0. Importing libraries and dataset

```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings( "ignore", module = "seaborn\..*" )
import seaborn as sns
from scipy import stats
from sklearn.feature_selection import SelectKBest,chi2

dataset = pd.read_csv('PEP1.csv')
```

1. Understanding the dataset

```
#a. Identify the shape of the dataset
dataset.shape

(1460, 81)

#b. Identify variables with null values
dataset.isnull().any()

Id                False
MSSubClass        False
MSZoning          False
LotFrontage        True
LotArea           False
                  ...
MoSold            False
YrSold            False
SaleType          False
SaleCondition     False
SalePrice         False
Length: 81, dtype: bool

#c. Identify variables with unique values
dataset.nunique()

Id                1460
MSSubClass          15
MSZoning             5
LotFrontage        110
LotArea           1073
                  ...
MoSold              12
YrSold               5
SaleType             9
SaleCondition        6
SalePrice          663
Length: 81, dtype: int64
```

2. Generating seperate datasets for numerical and categorical variables

```
num_cols =
['Id','MSSubClass','LotFrontage','LotArea','OverallQual','OverallCond'
,'YearBuilt','YearRemodAdd','MasVnrArea','BsmtFinSF1','BsmtFinSF2','Bs
mtUnfSF','TotalBsmtSF','1stFlrSF','2ndFlrSF','LowQualFinSF','GrLivArea
','BsmtFullBath','BsmtHalfBath','FullBath','HalfBath','BedroomAbvGr','
KitchebvGr','TotRmsAbvGrd','Fireplaces','GarageYrBlt','GarageCars','Ga
rageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3SsnPorch','Scre
enPorch','PoolArea','MiscVal','MoSold','YrSold','SalePrice']
num_var = dataset[num_cols].copy()
cat_var = dataset.drop(columns = num_cols).copy()

num_var.head()
```

```
   Id  MSSubClass  LotFrontage  LotArea  OverallQual  OverallCond
YearBuilt  \
0   1          60         65.0     8450            7            5
2003
1   2          20         80.0     9600            6            8
1976
2   3          60         68.0    11250            7            5
2001
3   4          70         60.0     9550            7            5
1915
4   5          60         84.0    14260            8            5
2000

   YearRemodAdd  MasVnrArea  BsmtFinSF1  ...  WoodDeckSF  OpenPorchSF
\
0          2003       196.0         706  ...           0           61

1          1976         0.0         978  ...         298            0

2          2002       162.0         486  ...           0           42

3          1970         0.0         216  ...           0           35

4          2000       350.0         655  ...         192           84


   EnclosedPorch  3SsnPorch  ScreenPorch  PoolArea  MiscVal  MoSold
YrSold  \
0              0          0            0         0        0       2
2008
1              0          0            0         0        0       5
2007
2              0          0            0         0        0       9
2008
3            272          0            0         0        0       2
2006
```

```
4              0          0            0         0        0      12
2008

   SalePrice
0     208500
1     181500
2     223500
3     140000
4     250000

[5 rows x 38 columns]
```

cat_var.head()

```
  MSZoning Street Alley LotShape LandContour Utilities LotConfig
LandSlope  \
0       RL   Pave   NaN      Reg         Lvl    AllPub    Inside
Gtl
1       RL   Pave   NaN      Reg         Lvl    AllPub       FR2
Gtl
2       RL   Pave   NaN      IR1         Lvl    AllPub    Inside
Gtl
3       RL   Pave   NaN      IR1         Lvl    AllPub    Corner
Gtl
4       RL   Pave   NaN      IR1         Lvl    AllPub       FR2
Gtl

  Neighborhood Condition1  ... GarageType GarageFinish GarageQual
GarageCond  \
0      CollgCr       Norm  ...     Attchd          RFn         TA
TA
1      Veenker      Feedr  ...     Attchd          RFn         TA
TA
2      CollgCr       Norm  ...     Attchd          RFn         TA
TA
3      Crawfor       Norm  ...     Detchd          Unf         TA
TA
4      NoRidge       Norm  ...     Attchd          RFn         TA
TA

  PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition
0          Y    NaN   NaN         NaN       WD        Normal
1          Y    NaN   NaN         NaN       WD        Normal
2          Y    NaN   NaN         NaN       WD        Normal
3          Y    NaN   NaN         NaN       WD       Abnorml
4          Y    NaN   NaN         NaN       WD        Normal

[5 rows x 43 columns]
```

3. EDA of numerical variables

```
num_var.isnull().any()
```

```
Id               False
MSSubClass       False
LotFrontage       True
LotArea          False
OverallQual      False
OverallCond      False
YearBuilt        False
YearRemodAdd     False
MasVnrArea        True
BsmtFinSF1       False
BsmtFinSF2       False
BsmtUnfSF        False
TotalBsmtSF      False
1stFlrSF         False
2ndFlrSF         False
LowQualFinSF     False
GrLivArea        False
BsmtFullBath     False
BsmtHalfBath     False
FullBath         False
HalfBath         False
BedroomAbvGr     False
KitchebvGr       False
TotRmsAbvGrd     False
Fireplaces       False
GarageYrBlt       True
GarageCars       False
GarageArea       False
WoodDeckSF       False
OpenPorchSF      False
EnclosedPorch    False
3SsnPorch        False
ScreenPorch      False
PoolArea         False
MiscVal          False
MoSold           False
YrSold           False
SalePrice        False
dtype: bool
```

```
#a. Missing value treatment
num_mean_na = ['LotFrontage','MasVnrArea']
for col1 in num_mean_na:
    num_var[col1] = num_var[col1].fillna(num_var[col1].mean())
for row in range(len(num_var['GarageYrBlt'])):
    if pd.isnull(num_var.loc[row,'GarageYrBlt']):
```

```
        num_var.loc[row,'GarageYrBlt'] = num_var.loc[row,'YearBuilt']
num_var.isnull().any()
```

```
Id                False
MSSubClass        False
LotFrontage       False
LotArea           False
OverallQual       False
OverallCond       False
YearBuilt         False
YearRemodAdd      False
MasVnrArea        False
BsmtFinSF1        False
BsmtFinSF2        False
BsmtUnfSF         False
TotalBsmtSF       False
1stFlrSF          False
2ndFlrSF          False
LowQualFinSF      False
GrLivArea         False
BsmtFullBath      False
BsmtHalfBath      False
FullBath          False
HalfBath          False
BedroomAbvGr      False
KitchebvGr        False
TotRmsAbvGrd      False
Fireplaces        False
GarageYrBlt       False
GarageCars        False
GarageArea        False
WoodDeckSF        False
OpenPorchSF       False
EnclosedPorch     False
3SsnPorch         False
ScreenPorch       False
PoolArea          False
MiscVal           False
MoSold            False
YrSold            False
SalePrice         False
dtype: bool
```

```
#b. Identify the skewness and distribution
num_var.skew()
```

```
Id                 0.000000
MSSubClass         1.407657
LotFrontage        2.384950
LotArea           12.207688
OverallQual        0.216944
```

```
OverallCond       0.693067
YearBuilt        -0.613461
YearRemodAdd     -0.503562
MasVnrArea        2.676412
BsmtFinSF1        1.685503
BsmtFinSF2        4.255261
BsmtUnfSF         0.920268
TotalBsmtSF       1.524255
1stFlrSF          1.376757
2ndFlrSF          0.813030
LowQualFinSF      9.011341
GrLivArea         1.366560
BsmtFullBath      0.596067
BsmtHalfBath      4.103403
FullBath          0.036562
HalfBath          0.675897
BedroomAbvGr      0.211790
KitchebvGr        4.488397
TotRmsAbvGrd      0.676341
Fireplaces        0.649565
GarageYrBlt      -0.694329
GarageCars       -0.342549
GarageArea        0.179981
WoodDeckSF        1.541376
OpenPorchSF       2.364342
EnclosedPorch     3.089872
3SsnPorch        10.304342
ScreenPorch       4.122214
PoolArea         14.828374
MiscVal          24.476794
MoSold            0.212053
YrSold            0.096269
SalePrice         1.882876
dtype: float64
```

```
num_var.drop(columns=['Id']).describe()
```

```
        MSSubClass  LotFrontage        LotArea  OverallQual
OverallCond  \
count  1460.000000  1460.000000    1460.000000  1460.000000
1460.000000
mean     56.897260    70.049958   10516.828082     6.099315
5.575342
std      42.300571    22.024023    9981.264932     1.382997
1.112799
min      20.000000    21.000000    1300.000000     1.000000
1.000000
25%      20.000000    60.000000    7553.500000     5.000000
5.000000
50%      50.000000    70.049958    9478.500000     6.000000
5.000000
```

```
75%      70.000000    79.000000   11601.500000     7.000000
6.000000
max     190.000000   313.000000  215245.000000    10.000000
9.000000

         YearBuilt  YearRemodAdd   MasVnrArea   BsmtFinSF1
BsmtFinSF2  ...  \
count  1460.000000   1460.000000  1460.000000  1460.000000
1460.000000  ...
mean   1971.267808   1984.865753   103.685262   443.639726
46.549315  ...
std      30.202904     20.645407   180.569112   456.098091
161.319273  ...
min    1872.000000   1950.000000     0.000000     0.000000
0.000000  ...
25%    1954.000000   1967.000000     0.000000     0.000000
0.000000  ...
50%    1973.000000   1994.000000     0.000000   383.500000
0.000000  ...
75%    2000.000000   2004.000000   164.250000   712.250000
0.000000  ...
max    2010.000000   2010.000000  1600.000000  5644.000000
1474.000000  ...

        WoodDeckSF  OpenPorchSF  EnclosedPorch    3SsnPorch
ScreenPorch  \
count  1460.000000  1460.000000    1460.000000  1460.000000
1460.000000
mean     94.244521    46.660274      21.954110     3.409589
15.060959
std     125.338794    66.256028      61.119149    29.317331
55.757415
min       0.000000     0.000000       0.000000     0.000000
0.000000
25%       0.000000     0.000000       0.000000     0.000000
0.000000
50%       0.000000    25.000000       0.000000     0.000000
0.000000
75%     168.000000    68.000000       0.000000     0.000000
0.000000
max     857.000000   547.000000     552.000000   508.000000
480.000000

          PoolArea       MiscVal       MoSold       YrSold
SalePrice
count  1460.000000   1460.000000  1460.000000  1460.000000
1460.000000
mean      2.758904     43.489041     6.321918  2007.815753
180921.195890
std      40.177307    496.123024     2.703626     1.328095
```

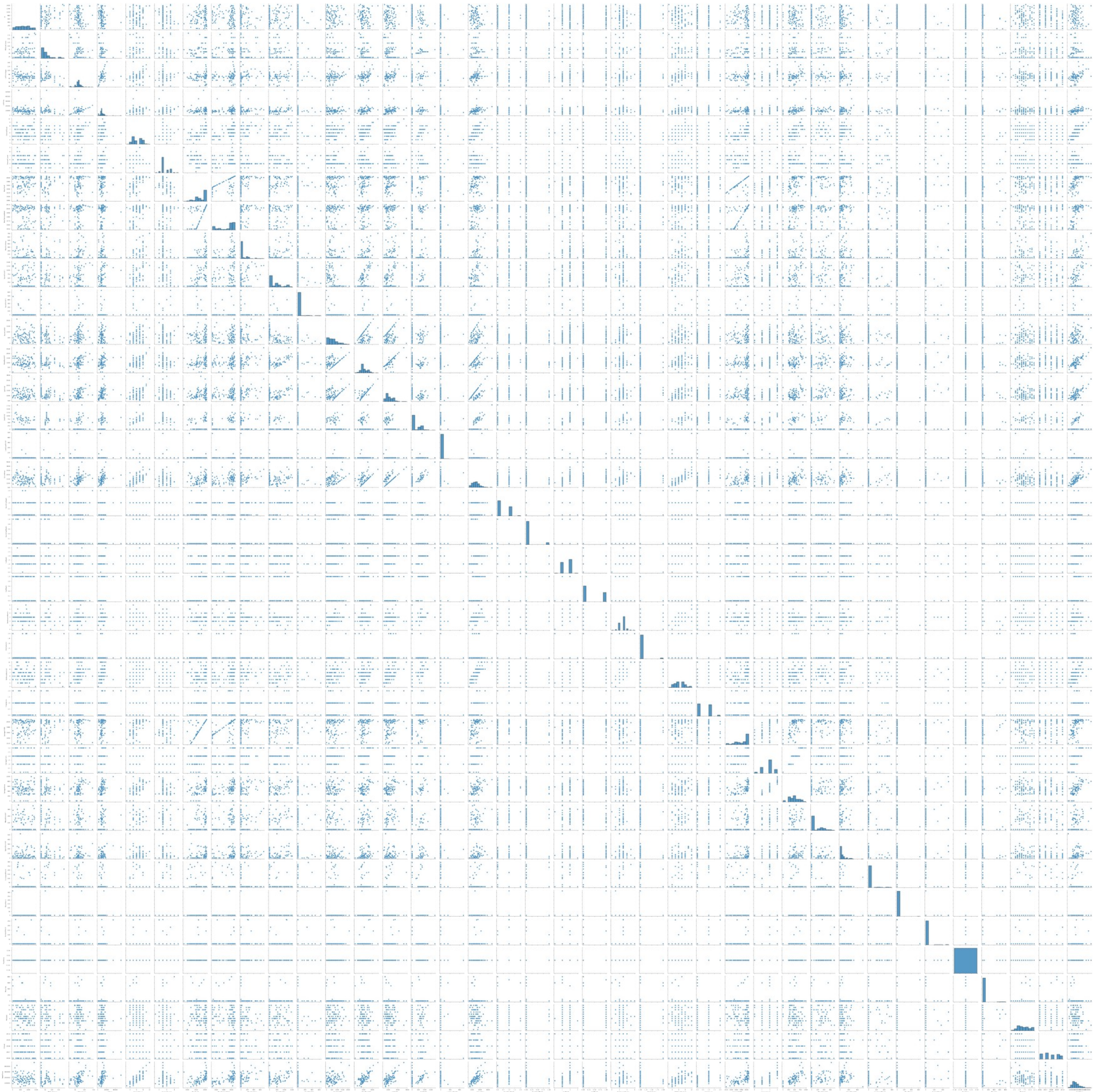
```
79442.502883
min       0.000000      0.000000     1.000000  2006.000000
34900.000000
25%       0.000000      0.000000     5.000000  2007.000000
129975.000000
50%       0.000000      0.000000     6.000000  2008.000000
163000.000000
75%       0.000000      0.000000     8.000000  2009.000000
214000.000000
max     738.000000  15500.000000    12.000000  2010.000000
755000.000000

[8 rows x 37 columns]
```

```
#c. Identify significant variables using a correlation matrix
f = plt.figure(figsize=(20,20))
corr = num_var.drop(columns=['Id']).corr()
corr.style.background_gradient(cmap='coolwarm',vmin=-1,vmax=1)
```

```
<pandas.io.formats.style.Styler at 0x1fa7bde2d10>

<Figure size 1440x1440 with 0 Axes>
```

```
#d. Pair plot for distribution and density
sns.pairplot(num_var.sample(100))
#we can zoom in and observe the relationship between any two variables
```

```
<seaborn.axisgrid.PairGrid at 0x1fa7bf6cc40>
```

