Milestone Report

1. Problem statement:

This project is to predict the housing prices in the given area considering various elements like whether the house contains car garage, swimming pool and how many bedrooms it contains and what is the dimensions of the building etc. The goal of this project is to create a regression model that are able to accurately estimate the price of the house given the features.

2. Data Wrangling:

This section describes the various data cleaning and data wrangling methods applied on the **House Price Prediction** to make it more suitable for further analysis. The following sections are divided based on the procedures followed.

Cleaning:

We have 18 categorical features with missing values. And 11 numerical features with missing values.

So, 18 categorical features and 10 numerical features to clean.

- We start with the numerical features, first thing to do is have a look at them to learn more about their distribution and decide how to clean them:
- Most of the features are going to be filled with mean values, because we assume that they don't exist, for example GarageArea, GarageCars with missing values are simply because the house lacks a garage.
- GarageYrBlt: Year garage was built can't be filled with 0s, so we fill with the median (1980).

And we have 18 Categorical features with missing values:

- Some features have just 1 or 2 missing values, so we will just use the forward fill method because they are obviously values that can't be filled with 'Missing'
- Features with many missing values are mostly basement and garage related (same as in numerical features) so as we did with numerical features (filling them with 0s), we will fill the categorical missing values with "Missing" assuming that the houses lack basements and garages.

Removing Unnecessary Features

This process was done in different ways

Number of missing values:

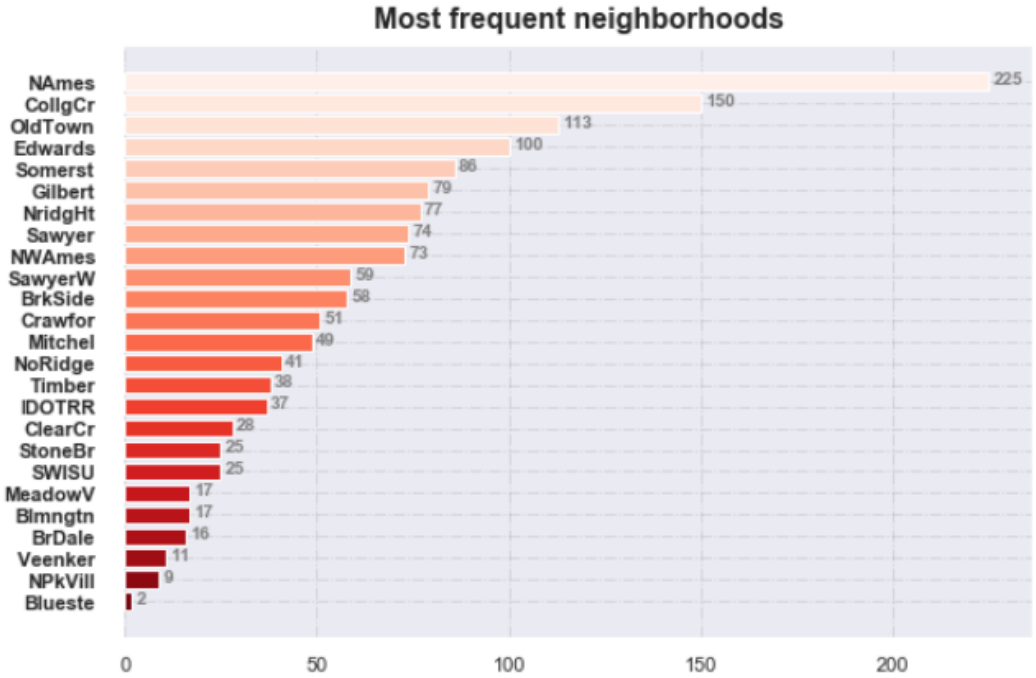
First thing to do is get rid of the features with more than 80% missing values. For example the PoolQC's missing values are probably due to the lack of pools in some buildings, which is very logical. But replacing those (more than 80%) missing values with "no pool" will leave us with a feature with low variance, and low variance features are uninformative for machine learning models. So we drop the features with more than 80% missing values.

Significance:

From co-relation we get to know features that are important for predicting the output and after performing co-relation features which has values greater than 0.7 are carried forward and remaining were dropped.

3. Data Story Telling:

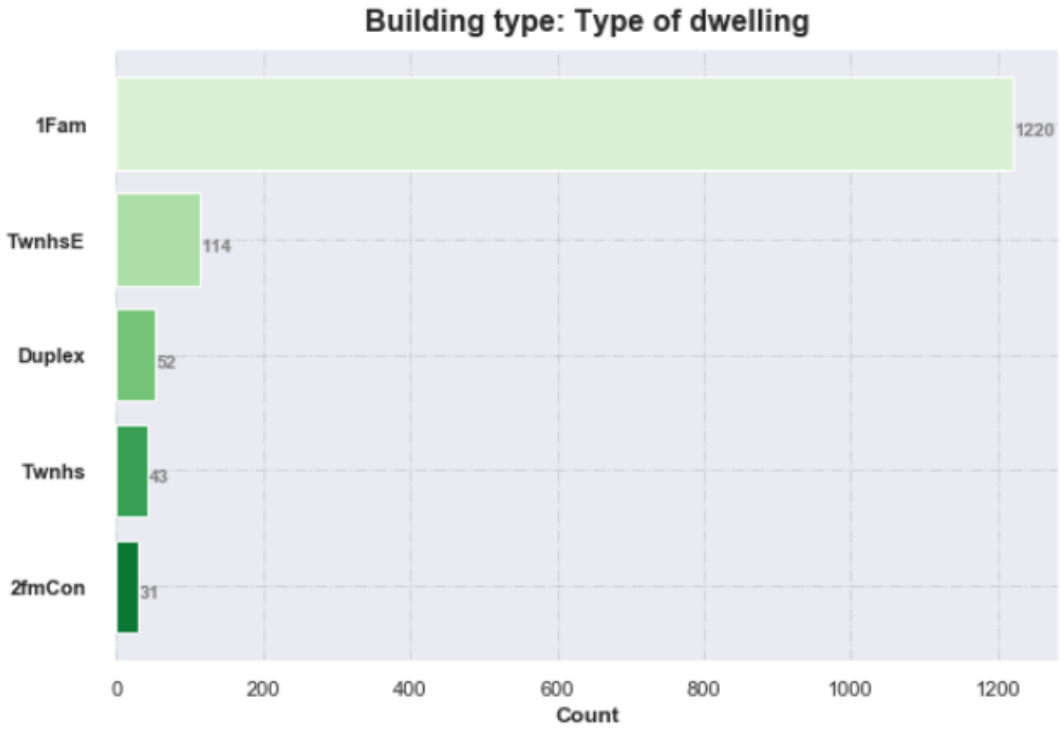
Most Frequent Neighbourhoods:



From the above graph, we noticed that most frequent neighbourhoods are Names with 225, CollgCr with 150, followed by old town, etc. and less frequent neighbourhoods are Blueste.

Type of Dwelling:

In the below bar plot, we noticed that we have 5 different types of dwellings are there. In that Most of the dwelling are of '1Fam' type. And '2fmcon' type of dwellings are very few.



