PSTAT126 Project

Liuqian Bao

2023-10-10

Data Descrption

Source

Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle.Obtained from https://kaggle.com/competitions/house-prices-advanced-regression-techniques (https://kaggle.com/competitions/house-prices-advanced-regression-techniques)

Observational unit:

Each property sampled in the data set.

Population:

All residential homes in Ames, Iowa.

Variable descriptions:

Id: The id number of the observation, from 1 to 1460.

Response variable:

SalePrice: The property's sale price in dollars. The dependent variable.

Predictors:

1. LotArea: Lot size in square feet
2. LotShape: General shape of property
3. LotConfig: Lot configuration
4. Neighborhood: Physical locations within Ames city limits
5. HouseStyle: Style of dwelling
6. RoofStyle: Type of roof
7. Exterior1st: Exterior covering on house
8. MasVnrType: Masonry veneer type
9. ExterQual:Evaluates the quality of the material on the exterior
10. ExterCond: Evaluates the present condition of the material on the exterior
11. Foundation: Type of foundation
12. BsmtQual: Evaluates the height of the basement
13. BsmtCond: Evaluates the general condition of the basement
14. BsmtFinType1: Rating of basement finished area
15. BsmtFinSF1: basement finished area square feet
16. BsmtFinType2: Rating of basement finished area (if multiple types)
17. BsmtUnfSF: Unfinished square feet of basement area
18. HeatingQC: Heating quality and condition
19. Electrical system

20. BsmtFullBath: Basement full bathrooms
21. FullBath: Full bathrooms above grade
22. HalfBath: Half baths above grade
23. KitchenQualKitchenQual: Kitchen quality
24. FireplaceQu: Fireplace quality
25. GarageType: Garage location
26. GarageFinish: Interior finish of the garage
27. GarageCars: Size of garage in car capacity
28. OpenPorchSF: Open porch area in square feet

Data Summary

Summary Statistics

We selected, in total, 28 predictors, 8 of which are numerical variables and 20 of which are categorical variables. Here is a summary table for all variables, including the response variable SalePrice and Id of the observational units:

Data summary

Name	train2
Number of rows	1460
Number of columns	30
_____	
Column type frequency:	
character	20
numeric	10
_____	
Group variables	None

Variable type: character

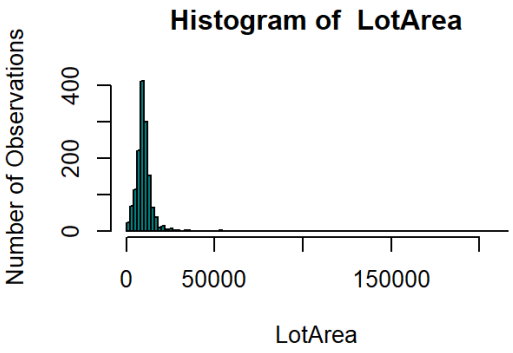
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
LotShape	0	1	3	3	0	4	0
LotConfig	0	1	3	7	0	5	0
Neighborhood	0	1	5	7	0	25	0
HouseStyle	0	1	4	6	0	8	0
RoofStyle	0	1	3	7	0	6	0
Exterior1st	0	1	5	7	0	15	0
MasVnrType	0	1	2	7	0	5	0
ExterQual	0	1	2	2	0	4	0
ExterCond	0	1	2	2	0	5	0
Foundation	0	1	4	6	0	6	0

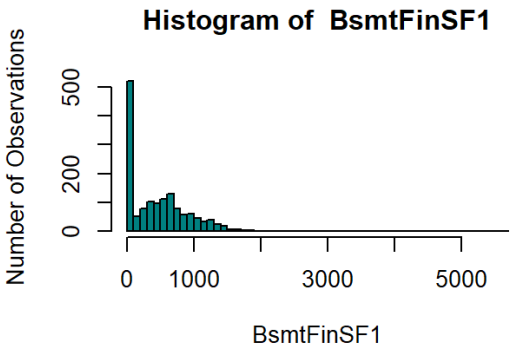
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
BsmtQual	0	1	2	2	0	5	0
BsmtCond	0	1	2	2	0	5	0
BsmtFinType1	0	1	2	3	0	7	0
BsmtFinType2	0	1	2	3	0	7	0
HeatingQC	0	1	2	2	0	5	0
Electrical	0	1	2	5	0	6	0
KitchenQual	0	1	2	2	0	4	0
FireplaceQu	0	1	2	2	0	6	0
GarageType	0	1	2	7	0	7	0
GarageFinish	0	1	2	3	0	4	0

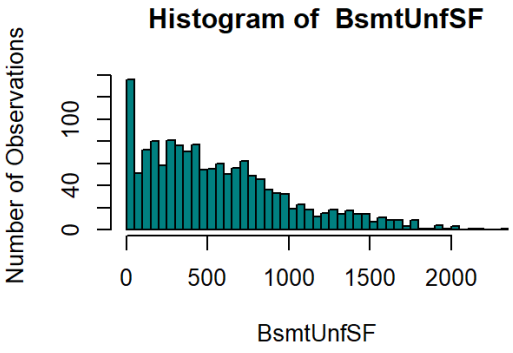
Variable type: numeric

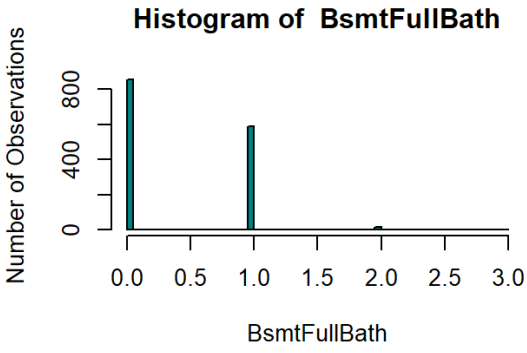
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Id	0	1	730.50	421.61	1	365.75	730.5	1095.25	1460	
LotArea	0	1	10516.83	9981.26	1300	7553.50	9478.5	11601.50	215245	
BsmtFinSF1	0	1	443.64	456.10	0	0.00	383.5	712.25	5644	
BsmtUnfSF	0	1	567.24	441.87	0	223.00	477.5	808.00	2336	
BsmtFullBath	0	1	0.43	0.52	0	0.00	0.0	1.00	3	
FullBath	0	1	1.57	0.55	0	1.00	2.0	2.00	3	
HalfBath	0	1	0.38	0.50	0	0.00	0.0	1.00	2	
GarageCars	0	1	1.77	0.75	0	1.00	2.0	2.00	4	
OpenPorchSF	0	1	46.66	66.26	0	0.00	25.0	68.00	547	
SalePrice	0	1	180921.20	79442.50	34900	129975.00	163000.0	214000.00	755000	

