Ames Housing price prediction Using Machine Learning techniques

Motivation

The price of a house is dependent on various factors like size or area, how many bedrooms, location, the price of other houses, and many other factors. Real estate investors would like to find out the actual cost of the house in order to buy and sell real estate properties. They will lose money when they pay more than the current market cost of the house and when they sell for less than current market cost. The banks also want to find the current market price for the house, when they use someone's house as collateral for loans. Sometimes loan applicant overvalued their house to borrow the maximum loan from the bank. Banks and financial institutions also provide mortgage loan to home buyers. Local home buyers can also predict the price of the house to find out if a seller is asking for too much. The local seller can also predict their house price and find out how much is a fair market price.

Why am I using House price dataset:

1. This is a good project because it is so well understood.
2. Attributes are numeric and categorical so you have to figure out how to load and handle data.
3. It is a Regression problem, allowing you to practice with perhaps an easier type of supervised learning algorithm.
4. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.
5. Creative feature engineering .

Data Source

Kaggle competition: Competition Link

Feature Description

The data set (Kaggle Dataset) consists of two spreadsheets - 1. train.csv, containing data to train the prediction algorithm and 2. test.csv, containing data to test the prediction algorithm. The data fields in the train.csv are enumerated below

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior

- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet

- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

Data Cleaning and Wrangling

Since the dataset was loaded from Kaggle, it was fairly clean. Nevertheless, inconsistencies were found in the data that needed to be corrected.

First, the data was examined for missing or Null values using the Pandas info method. Where realized many column with missing values with significant number of missing value in column like Alley , Pool QC and MiscFeature.

Then tried to understand values with describe function where realized that all column have different range of distribution thus data will be needed to normalize (For algorithms which does not have pre installed normalizer) . After that understanding relation of dependant variable to get important independent variable(numerical) information, which were OverallQual , garage , area etc. Then data was divided into two parts one with numerical variable and other with categorical variables

While working with numerical missing columns , important column which were identified with high percent of missing values were lot frontage, GarageYrBlt and MasVnrArea as seen below(note: sale price missing values are basically test data without sale price value)

	Total	Percent
SalePrice	1459	49.982871
LotFrontage	486	16.649538
GarageYrBlt	159	5.447071
MasVnrArea	23	0.787941
BsmtHalfBath	2	0.068517
BsmtFullBath	2	0.068517

Here as Lotfrontage as about 17 % missing value , dropping wont be a bad decision as imputing lots of data can make prediction inaccurate. For column with misssing numerical value in basement , garage and masonry, firstly analysis were done of those missing rows with respect to other column with same part of house(i.e all garage columns were analysed together to understand missing row data) . after analysing it was concluded that MasVnrArea is missing when MasVnrType is none thus MasVnrArea missing data were imputed as 0 (Masonry veneer Is missing in

hose). similarly , for Garage Data it was realized that all values were missing when when GarageType is null and according to data source documentation all values with missing GarageType are house with no Garage thus missing values here will be imputed as 0 and when working with categorical data GarageType will be imputed as missing. finally basement has same case thus basement missing numerical variable will be replaced with 0 and categorical missing data with missing category .

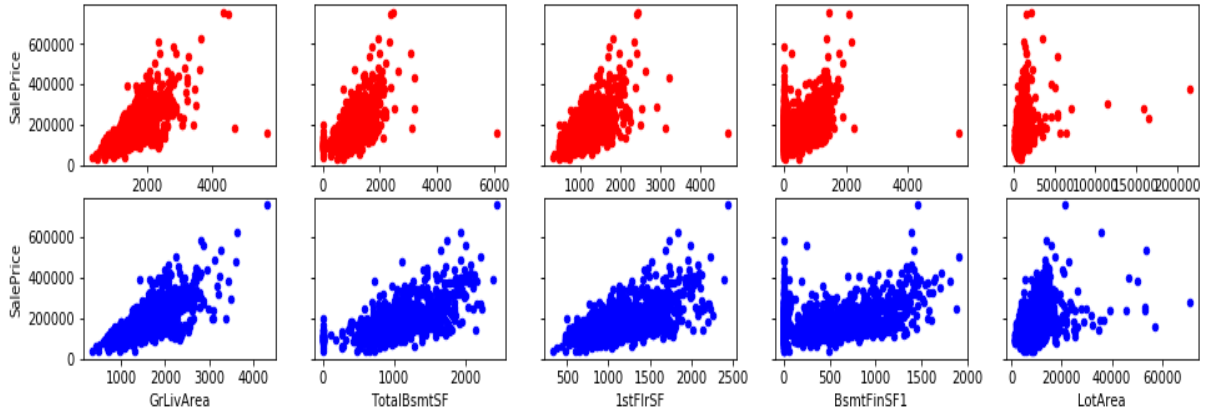
After analysising numerical missing data catagorical missing data were analysed , where certain columns related to basement , garage and masonry will be imputed based on anlysis done above. Later similar to numerical data missing column and their percent of missing were understood

	Total	Percent
PoolQC	2909	99.657417
MiscFeature	2814	96.402878
Alley	2721	93.216855
Fence	2348	80.438506
FireplaceQu	1420	48.646797
GarageCond	159	5.447071

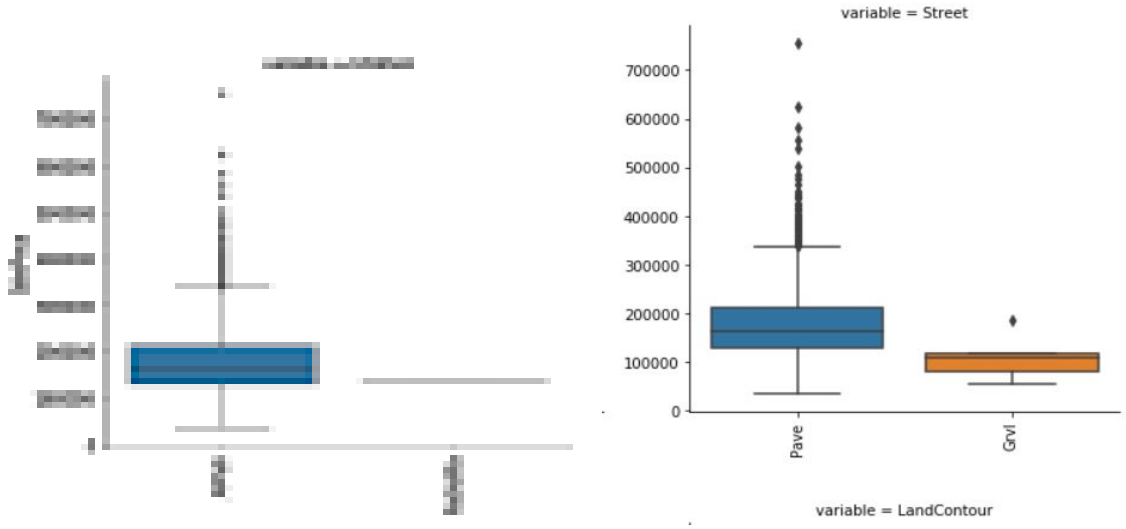
Based on above information columns like PoolQC , MiscFeature , Alley , Fence and FireplaceQu with most of values missing were dropped. After further analysis it was concluded that most houses do not have pool and fence thus it was decided that new variable signifying presence of pool or not will be created. Later other columns with missing data were analysed and it was identified that Functional column with missing data will be replaced with Typ as per document, similarly for fireplace it will be missing fireplace. Finally remaining columns with missing values were with 1 or 2 rows of missing data thus mode of respective column will be used to replace those column .

outlier detection

Outlier detection were also performed with similar procedure as missing value imputation which is separating numerical and categorical data with difference being only training data will be used here over full data. For numerical variable outlier , all variabele were plotted against sale price to find outlier. GrLiving area , LotArea,TotalBsmtSF,1stFlrSF,BsmtFinSF1 columns were columns which were further analysed as seen below . where 1 st row of chart represent original data and 2nd row of column represent data after removing outlier



For Categorical variable outlier , box plot were used to identify outliers . where categorical distribution with respect to sale price were plotted



As seen above it was concluded Utilities and Street does not deliver any information as most of house have same utilities and street type

Data story and EDA

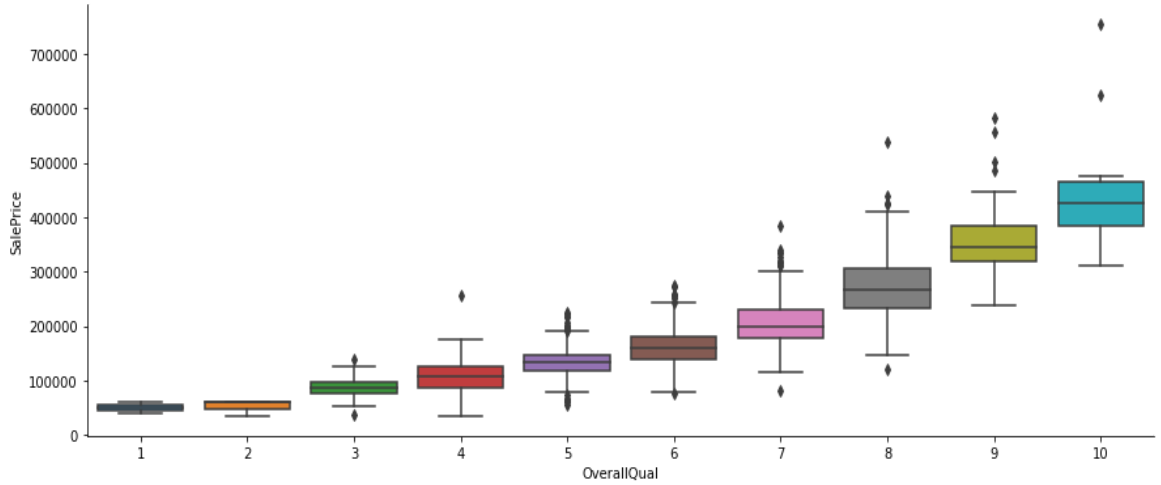
Data storya nd EDA were performed in mainly 3 parts firstly data were looked as different section of house , secondly data distribution understand performed and finally some statistical test were performed on available data

