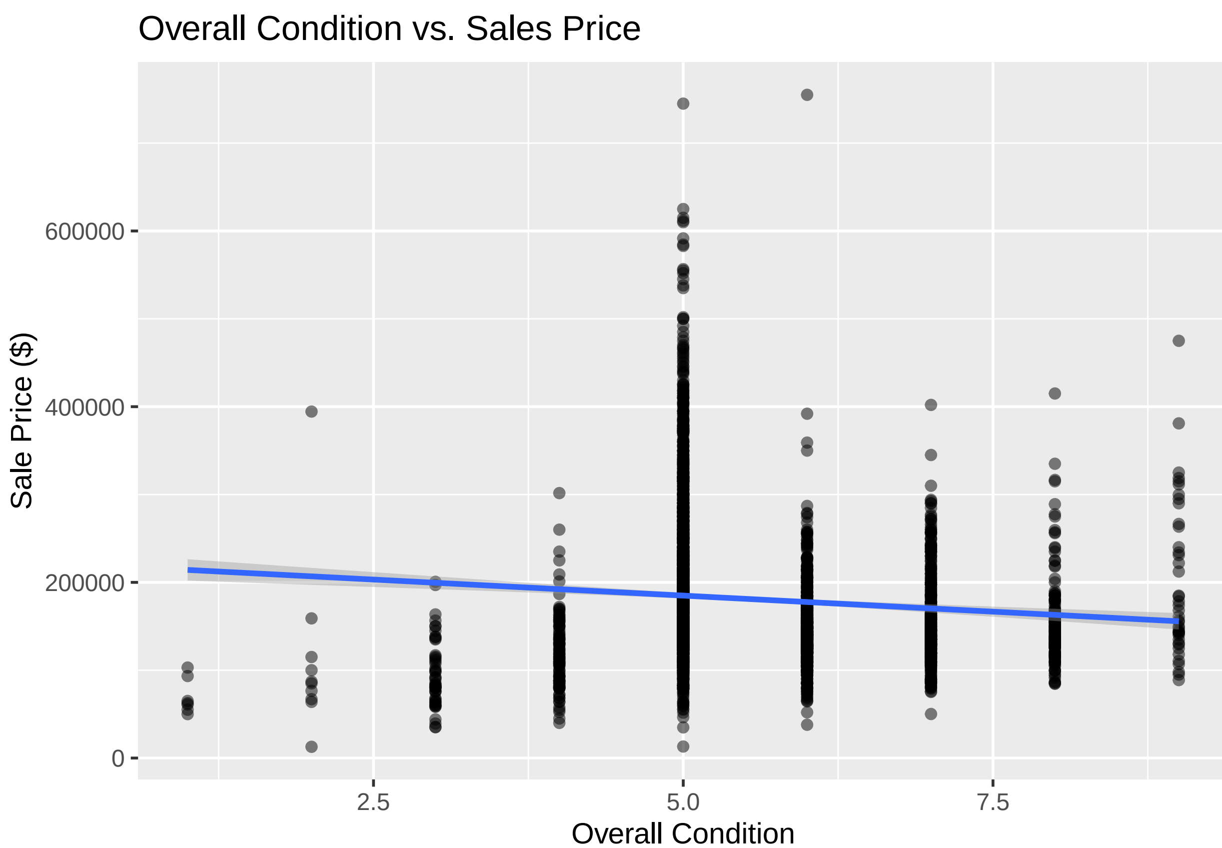
```
##
## Corelation between Age of House and Sale Price : -0.5589068
```

```
##
## Corelation between Overall Condition and Sale Price : -0.1016969
```