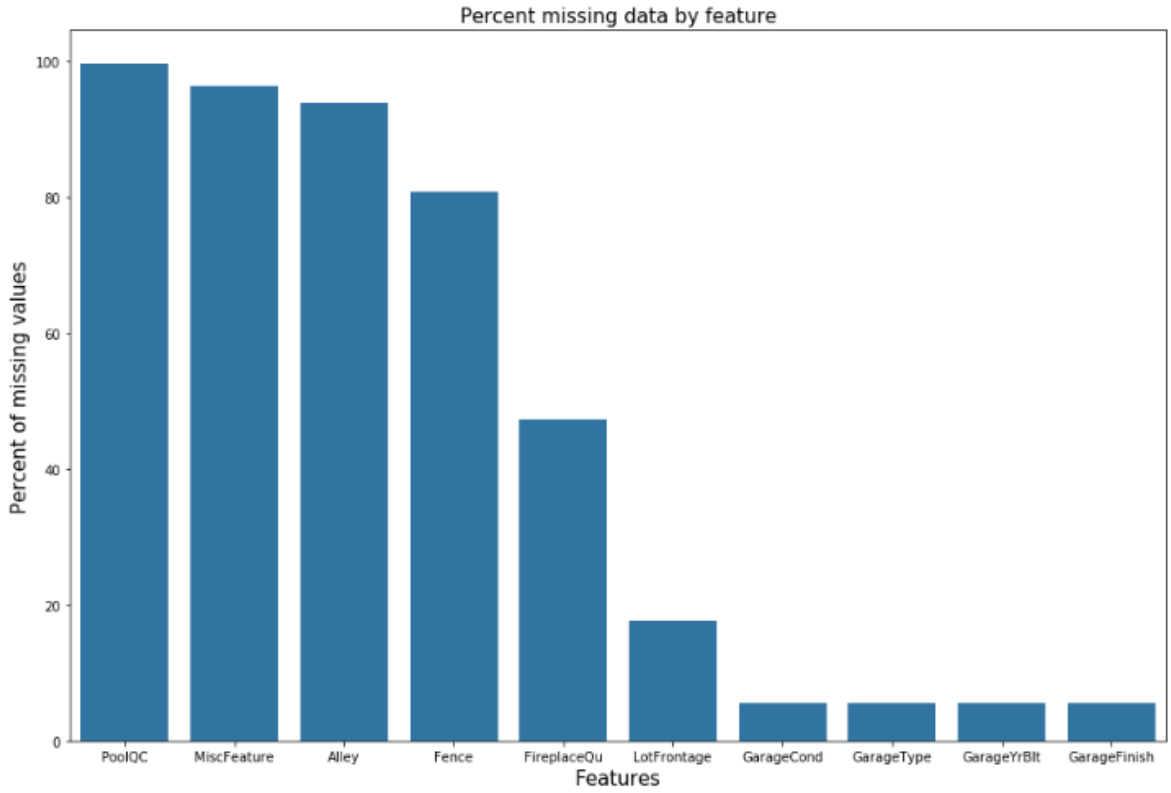
4. What is Exploratory Data Analysis (EDA)?

- How to ensure you are ready to use machine learning algorithms in a project?
- How to choose the most suitable algorithms for your data set?
- How to define the feature variables that can potentially be used for machine learning?

Exploratory Data Analysis (EDA) helps to answer all these questions, ensuring the best outcomes for the project. It is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a data set

Missing Values:

In the fig below, I'm basically finding out the percentage of missing values in each and every feature. And also we observed that 'PoolQC', 'MiscFeature', 'Alley' and 'Fence' these 4 features are having above 50% of missing values.



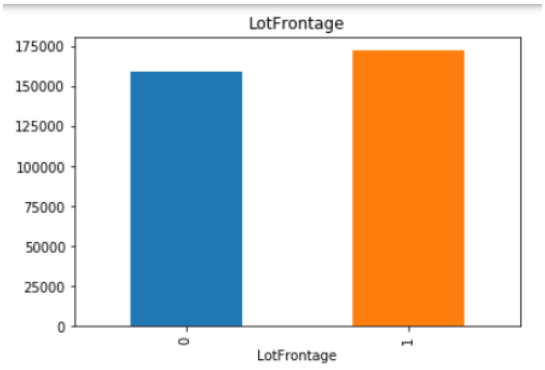
Since they are many missing values, we need to find the relationship between missing values and Sales Price:

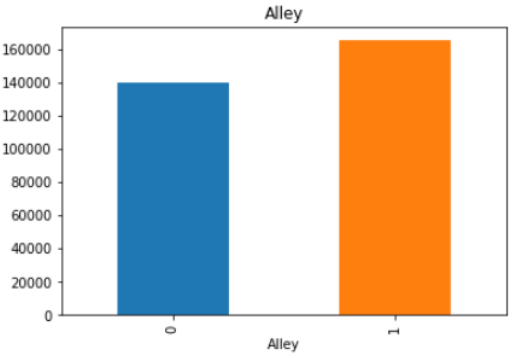
Understanding, whether that missing values has some dependency or is there any relationship with the dependent feature which is the SalesPrice. For that we are plotting. Let's make a variable that indicates 1 if the observation was missing or zero otherwise.

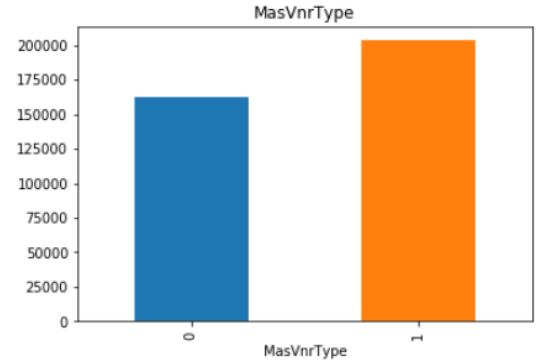
Suppose if a feature had a null value I'm converting it as '1' or else '0'. The reason I'm doing this is to create count plot that will help me to understand that the missing values plays an important role or not.

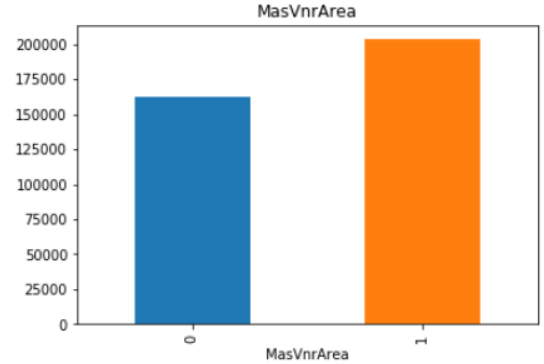
For ex, in the below bar plot LotFrontage feature had lot NaN values. And I convert these Nan values into '1'. Because of this NaN values Salesprice also increases. That means LotFrontage feature plays an important role. Since there are many missing values, we need to find the relationship between missing values and Sales Price.

