Housing prices in Ames,Iowa

Luis Fernando Romo

January 2021

Introduction

Heterogeneous is a key word to describe the housing market nowadays. Houses can have a vast variety of characteristics to set them apart and give them that personal touch. This wide variety of elements becomes a problem when you are trying to assess home values. How can you accurately determine the fair price of a property when it has such a unique set of characteristics.

Data acquisition

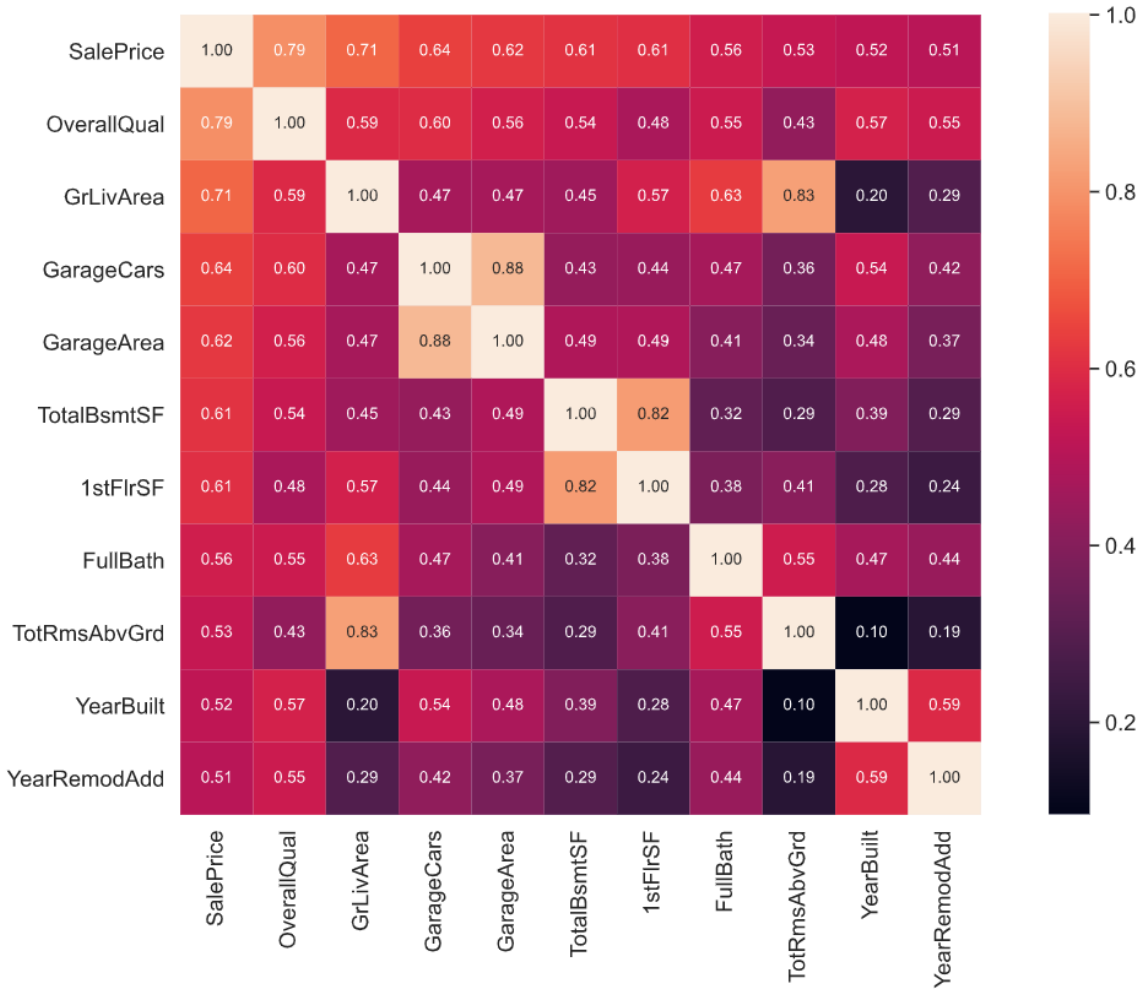
To tackle this problem and find a model that can accurately predict the selling price of homes I will analyze and compare explanatory variables describing different aspects and characteristics of residential property in Ames, Iowa from 2006 to 2010. There are 2930 observations in the dataset with 79 variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) such as land slope, lot shape, the general zoning classification of the sale, linear feet of street connected to property and many more. This dataset was compiled by Dean De Cock (http://jse.amstat.org/v19n3/decock.pdf).

