Project Name:Advanced House Price Prediction

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.pandas.set_option('display.max_columns',None)
```

In [3]:
```
ds=pd.read_csv('train.csv')
```

In [4]:
```
print(ds.shape)
```

```
(1460, 81)
```

In [5]:
```
ds.head()
```

Out[5]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContou
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lv
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lv
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lv
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lv
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lv

In Data Analysis we will analyze to find out Missing values All the numerical variables Distribution of the numerical variables Categorical variables cardinality of categorical variables Outliers Relationship between independent and dependent features

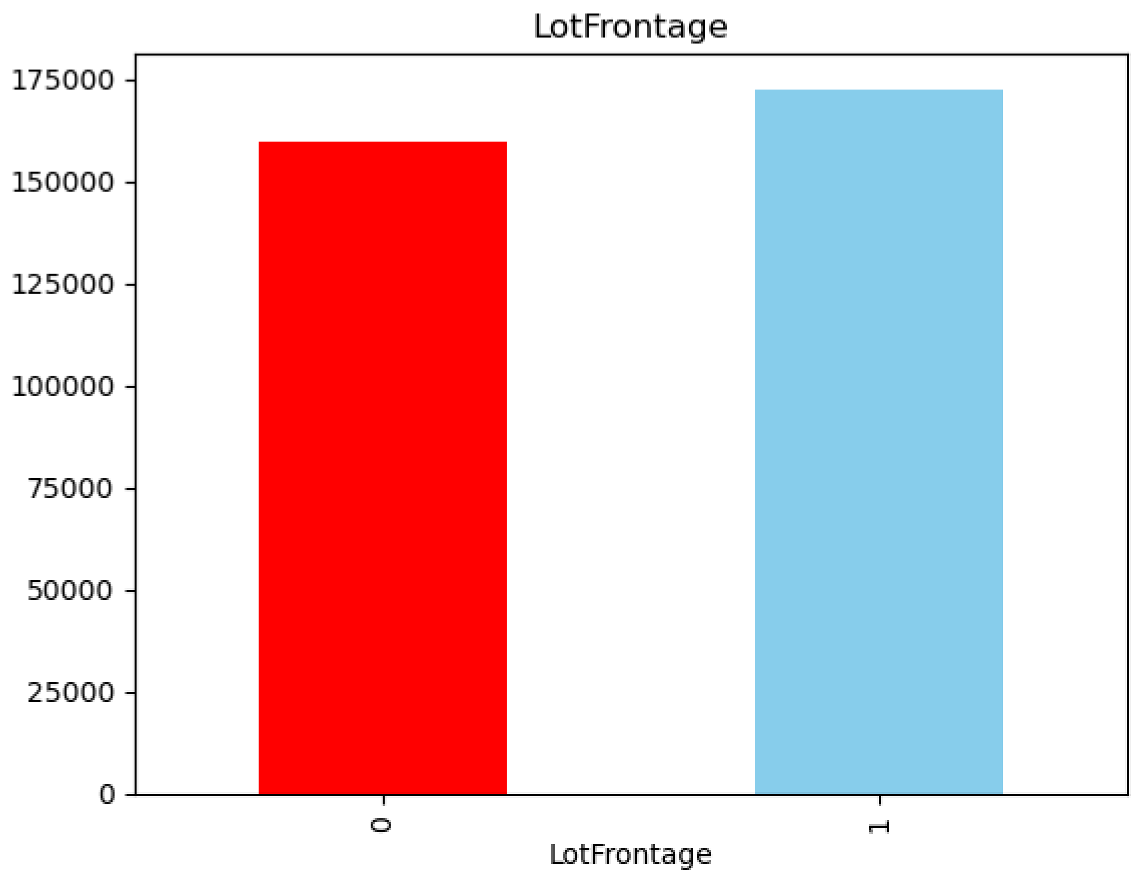
Missing Values

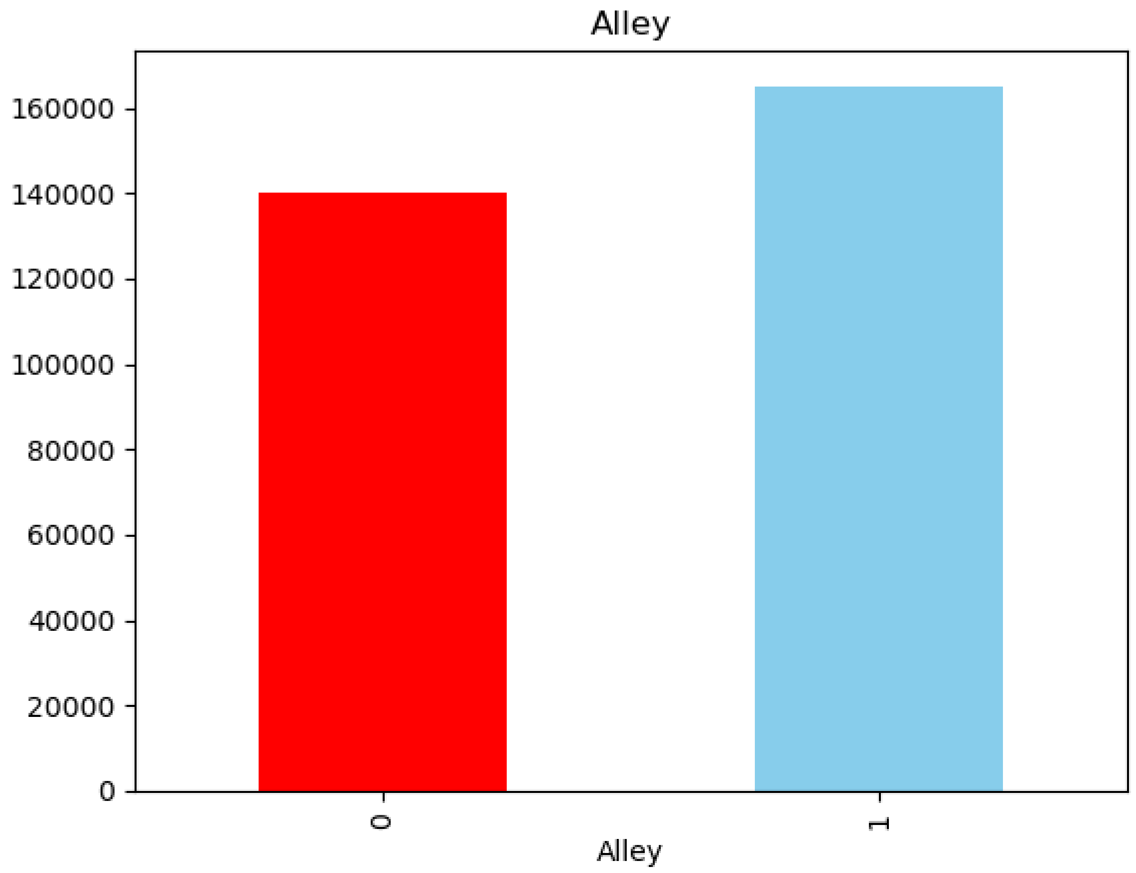
In [6]:
```
#check the percentage of nan values present in each feature
#step 1 -the list of features which has missing values
features_with_na=[feature for feature in ds.columns if ds[feature].isnull().sum()>1
##step-2 print the features name and the percentage of misssing values