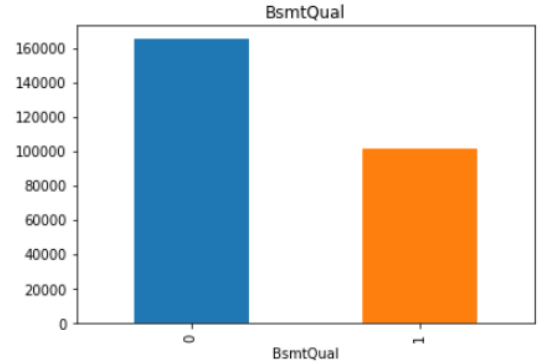
Let's plot some diagram for this relationship

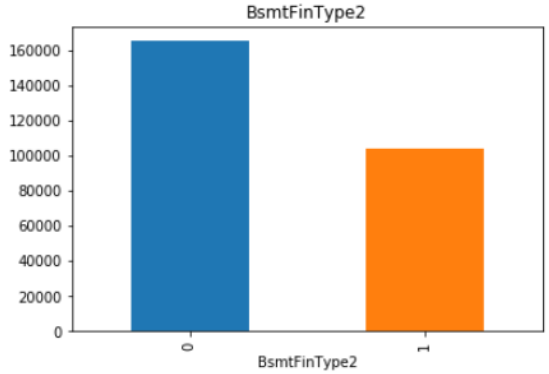

LotFrontage
175000
150000
125000
100000
75000
50000
25000
0
0
1
LotFrontage

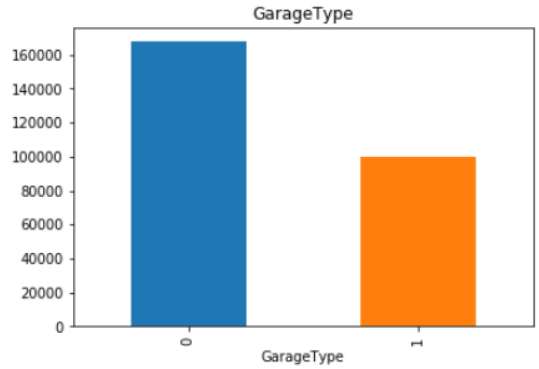

Alley
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
Alley

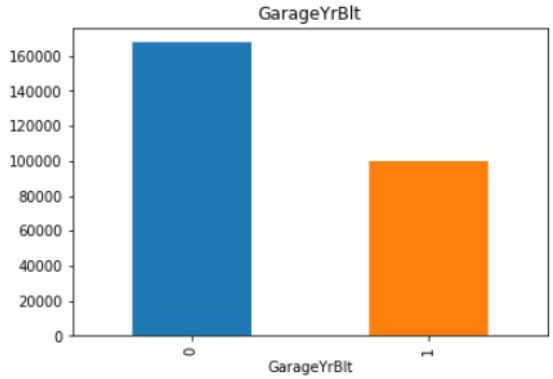

MasVnrType
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
MasVnrType

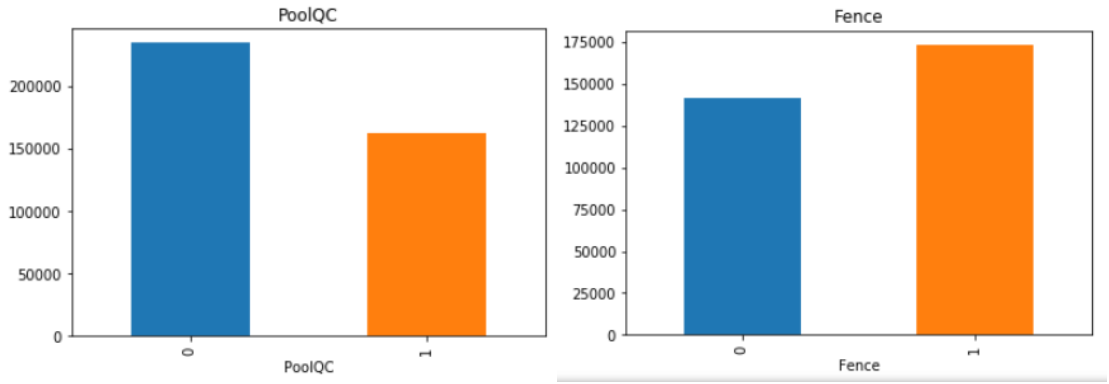

MasVnrArea
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
MasVnrArea


BsmtQual
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtQual


BsmtFinType2
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtFinType2

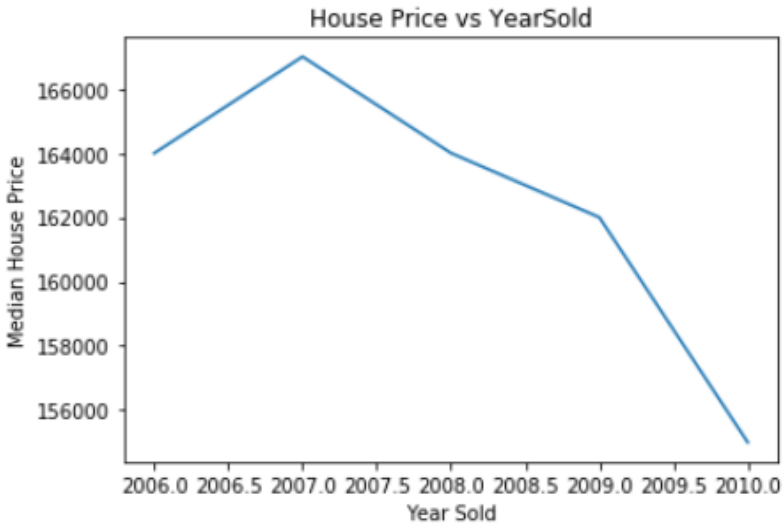

GarageType
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
GarageType


GarageYrBlt
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
GarageYrBlt

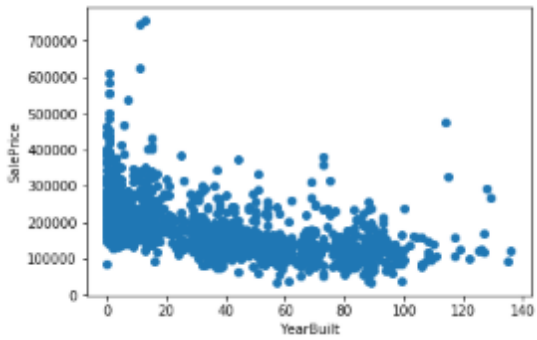


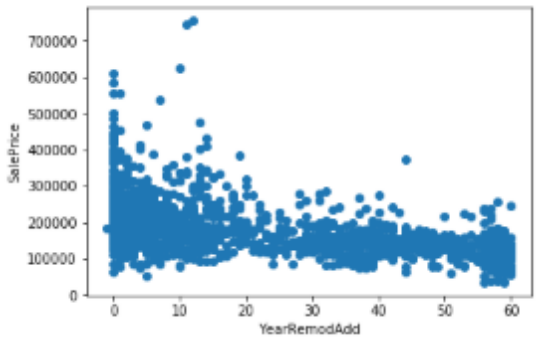
Let's analyze the Temporal Date-time Variables:

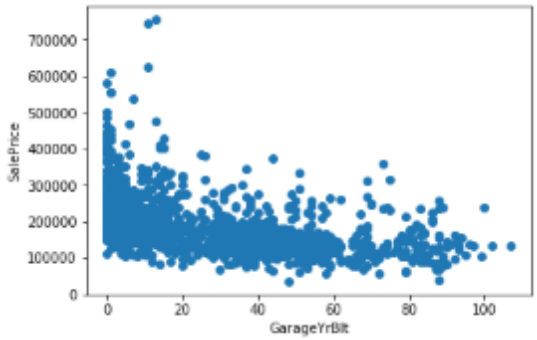
We will check whether there is a relation between year the house is sold and the sales price. As we see in the below fig, as the year sold is going on the price is decreasing this cannot be just true. So we'll try to find out some more information from this.



