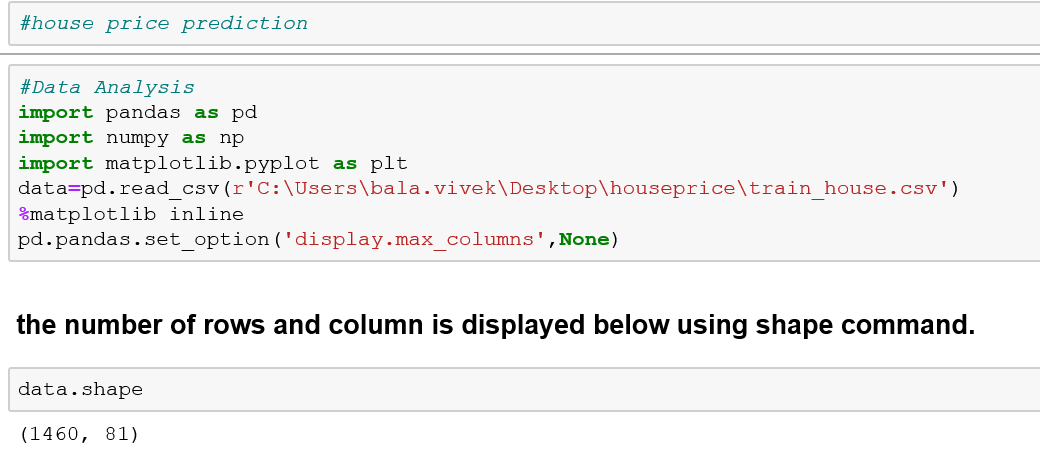
**Exploratory Data Analysis for Machine learning**

The are data set which Im interested to pick is House Price Prediction which has 81 feature columns and one target field.



**Description of Data set** :

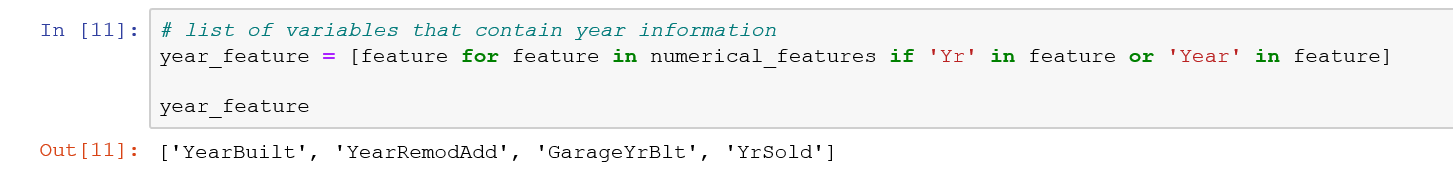
The target column is ‘**SalePrice**’ which has to be predicted based on the other input features.

I can run the data.describe or data.info() method but as the feature list is huge we cannot visualize it quickly.

The first step is to separate the **numerical features** and the Object features from the dataset.

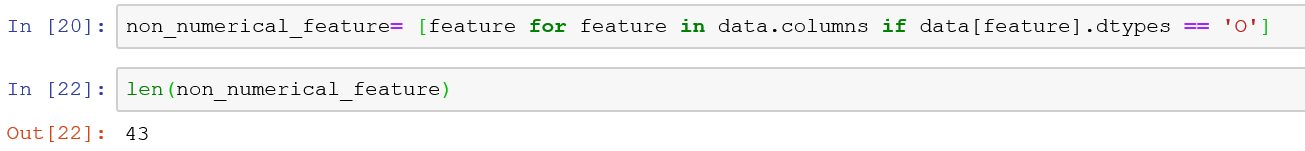


The dtypes which are **not equal to O** are considered as numerical features. And I will separate the features which are Object type as below.



The above code represent to segregate the **Year or timedate features**. By seperating the based on the data type we can infere a lot of data from data set.

The feature which has dtype as Object are separated and stored in another variable as non\_numerical\_feature.