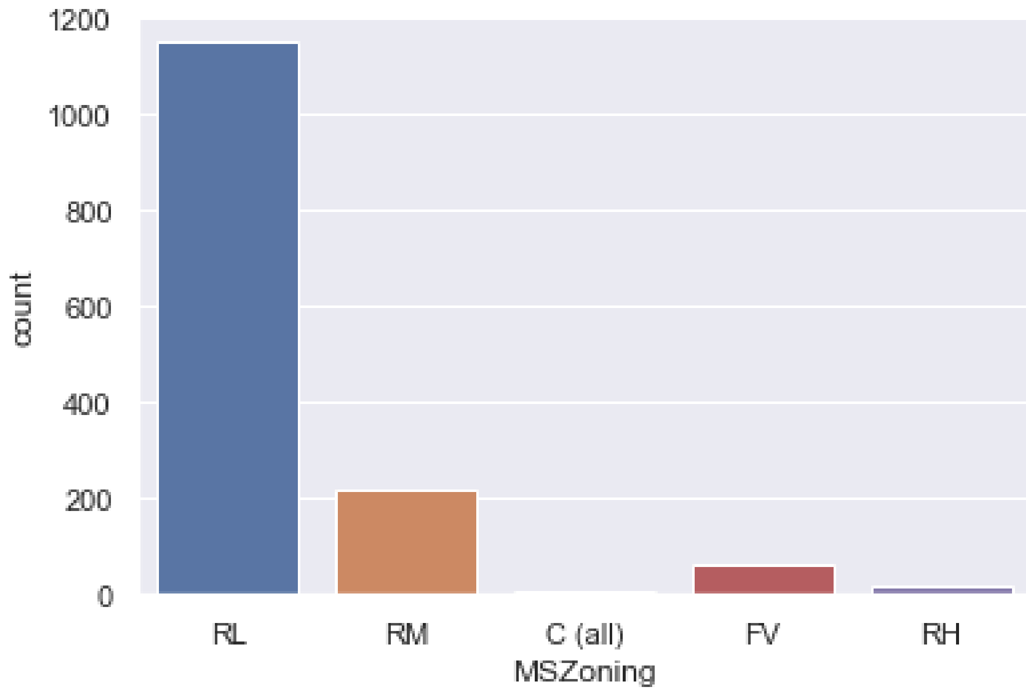


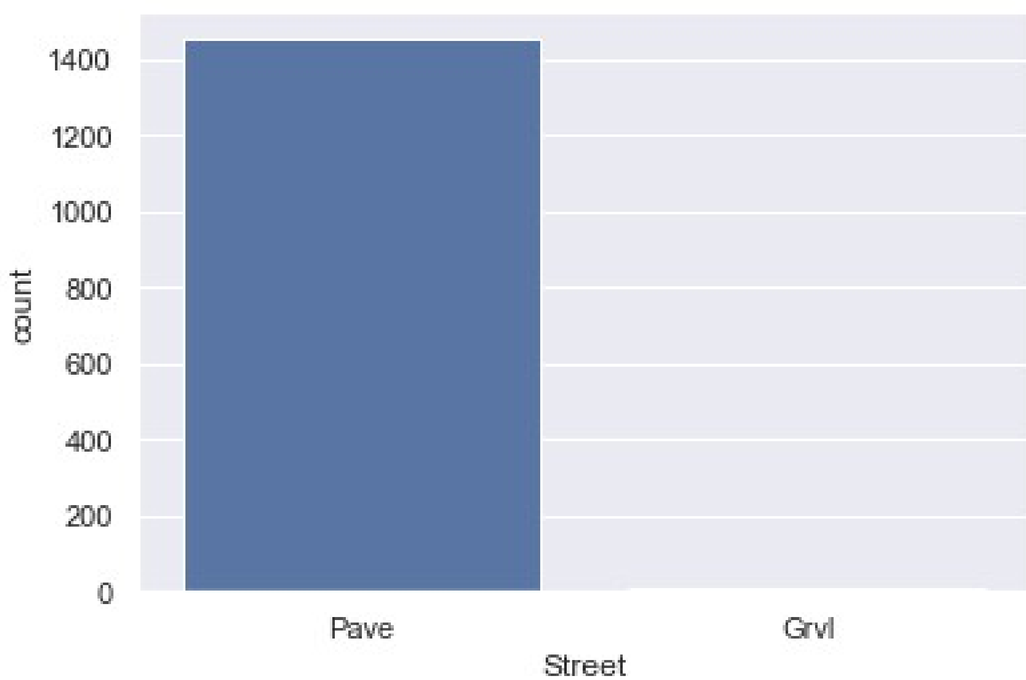
3. EDA of categorical variables

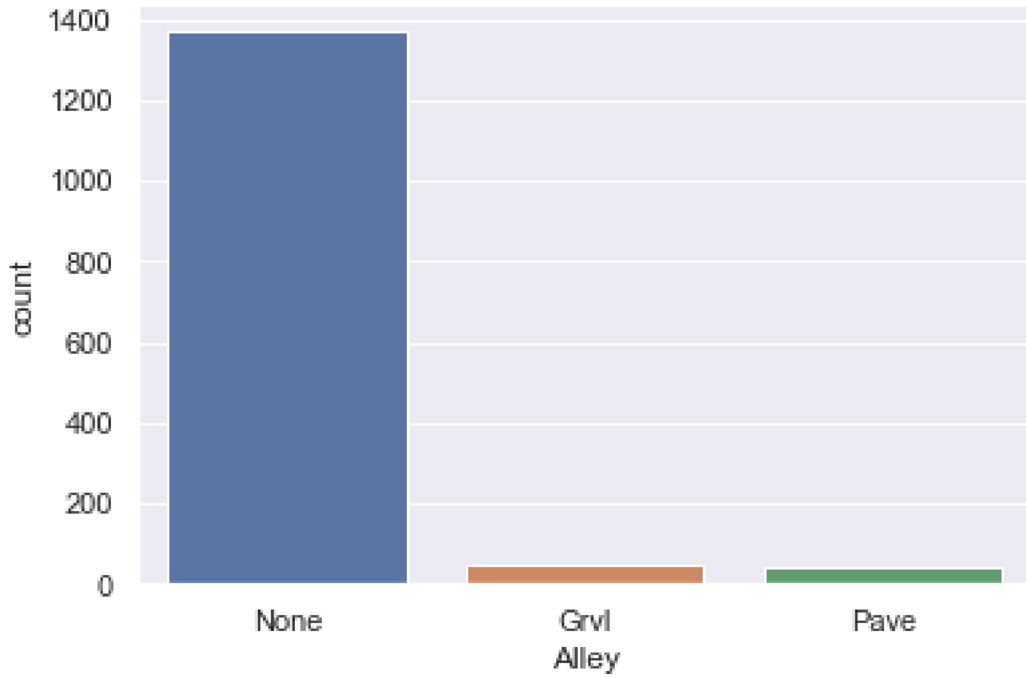
```
#a. Missing value treatment
cat_none_na =
['Alley','MasVnrType','FireplaceQu','PoolQC','MiscFeature']
cat_mode_na =
['BsmtQual','BsmtCond','BsmtExposure','BsmtFinType1','BsmtFinType2','E
lectrical','GarageType','GarageFinish','GarageQual','GarageCond','Fenc
e']
for col2 in cat_none_na:
    cat_var[col2] = cat_var[col2].fillna('None')
for col3 in cat_mode_na:
    cat_var[col3] = cat_var[col3].fillna(cat_var[col3].mode()[0])
cat_var.isnull().any()

MSZoning         False
Street           False
```

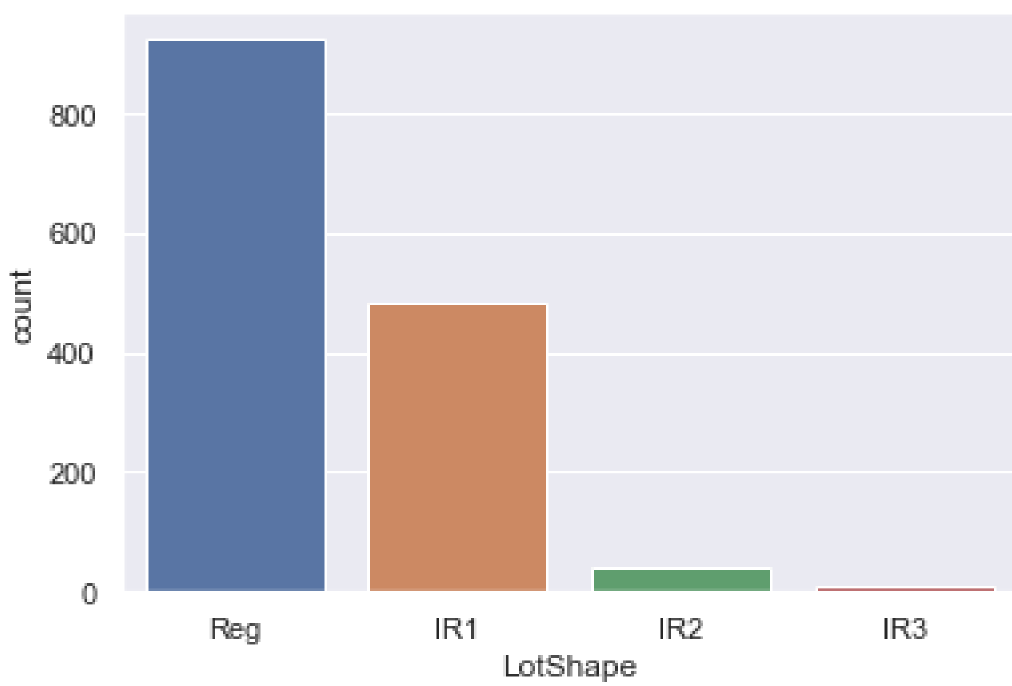
```
Alley            False
LotShape         False
LandContour      False
Utilities        False
LotConfig        False
LandSlope        False
Neighborhood     False
Condition1       False
Condition2       False
BldgType         False
HouseStyle       False
RoofStyle        False
RoofMatl         False
Exterior1st      False
Exterior2nd      False
MasVnrType       False
ExterQual        False
ExterCond        False
Foundation       False
BsmtQual         False
BsmtCond         False
BsmtExposure     False
BsmtFinType1     False
BsmtFinType2     False
Heating          False
HeatingQC        False
CentralAir       False
Electrical       False
KitchenQual      False
Functiol         False
FireplaceQu      False
GarageType       False
GarageFinish     False
GarageQual       False
GarageCond       False
PavedDrive       False
PoolQC           False
Fence            False
MiscFeature      False
SaleType         False
SaleCondition    False
dtype: bool
```

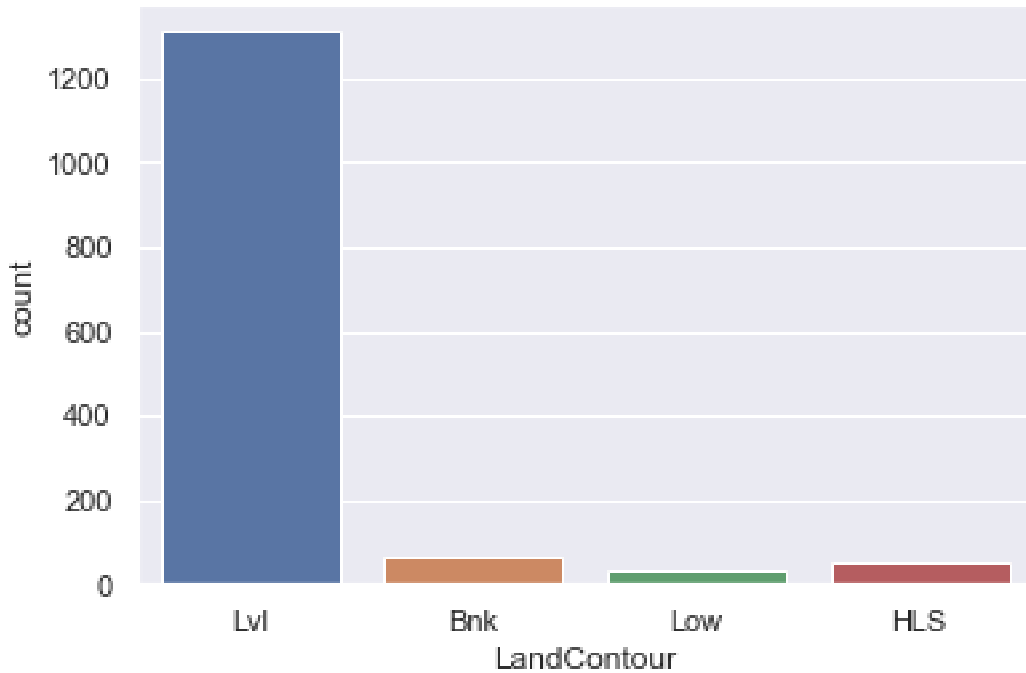
```python
#b. Count plot and box plot for bivariate analysis
sns.set()
cols = cat_var.columns.values.tolist()
for col in cols:
    sns.countplot(cat_var[col])
    plt.show()
```