Here we will compare the difference between all years feature with SalePrice. We will capture the difference between year variable and year the house was sold for. That difference I'm trying to shown in the below plot. It tell us that suppose if the house was 140 years old then the price of the house was decreasing. If the house was newly built in that case the price of the house was too high. Likewise if you see the GarageYrBlt vs SalePrice, if the GarageYrBlt year is too old then the SalePrice decreases or else the SalePrice increases.







Numerical variables are usually of 2 types:

1. Discrete Variables

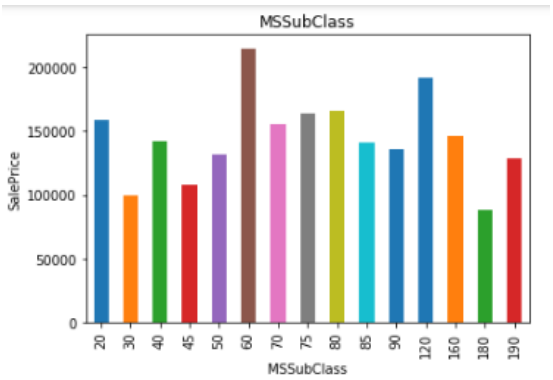
2. Continuous Variables

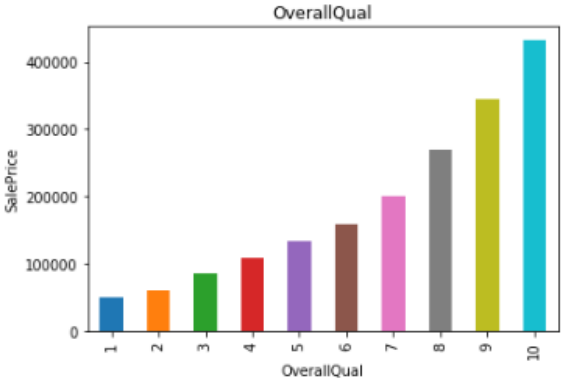
There are 38 Numerical variables, in that 17 features are Discrete Variables, 16 features are Continuous Variables, 4 temporal date-time Variables and Id variable.

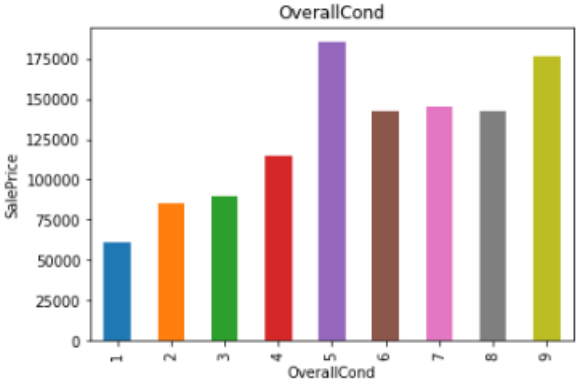
1. Discrete Variables:

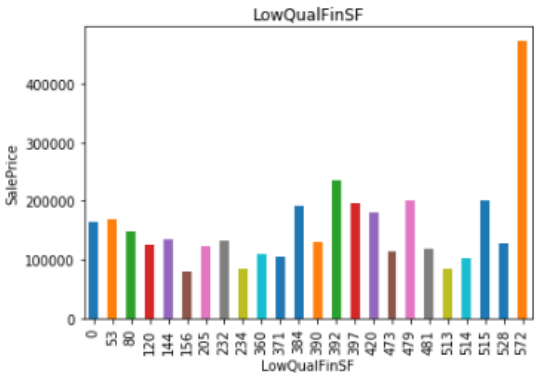
Let’s find the relationship between Discrete Variables and SalePrice.

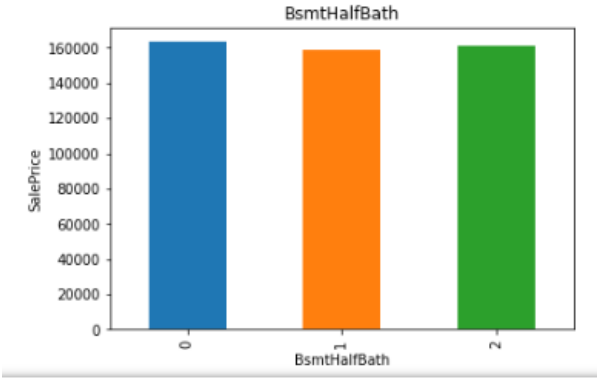
Suppose if we take the OverallQual vs SalePrice, As overall quality increases then saleprice is also increases. It tells us based on the overall quality our saleprice also increases. Likewise we’ll see the other Discrete Variables (vs) SalePrice.

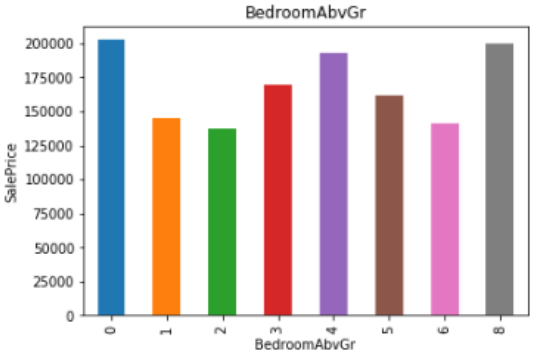

MSSubClass
SalePrice
MSSubClass
0
50000
100000
150000
200000
20
30
40
45
50
60
70
75
80
85
90
120
160
180
190

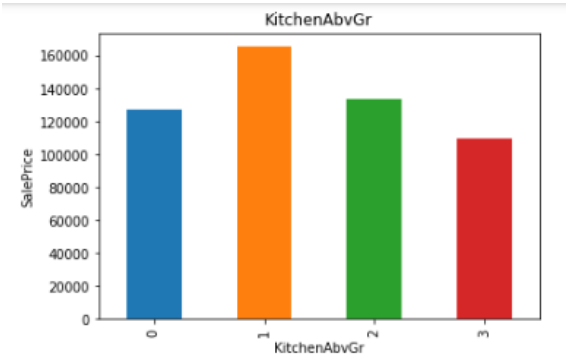

OverallQual
SalePrice
OverallQual
0
100000
200000
300000
400000
1
2
3
4
5
6
7
8
9
10

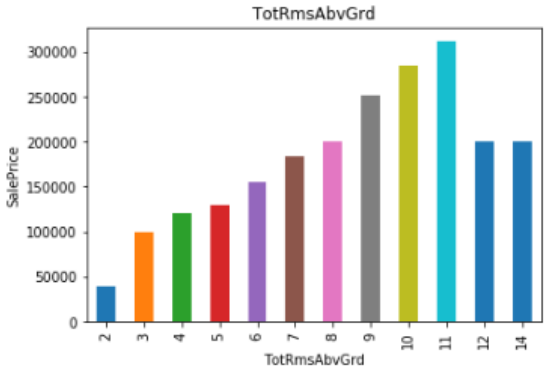

OverallCond
SalePrice
OverallCond
0
25000
50000
75000
100000
125000
150000
175000
1
2
3
4
5
6
7
8
9