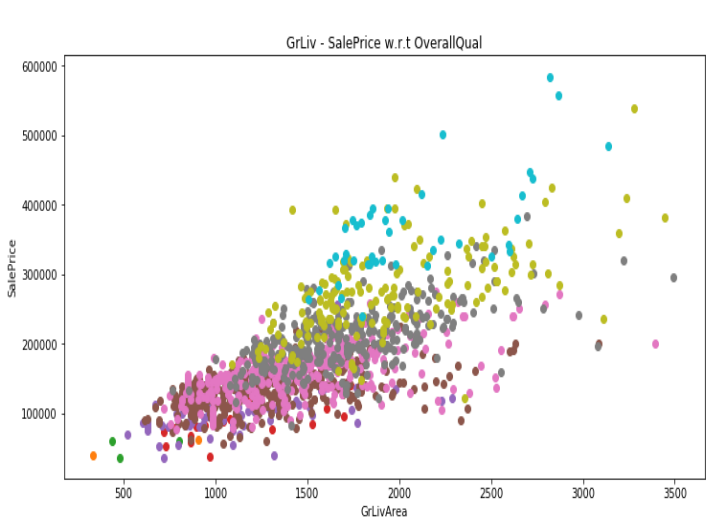
House Section Understanding

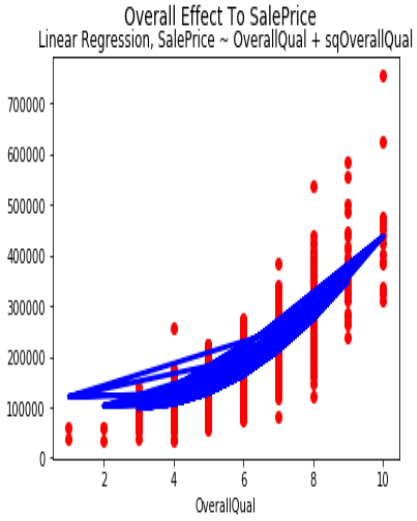
This section firstly most important attribute overall quality and living area were analysed , later basement and 2nd floor were analysed , further analysis based on bathroom , technical rooms , outside area , season effect and finally garage effect were visualized

Overall Quality and Living area

First overall quality was plotted with respect to sale price , later living area was plotted with respect to sale price and finally linear relationship were analyzed





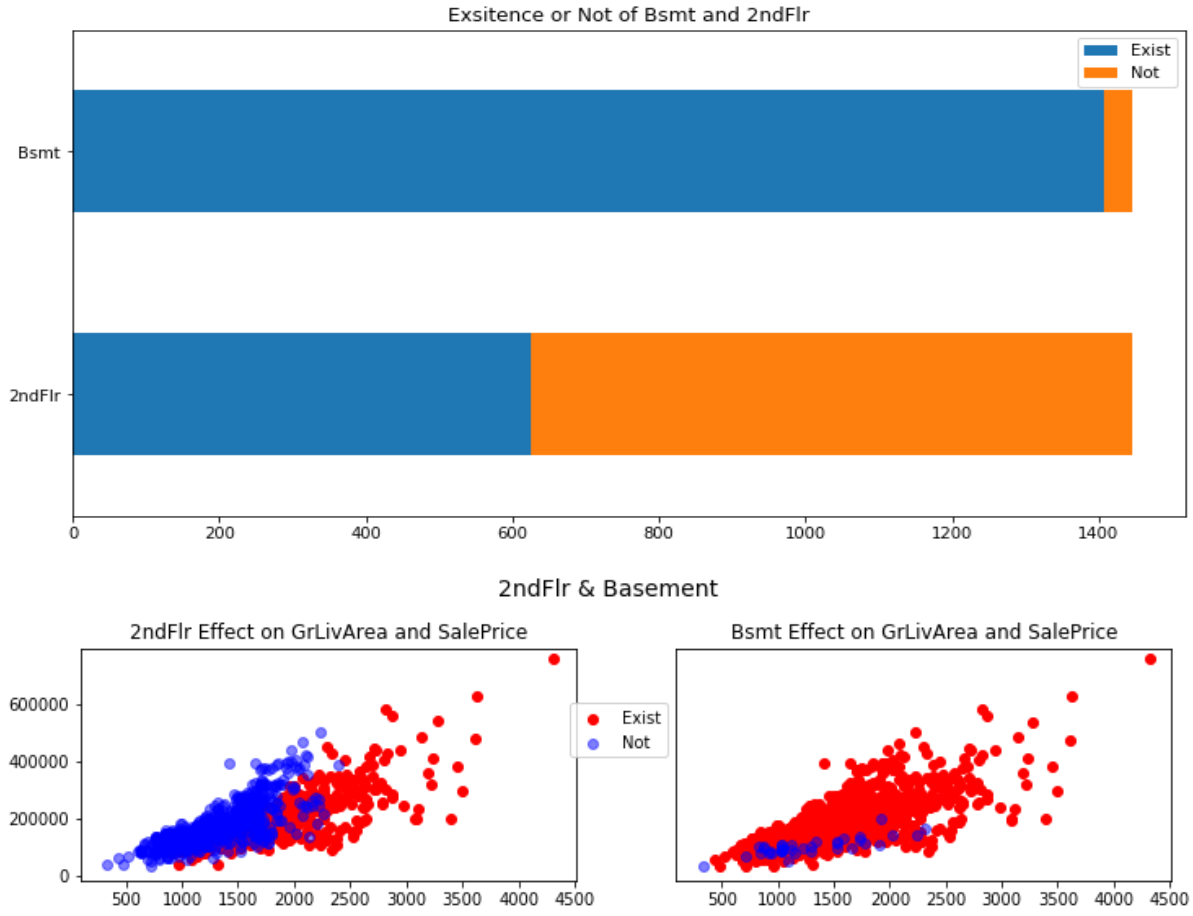


From the above figures it was conclude that

- OverallQual is the very good variables Variables regard of explaining SalePrice
- OverallQual causes different SalePrice where having same "GrLivArea". We have to put a strong attention on OverallQual
- OverallQual was proportional to SalePrice
- Even though Living area does look to have liner realtionship with sale price overall quality matter more

Basement and Upper floor

Here firstly existence of basement and 2nd floor were plotted later affect of basement and 2nd floor with respect to living area to sale price were plotted

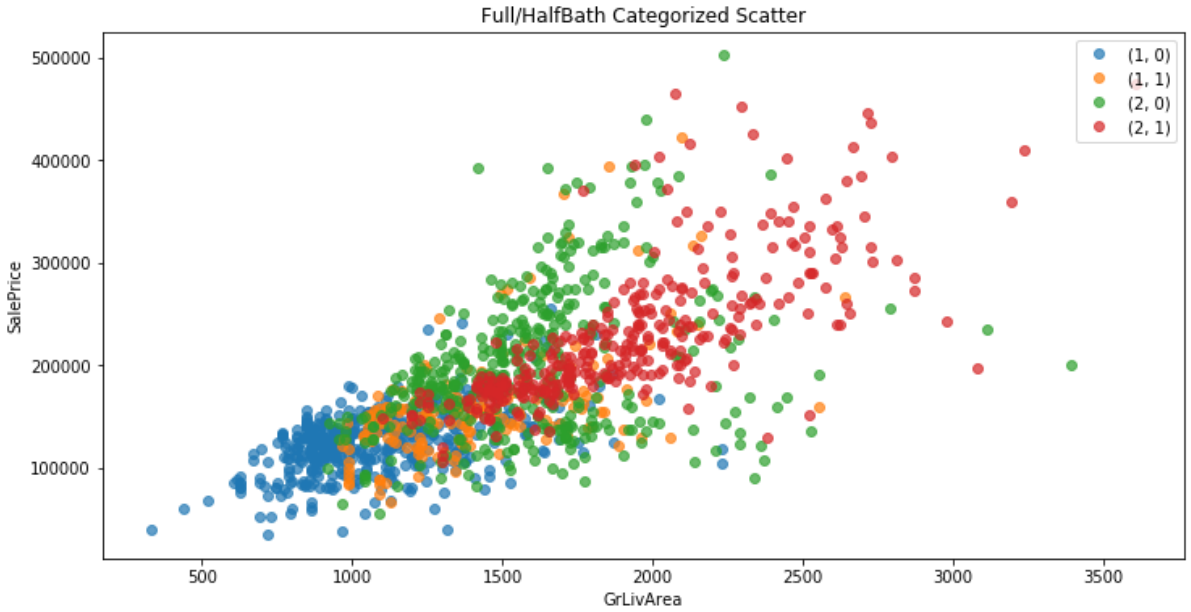


From the above figures it was conclude that

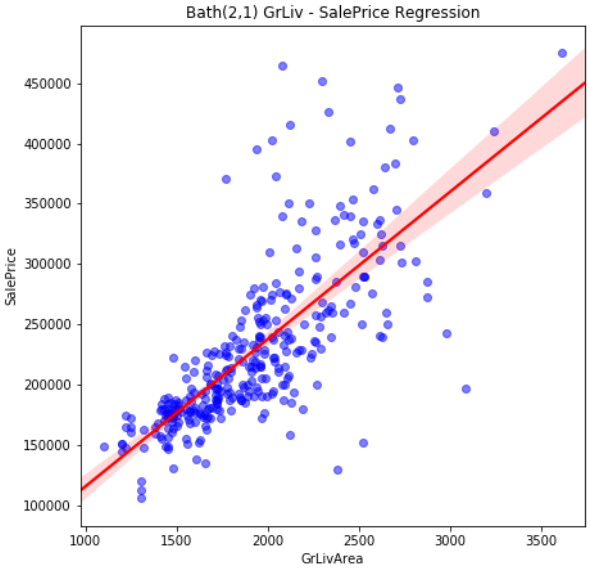
- 2ndFlrSF depressed the power of GrLivArea toward SalePrice
- Bsmt has nothing related to the price

