Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock Truman State University Journal of Statistics Education Volume 19, Number 3(2011)

-Rishabh Mehta (801257231)

Problem Statement:

Every day, thousands of residences are sold. Every buyer has various queries, such as: What is the genuine price that this home deserves? Is the price I'm paying reasonable? A machine learning approach is developed in this work to forecast a house price based on data about the house (its size, the year it was built in, etc.).

Data Description:

1. The dataset contains 2390 records and 82 features.

The below image is a sample image of the dataset. Detailed dataset can be seen at: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data?select=train.csv

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandCont...	Utilities	LotConfig	LandSlope
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl
7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl
8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl
9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl
10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl
11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl
12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl

Getting a feel of the dataset

1. Mean, SD, min, max, and 50th percentile for each numeric column in the dataset are calculated.

	mean	std	min	50%	max
Order	1465.50	845.96	1.00	1465.50	2930.00
PID	714464496.99	188730844.65	526301100.00	535453620.00	1007100110.00
MS SubClass	57.39	42.64	20.00	50.00	190.00
Lot Frontage	69.22	23.37	21.00	68.00	313.00
Lot Area	10147.92	7880.02	1300.00	9436.50	215245.00
Overall Qual	6.09	1.41	1.00	6.00	10.00
Overall Cond	5.56	1.11	1.00	5.00	9.00
Year Built	1971.36	30.25	1872.00	1973.00	2010.00
Year Remod/Add	1984.27	20.86	1950.00	1993.00	2010.00
Mas Vnr Area	101.90	179.11	0.00	0.00	1600.00
BsmtFin SF 1	442.63	455.59	0.00	370.00	5644.00
BsmtFin SF 2	49.72	169.17	0.00	0.00	1526.00
Bsmt Unf SF	559.26	439.49	0.00	466.00	2336.00
Total Bsmt SF	1051.61	440.62	0.00	990.00	6110.00
1st Flr SF	1159.56	391.89	334.00	1084.00	5095.00
2nd Flr SF	335.46	428.40	0.00	0.00	2065.00
Low Qual Fin SF	4.68	46.31	0.00	0.00	1064.00
Gr Liv Area	1499.69	505.51	334.00	1442.00	5642.00
Bsmt Full Bath	0.43	0.52	0.00	0.00	3.00
Bsmt Half Bath	0.06	0.25	0.00	0.00	2.00
Full Bath	1.57	0.55	0.00	2.00	4.00
Half Bath	0.38	0.50	0.00	0.00	2.00
Bedroom AbvGr	2.85	0.83	0.00	3.00	8.00
Kitchen AbvGr	1.04	0.21	0.00	1.00	3.00
TotRms AbvGrd	6.44	1.57	2.00	6.00	15.00
Fireplaces	0.60	0.65	0.00	1.00	4.00
Garage Yr Blt	1978.13	25.53	1895.00	1979.00	2207.00
Garage Cars	1.77	0.76	0.00	2.00	5.00
Garage Area	472.82	215.05	0.00	480.00	1488.00
Wood Deck SF	93.75	126.36	0.00	0.00	1424.00
Open Porch SF	47.53	67.48	0.00	27.00	742.00

Getting a feel of the dataset

2. Statistical information about non-numerical dataset. unique represents the number of unique values, top represents the most frequent element, and freq represents the frequency of the most frequent element.

	unique	top	freq
MS Zoning	7	RL	2273
Street	2	Pave	2918
Alley	2	Grvl	120
Lot Shape	4	Reg	1859
Land Contour	4	Lvl	2633
Utilities	3	AllPub	2927
Lot Config	5	Inside	2140
Land Slope	3	Gtl	2789
Neighborhood	28	NAmes	443
Condition 1	9	Norm	2522
Condition 2	8	Norm	2900
Bldg Type	5	1Fam	2425
House Style	8	1Story	1481
Roof Style	6	Gable	2321
Roof Matl	8	CompShg	2887
Exterior 1st	16	VinylSd	1026
Exterior 2nd	17	VinylSd	1015
Mas Vnr Type	5	None	1752
Exter Qual	4	TA	1799
Exter Cond	5	TA	2549
Foundation	6	PConc	1310
Bsmt Qual	5	TA	1283
Bsmt Cond	5	TA	2616
Bsmt Exposure	4	No	1906
BsmtFin Type 1	6	GLQ	859
BsmtFin Type 2	6	Unf	2499
Heating	6	GasA	2885
Heating QC	5	Ex	1495
Central Air	2	Y	2734

