HOUSING: PRICE PREDICTION

Submitted by:

SHAILAJ JOSHI

INTRODUCTION

- Problem Statement

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house

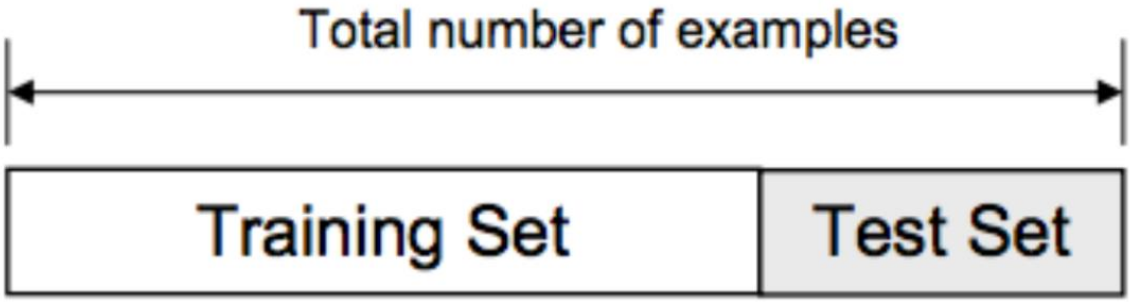
Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Train/Test Split:

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset



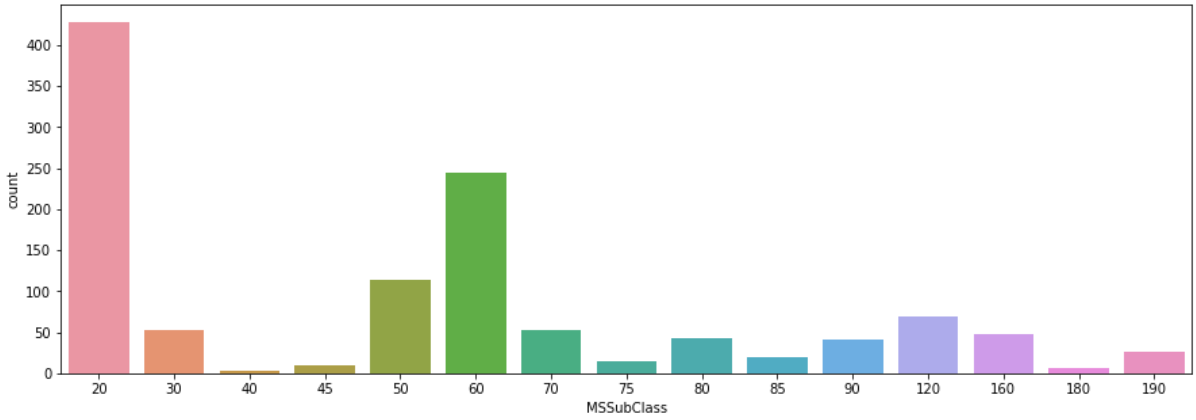
Using the random state, we divide the training and testing data in the ratio of 0.8 and 0.2 meaning 20% of the data is the testing data and training data is 80%

- Data Sources and their formats

MSSubClass: Identifies the type of dwelling involved in the sale.

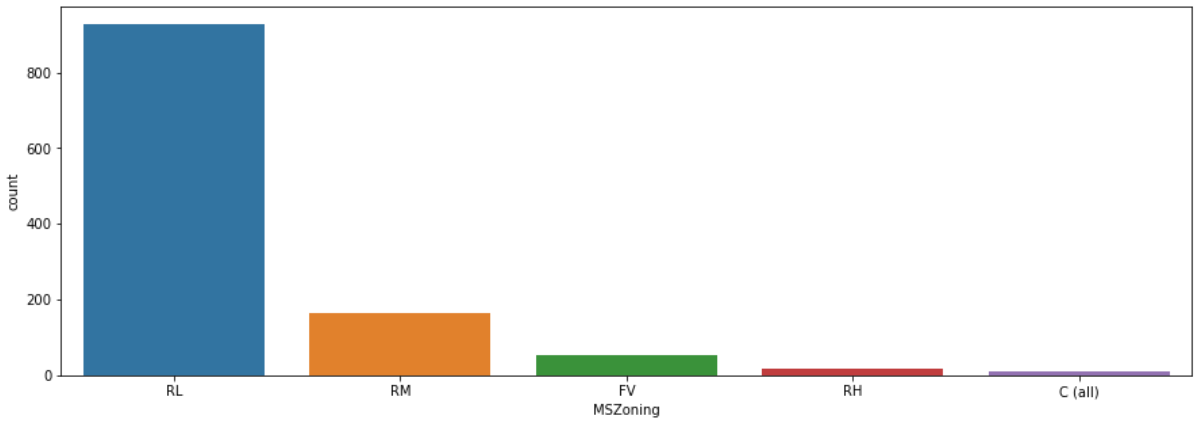
20 1-STORY 1946 & NEWER ALL STYLES
30 1-STORY 1945 & OLDER
40 1-STORY W/FINISHED ATTIC ALL AGES
45 1-1/2 STORY - UNFINISHED ALL AGES
50 1-1/2 STORY FINISHED ALL AGES
60 2-STORY 1946 & NEWER
70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES
80 SPLIT OR MULTI-LEVEL
85 SPLIT FOYER
90 DUPLEX - ALL STYLES AND AGES
120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150 1-1/2 STORY PUD - ALL AGES
160 2-STORY PUD - 1946 & NEWER
180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190 2 FAMILY CONVERSION - ALL STYLES AND AGES



MSZoning: Identifies the general zoning classification of the sale.

A Agriculture
C Commercial
FV Floating Village Residential
I Industrial
RH Residential High Density
RL Residential Low Density
RP Residential Low Density Park
RMResidential Medium Density

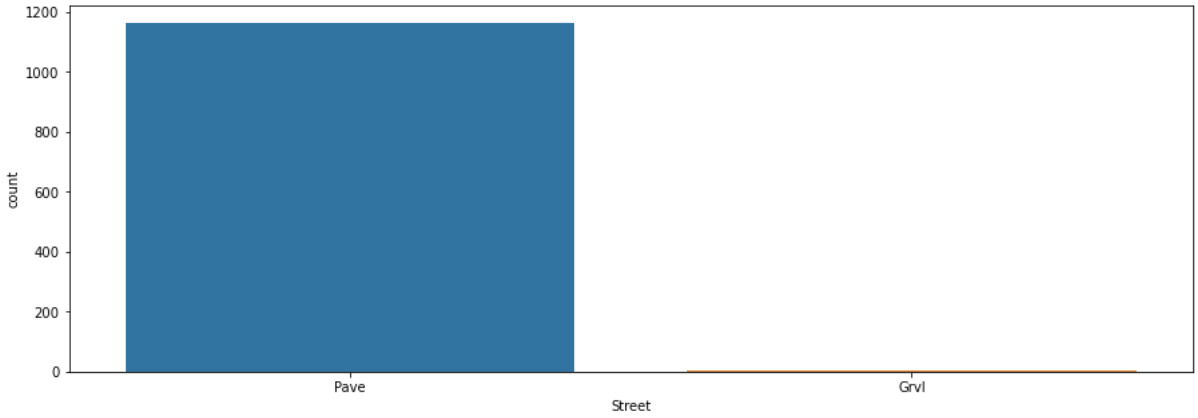


LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

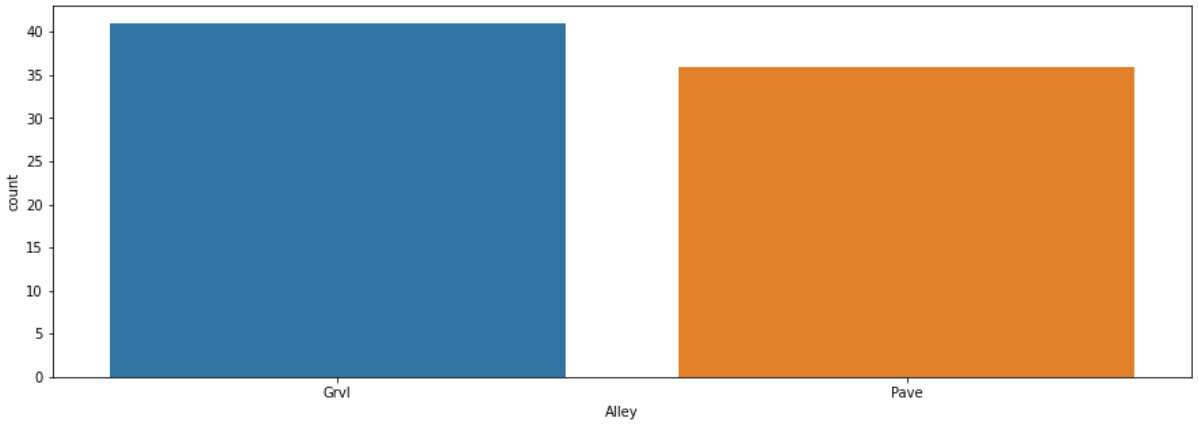
Street: Type of road access to property

Grvl	Gravel
Pave	Paved



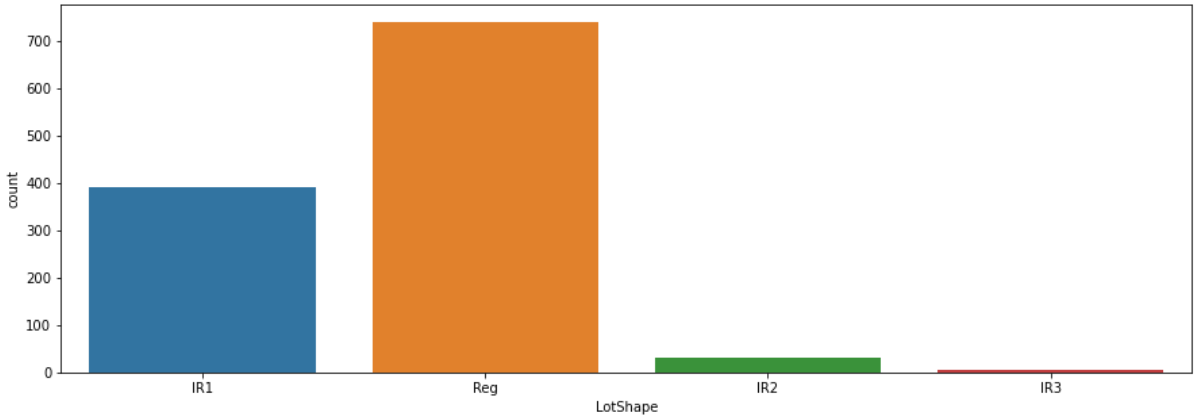
Alley: Type of alley access to property

Grvl	Gravel
Pave	Paved
NA	No alley access



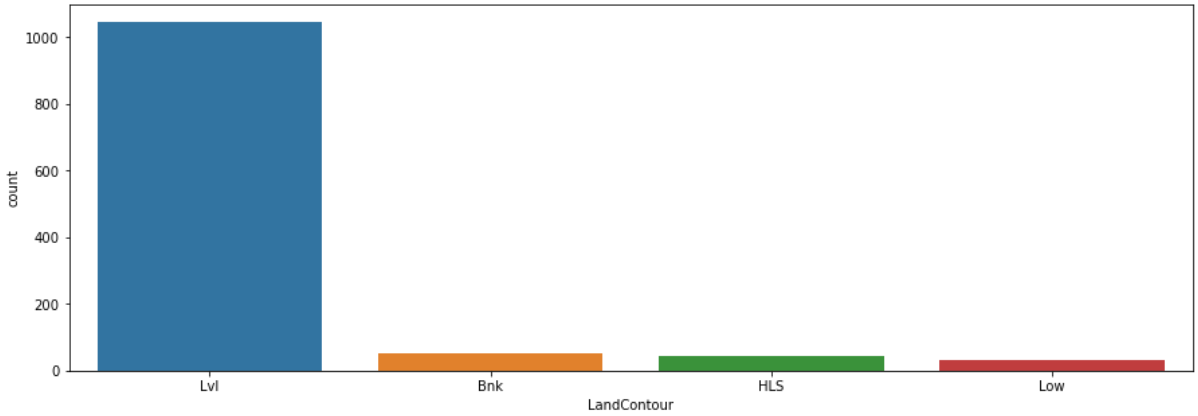
LotShape: General shape of property

Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular



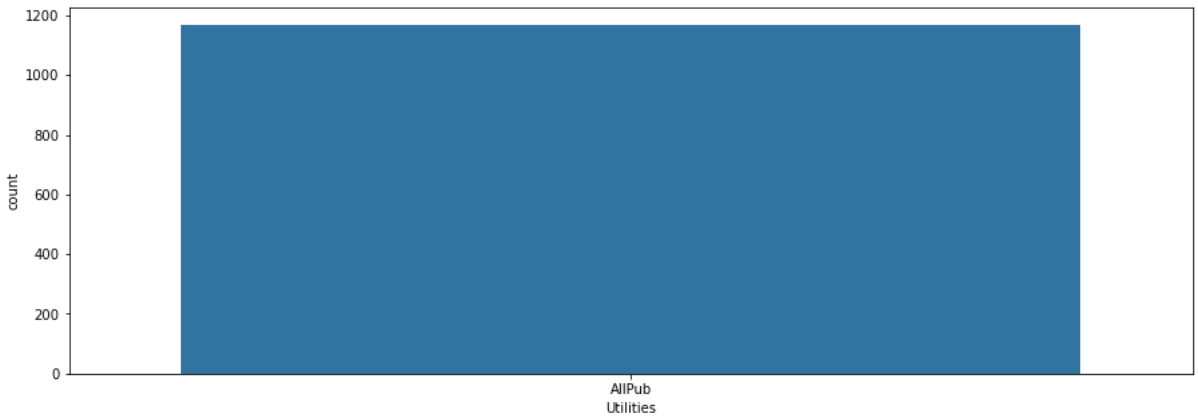
LandContour: Flatness of the property

Lvl	Near Flat/Level
Bnk	Banked - Quick and significant rise from street grade to building
HLS	Hillside - Significant slope from side to side
Low	Depression



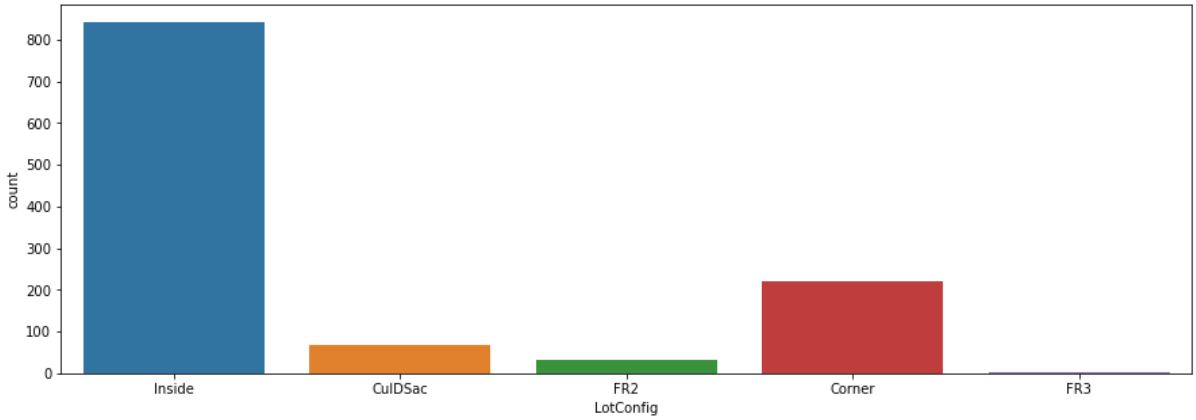
Utilities: Type of utilities available

AllPub	All public Utilities (E,G,W,& S)
NoSewr	Electricity, Gas, and Water (Septic Tank)
NoSeWa	Electricity and Gas Only
ELO	Electricity only