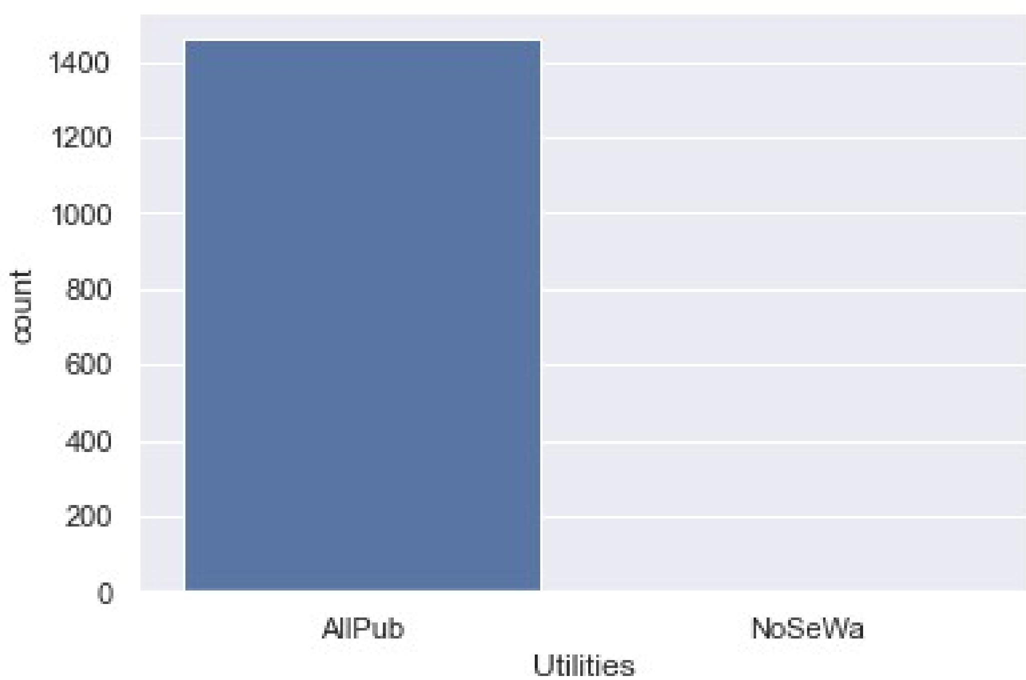

1200
1000
800
600
400
200
0
count
RL
RM
C (all)
FV
RH
MSZoning

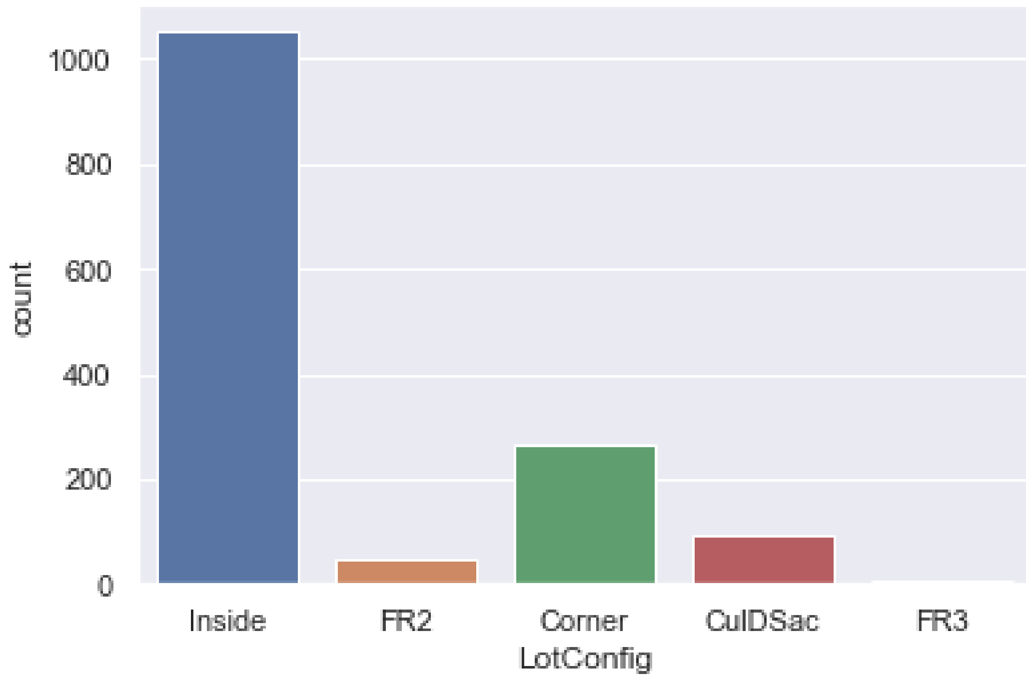

1400
1200
1000
800
600
400
200
0
count
Pave
Grvl
Street

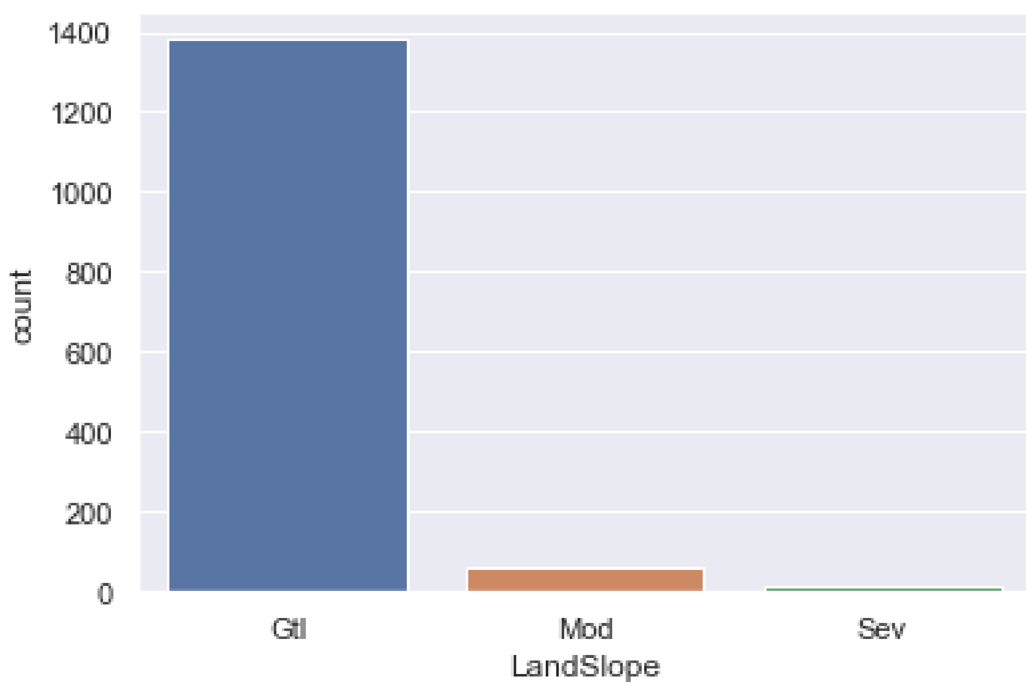

1400
1200
1000
800
600
400
200
0
count
None
Grvl
Pave
Alley

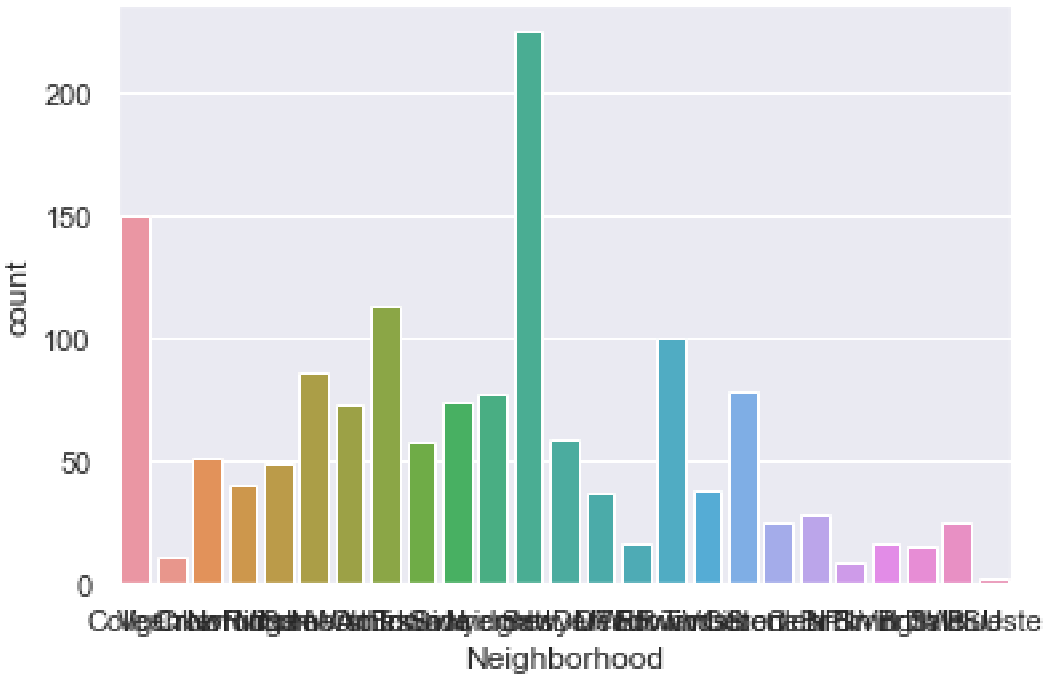

800
600
400
200
0
count
Reg
IR1
IR2
IR3
LotShape

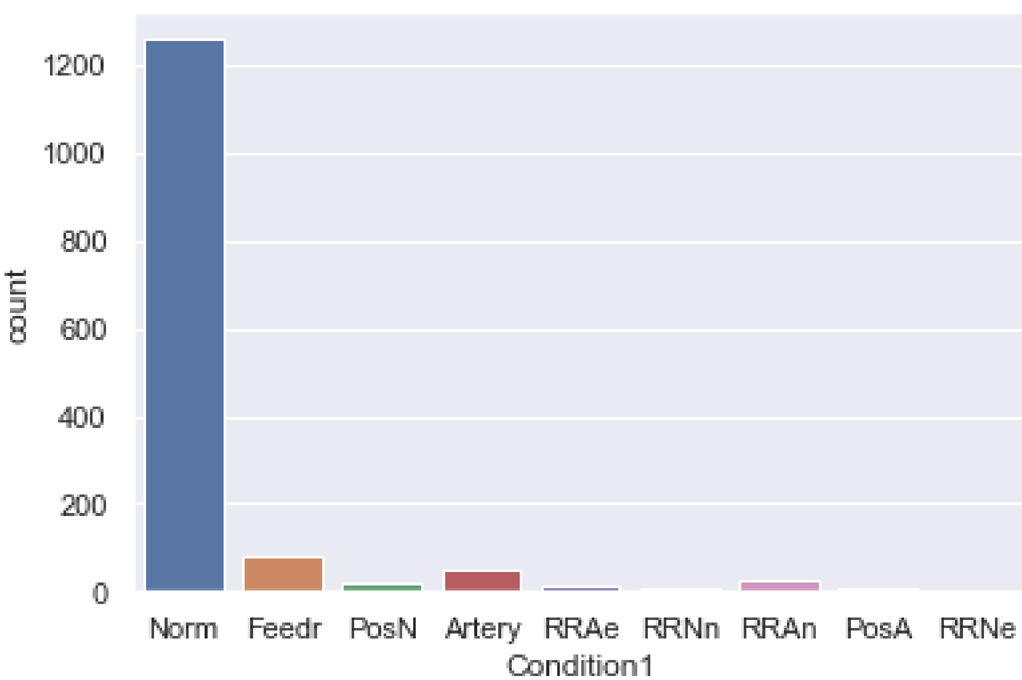

count
1200
1000
800
600
400
200
0
Lvl
Bnk
Low
HLS
LandContour

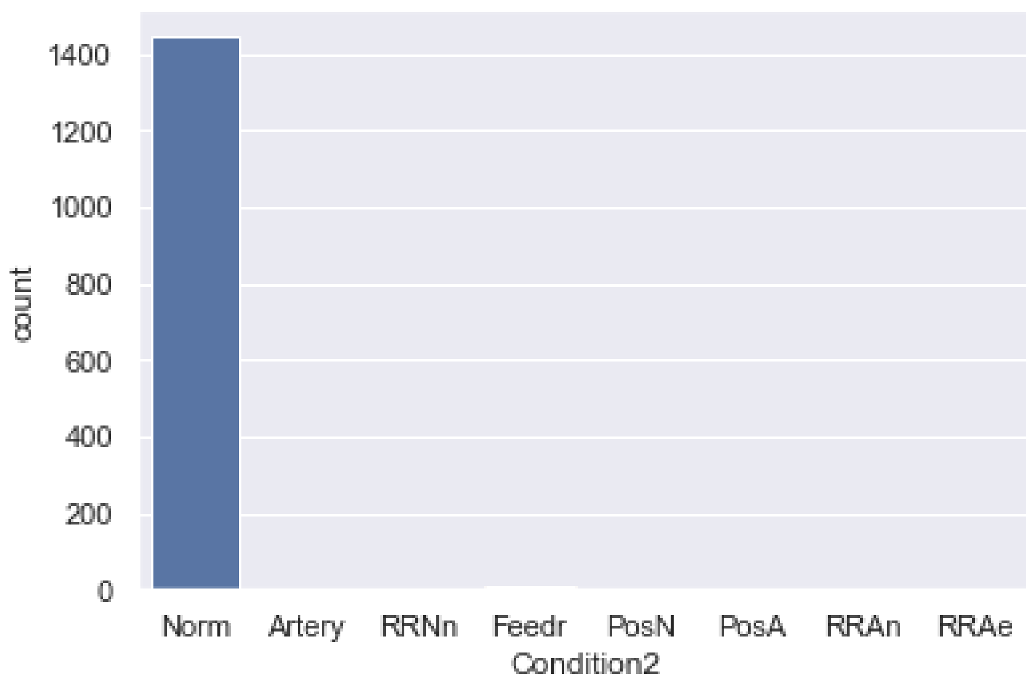

count
1400
1200
1000
800
600
400
200
0
AllPub
NoSeWa
Utilities

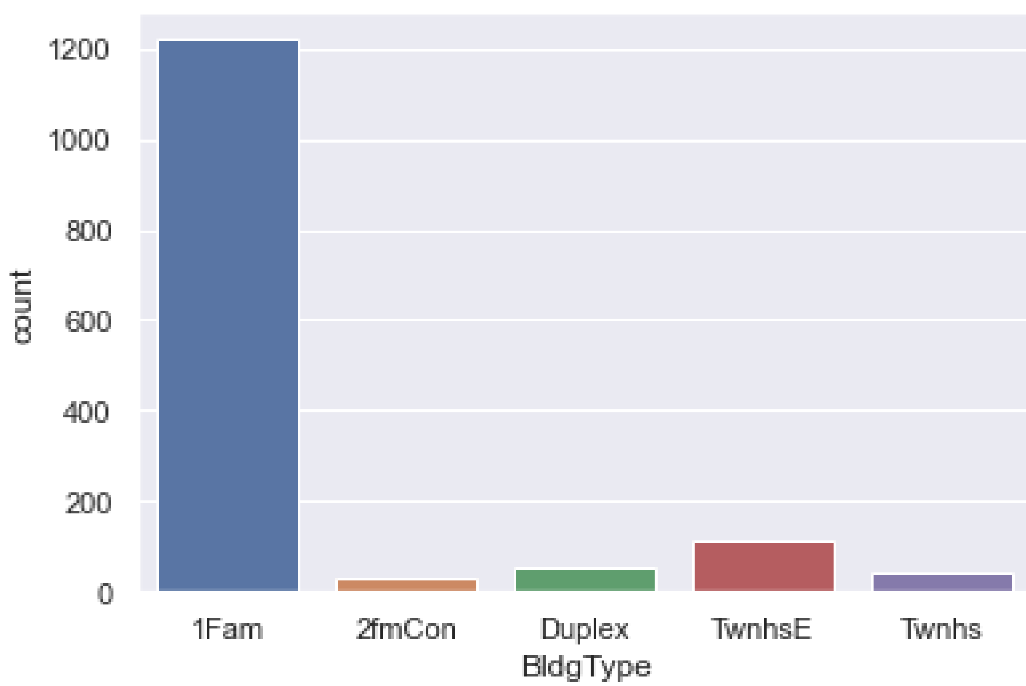

1000
800
600
400
200
0
count
Inside
FR2
Corner
CulDSac
FR3
LotConfig

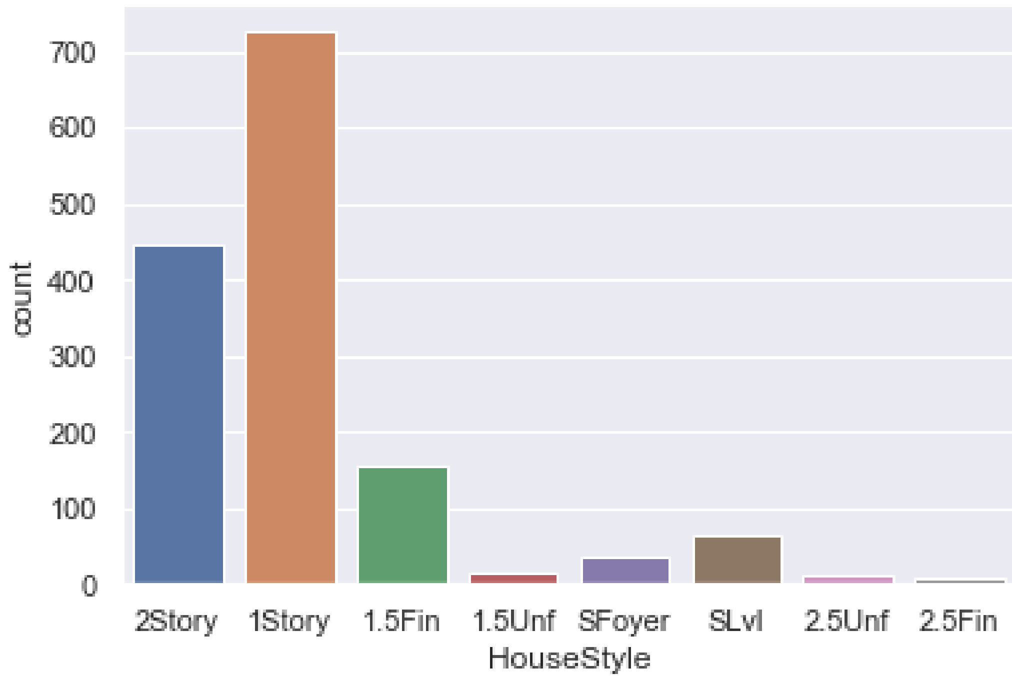

1400
1200
1000
800
600
400
200
0
count
Gtl
Mod
Sev
LandSlope