Methodology

A good starting point for the analysis is to get the descriptive statistics of the sample and plotting a heatmap to visualize the variables with the highest correlation to sale price.

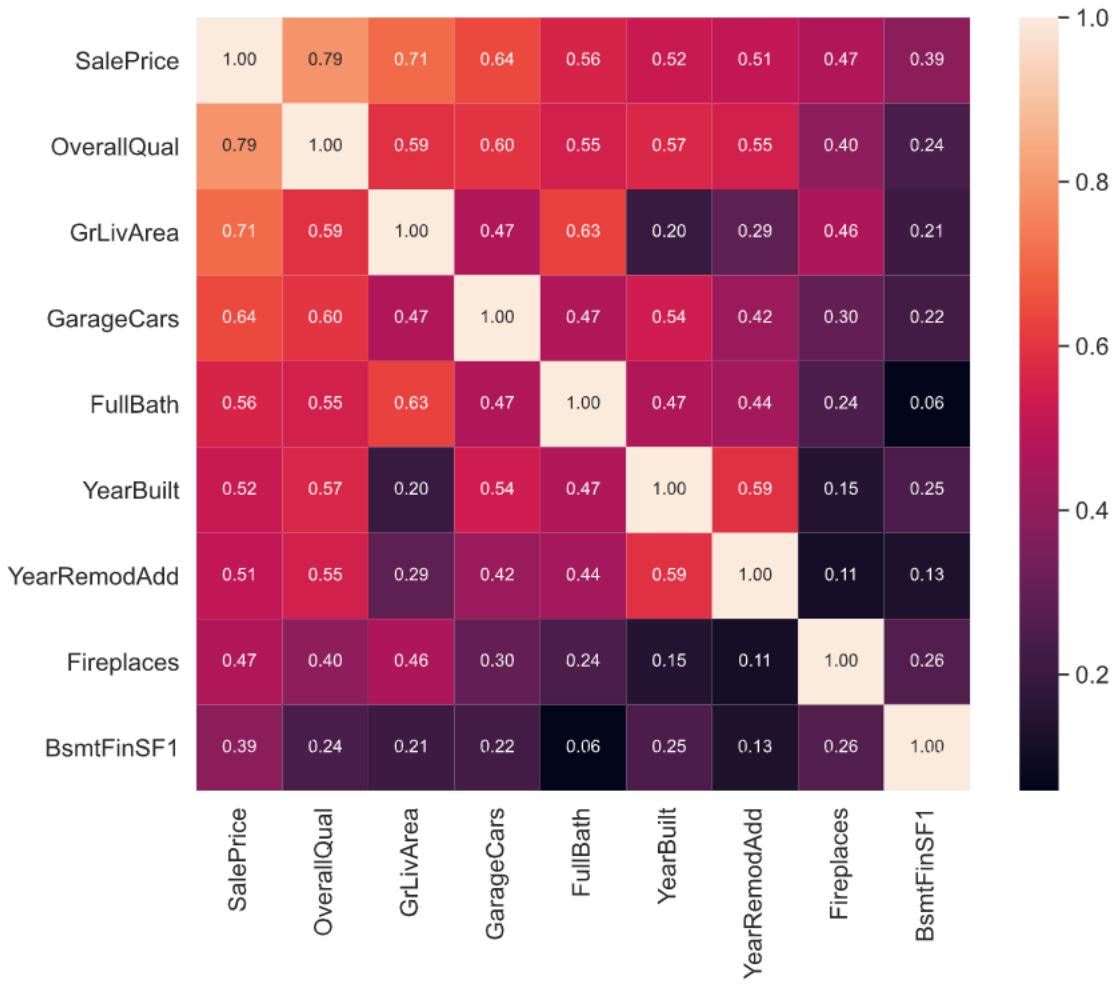
	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeckSF	OpenPorchSF	EnclosedPorch
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1452.000000	1460.000000	...	1460.000000	1460.000000	1460.000000
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.865753	103.685262	443.639726	...	94.244521	46.660274	21.954110
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799	30.202904	20.645407	181.066207	456.098091	...	125.338794	66.256028	61.119149
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.000000	0.000000	383.500000	...	0.000000	25.000000	0.000000
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.000000	166.000000	712.250000	...	168.000000	68.000000	0.000000
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	857.000000	547.000000	552.000000