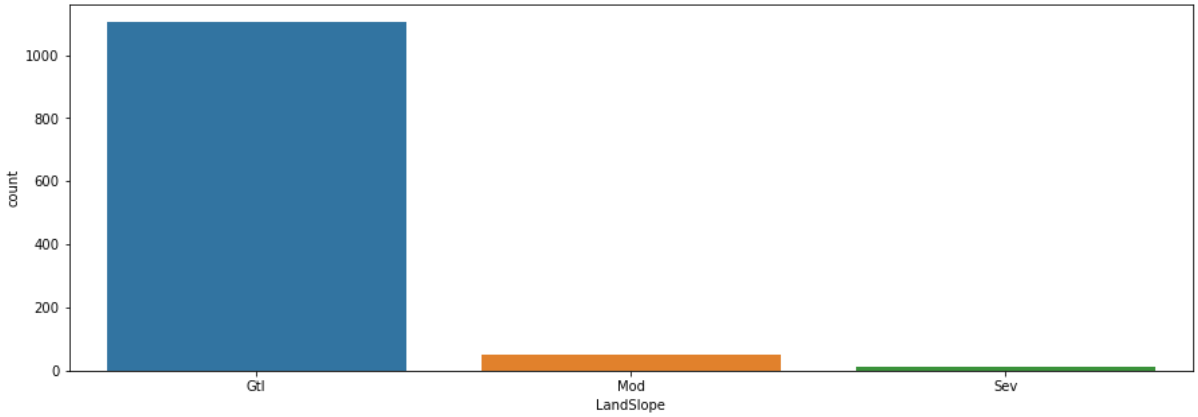
LotConfig: Lot configuration

Inside Inside lot
Corner Corner lot
CulDSac Cul-de-sac
FR2 Frontage on 2 sides of property
FR3 Frontage on 3 sides of property



LandSlope: Slope of property

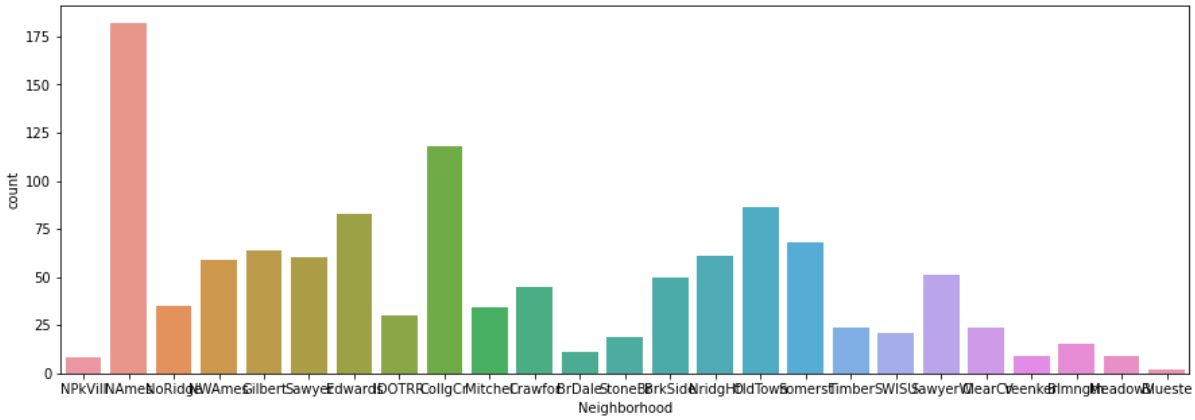
Gtl Gentle slope
Mod Moderate Slope
Sev Severe Slope



Neighborhood: Physical locations within Ames city limits

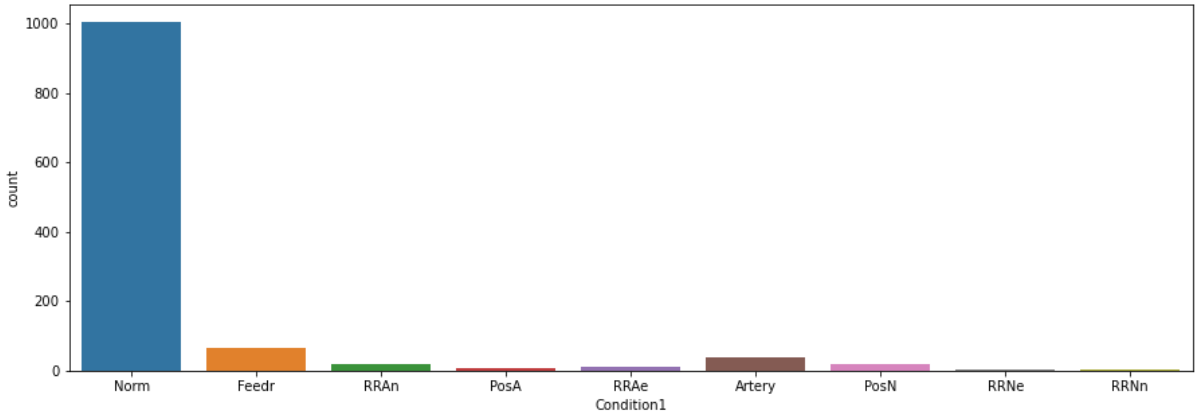
Blmngtn Bloomington Heights
Blueste Bluestem
BrDale Briardale
BrkSide Brookside
ClearCr Clear Creek
CollgCr College Creek
Crawfor Crawford
Edwards Edwards

Gilbert Gilbert
IDOTRR Iowa DOT and Rail Road
MeadowV Meadow Village
Mitchel Mitchell
Names North Ames
NoRidge Northridge
NPkVill Northpark Villa
NridgHt Northridge Heights
NWAmes Northwest Ames
OldTown Old Town
SWISU South & West of Iowa State University
SawyerSawyer
SawyerW Sawyer West
Somerst Somerset
StoneBr Stone Brook
Timber Timberland
Veenker Veenker



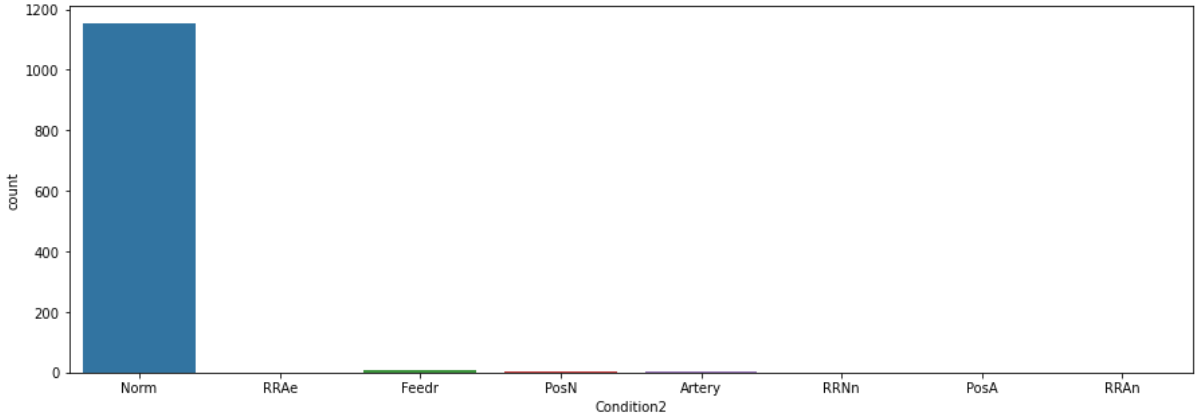
Condition1: Proximity to various conditions

Artery Adjacent to arterial street
Feedr Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad



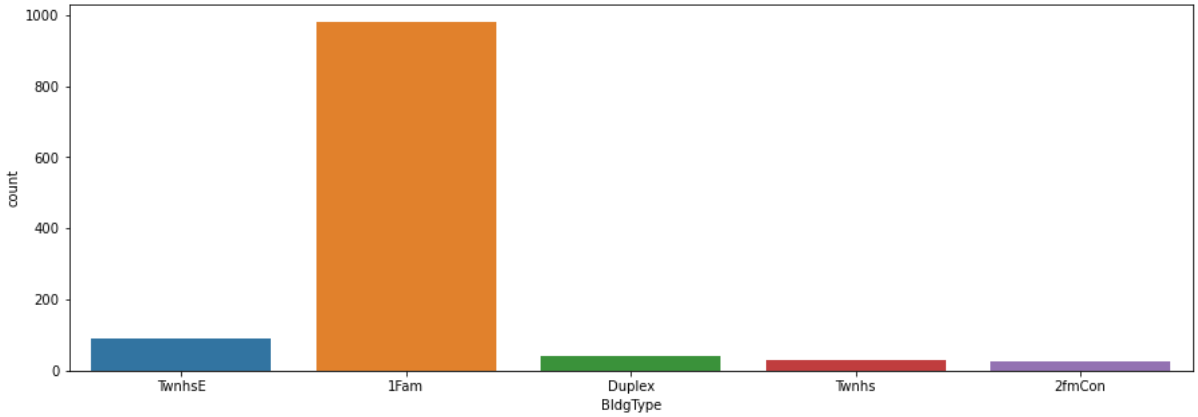
Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street
Feedr Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad



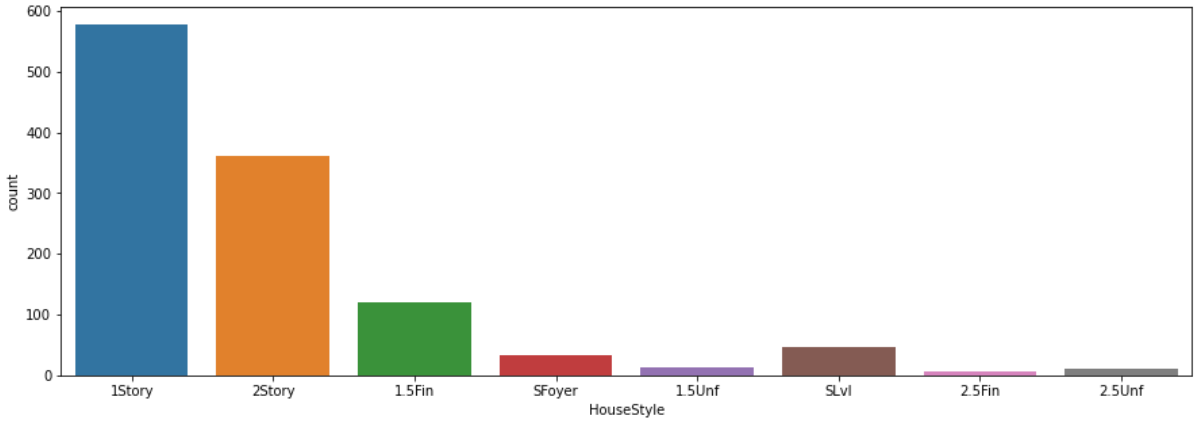
BldgType: Type of dwelling

1Fam Single-family Detached
2FmCon Two-family Conversion; originally built as one-family dwelling
Duplx Duplex
TwnhsE Townhouse End Unit
TwnhsI Townhouse Inside Unit



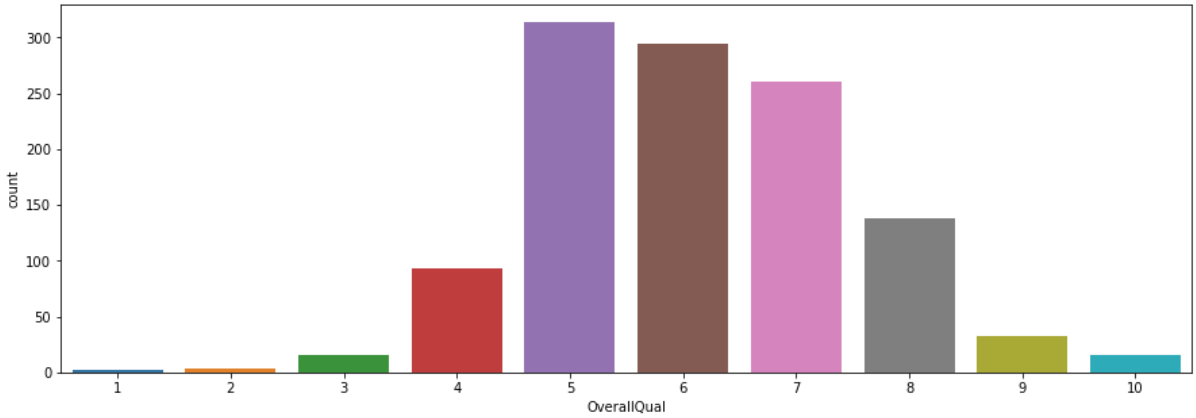
HouseStyle: Style of dwelling

1Story One story
1.5Fin One and one-half story: 2nd level finished
1.5Unf One and one-half story: 2nd level unfinished
2Story Two story
2.5Fin Two and one-half story: 2nd level finished
2.5Unf Two and one-half story: 2nd level unfinished
SFoyer Split Foyer
SLvl Split Level



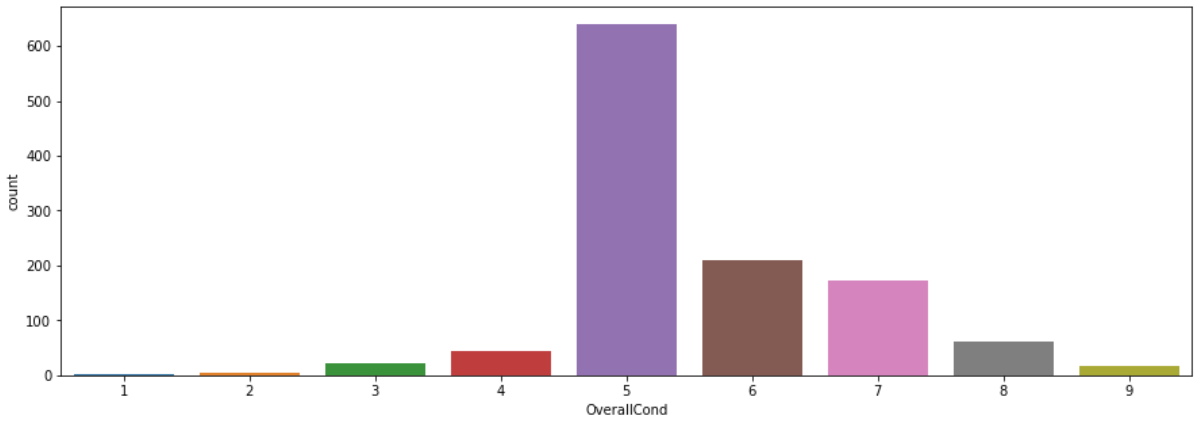
OverallQual: Rates the overall material and finish of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor



OverallCond: Rates the overall condition of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor



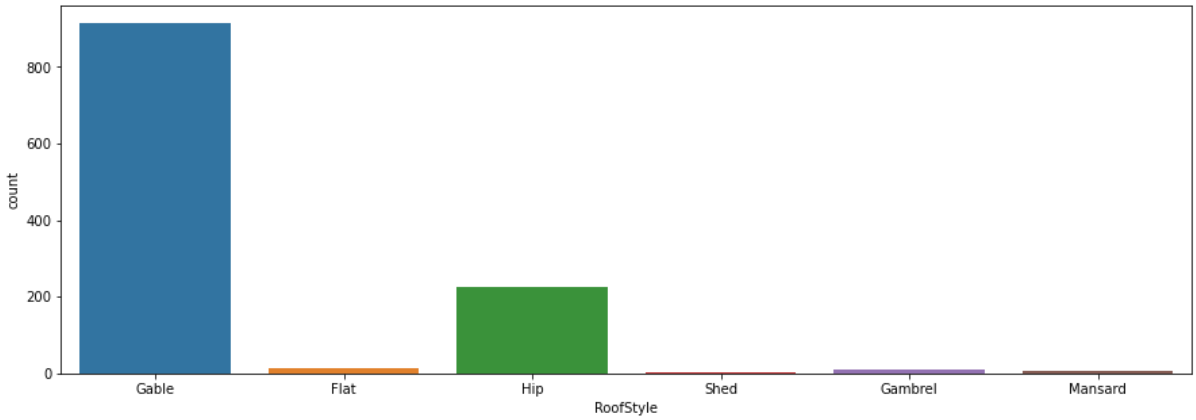
YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

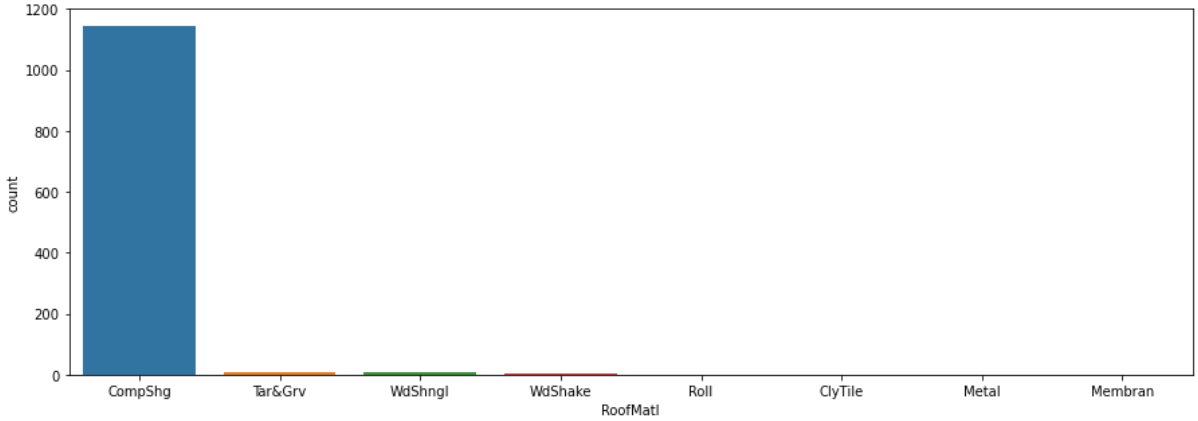
Flat Flat
Gable Gable
Gambrel Gabrel (Barn)
Hip Hip
Mansard Mansard

Shed Shed



RoofMatl: Roof material

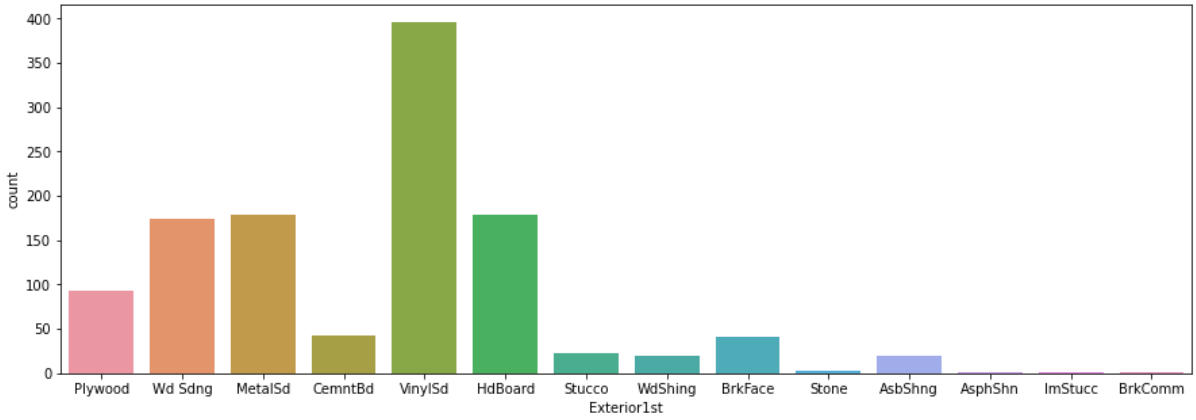
ClyTile Clay or Tile
CompShg Standard (Composite) Shingle
Membran Membrane
Metal Metal
Roll Roll
Tar&Grv Gravel & Tar
WdShake Wood Shakes
WdShngl Wood Shingles



Exterior1st: Exterior covering on house

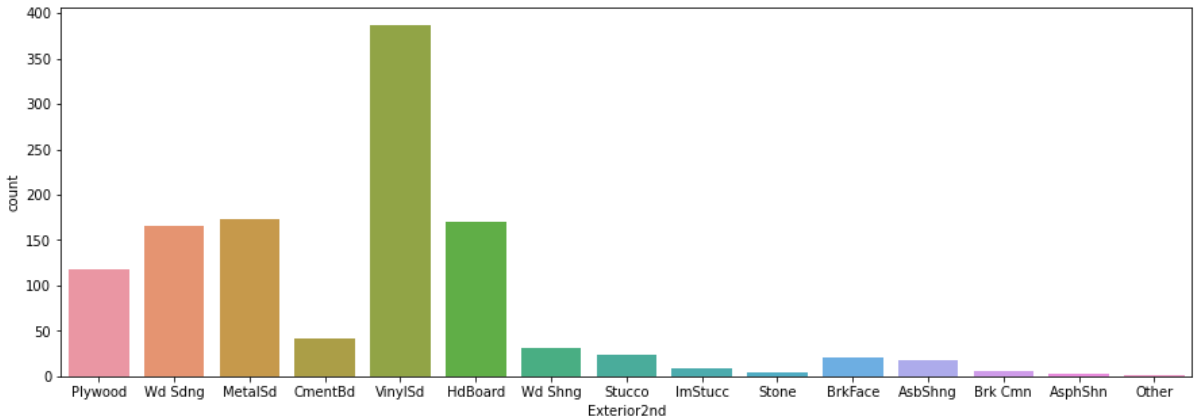
AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast

Stone Stone
Stucco Stucco
VinylSdVinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles



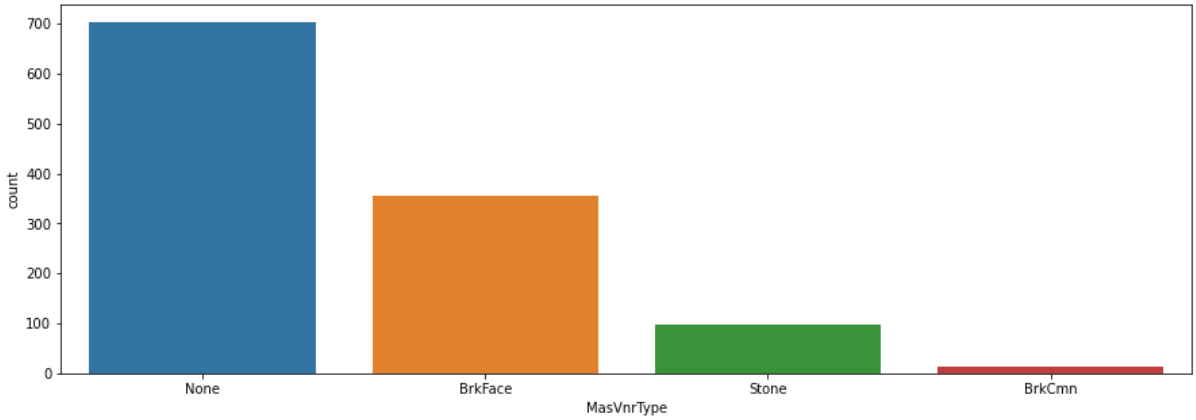
Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSdVinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles



MasVnrType: Masonry veneer type

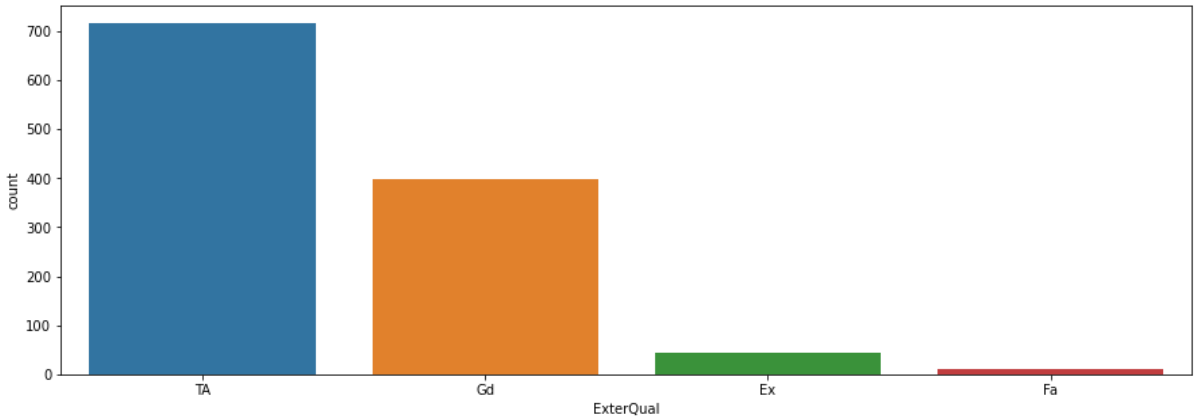
BrkCmn Brick Common
BrkFace Brick Face
CBlock Cinder Block
None None
Stone Stone



MasVnrArea: Masonry veneer area in square feet

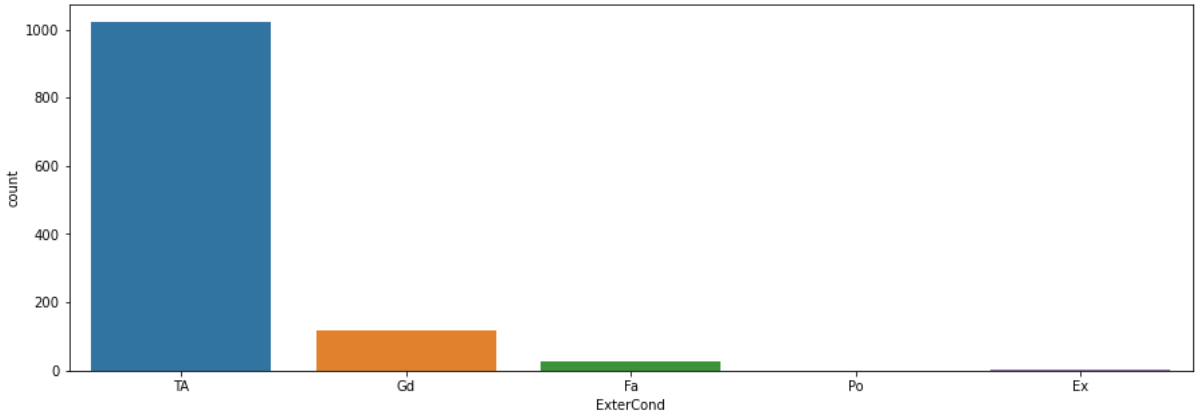
ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor



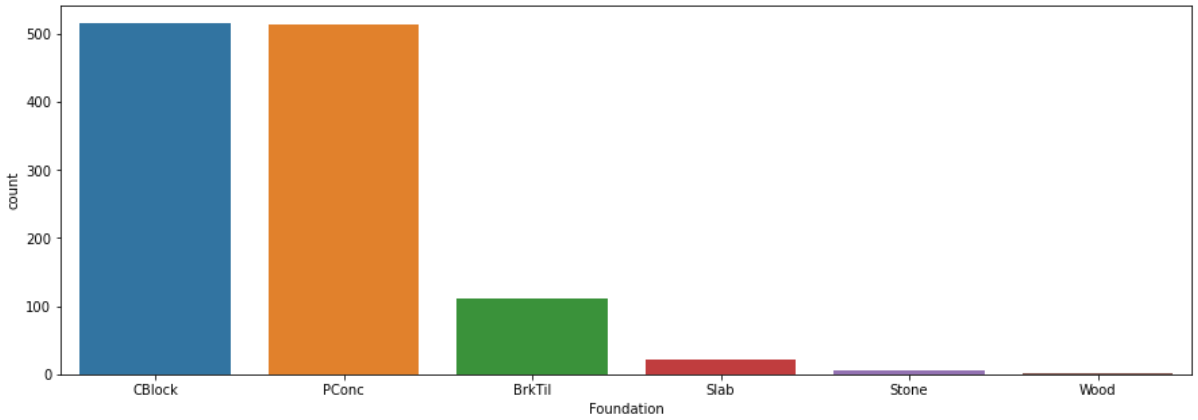
ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor



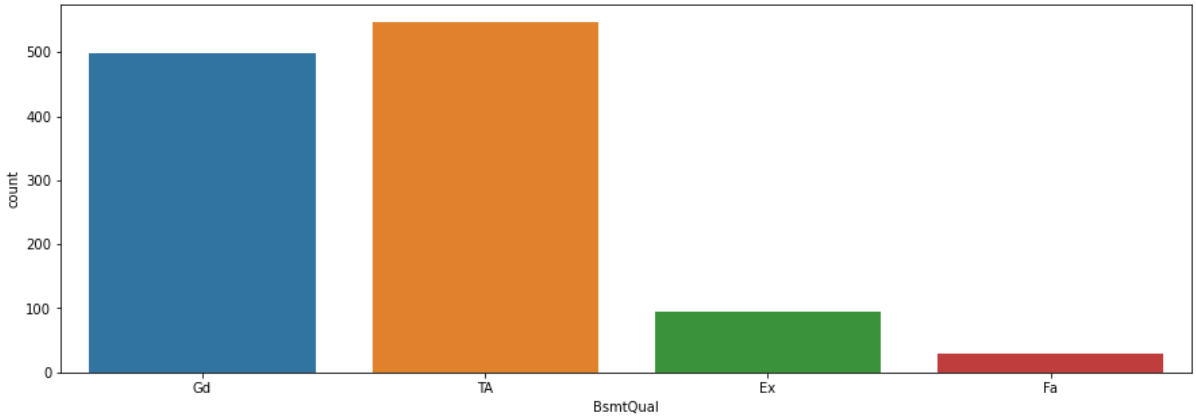
Foundation: Type of foundation

BrkTil Brick & Tile
CBlock Cinder Block
PConc Poured Contrete
Slab Slab
Stone Stone
Wood Wood