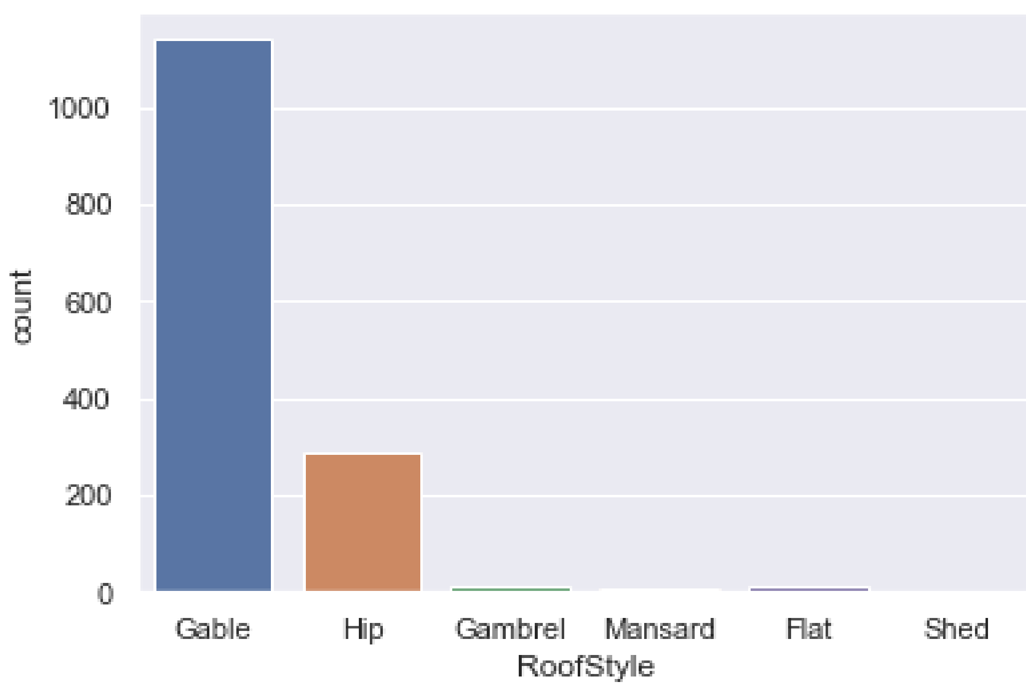

count
0
50
100
150
200
Neighborhood

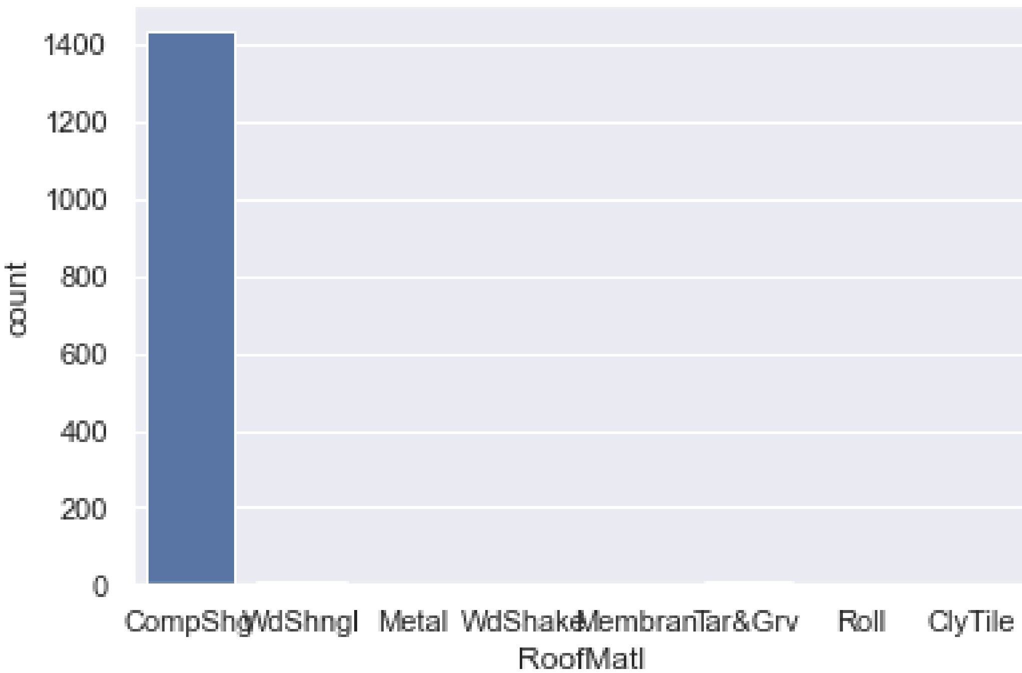

count
0
200
400
600
800
1000
1200
Norm
Feedr
PosN
Artery
RRAe
RRNn
RRAn
PosA
RRNe
Condition1

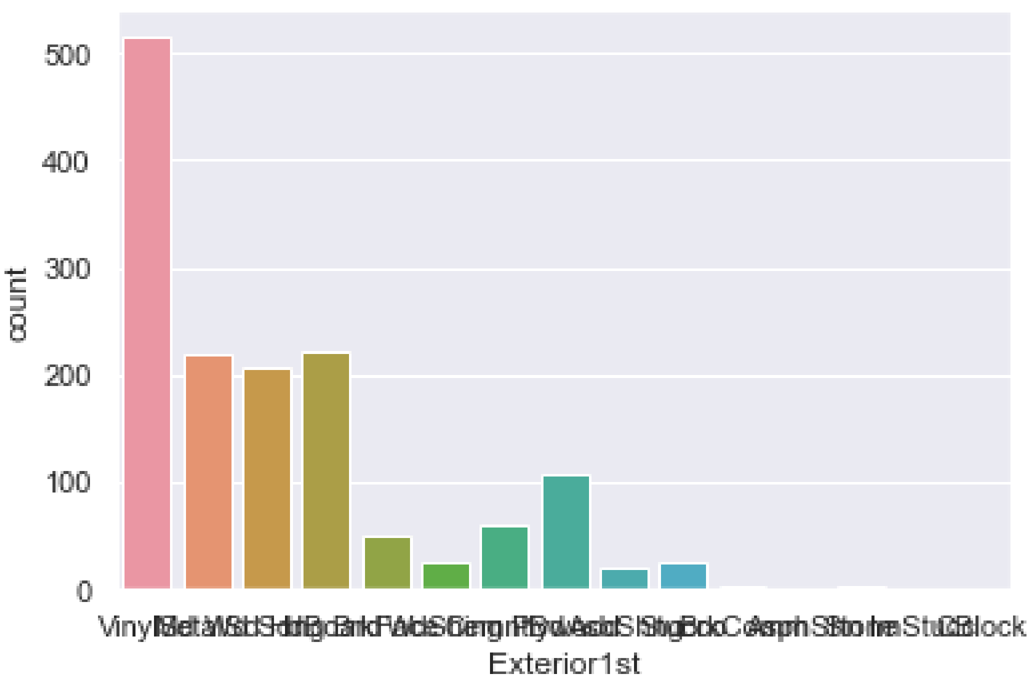

count
1400
1200
1000
800
600
400
200
0
Norm
Artery
RRNn
Feedr
PosN
PosA
RRAn
RRAe
Condition2

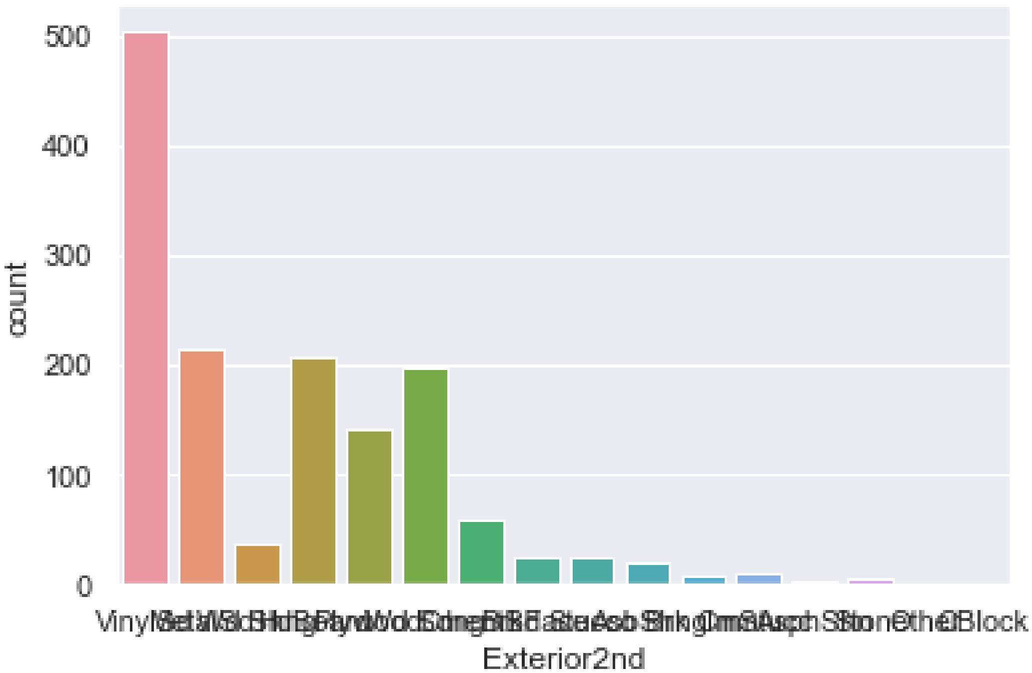

count
1200
1000
800
600
400
200
0
1Fam
2fmCon
Duplex
TwnhsE
Twnhs
BldgType

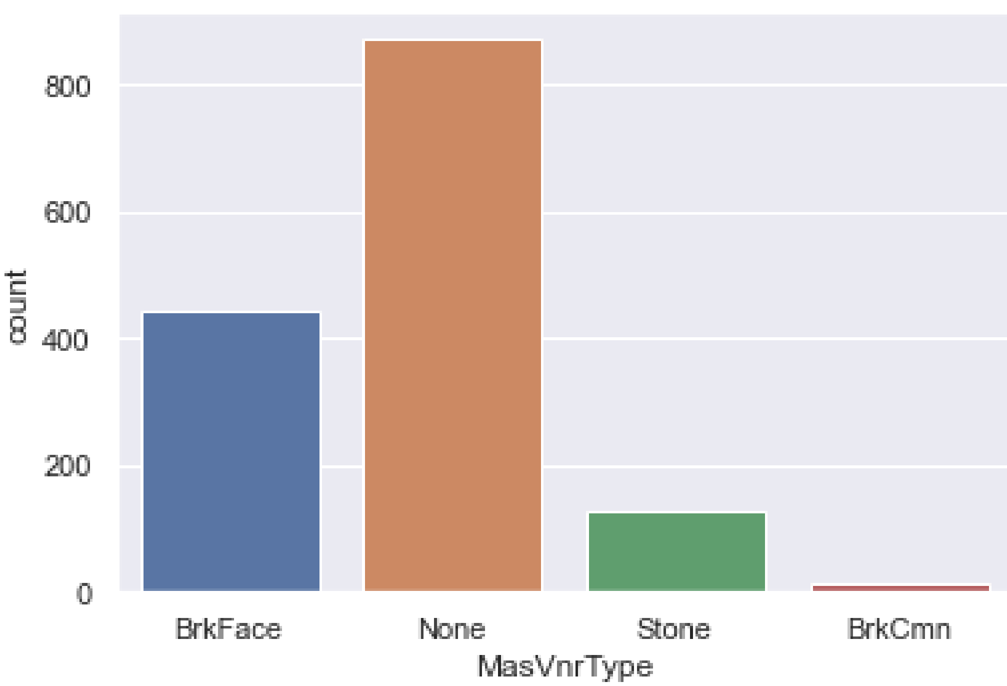

700
600
500
400
300
200
100
0
count
2Story
1Story
1.5Fin
1.5Unf
SFoyer
SLvl
2.5Unf
2.5Fin
HouseStyle

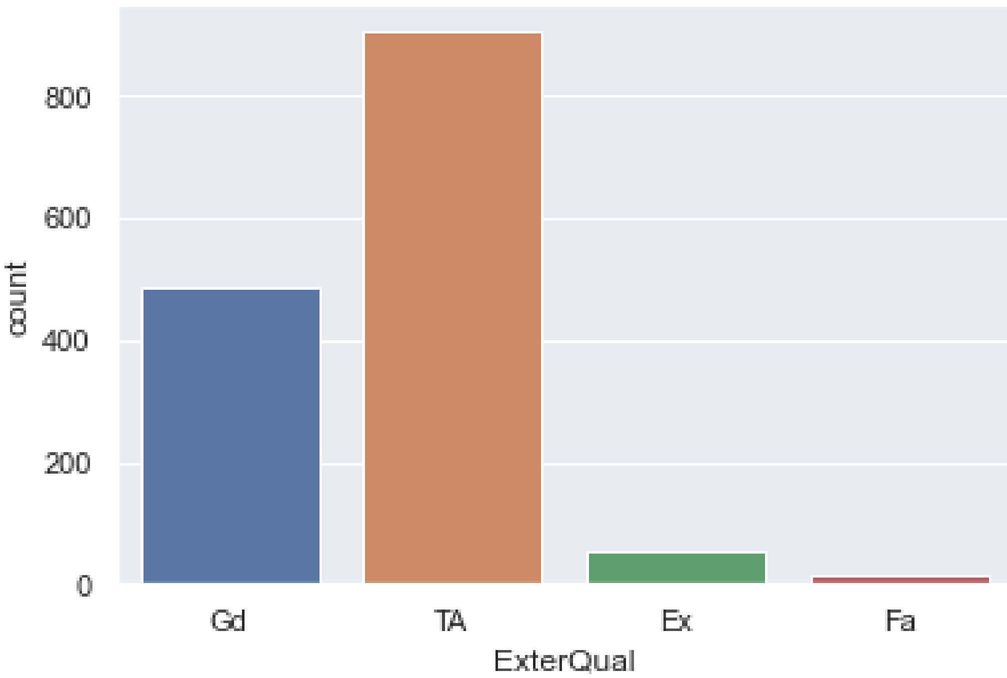

1000
800
600
400
200
0
count
Gable
Hip
Gambrel
Mansard
Flat
Shed
RoofStyle


count
1400
1200
1000
800
600
400
200
0
CompShg
WdShngl
Metal
WdShake
Tar&Grv
Roll
ClyTile
RoofMatl