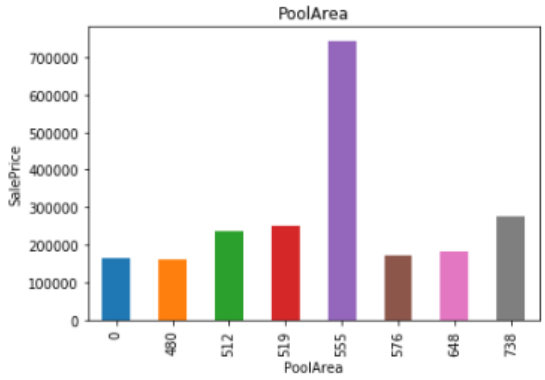

LowQualFinSF
SalePrice
LowQualFinSF
0
100000
200000
300000
400000
0
53
80
120
144
156
205
232
234
360
371
384
390
392
397
420
473
479
481
513
514
515
528
572

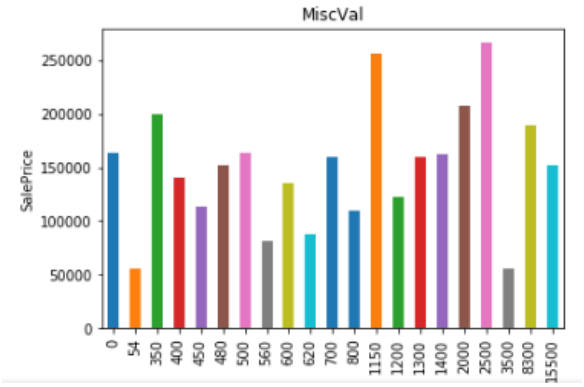

BsmtHalfBath
SalePrice
BsmtHalfBath
0
20000
40000
60000
80000
100000
120000
140000
160000
0
1
2


BedroomAbvGr
SalePrice
BedroomAbvGr
0
25000
50000
75000
100000
125000
150000
175000
200000
0
1
2
3
4
5
6
8


KitchenAbvGr
SalePrice
KitchenAbvGr
0
20000
40000
60000
80000
100000
120000
140000
160000
0
1
2
3


TotRmsAbvGrd
SalePrice
TotRmsAbvGrd
0
50000
100000
150000
200000
250000
300000
2
3
4
5
6
7
8
9
10
11
12
14



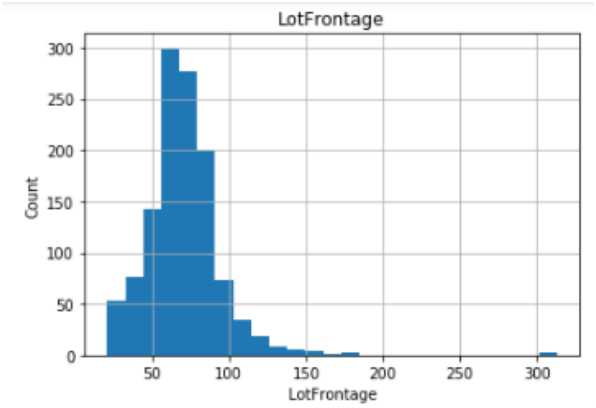


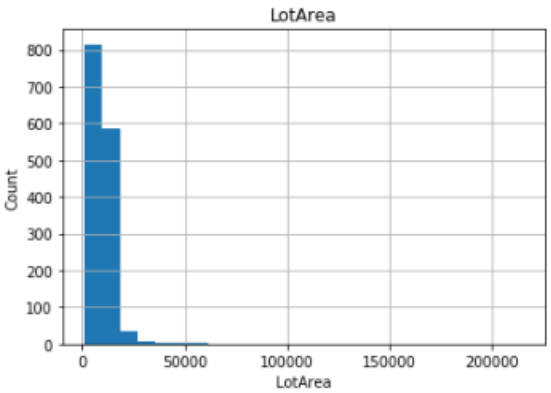
2. Continuous Variable:

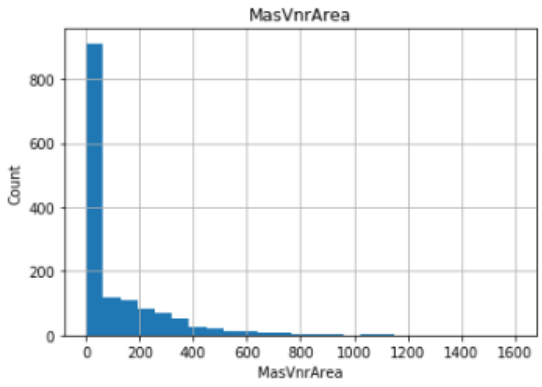
Let's analyse the continuous values by creating histograms to understand the distribution.

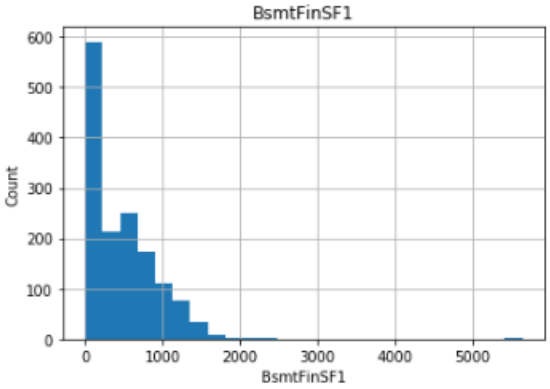
Here in this case we need to find out the distribution of continuous values. For that we are creating the histogram. Here some of the features doesn't having Gaussians Distribution, these are skewed data. So we need to convert those skewed data into Gaussian's Distribution using log normalization, because that will be helpful for linear model prediction, that is pretty much important.

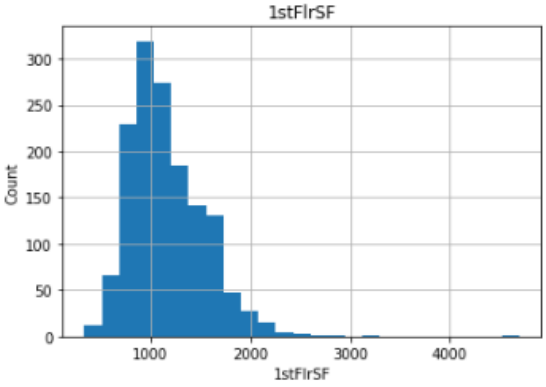
For example in the below, 1stFlrSf feature is not a Gaussian's Distribution. So we need to convert that feature into Gaussian distribution.