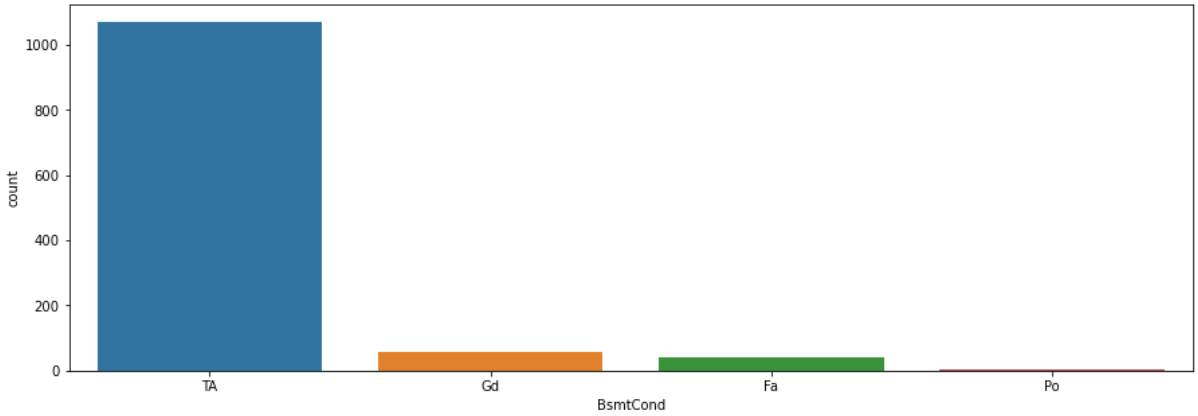
BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)
Gd Good (90-99 inches)
TA Typical (80-89 inches)
Fa Fair (70-79 inches)
Po Poor (<70 inches
NA No Basement



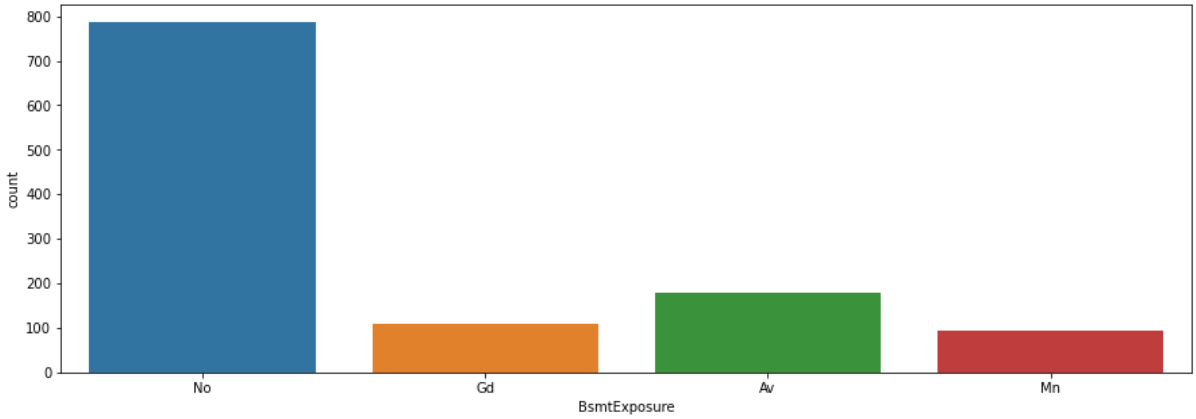
BsmtCond: Evaluates the general condition of the basement

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement



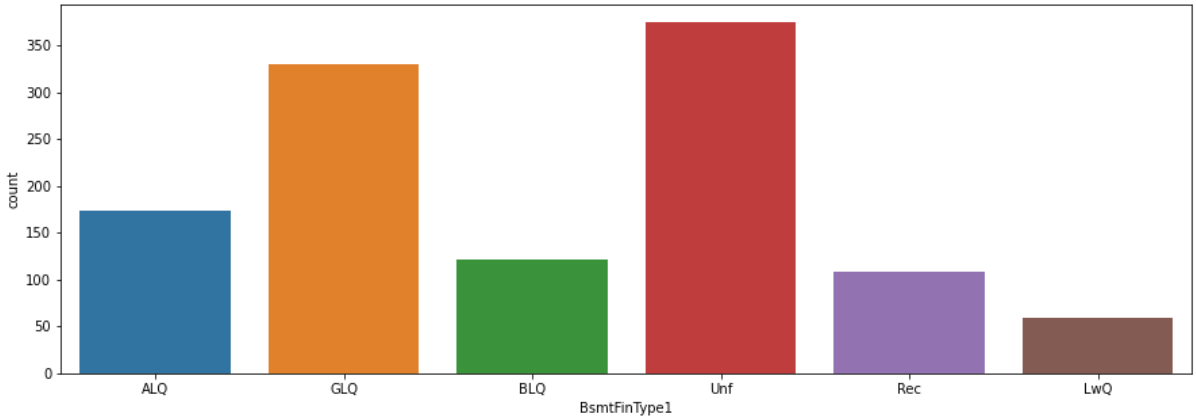
BsmtExposure: Refers to walkout or garden level walls

Gd	Good Exposure
Av	Average Exposure (split levels or foyers typically score average or above)
Mn	Mimimum Exposure
No	No Exposure
NA	No Basement



BsmtFinType1: Rating of basement finished area

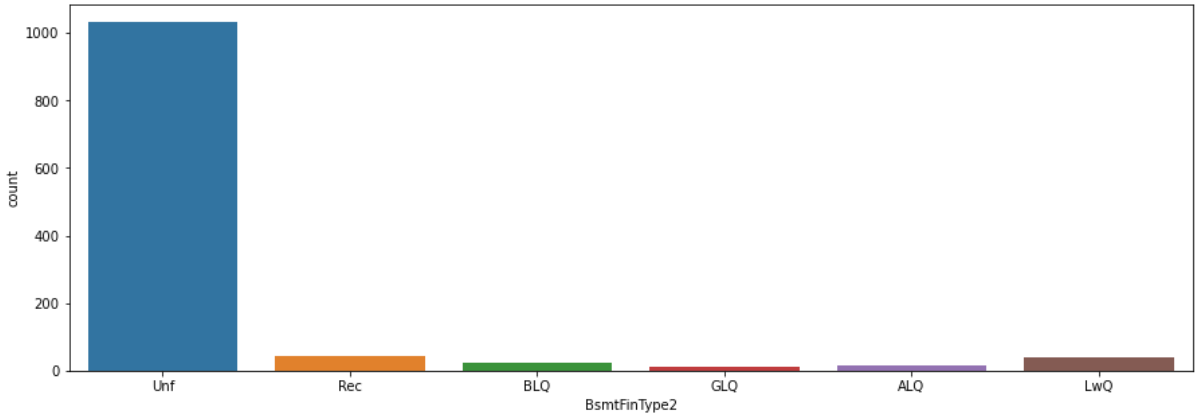
GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement



BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ	Good Living Quarters
ALQ	Average Living Quarters
BLQ	Below Average Living Quarters
Rec	Average Rec Room
LwQ	Low Quality
Unf	Unfinshed
NA	No Basement



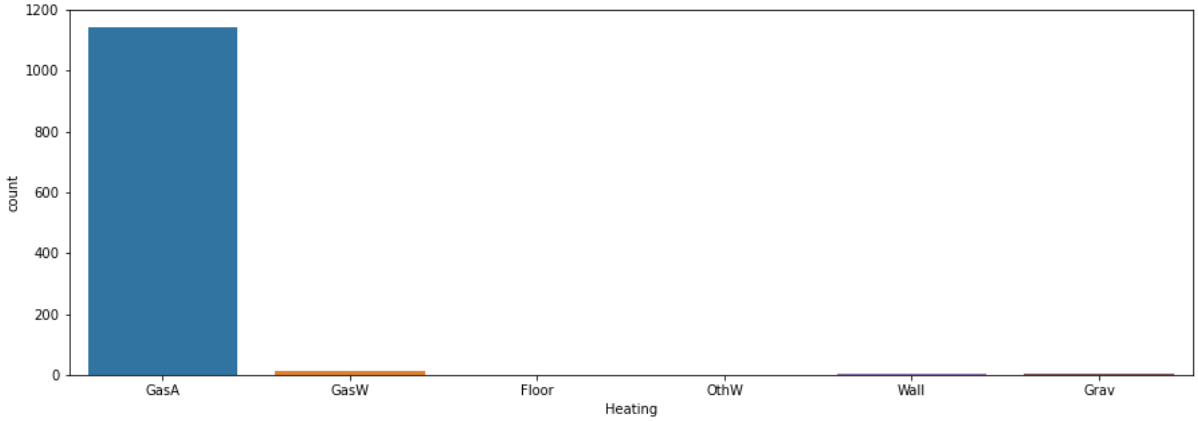
BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

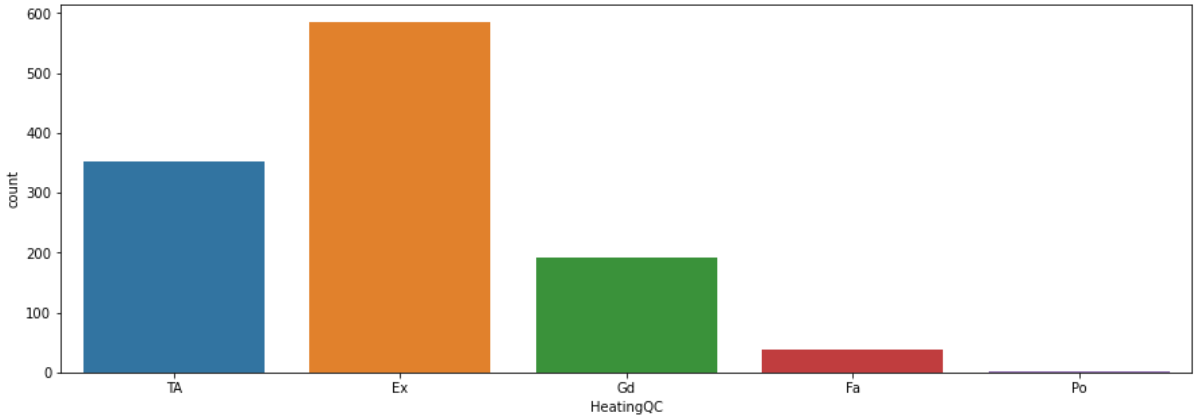
Heating: Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace



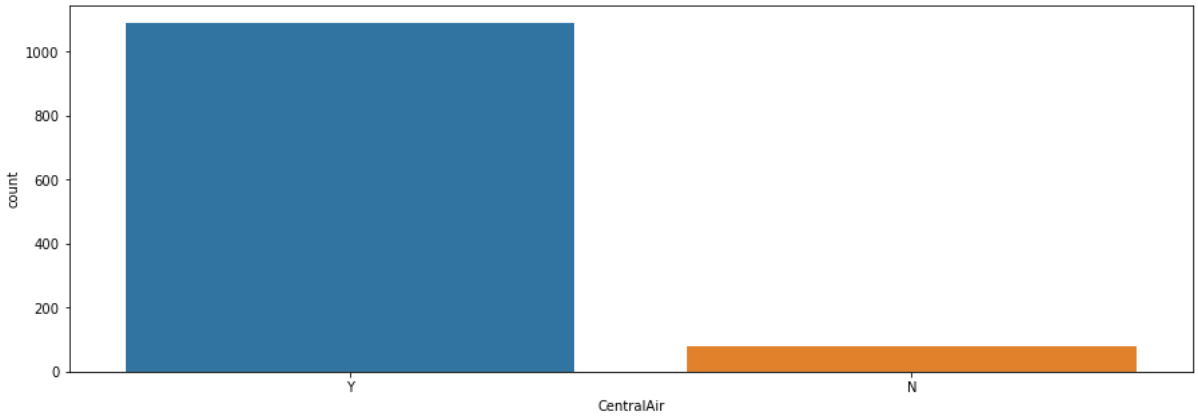
HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor



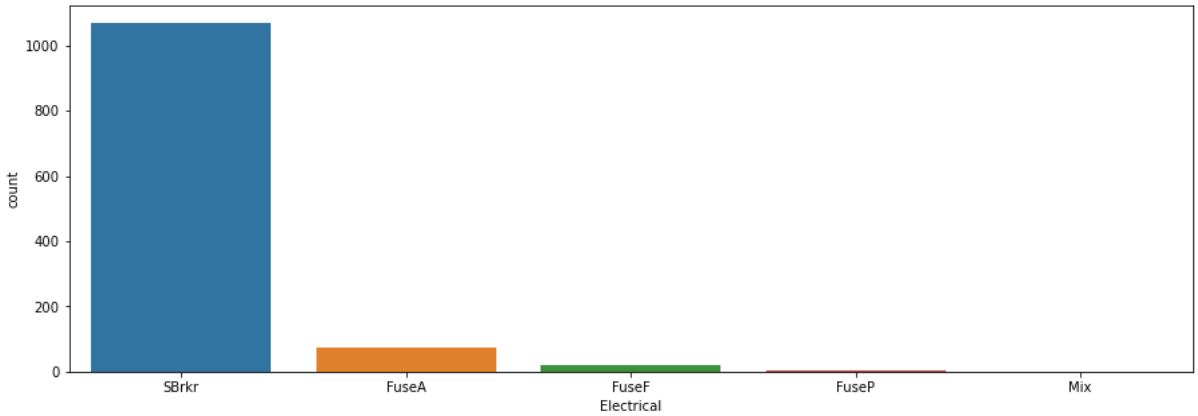
CentralAir: Central air conditioning

N No
Y Yes



Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex
FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix Mixed



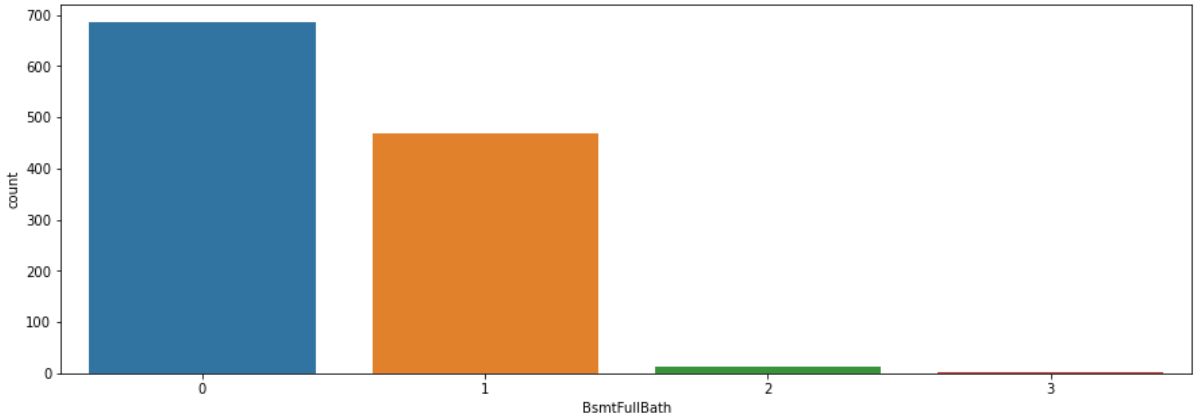
1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