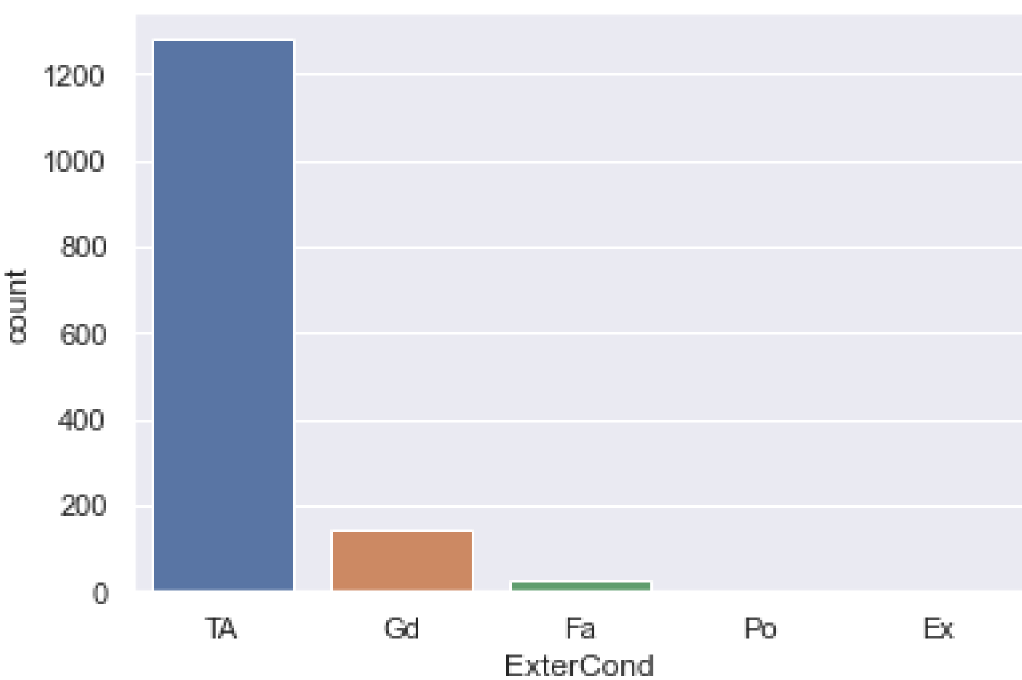

count
500
400
300
200
100
0
CBlock
Exterior1st


500
400
300
200
100
0
count
Exterior2nd

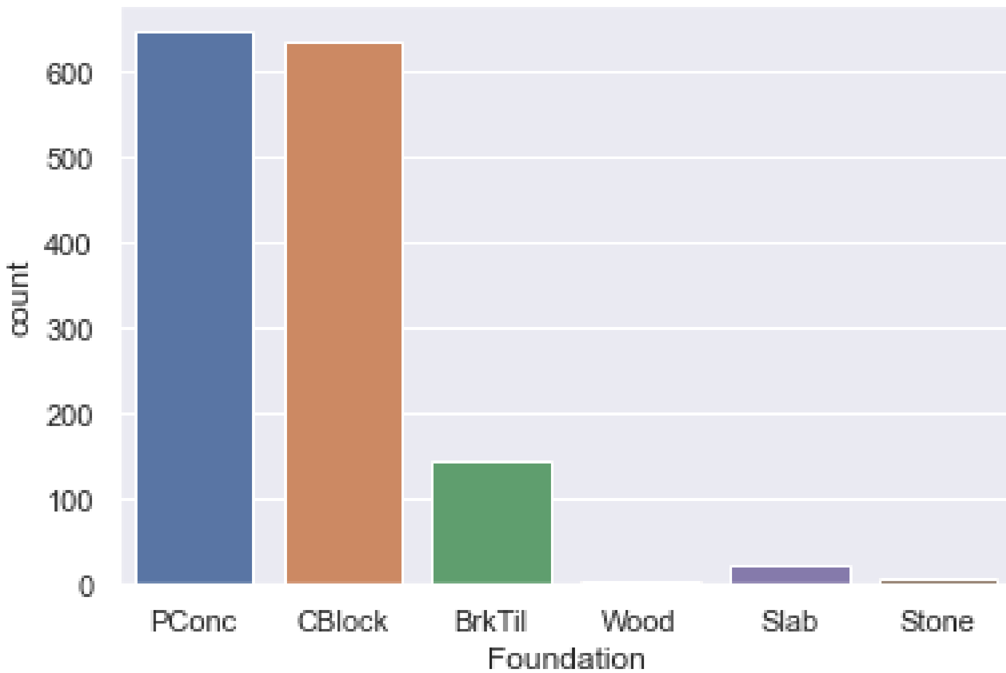

800
600
400
200
0
count
BrkFace
None
Stone
BrkCmn
MasVnrType



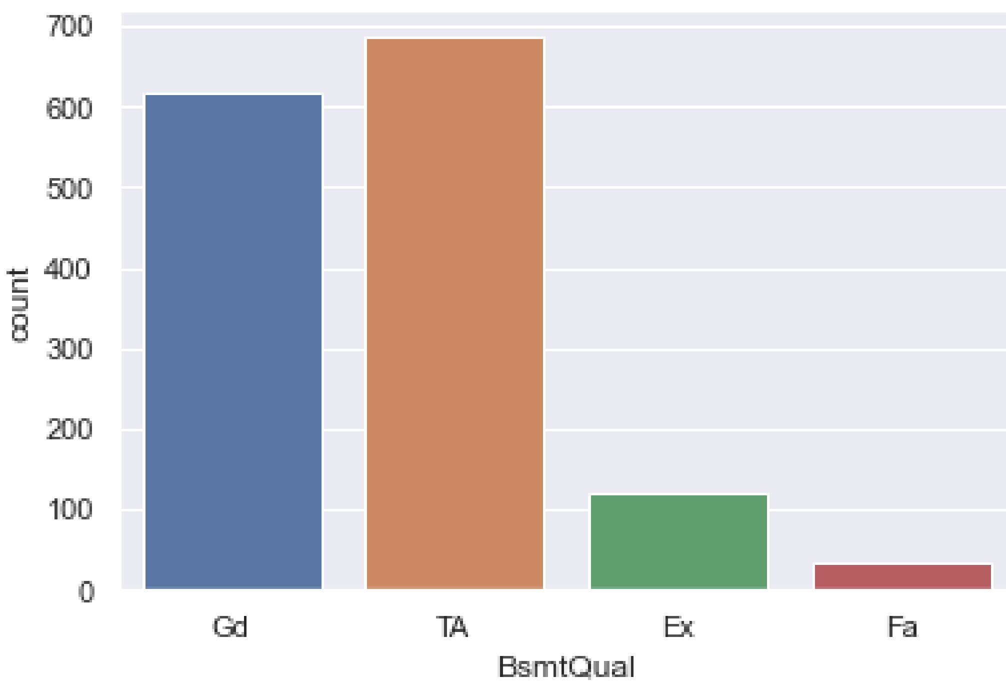
800
600
400
200
0
count
Gd
TA
Ex
Fa
ExterQual



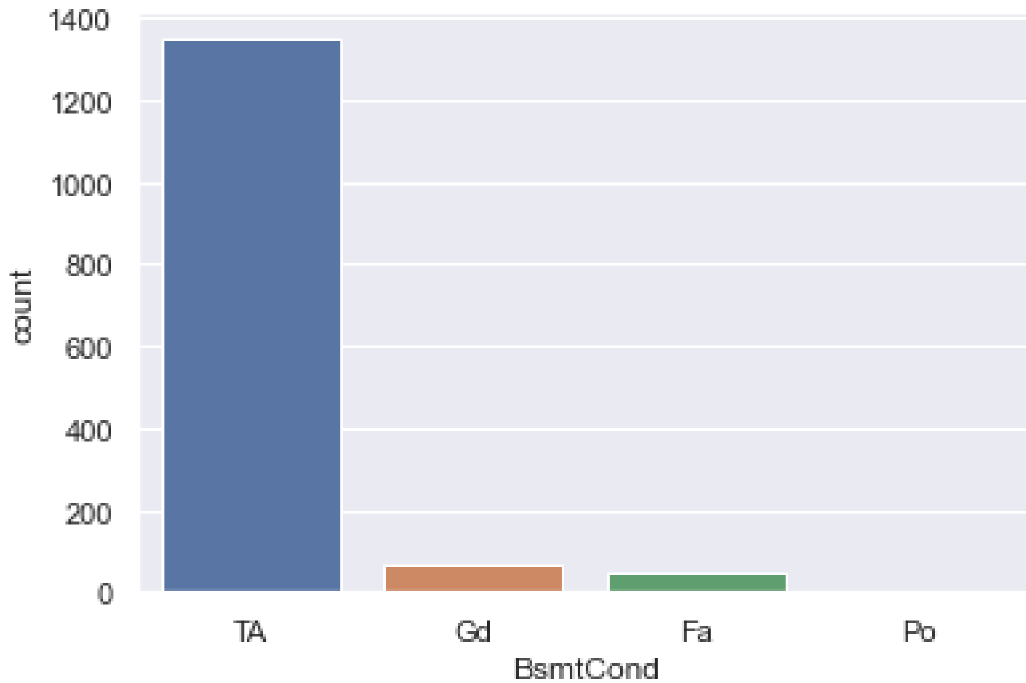
1200
1000
800
600
400
200
0
count
TA
Gd
Fa
Po
Ex
ExterCond



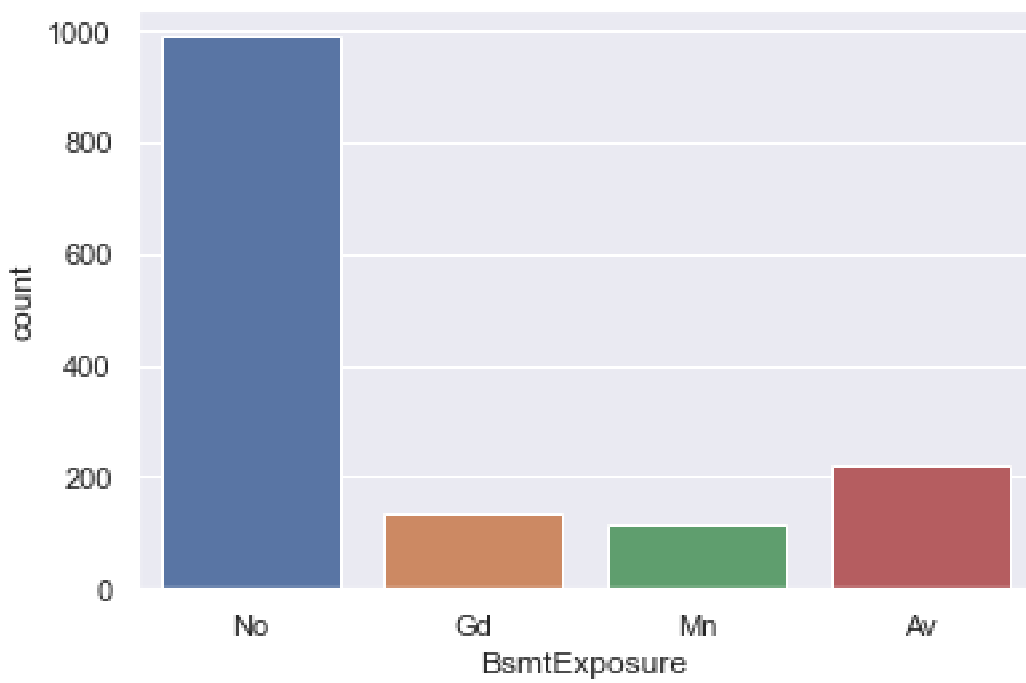
count
600
500
400
300
200
100
0
PConc
CBlock
BrkTil
Wood
Slab
Stone
Foundation



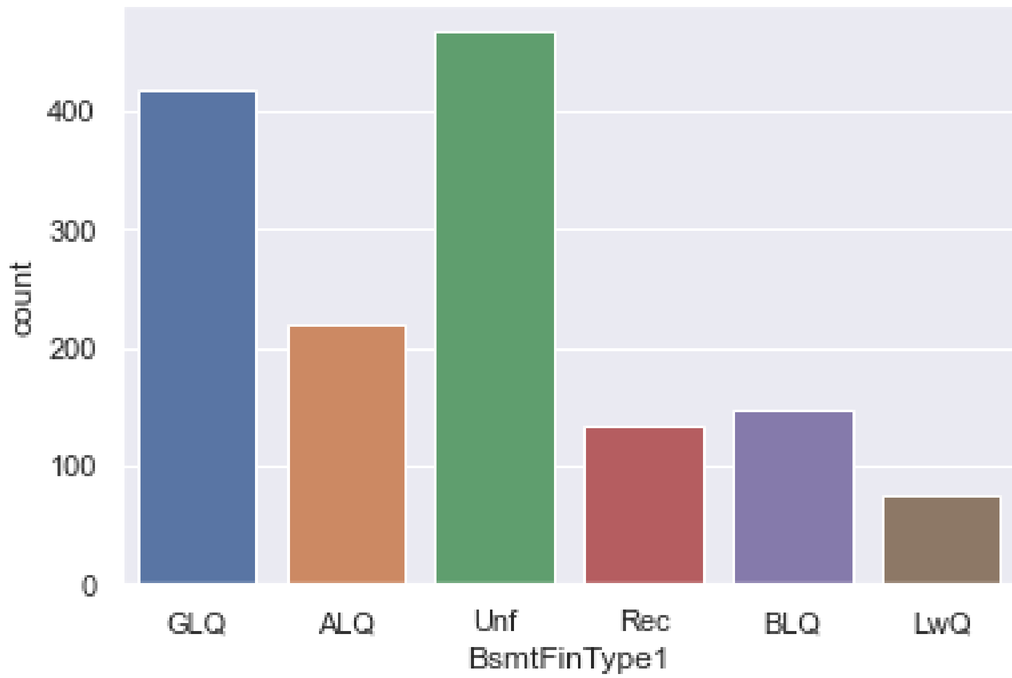
count
700
600
500
400
300
200
100
0
Gd
TA
Ex
Fa
BsmtQual

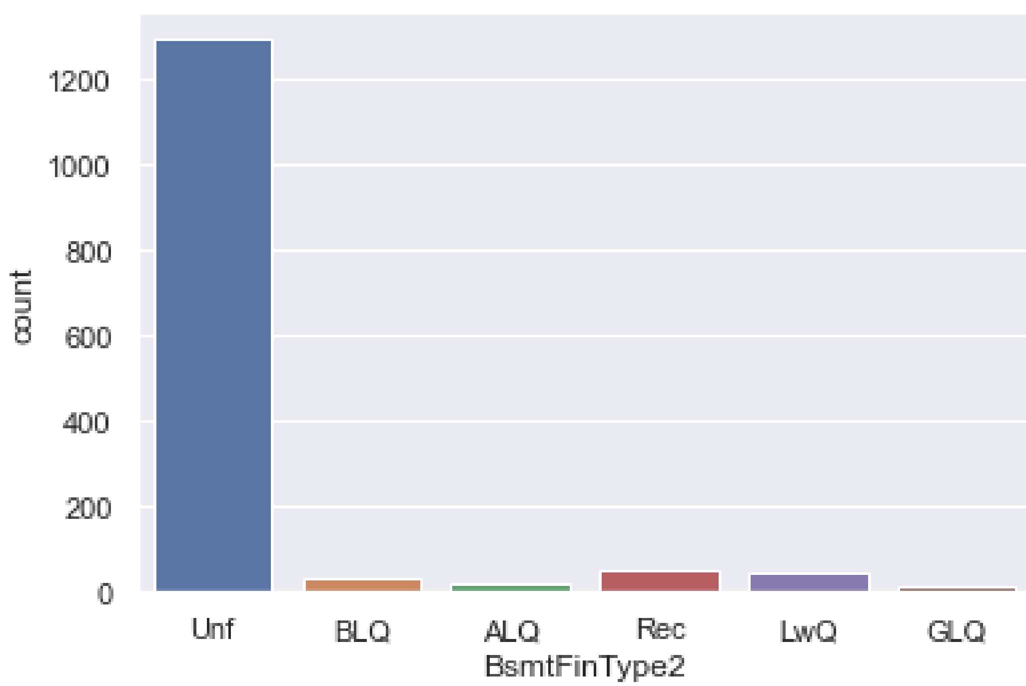


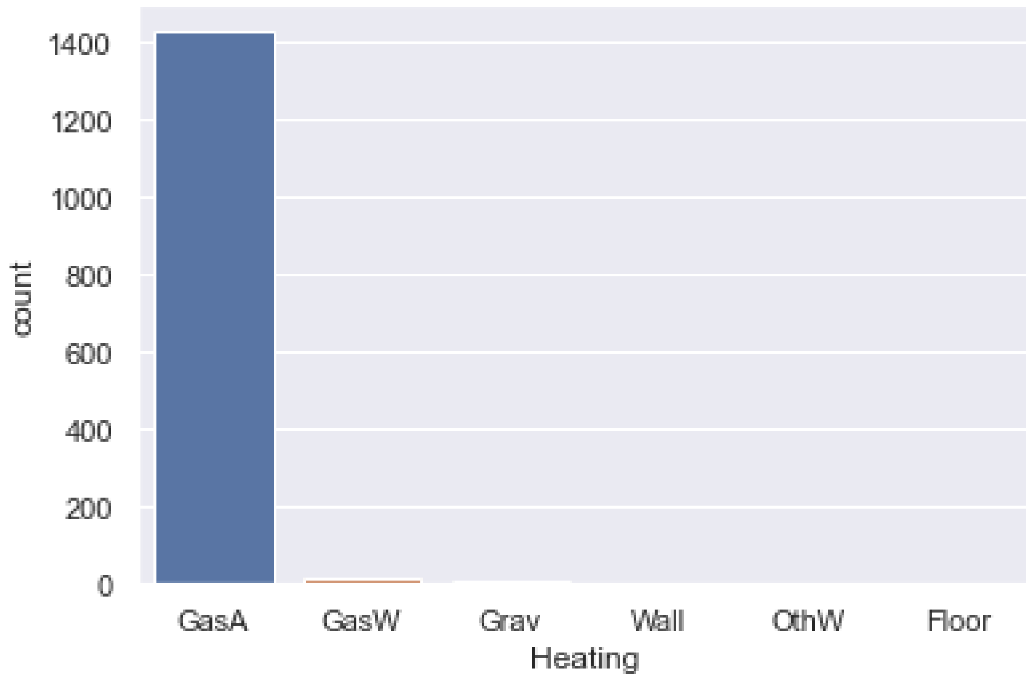
1400
1200
1000
800
600
400
200
0
count
TA
Gd
Fa
Po
BsmtCond