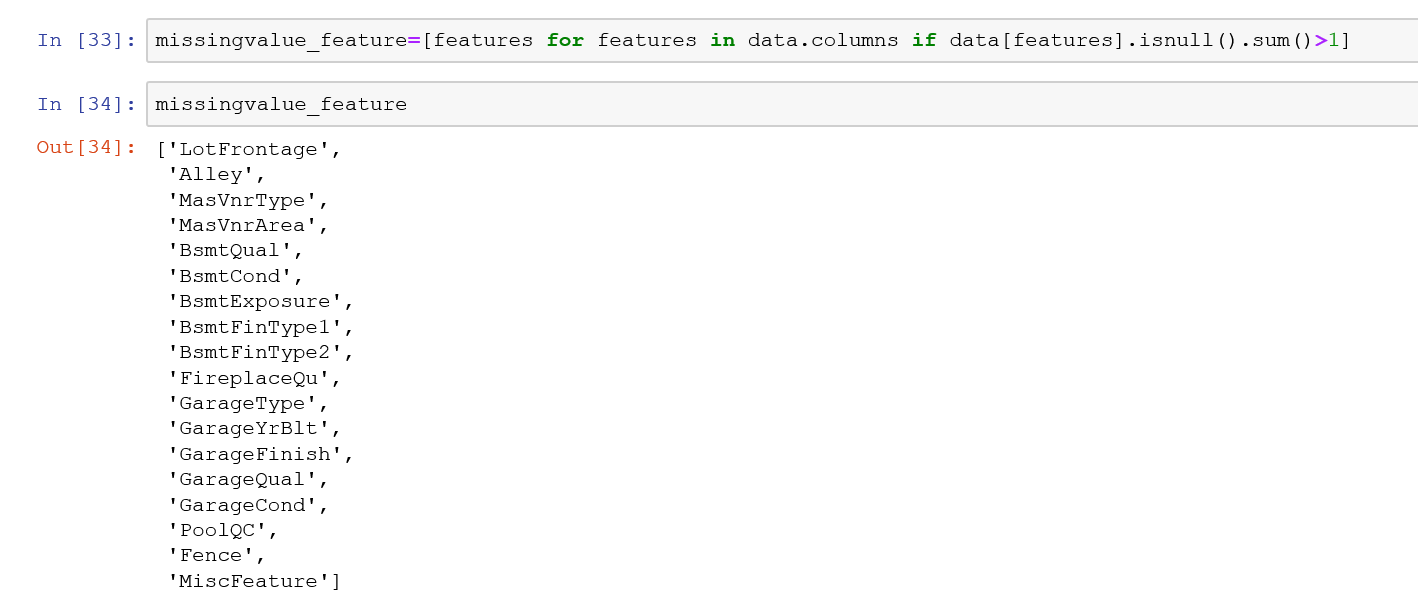


**In Data Analysis We will Analyze To Find out the below stuff**

1. Missing Values
2. All The Numerical Variables
3. Distribution of the Numerical Variables
4. Categorical Variables
5. Cardinality of Categorical Variables
6. Outliers
7. Relationship between independent and dependent feature(SalePrice)

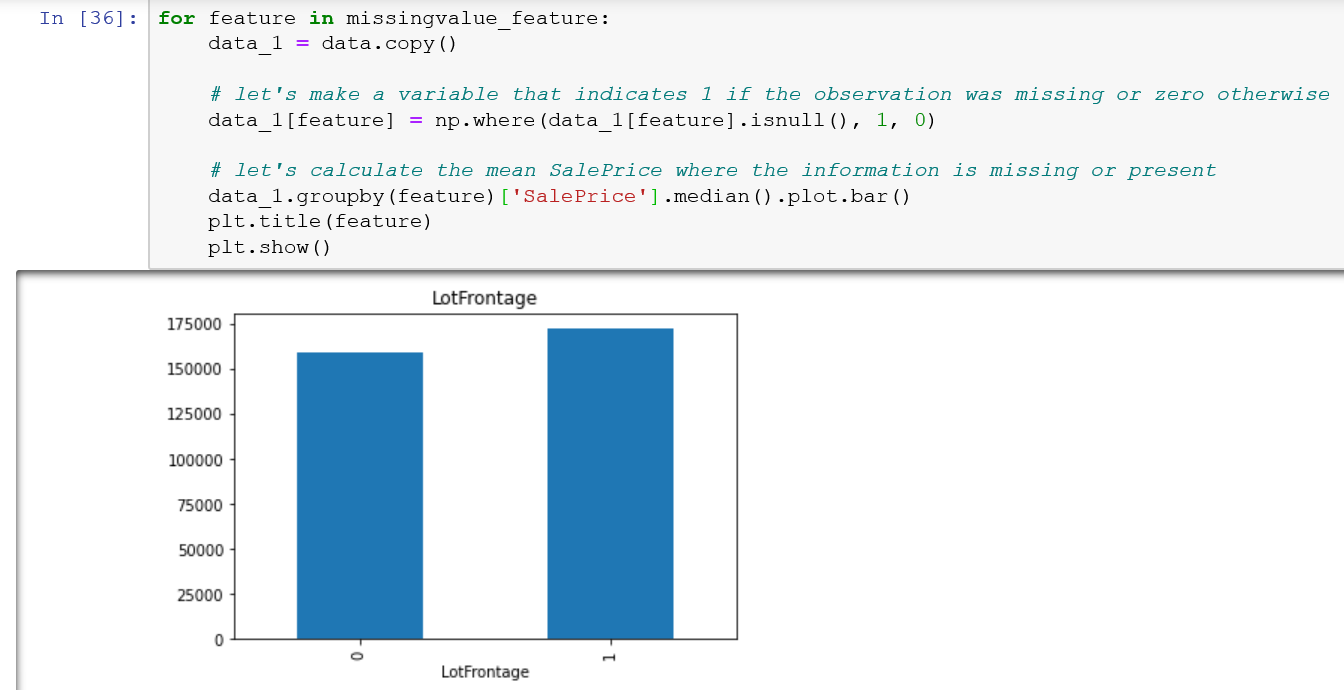
**Missing Values :-**

The missing value in the data set can be found by the isnull method. I have created a list comphrension to find the missing value for a feature whose sum is greater than 1. The below list of features has the missing value which is more than 1.



