EDA of the Ames Housing Report

Joel Riesen, SPS MSDS Northwestern University

04/11/2021

Section 1: Introduction

The Ames data set submitted by Dean De Cock of Truman State University. The original data contains individual residential properties sold in Ames, IA, from 2006 to 2010. The data includes 82 columns, 23 nominal, 23 ordinals, 14 discrete, and 20 continuous variables. It contains 2930 observations.

Table 1 : Ames Housing descriptions from http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

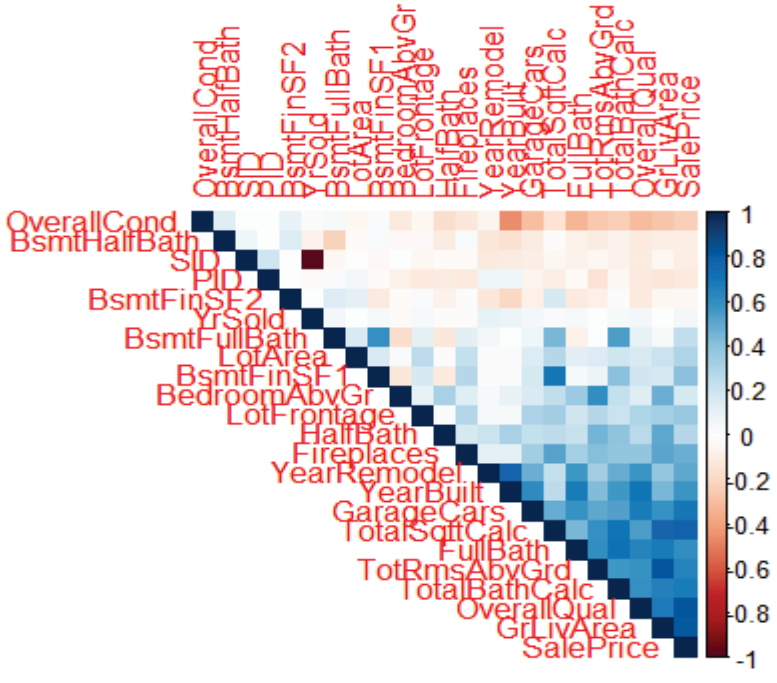
1	Order (Discrete)	Observation number	42	HeatingQC (Ordinal)	Heating quality and condition
2	PID (Nominal)	Parcel identification number - can be used with city website for parcel review.	43	Central Air (Nominal)	Central air conditioning
3	MS SubClass (Nominal)	Identifies the type of dwelling involved in the sale.	44	Electrical (Ordinal)	Electrical system
4	MS Zoning (Nominal)	Identifies the general zoning classification of the sale.	45	1st Flr SF (Continuous)	First Floor square feet
5	Lot Frontage (Continuous)	Linear feet of street-connected to property	46	2nd Flr SF (Continuous)	Second-floor square feet
6	Lot Area (Continuous)	Lot size in square feet	47	Low Qual Fin SF (Continuous)	Low quality finished square feet (all floors)
7	Street (Nominal)	Type of road access to property	48	Gr Liv Area (Continuous)	Above grade (ground) living area square feet
8	Alley (Nominal)	Type of alley access to property	49	Bsmt Full Bath (Discrete)	Basement full bathrooms
9	Lot Shape (Ordinal)	General shape of property	50	Bsmt Half Bath (Discrete)	Basement half bathrooms
10	Land Contour (Nominal)	Flatness of the property	51	Full Bath (Discrete)	Full bathrooms above grade
11	Utilities (Ordinal)	Type of utilities available	52	Half Bath (Discrete)	Half baths above grade
12	Lot Config (Nominal)	Lot configuration	53	Bedroom (Discrete)	Bedrooms above grade (does NOT include basement bedrooms)
13	Land Slope (Ordinal)	Slope of property	54	Kitchen (Discrete)	Kitchens above grade
14	Neighborhood (Nominal)	Physical locations within Ames city limits (map available)	55	KitchenQual (Ordinal)	Kitchen quality
15	Condition 1 (Nominal)	Proximity to various conditions	56	TotRmsAbvGrd (Discrete)	Total rooms above grade (does not include bathrooms)
16	Condition 2 (Nominal)	Proximity to various conditions (if more than one is present)	57	Functional (Ordinal)	Home functionality (Assume typical unless deductions are warranted)
17	Bldg Type (Nominal)	Type of dwelling	58	Fireplaces (Discrete)	Number of fireplaces
18	House Style (Nominal)	Style of dwelling	59	FireplaceQu (Ordinal)	Fireplace quality
19	Overall Qual (Ordinal)	Rates the overall material and finish of the house	60	Garage Type (Nominal)	Garage location
20	Overall Cond (Ordinal)	Rates the overall condition of the house	61	Garage Yr Blt (Discrete)	Year garage was built
21	Year Built (Discrete)	Original construction date	62	Garage Finish (Ordinal)	Interior finish of the garage
22	Year Remod/Add (Discrete)	Remodel date (same as construction date if no remodeling or additions)	63	Garage Cars (Discrete)	Size of garage in car capacity
23	Roof Style (Nominal)	Type of roof	64	Garage Area (Continuous)	Size of garage in square feet
24	Roof Matl (Nominal)	Roof material	65	Garage Qual (Ordinal)	Garage quality