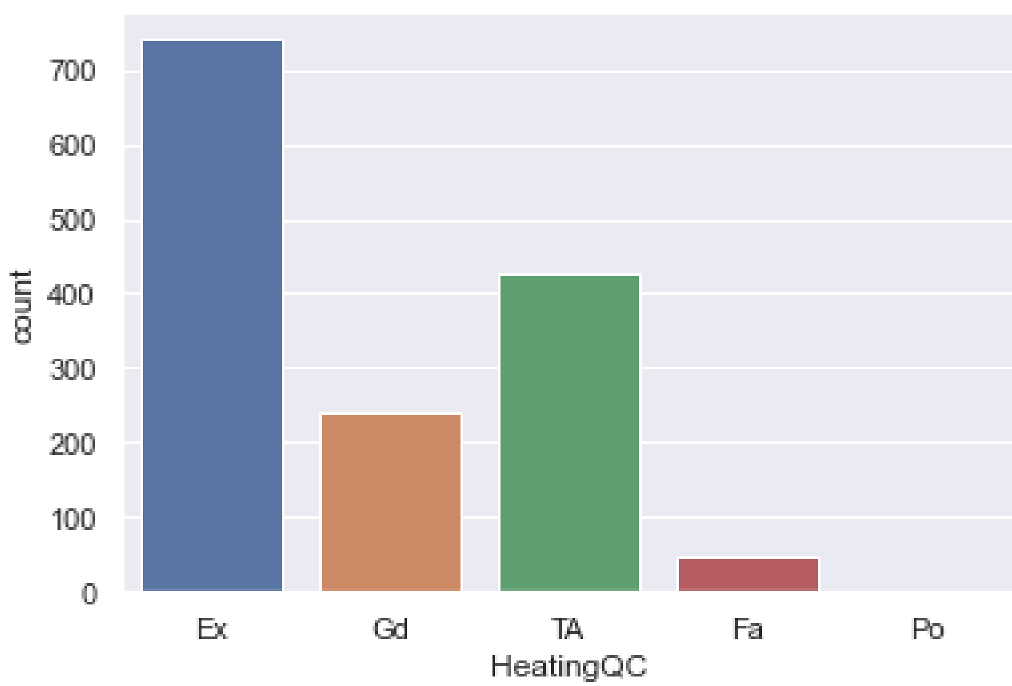
1000
800
600
400
200
0
count
No
Gd
Mn
Av
BsmtExposure


count
400
300
200
100
0
GLQ
ALQ
Unf
Rec
BLQ
LwQ
BsmtFinType1

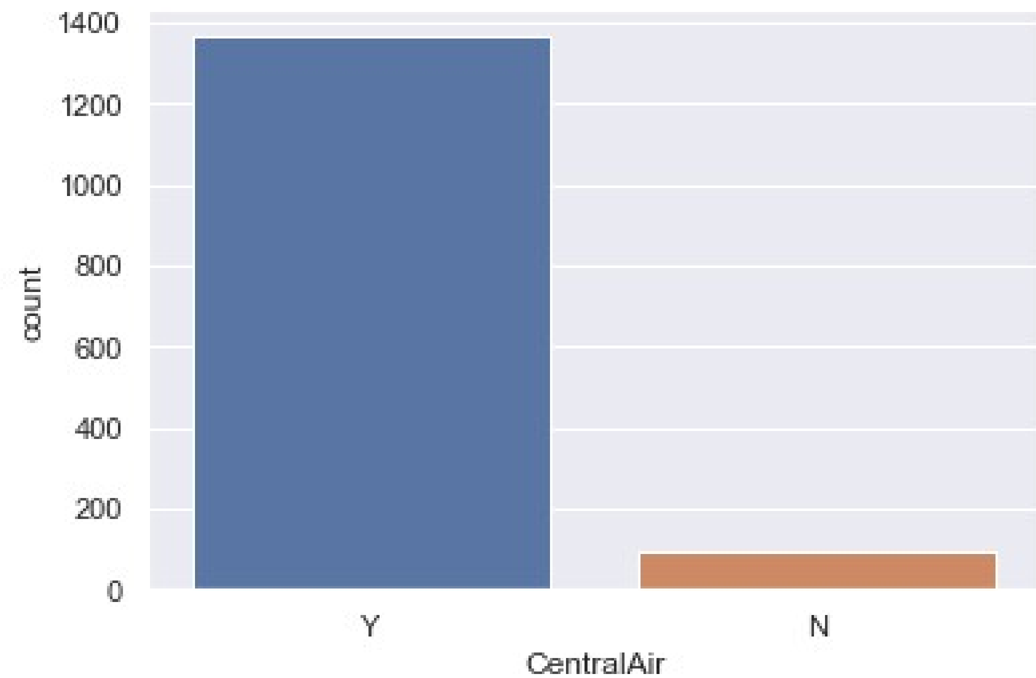

count
1200
1000
800
600
400
200
0
Unf
BLQ
ALQ
Rec
LwQ
GLQ
BsmtFinType2

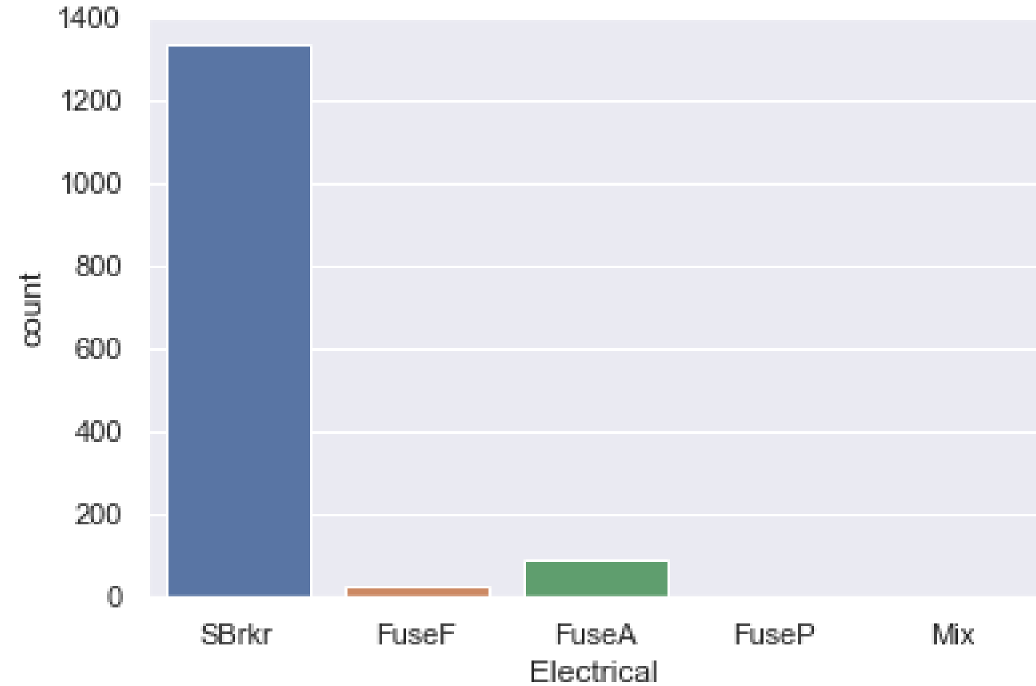


1400
1200
1000
800
600
400
200
0
count
GasA
GasW
Grav
Wall
OthW
Floor
Heating

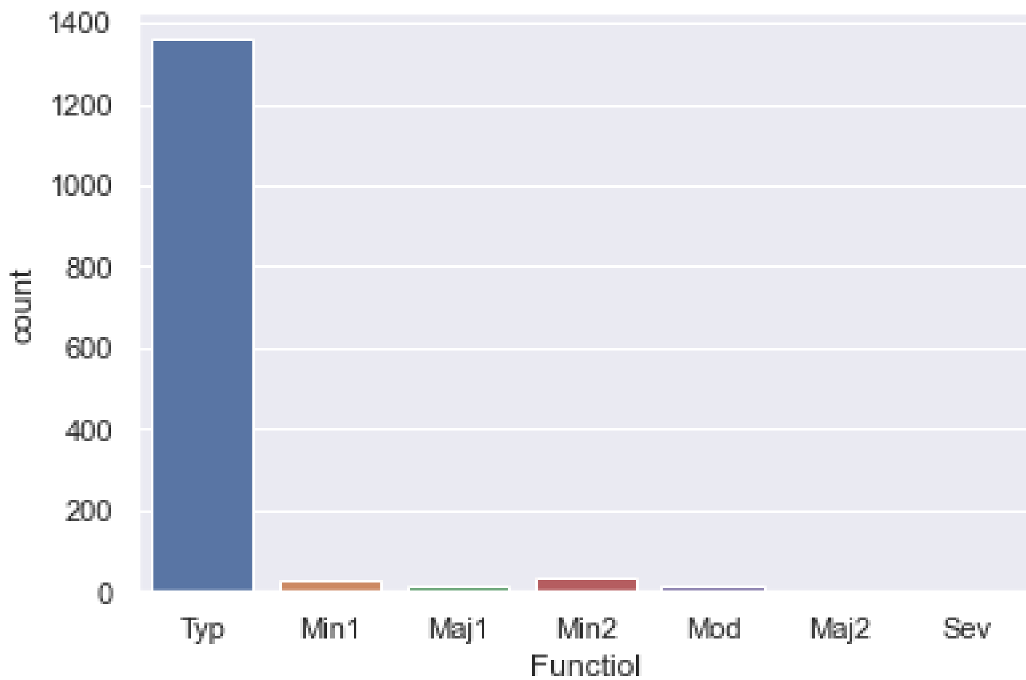


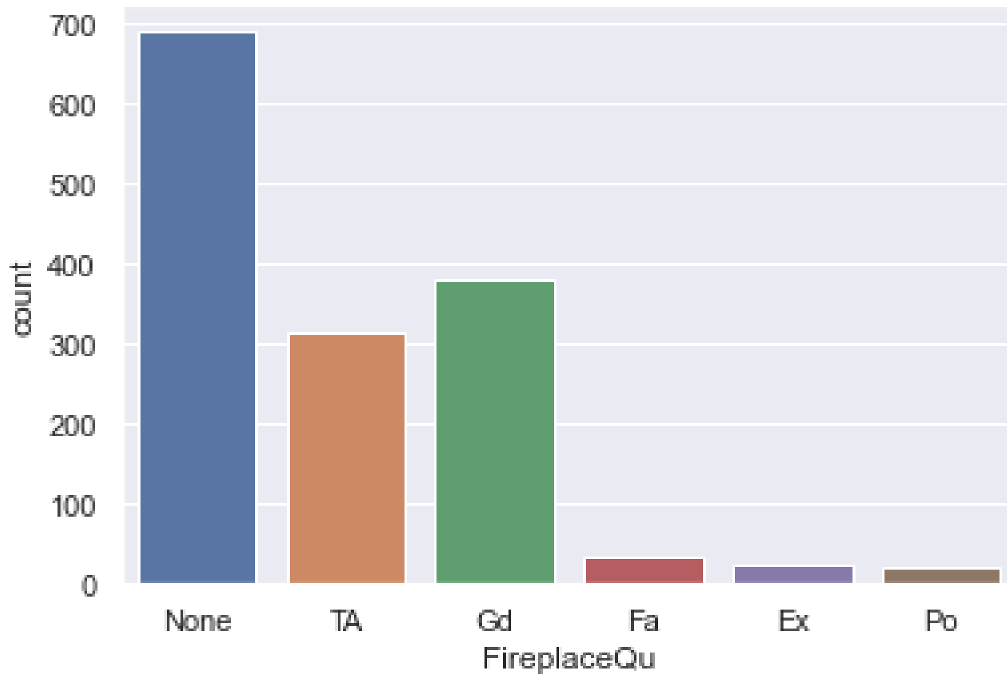
700
600
500
400
300
200
100
0
count
Ex
Gd
TA
Fa
Po
HeatingQC