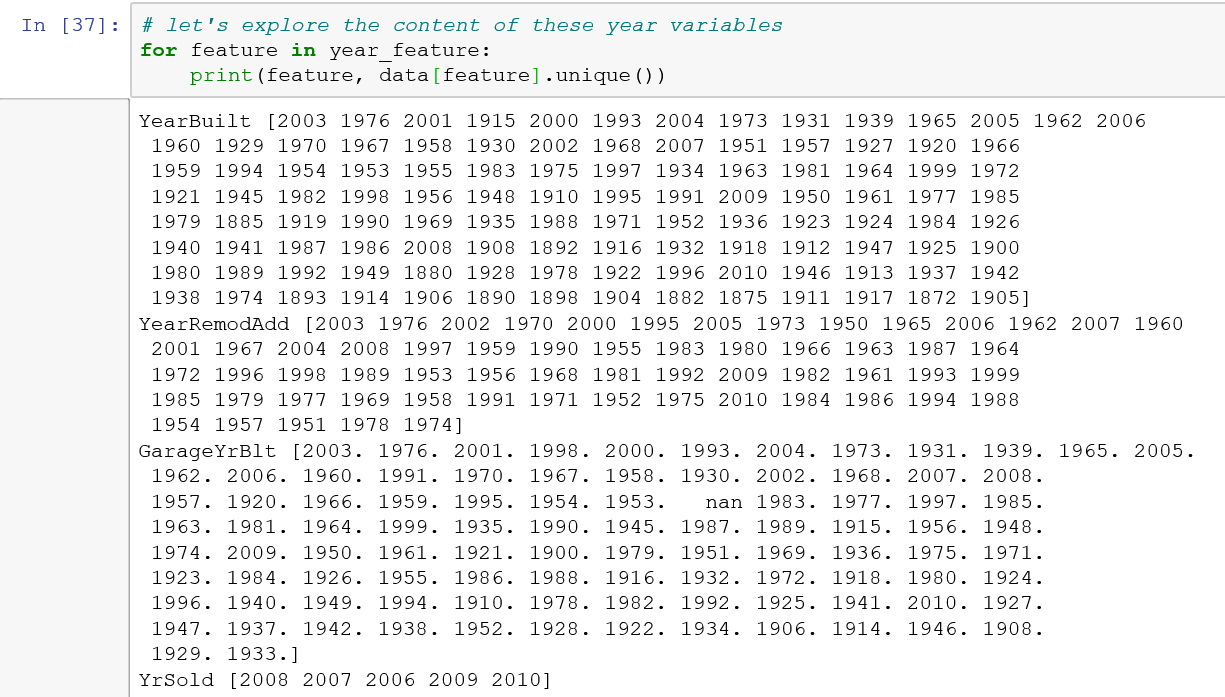
1. Needed to find the relation between the missing value and the imapct of the sales price.

The below bar chart will help us to show the relation between the missing values for each feature with respect to the increase or decrease in sale price. The below piece of code the null values is converted as 1 if there is a value present then it is considered as 0.

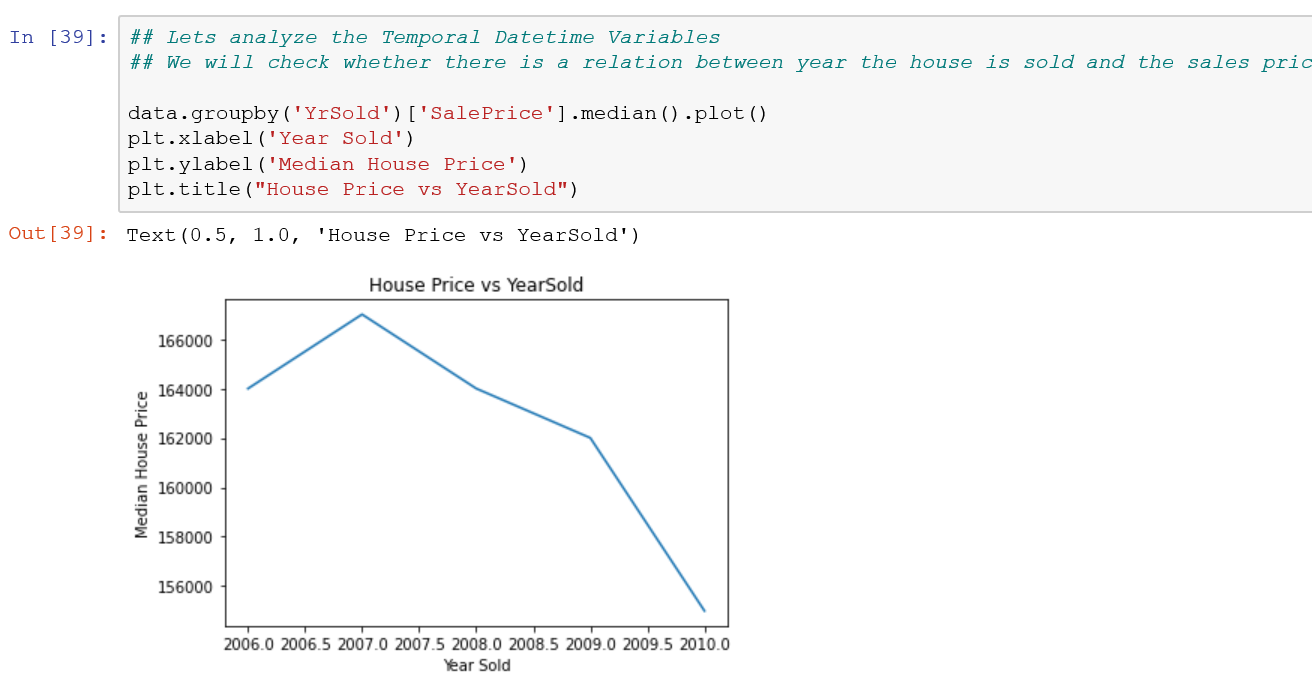
w

When the null or missing values are high the sale price seems to be high for the features like – MasvnrType, MasvnrArea, Fence, MiscFeature. Its clearly visible that we need to replace the missing value with meaningful value.