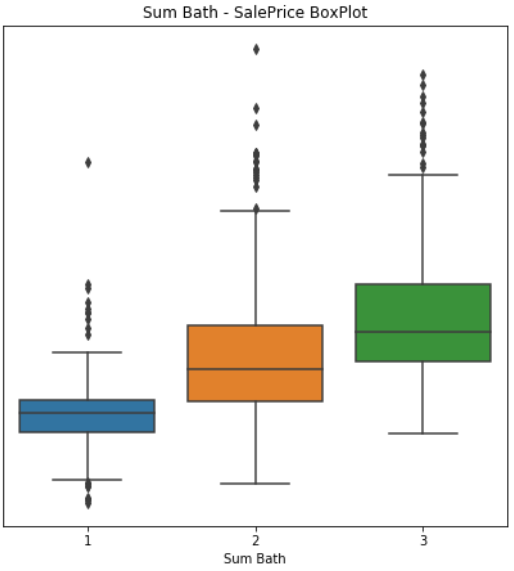
Bathroom

Here no of full and half bathroom relation with sale price and living area were analysed, since most high price house have 2 full and 1 half bathroom its linearity was plotted and finally sale price and bathroom relation was plotted with box plot



Bath Relationship with SalePrice



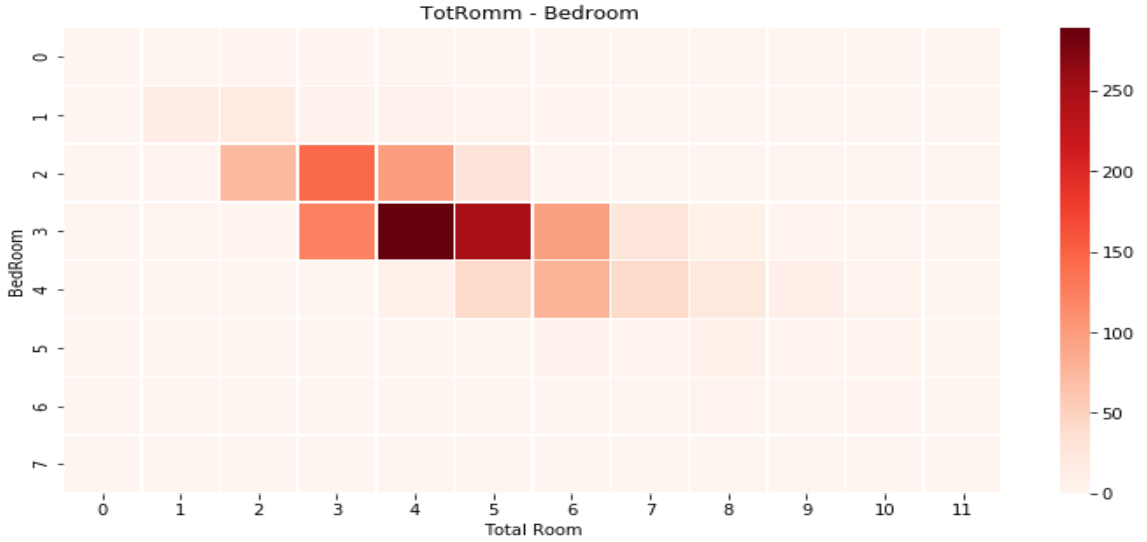


From the above figures it was conclude that

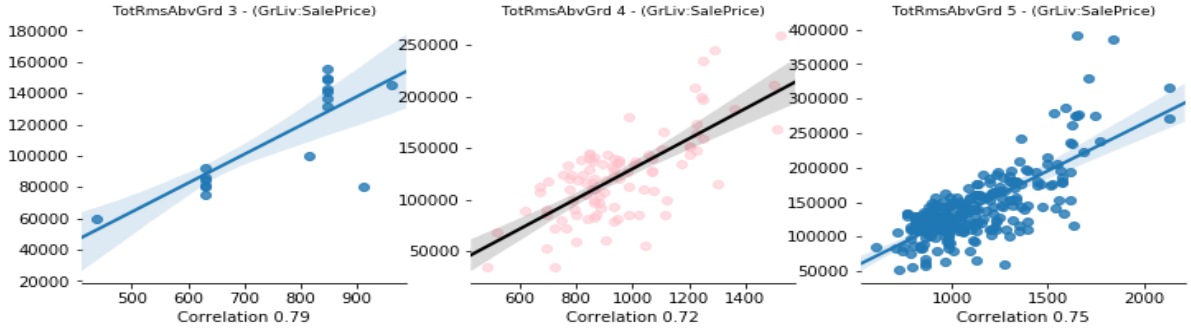
- The Number of Bath usually increased the SalePrice
- combination of (Full 2, Half1) improved the linearity and decreased the spreadness of SalePrice - GrLivArea.

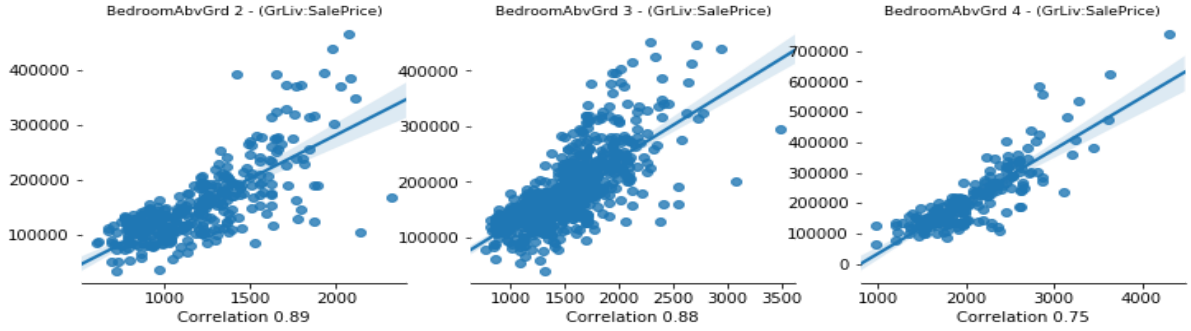
Technical Room

Here no of bed room and total room relation was plotted then all important combination of bed room relationship with sale price were plotted



Room's Usage with SalePrice



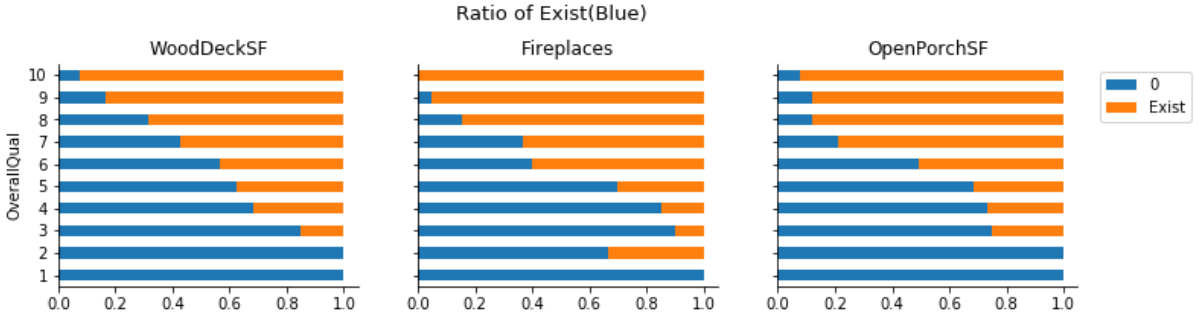


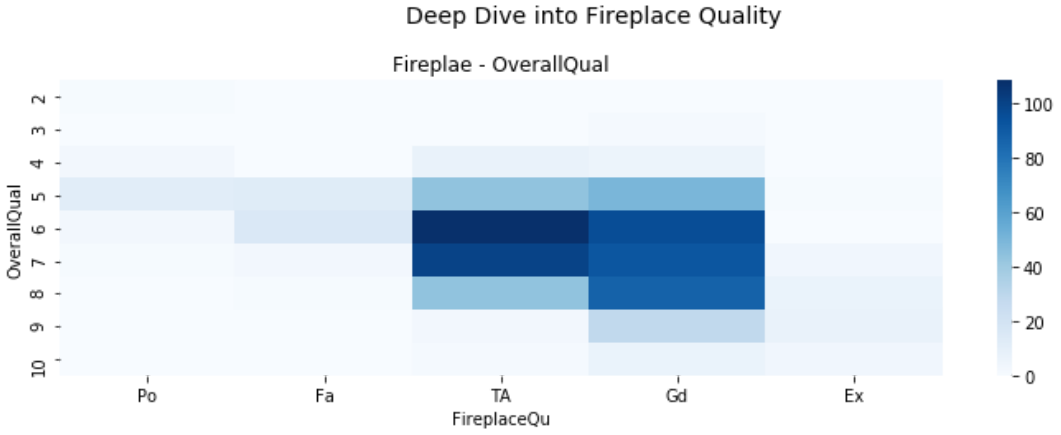
From the above figures it was conclude that