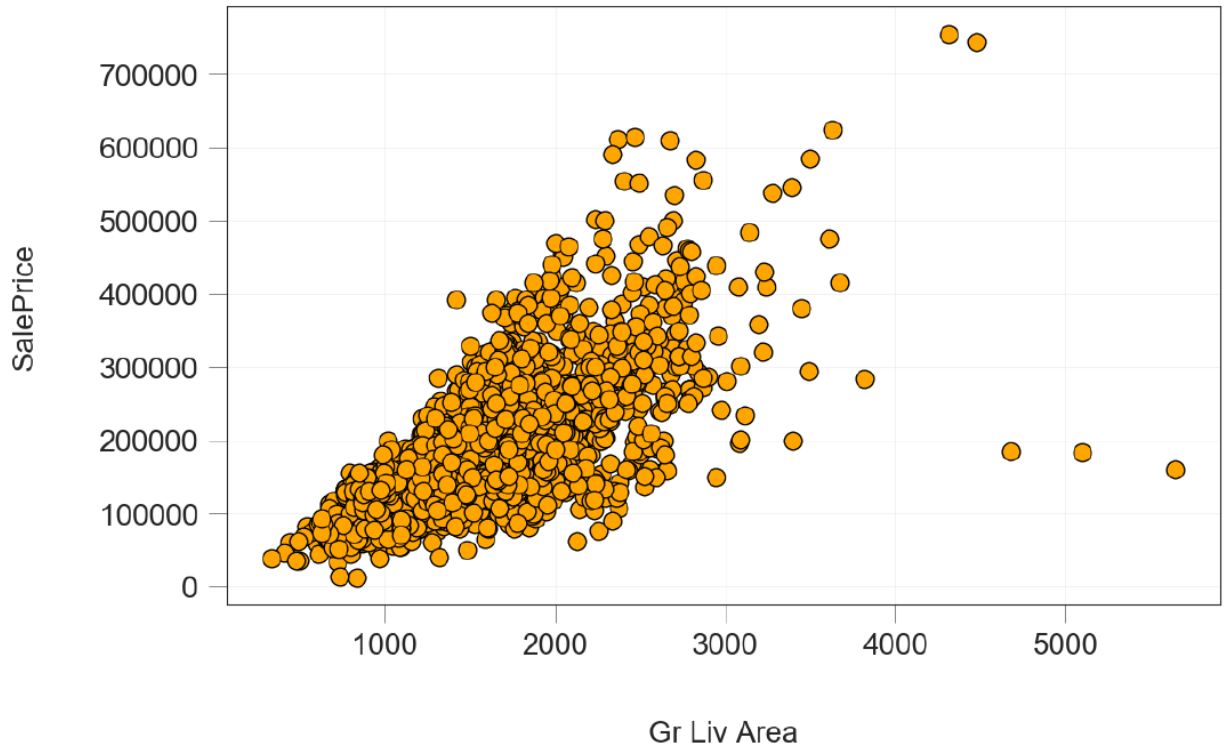
Data Cleaning

1. Dealing with missing values – we see the missing values in dataset as some machine learning don't accept missing values.