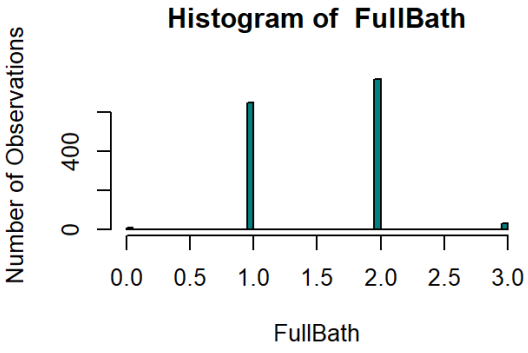
Distribution of individual variables

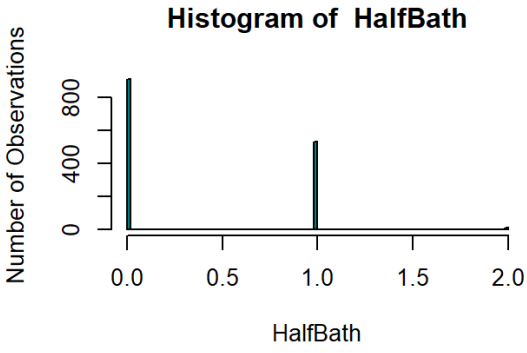

Histogram of LotArea
Number of Observations
0
200
400
0
50000
150000
LotArea

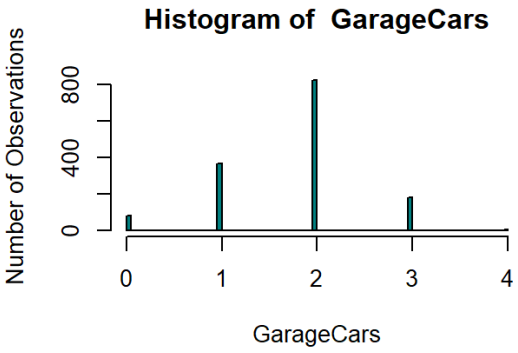

Histogram of BsmtFinSF1
Number of Observations
0
200
500
0
1000
3000
5000
BsmtFinSF1

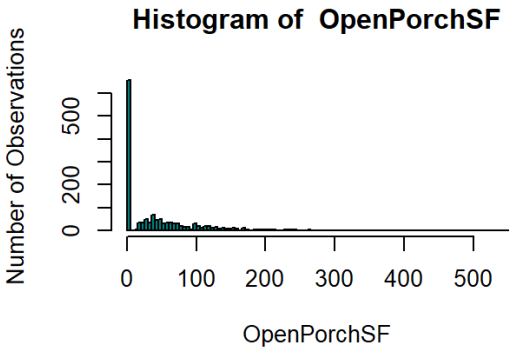

Histogram of BsmtUnfSF
Number of Observations
0
40
100
0
500
1000
1500
2000
BsmtUnfSF

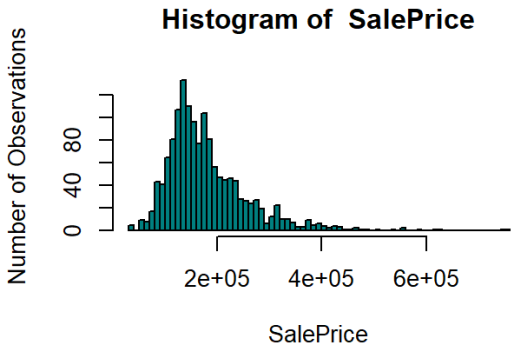

Histogram of BsmtFullBath
Number of Observations
0
400
800
0.0
0.5
1.0
1.5
2.0
2.5
3.0
BsmtFullBath


Histogram of FullBath
Number of Observations
0
400
0.0
0.5
1.0
1.5
2.0
2.5
3.0
FullBath


Histogram of HalfBath
Number of Observations
0
400
800
0.0
0.5
1.0
1.5
2.0
HalfBath


Histogram of GarageCars
Number of Observations
0
400
800
0
1
2
3
4
GarageCars


Histogram of OpenPorchSF
Number of Observations
0
200
500
0
100
200
300
400
500
OpenPorchSF



1. The LotArea has a mean of 10,516.83, with the highest bar (frequency over 400) occurring in the 1000-1300 range. While the shape approximates a bell curve, it does not form a true bell-shaped distribution due to outliers exceeding 500,000 in lot area.
2. The histogram for BsmtFinSF1 has a mean that is not easy to tell from first glance. The actual mean is 443.64. The shape of the distribution is right-skewed with high frequency around zero.
3. The histogram for BsmtFinSF has a mean that is not easy to tell from first glance compared to the actual mean of 567.24. The shape is of a strange type where when BsmtFinSF is 0 the frequency is the highest reaching over 120 but afterwards it has a distribution that is even as BsmtFinSF increases, slowly decreasing in frequency along the way.
4. The histogram for BsmtFullBath has a mean that is not easy to tell from first glance compared to the actual mean of 0.43. The shape is of a strange type where when BsmtFullBathF is 0.0 the frequency is the highest reaching well over 800 but afterwards it has no frequency at all until BsmtFullBathF reaches 1.0 where frequency hits around 600.
5. The histogram for FullBath has a mean that is not easy to tell from first glance compared to the actual mean of 1.57. The shape is of a strange type where compared with BsmtFullBathF at 0.0 the frequency is very low sitting around 0 and afterwards it has no frequency at all. However, at 1.0 and 2.0, the frequency increases so dramatically at that 2.0, the frequency hits over 600. Afterwards, again, there is no frequency at all.
6. The histogram for HalfBath has a mean that is not easy to tell from first glance compared to the actual mean of 0.38. The shape is similar to that of the histogram for FullBath, however, the most striking difference is that when HalfBath is 0.0, it has the highest frequency as it tops well over 800. When HalfBath is at 1.0, the frequency is around 600, with the frequency being 0 the rest of the times.
7. The histogram for GarageCars has a mean that is not easy to tell from first glance compared its actual mean of 1.77. The shape is similar to that of the histogram for FullBath, but it is more sophisticated. There is data for every number of garage car, starting from 0 all the way to 4. At 2, the frequency is by far the highest, topping over 800, with 1 and 3 having the next highest frequencies in the plot.
8. The histogram for OpenPorchSF has a mean that is not easy to tell from first glance compared its actual mean of 46.66. The shape is similar to that of the histogram for BsmtFinSF1, but OpenPorchSF has an even greater difference between the frequency of when OpenPorchSF is 0 and when OpenPorchSF is at different values. At 0, the frequency sits well above 500, while none of the rest of the freqencies hit 100.