25	Exterior 1 (Nominal)	Exterior covering on house	66	Garage Cond (Ordinal)	Garage condition
26	Exterior 2 (Nominal)	Exterior covering on house (if more than one material)	67	Paved Drive (Ordinal)	Paved driveway
27	Mas Vnr Type (Nominal)	Masonry veneer type	68	Wood Deck SF (Continuous)	Wood deck area in square feet
28	Mas Vnr Area (Continuous)	Masonry veneer area in square feet	69	Open Porch SF (Continuous)	Open porch area in square feet
29	Exter Qual (Ordinal)	Evaluates the quality of the material on the exterior	70	Enclosed Porch (Continuous)	Enclosed porch area in square feet
30	Exter Cond (Ordinal)	Evaluates the present condition of the material on the exterior	71	3-Ssn Porch (Continuous)	Three season porch area in square feet
31	Foundation (Nominal)	Type of foundation	72	Screen Porch (Continuous)	Screen porch area in square feet
32	Bsmt Qual (Ordinal)	Evaluates the height of the basement	73	Pool Area (Continuous)	Pool area in square feet
33	Bsmt Cond (Ordinal)	Evaluates the general condition of the basement	74	Pool QC (Ordinal)	Pool quality
34	Bsmt Exposure (Ordinal)	Refers to walkout or garden level walls	75	Fence (Ordinal)	Fence quality
35	BsmtFin Type 1 (Ordinal)	Rating of basement finished area	76	Misc Feature (Nominal)	Miscellaneous feature not covered in other categories
36	BsmtFin SF 1 (Continuous)	Type 1 finished square feet	77	Misc Val (Continuous)	$Value of miscellaneous feature
37	BsmtFinType 2 (Ordinal)	Rating of basement finished area (if multiple types)	78	Mo Sold (Discrete)	Month Sold (MM)
38	BsmtFin SF 2 (Continuous)	Type 2 finished square feet	79	Yr Sold (Discrete)	Year Sold (YYYY)
39	Bsmt Unf SF (Continuous)	Unfinished square feet of basement area	80	Sale Type (Nominal)	Type of sale
40	Total Bsmt SF (Continuous)	Total square feet of basement area	81	Sale Condition (Nominal)	Condition of sale
41	Heating (Nominal)	Type of heating	82	SalePrice (Continuous)	Sale price $$

The reason for the analysis is to use the data to fit specific models. We will start with a Simple Linear Regression (SLR). SalesPrice will be our response variable. We will conduct an EDA to find two predictor variables to help us understand the data better. Next, we will complete a Multiple Linear Regression (MLR) using the two predictor variables from the SLR. We will add two more predictor variables which can be continuous or discrete.