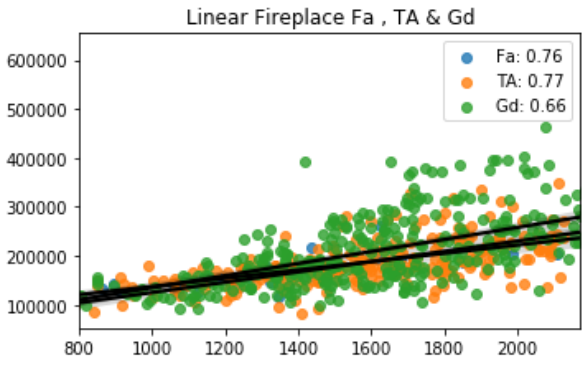
- Total room above ground and no of bedroom are linearly related
- total room above ground also have very good correlation above 0.7
- bedroom has very high correlation with sale price and living area(usually above 0.75)

Outside area

Here existence of wood deck , fireplace and open porch were plotted and since fireplace is important attribute its relation with sale price was analysed







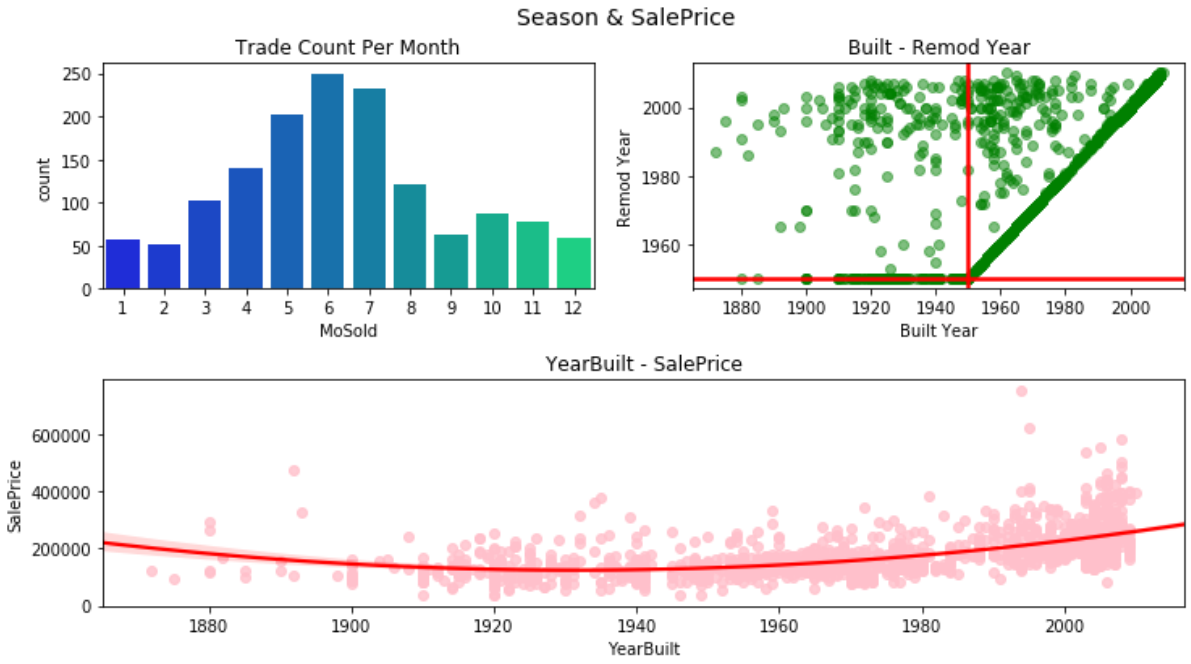
From the above figures it was conclude that

- Good Quality House has more outside instrumental places.
- PoolArea, ScreenPorch, 3SsnPorch were almost negligible thus not plotted

- Fireplace is linear related to overall quality
- only Fa,TA and Gd has some decent linear relationship
- Ex and Po does not have good linear relationship, which is clearly visible by looking where Ex are so spread around

Effect of Season

Here month sold year built are plotted with respect to sale price

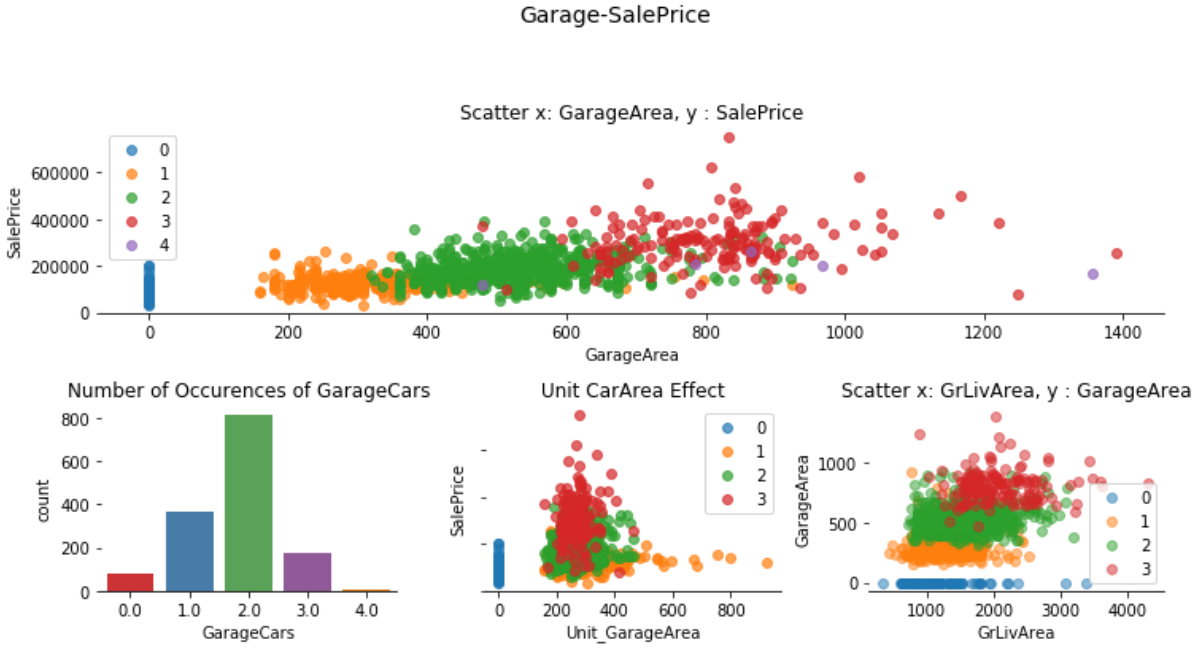


From the above figures it was conclude that

- The amount of trade was increased by rising temperature(Less trend in winter)
- The part of house, built after 1950, was not remodeled yet
- YearBuilt^2 will be proper if the variables is used to predict

Garage

Garage quality , no of car space and area were analysed with respect to sale price



From the above figures it was conclude that

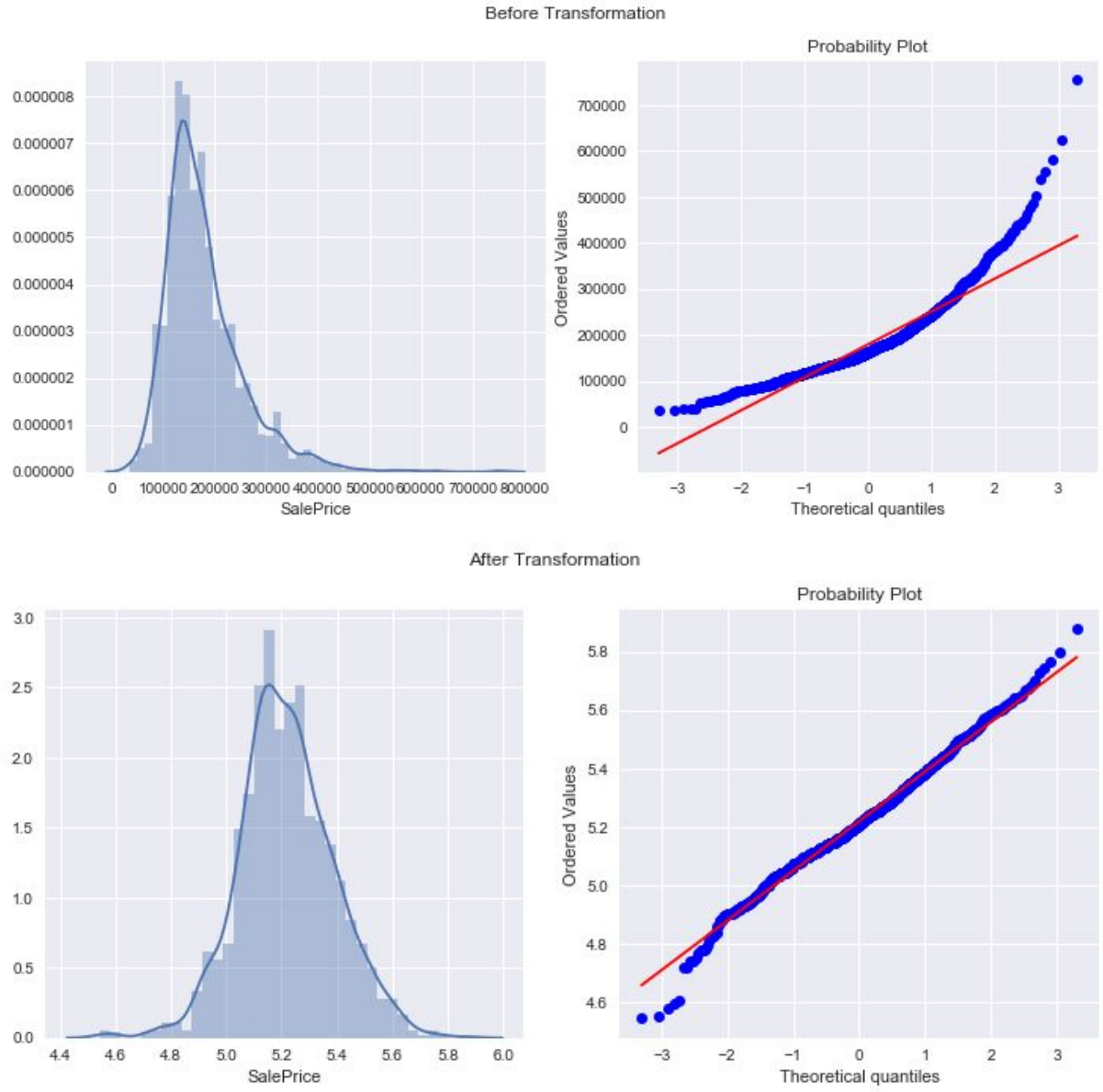
- Most of houses have two cars
- GarageArea makes Chunk having small linearity with SalePrice
- 0 Cars and 1 Cars has no difference in SalePrice
- 4 Cars are similar with 3 Cars house. Merge them
- (Update) Unt_Garage Area said that "Expensive house sustain the proper line of the area!"
- (Update) GrLivArea is a good variable not related to GarageArea. Those two variables enforces the prediction power.