9. The histogram for SalePrice has a mean that is around 50, which is pretty close to the true mean that is around 46.66. The shape of the histogram is close to and has parts that resembles that of a bell curve distribution. The highest frequency occurs when the SalePrice is near 2e+05, with a frequency topping 120. However, we cannot say that it is a bell curve distribution since there are outliers.

Correlation between the numerical variables and the response

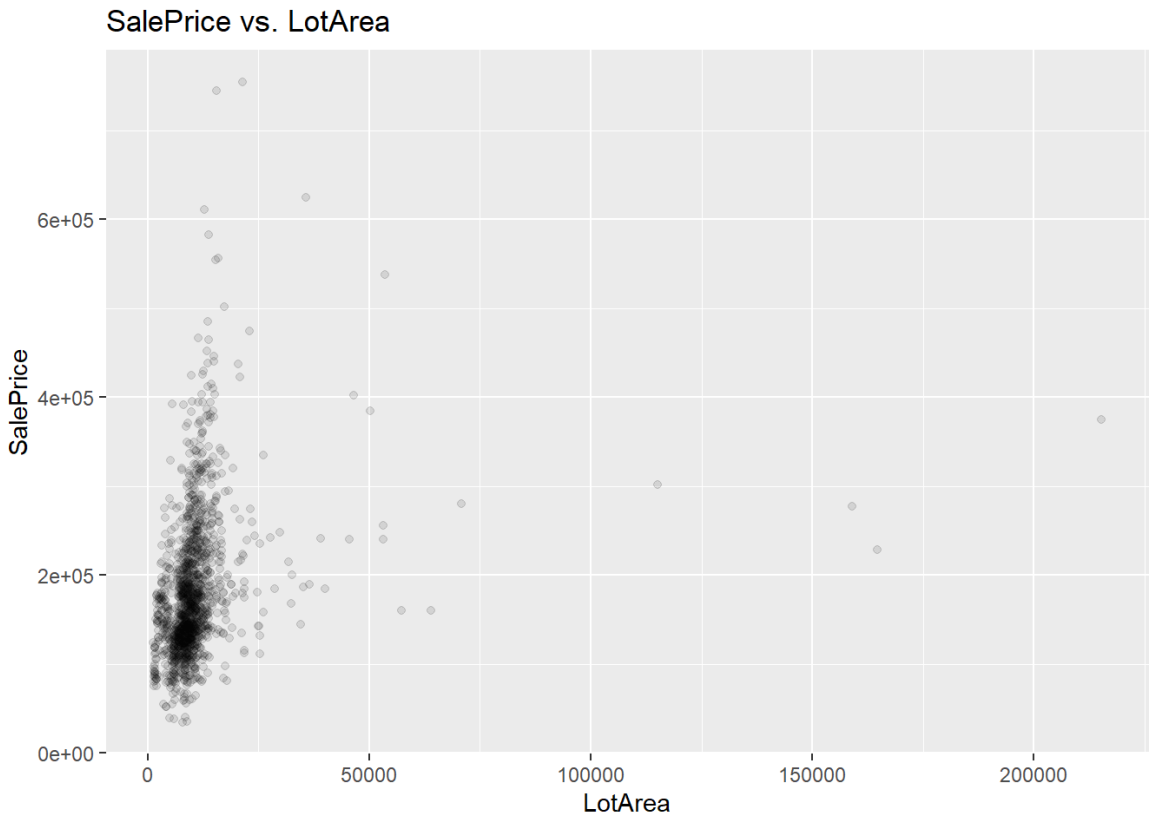
correlation matrix

The correlation matrix of all numerical variables:

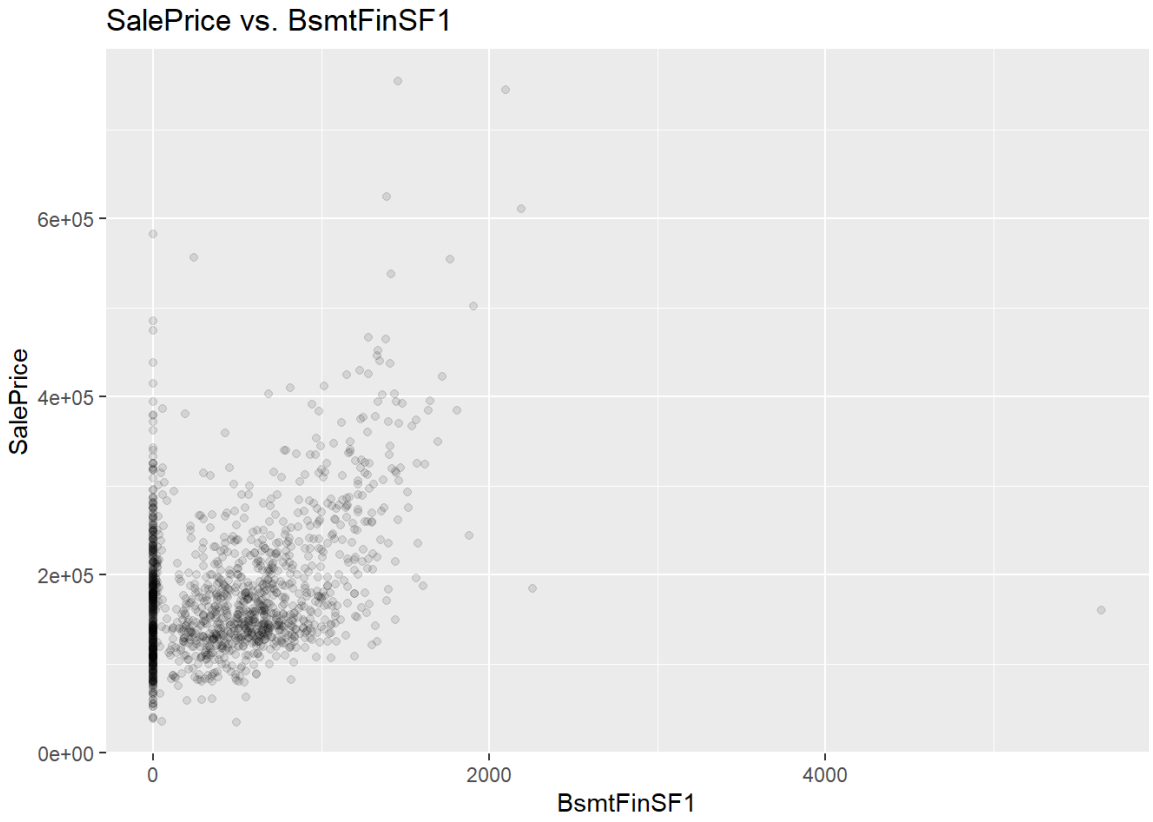
```
##                         Id     LotArea   BsmtFinSF1    BsmtUnfSF BsmtFullBath
## Id            1.0000000000 -0.03322552 -0.005024049 -0.007939703  0.002288556
## LotArea      -0.0332255186  1.00000000  0.214103131 -0.002618360  0.158154531
## BsmtFinSF1   -0.0050240490  0.21410313  1.000000000 -0.495251469  0.649211754
## BsmtUnfSF    -0.0079397034 -0.00261836 -0.495251469  1.000000000 -0.422900477
## BsmtFullBath  0.0022885556  0.15815453  0.649211754 -0.422900477  1.000000000
## FullBath      0.0055874529  0.12603063  0.058543137  0.288886055 -0.064512049
## HalfBath      0.0067838113  0.01425947  0.004262424 -0.041117530 -0.030904959
## GarageCars    0.0165696771  0.15487074  0.224053522  0.214175190  0.131881224
## OpenPorchSF  -0.0004769113  0.08477381  0.111760613  0.129005415  0.067341461
## SalePrice    -0.0219167194  0.26384335  0.386419806  0.214479106  0.227122233
##                  FullBath     HalfBath GarageCars   OpenPorchSF   SalePrice
## Id            0.005587453  0.006783811 0.01656968 -0.0004769113 -0.02191672
## LotArea       0.126030627  0.014259469 0.15487074  0.0847738088  0.26384335
## BsmtFinSF1    0.058543137  0.004262424 0.22405352  0.1117606134  0.38641981
## BsmtUnfSF     0.288886055 -0.041117530 0.21417519  0.1290054146  0.21447911
## BsmtFullBath -0.064512049 -0.030904959 0.13188122  0.0673414614  0.22712223
## FullBath      1.000000000  0.136380589 0.46967204  0.2599774255  0.56066376
## HalfBath      0.136380589  1.000000000 0.21917815  0.1997401475  0.28410768
## GarageCars    0.469672043  0.219178152 1.00000000  0.2135694456  0.64040920
## OpenPorchSF   0.259977425  0.199740148 0.21356945  1.0000000000  0.31585623
## SalePrice     0.560663763  0.284107676 0.64040920  0.3158562271  1.00000000
```