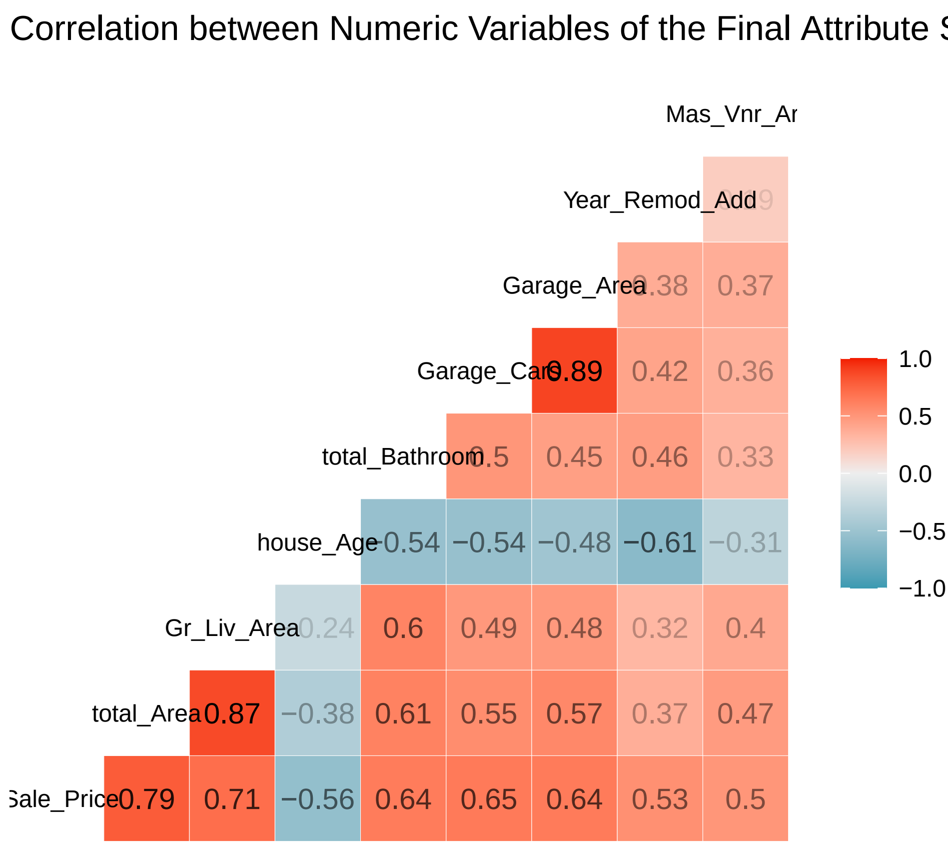



Total Area vs. Sales Price
Sale Price ($)
750000
500000
250000
0
0
3000
6000
9000
12000
Total Area (sqft)


Age of the house vs. Sales Price
Sale Price ($)
600000
400000
200000
0
0
50
100
Age




Looking at the negative correlation between overall condition of the house and sales price I felt that there is something incorrect about the data. I excluded the overall condition from the final parameter set
# Create Final Set with Parameters ## Numeric - Sale_Price,total_Area, Gr_Liv_Area, house_Age, total_Bathroom ,Garage_Cars,Garage_Area, Year_Remod_Add, Mas_Vnr_Area,
## Non-Numeric - House Attributes - Lot_Shape, Foundation, Sale_Condition , Garage_Finish, House_Style, Heating_QC, External Attributes - MS_Zoning, Neighborhood




# Create Test Set and Training set for building Linear Models

## Test set will be 20% of housing_data data

Table 4: Ames Housing Dataset dimension

| x |
|---|
| 2930 |
| 17 |

Table 5: Ames Housing Dataset Summary

| Sale_Price | total_Area | Gr_Liv_Area | house_Age | total_Bathroom | Garage_Cars | Garage_Area | Year_Remod_Add | Mas_Vnr_Area | Lot_Shape | Foundation | Sale_Condition | Garage_Finish | House_Style | Heating_QC | MS_Zoning | Neighborhood |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Min. : 12789 | Min. : 334 | Min. : 334 | Min. : -1.00 | Min. :1.000 | Min. :0.000 | Min. : 0.0 | Min. :1950 | Min. : 0.0 | Regular :1859 | BrkTil: 311 | Abnorml: 190 | Fin : 728 | One_Story :1481 | Excellent:1495 | Floating_Village_Residential: 139 | North_Ames : 443 |
| 1st Qu.:129500 | 1st Qu.: 2000 | 1st Qu.:1126 | 1st Qu.: 7.00 | 1st Qu.:1.500 | 1st Qu.:1.000 | 1st Qu.: 320.0 | 1st Qu.:1965 | 1st Qu.: 0.0 | Slightly_Irregular : 979 | CBlock:1244 | AdjLand: 12 | No_Garage: 159 | Two_Story : 873 | Fair : 92 | Residential_High_Density : 27 | College_Creek : 267 |