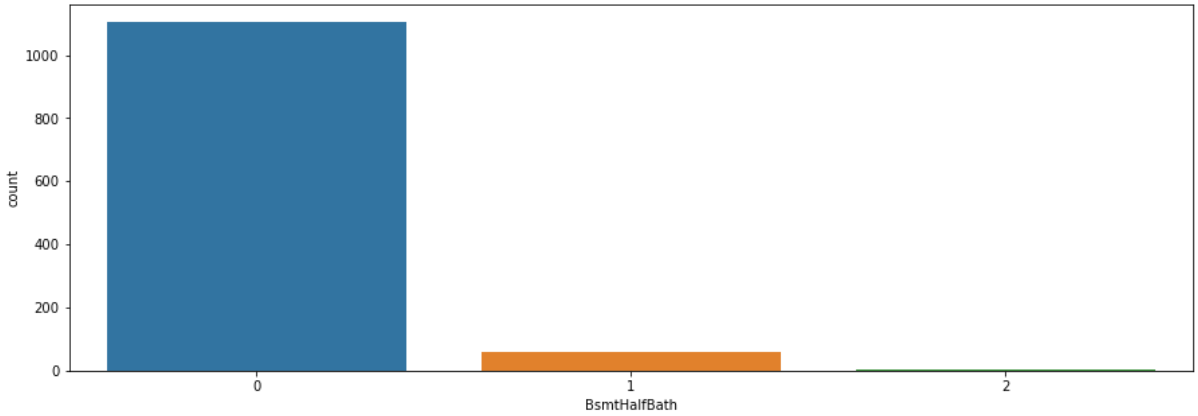
LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

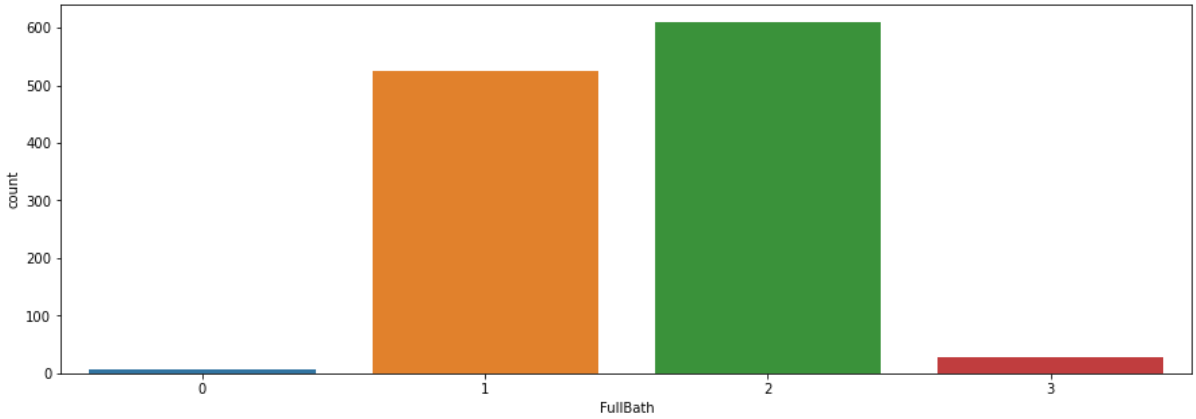
BsmtFullBath: Basement full bathrooms



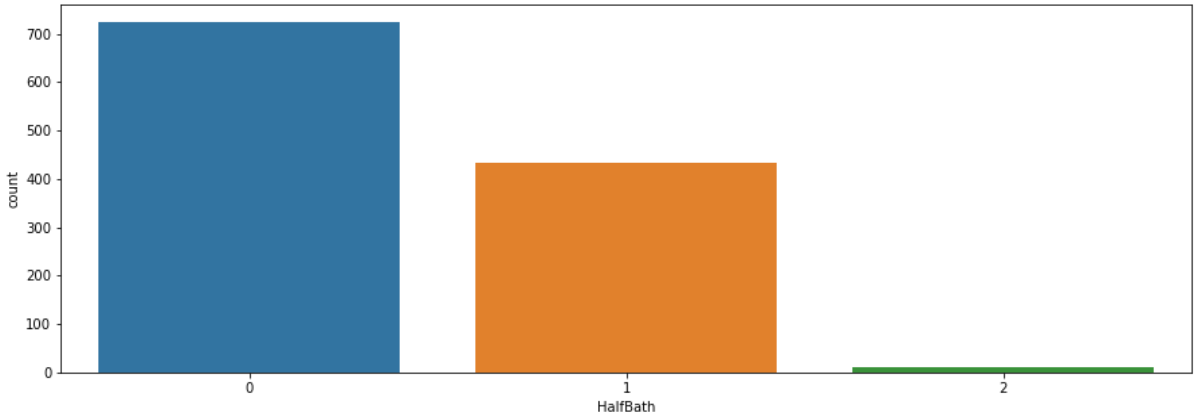
BsmtHalfBath: Basement half bathrooms



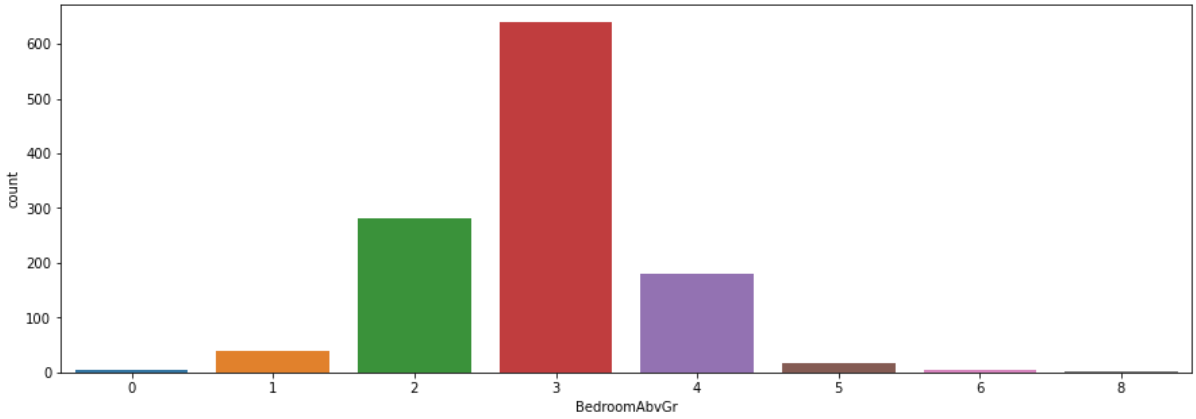
FullBath: Full bathrooms above grade



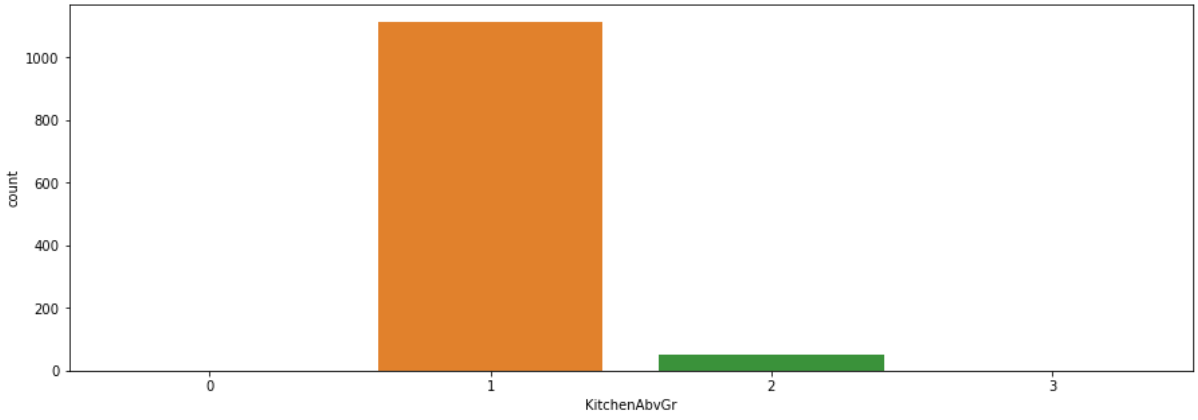
HalfBath: Half baths above grade



Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

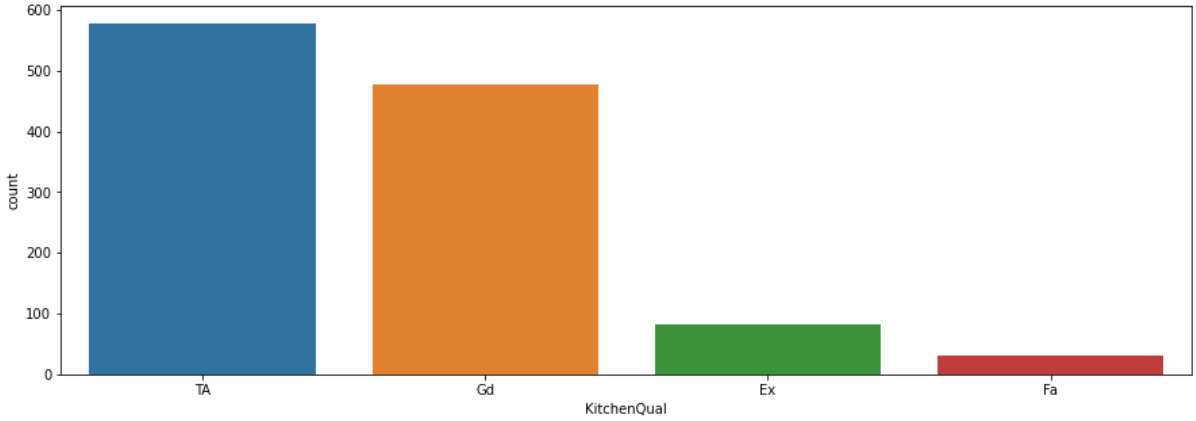


Kitchen: Kitchens above grade

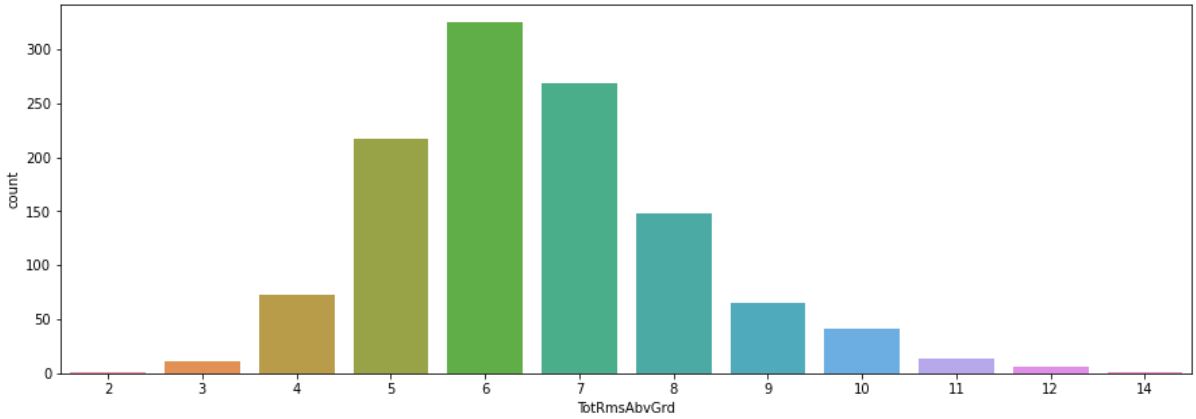


KitchenQual: Kitchen quality

Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair
Po	Poor

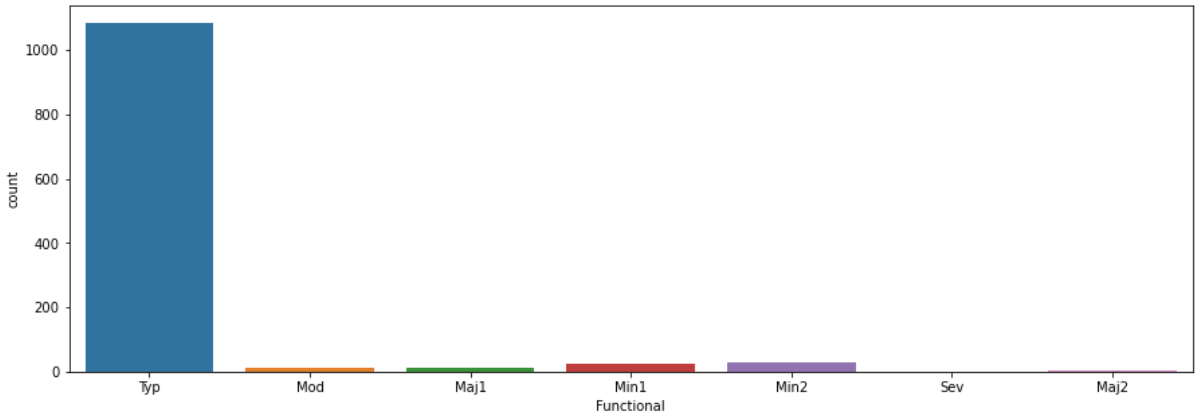


TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

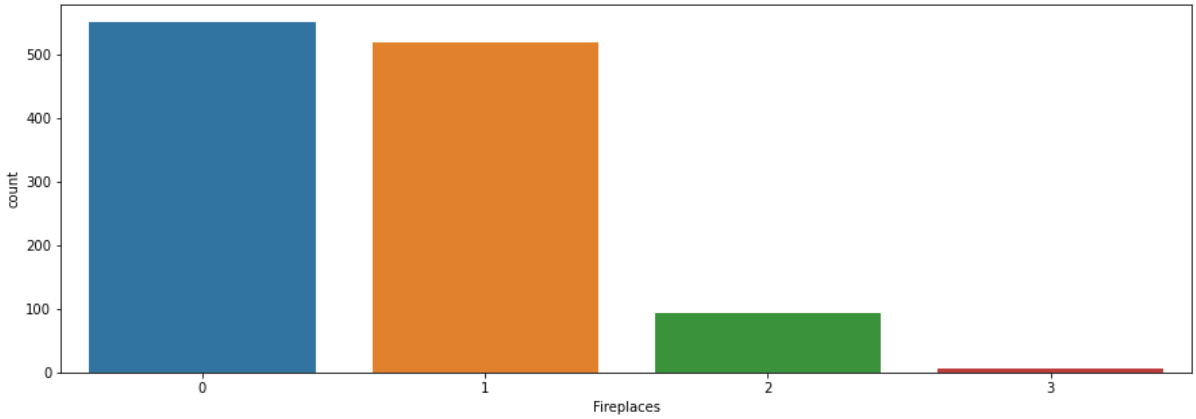


Functional: Home functionality (Assume typical unless deductions are warranted)

Typ	Typical Functionality
Min1	Minor Deductions 1
Min2	Minor Deductions 2
Mod	Moderate Deductions
Maj1	Major Deductions 1
Maj2	Major Deductions 2
Sev	Severely Damaged
Sal	Salvage only