for feature in features_with_na:
    print(feature, np.round(ds[feature].isnull().mean(), 4), '%missing values')
```

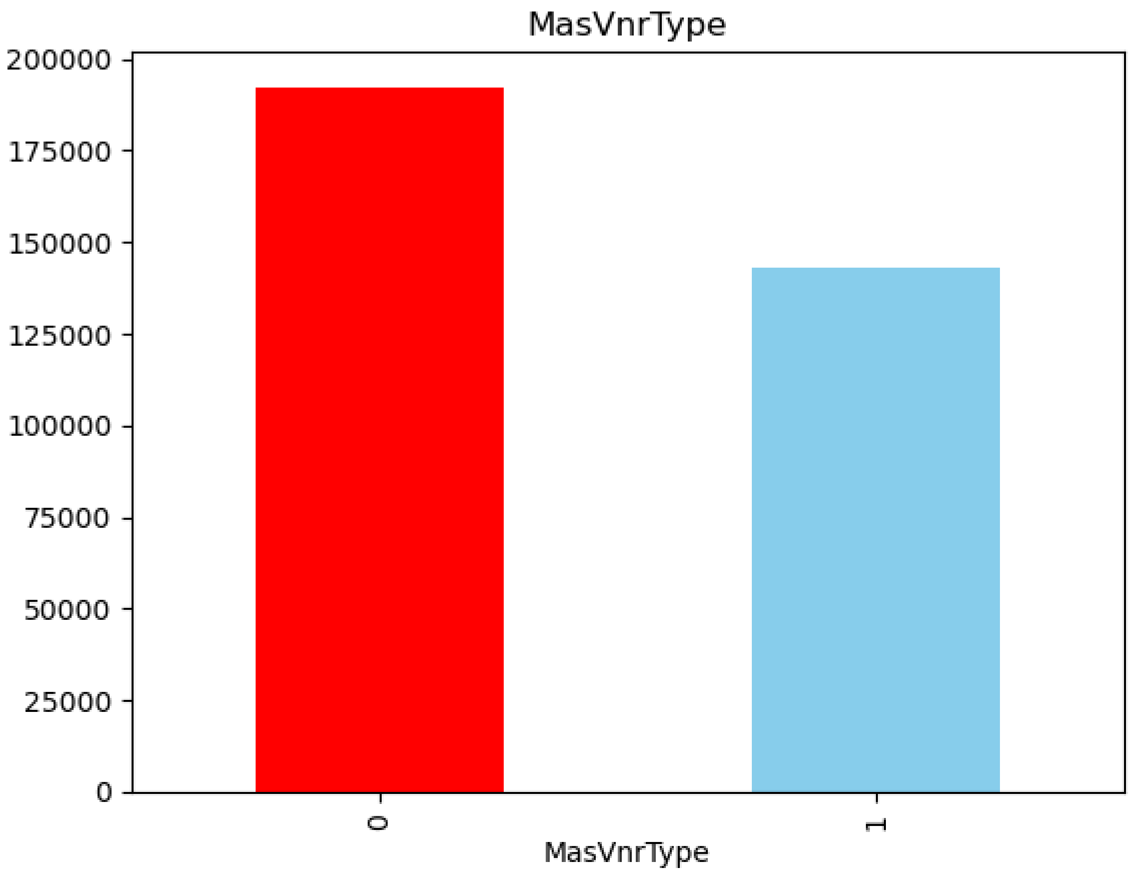
```
LotFrontage 0.1774 %missing values
Alley 0.9377 %missing values
MasVnrType 0.5973 %missing values
MasVnrArea 0.0055 %missing values
BsmtQual 0.0253 %missing values
BsmtCond 0.0253 %missing values
BsmtExposure 0.026 %missing values
BsmtFinType1 0.0253 %missing values
BsmtFinType2 0.026 %missing values
FireplaceQu 0.4726 %missing values
GarageType 0.0555 %missing values
GarageYrBlt 0.0555 %missing values
GarageFinish 0.0555 %missing values
GarageQual 0.0555 %missing values
GarageCond 0.0555 %missing values
PoolQC 0.9952 %missing values
Fence 0.8075 %missing values
MiscFeature 0.963 %missing values
```

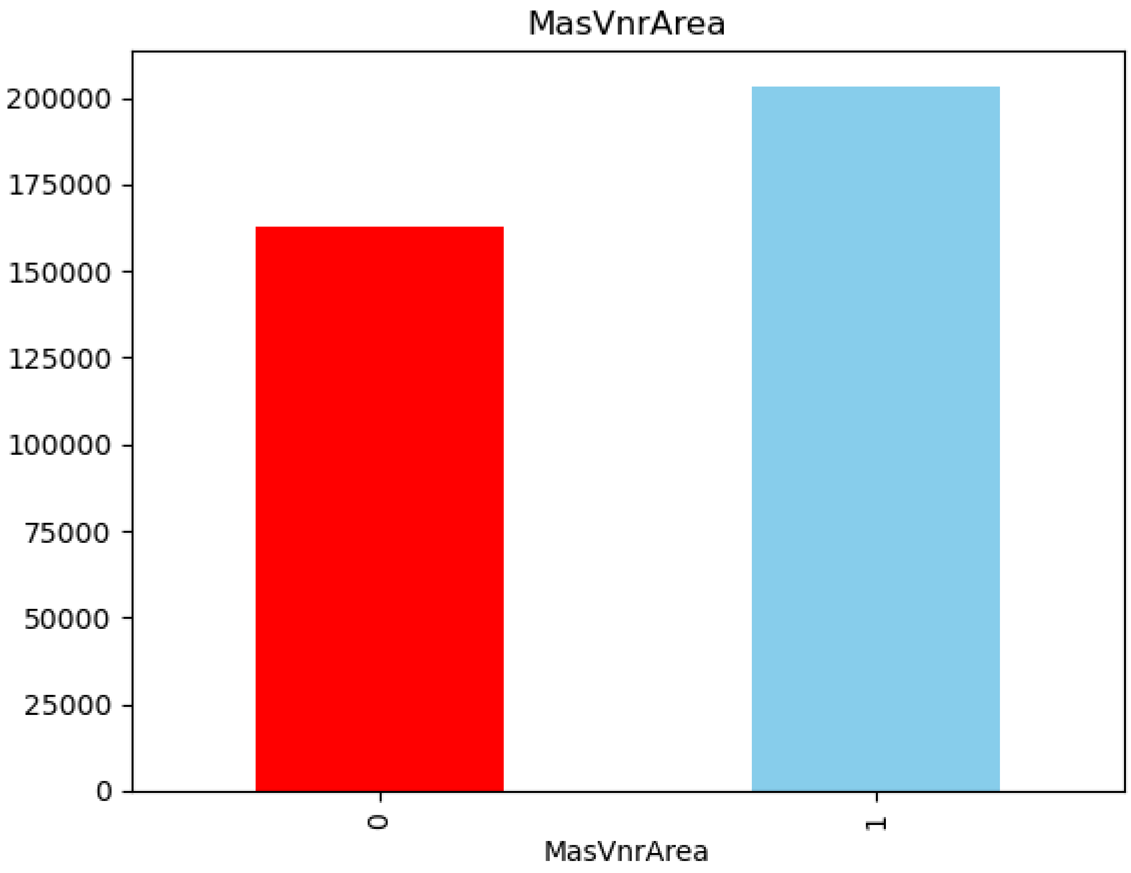
Since they are many missing values ,we need to find the relationship between missing values and sales price

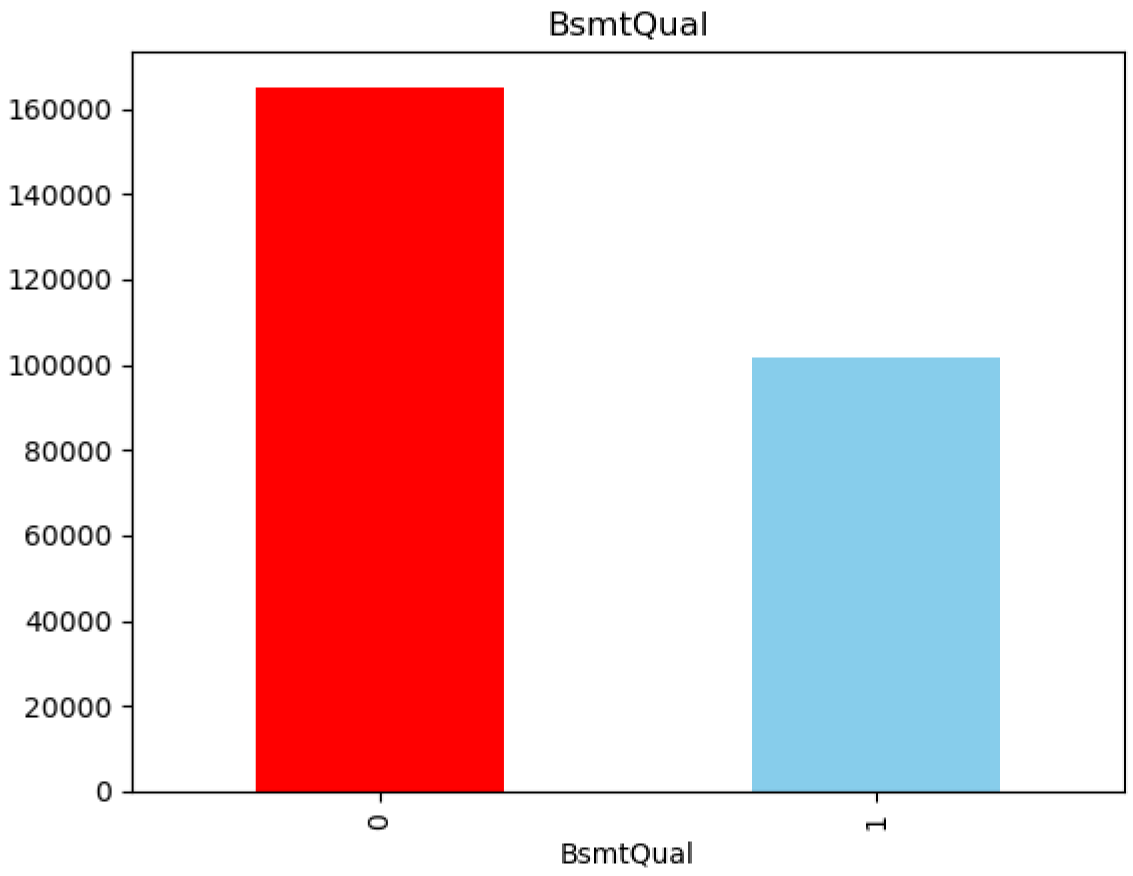
In [7]:
```
for feature in features_with_na:
    data=ds.copy()


    #lets make a variable that indicates 1 if the observation was missing or zero
    data[feature]=np.where(data[feature].isnull(),1,0)