| Sale_Price | total_Area | Gr_Liv_Area | House_Age | total_Bathroom | Garage_Cars | Garage_Area | Year_Remod_Add | Mas_Vnr_Area | Lot_Shape | Foundation | Sale_Condition | Garage_Finish | House_Style | Heating_QC | MS_Zoning | Neighborhood |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Median :160000 | Median : 2450 | Median :1442 | Median : 34.00 | Median :2.000 | Median :2.000 | Median : 480.0 | Median :1993 | Median : 0.0 | Moderately_Irregular: 76 | PConc :1310 | Abnorml : 24 | RFn : 812 | One_and_Half_Fin: 314 | Good : 476 | Residential_Low_Density :2273 | Old_Town : 239 |
| Mean :180796 | Mean : 2546 | Mean :1500 | Mean : 36.43 | Mean :2.218 | Mean :1.766 | Mean : 472.7 | Mean :1984 | Mean : 101.1 | Irregular : 16 | Slab : 49 | Family : 46 | Unf :1231 | SLvl : 128 | Poor : 3 | Residential_Medium_Density : 462 | Edwards : 194 |
| 3rd Qu.:213500 | 3rd Qu.: 2990 | 3rd Qu.:1743 | 3rd Qu.: 54.00 | 3rd Qu.:2.500 | 3rd Qu.:2.000 | 3rd Qu.: 576.0 | 3rd Qu.:2004 | 3rd Qu.: 162.8 | NA | Stone : 11 | Normal :2413 | NA | SFoyer : 83 | Typical : 864 | A_agr : 2 | Somerset : 182 |
| Max. :755000 | Max. :11752 | Max. :5642 | Max. :136.00 | Max. :7.000 | Max. :5.000 | Max. :1488.0 | Max. :2010 | Max. :1600.0 | NA | Wood : 5 | Partial : 245 | NA | Two_and_Half_Unf: 24 | NA | C_all : 25 | Northridge_Heights: 166 |
| NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | (Other) : 27 | NA | I_all : 2 | (Other) :1439 |

# House Price Prediction Model - develop, train and test

# Average House Price

## Build Linear Models

I started with linear model and some selected set of parameters/attributes and evaluated the performaces of the models use RMSE. In the first Linear Model we used "Age of the House" and "Total Bathroom" I enhanced the model and added "Age of the House", Garage_Cars + Garage_Area + Year_Remod_Add + Mas_Vnr_Area

```
##
## Naive RMSE in ,000 : 80001.97
```

## 1st Leaner Model - Sale Price ~ total area + total bathroom

Build the model using training set data Predict the Sale Prices of the test set Calculate RMSA

```
##
## Call:
## lm(formula = Sale_Price ~ total_Area + total_Bathroom, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -688177  -20125    -362   19674  258157
##
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    -38520.900   3278.247  -11.75 <0.0000000000000002 ***
## total_Area         66.937      1.495   44.76 <0.0000000000000002 ***
## total_Bathroom  22233.765   1463.347   15.19 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 45410 on 2339 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6766
## F-statistic:  2450 on 2 and 2339 DF,  p-value: < 0.00000000000000022
```

### 1st Linear Model coefficients and RMSE

| method | RMSE |
|---|---|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |

```
##   (Intercept)     total_Area total_Bathroom
##  -38520.89969       66.93719    22233.76524
```