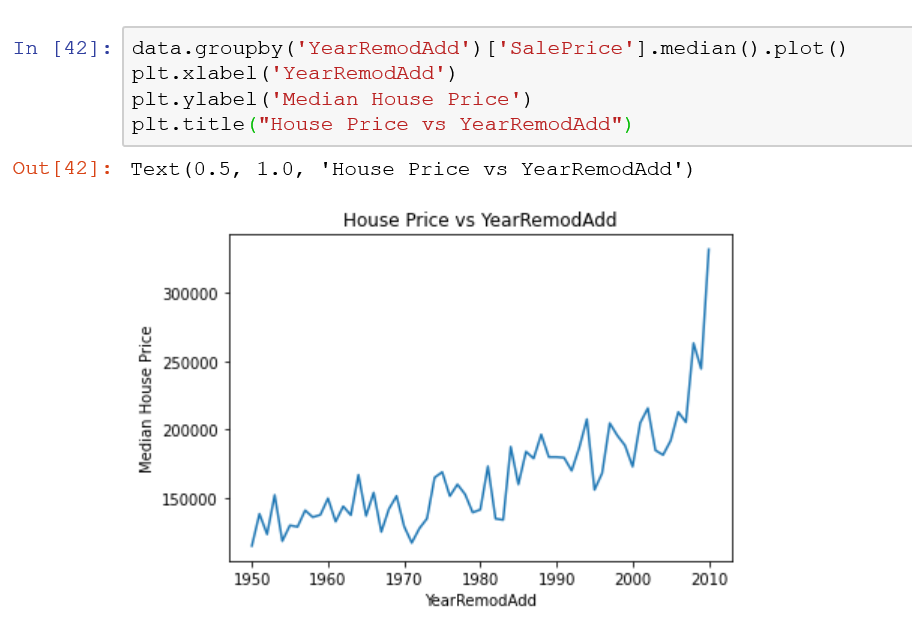
* The feature ID is holding the unique occurences of values which is not required for the model building so I will dropping the feature by using the “drop” command with axis=1.
* In the ‘year\_feature’ where the year build , year RemodAdd and GarageYrBlt and year Sold has many unique year values.



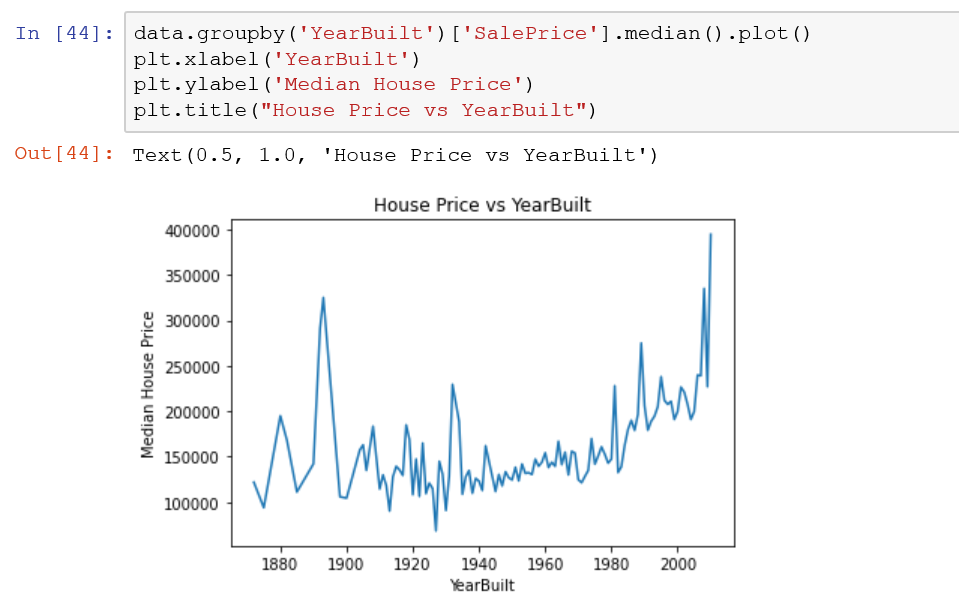
* Lets see the relation between the year sold and Sale price of the house below, its clear that the price of the house decreases if the year sold feature is more.



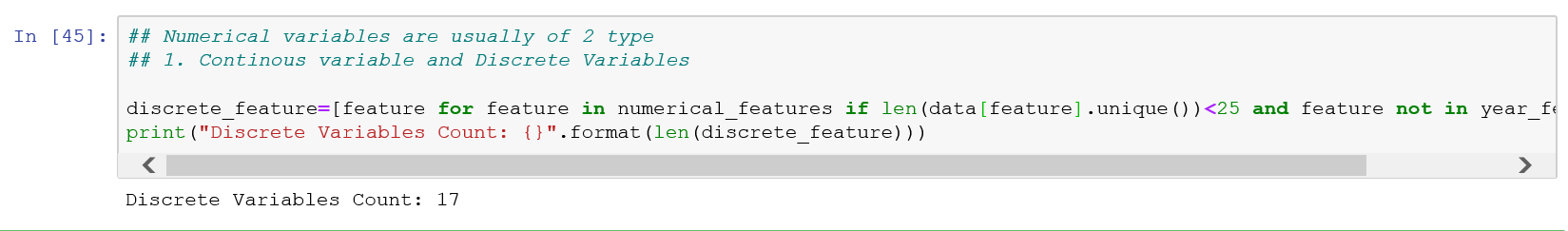
* From the relation between YearRemodAdd feature and the sale price they seems to be directly related to each other. As the year value increases the Price started to gradually increase.