    #lets calculate the mean Saleprice where the information missing or present
    data.groupby(feature)['SalePrice'].median().plot.bar(color=['red','skyblue']  #
                                                        )
    plt.title(feature)
    plt.show()
```

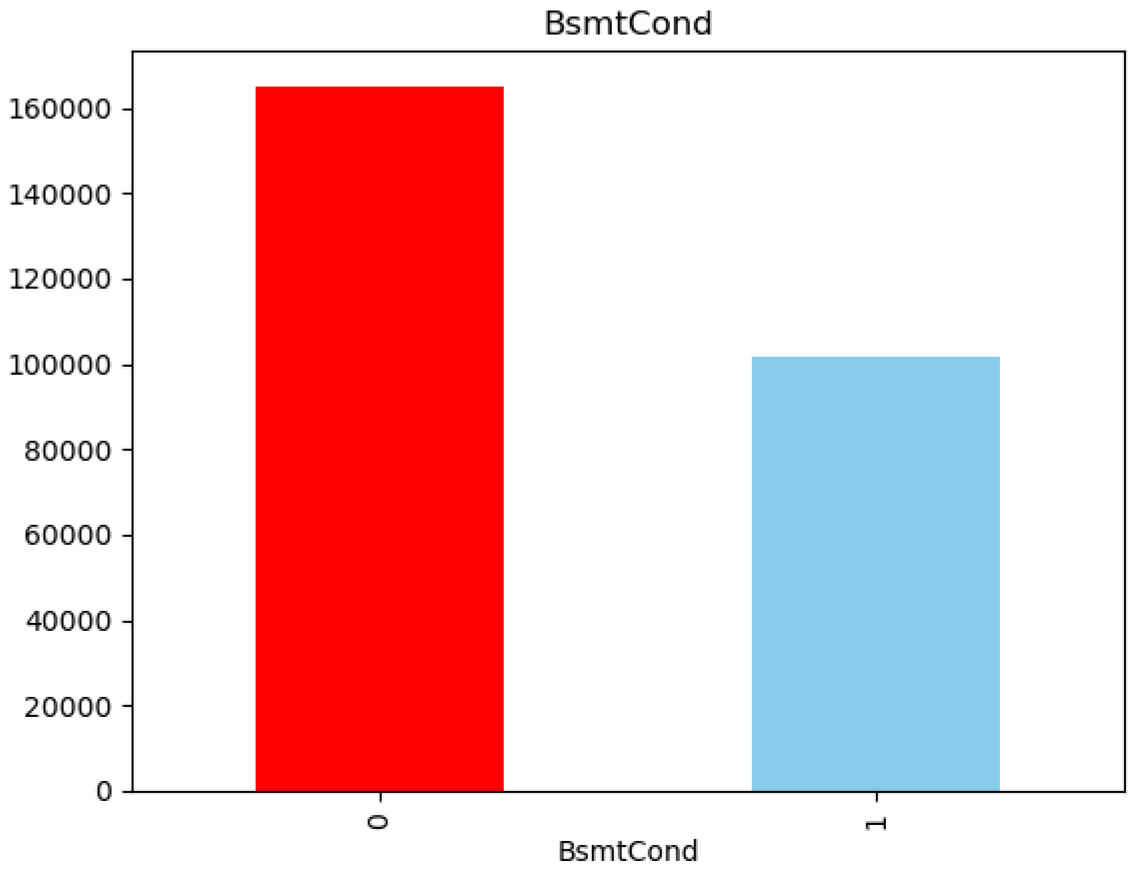

LotFrontage
175000
150000
125000
100000
75000
50000
25000
0
0
1
LotFrontage

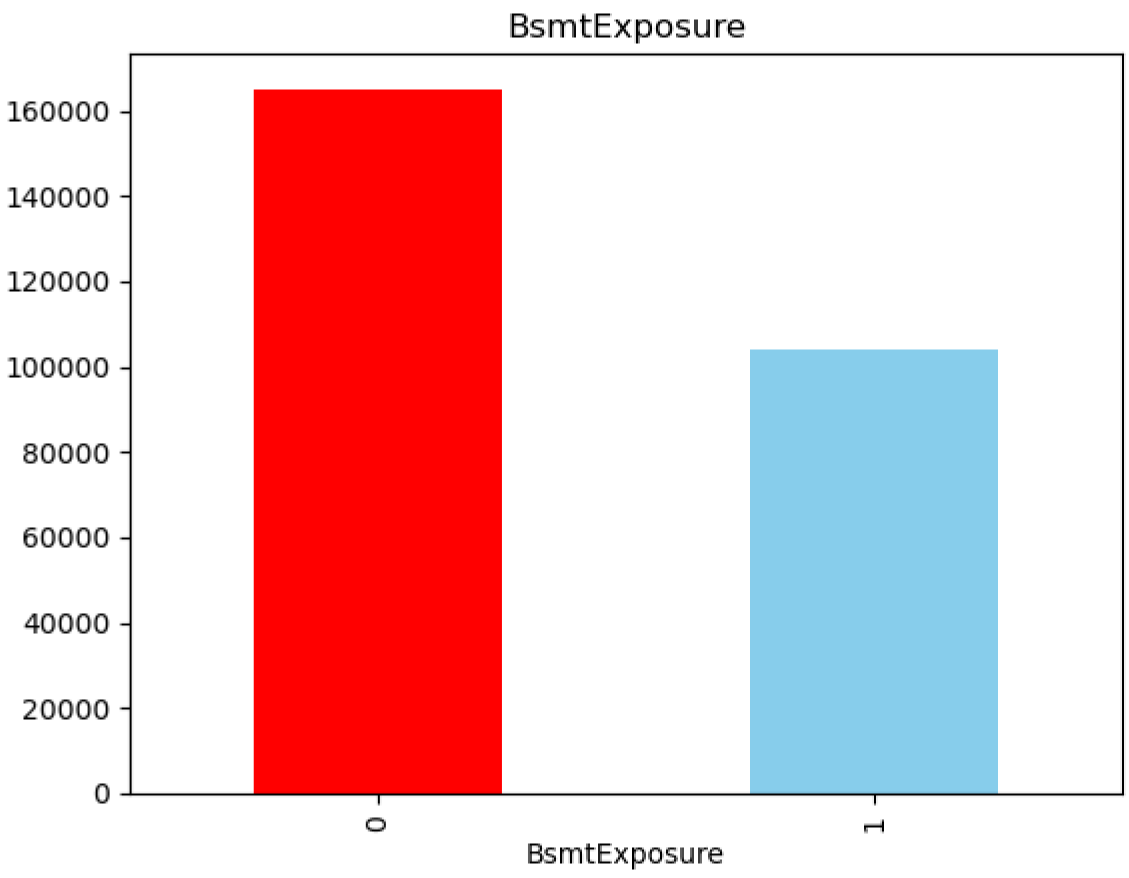

Alley
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
Alley

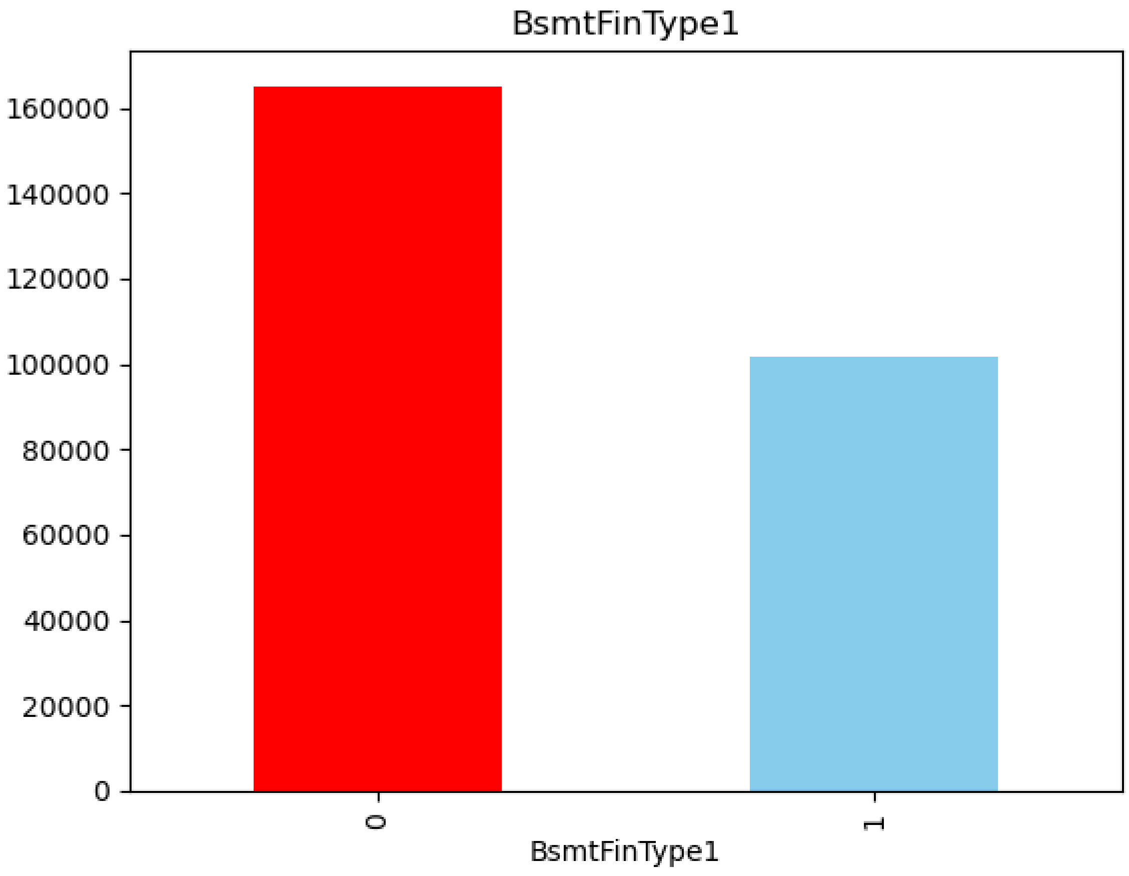

MasVnrType
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
MasVnrType

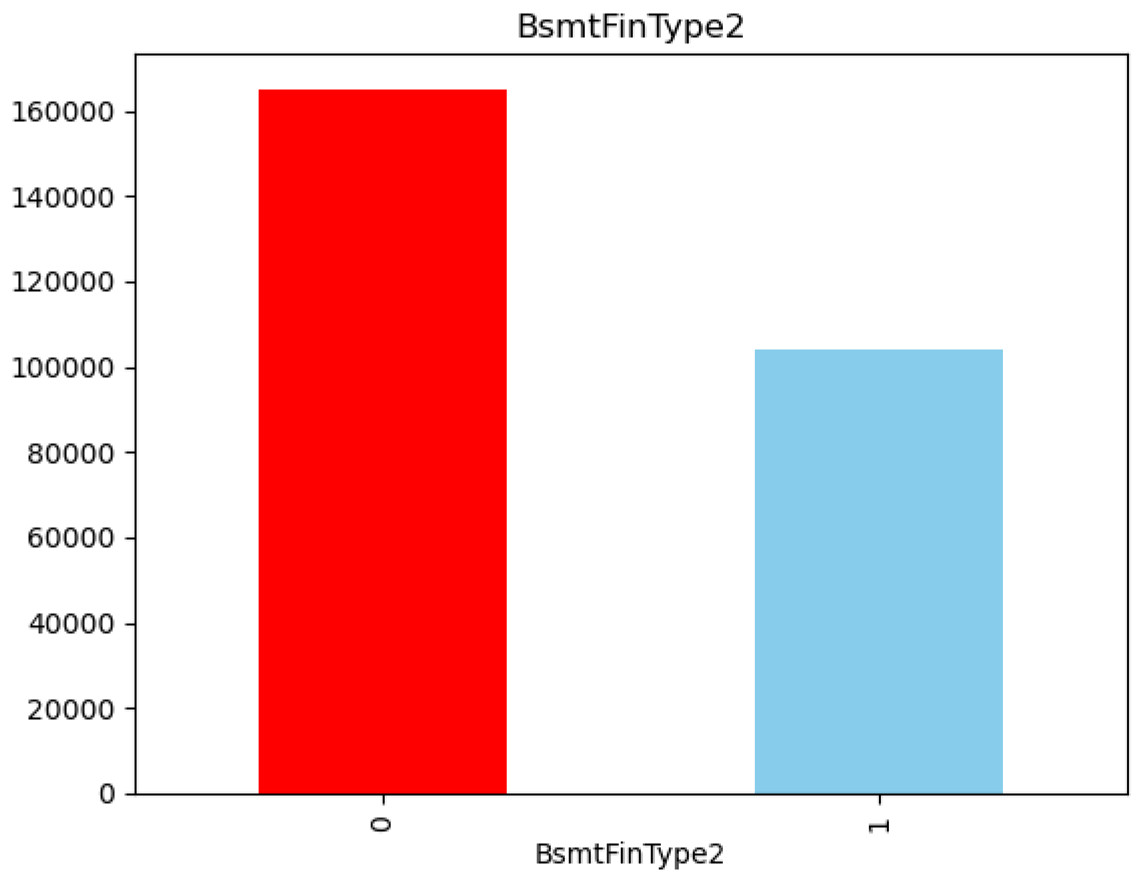

MasVnrArea
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
MasVnrArea

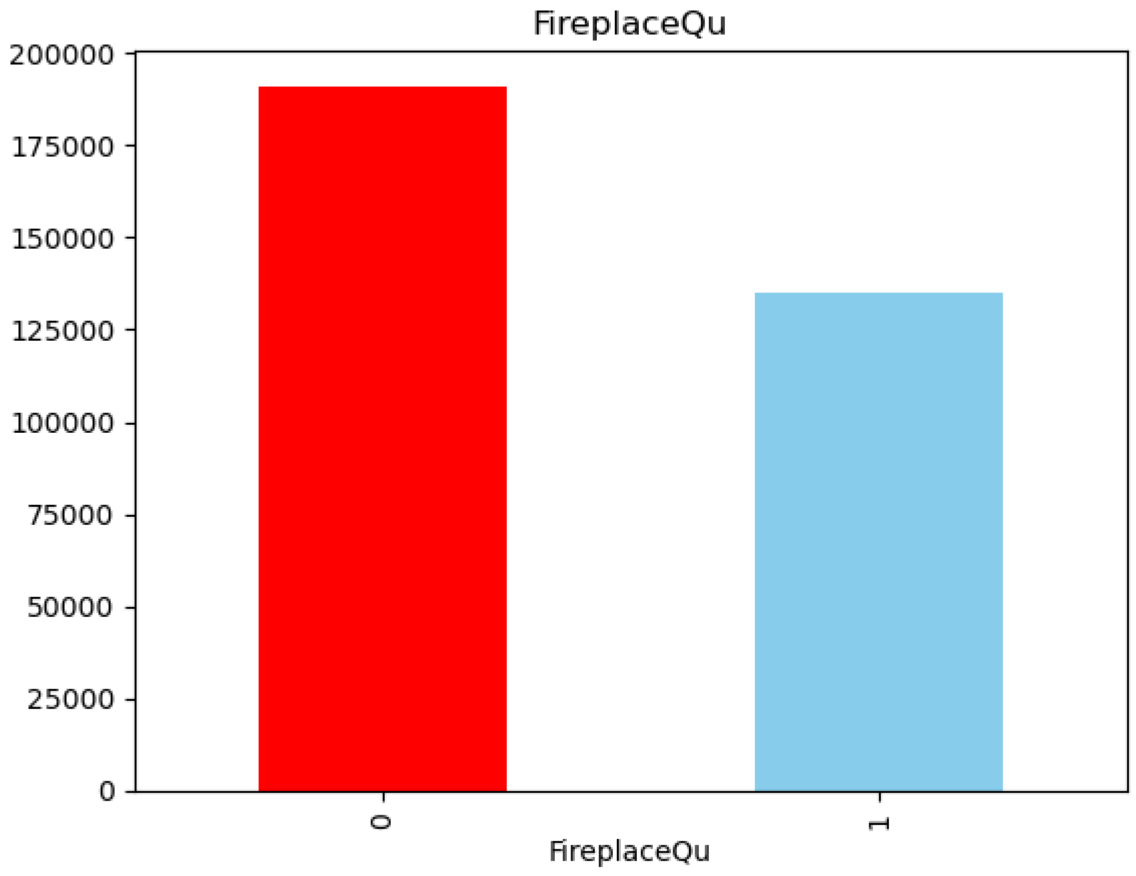