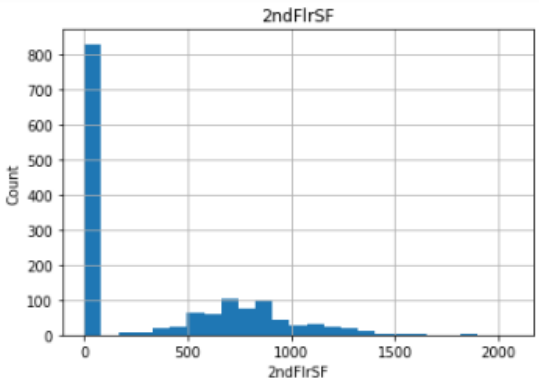


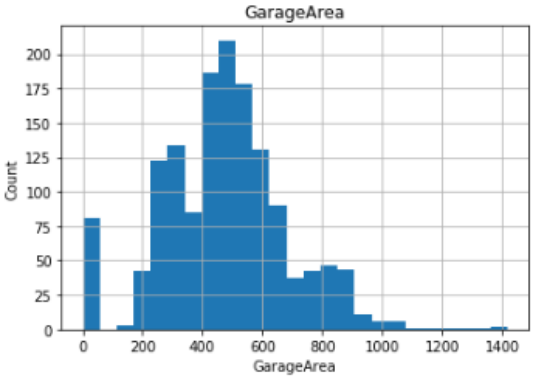


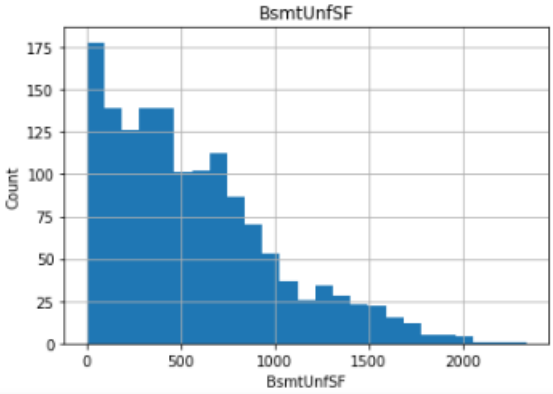






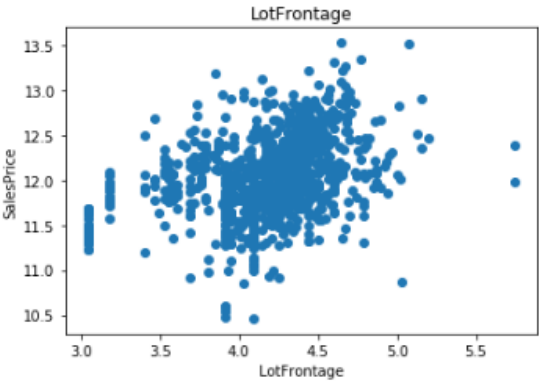


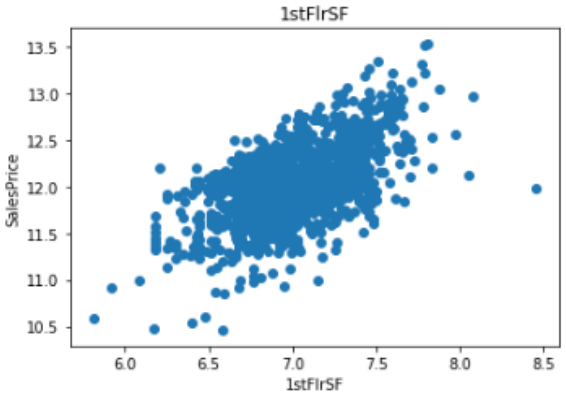


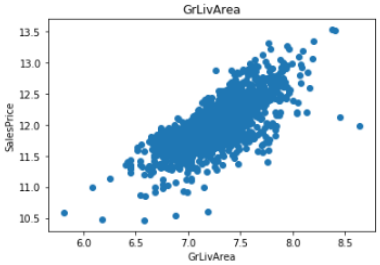


Using logarithmic transformation:

From the above continuous variables we saw that some of the features are not Gaussians Distribution. So it is very important to convert that features into Gaussian's Distribution that's why we are using this logarithmic transformation.