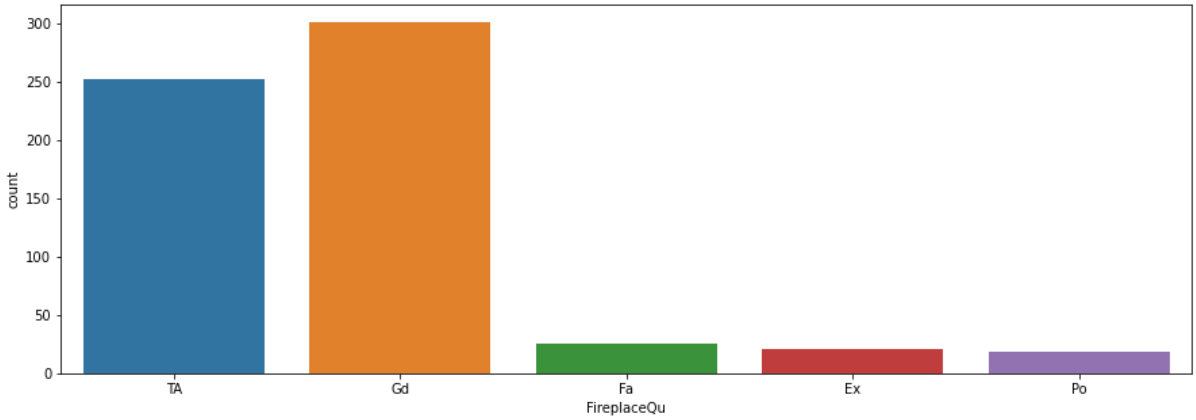


Fireplaces: Number of fireplaces



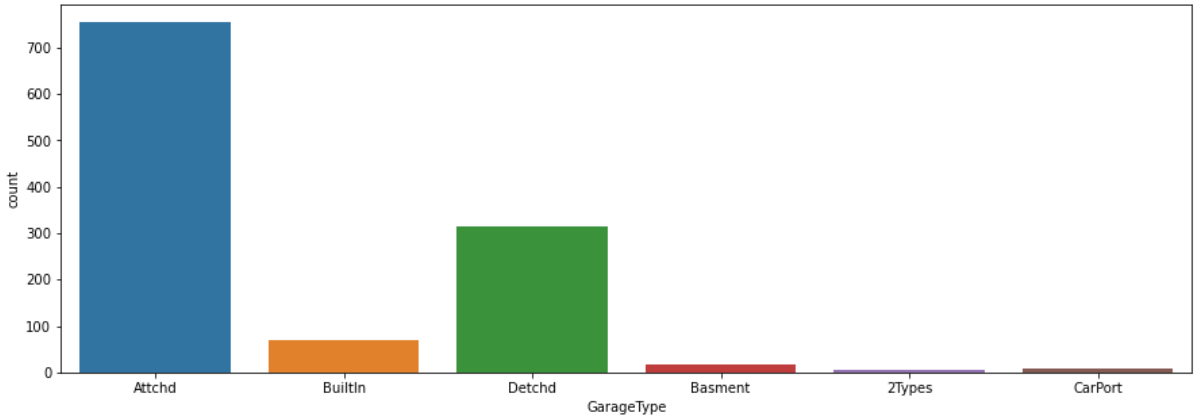
FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace



GarageType: Garage location

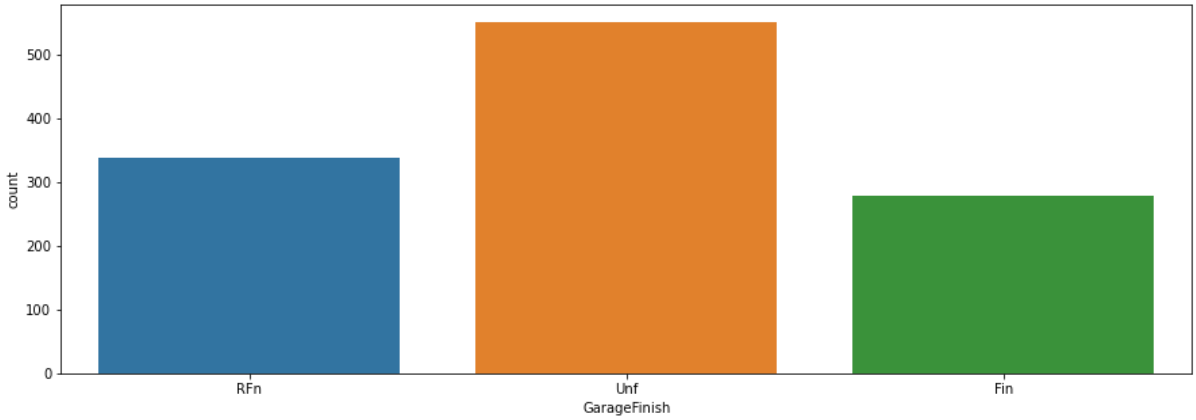
2Types More than one type of garage
Attchd Attached to home
Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port
Detchd Detached from home
NA No Garage



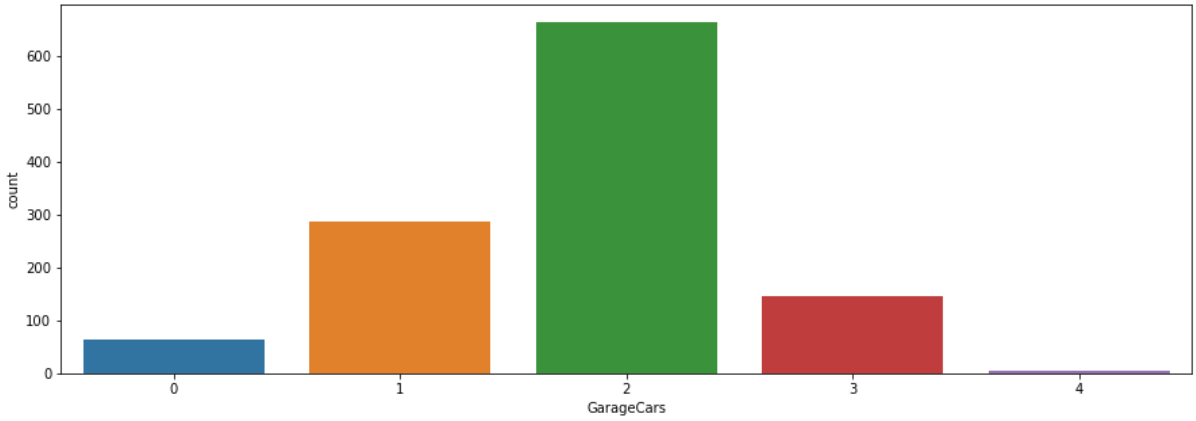
GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage



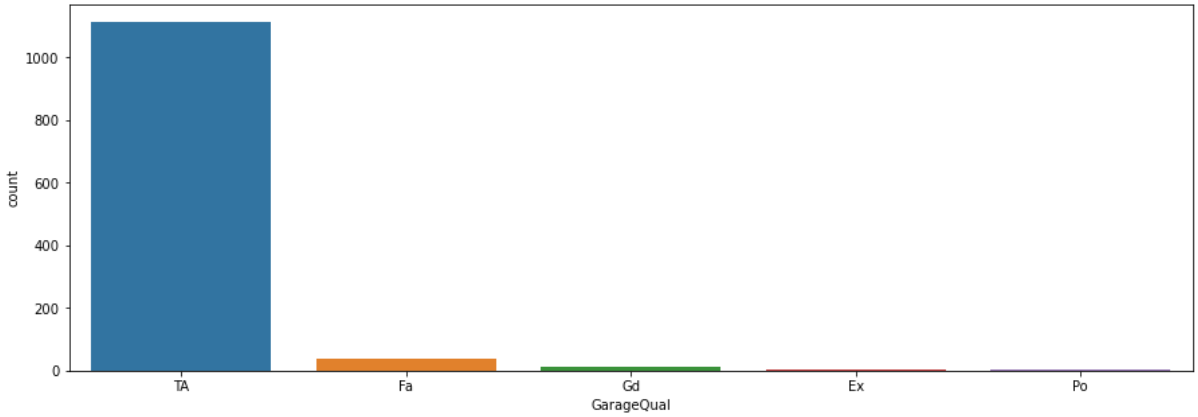
GarageCars: Size of garage in car capacity



GarageArea: Size of garage in square feet

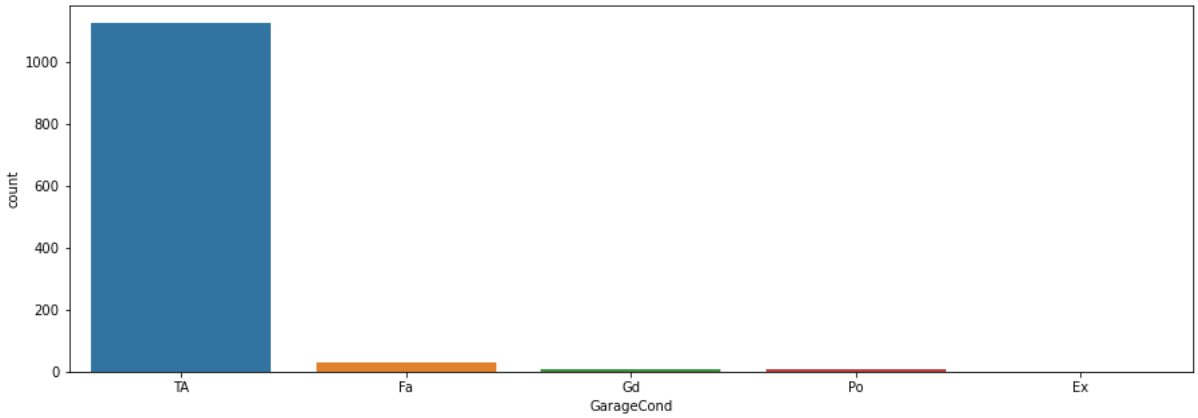
GarageQual: Garage quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage



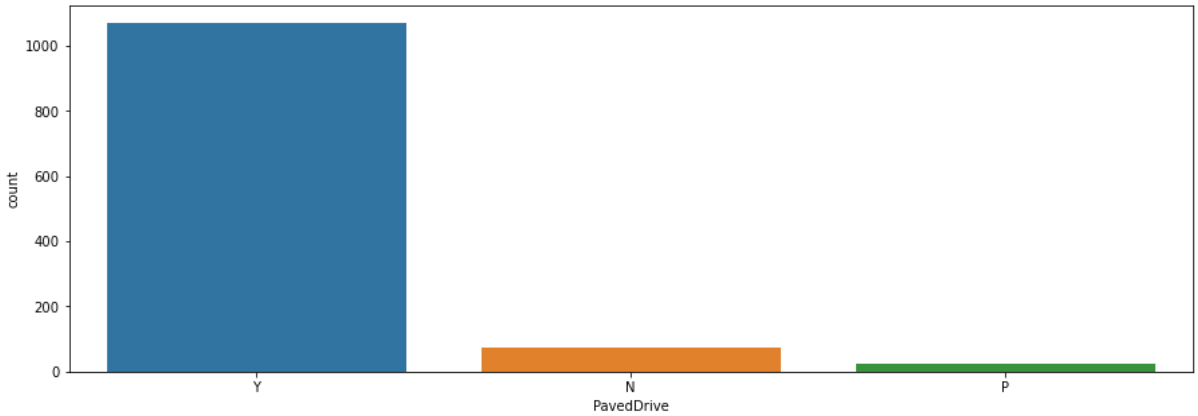
GarageCond: Garage condition

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage



PavedDrive: Paved driveway

Y Paved
P Partial Pavement
N Dirt/Gravel

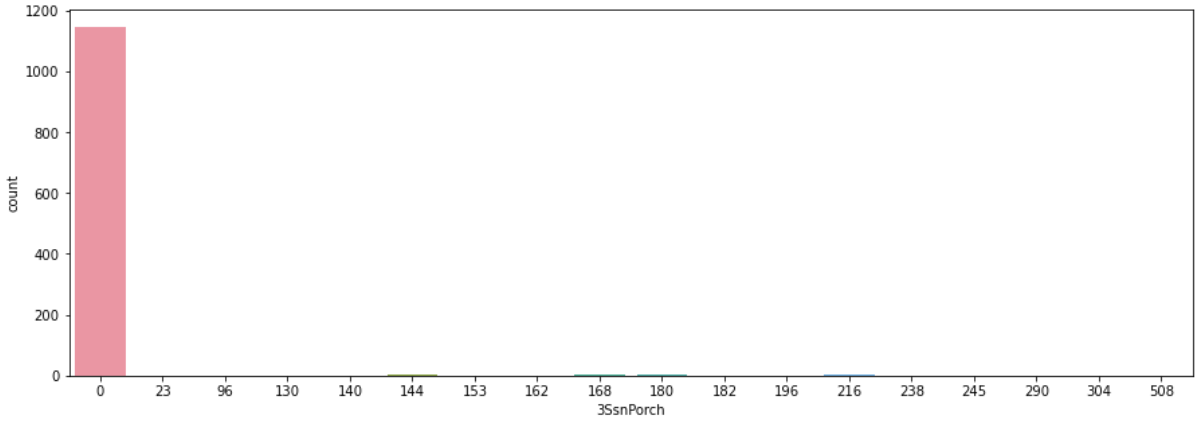


WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

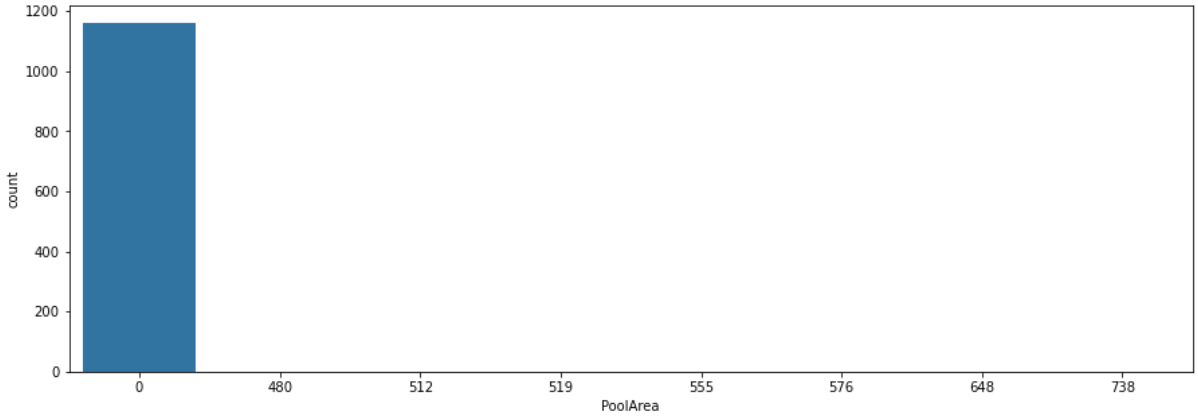
EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet



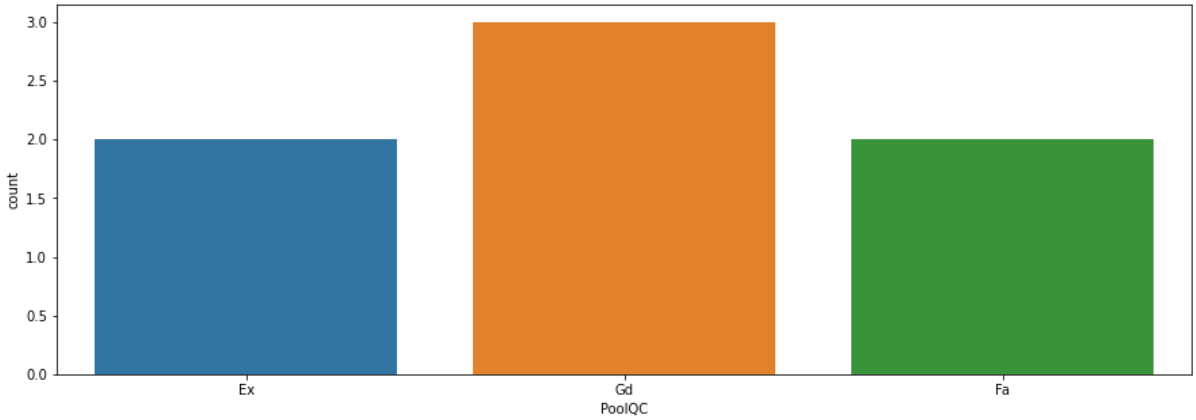
ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet



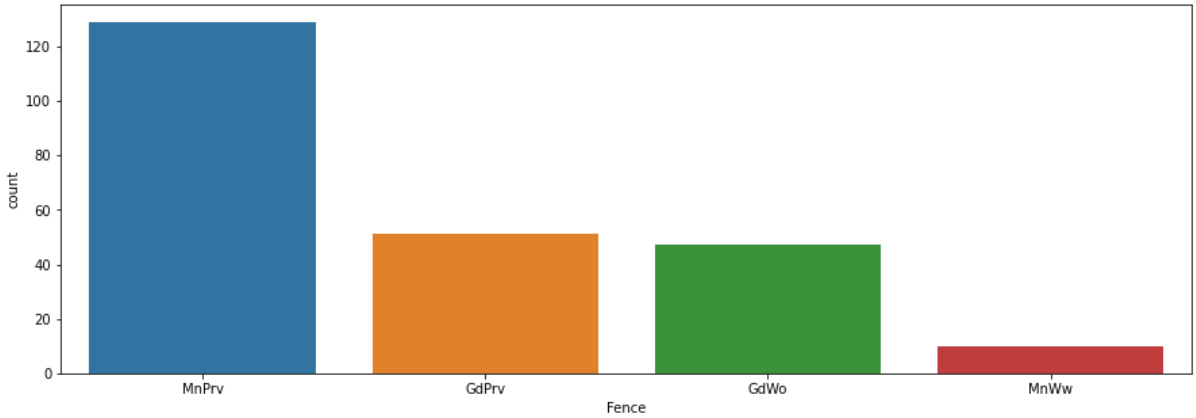
PoolQC: Pool quality

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool



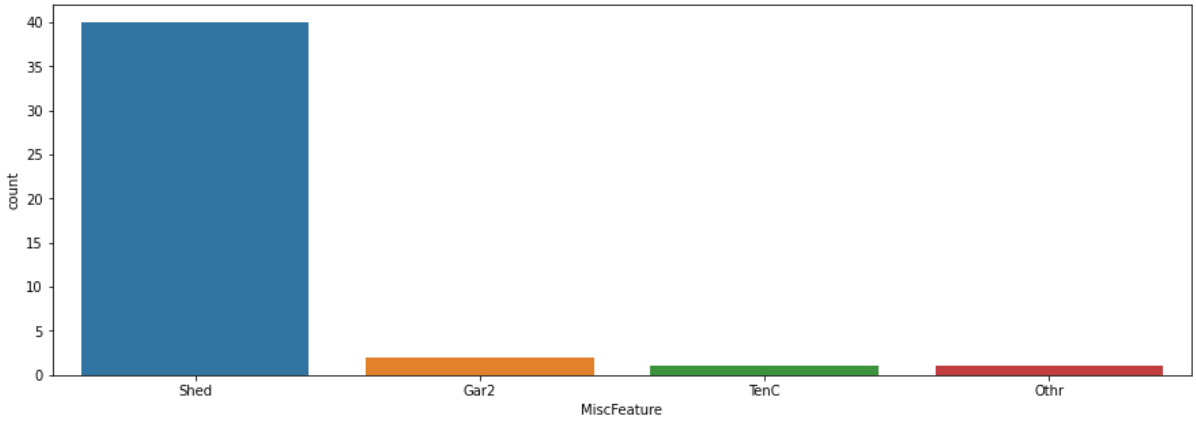
Fence: Fence quality

GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence



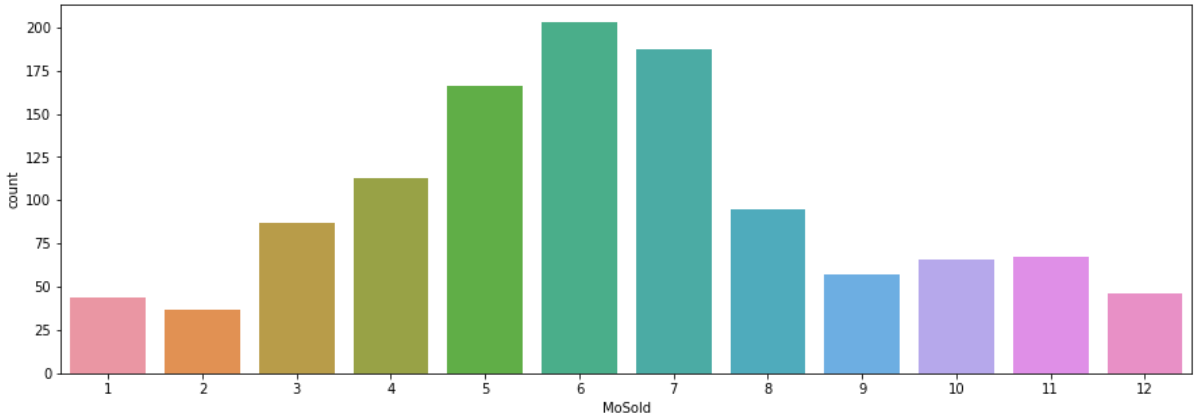
MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court
NA None

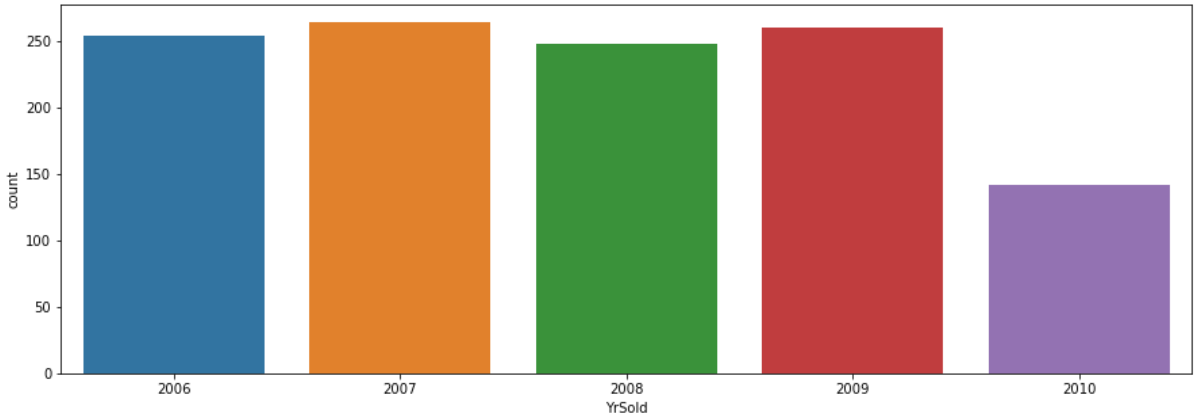


MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)



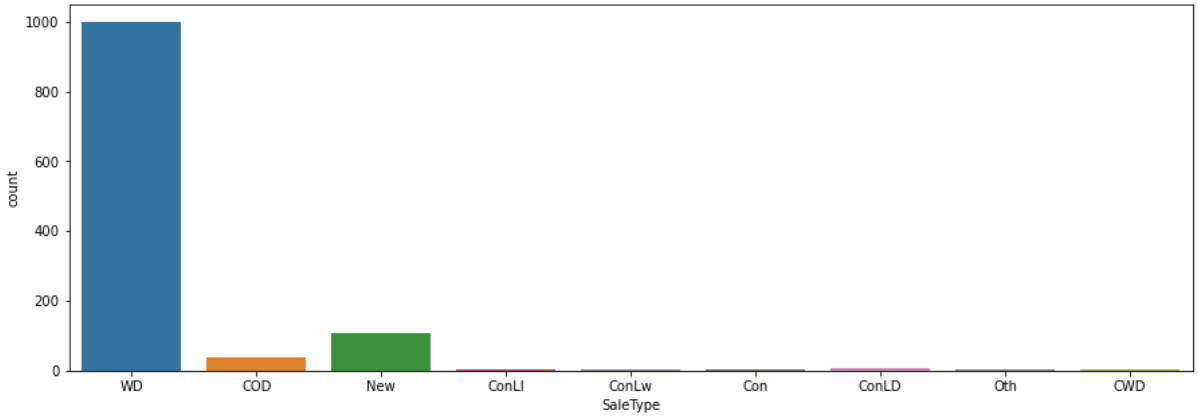
YrSold: Year Sold (YYYY)



SaleType: Type of sale

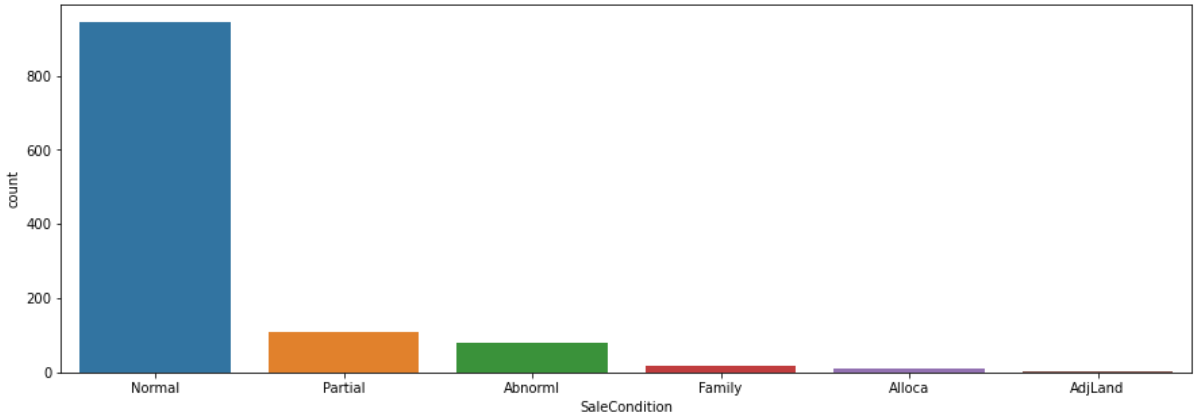
WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest

ConLI Contract Low Interest
ConLD Contract Low Down
Oth Other



SaleCondition: Condition of sale

Normal Normal Sale
Abnorml Abnormal Sale - trade, foreclosure, short sale
AdjLand Adjoining Land Purchase
Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family Sale between family members
Partial Home was not completed when last assessed (associated with New Homes)



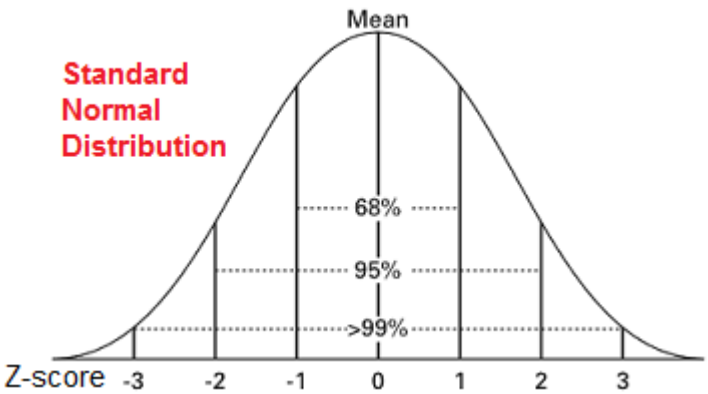
- Data Pre-processing Done

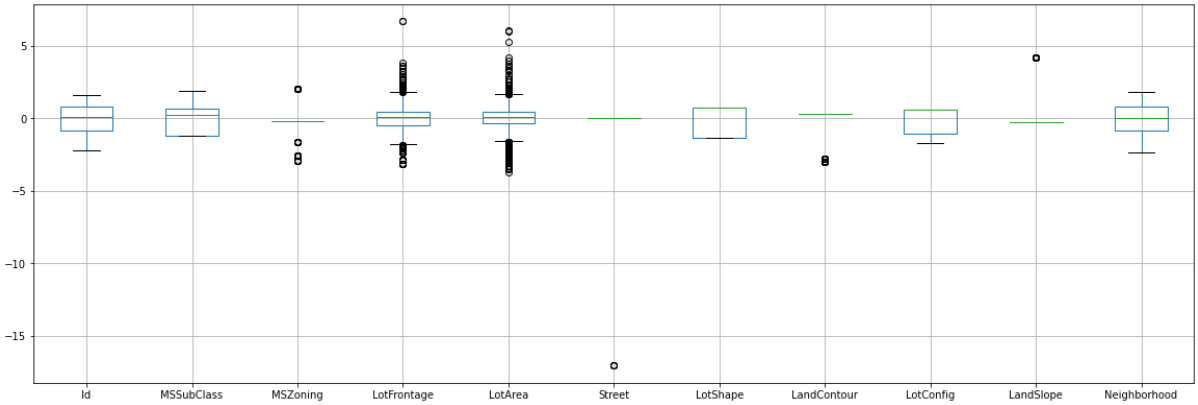
Column/Independent Variables	Null Values	Data Type
Id	0	int64
MSSubClass	0	int64
MSZoning	0	object
LotFrontage	214	float64
LotArea	0	int64
Street	0	object
Alley	1091	object
LotShape	0	object
LandContour	0	object
Utilities	0	object
LotConfig	0	object
LandSlope	0	object
Neighborhood	0	object
Condition1	0	object
Condition2	0	object
BldgType	0	object
HouseStyle	0	object
OverallQual	0	int64
OverallCond	0	int64
YearBuilt	0	int64

YearRemodAdd	0	int64
RoofStyle	0	object
RoofMatl	0	object
Exterior1st	0	object
Exterior2nd	0	object
MasVnrType	7	object
MasVnrArea	7	float64
ExterQual	0	object
ExterCond	0	object
Foundation	0	object
BsmtQual	30	object
BsmtCond	30	object
BsmtExposure	31	object
BsmtFinType1	30	object
BsmtFinSF1	0	int64
BsmtFinType2	31	object
BsmtFinSF2	0	int64
BsmtUnfSF	0	int64
TotalBsmtSF	0	int64
Heating	0	object
HeatingQC	0	object
CentralAir	0	object
Electrical	0	object
1stFlrSF	0	int64
2ndFlrSF	0	int64
LowQualFinSF	0	int64
GrLivArea	0	int64
BsmtFullBath	0	int64
BsmtHalfBath	0	int64
FullBath	0	int64
HalfBath	0	int64
BedroomAbvGr	0	int64
KitchenAbvGr	0	int64
KitchenQual	0	object
TotRmsAbvGrd	0	int64
Functional	0	object
Fireplaces	0	int64
FireplaceQu	551	object
GarageType	64	object
GarageYrBlt	64	float64
GarageFinish	64	object
GarageCars	0	int64
GarageArea	0	int64
GarageQual	64	object
GarageCond	64	object

PavedDrive	0	object
WoodDeckSF	0	int64
OpenPorchSF	0	int64
EnclosedPorch	0	int64
3SsnPorch	0	int64
ScreenPorch	0	int64
PoolArea	0	int64
PoolQC	1161	object
Fence	931	object
MiscFeature	1124	object
MiscVal	0	int64
MoSold	0	int64
YrSold	0	int64
SaleType	0	object
SaleCondition	0	object
SalePrice	0	int64

- Now as we see there are null values, thus we remove the null values. To remove the null values, for float data types we use mean of remaining data and for object data types we use mode of remaining data.
- Also, the object data types are encoded as then it is easy to operate on the variables.
- Now we need to check whether outliers are present in the data or not. For that we need to check if the z value/z score of all the factors is exceeding the range (-3,3). As we want the data to be normally distributed in the range of (-3,3) as the data be in the 99% domain. Thus, this will be the best data to work with.