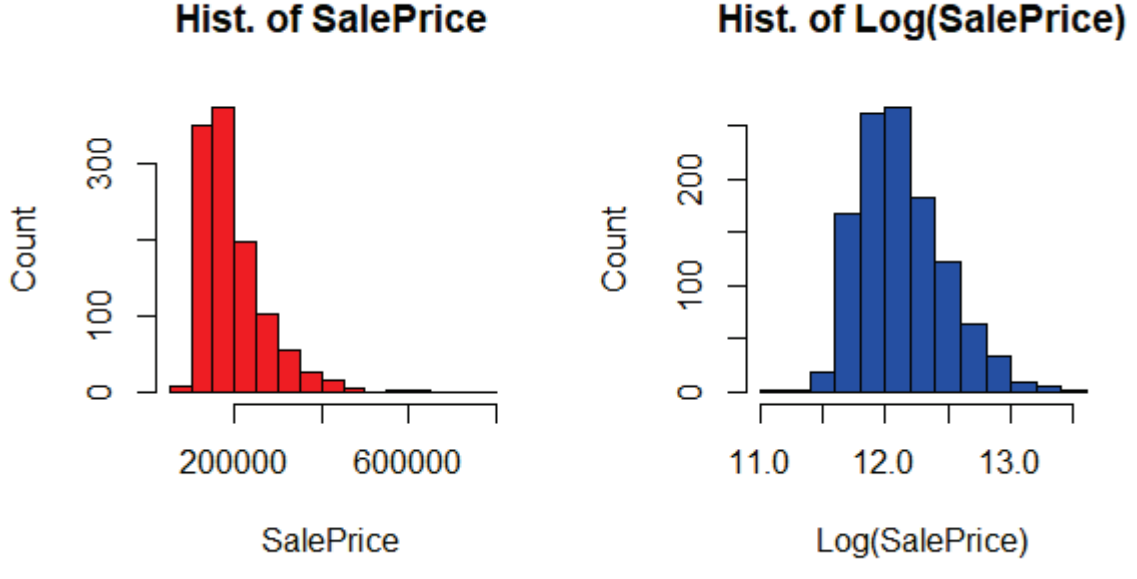
To find the most vital relationship with SalePrice, we will run a correlation matrix heat map to narrow down top relationships. We need to take out all the non-numeric columns first.

Fig1 Correlation Matrix



From this, we can see that GrLivArea, OverallQual, TotalBathCalc, and TotRmsAbvGrd. Looking at the data, we should see if Logging SalePrice will help to smooth out the data.