Scatter plots of SalePrice against the numerical variables:

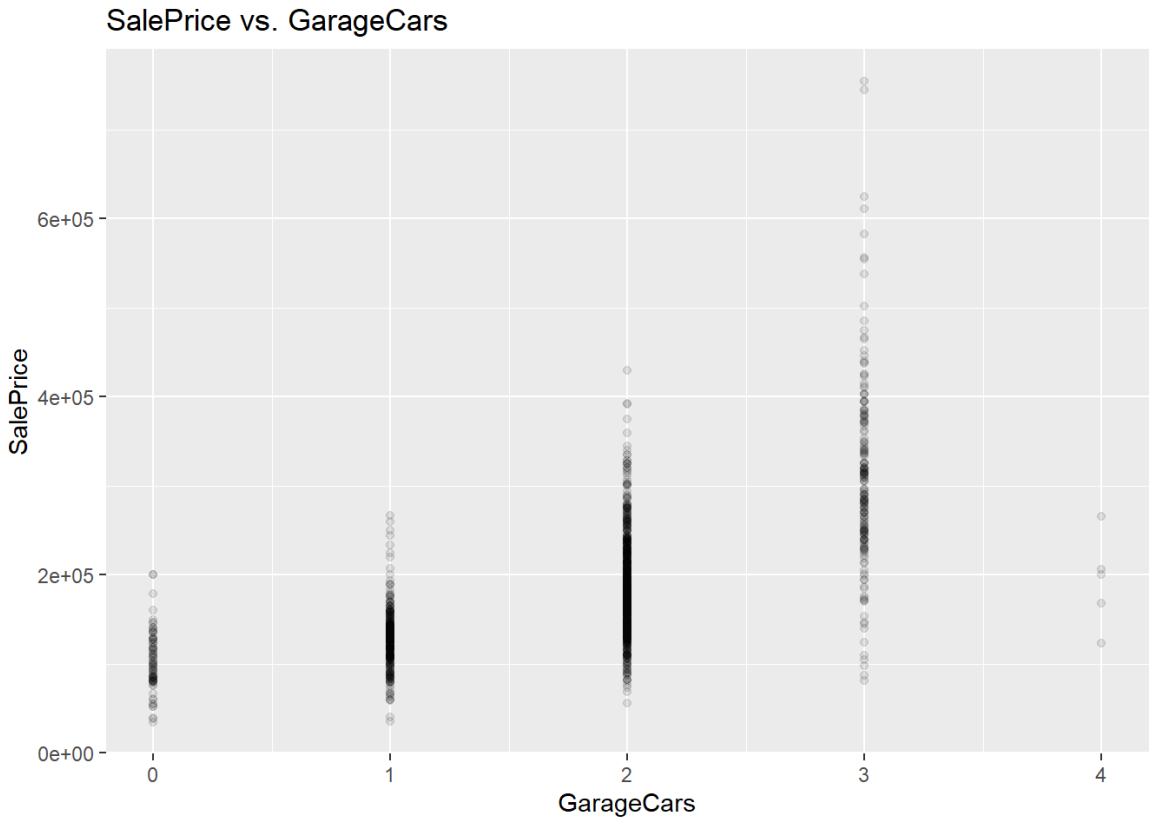
1. The scatterplot for Lot Area versus SalePrice reveals a connection between the size of the lot and the selling price. Generally, when the lot size increases, the sale price also tends to go up. However, this relationship isn't very strong because most data points are clustered near the lower end of the lot sizes, as shown below:

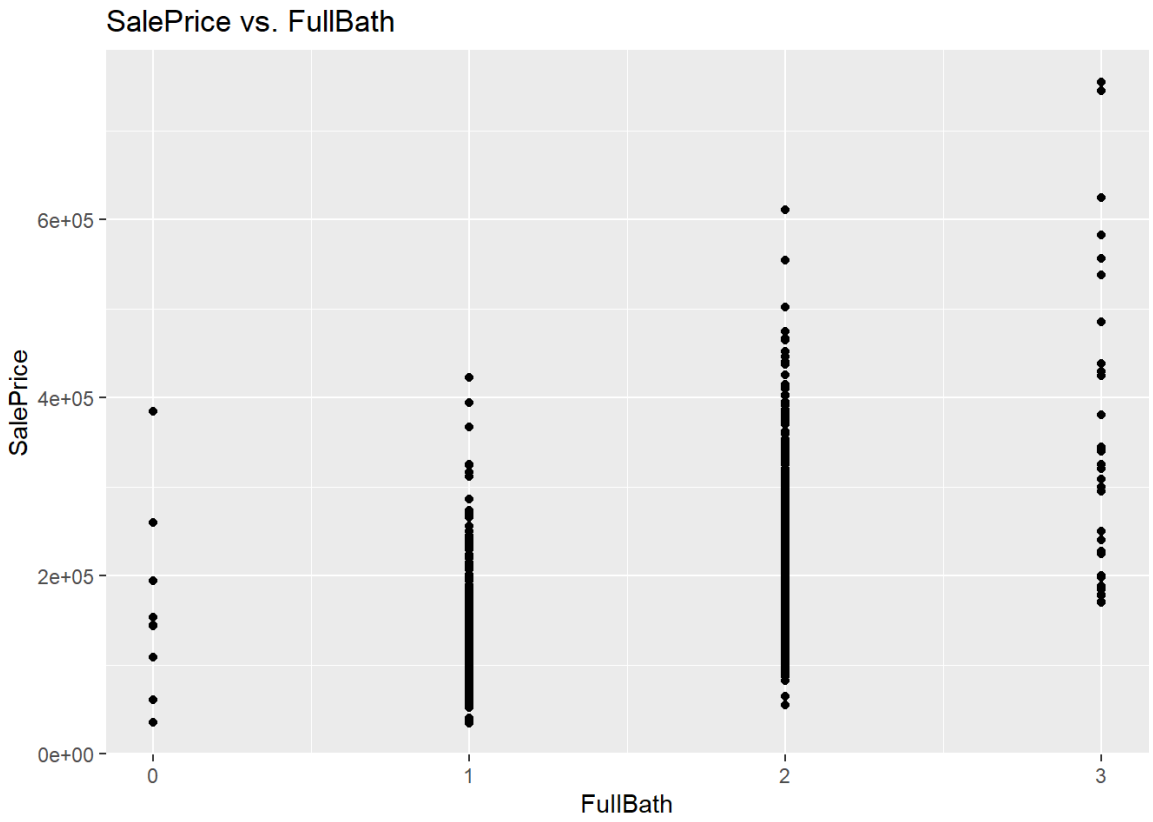


2. The scatterplot of SalePrice versus BsmtFinSF1 indicates a strong association between BsmtFinSF1 and SalePrice. When BsmtFinSF1 increases, there is a clear corresponding increase in SalePrice, as indicated by the tightly packed data points. Similarly to the previous plot, there is also a clustering of data points at the 0 value for BsmtFinSF1, suggesting many properties without finished basement space.



3. The GarageCars and FullBath variables show similar and strong associations with the SalePrice. Generally, the greater the number of garage cars and full baths, the higher the sale price of the property, as shown below:

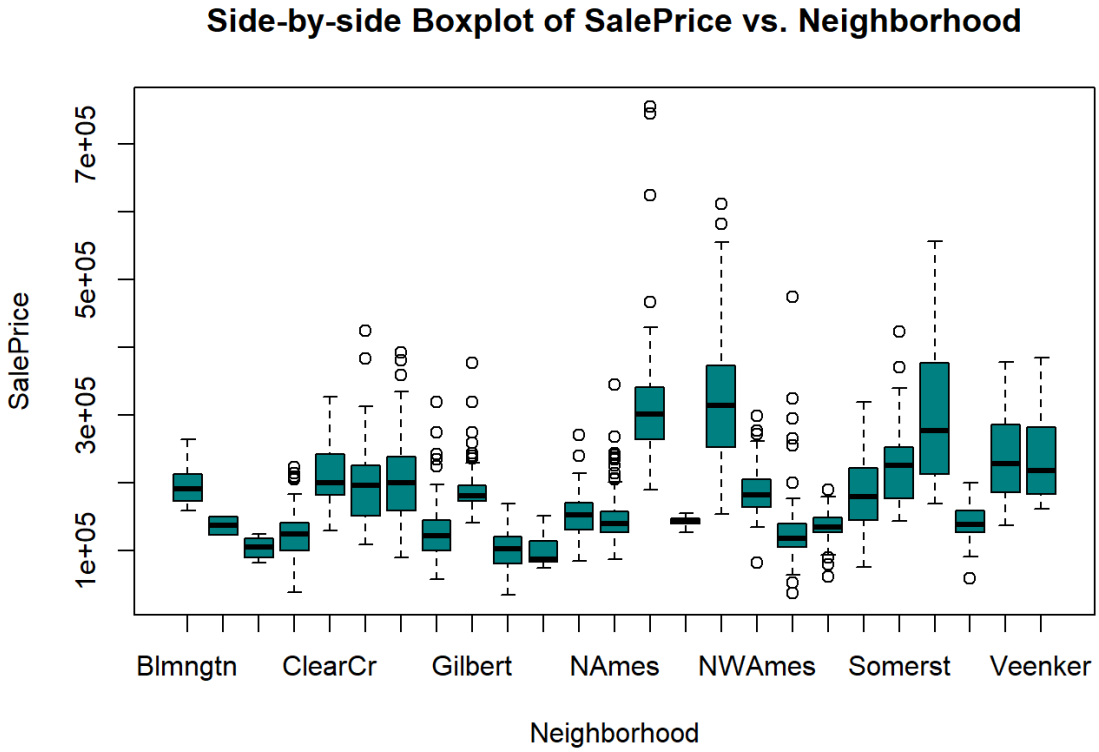




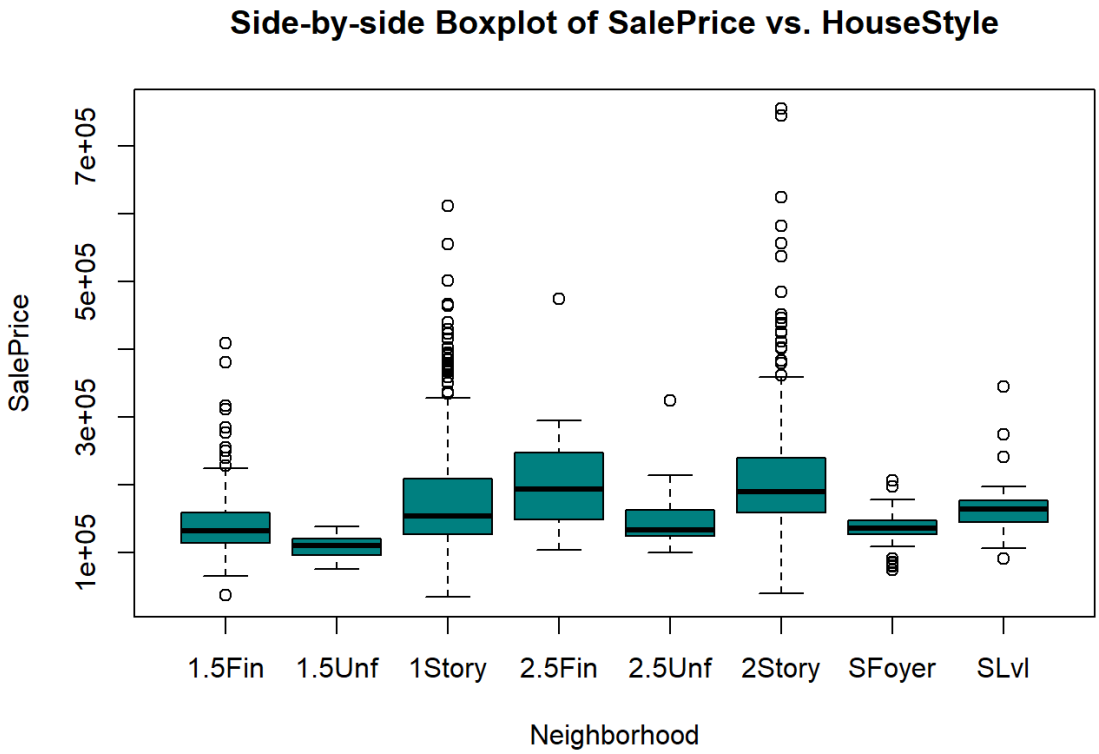
Correlation between the categorical variables and the response