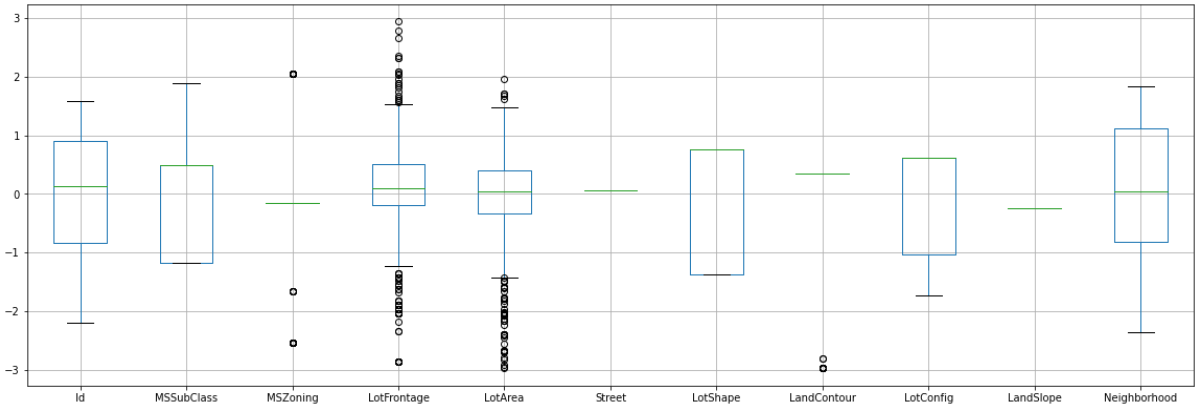




- Here we see there are a few outliers, thus removing them and bringing the data in the range of (-3,3)

  from scipy.stats import zscore
  z=np.abs(zscore(new_df))
  new_df=new_df[(z<3).all(axis=1)]

- Using the above code, we were successfully able to remove the outliers and we get the below range of data



- Now we can proceed with modelling
- As maximum values for the below columns are null, we cannot use the mode as it will be biased data, thus dropping these columns

  Alley 1091
  PoolQC 1161
  Fence 931
  MiscFeature 1124
  FireplaceQu 551

  And Utilities have same value for all rows thus dropping it as it will not contribute to the modelling

- Data Inputs- Logic- Output Relationships

Now here we know that there are 80 independent variables which are contributing to predict the dependent variable that is the Sale Price of House. Here the relationship is of Linear Regression.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Thus, assuming they follow a linear regression model, the relationship between the input variables or independent variables and the output variable or dependent variable is:

$$Y_i = f(X_i, \beta) + e_i$$

where Xi is the explanatory/80 independent variables and Yi is the dependent variable/Sale Price. f is the function, Beta is the unknown parameters and ei is the error terms.

- Correlation: Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation. Here we see the below relationship.

- Positive Correlation important factors:

  - SalePrice 1.000000
  - OverallQual 0.789185
  - GrLivArea 0.707300
  - GarageCars 0.628329
  - GarageArea 0.619000

- Negative Correlation important factors:

  - HeatingQC -0.406604
  - GarageFinish -0.537121

- KitchenQual -0.592468
- ExterQual -0.624820
- BsmtQual -0.626850

We see that the factors affecting the House Sale Price are OverallQual,GrLivArea,GarageCars,GarageArea But the most important positively affecting factor is OverallQual as it Rates the overall material and finish of the house which is important factor when a person is buying a house. Also, the negative factors affecting the Sale Price is the BsmtQual, ExterQual, KitchenQual and GarageFinish

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  - Cross-validation:

    It is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

  - Hyper Parameter Testing:

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

There are two types of search used for hyper parameter testing: Research and RandomizedSearch CV.

With small data sets and lots of resources, Grid Search will produce accurate results. However, with large data sets, the high dimensions will significantly slow down computation time and be very expensive. In this instance, it is advised to use Randomized Search. Thus, here we have used Grid Search CV as the data is too little thus better accuracy.

- Testing of Identified Approaches (Algorithms)

  - First importing all the libraries and models for train test split and linear regression

    from sklearn.model_selection import train_test_split

```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()

from sklearn.preprocessing import MinMaxScaler

mms=MinMaxScaler()

from sklearn.metrics import r2_score:
```

- Now testing the random state for maximum accuracy for dividing the data into test and train with the logic of 80:20 which is 80% of data is training data and 20% is testing data

```
for i in range (0,1000) :
    x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=i)
    lr.fit(x_train,y_train)
    pred_test=lr.predict(x_test)
    pred_train=lr.predict(x_train)
    print(f"at random state {i},the training accuracy is :{r2_score(y_train,pred_train)}")
    print(f"at random state {i},the testing accuracy is :{r2_score(y_test,pred_test)}")
    print("\n")
```

Output:

```
at random state 418, the training accuracy is :0.8941954867966609
at random state 418, the testing accuracy is :0.8948933974111771
```

We chose the random state as 418 as the testing and training accuracy were the closest to 89.4%

- Now diving the training and test data for random state=418

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=418)
```

As the total data is of 564 data points,

Train data=451

Test data=113

- Now to check which model is fit for the train and test data

```
lr.fit(x_train,y_train)
```

Output:

```
LinearRegression()
```

- Now to calculate the model accuracy

pred_test=lr.predict(x_test)

print(r2_score(y_test,pred_test))

Output:

```
0.8948933974111771
```

Thus, the model accuracy came out to be 89.49%

- Now to test for cross validation or cross fold no

Train_accuracy=r2_score(y_train,pred_train)

Test_accuracy=r2_score(y_test,pred_test)

from sklearn.model_selection import cross_val_score

for j in range(2,10):

cv_score=cross_val_score(lr,x,y,cv=j)

cv_mean=cv_score.mean()

print(f'At cross fold {j} the cv score is {cv_mean} and accuracy score for training is {Train_accuracy}and accuracy score for testing is {Test_accuracy}')

print('\n')

Output:

```
At cross fold 2 the cv score is 0.8492812637057208 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 3 the cv score is 0.8582066582145197 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 4 the cv score is 0.8597380128395549 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 5 the cv score is 0.8686996163383688 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 6 the cv score is 0.8580790685520667 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771
```

```
At cross fold 7 the cv score is 0.8676688702800658 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 8 the cv score is 0.8644717625194005 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771


At cross fold 9 the cv score is 0.8654340066161105 and accuracy score for t
raining is -0.9630898330175612and accuracy score for testing is 0.894893397
4111771
```

Here we cross validate that at which cv is the cv score maximum. Thus, we see that at cv=5 the cv score is maximum of 86.87% and the testing accuracy is 89.48% and training accuracy is 96.30% which should ideally be more than the testing accuracy

- Run and evaluate selected models

From the above testing we got the data which needs to be worked upon and that the best fit model is the Linear Regression Model with accuracy of 89.49% and the cv score is 85.9% and the model accuracy improved to 89.64% using the hyper parameter testing

- Lasso Regression

```
from sklearn.model_selection import GridSearchCV

from sklearn.model_selection import cross_val_score

import warnings

warnings.filterwarnings('ignore')

from sklearn.linear_model import Lasso

parameters={'alpha':[.0001,.001,.01,.1,1,10],'random_state':list(range(0,10))}

ls=Lasso()

clf=GridSearchCV(ls,parameters)

clf.fit(x_train,y_train)

print(clf.best_params_)
```

Output:

```
{'alpha': 10, 'random_state': 0}
```

```
ls=Lasso(alpha=10,random_state=0)
```

```
ls.fit(x_train,y_train)

ls.score(x_train,y_train)

pred_ls=ls.predict(x_test)

lss=r2_score(y_test,pred_ls)

lss
```

Output:

```
0.8964598211473586
```

```
cv_score=cross_val_score(ls,x,y,cv=3)

cv_mean=cv_score.mean()

cv_mean
```

Output:

```
0.8593456878449621
```

Here we see that the model accuracy is 89.64% and the cross validation score is 85.9%

- Random Forest Regression

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor

parameters = {'criterion':['mse','mae'],'max_features':["auto","sqrt","log2"]}
rf=RandomForestRegressor()
clf=GridSearchCV(rf,parameters)
clf.fit(x_train,y_train)
print(clf.best_params_)
```

Output:

```
{'criterion': 'mse', 'max_features': 'sqrt'}
```

```
rf=RandomForestRegressor(criterion='mse',max_features='sqrt')
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
pred_decision=rf.predict(x_test)

rfs=r2_score(y_test,pred_decision)
print('R2 score:',rfs*100)

rfscore=cross_val_score(rf,x,y,cv=5)
rfc=rfscore.mean()
print('Cross Val Score:',rfc*100)
```

Output:

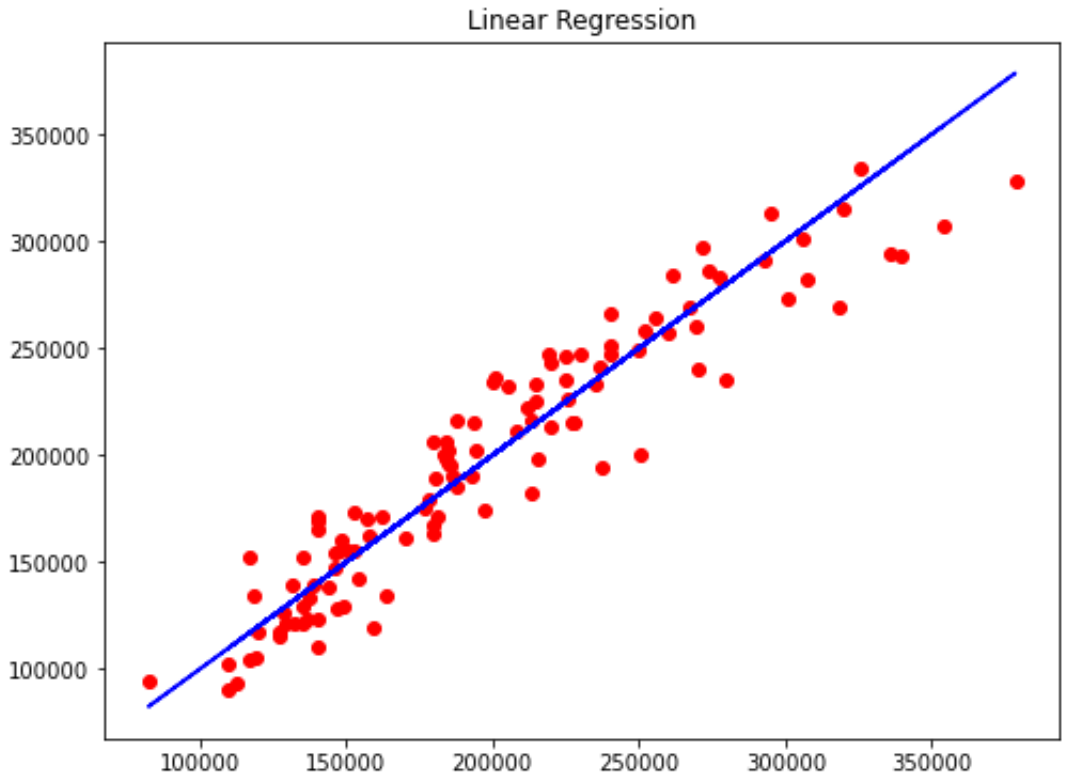
```
R2 score: 89.05735508460583
Cross Val Score: 87.34008584745702
```

- Key Metrics for success in solving problem under consideration

  The key metrics for improving the model accuracy are:
  Cross Validation
  Hyper Parameter testing

- Visualizations

```
import matplotlib.pyplot as plt
plt.figure(figsize=(8,6))
plt.scatter(x=y_test,y=pred_test,color='r')
plt.plot(y_test,y_test,color='b')
plt.title("Linear Regression")
plt.show()
```



- Interpretation of the Results

  Here we see that our data points are in line with the blue line which is the ideal line or slope of the linear regression equation. Thus, this

shows that the model is perfectly fitted, neither underfitted nor overfitted.

Here the model accuracy is almost the same to 89.05% but the cross-validation score is increased to 87.34% which is close to the model accuracy which means that the model is not overfitted.

- Predicting the Sale Price for the provided test data

  First, we have to handle the missing/null values and data pre-processing is done then this data is then fed into the trained model and we get the output.

```
model_test=pd.DataFrame(loaded_model.predict(Sales7))
model_test
result = pd.DataFrame()
target_var=loaded_model.predict(Sales7)

result["Sale Price"] = target_var
result = result.sort_index()
result.to_csv('Housing_Price_Prediction.csv',index = False)
```

The predicted Sale Price are stored in a csv as below:

Housing_Price_Prediction.csv

CONCLUSION

- Key Findings and Conclusions of the Study

We did Exploratory information Investigation on the highlights of this dataset and saw how each include is distributed.

We dissected each variable to check in the event that information is cleaned and ordinarily distributed. We cleaned
the information and evacuated NA values. We tried to find the correlation and based on the outcomes, we accepted whether or not there's a relation between the Sale Price of the house and the other factors, we see that the factors affecting the House Sale Price are OverallQual,GrLivArea,GarageCars,GarageArea But the most important positively affecting factor is OverallQual as it Rates the overall material and finish of the house which is important factor when a person is buying a house. Also, the negative factors affecting the Sale Price is the BsmtQual, ExterQual, KitchenQual and GarageFinish. We used the linear regression model to predict the Sale Price of House.

We got a very good model with an 89% accuracy which is very good. Just a disadvantage that the data available to train the model is little less and thus needs to be improved. The more the data the better the model will be trained.

This model will be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

- Learning Outcomes of the Study in respect of Data Science

➢ Able to demonstrate proficiency with statistical analysis of data.

- Ability to build and assess data-based models.
- Data management.
- Able to do basic data cleaning, and can transform variables to facilitate analysis.
- How variables are connected and changing the independent variables how the dependent variable changes.
- To perform hyper parameter testing and cross validation.
- Interpretation of data and results
- How to reach the outputs

- Limitations of this work and Scope for Future Work

  The limitations were that we had limited data for both test and train data, also we did not know the sources of data so to handle the missing values and null values was done using our own assumptions. Also, lack of reliable data such as self-reported data, missing data, and deficiencies in data measurements (such as a questionnaire item not asked that could have been used to address a specific issue).

  We can improve the results by taking more samples and correct data. Also, for the 4-5 variables which have very high (more than 80%) missing/null values, if we could again get data then we could build a more robust model.