Fig 2: Histogram of logged SalesPrice v SalePrice

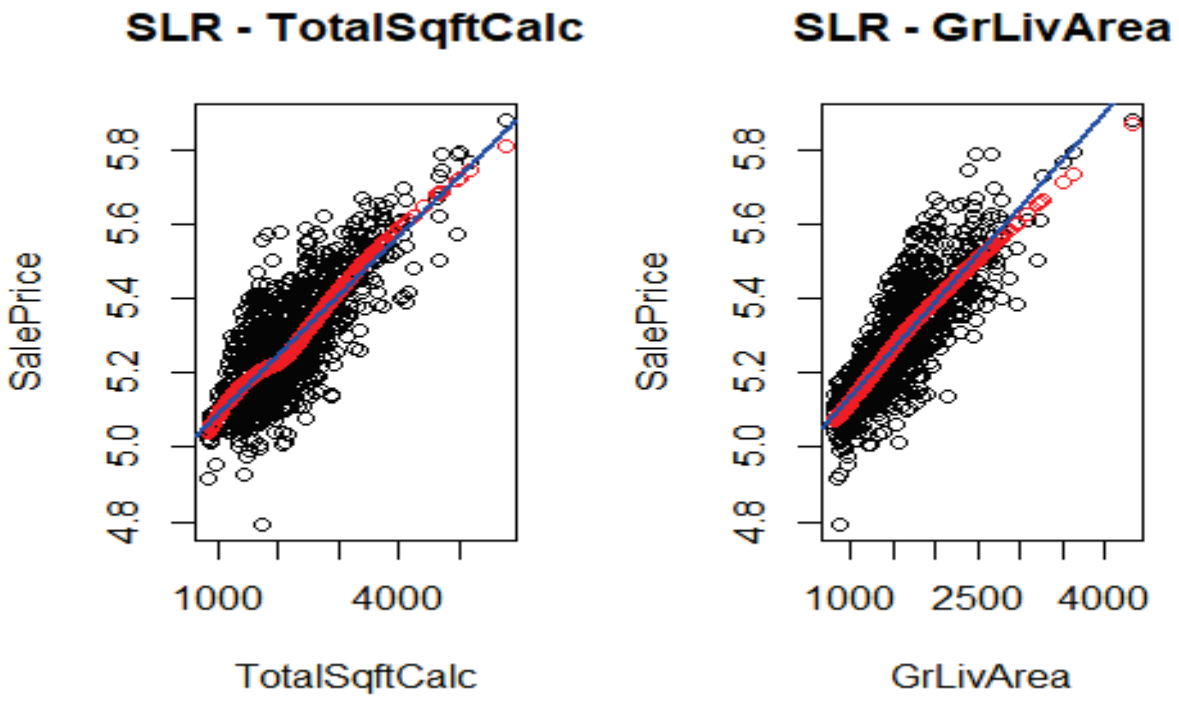


Looking at the graphs, the logged SalePrice looks more like a normal distribution than the non-logged SalePrice. So we will use the logged SalesPrice; the predictor variables will be GrLivArea, and TotalSqftCalc.

Section 2: Simple Linear Regression Models

The predictor variables are GrLivArea, and TotalSqftCalc. Both deal with the square feet of the house. The TotalSqftCalc also looks to include the basement.

Fig 3: SLR of the two predictor variables



From the graphs, the blue line is the fitted linear model, and the red is the Loess local fit. TotalSqftCalc is a little more accurate in helping to predict the SalePrice.

Section 2.3 Model Comparison

Comparing the two predictors and how they relate to SalePrice

Table 2: Linear Regression Results

Fitted SLR

	Dependent variable:	
	SalePrice	
	TotalSqftCalc	GrLivArea
	(1)	(2)
TotalSqftCalc	0.0002*** (0.00000)	
GrLivArea		0.0003*** (0.00001)
Constant	4.930*** (0.009)	4.879*** (0.008)
Observations	1,135	1,135
R^2	0.589	0.694
Adjusted R^2	0.588	0.694
Residual Std. Error (df = 1133)	0.095	0.082
F Statistic (df = 1; 1133)	1,620.627***	2,574.373***
Note:		*p<0.1; **p<0.05; ***p<0.01

Both have an R^2 above 50%; however, GrLivArea is almost 70%. So I would make sure to include it in the prediction.

Section 4: Multiple Linear Regression Model

Putting both the TotalSqftCalc and GrLivArea for a multiple linear regression model. It does help a little bit pushing the R^2 up over 70%.