BsmtQual
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtQual

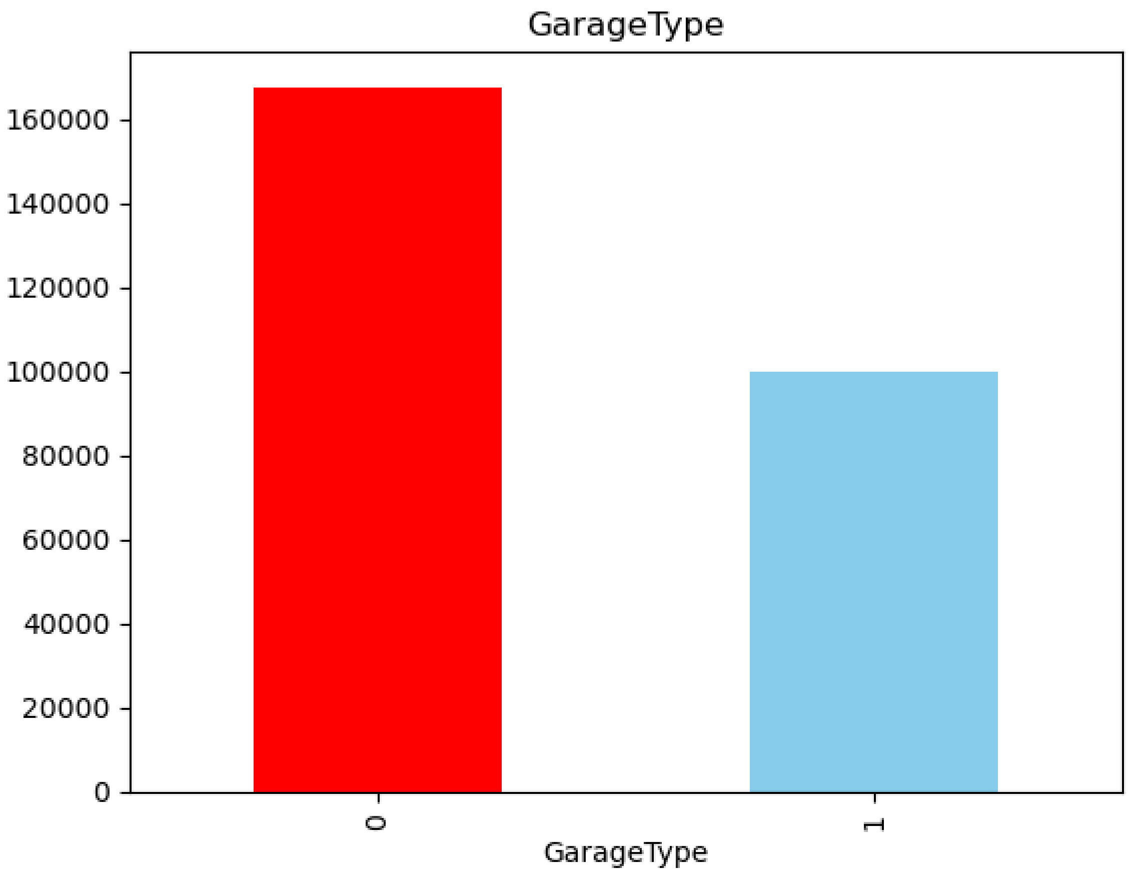

BsmtCond
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtCond

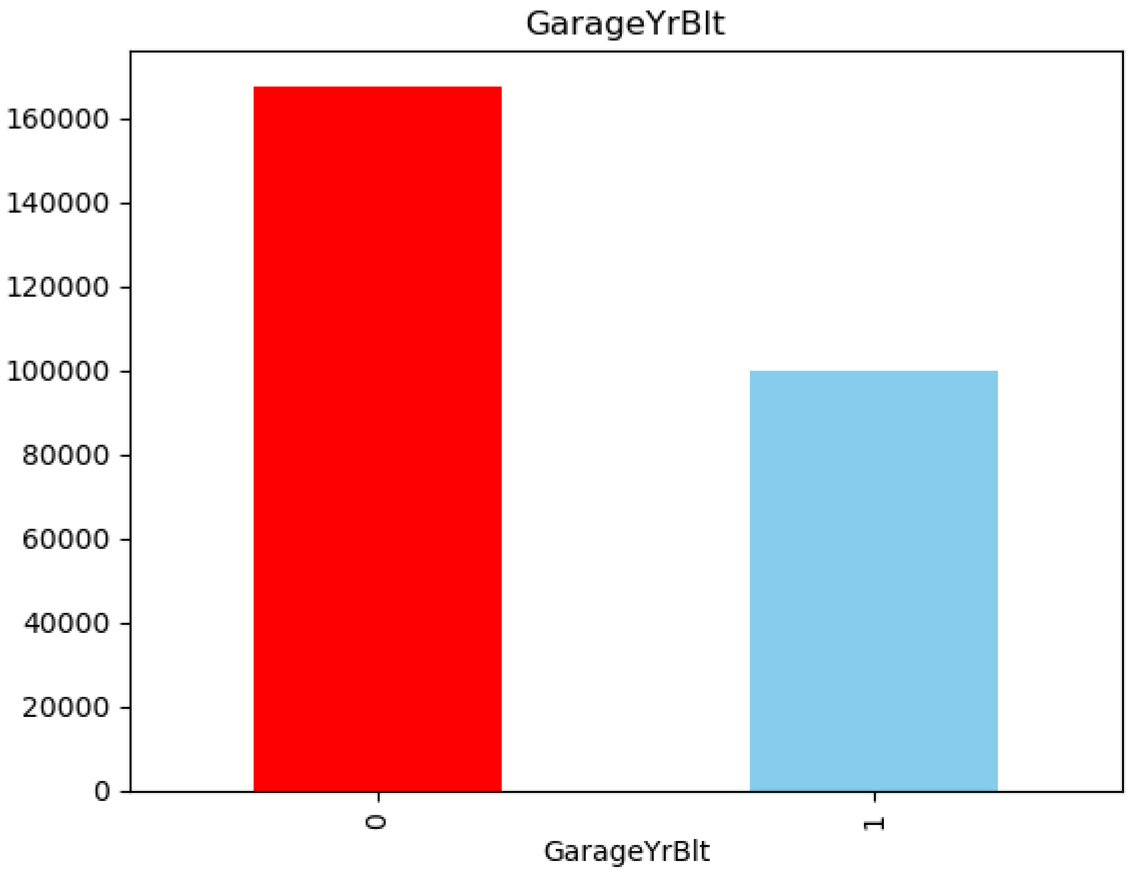

BsmtExposure
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtExposure

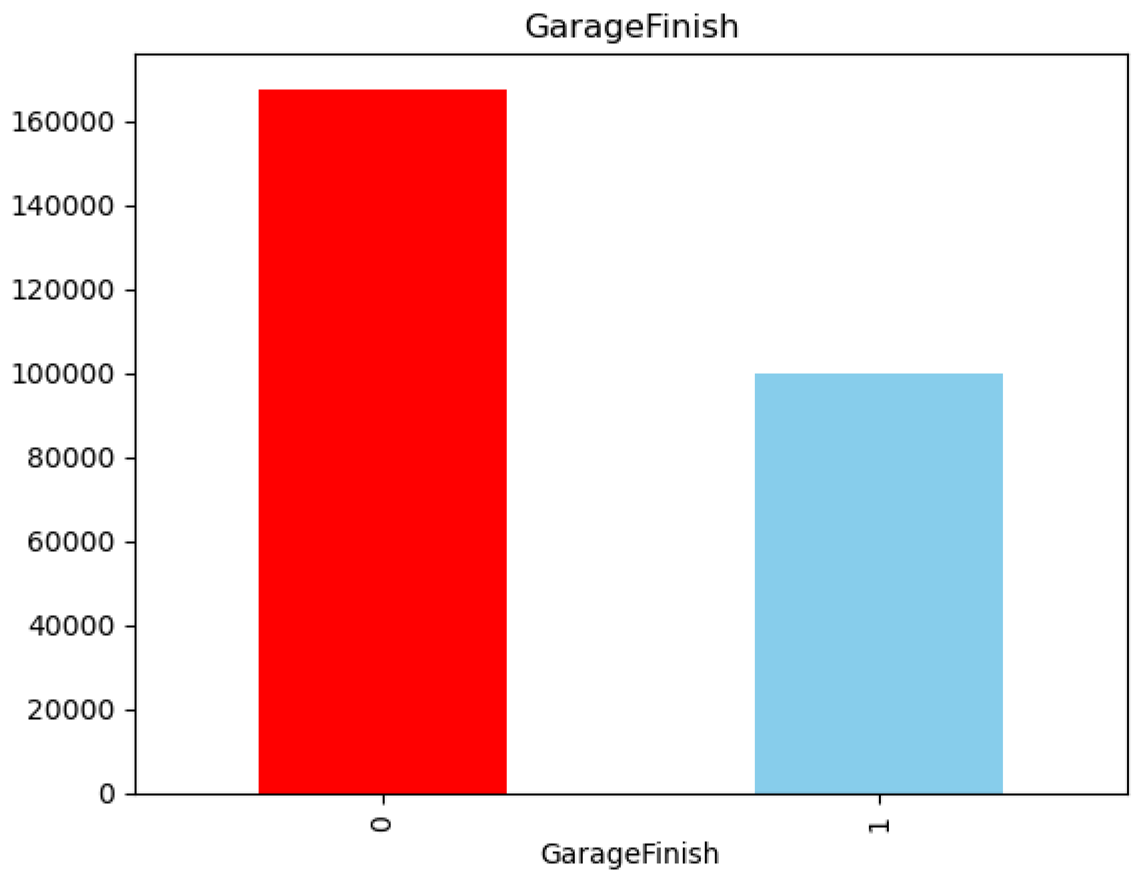

BsmtFinType1
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtFinType1

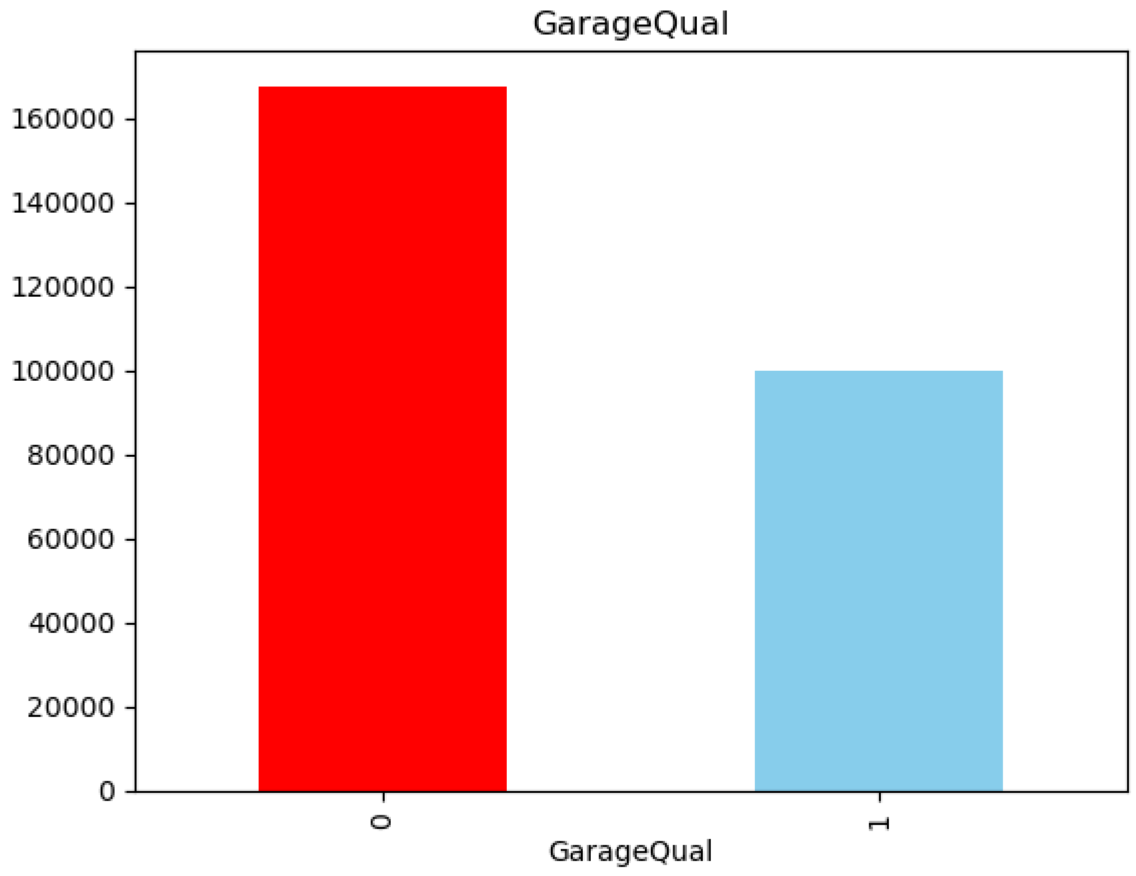

BsmtFinType2
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
BsmtFinType2

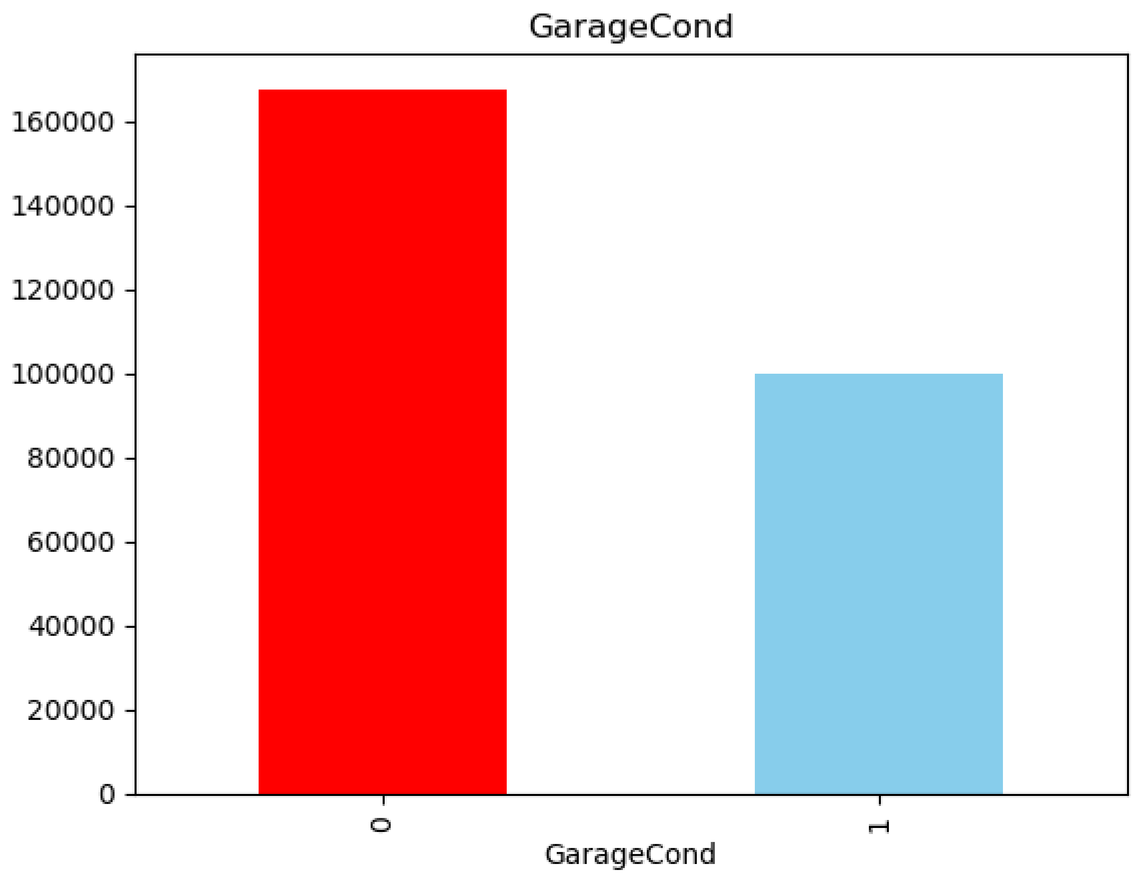

FireplaceQu
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
FireplaceQu

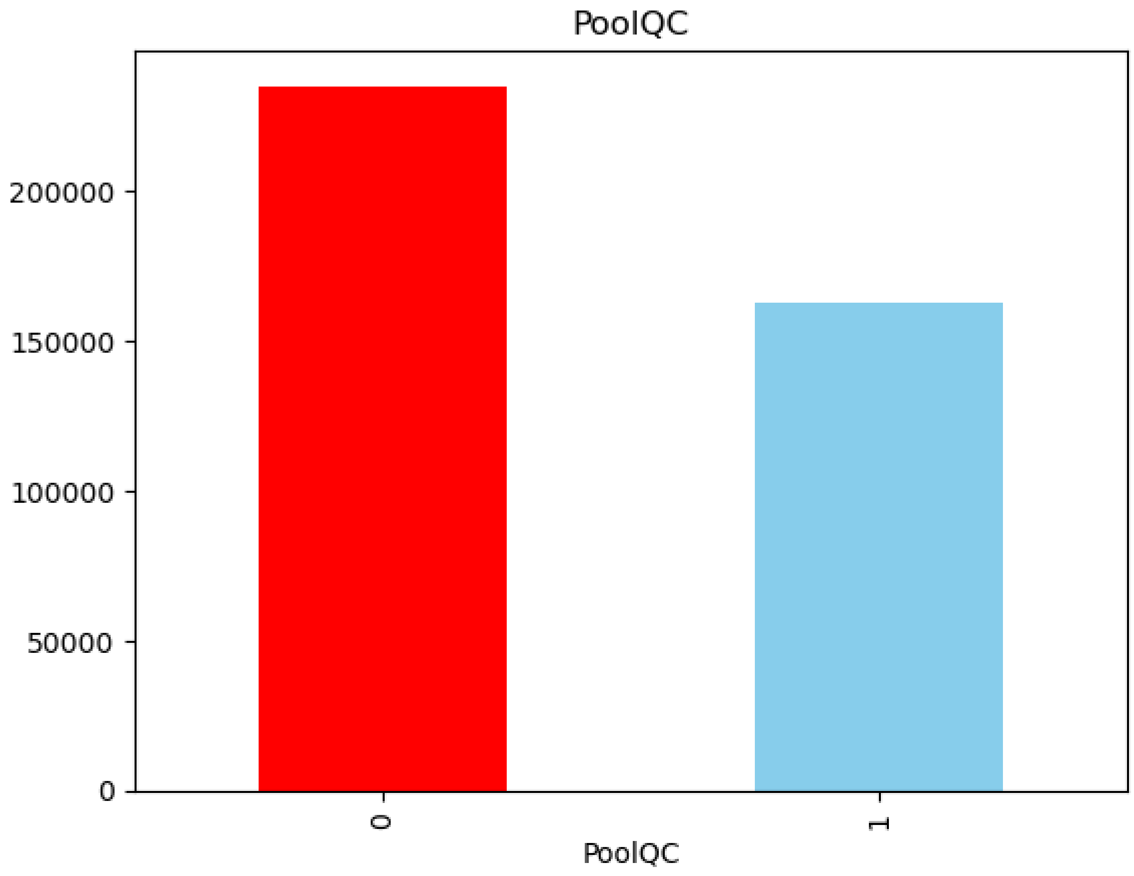

GarageType
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