8 rows x 38 columns

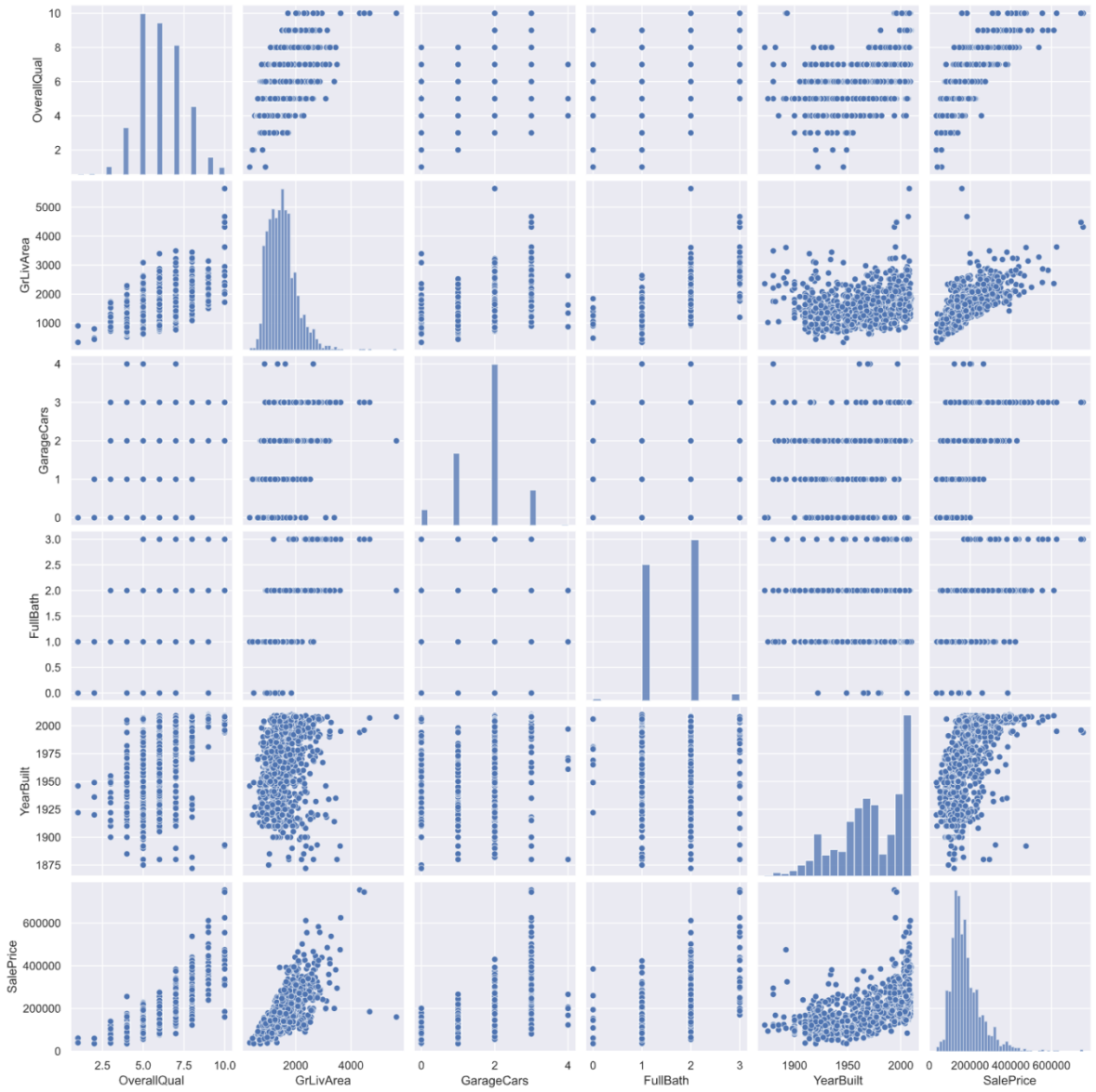


We can see that there is a strong correlation between some of the variables such as garage cars and garage área or total basement square foot and first floor square foot. This implies that if we use this correlated variables as predictors we would have multicollinearity in our model.

