Determine which factors influence housing Prices

CUNY 621 – Final Project

Spring 2020 Semester – 17-May-2020

Author : Deepak Mongia

This is the report for the final project for Data 621 for Regression Analysis on a real life dataset. For this project, I have used the dataset for the Housing prices given at the Kaggle competition web page: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

As a part of the project, we have analyzed the dataset which has 80 independent features which might be helpful in determining the sale price of a house.

Abstract:

The problem we had at hand was to analyze and understand which are the major important features which determine a house price. Buying a house is a very important achievement in everyone's life. We wanted to explore what all factors are important in determination of the price of a house.

We cleaned up our data, and made it ready for our predictive models. We followed the linear regression approach to predict the sale price of a house based on multiple independent features.

Based on our research and analysis, we found out that the most important features that determine the price of a house are:

Overall quality and condition

Neighborhood Pricing

Year built

Keywords:

SalePrice | Quality | Construction

Year built | Living Area

Introduction:

The dataset was built by Dean De Cock of Truman State University to build a real life dataset for advanced regression projects for his university. The data describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. Ames is a city in Story County, Iowa, United States, located approximately 30 miles (48 km) north of Des Moines in central Iowa. It is best known as the home of Iowa State University (ISU), with leading agriculture, design, engineering, and veterinary medicine colleges.

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. The dataset is on the same lines on the very famous Boston housing dataset that has been used in various data science projects and discussions.

The original paper by De Cock about this dataset is here: http://jse.amstat.org/v19n3/decock.pdf

Problem Motivation:

We wanted to explore the numerous features which buyers consider before they decide to buy a house. Buying a house is a very critical task, and anyone who is buying a new house wants to ensure that they make a well informed decision so that they don't have to face troubles later on. To get one's dream house, one must consider all these important features.

My motivation to do this project was to see how predictive modeling can help the new buyers by giving them a good prediction of the amount any house is worth, so that they can make intelligent decisions to get their dream home.

Literature Review:

De Cock in his paper given above mentions a good deal about how this dataset can be approached to build predictive models by students to predict the sale price of a house. This dataset has been created for academic work primarily. There have been multiple papers written on this. As the dataset has a vast number of features, it has been used by many data scientists to work on feature engineering.

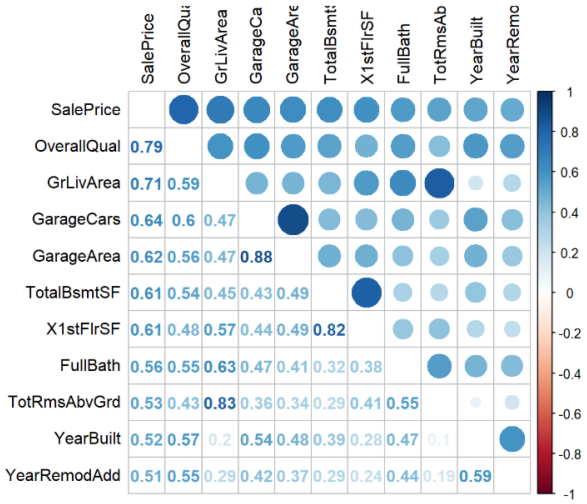
As the dataset is to predict the sale price of the house, which is a continuous variable, based on multiple independent features, the linear regression is the best option here.

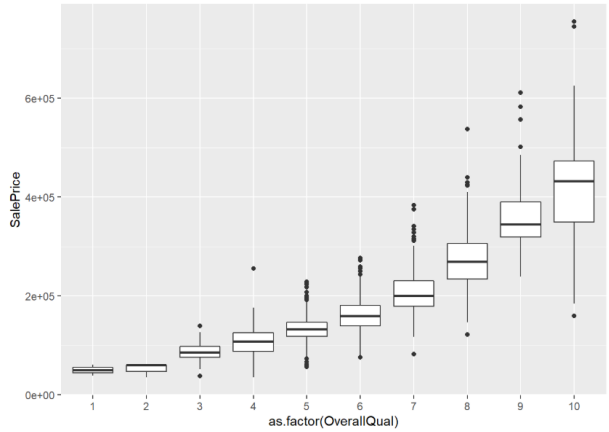
As a part of our project, we have worked on detailed feature engineering so that we have the data ready to be fed into the machine learning models.