GarageType

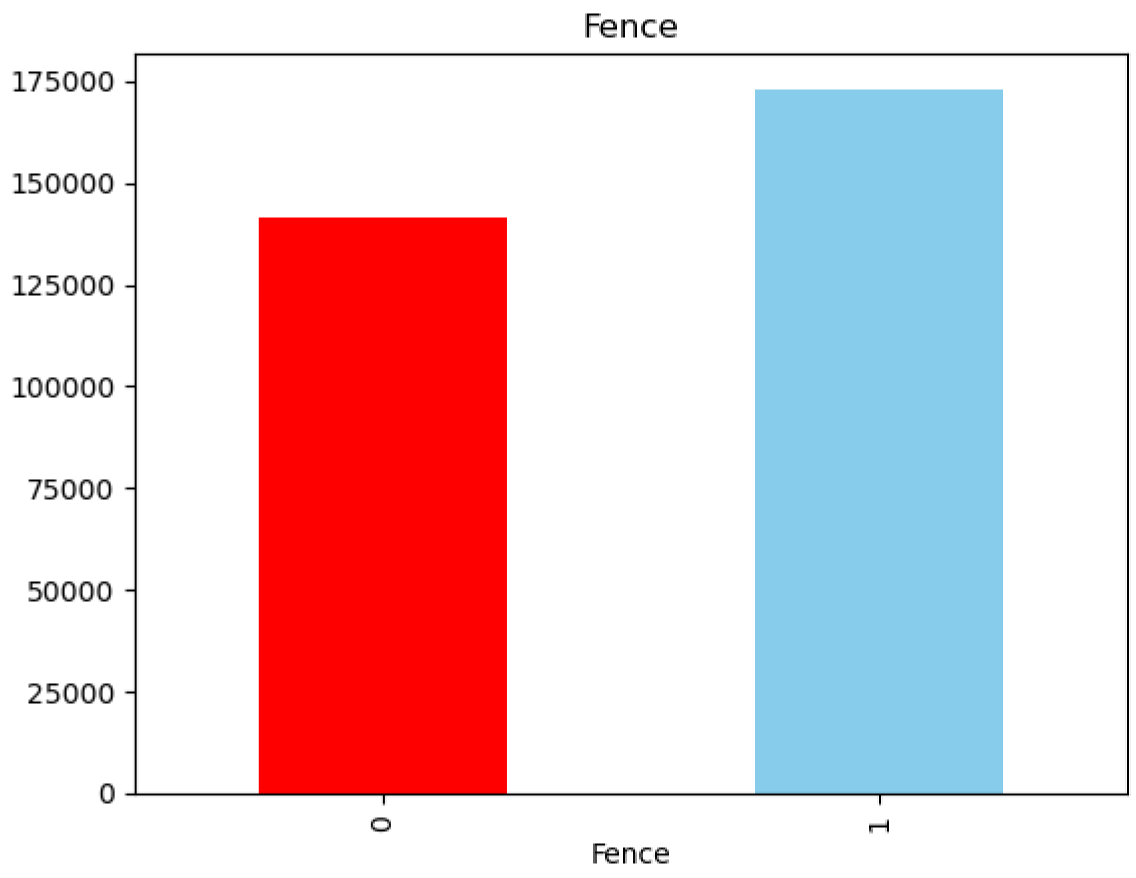

GarageYrBlt
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
GarageYrBlt

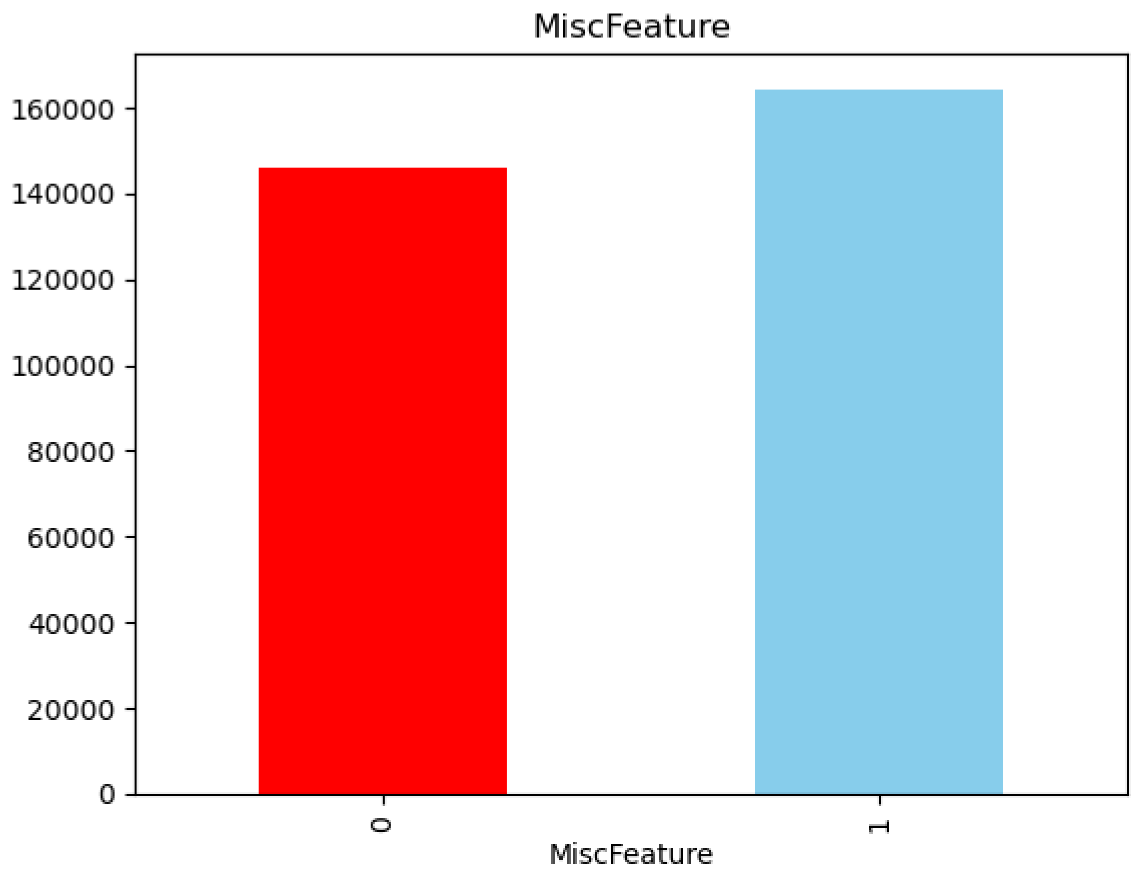

GarageFinish
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
GarageFinish


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


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


PoolQC
200000
150000
100000
50000
0
0
1
PoolQC





Here with relation between the missing values and the dependent variables is clearly visible.So we need to replace these nan values with something meaningfull which we will do in the Feature Engineering selection From the above dataset soome of the features like id i not required

In [8]:
```
print("Id of Houses {}".format(len(ds.Id)))
```