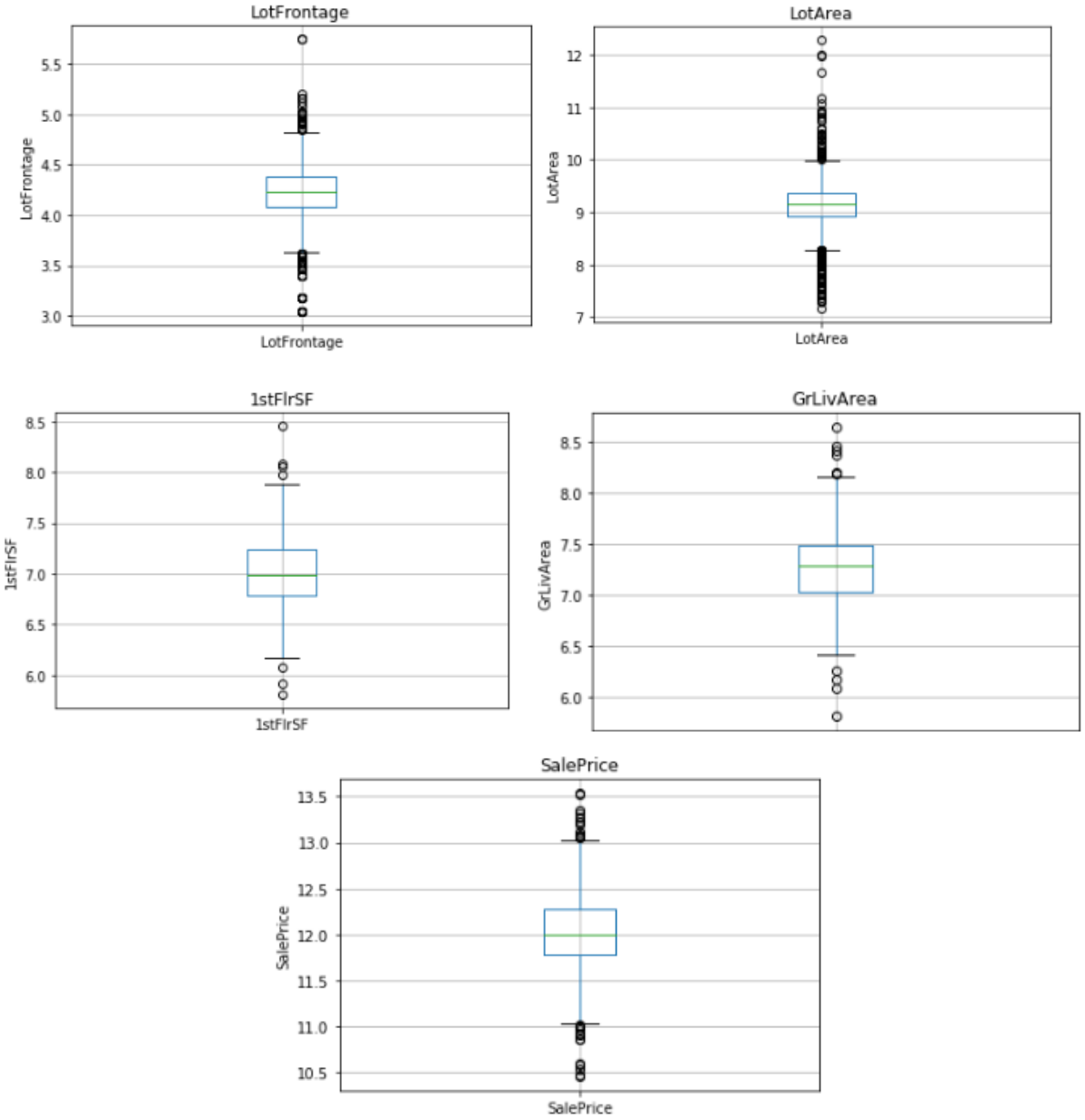




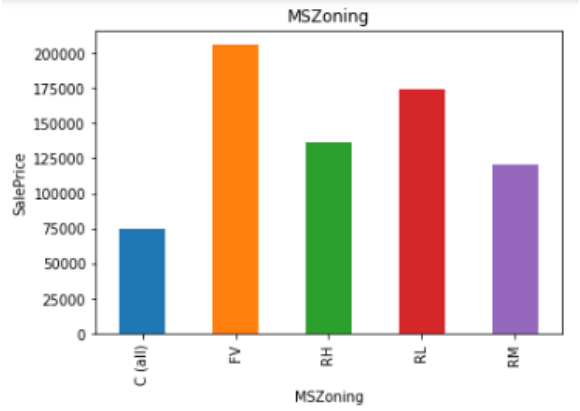


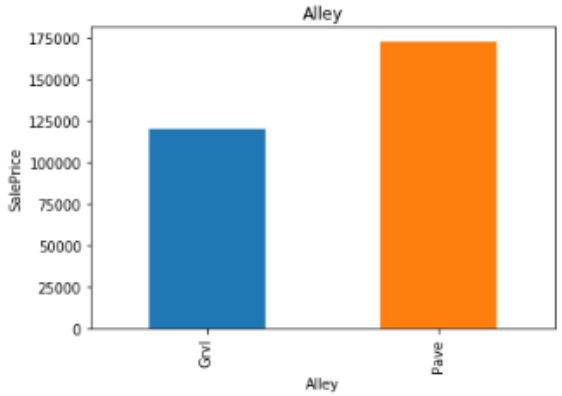
Outliers:

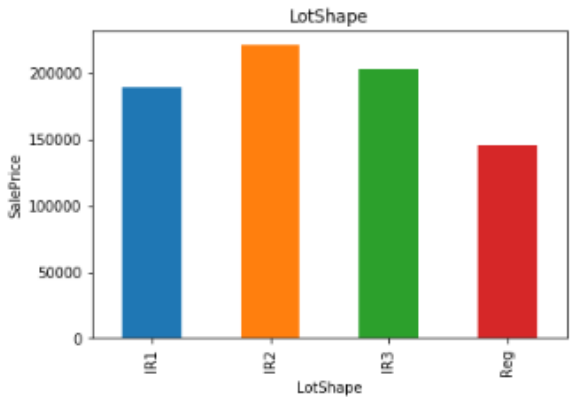
By using Boxplot, in each and every continuous variable we can actually find out the outliers.

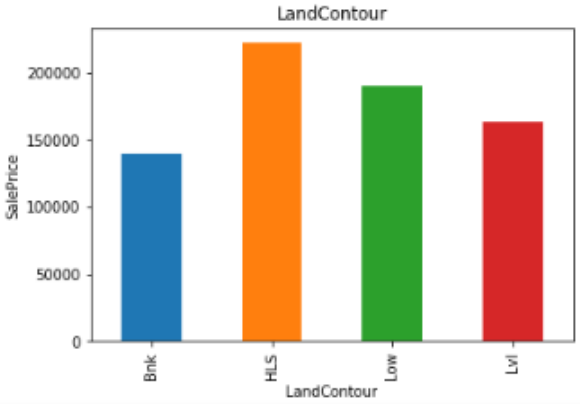


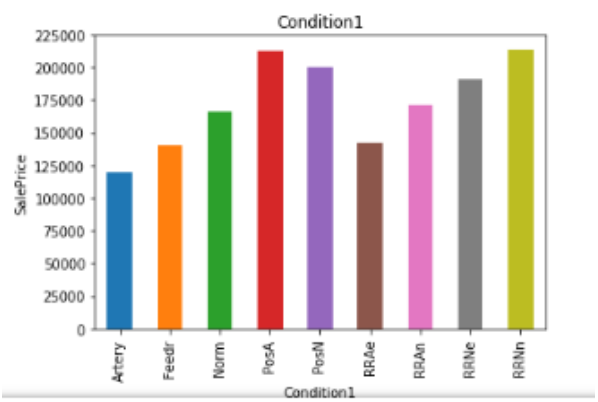
Find out the relationship between categorical variable and dependent feature SalesPrice