### Second Linear Model using all selected Numeric Attributes

```
## [1] 43410.61
```

```
## # A tibble: 8 x 7
##   term             estimate std.error statistic   p.value   conf.low  conf.high
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>      <dbl>      <dbl>
## 1 (Intercept)    -1191465.   89824.      -13.3  4.89e- 39 -1367589.  -1015341.
## 2 total_Area           48.2      1.32     36.4  8.70e-240       45.6       50.8
## 3 total_Bathroom     8901.    1266.        7.03 2.57e- 12     6418.     11384.
## 4 house_Age          -259.      34.4      -7.53 6.87e- 14     -326.      -192.
## 5 Garage_Cars       10704.    2208.        4.85 1.31e-  6     6376.     15033.
## 6 Garage_Area          29.9      7.65      3.91 9.51e-  5       14.9       44.9
## 7 Year_Remod_Add      605.      45.3      13.4  1.55e- 39      516.       694.
## 8 Mas_Vnr_Area         52.7      4.71     11.2  1.66e- 28       43.5       61.9
```

```
##
## Call:
## lm(formula = Sale_Price ~ total_Area + total_Bathroom + house_Age +
##     Garage_Cars + Garage_Area + Year_Remod_Add + Mas_Vnr_Area,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -575711  -18856   -2996   16275  303778
##
## Coefficients:
##                   Estimate   Std. Error t value           Pr(>|t|)
## (Intercept)    -1191465.415    89823.652 -13.264 < 0.0000000000000002 ***
## total_Area           48.171        1.323  36.420 < 0.0000000000000002 ***
## total_Bathroom     8900.915     1266.254   7.029   0.0000000000025728 ***
## house_Age          -259.007       34.410  -7.527   0.0000000000000687 ***
## Garage_Cars       10704.480     2207.561   4.849   0.0000013057274830 ***
## Garage_Area          29.881        7.646   3.908   0.0000951141288654 ***
## Year_Remod_Add      605.234       45.317  13.356 < 0.0000000000000002 ***
## Mas_Vnr_Area         52.678        4.706  11.193 < 0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39370 on 2922 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:  0.7572
## F-statistic:  1306 on 7 and 2922 DF,  p-value: < 0.00000000000000022
```

With linear model and with a set of attributes I was able to tune the model and reduce RMSE.

## Final Linear Model coefficients and model improvements

| method | RMSE |
|---|---:|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |
| Linear Model based on selected Numeric attributes of the dataset: | 43410.61 |

```
##    (Intercept)     total_Area total_Bathroom      house_Age    Garage_Cars
## -1191465.41520       48.17062     8900.91505     -259.00726    10704.47954
##    Garage_Area Year_Remod_Add   Mas_Vnr_Area
##       29.88056      605.23423       52.67801
```