```
Id of Houses 1460
```

Numerical Variables

In [9]:
```
#list of numerical variables
```

In [10]:
```
numerical_features=[feature for feature in ds.columns if ds[feature].dtypes != '0']
print('Number of numerical variables: ', len(numerical_features))

#visualise the numerical variables
ds[numerical_features].head()
```

```
Number of numerical variables:  81
```

Out[10]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContou
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lv
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lv
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lv
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lv
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lv

Temporal variables (Eg:Datatime Variables)

From the Dataset we have 4 year variables.We have exact information from the datatime variables like no of yeard or no of dates.One example in this specific scenario can be difference in years between the year the house was built and the year house was sold. We will be performing this analysis in the Feature Engineering which is the next video

In [11]:
```
# List of variables that contain year information
year_feature=[feature for feature in numerical_features if 'Yr' in feature or 'Year
year_feature
```

Out[11]:
```
['YearBuilt', 'YearRemodAdd', 'GarageYrBlt', 'YrSold']
```

In [12]:
```
# let's explore the content of these year variables
for feature in year_feature:
    print(feature,ds[feature].unique())
```

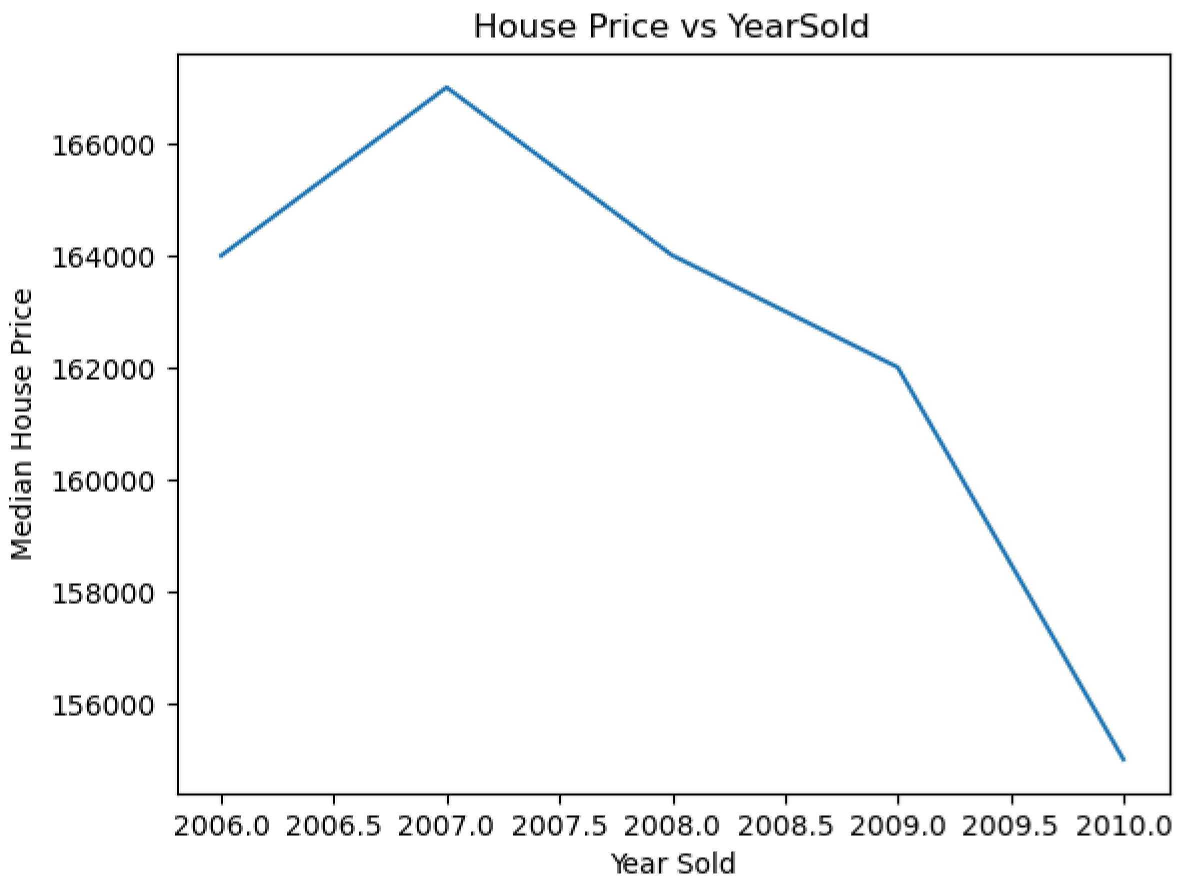
```
YearBuilt [2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 1965 2005 1962 2006
 1960 1929 1970 1967 1958 1930 2002 1968 2007 1951 1957 1927 1920 1966
 1959 1994 1954 1953 1955 1983 1975 1997 1934 1963 1981 1964 1999 1972
 1921 1945 1982 1998 1956 1948 1910 1995 1991 2009 1950 1961 1977 1985
 1979 1885 1919 1990 1969 1935 1988 1971 1952 1936 1923 1924 1984 1926
 1940 1941 1987 1986 2008 1908 1892 1916 1932 1918 1912 1947 1925 1900
 1980 1989 1992 1949 1880 1928 1978 1922 1996 2010 1946 1913 1937 1942
 1938 1974 1893 1914 1906 1890 1898 1904 1882 1875 1911 1917 1872 1905]
YearRemodAdd [2003 1976 2002 1970 2000 1995 2005 1973 1950 1965 2006 1962 2007 1960
 2001 1967 2004 2008 1997 1959 1990 1955 1983 1980 1966 1963 1987 1964
 1972 1996 1998 1989 1953 1956 1968 1981 1992 2009 1982 1961 1993 1999
 1985 1979 1977 1969 1958 1991 1971 1952 1975 2010 1984 1986 1994 1988
 1954 1957 1951 1978 1974]
GarageYrBlt [2003. 1976. 2001. 1998. 2000. 1993. 2004. 1973. 1931. 1939. 1965. 2005.
 1962. 2006. 1960. 1991. 1970. 1967. 1958. 1930. 2002. 1968. 2007. 2008.
 1957. 1920. 1966. 1959. 1995. 1954. 1953.   nan 1983. 1977. 1997. 1985.
 1963. 1981. 1964. 1999. 1935. 1990. 1945. 1987. 1989. 1915. 1956. 1948.
 1974. 2009. 1950. 1961. 1921. 1900. 1979. 1951. 1969. 1936. 1975. 1971.
 1923. 1984. 1926. 1955. 1986. 1988. 1916. 1932. 1972. 1918. 1980. 1924.
 1996. 1940. 1949. 1994. 1910. 1978. 1982. 1992. 1925. 1941. 2010. 1927.
 1947. 1937. 1942. 1938. 1952. 1928. 1922. 1934. 1906. 1914. 1946. 1908.
 1929. 1933.]
YrSold [2008 2007 2006 2009 2010]
```

In [13]:

```python
 ## lets analyze the temporal datetime variables
## we will check whether there is a relation between year the house is sold and

ds.groupby('YrSold')['SalePrice'].median().plot()
plt.xlabel('Year Sold')
plt.ylabel('Median House Price')
plt.title('House Price vs YearSold')
```

Out[13]:

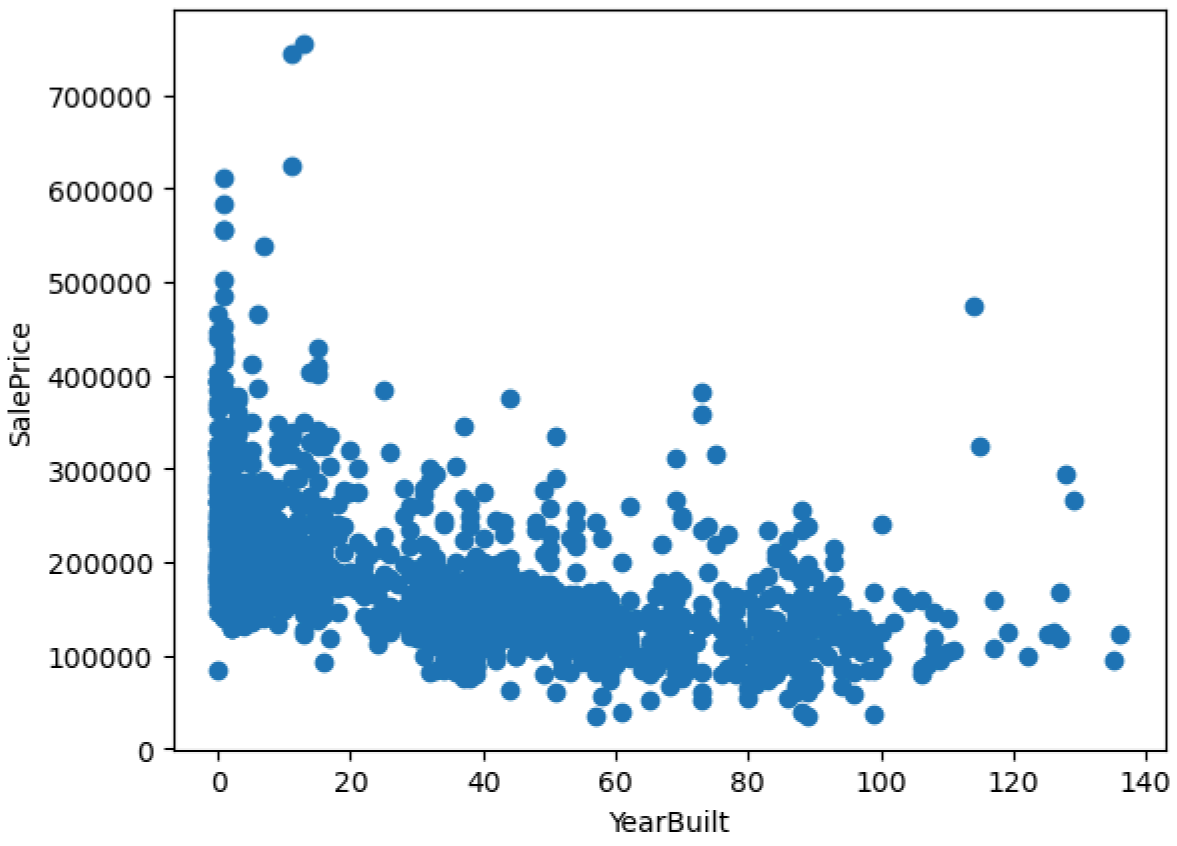
```
Text(0.5, 1.0, 'House Price vs YearSold')
```

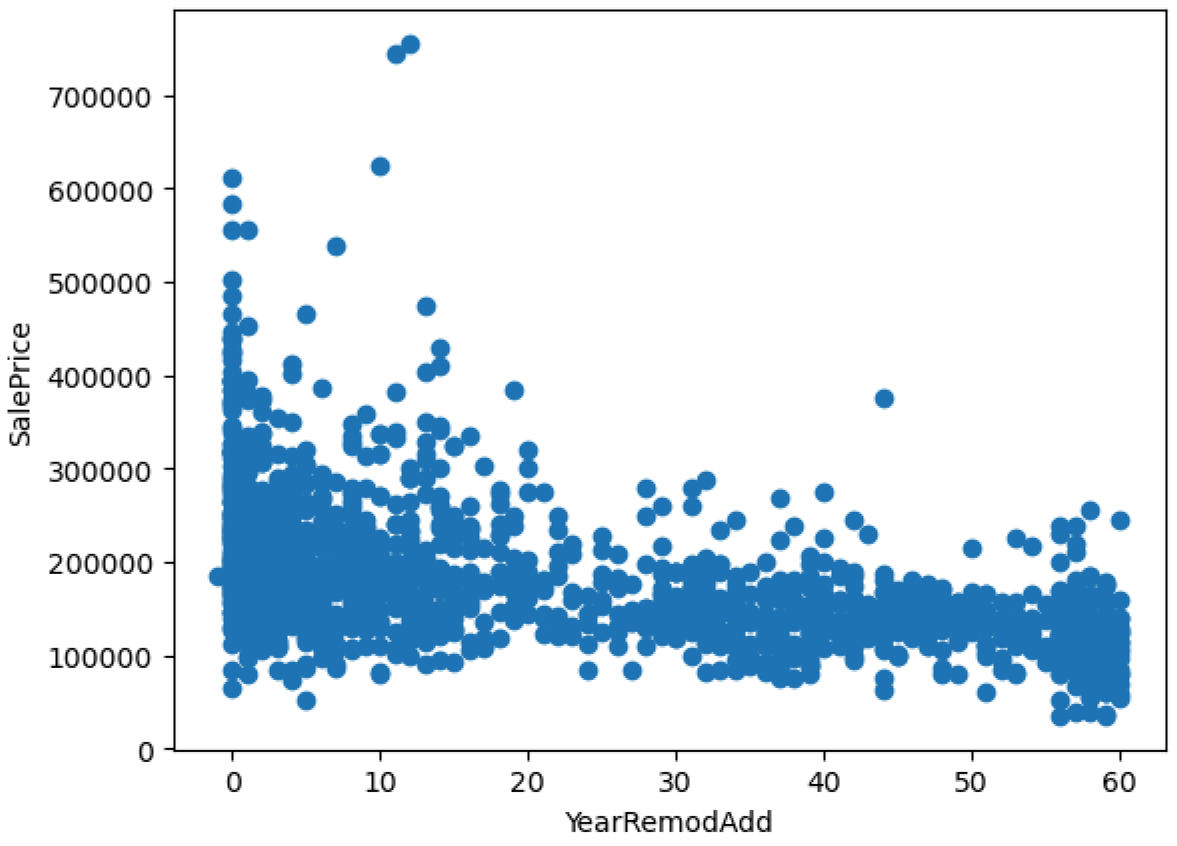


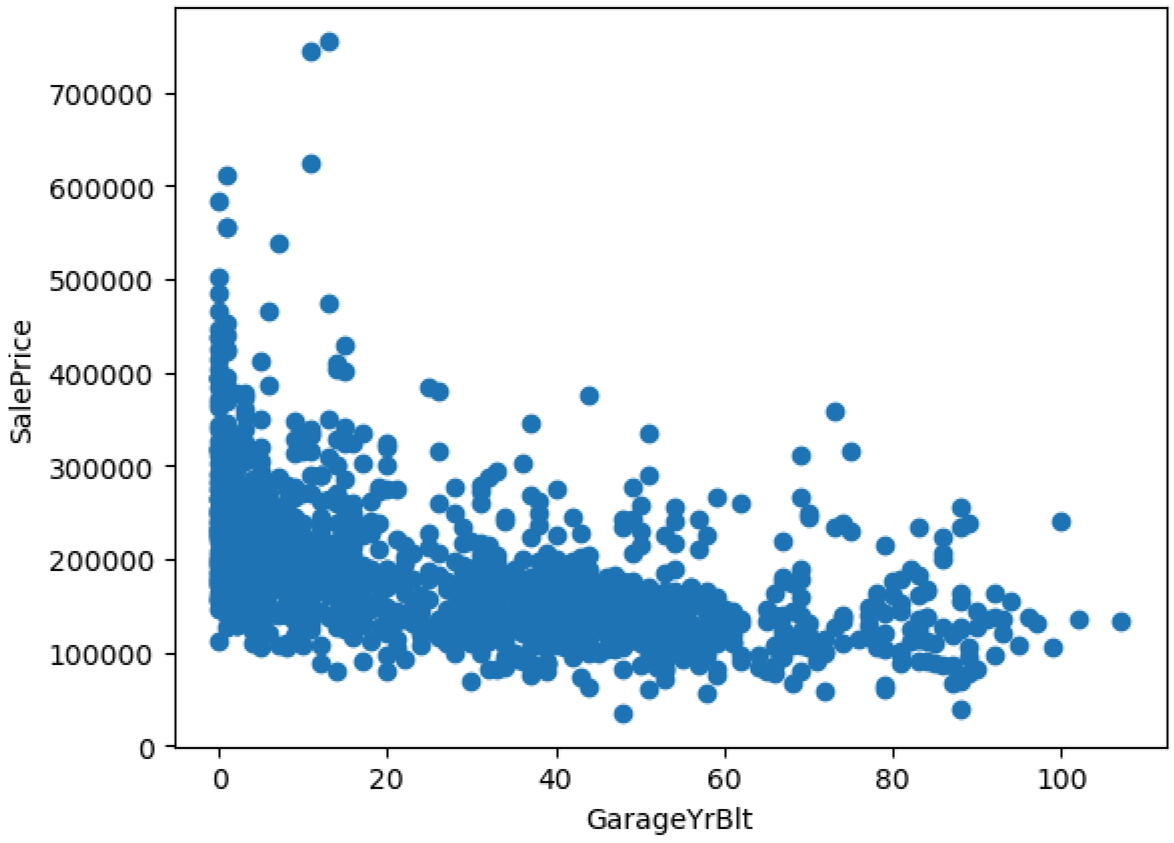
```
In [14]: year_feature
```

```
Out[14]: ['YearBuilt', 'YearRemodAdd', 'GarageYrBlt', 'YrSold']
```

```
In [15]: ## Here we will compare the differnece between all years feature with SalePrice

         for feature in year_feature:
             if feature!='YrSold':
                 data=ds.copy()
                 ##we will capture the difference between your variables and year of the hou
                 data[feature]=data['YrSold']-data[feature]

                 plt.scatter(data[feature],data['SalePrice'])
                 plt.xlabel(feature)
                 plt.ylabel('SalePrice')
                 plt.show()
```


SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
20
40
60
80
100
120
140
YearBuilt


SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
10
20
30
40
50
60
YearRemodAdd



```
In [16]: ## Numerical Variables are usually of 2 types
         ## 1.Continous variables and Discrete variables

         discrete_feature=[feature for feature in numerical_features if len(ds[feature].uniq
         print("Discrete Variables Count: {}".format(len(discrete_feature)))
```