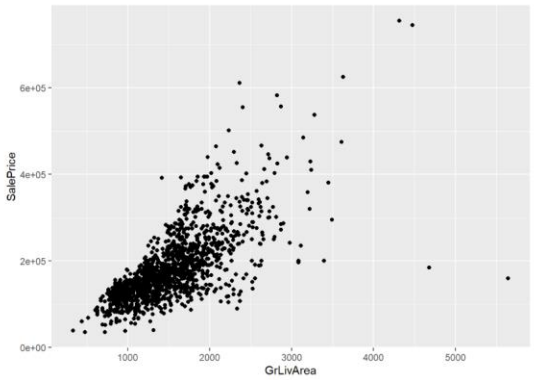
In the papers I got a chance to review like this one: https://www.lexjansen.com/scsug/2017/MK29.pdf, the authors dealt with this data thru detailed feature engineering, outlier removal and then doing the linear regression on the data to predict the sale price of a house.

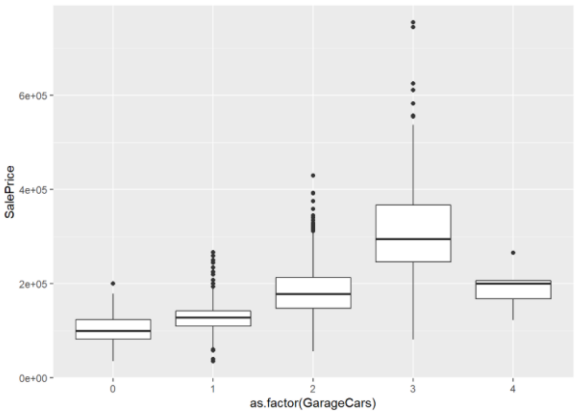
Methodology:

First things first, we first did a detailed exploratory data analysis to understand the data at hand. From the correlation plot, it is very clear that Sale Price is highly related to Garage Area, First floor area and overall quality:

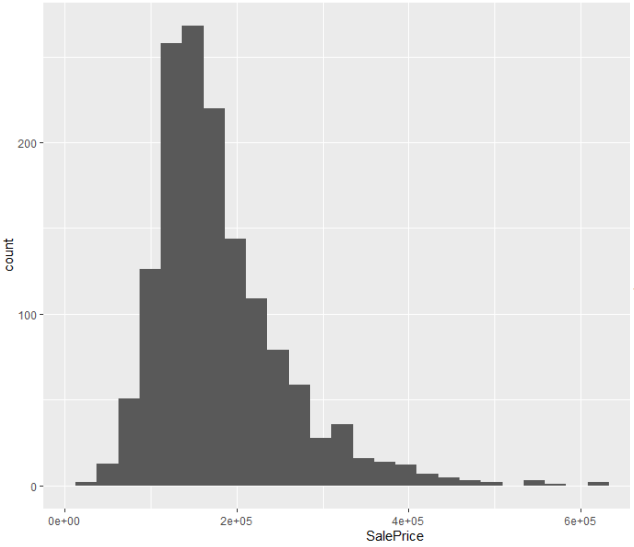


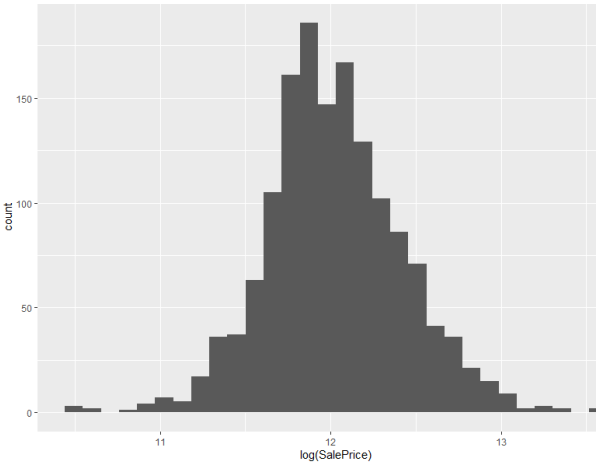




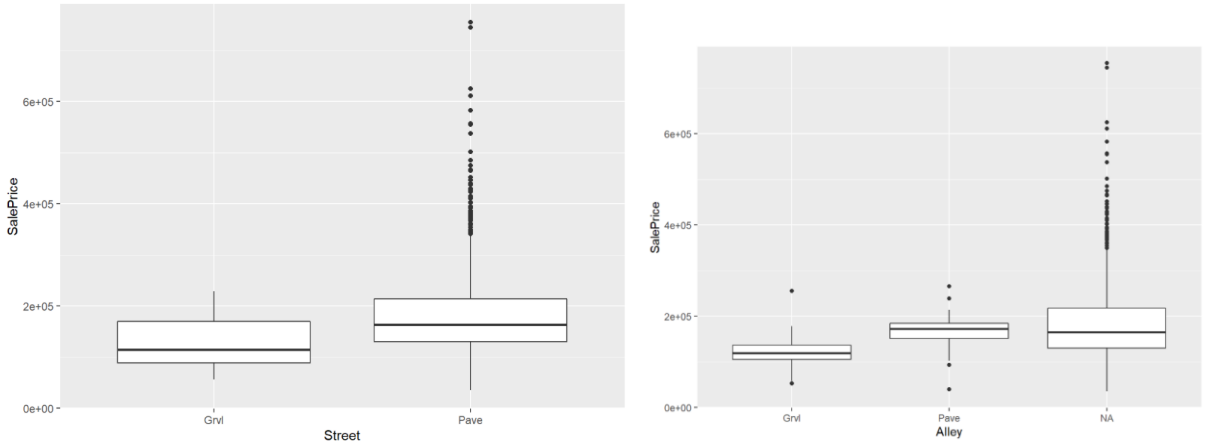


We also explored each numeric variable to see the distribution. The target variable – sale proce has a right skewed distribution, which we then took a log of, and the log shows a normal distribution.





Some of the categorical features showed a very clear relationship with the target variable.



We had to convert all the categorical features to numerical encoded variables so that those can be used to feed into prediction models. Like Street which is 'Paved' took value 1 and which is 'Gravel' took 0.

Experimentation and Results:

To build proper models, we must be having some test data so that we can test our results. For that we used the test-train split method. From 1500 observations, we took 80% for training models and 20% for testing the models.

Test-Train split

Splitting the data into test and train datasets

```
n <- nrow(housing2_raw)
set.seed(123)
housing2_random_index <- housing2_raw[sample(n), ]

housing2.train.df <- housing2_raw[1:as.integer(0.8*n),]

housing2.test.df <- housing2_raw[as.integer(0.8*n +1):n, ]
```

As the target variable is a continuous numeric variable, we used linear regression.

Model-1 : We built a model with all the variables to start with. With this, we got many statistically insignificant features. So we went further to built further models.

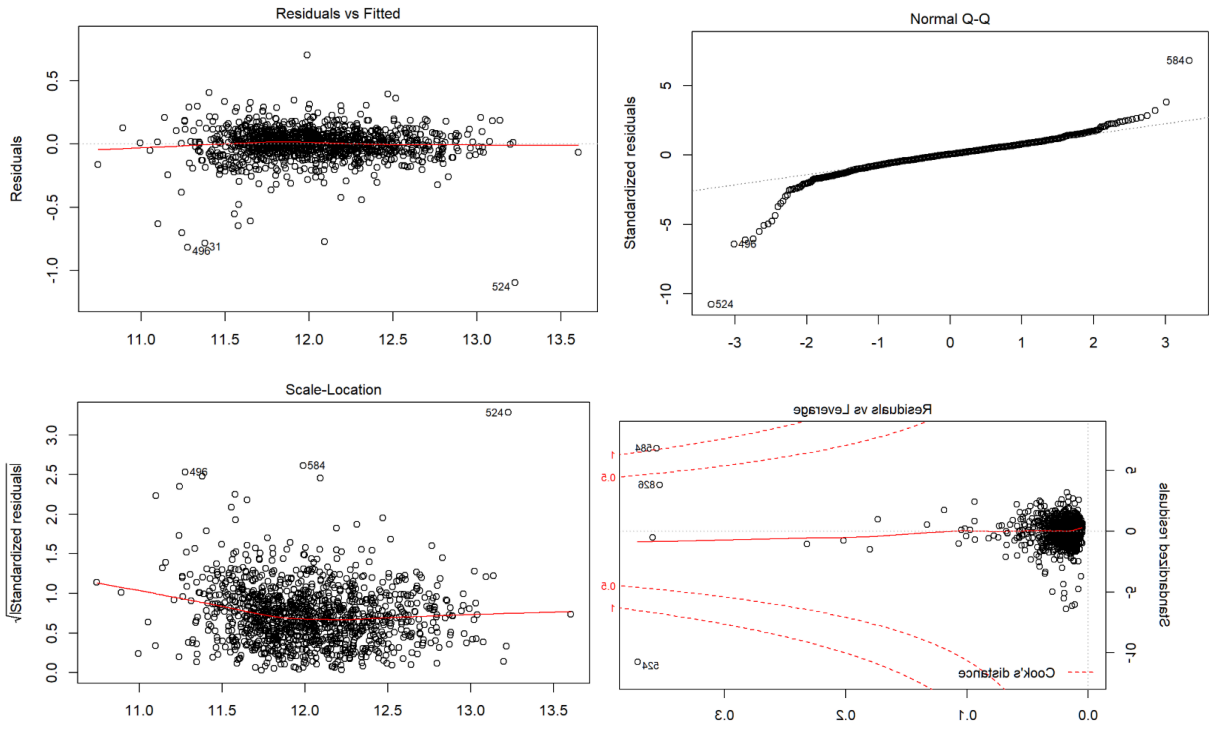
Till Model-3, we were getting a curved residual plot, so we decided to try models with log of the target variable – Sale Price and that gave us straight residual plots. Thus all the assumptions for linear model with the model having log of Sale Price held true.