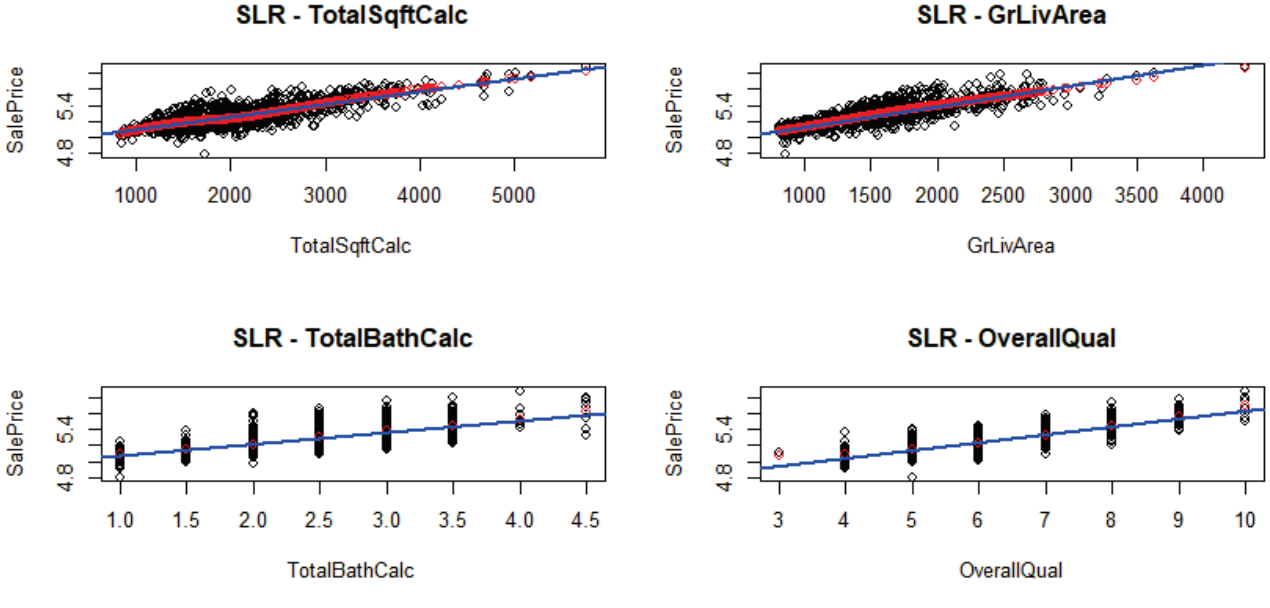
Table 3: The two predictors combined

Fitted MLR

	Dependent variable:		
	SalePrice		
	TotalSqftCalc	GrLivArea	Combined
	(1)	(2)	(3)
TotalSqftCalc	0.0002*** (0.00000)		0.0001*** (0.00001)
GrLivArea		0.0003*** (0.00001)	0.0002*** (0.00001)
Constant	4.930*** (0.009)	4.879*** (0.008)	4.854*** (0.008)
Observations	1,135	1,135	1,135
R^2	0.589	0.694	0.733
Adjusted R^2	0.588	0.694	0.733
Residual Std. Error	0.095 (df = 1133)	0.082 (df = 1133)	0.076 (df = 1132)
F Statistic	1,620.627*** (df = 1; 1133)	2,574.373*** (df = 1; 1133)	1,554.249*** (df = 2; 1132)
Note:			*p<0.1; **p<0.05; ***p<0.01

To see if we can increase this by adding more predictor variables. Using the subsequent two strongest correlations, we will add OverallQual and TotalBathCalc.

Fig 4: The four predictors



The lines are in the above charts are very close together. Adding the extra two predictors for four predictors, we now have an R^2 of almost 90%.

Table 4: Linear regression results of the four predictors

Fitted MLR

	Dependent variable:				
			SalePrice		
	TotalSqftCalc	GrLivArea	TotalBathCalc	OverallQual	
	(1)	(2)	(3)	(4)	(5)
TotalSqftCalc	0.0002*** (0.00000)				0.00005*** (0.00000)
GrLivArea		0.0003*** (0.00001)			0.0001*** (0.00001)
TotalBathCalc			0.144*** (0.004)		0.021*** (0.003)
OverallQual				0.097*** (0.002)	0.057*** (0.002)
Constant	4.930*** (0.009)	4.879*** (0.008)	4.931*** (0.010)	4.655*** (0.011)	4.644*** (0.007)
Observations	1,135	1,135	1,135	1,135	1,135
R^2	0.589	0.694	0.531	0.739	0.890
Adjusted R^2	0.588	0.694	0.530	0.739	0.890
Residual Std. Error	0.095 (df = 1133)	0.082 (df = 1133)	0.101 (df = 1133)	0.075 (df = 1133)	0.049 (df = 1130)
F Statistic	1,620.627*** (df = 1; 1133)	2,574.373*** (df = 1; 1133)	1,280.411*** (df = 1; 1133)	3,215.188*** (df = 1; 1133)	2,283.792*** (df = 4; 1130)
Note:					*p<0.1; **p<0.05; ***p<0.01

Section 5: Transformed MLR Model

So did we make the right call by transforming SalePrice and using it for this report?

Table 5: Comparison of logged v non logged.

SalesPrice not logged(transformed)

	SID	PID	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodel	BsmtFinSF1	BsmtFinSF2	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	TotRmsAbvGrd
SalePrice	-0.10	-0.12	0.36	0.29	0.82	-0.24	0.58	0.50	0.40	-0.05	0.81	0.24	-0.10	0.61	0.27	0.18	0.65

SalesPrice logged

	SID	PID	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodel	BsmtFinSF1	BsmtFinSF2	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	TotRmsAbvGrd
SalePrice	-0.10	-0.10	0.36	0.28	0.86	-0.26	0.65	0.58	0.36	-0.07	0.83	0.22	-0.11	0.69	0.31	0.21	0.68

The graphs confirm the log, so I think it was a good choice.