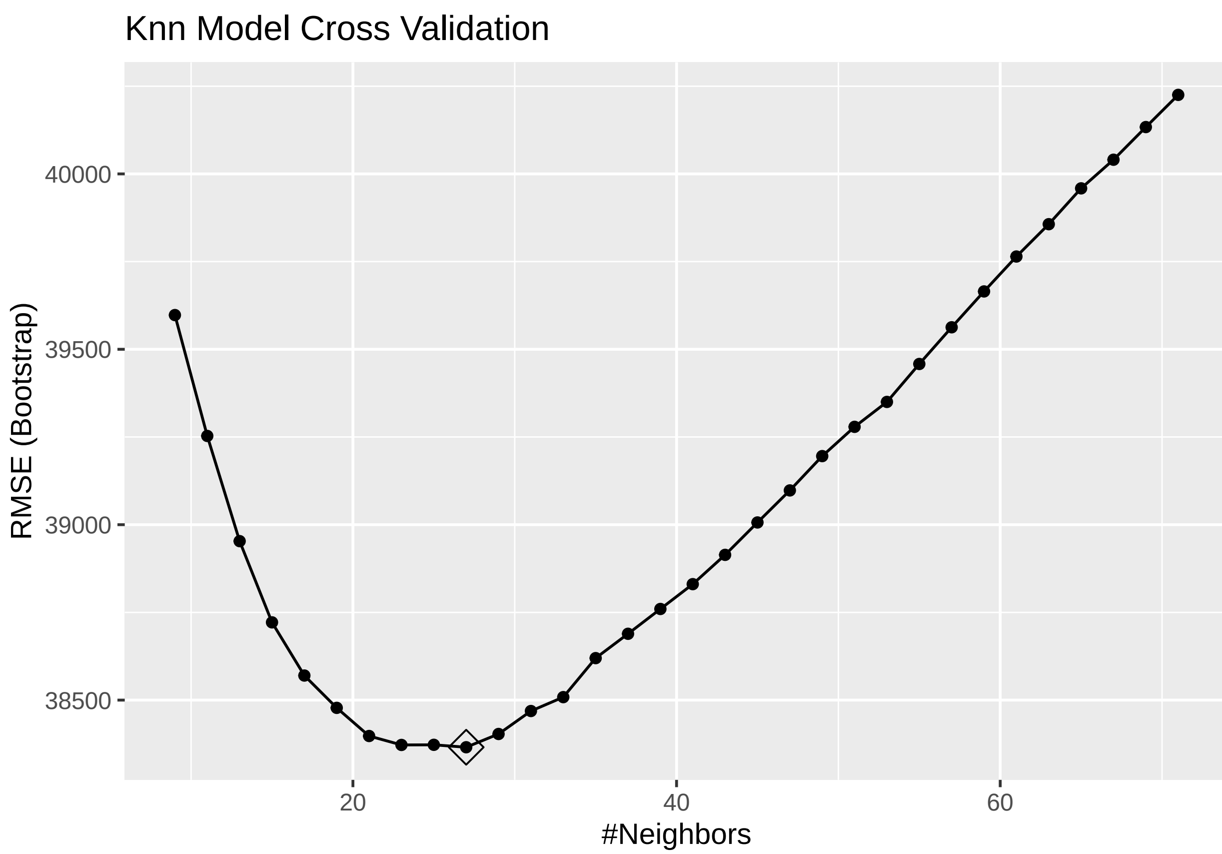
## Non-linear Models

I wanted to further tune the model and enhance the accuracy. I planned to use "kNN", "Classification and regression trees (CART)" and Random Forrest. I added the non-linear parameters with the linear ones. Some of the non-linear ones are attributes of the house and some are external External attributes - Zoning and Neighborhood

### k Nearest Neighbor (kNN) Model

Build th model and find out the predicted Sale Prices Calculate RMSE

```
##             Length Class      Mode
## learn        2     -none-     list
## k            1     -none-     numeric
## theDots      0     -none-     list
## xNames      69     -none-     character
## problemType  1     -none-     character
## tuneValue    1     data.frame list
## obsLevels    1     -none-     logical
## param        0     -none-     list
```