Variable Distribution

Here we will understand distribution of data, understanding of distribution is very important since many of algorithm does assume normal or some distribution and having data in proper form always help algorithm to be efficient ,thus firstly dependant variable was analyzed and later understanding of distribution of independent variable was carried out.Also after looking at distribution we also looked at possible transformation and its effect on distribution

Sale Price

Here we have plotted distribution of sale price from given training data



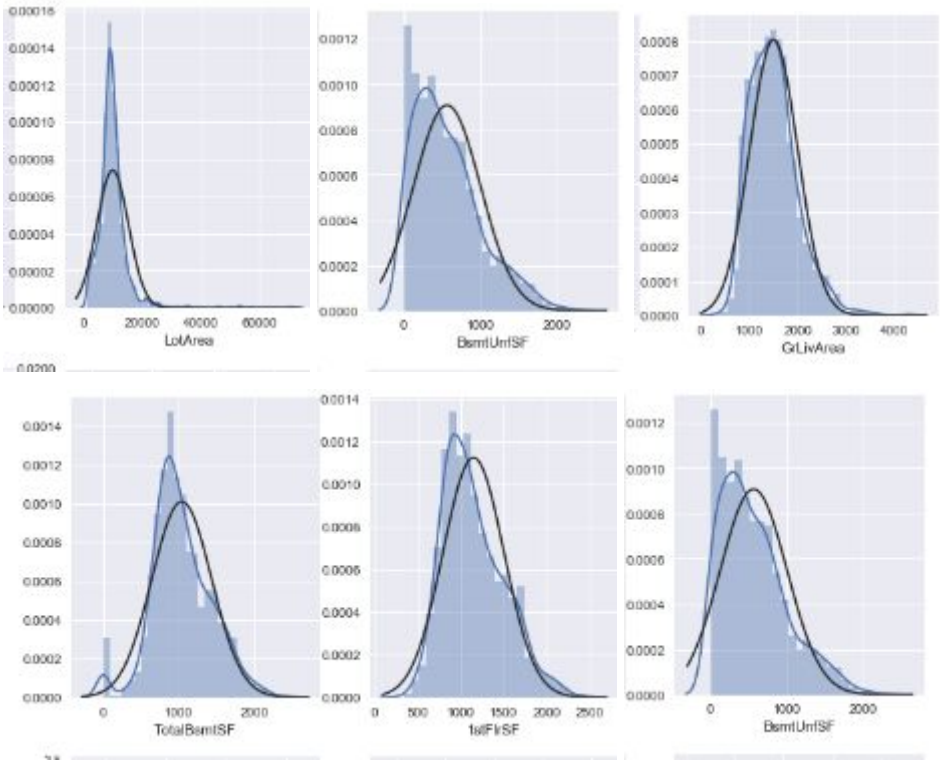
From the above figures it was conclude that

- sale price is not not normally distributed(Right Skewed)
- After applying log transform to sale price data seems to be normally distributed

Independent variables

Here we plotted all variable with normal distrubution curve

After looking at all column it was realized that many variable are normally distributed and variable like LotArea, BsmtUnfSF, 2ndFlrSF, MasVnrArea, BsmtFinSF1 , TotalBsmtSF, 1stFlrSF ,GrLivArea ,TotalBsmtSF are nearly normally distributed

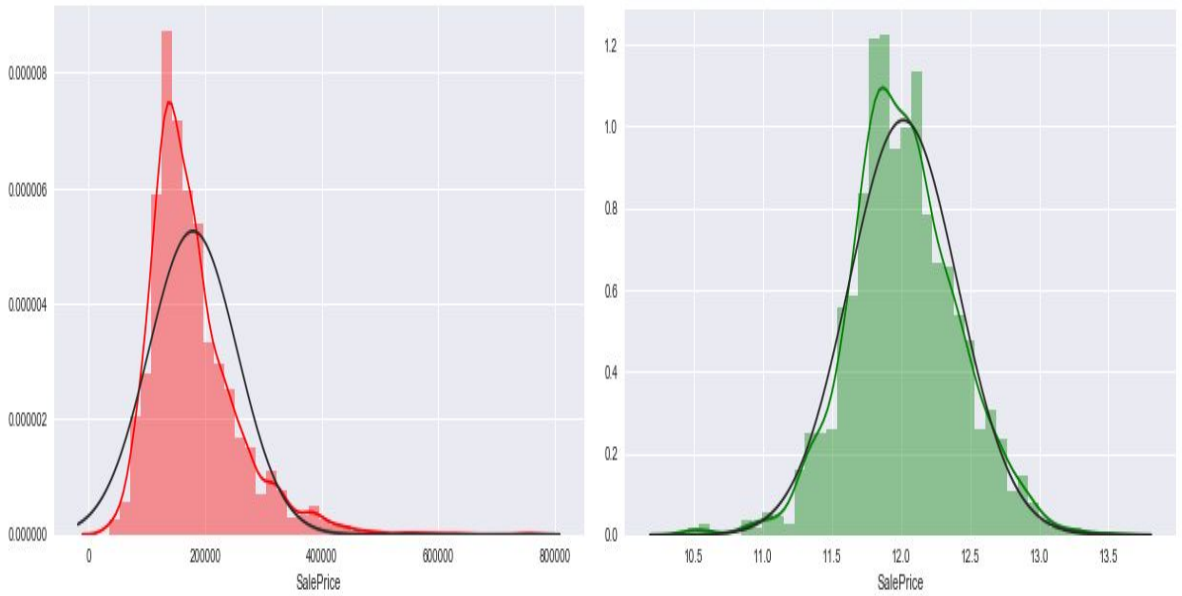


Skewness and Kurtosis

Skewness measures the distortion of the bell shape of the distribution. Positive values indicate that the distribution tail is extended to the right, vice versa if negative. From -0.5 to 0.5: fairly symmetrical curve. From -1 to -0.5 or +0.5 to +1.0 moderate asymmetry. Beyond this range: highly asymmetric. Kurtosis measures the extreme tails of distribution; it does not concern the shape of the figure. A high value indicates that the distribution has many extreme values (probably anomalous)

SalePrice

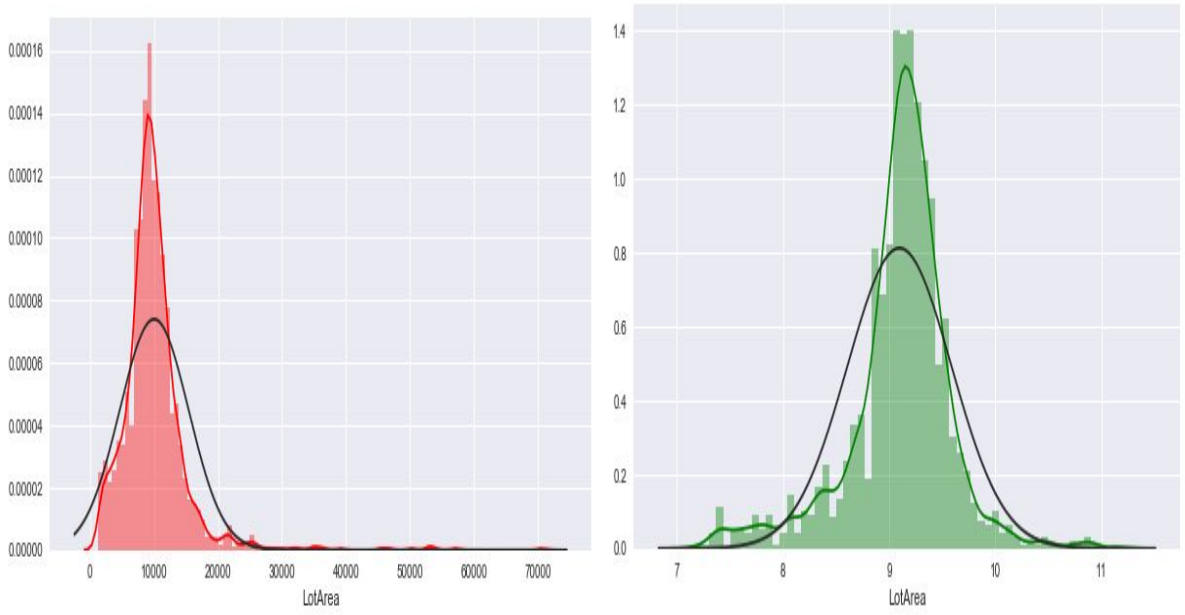
Here SalePrice Skewness and Kurtosis were analyzed and their effect after transformation of variables