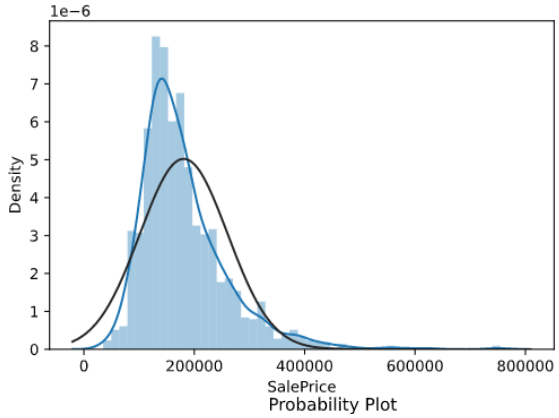
To prevent this from happening I eliminated `garage area`, `total basement SF`, `total rooms above ground`, `1st floor SF`, `garage year built` and `masonry veneer area` from the training data and replotted the correlation heatmap.

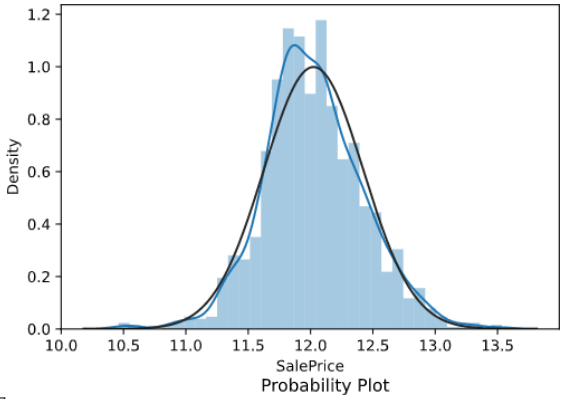


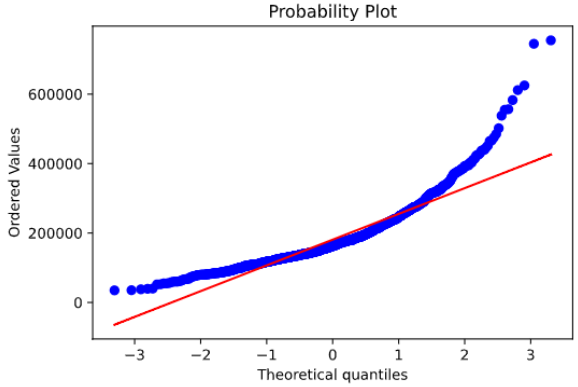
Now we can plot the sale price against the chosen predictors.

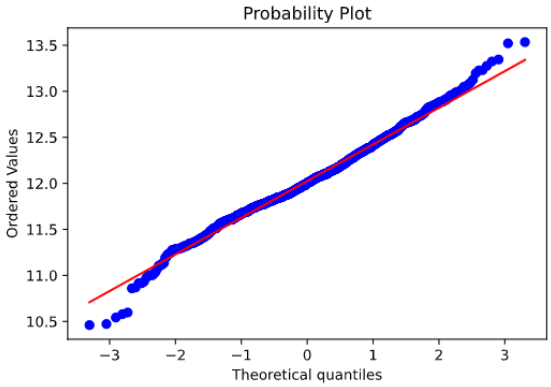


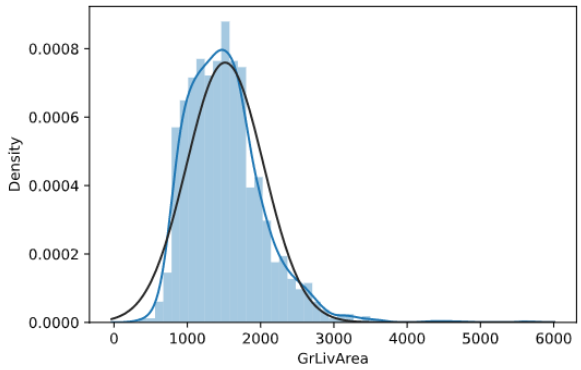
We can observe that sale price is skewed and that ground living area becomes wider as the values grow. Furthermore, the dristibution and probability plots show us that they do no follow a normal distribution. This can cause problems such as heteroscedasticity in our model so I applied a logaritmic tranformation to the sale price and ground living área.