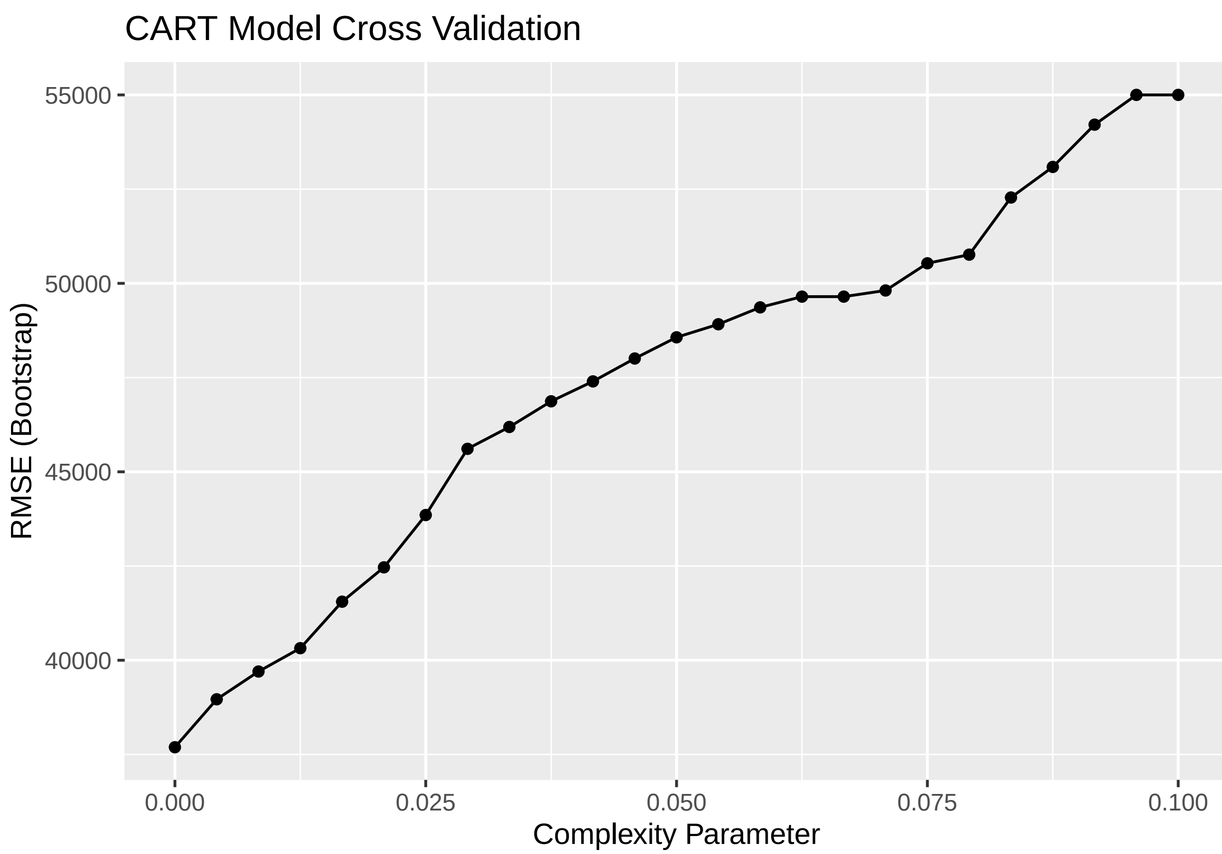



| method | RMSE |
|---|---|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |
| Linear Model based on selected Numeric attributes of the dataset: | 43410.61 |
| Knn Model: | 42323.86 |

Next I am using Classification and regression trees (CART) model to see whether it reduces the RMSE value