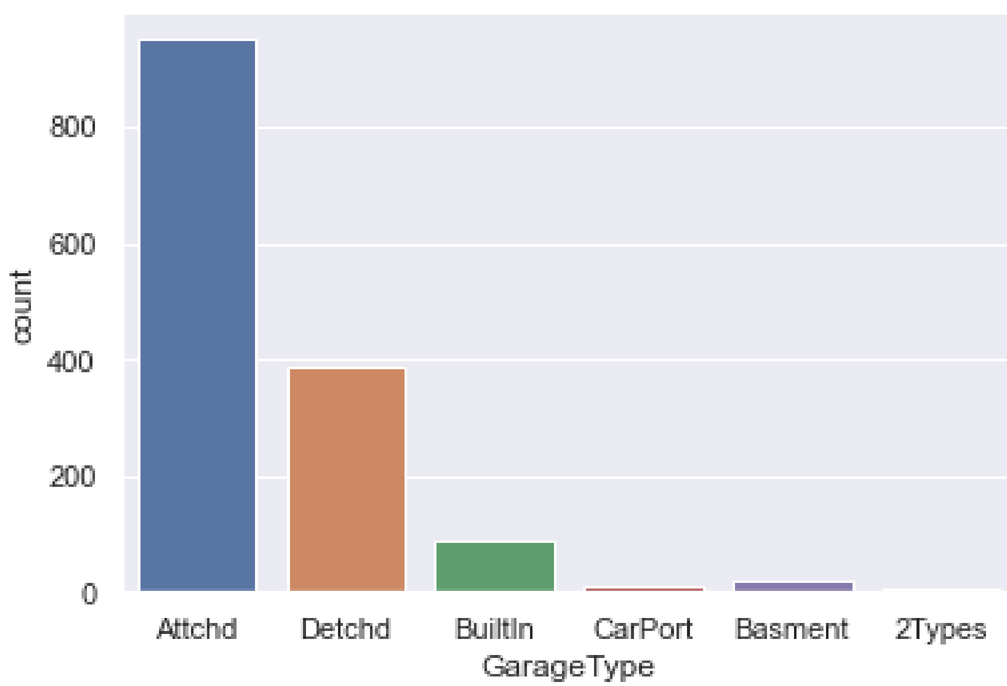

1400
1200
1000
800
600
400
200
0
count
Y
N
CentralAir

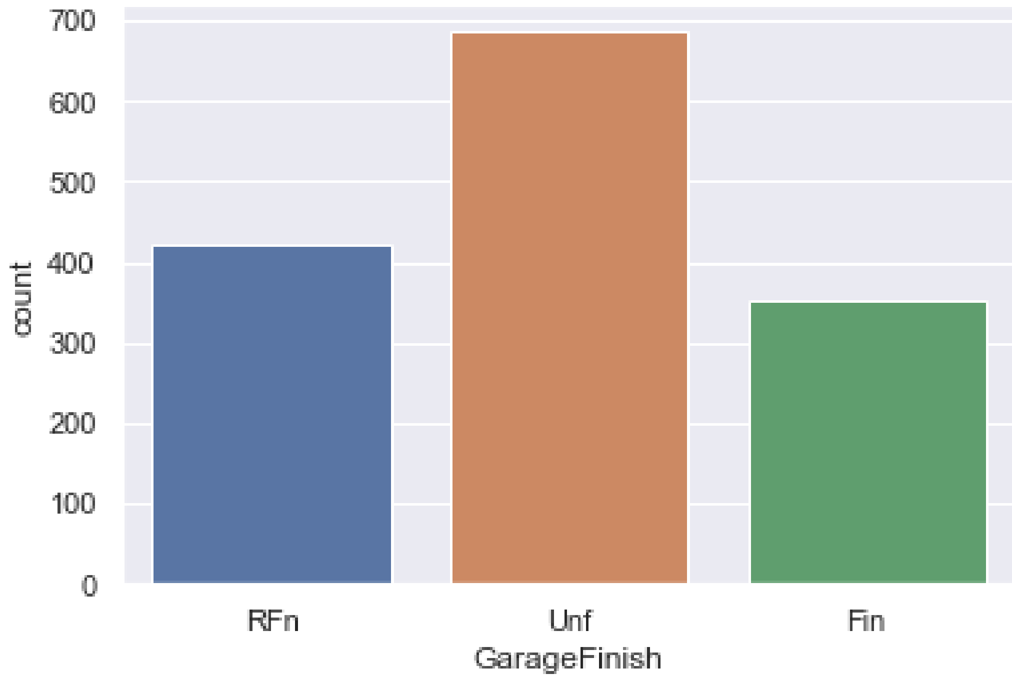

1400
1200
1000
800
600
400
200
0
count
SBrkr
FuseF
FuseA
FuseP
Mix
Electrical

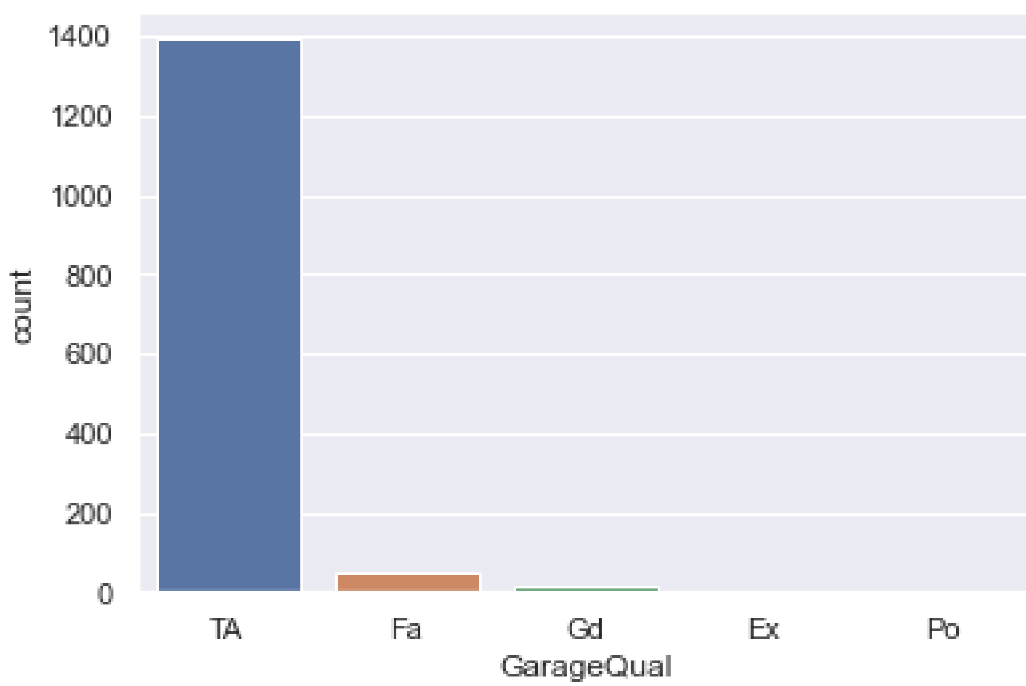

700
600
500
400
300
200
100
0
count
Gd
TA
Ex
Fa
KitchenQual

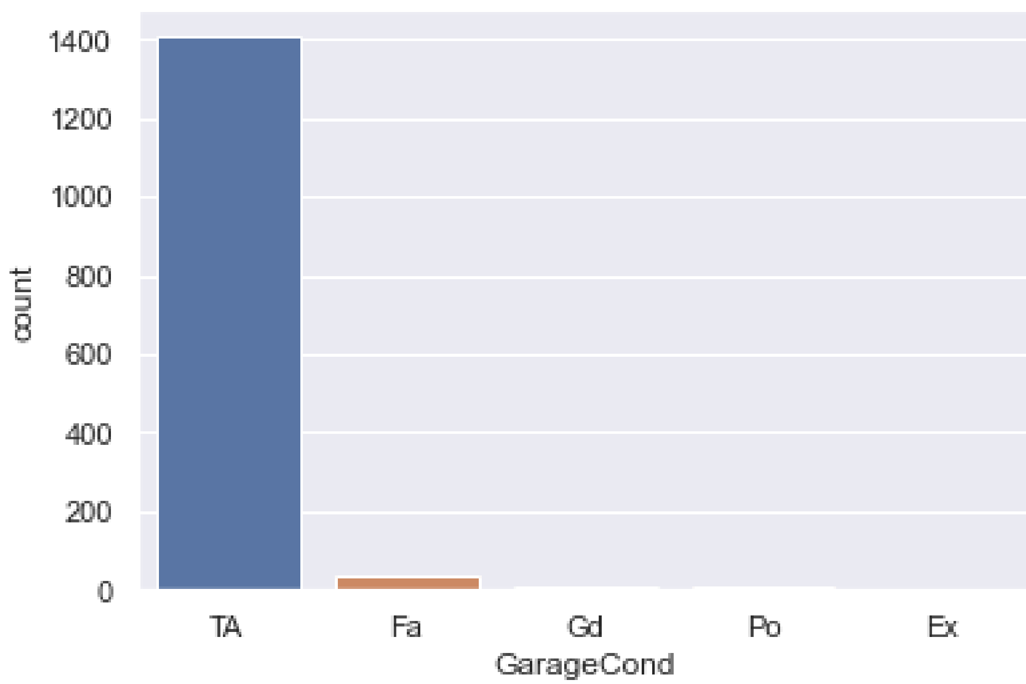

1400
1200
1000
800
600
400
200
0
count
Typ
Min1
Maj1
Min2
Mod
Maj2
Sev
Functiol

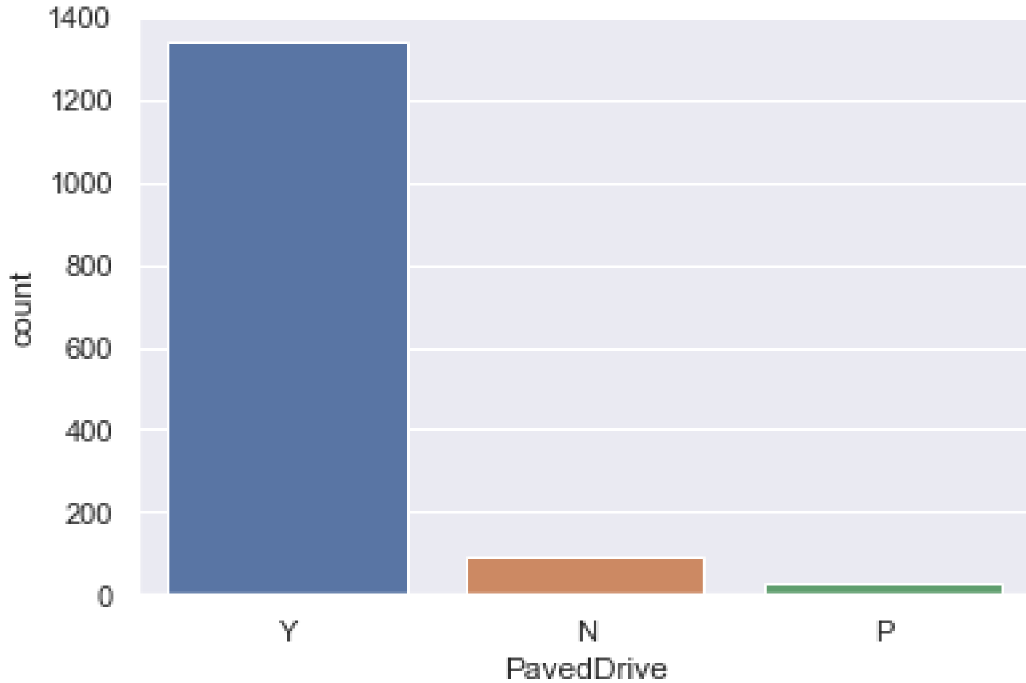

700
600
500
400
300
200
100
0
count
None
TA
Gd
Fa
Ex
Po
FireplaceQu

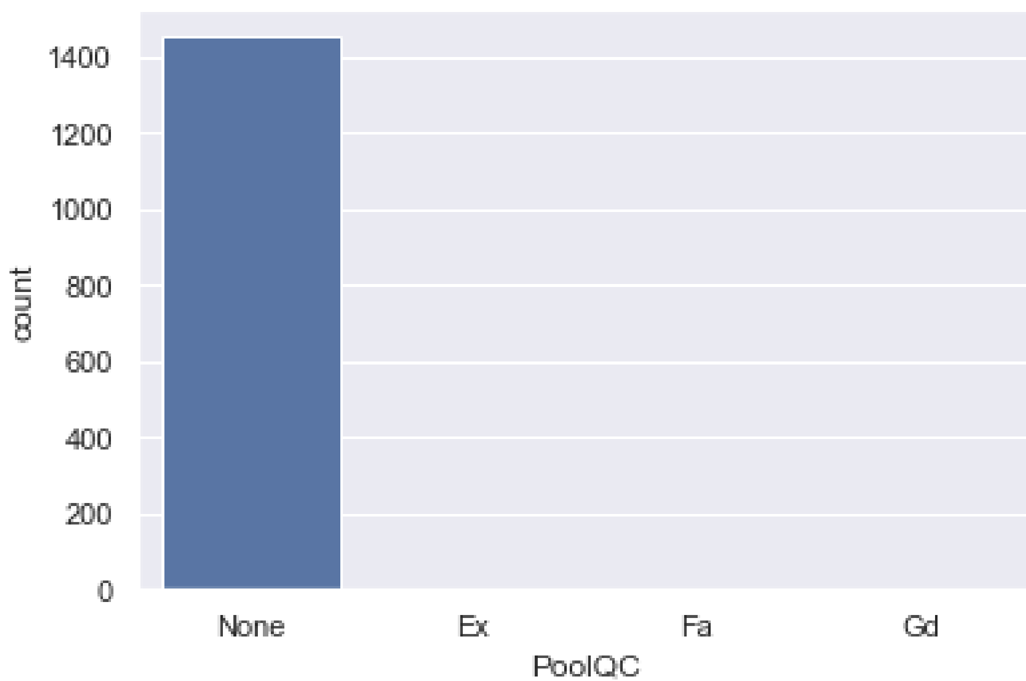

800
600
400
200
0
count
Attchd
Detchd
BuiltIn
CarPort
Basment
2Types
GarageType

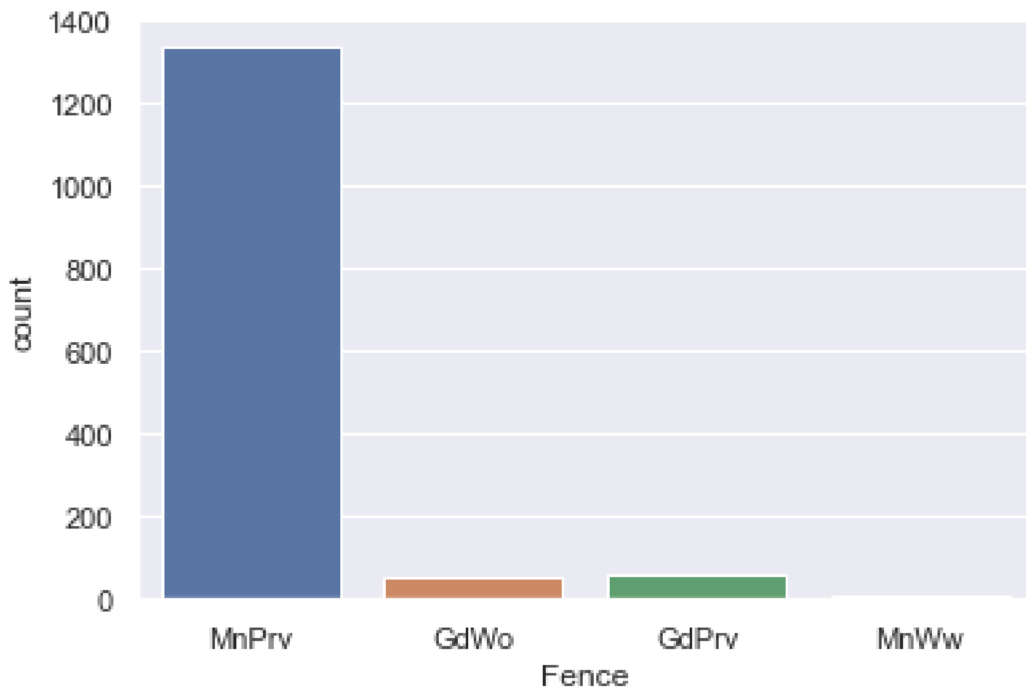

700
600
500
400
300
200
100
0
count
RFn
Unf
Fin
GarageFinish

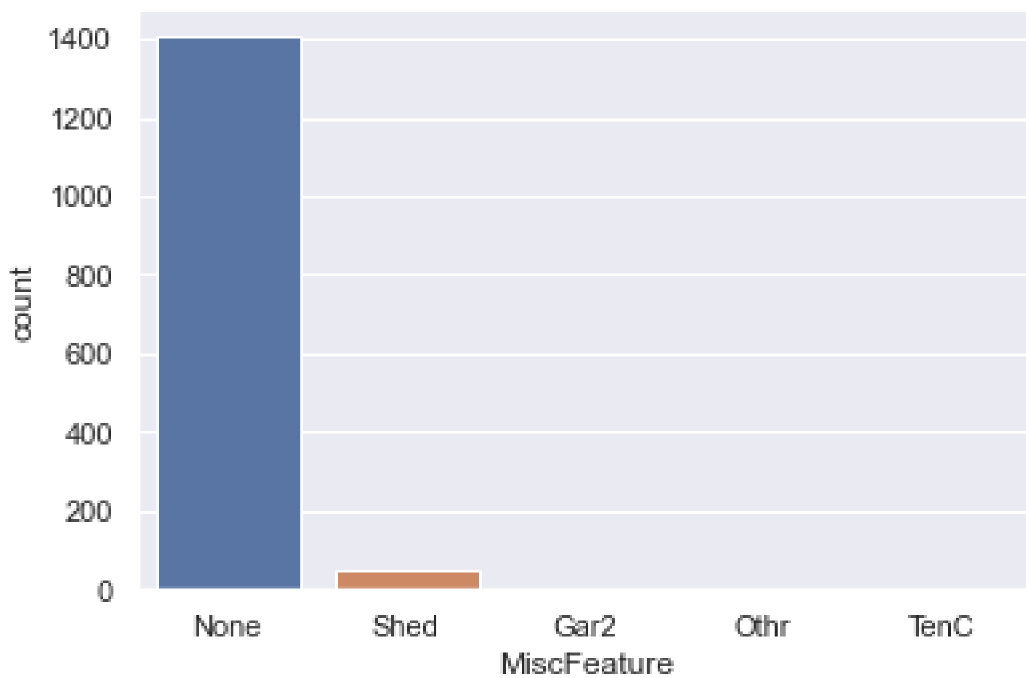

1400
1200
1000
800
600
400
200
0
count
TA
Fa
Gd
Ex
Po
GarageQual

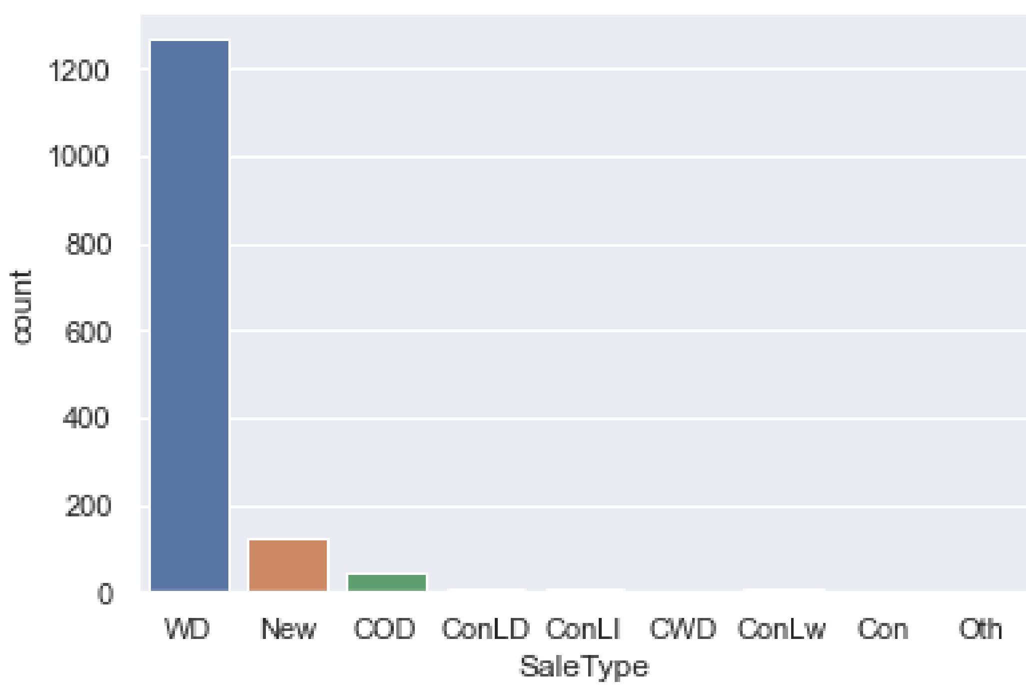

1400
1200
1000
800
600
400
200
0
count
TA
Fa
Gd
Po
Ex
GarageCond

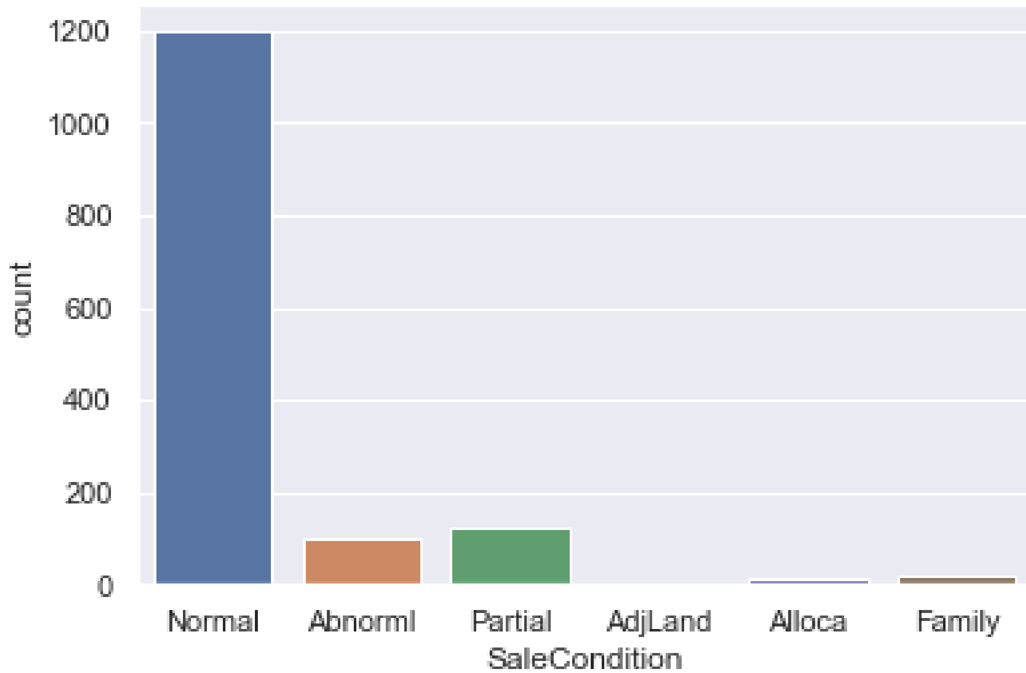

1400
1200
1000
800
600
400
200
0
count
Y
N
P
PavedDrive


1400
1200
1000
800
600
400
200
0
count
None
Ex
Fa
Gd
PoolQC

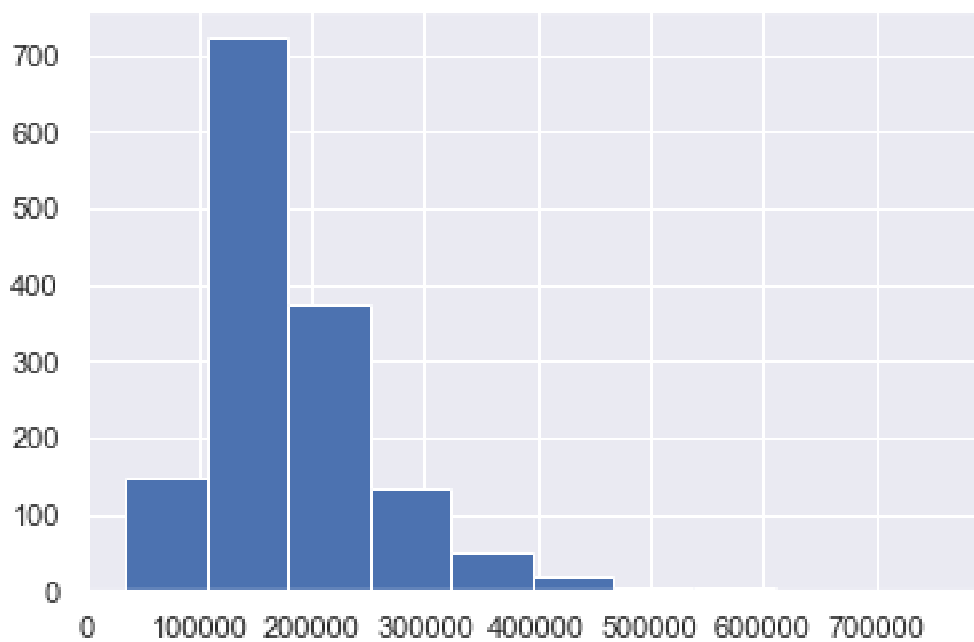

1400
1200
1000
800
600
400
200
0
count
MnPrv
GdWo
GdPrv
MnWw
Fence


1400
1200
1000
800
600
400
200
0
count
None
Shed
Gar2
Othr
TenC
MiscFeature


1200
1000
800
600
400
200
0
count
WD
New
COD
ConLD
ConLI
CWD
ConLw
Con
Oth
SaleType




```
num_var['SalePrice'].hist()
SalePriceSegmentation = num_var['SalePrice'].copy()
```




```
#Grouping the price to carrry out the cat-cat test
for price in range(len(SalePriceSegmentation)):
    if SalePriceSegmentation[price]<=100000:
        SalePriceSegmentation[price] = 'P<=100000'
    elif ((SalePriceSegmentation[price]>100000) and
(SalePriceSegmentation[price]<=200000)):
```

```
        SalePriceSegmentation[price] = '100000<P<=200000'
    elif ((SalePriceSegmentation[price]>200000) and
(SalePriceSegmentation[price]<=300000)):
         SalePriceSegmentation[price] = '200000<P<=300000'
    elif ((SalePriceSegmentation[price]>300000) and
(SalePriceSegmentation[price]<=400000)):
        SalePriceSegmentation[price] = '300000<P<=400000'
    elif ((SalePriceSegmentation[price]>100000) and
(SalePriceSegmentation[price]<=200000)):
        SalePriceSegmentation[price] = '400000<P<=500000'
    else:
        SalePriceSegmentation[price] = '500000<P'

#c. Identify significant variables using p-values and Chi-Square
values
Y = SalePriceSegmentation.astype(str)
alpha  = 0.05
for col in cols:
    X = cat_var[col].astype(str)
    dfObserved = pd.crosstab(Y, X)
    chi2, p, dof, expected = stats.chi2_contingency(dfObserved.values)
    result = ""
    if p < alpha:
        result = "{:15s} {} is IMPORTANT for Prediction".format(col,
p)
    else:
        result = "{:15s} {} is NOT an important predictor. (Discard {}
from model)".format(col, p, col)
    print(result)
```

```
MSZoning        2.5282273407169786e-39 is IMPORTANT for Prediction
Street          0.2751226012422085 is NOT an important predictor.
(Discard Street from model)
Alley           2.655900620446592e-05 is IMPORTANT for Prediction
LotShape        1.429883200516838e-21 is IMPORTANT for Prediction
LandContour     1.4009736755306051e-05 is IMPORTANT for Prediction
Utilities       0.9625277056368384 is NOT an important predictor.
(Discard Utilities from model)
LotConfig       9.170348236154635e-05 is IMPORTANT for Prediction
LandSlope       0.3600405418689199 is NOT an important predictor.
(Discard LandSlope from model)
Neighborhood    9.800523695735449e-178 is IMPORTANT for Prediction
Condition1      0.0003638430970265861 is IMPORTANT for Prediction
Condition2      0.06665143224538962 is NOT an important predictor.
(Discard Condition2 from model)
BldgType        1.2566195087358137e-09 is IMPORTANT for Prediction
HouseStyle      1.0600142036940675e-16 is IMPORTANT for Prediction
RoofStyle       5.556433352503948e-14 is IMPORTANT for Prediction
RoofMatl        0.0028622590308057952 is IMPORTANT for Prediction
Exterior1st     1.3904485257029132e-50 is IMPORTANT for Prediction
Exterior2nd     8.7500422503955365e-47 is IMPORTANT for Prediction
```

```
MasVnrType      5.196399589724438e-49 is IMPORTANT for Prediction
ExterQual       1.55420777592641e-197 is IMPORTANT for Prediction
ExterCond       3.7011459223473924e-12 is IMPORTANT for Prediction
Foundation      1.1785264199530343e-72 is IMPORTANT for Prediction