The final model we selected is:

```
##
## Call:
## lm(formula = log(SalePrice) ~ MSSubClass + LotFrontage + LotArea +
##     OverallQual + OverallCond + YearBuilt + BsmtFinSF1 + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + LowQualFinSF + KitchenAbvGr +
##     TotRmsAbvGrd + GarageArea + WoodDeckSF + ScreenPorch + Neighborhood_highprice +
##     Condition2_good + BsmtQual_level + BsmtExposure_level + KitchenQual_level +
##     SaleType_value + GarageCond_status, data = housing2.train.df)
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.587e+00  4.021e-01  11.406  < 2e-16 ***
## MSSubClass             -4.306e-04  1.181e-04  -3.645 0.000280 ***
## LotFrontage             8.575e-04  2.185e-04   3.924 9.22e-05 ***
## LotArea                 1.692e-06  3.938e-07   4.296 1.88e-05 ***
## OverallQual             6.884e-02  5.162e-03  13.335  < 2e-16 ***
## OverallCond             5.109e-02  3.909e-03  13.072  < 2e-16 ***
## YearBuilt               2.887e-03  2.104e-04  13.722  < 2e-16 ***
## BsmtFinSF1              1.421e-04  2.024e-05   7.021 3.76e-12 ***
## BsmtFinSF2              9.506e-05  2.938e-05   3.236 0.001248 **
## BsmtUnfSF               5.479e-05  1.961e-05   2.794 0.005287 **
## X1stFlrSF               2.784e-04  2.368e-05  11.753  < 2e-16 ***
## X2ndFlrSF               2.472e-04  1.758e-05  14.064  < 2e-16 ***
## LowQualFinSF            1.494e-04  7.749e-05   1.929 0.054020 .
## KitchenAbvGr           -4.621e-02  2.141e-02  -2.158 0.031108 *
## TotRmsAbvGrd            1.319e-02  4.674e-03   2.821 0.004864 **
## GarageArea              1.115e-04  2.803e-05   3.977 7.42e-05 ***
## WoodDeckSF              8.091e-05  3.363e-05   2.406 0.016276 *
## ScreenPorch             3.091e-04  7.130e-05   4.336 1.58e-05 ***
## Neighborhood_highprice  9.351e-02  1.097e-02   8.524  < 2e-16 ***
## Condition2_good        -6.052e-01  7.775e-02  -7.784 1.57e-14 ***
## BsmtQual_level          2.151e-02  8.507e-03   2.529 0.011568 *
## BsmtExposure_level      8.058e-03  4.298e-03   1.875 0.061032 .
## KitchenQual_level       3.169e-02  8.297e-03   3.819 0.000141 ***
## SaleType_value          5.696e-02  1.162e-02   4.903 1.08e-06 ***
## GarageCond_status       2.582e-02  9.331e-03   2.767 0.005741 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1286 on 1143 degrees of freedom
## Multiple R-squared:  0.8996, Adjusted R-squared:  0.8975
## F-statistic: 426.6 on 24 and 1143 DF,  p-value: < 2.2e-16
```

Plots:



Discussions and Conclusions:

1. There is a very clear Log linear relationship between the target feature - Sale Price and the following independent features.

MSSubClass	LotFrontage	LotArea
OverallQual	OverallCond	YearBuilt
BsmtFinSF1	BsmtFinSF2	BsmtUnfSFX1stFlrSF
X2ndFlrSF	LowQualFinSF	KitchenAbvGr
TotRmsAbvGrd	GarageArea	WoodDeckSF
ScreenPorch	Neighborhood_highprice	Condition2_good
BsmtQual_level	BsmtExposure_level	KitchenQual_level
SaleType_value	GarageCond_status	

2. The most relevant features are:

OverallQual	OverallCond	YearBuilt
BsmtFinSF1	X1stFlrSF	X2ndFlrSF

Condition2_good

3. The last model is our final model and can be used to deploy for the prediction.

References:

1. Data source: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
2. http://jse.amstat.org/v19n3/decock.pdf

Appendices:

1. R code - https://raw.githubusercontent.com/deepakmongia/Data621/master/Final%20Project/Data_621_Final_Project.Rmd