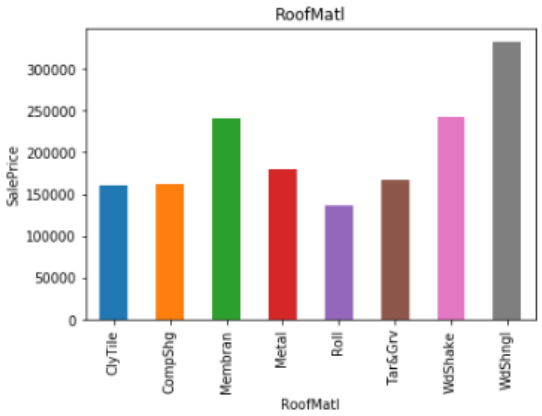




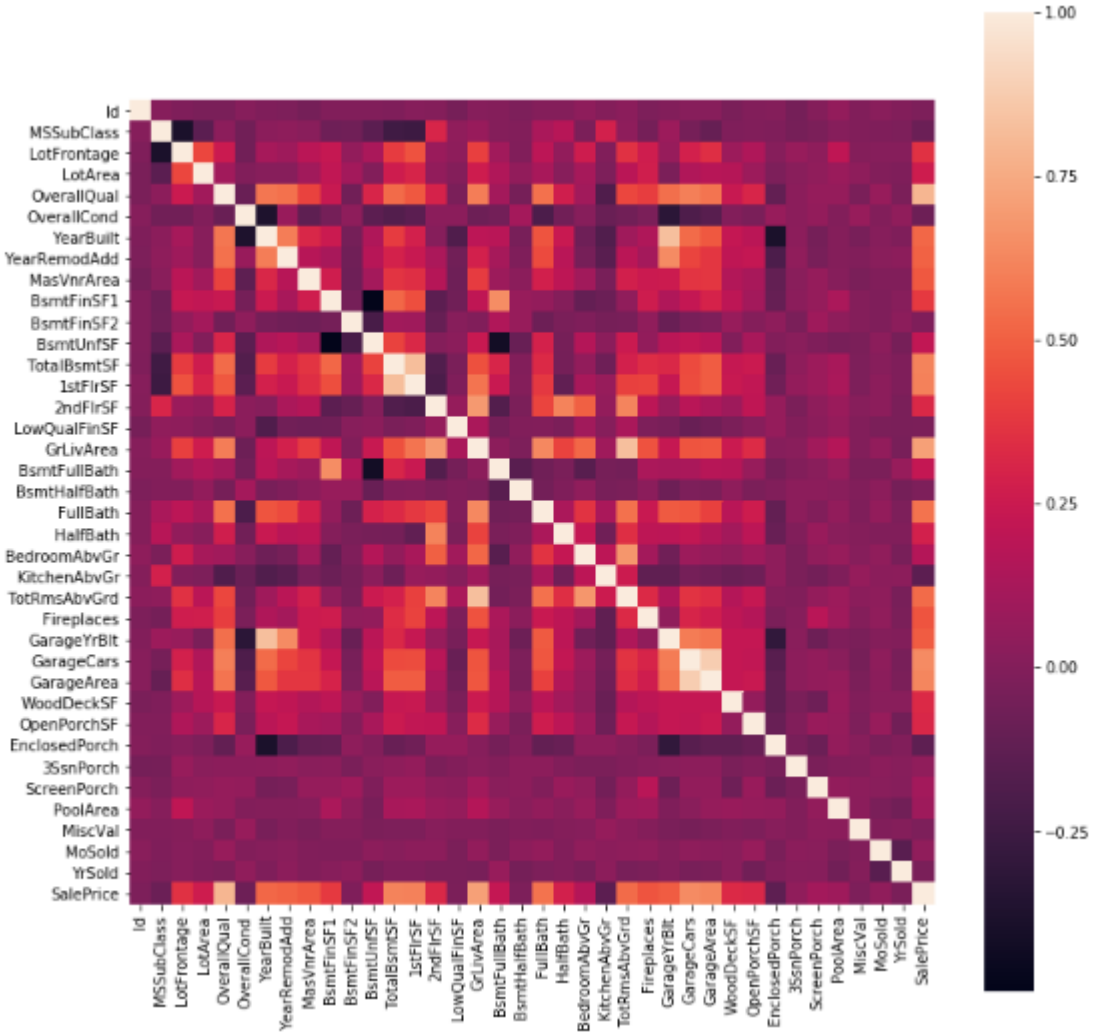








Finding Correlation between different features:



The heat map is the best way to get a quick overview of correlated features thanks to seaborn!

At initial glance it is observed that there are two red coloured squares that get my attention.

The first one refers to the 'TotalBsmtSF' and '1stFlrSF' variables. Second one refers to the 'GarageX' variables. Both cases show how significant the correlation is between these variables. Actually, this correlation is so strong that it can indicate a situation of multicollinearity. If we think about these variables, we can conclude that they give almost the same information so multicollinearity really occurs. Heat maps are great to detect this kind of multicollinearity situations and in problems related to feature selection like this project, it comes as an excellent exploratory tool.

Another aspect I observed here is the 'SalePrice' correlations.As it is observed that 'GrLivArea', 'TotalBsmtSF', and 'OverallQual' saying a big 'Hello !' to SalePrice, however we cannot exclude

the fact that rest of the features have some level of correlation to the SalePrice. To observe this correlation closer let us see it in Zoomed Heat Map

SalePrice Correlation matrix:

The above heat map shows that

1 OverallQual', 'GrLivArea' and 'TotalBsmtSF' are strongly correlated with 'SalePrice'.

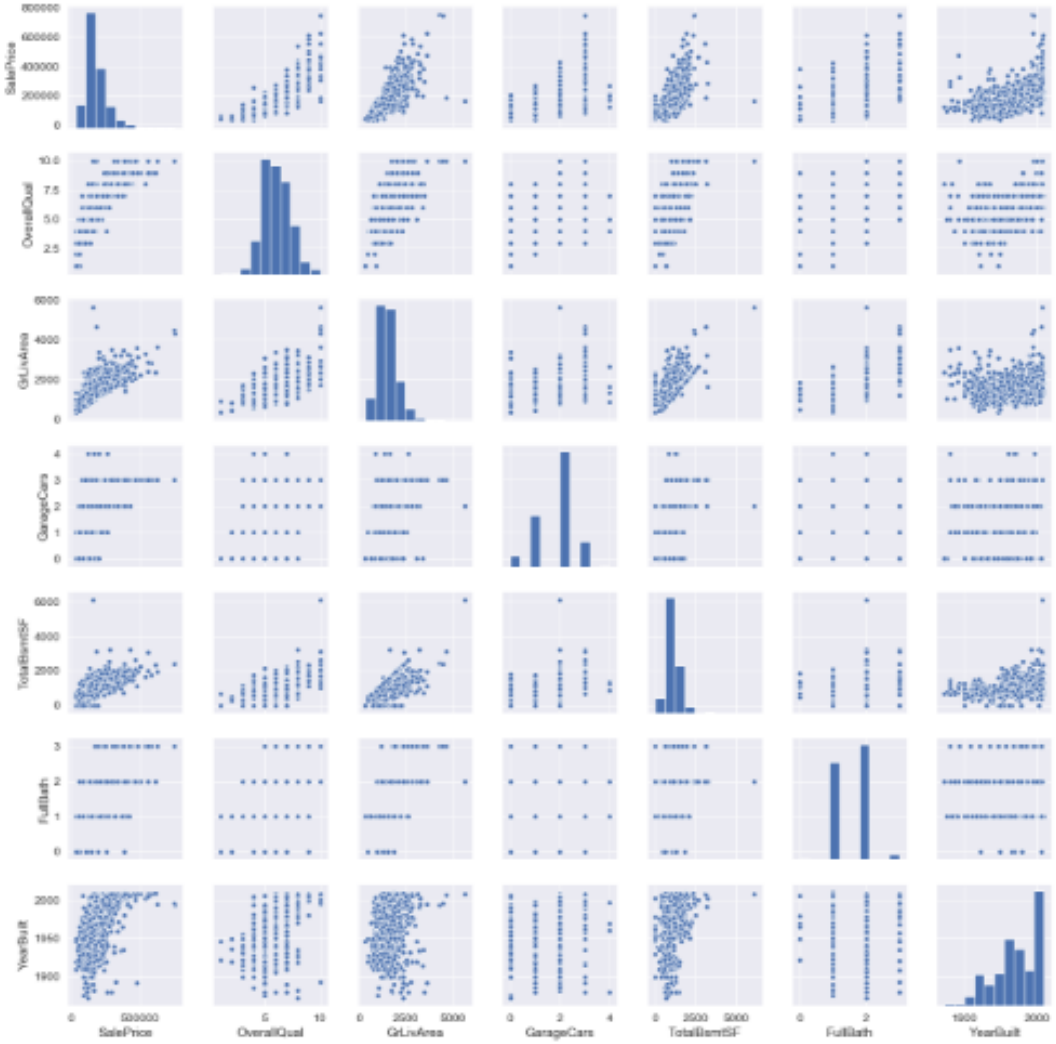
2 'GarageCars' and 'GarageArea' are strongly correlated variables.

3 'TotalBsmtSF' and '1stFloor' seem to be correlated with each other.

4 'TotRmsAbvGrd' and 'GrLivArea' also seem to correlated with each other.

Pair Plot:

Although we already know some of the main figures, this pair plot gives us a reasonable overview insight about the correlated features.



1. One interesting observation is between 'TotalBsmtSF' and 'GrLiveArea'. In this figure we can see the dots drawing a linear line, which almost acts like a border. It totally makes sense that the majority of the dots stay below that line. Basement areas can be equal to the above ground living area, but it is not expected a basement area bigger than the above ground living area.
2. One more interesting observation is between 'SalePrice' and 'YearBuilt'. In the bottom of the 'dots cloud', we see what almost appears to be an exponential function. We can also see this same tendency in the upper limit of the 'dots cloud'
3. Last observation is that prices are increasing faster now with respect to previous years.