	Missing Values	Percentage
Pool QC	2917	99.56
Misc Feature	2824	96.38
Alley	2732	93.24
Fence	2358	80.48
Fireplace Qu	1422	48.53
Lot Frontage	490	16.72
Garage Cond	159	5.43
Garage Qual	159	5.43
Garage Finish	159	5.43
Garage Yr Blt	159	5.43
Garage Type	157	5.36
Bsmt Exposure	83	2.83
BsmtFin Type 2	81	2.76

- fill the missing values in Misc Feature column with "No Feature“
- fill the missing values in Pool QC column with "No Pool“
- fill in the missing values in Alley, Fence, and Fireplace Qu columns with "No Alley", "No Fence", and "No Fireplace“
- fill in the Lot Frontage missing values with 0, etc…

Outlier Removal



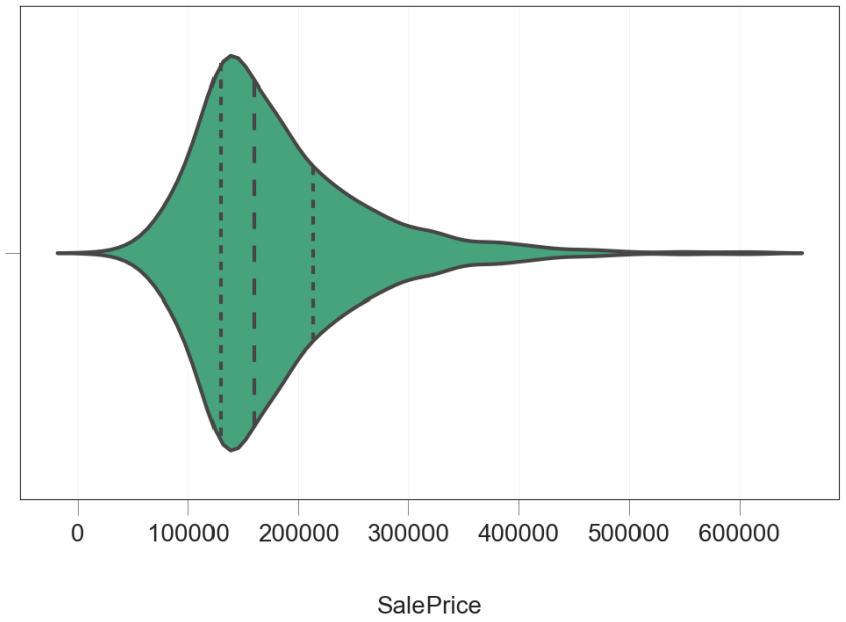
- We keep data points that have Gr Liv Area less than 4,000.
- Delete columns that are not useful in our analysis. The columns to be deleted are Order and PID

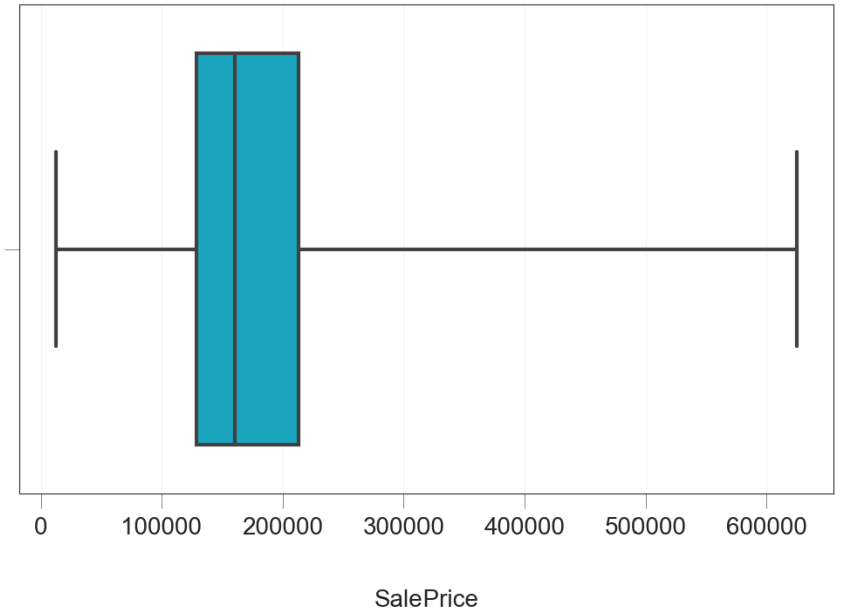
Exploratory Data Analysis

We will explore the data using visualizations. This will allow us to understand the data and the relationships between variables better, which will help us build a better.