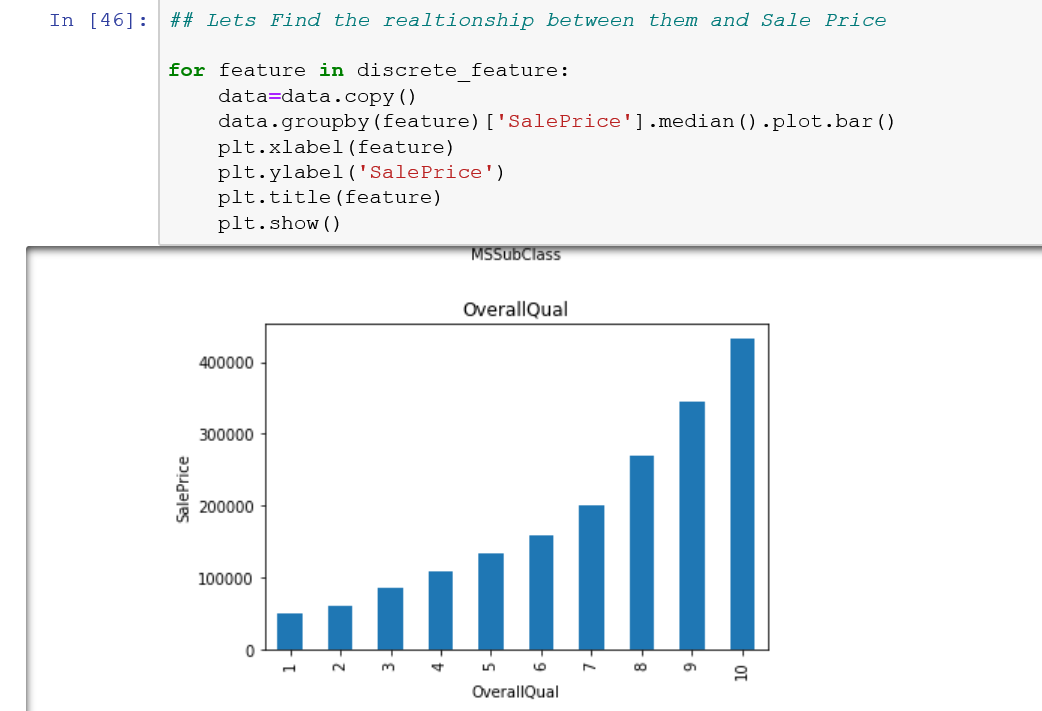
* The relation between the year built and the sale price is given below. We can clearly notice that the new the building more the Sale price.



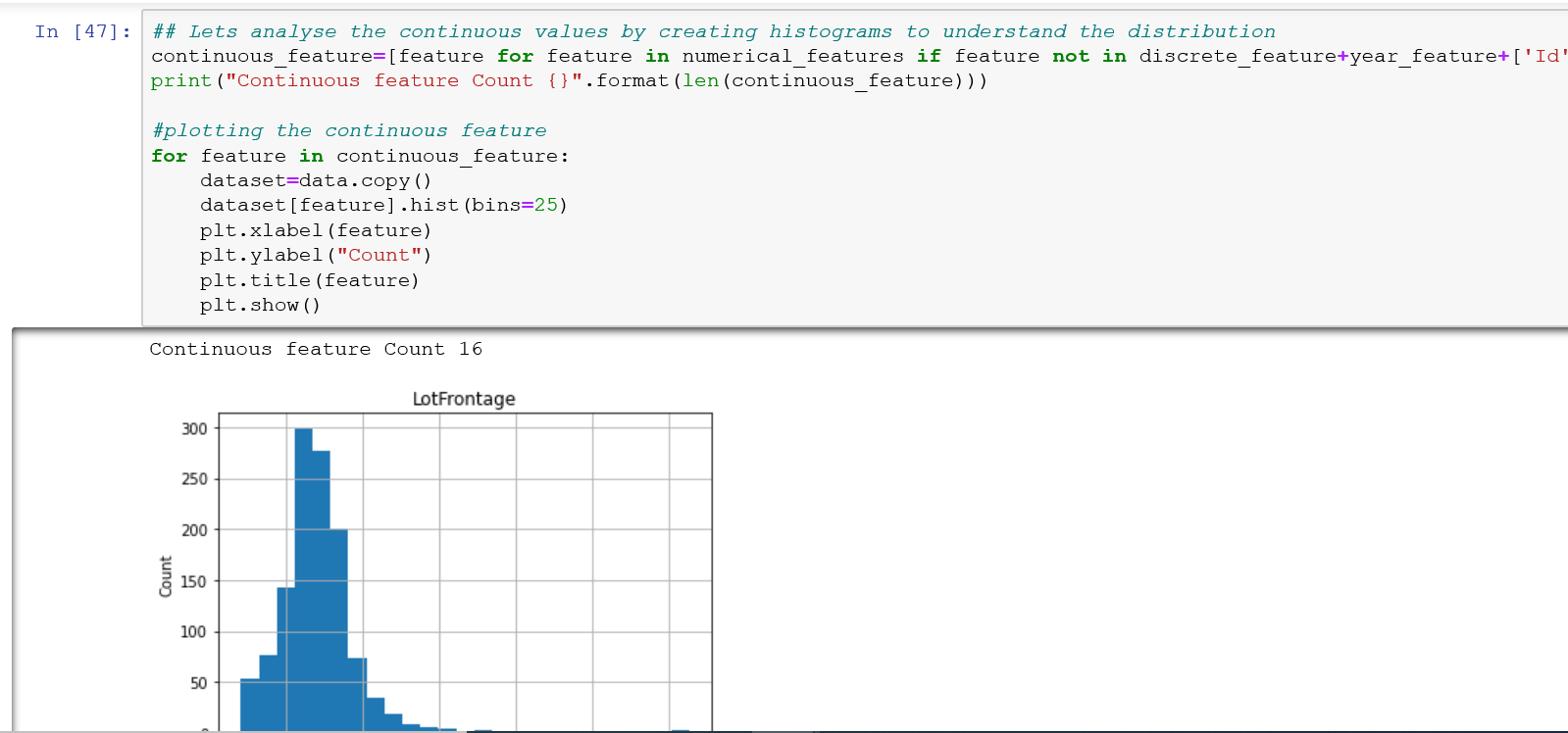
* I will separate the discrete features from the numerical\_feature column which is already created. I will be considering a feature as Discrete Feature when it has max of 25 unique values present. By this way we can filter the discrete feature.