Side by side boxplot for SalePrice against the categorical variables:

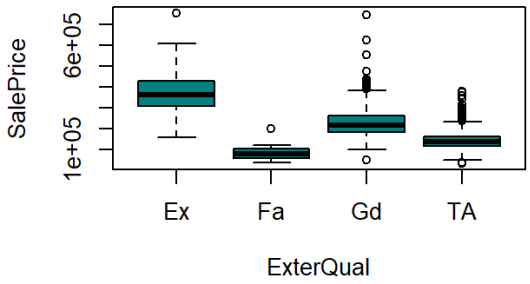
1. The Neighborhood variable has a strong correlation with the SalePrice, because we can see from the side-by-side box plot that the mean of the distribution of SalePrice for different neighborhoods are very different. Also, different neighborhoods have very different variance in SalPrice. As shown below:

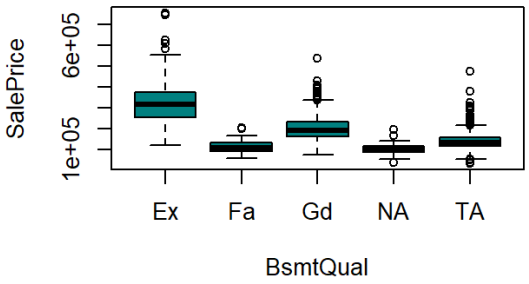


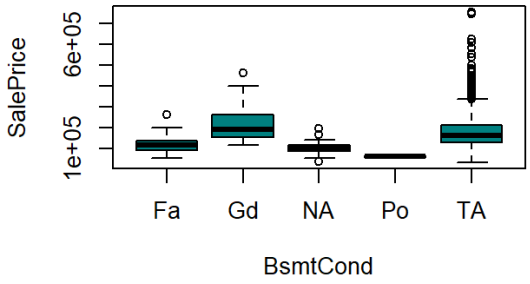
2. The HouseStyle variable has a moderate correlation with the SalePrice, and different house styles also show great differences in variance of the SalePrice distribution. The 1 story, 2.5Fin, and 2 story styles, in particular, have much greater variance than other styles. As shown below:

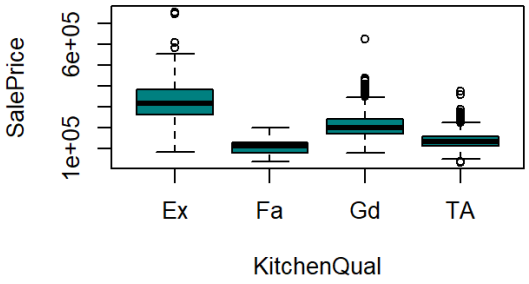


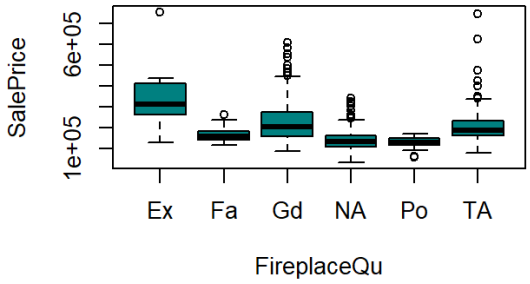
3. The ExterQual, BsmtQual, BsmtCond, KitchenQual, and FireplaceQu, which are the condition and qualities of the exterior, basement, kitchen and fireplace of the properties, all have similar relationships with the SalePrice. In general, the better the condition/quiality, the higher the SalePrice. As shown below:



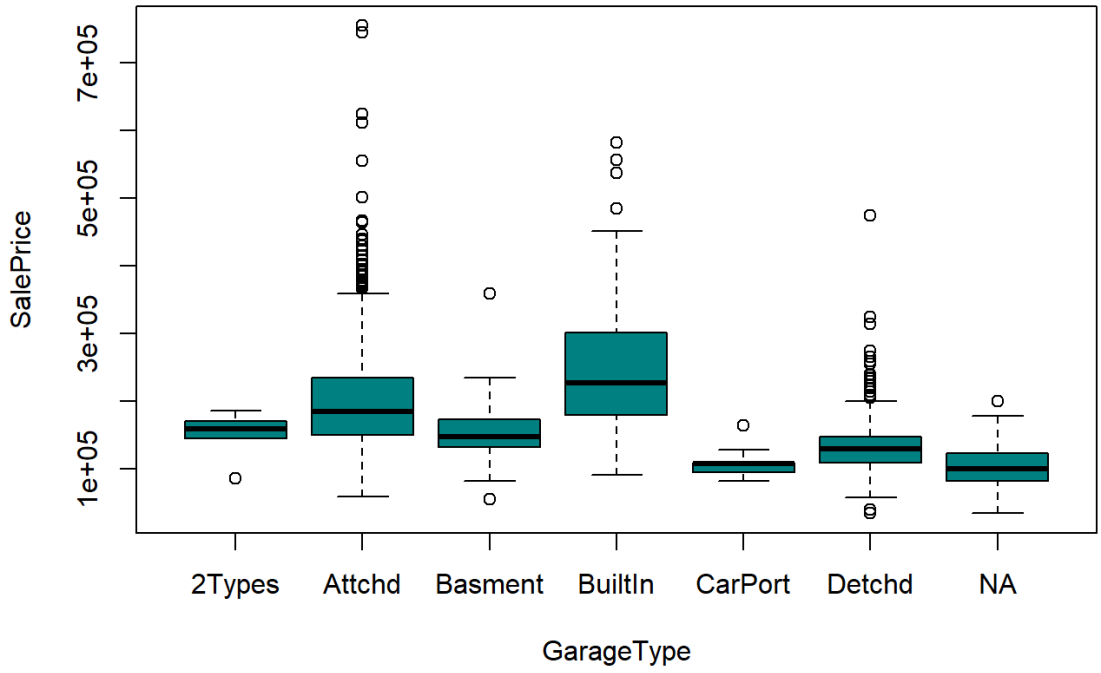


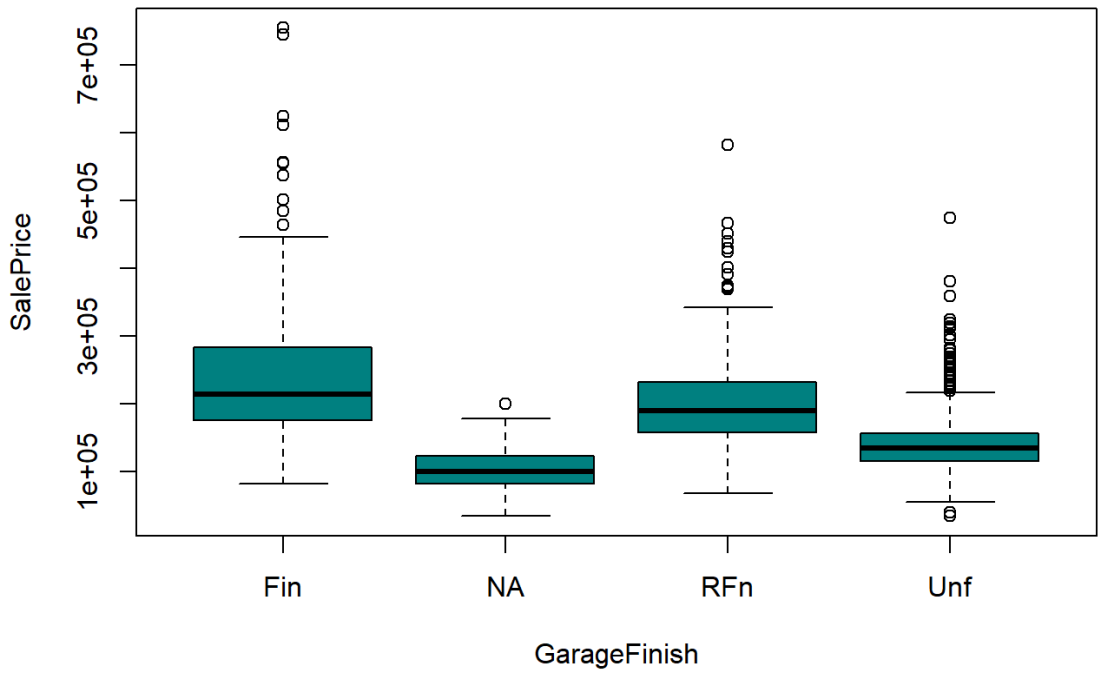






4. GarageType, GarageFinish variables also show some association with the SalePrice. In particular, built-in and attached garage types have higher sale prices then other categories. Properties with finished and rough finished garages have higher price, on average, than those with unfinished or no garage, as shown below:





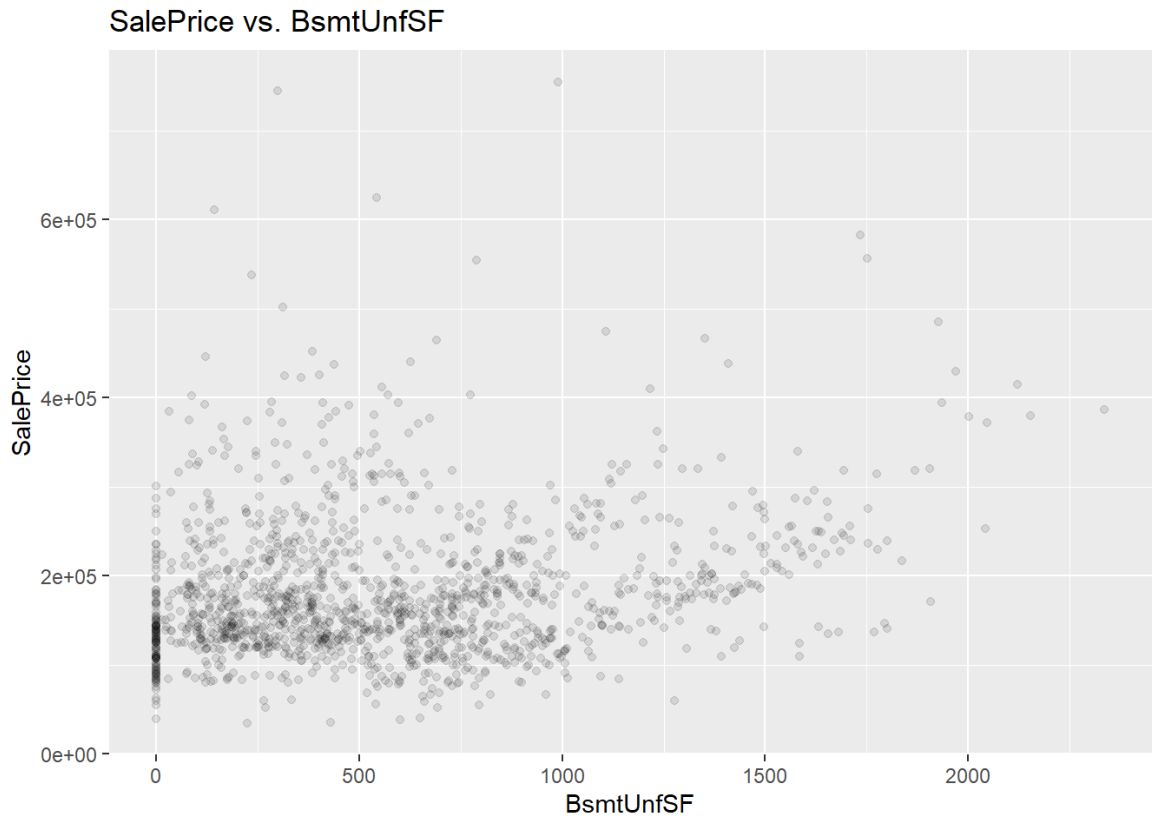
Interesting features

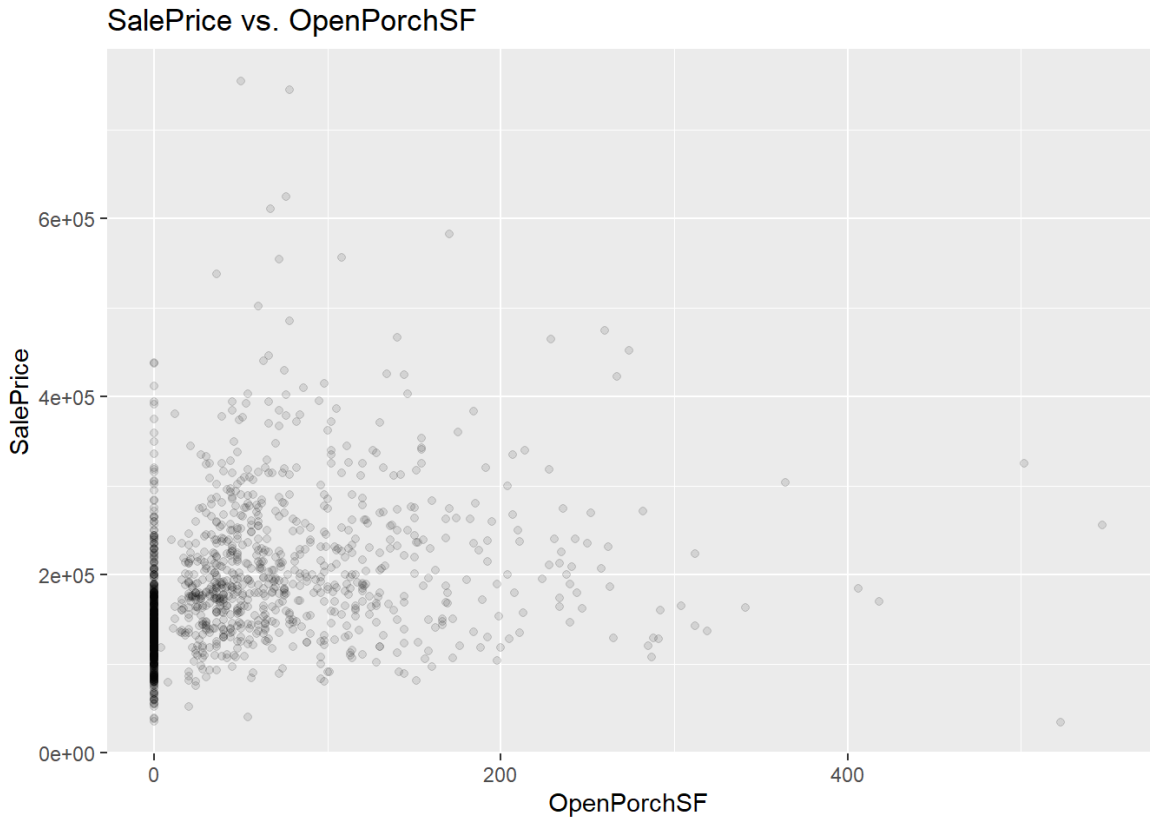
1. The ExterCond variable, which is the the present condition of the material on the exterior, surprisingly, has very little correlation with the SalePrice based on the side-by-side box plot.
2. There is a very weak correlation between unfinished basement and sale proce. In general, it doesn't seem like having a larger unfinished basement (BsmtUnfSF) results in a notably higher sale price. Interestingly, many data points bunch up around the 0 value for BsmtUnfSF, suggesting a common occurrence of properties without much unfinished space.

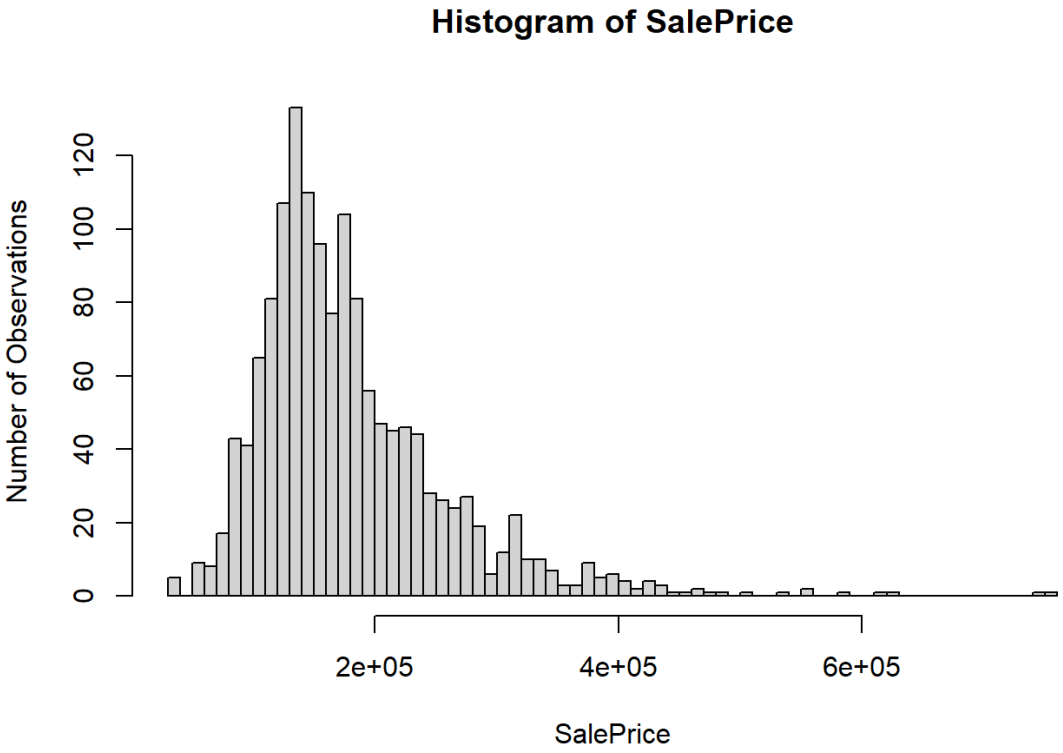
3. Also, there is a very weak relationship between OpenPorchSF and SalePrice, which is surprising because we would imagine that properties with greater open porch area will have higher sale price.
4. The distribution of SalePrice for properties in Ames, Iowa shows a very right-skewed shape, which implies that a significant number of houses have lower sale prices, and there are relatively fewer houses with very high sale prices. This may reflect income disparities or variations in property values within the area, with a concentration of less expensive homes and only a few high-end properties.

Graphics of the interesting features:


Side-by-side Boxplot of SalePrice vs. ExterCond
SalePrice
1e+05
3e+05
5e+05
7e+05
Ex
Fa
Gd
Po
TA
ExterCond


SalePrice vs. BsmtUnfSF
SalePrice
0e+00
2e+05
4e+05
6e+05
0
500
1000
1500
2000
BsmtUnfSF


SalePrice vs. OpenPorchSF
SalePrice
6e+05
4e+05
2e+05
0e+00
0
200
400
OpenPorchSF


Histogram of SalePrice
Number of Observations
0
20
40
60
80
100
120
2e+05
4e+05
6e+05
SalePrice