Target value distribution:

We plot the violin plot for the target variable. The width of the violin represents the frequency. This means that if a violin is the widest between 300 and 400, then the area between 300 and 400 contains more data points than other

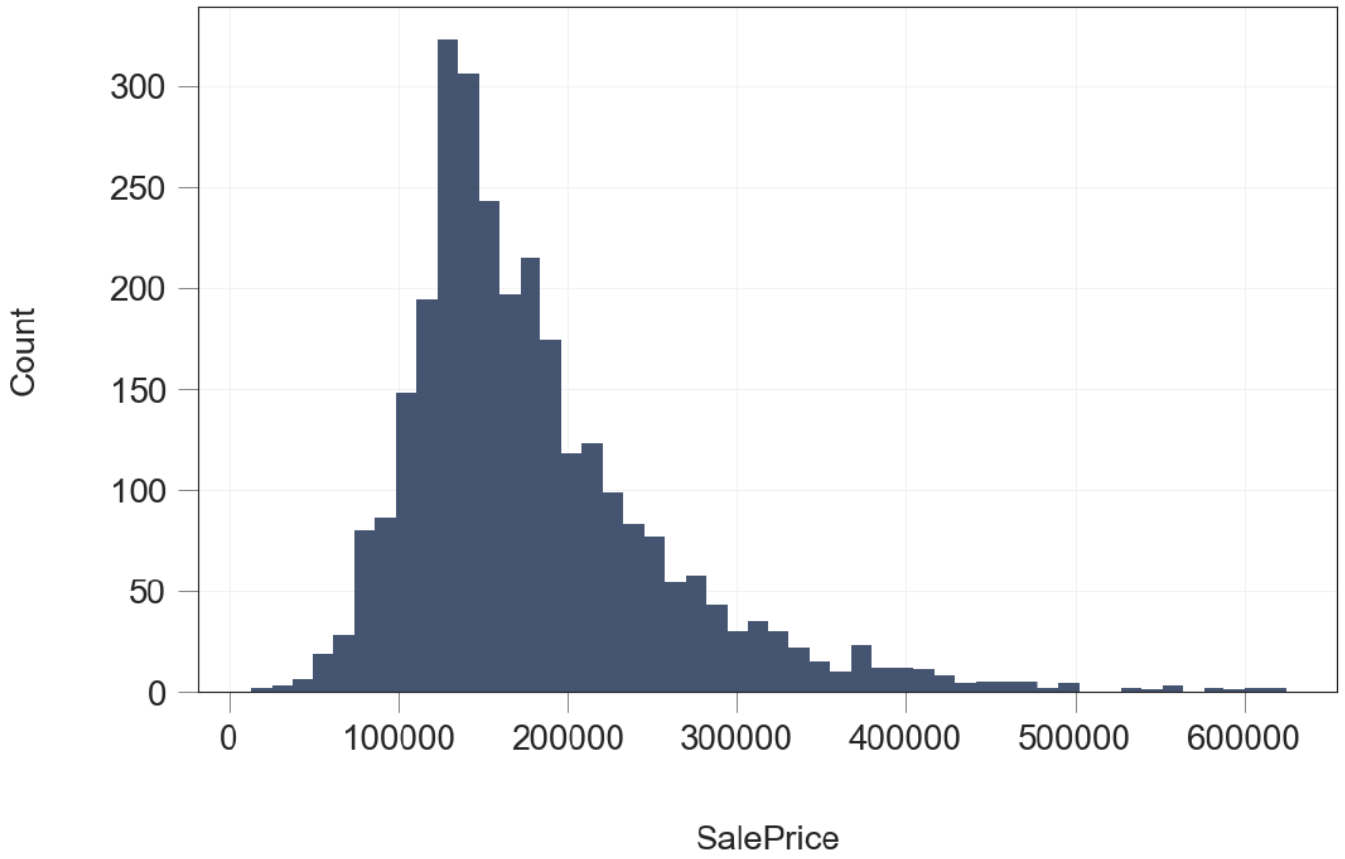




The box plot shows us the minimum and maximum values of SalePrice. It shows us also the three quartiles represented by the box and the vertical line inside of it.