BsmtQual        3.880577555736088e-175 is IMPORTANT for Prediction
BsmtCond        3.3994117018631147e-11 is IMPORTANT for Prediction
BsmtExposure    3.401498106568333e-33 is IMPORTANT for Prediction
BsmtFinType1    2.889580441813886e-53 is IMPORTANT for Prediction
BsmtFinType2    0.02814393466453002 is IMPORTANT for Prediction
Heating         4.336687510043468e-15 is IMPORTANT for Prediction
HeatingQC       3.012633020937661e-62 is IMPORTANT for Prediction
CentralAir      2.448545148671904e-53 is IMPORTANT for Prediction
Electrical      5.489750623289287e-21 is IMPORTANT for Prediction
KitchenQual     6.78946268698469e-173 is IMPORTANT for Prediction
Functiol        0.00208969002553416 is IMPORTANT for Prediction
FireplaceQu     3.605469235823506e-82 is IMPORTANT for Prediction
GarageType      3.9876796004607706e-36 is IMPORTANT for Prediction
GarageFinish    3.34059578076278e-82 is IMPORTANT for Prediction
GarageQual      9.543903023088993e-10 is IMPORTANT for Prediction
GarageCond      1.796654195081666e-08 is IMPORTANT for Prediction
PavedDrive      1.5663434213380232e-29 is IMPORTANT for Prediction
PoolQC          0.005010663275726978 is IMPORTANT for Prediction
Fence           0.0003956948783349849 is IMPORTANT for Prediction
MiscFeature     0.21585761169156817 is NOT an important predictor.
(Discard MiscFeature from model)
SaleType        3.9351164557013877e-32 is IMPORTANT for Prediction
SaleCondition   8.507191296689659e-37 is IMPORTANT for Prediction
```

5. Combining significant variables

```
#Dropping all numerical features below 5% coorelation with the target
and all Categorical features that failed the test or have 1200+ of the
same category.
dropped_num_cols =
['TotRmsAbvGrd','YrSold','MoSold','MiscVal','3SsnPorch','BsmtHalfBath'
,'LowQualFinSF','BsmtFinSF2']
dropped_cat_cols =
['Street','Alley','LandContour','Utilities','LandSlope','Condition1','
Condition2','BldgType','RoofMatl','BsmtCond','BsmtFinType2','Heating',
'CentralAir','Electrical','Functiol','GarageQual','GarageCond','PavedD
rive','PoolQC','Fence','MiscFeature','SaleType','SaleCondition']
dataset = pd.concat([num_var.drop(columns=dropped_num_cols),
cat_var.drop(columns=dropped_cat_cols)], axis=1)
dataset.head()
```

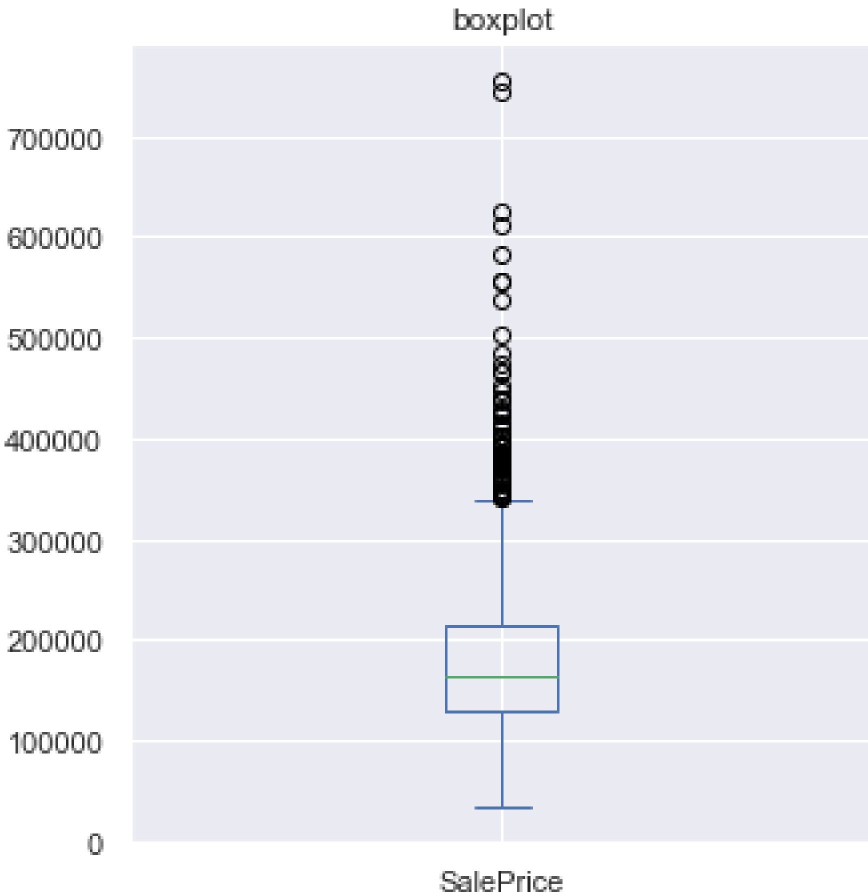
```
   Id  MSSubClass  LotFrontage  LotArea  OverallQual  OverallCond
YearBuilt  \
0   1          60         65.0     8450            7            5
2003
1   2          20         80.0     9600            6            8
1976
```

```
2   3          60          68.0    11250             7             5
2001
3   4          70          60.0     9550             7             5
1915
4   5          60          84.0    14260             8             5
2000

   YearRemodAdd  MasVnrArea  BsmtFinSF1  ...  ExterCond  Foundation
BsmtQual  \
0          2003       196.0         706  ...         TA       PConc
Gd
1          1976         0.0         978  ...         TA      CBlock
Gd
2          2002       162.0         486  ...         TA       PConc
Gd
3          1970         0.0         216  ...         TA      BrkTil
TA
4          2000       350.0         655  ...         TA       PConc
Gd

   BsmtExposure  BsmtFinType1  HeatingQC  KitchenQual  FireplaceQu  \
0            No           GLQ         Ex           Gd         None
1            Gd           ALQ         Ex           TA           TA
2            Mn           GLQ         Ex           Gd           TA
3            No           ALQ         Gd           Gd           Gd
4            Av           GLQ         Ex           Gd           TA

   GarageType  GarageFinish
0      Attchd           RFn
1      Attchd           RFn
2      Attchd           RFn
3      Detchd           Unf
4      Attchd           RFn

[5 rows x 50 columns]
```

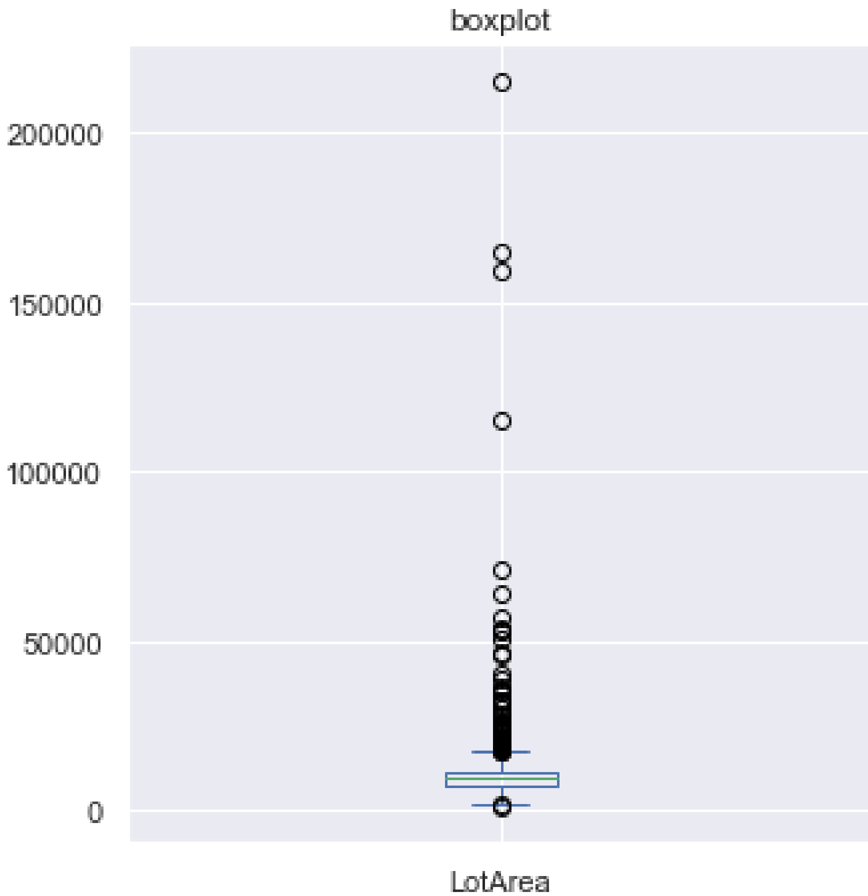
6. Plotting box plot for the new dataset

```
cols = dataset.columns.values.tolist()
#Removing some features for visualization purposes
dropcols = ['SalePrice','LotArea']
for col in dropcols:
    cols.remove(col)
ax = dataset[cols].plot(kind='box', title='boxplot')
plt.rcParams["figure.figsize"] = (200,5.5)
plt.show()
```

```
ax = dataset['SalePrice'].plot(kind='box', title='boxplot')
plt.rcParams["figure.figsize"] = (5,5.5)
plt.show()
```




```
ax = dataset['LotArea'].plot(kind='box', title='boxplot')
plt.rcParams["figure.figsize"] = (8,5.5)
plt.show()
```


boxplot
200000
150000
100000
50000
0
LotArea