## Using Model - Classification and regression trees (CART)

Build th model and find out the predicted Sale Prices Calculate RMSE

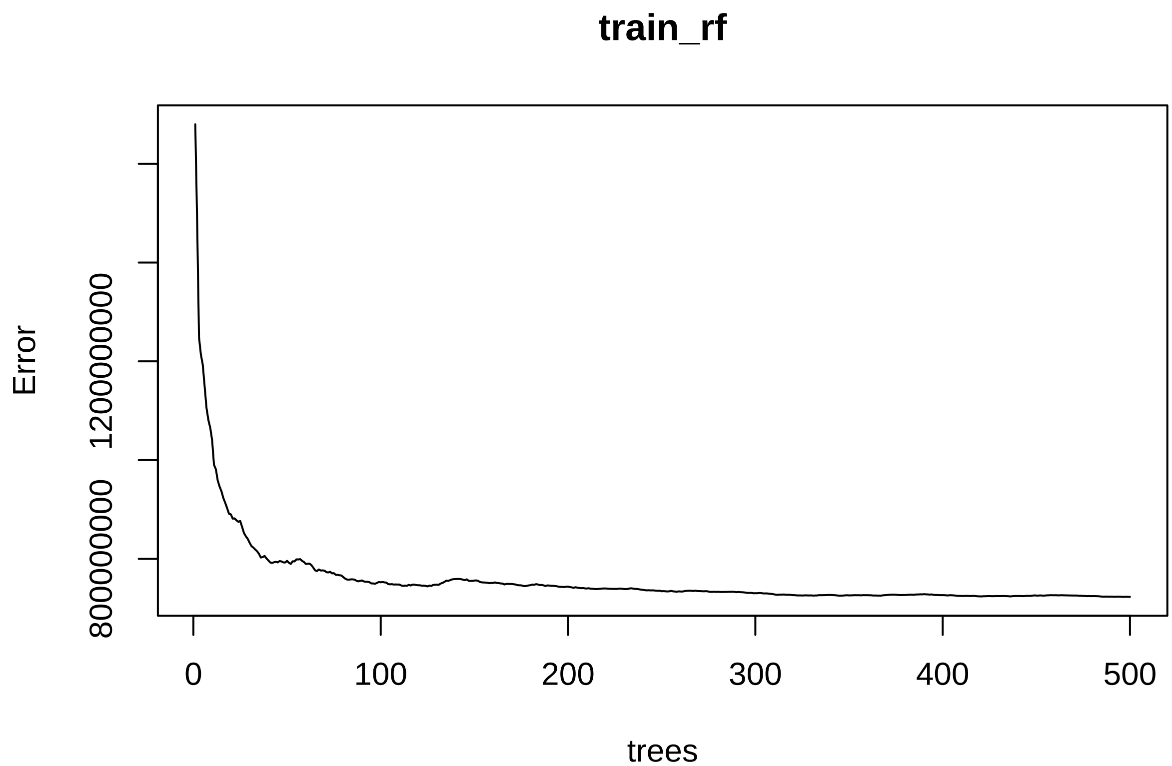



| method | RMSE |
|---|---|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |
| Linear Model based on selected Numeric attributes of the dataset: | 43410.61 |
| Knn Model: | 42323.86 |
| Classification and regression trees (CART) Model: | 38968.58 |

## Random Forrest -

Train the model and find out the predicted Sale Prices Calculate RMSE

```
## [1] "Error vs. Trees"
```




| method | RMSE |
|---|---|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |
| Linear Model based on selected Numeric attributes of the dataset: | 43410.61 |
| Knn Model: | 42323.86 |
| Classification and regression trees (CART) Model: | 38968.58 |
| Random Forrest Model: | 27818.86 |

I got the best result when I used the Classification and regression trees (CART). I wanted to use the Confusion Matrix to calculate the accuracy for in the case of kNN and Classification and regression trees (CART). But because Sale Price is a continuous variable, I could not use Confusion Matrix function directly. When I converted Sale Price ( both predicted and original) into factor, I got extremely low accuracy. After doing further research I found out that this is not a ideal situation to use Confusion Matrix to calculate the accuracy.

# Final Result and Model Performances

RMSEs over Models

| method | RMSE |
|---|---|
| Just the average: | 80001.97 |
| Linear Model based on Total Area and Total Bathroom: | 49157.19 |

| method | RMSE |
|---|---|
| Linear Model based on selected Numeric attributes of the dataset: | 43410.61 |
| Knn Model: | 42323.86 |
| Classification and regression trees (CART) Model: | 38968.58 |
| Random Forrest Model: | 27818.86 |

# Conclusion

To build the House Price Prediction model I started with building Linear model with a set of numeric variables. I identified those variables by observing strong correlation with the "Sale Price" ## Parameters used in the Linear Model Sale_Price,total_Area, Gr_Liv_Area, house_Age, total_Bathroom ,Garage_Cars,Garage_Area, Year_Remod_Add, Mas_Vnr_Area I used RMSE to calculate the efficiency

Next to reduce th error margin , I looked at three other Models kNN, Classification and regression trees (CART) and Random Forrest. I identified some non-numeric attributes looking at their correlation with the Sale Price ## Non-Numeric - House Attributes - Lot_Shape, Foundation, Sale_Condition , Garage_Finish, House_Style, Heating_QC, External Attributes - MS_Zoning, Neighborhood

Finally with Random Forrest Model I got the lowest RMSE.

I am sure doing some additional Feature Engineering and combining more than one models I will be able to build a better House Prediction Model.

# Reference -

Introduction to Data Science by Rafael A. Irizarry

https://jse.amstat.org/v19n3/decock.pdf - Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project - Dean De Cock

https://modeldata.tidymodels.org/reference/ames.html - Ames Housing Data

https://www.investopedia.com