Exploratory Data Analysis

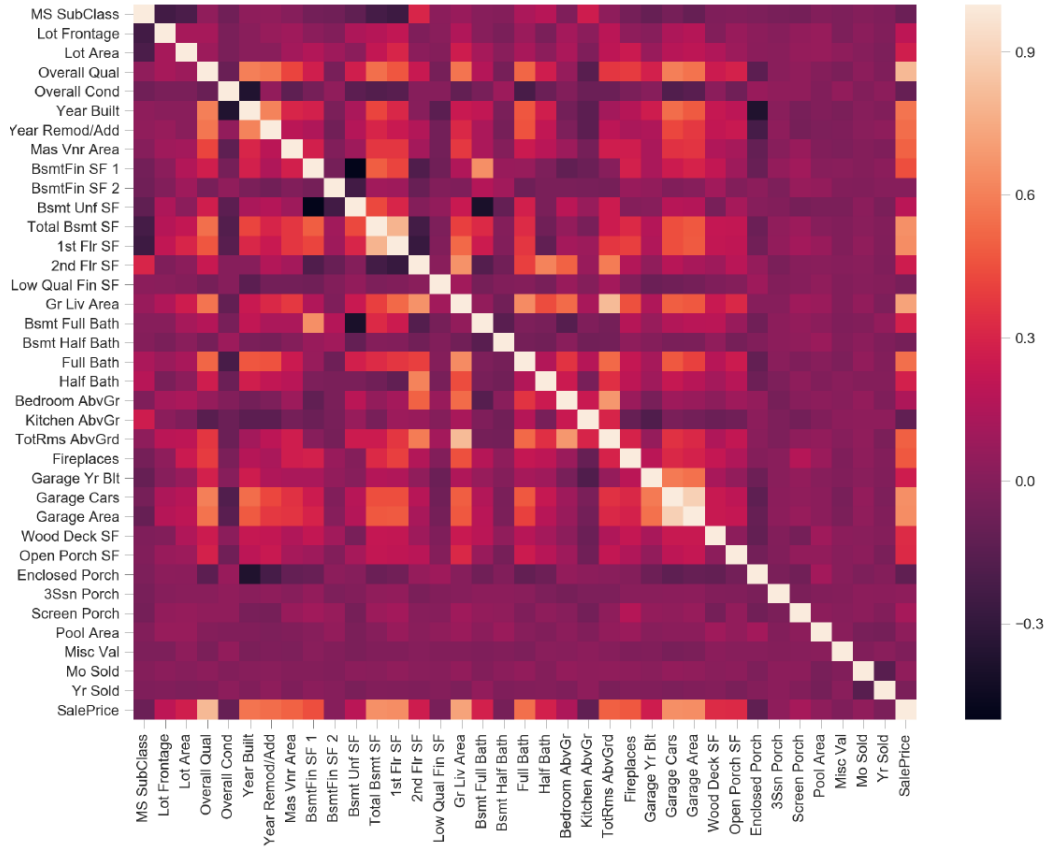
The histogram of the variable shows a more detailed view of the distribution:



Exploratory Data Analysis

Correlation Between Variables:
We want to see how the dataset variables are correlated with each other and how predictor variables are correlated with the target variable. For example, we would like to see how Lot Area and SalePrice are corelated.
Correlation is represented as a value between -1 and +1 where +1 denotes the highest positive correlation, -1 denotes the highest negative correlation, and 0 denotes that there is no correlation.

A heat map will be used to show the correlation:



Target variable is highly positively correlated with Overall Qual and Gr Liv Area. We see also that the target variable is positively correlated with Year Built, Year Remod/Add, Mas Vnr Area, Total Bsmt SF, 1st Flr SF, Full Bath, Garage Cars, and Garage Area.

Exploratory Data Analysis

The overall quality increases, the sale price increases too.
There is strong positive correlation between Gr Liv Area and SalePrice verifying what we found with the heatmap.