Here after plotting distribution it was clear that sale price has a tail on the right thus Log transformation was applied, After applying transformation Skewness of Saleprice:1.6773 was reduced to 0.0608 similarly kurtosis: from 5.2079 to 0.7350 thus it was clear that after applying Log Transform near to normal distribution was obtained

LotArea

Here LotArea Skewness and Kurtosis were analyzed and their effect after transformation of variable

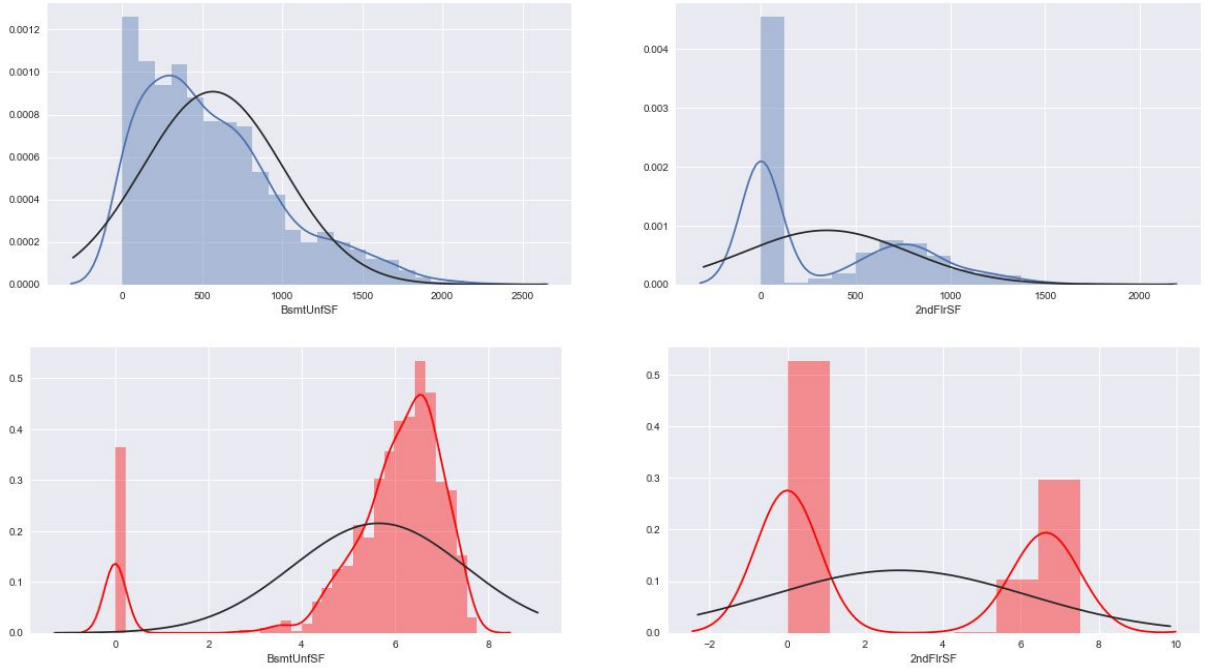


Here after plotting distribution it was clear that LotArea has a tail on the right thus Log transformation was applied, After applying transformation Skewness of LotArea :

3.9759 was reduced to -0.7238 similarly kurtosis: from 29.7375 to 2.8932 thus it was clear that after applying Log Transform near to normal distribution was obtained

BsmtUnfSF and 2ndFlrSF (Unfinished area of basement / Second floor area)

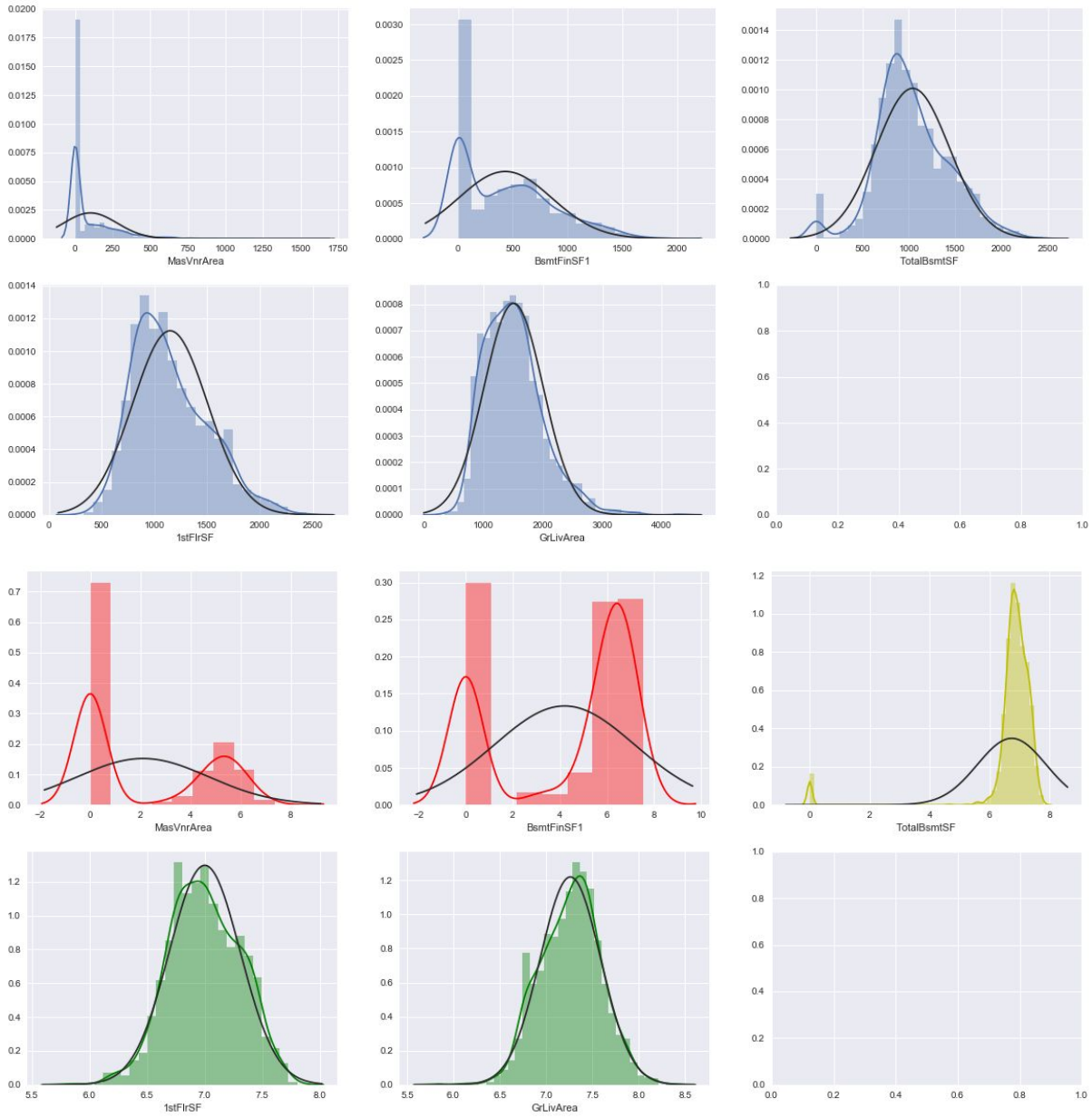
Here BsmtUnfSF and 2ndFlrSF Skewness and kurtosis were analyzed and their effect after transformation of variables



Here after plotting distribution it was clear that BsmtUnfSF and 2ndFlrSF has a tail on the right thus Log transformation was applied. Due to which Skewness of BsmtUnfSF and 2ndFlrSF : 0.918856 and 0.781220 was reduced to -2.188213 and 0.284881 similarly kurtosis: from 0.480840 and 0.0271 to 4.137715 and -1.9024 here for BsmtUnfSF even though skewness went out of range from ideal normal distribution range it was clear that it has better distribution was obtained as single line of zero were creating increase in skewness but overall distribution is much much better, similarly for 2ndFlrSF values were little out of range because of 0 but distribution looked better while looking at graph .

Other columns

Here MasVnrArea, BsmtFinSF1,TotalBsmtSF,1stFlrSF,GrLivArea Skewness and Kurtosis were analyzed and their effect after transformation of variables



Here after plotting distribution it was clear that GrLivArea and 1stFlrSF has a tail on the right thus Log transformation was applied. Due to which Skewness of GrLivArea and 1stFlrSF : 0.922740 and 0.662811 was reduced to-0.097386 and -0.06940 after similarly kurtosis: from 1.563783 and 0.027181 to 0.063599 and -0.22448 thereby near to normal distribution was obtained

For TotalBsmtSF distribution was quite near to normal even though graph had shown skewness and kurtosis line bit far. But understanding data it was clear that it is due to zeros presence and finally for MasVnrArea and BsmtFinSF1 skewness and kurtosis values are not good but at the same time distribution is much better ignoring zeros thus log transformation of these columns were performed. Also to support those zero value distribution new column are created to represent presence of those position of house thereby reducing effect of zeros during algorithm process . finally it

is important to note that scaling of variable still need to be performed as there are many variable with very different range are still present