* Finiding the relation between the Sale price and the Discrete feature by ploting bar chart as below. The overall qual , FullBath , GarageCars has direct relation with the saleprice. They are directly propotional.

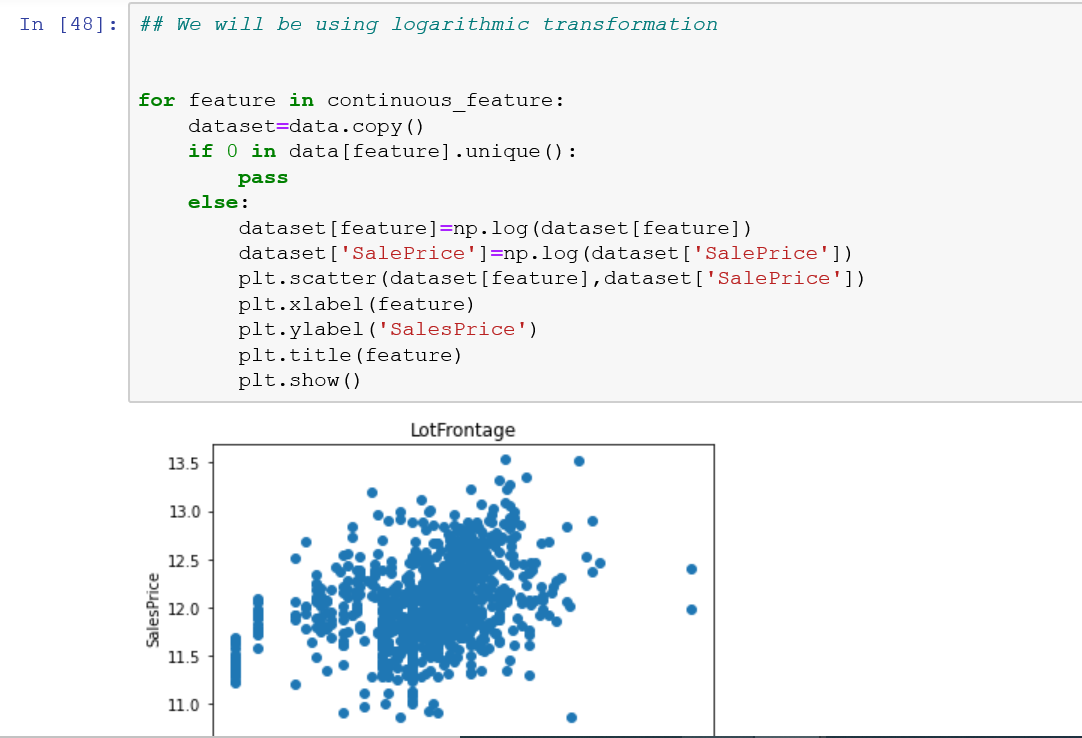


* One more category which has to be identified is the continuous feature value and understanding its distribution.



It is visible that many features like MasVnrArea, LotArea, BsmtFinSF1 , BsmtFinSF2, BsmtUnfSF, WoodDeckSF, OpenPorchSF , ScreenPorch, EnclosedPorch seems to be right skewed ! it’s always recommented to make the distrubution to noraml using transformations like Box cox or log normal transform.