Fig 5: Log vs. non-log

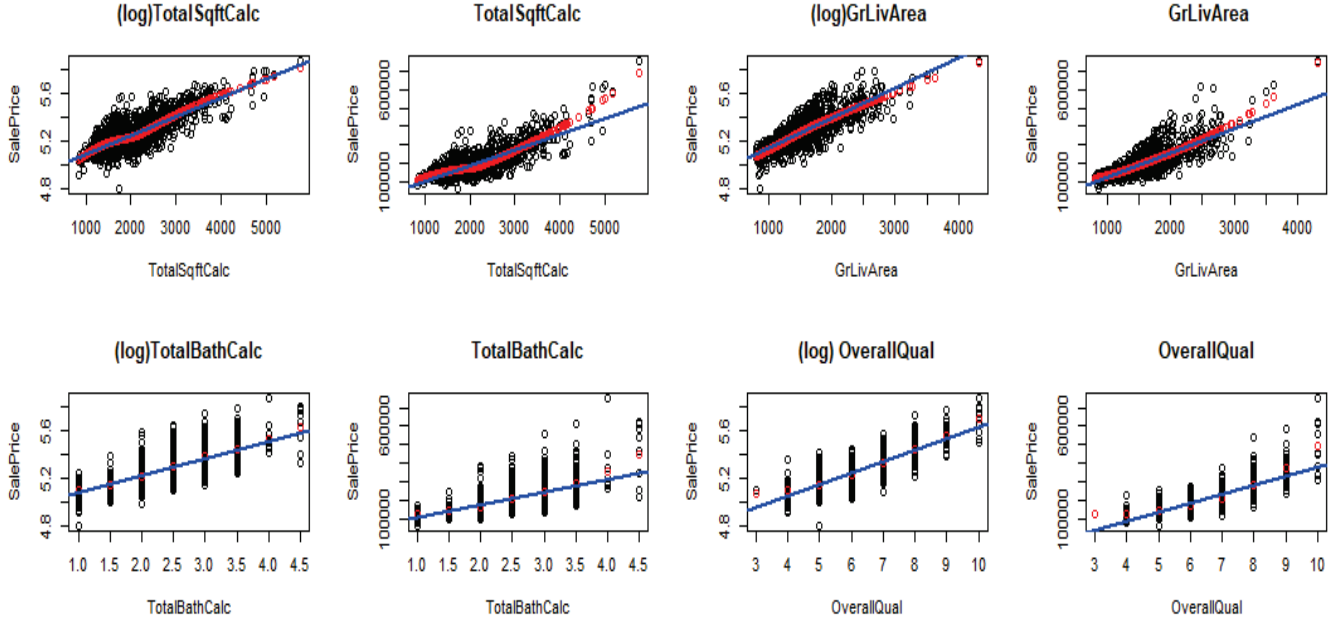


Table 6: The linear comparison

Fitted MLR

	Dependent variable:	
	SalePrice	
	log-SalePrice	SalePrice
	(1)	(2)
TotalSqftCalc	0.00005*** (0.00000)	41.437*** (2.152)
GrLivArea	0.0001*** (0.00001)	30.377*** (3.328)
TotalBathCalc	0.021*** (0.003)	-1,878.221 (1,807.801)
OverallQual	0.057*** (0.002)	29,558.380*** (968.756)
Constant	4.644*** (0.007)	-119,037.900*** (4,437.400)
Observations	1,135	1,135
R^2	0.890	0.851
Adjusted R^2	0.890	0.850
Residual Std. Error (df = 1130)	0.049	29,913.560
F Statistic (df = 4; 1130)	2,283.792***	1,611.802***
Note:		*p<0.1; **p<0.05; ***p<0.01

Conclusion

Running through the EDA process, I found a lot of data that I didn't lose. I also found a few outliers that I left in, and I am happy that I did. From the graph of the GrLivArea, vs. the log, you can see the outliers. I think adding a few more predictor variables would help a little bit more in explaining the SalePrice, but I think it would help with increasing its accuracy as well.