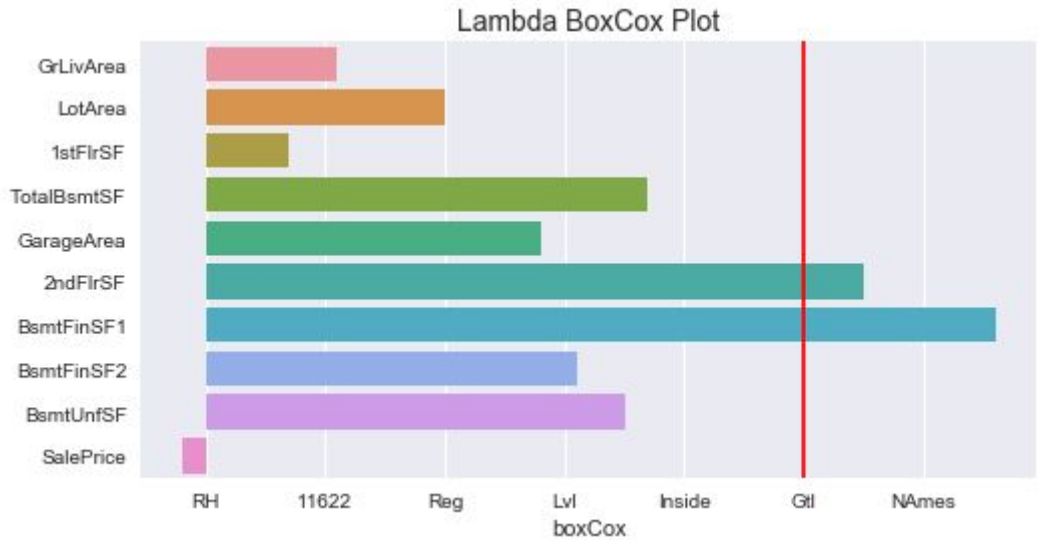
Categorical distribution
After observing Categorical distribution following points were noted:

- Some categories seem to more diverse with respect to SalePrice than others.
- Neighborhood has big impact on house prices.
- Most expensive seems to be Partial SaleCondition.
- Having pool on property seems to improve price substantially.
- There are also differences in variabilities between category values.

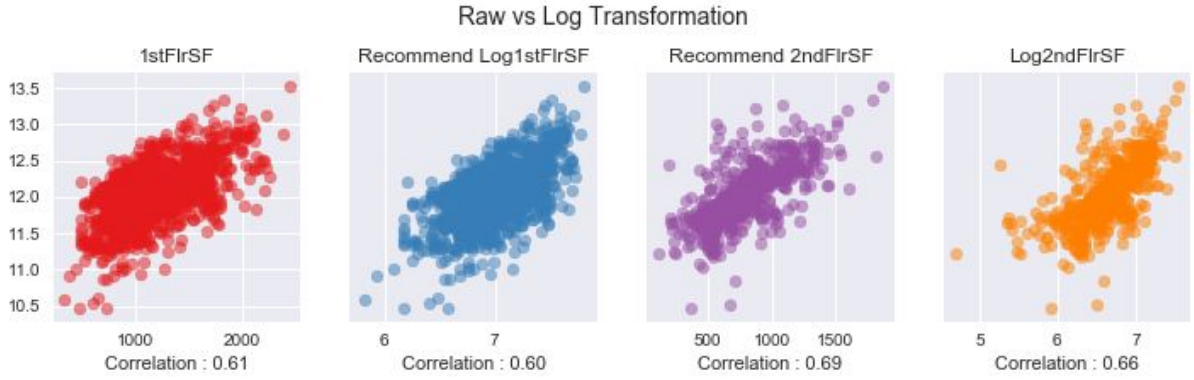
Statistics

Boxcox

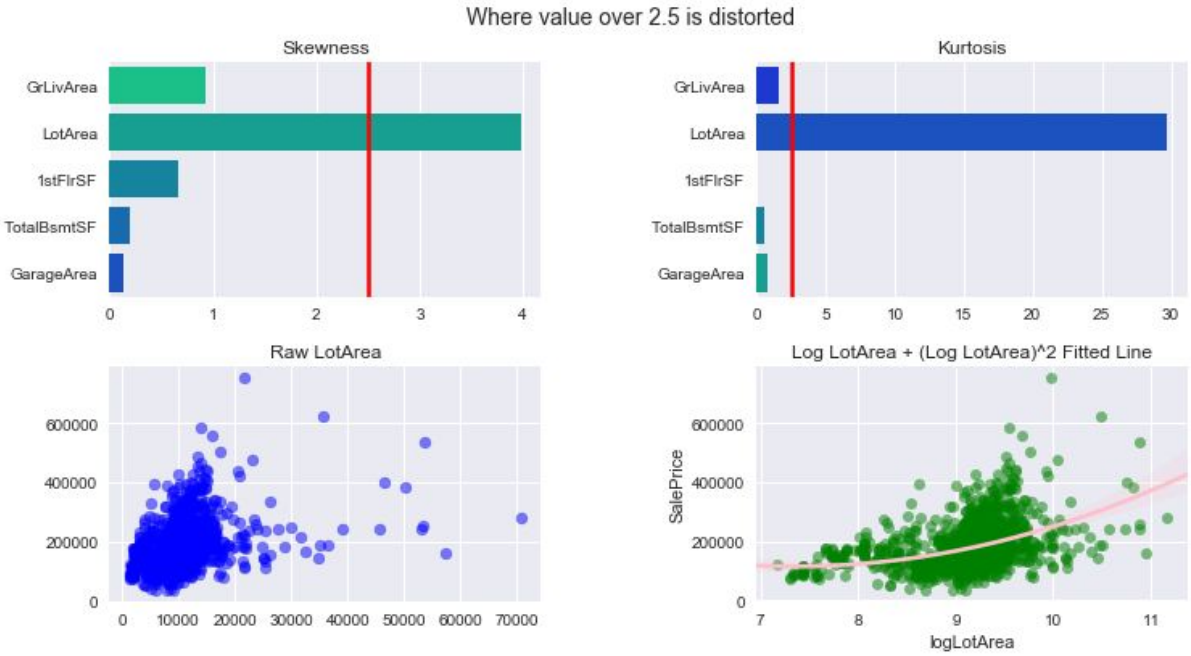
calculate the proper lambda, which inverse or power.
For example, if one variables have round(lambda) = 0, Log transform needed. And in round(lambda) = 2 case, the variable changed as variable^2. (It is useful in the range(-5,5). But some material said (-2,2) transformation is enough)



Except 2ndFlrSF, BsmtFinSF1, the other variables need to deal with by Log Transformation



Recommend Log transformation of 1stFlrSF is worse than the prior. But it has an advantage to decrease the spreadness in the large 1stFlrSF area. Even if Log2ndFlrSF is not recommend, the distribution has merit regard of a large 2ndFlrSF value. Both of them has merit.

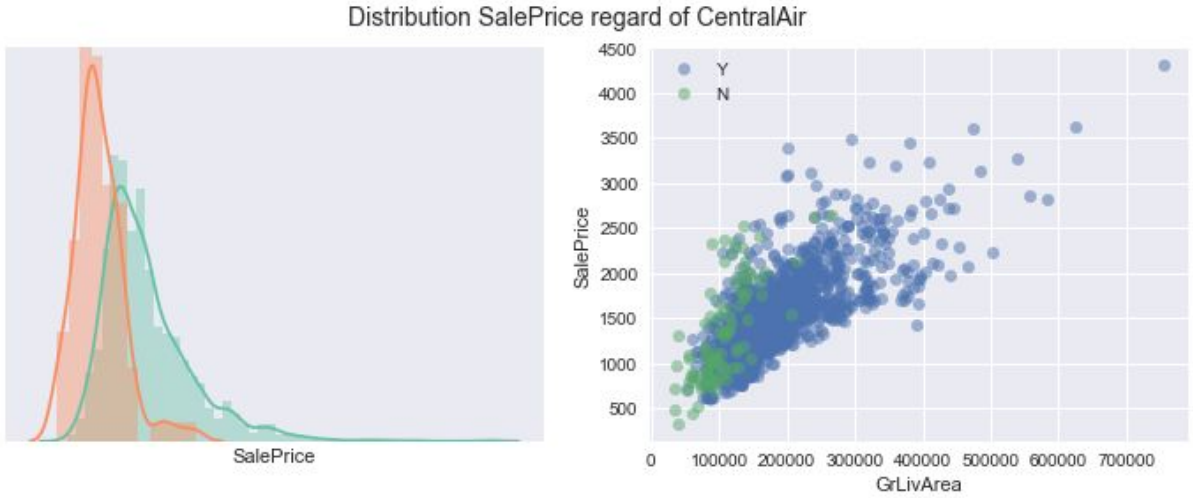


From above analysis it can be said that:

- The meaning of high kurtosis value is most of value gathered in just one part
- LowArea was densed in low value Part
- Just Log_LotArea was not good variables, but with (Log_LotArea)^2, the fitted line w.r.t SalePrice was better

Nonparameteric Test /Wilxoc-rank Sum test

Wilxoc-rank Sum test calculate the p-value on the hypothesis, two distribution has same distribution. So if the p-value < p-criterion, then two distribution is useful to predict.



P_value : 0.00 SalePrice according to CentralAir was changed
Thus CentralAir can be used to predict

Kruskal Will Test

Kruskal Will Test calculate the p-value that the null hypothesis does not indicate which of the groups differs. So if the p-value < p-criterion, then the variable was useful to predict.