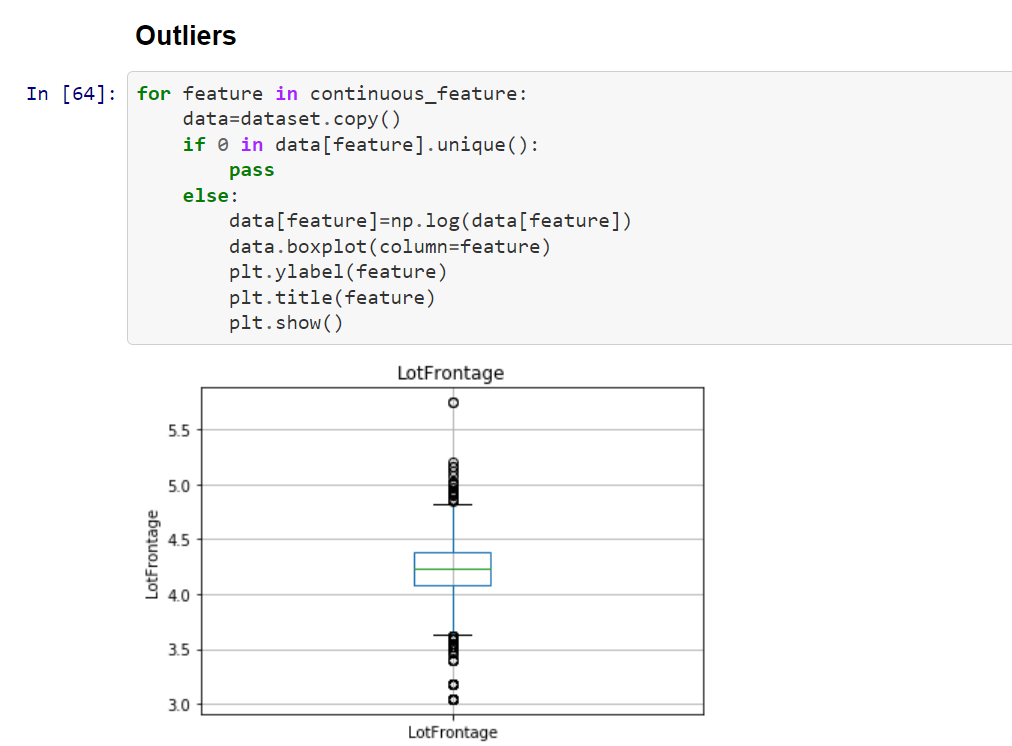
For this scenario of continuous feature Im taking the logarathmic transformation so the skewed data will be converted to the Normal transformation !



**Outliers :**

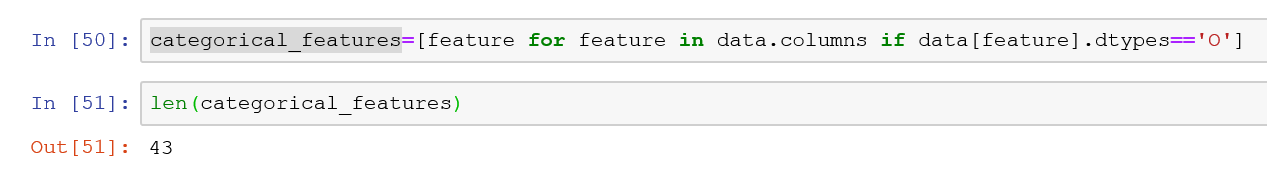
The intention is to find the outliers in the Numerical data features i.e the continuous\_feature which is dervied above. Im plotting the Box plot in order to identify the outliers in each of the continuous feature.

From the observation I can see clearly that features like – LotFrontage, LotArea, 1stFlrSF, GrLivArea all the numerical features will be fetched in the box plot.