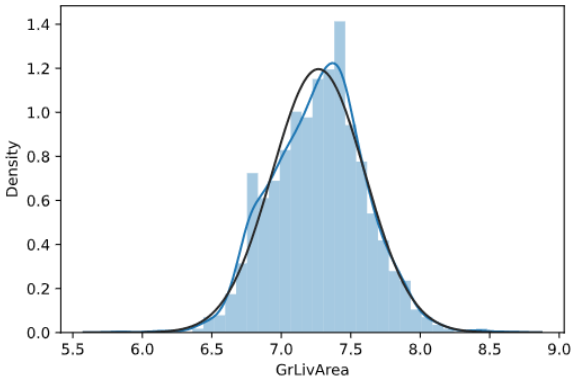

1e−6
Density
SalePrice

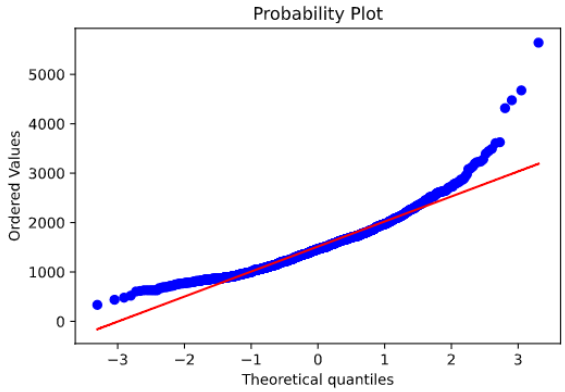

Density
SalePrice

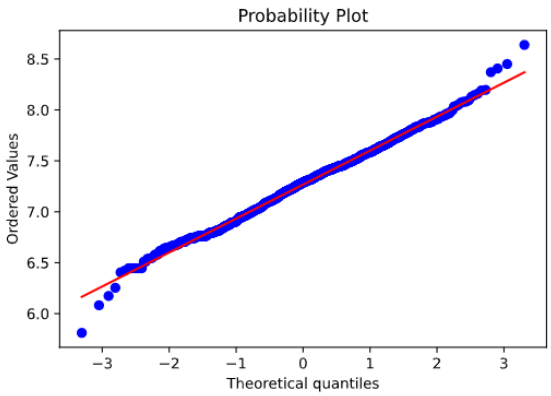

Probability Plot
Ordered Values
Theoretical quantiles


Probability Plot
Ordered Values
Theoretical quantiles

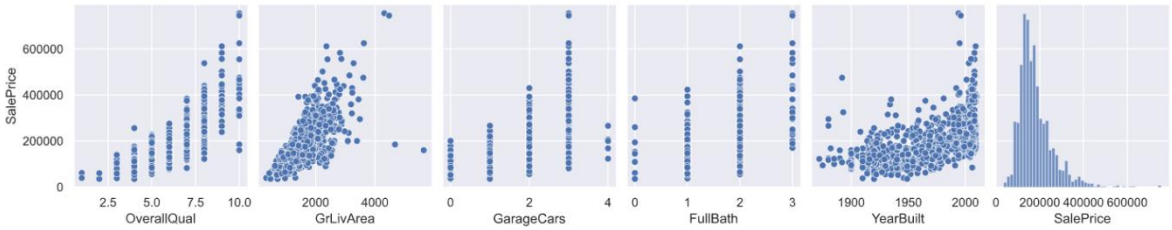

Density
GrLivArea


Density
GrLivArea

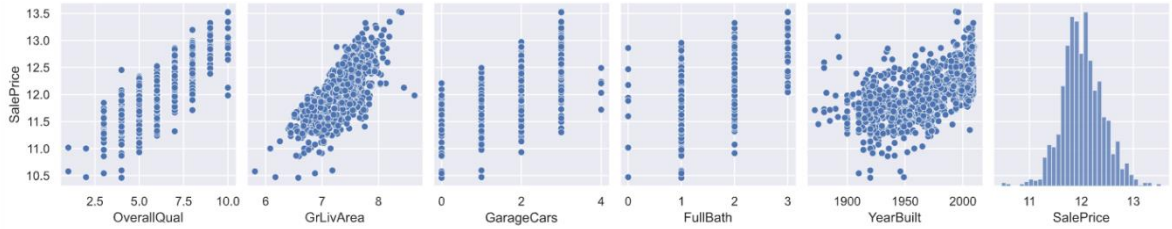

Probability Plot
Ordered Values
Theoretical quantiles


Probability Plot
Ordered Values
Theoretical quantiles

Before applying logarithmic transformation:

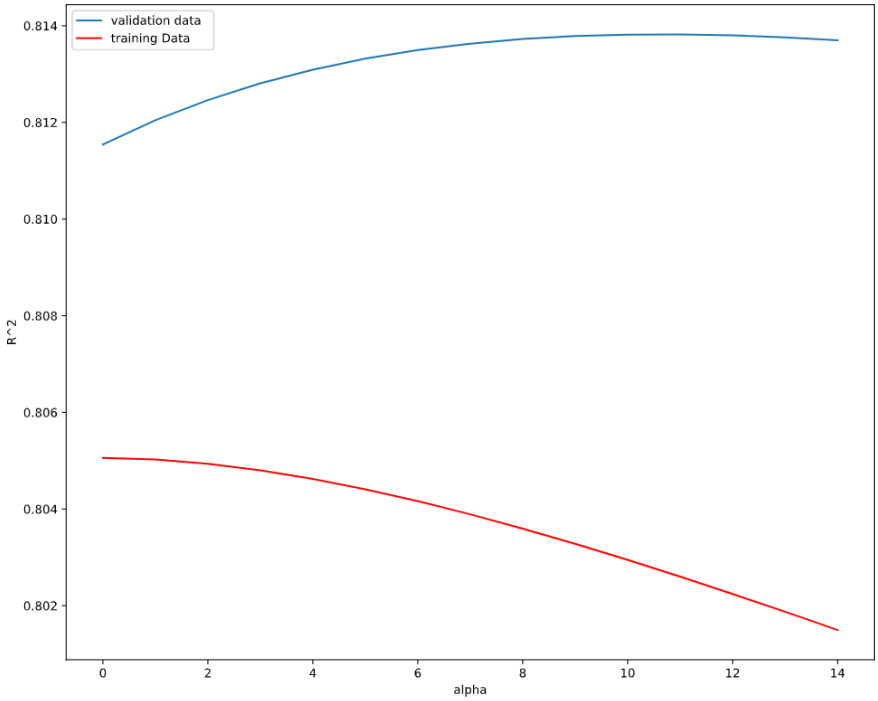


Applying log transformation to saleprice and grlivarea:



Now we can see how all the predictor variables have a linear behaviour when plotted against sale price.

After selecting and transforming the predictor variables we can start testing some models. The first model I decided to test was ridge regression. To visualize the outcomes of the model I plotted the R2 outcomes for diferente values of Alpha using the train and test datasets.



As you can see in the graph the train and test outcomes contradict themselves. We get lower values of R2 in the training set as Alpha grows and higher values in the test set.

After this confusing outcome I decided to test the multiple linear regression model.

To complement the linear regression analysis I used a machine learning k-means cluster model. The model uses all the continuous variables to group and label each individual observation to a cluster. Unfortunately, the labeling process was inconsistent after multiple runs and when added to the linear regression the precisión metrics were not significantly improved.

Last but not least, the R^2 and root mean squared error (RMSE) of the multiple linear regression closely resemble the ones obtained using ridge regression so I decided to go for the simple solution and keep this model for the final test results.

Results

The following results were obtained using the test dataset.

Ridge regression:

```
BestRR = Ridge(alpha=1)
Ridge R^2 = 0.8120441945914397
Ridge RMSE = 0.17041497822092613
Range = 2.652571048781981
```

K means clustering and linear regression:

```
Cluster R^2 = 0.8137879307033405
Cluster RMSE = 0.16962263456404542
Range = 2.652571048781981
```

Multiple linear regression:

```
MLR R^2 = 0.8115446401767943
MLR RMSE = 0.17064129488895512
Range = 2.652571048781981
```

Discussion

I compared my results with the ones obtained by other people using the same sample to create a prediction model and there isnt a big precision improvement and their models tend to get really complicated. Other models use 15> variables to get a .03 improvement which is not efficient nor convenient considering the strong correlation between variables in the dataset.

Conclusion

`Overall quality, ground living área, garage cars, full baths` and `year built` are the best variables when trying to accuaretely predict the sale price of a property in Ames, Iowa. Adding more predictor variables to our model would only complicate things and the improvement would be marginal. A simple model with relatively good accuracy is almost always a better option.