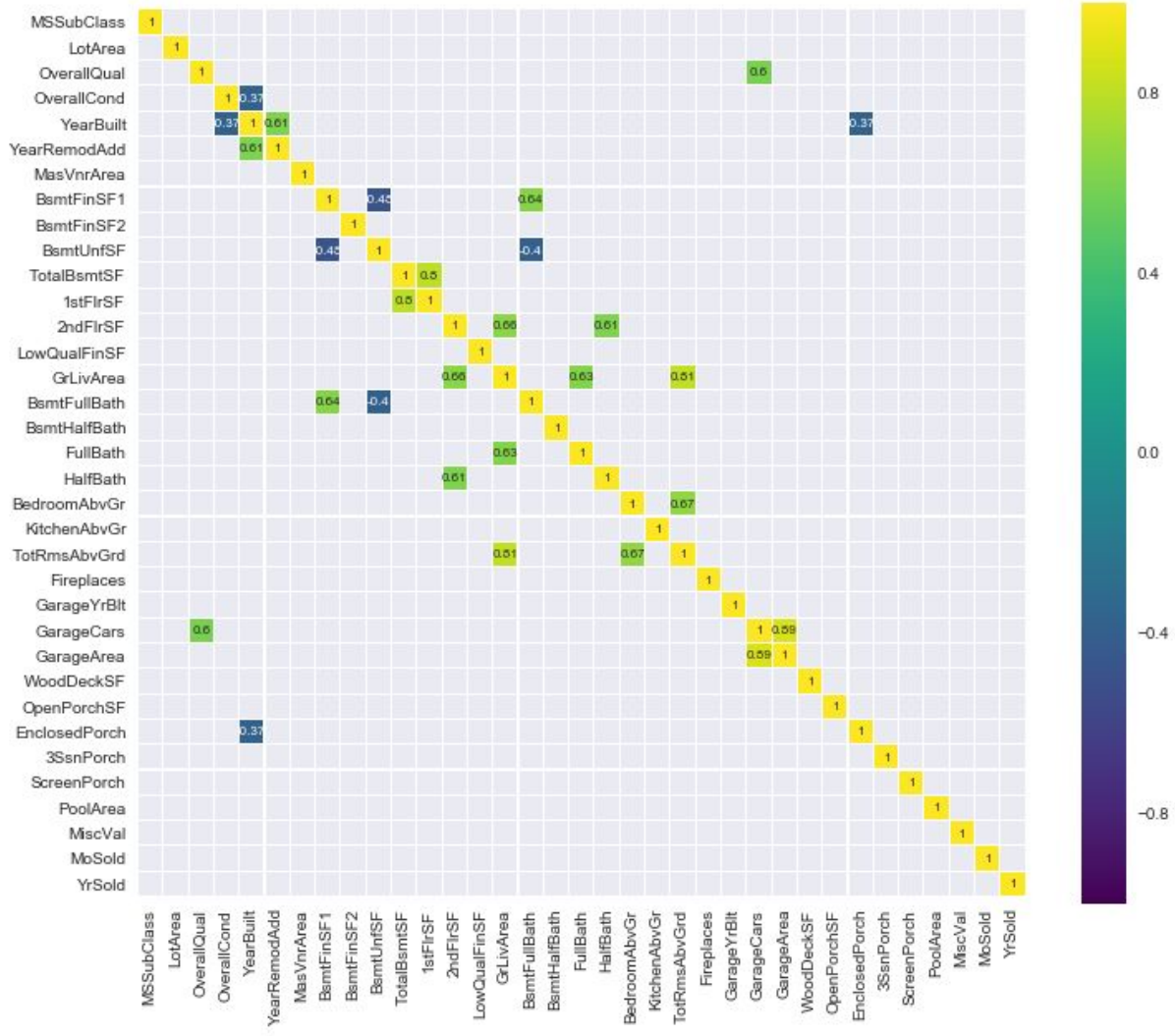
The Kruskal–Wallis test by ranks, Kruskal–Wallis H test (named after William Kruskal and W. Allen Wallis), or one-way ANOVA on ranks is a non-parametric method for testing whether samples originate from the same distribution

P-value 0.00 According to Fireplaces, SalePrice distribution changed
P-value 0.88 According to BsmtHalfBath, SalePrice distribution didn't changed

- Fireplaces is good at predicting SalePrice, but Bsmt is not as good
- Most of Variable are useful, except MoSold, PoolQC, PoolArea, LandSlope, LotConfig, BmstHalfBath, YrSold

Correlation

It is clear that important factors to predict Sale Price are quality,Living Area,Garage and bathrooms and year renovated, here correlation between dependant variable are analysed



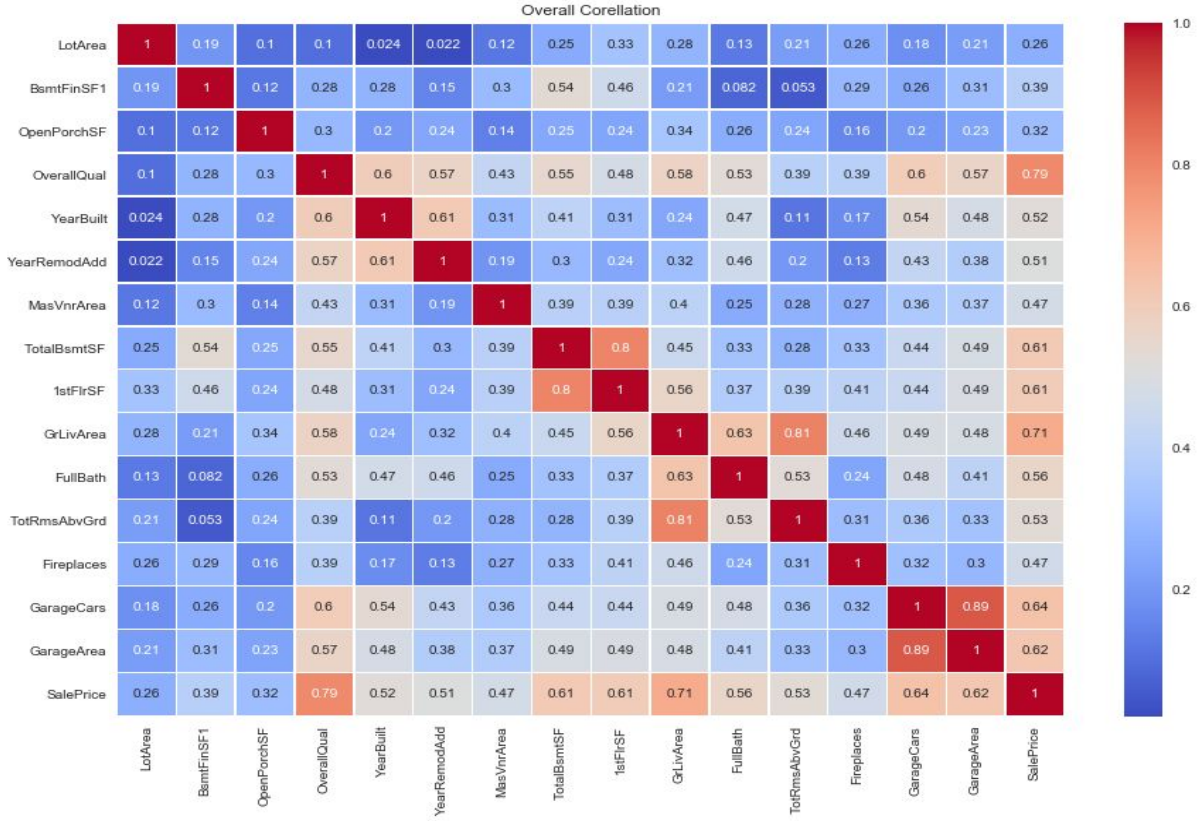
From above figure it following information were revealed:

- A lot of features seems to be correlated between each other but some of them such as YearBuild/GarageYrBlt may just indicate a price inflation over the years.
- As for 1stFlrSF/TotalBsmtSF, it is normal that the more the 1st floor is large (considering many houses have only 1 floor), the more the total basement will be large.
- Now for the ones which are less obvious we can see that:

- There is a strong negative correlation between BsmtUnfSF (Unfinished square feet of basement area) and BsmtFinSF2 (Type 2 finished square feet). There is a definition of unfinished square feet here but as for a house of "Type 2", I can't tell what it really is.
- similar to basement and garage .. above ground variable have relation
- living area is related to all area related variabele and no of rooms/bathrooms
- year built vs area
- year built vs condition
- HalfBath/2ndFlrSF is interesting and may indicate that people gives an importance of not having to rush downstairs in case of urgently having to go to the bathroom (I'll consider that when I'll buy myself a house uh...)
- There is of course a lot more to discover but I can't really explain the rest of the features except the most obvious ones.
- It can be said that, by essence, some of those features may be combined between each other in order to reduce the number of features (1stFlrSF & TotalBsmtSF, GarageCars & GarageArea) and others indicates that people expect multiples features to be packaged together.
- might get more information after one hot encoding

Plot important feature correlation matrix

Below figure indicate correlation value between some important variable



Feature Engineering

Here mainly three type of feature engineering were implemented

Feature created based combining many features

Here by creating new column which can deliver more information
Example being 'YearBuiltemodel' which is created from coluumn called 'YearBuilt' and 'YearRemodAdd' . which can be clearly seen in table below which show co-relation between sale price and existing as well as new feature

Correlation	YearBuilt	YearRemodAdd	YearBuiltemodel
SalePrice	0.534168	0.521317	0.590277

Similarly other column like TotalSF which is basically total are of house can be a wonderful feature to have

Correlation	TotalBsmtSF	1stFlrSF	2ndFlrSF	'TotalSF'
SalePrice	0.632441	0.613275	0.333395	0.826080

With the same approach as other features were created called
'Total_sqr_footage' which as sum of surface area of house surface area
'Total_Bathrooms' which is sum of total bathroom where half bathroom is multiplied with 0.5 and 'Total_porch_sf' total Porch area

Feature created from existing feature

These are based on analysis done previously where feature which representing presence of a particular part of house(i.e do the exist or not). Thus here features like 'haspool' , 'has2ndfloor' , 'hasgarage' ,'hasbsmt' , 'hasfireplace' are created when these feature have any area assigned to it

Feature created from transforming feature

SalePrice,LotArea,BsmtUnfSF,MasVnrArea,TotalBsmtSF,1stFlrSF,GrLivArea are all log transformed