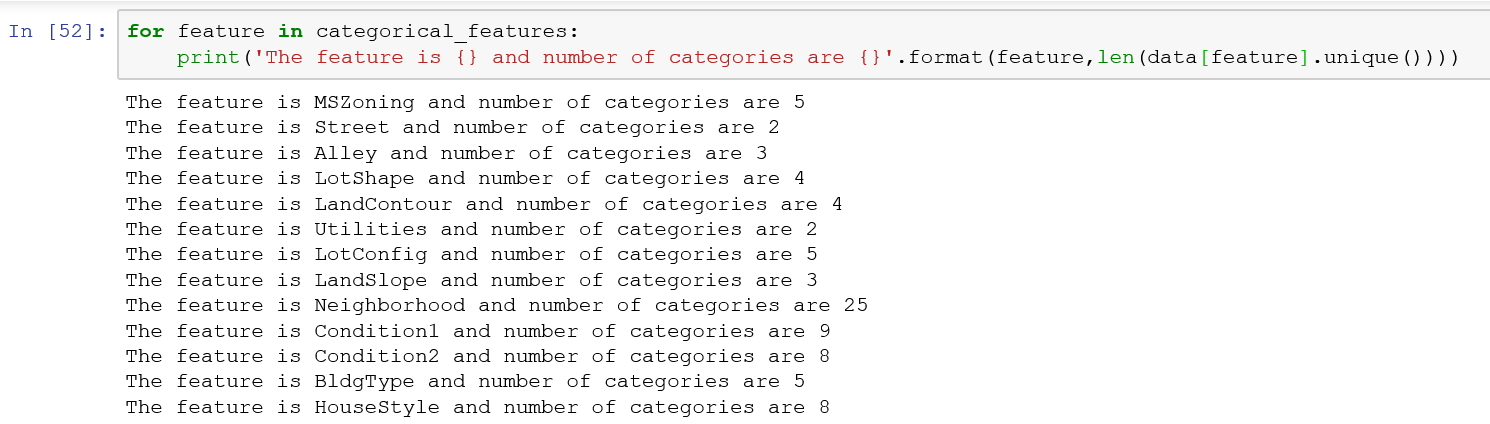


**Categorical Variable :**

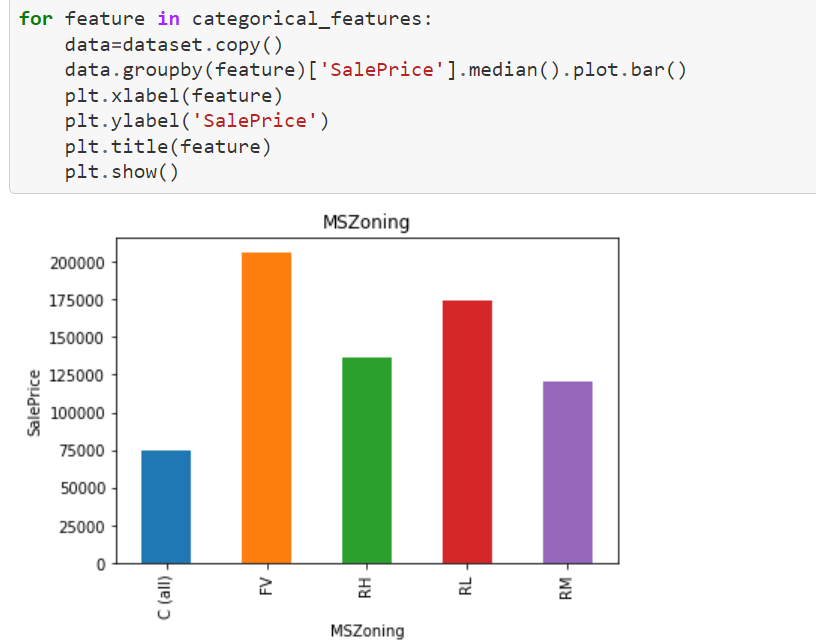
The categorical variable are seperated with the below following method. Im trying to fetch out the number of unique category available for each of the category.



The below method shows the number of unique category available at each feature.

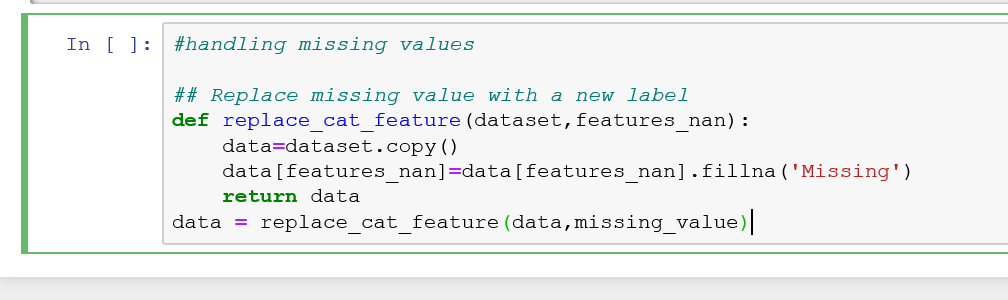


The relation between each categorical feature with the target sale price is calculated in order to know the impact on the target sale price.



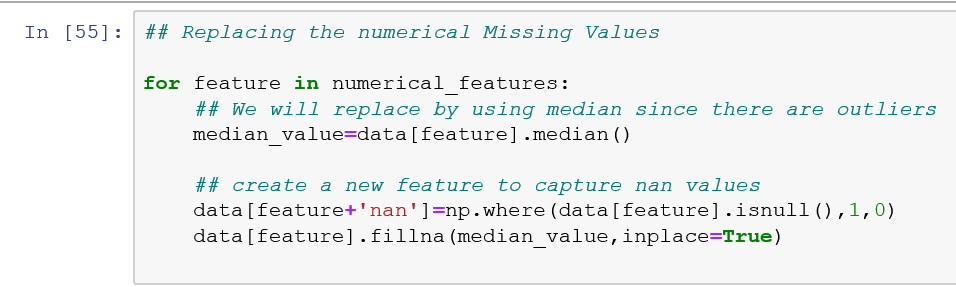
Feature Engineering :

Handling missing value for the Categorical features, I will be replacing the NAN with “Missing” keyword so that I can avoid having any NAN values.



The above method will make sure all the NAN values are replaced.

* The numerical\_features which is created already has the NAN values which is removed by replacing it with the Median values.



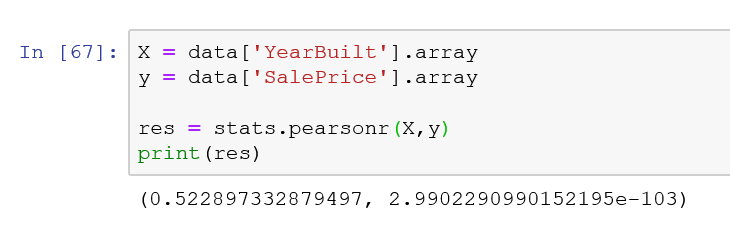
Hypothesis Testing :

* We have seen that the Year Build and the Sale price has a direct impact in the previous sections. So Im considering these two features for the hypothesis testing.

**H0:** The earlier built date of the house result in the lower house price  
**H1:**The earlier built date of the house not result in the lower house price.

The significance level is kept as 0.05 which is likely to have 95% confidence.

The pearson coefficient is given as 0.522897 and two tailed p value is : 2.99022 **so the p value which we assumed is very low so we are rejecting the H0.**



In the below scenario I have considered the feature OverallQual and the Sale Price.

Ho – Less the quality of the Quality of home more the price.

H1 – More the quality of home build more the sale price. From the p value which is shown we can reject the H0 hypothesis and accept the H1.