```
Discrete Variables Count: 59
```

```
In [17]: discrete_feature
```

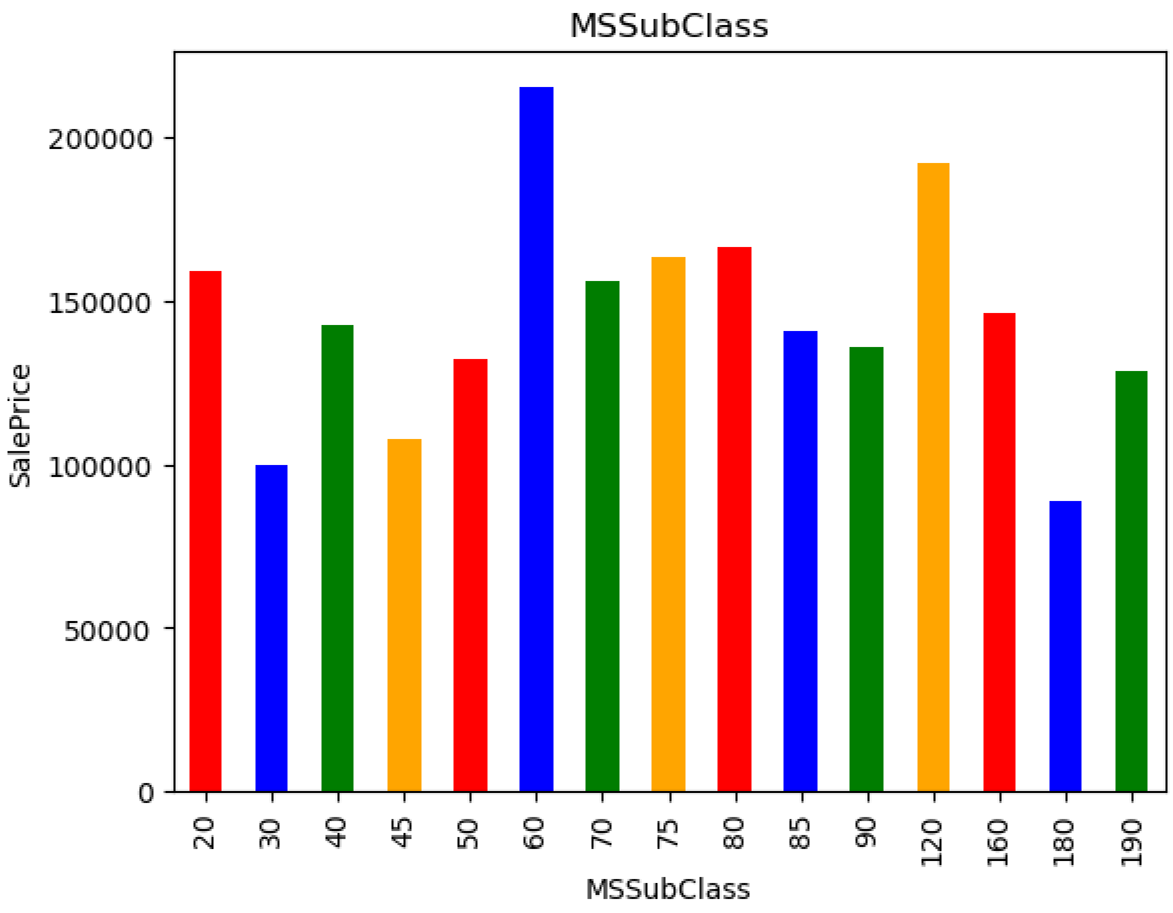
```
Out[17]: ['MSSubClass',
          'MSZoning',
          'Street',
          'Alley',
          'LotShape',
          'LandContour',
          'Utilities',
          'LotConfig',
          'LandSlope',
          'Condition1',
          'Condition2',
          'BldgType',
          'HouseStyle',
          'OverallQual',
          'OverallCond',
          'RoofStyle',
          'RoofMatl',
          'Exterior1st',
          'Exterior2nd',
          'MasVnrType',
          'ExterQual',
          'ExterCond',
          'Foundation',
          'BsmtQual',
          'BsmtCond',
          'BsmtExposure',
          'BsmtFinType1',
          'BsmtFinType2',
          'Heating',
          'HeatingQC',
          'CentralAir',
          'Electrical',
          'LowQualFinSF',
          'BsmtFullBath',
          'BsmtHalfBath',
          'FullBath',
          'HalfBath',
          'BedroomAbvGr',
          'KitchenAbvGr',
          'KitchenQual',
          'TotRmsAbvGrd',
          'Functional',
          'Fireplaces',
          'FireplaceQu',
          'GarageType',
          'GarageFinish',
          'GarageCars',
          'GarageQual',
          'GarageCond',
          'PavedDrive',
          '3SsnPorch',
          'PoolArea',
          'PoolQC',
          'Fence',
          'MiscFeature',
          'MiscVal',
 'MoSold',
 'SaleType',
 'SaleCondition']
```

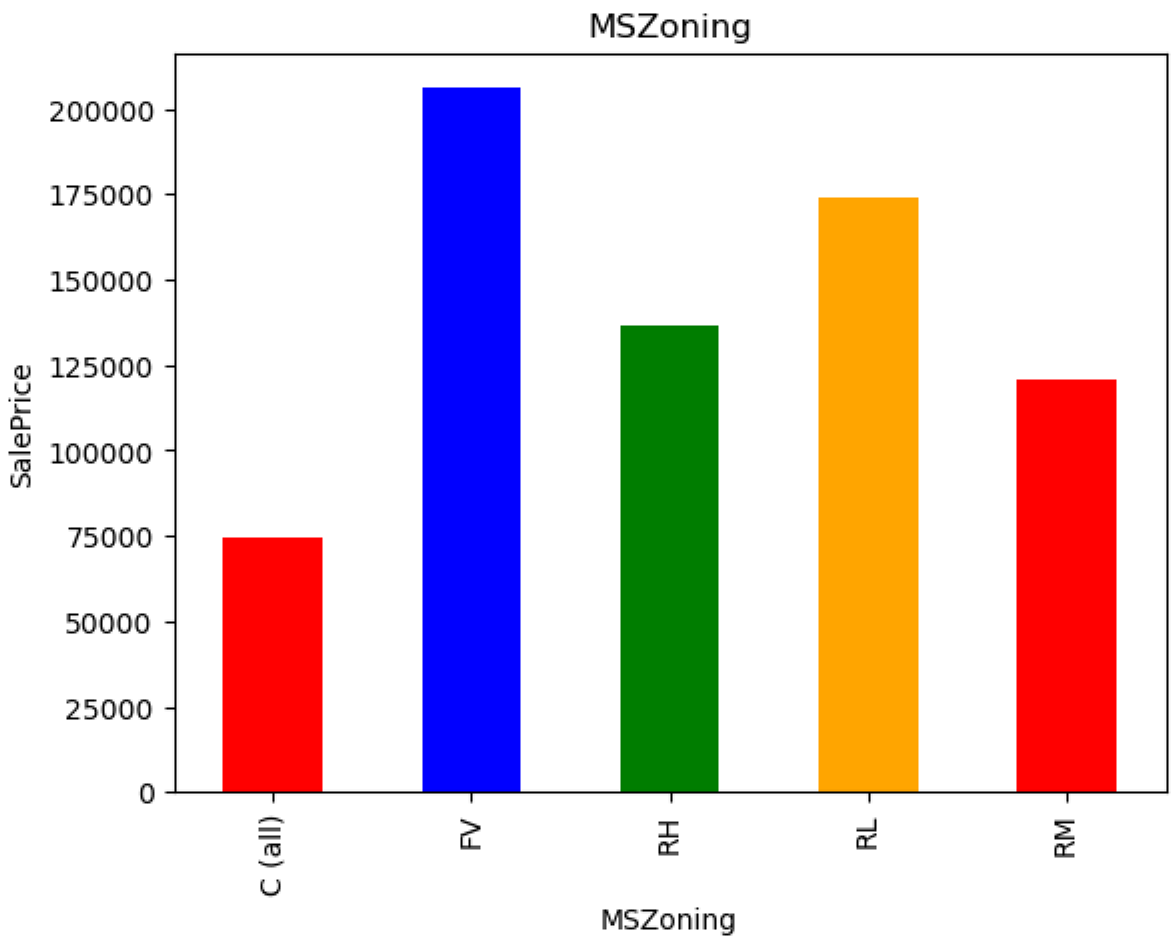
In [18]:
```
ds[discrete_feature].head()
```

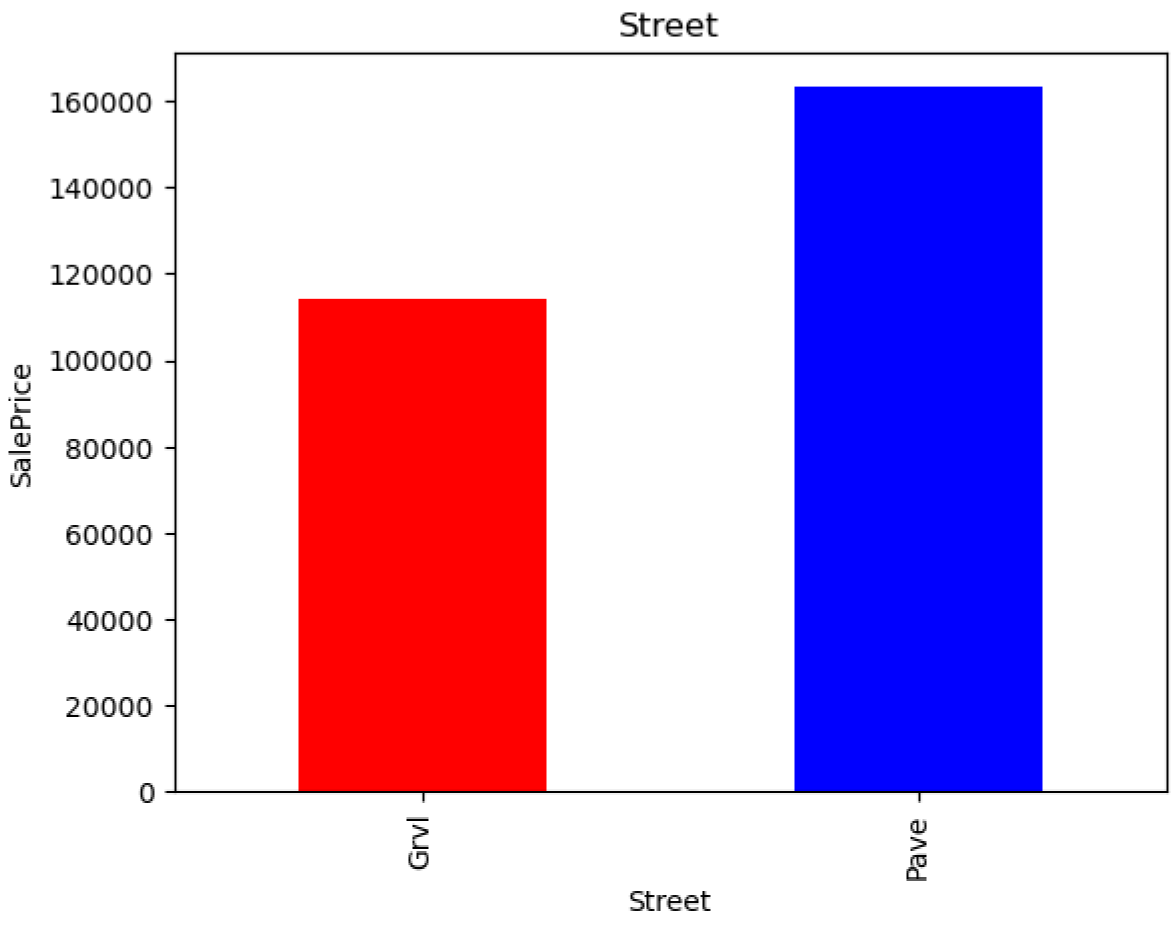
Out[18]:

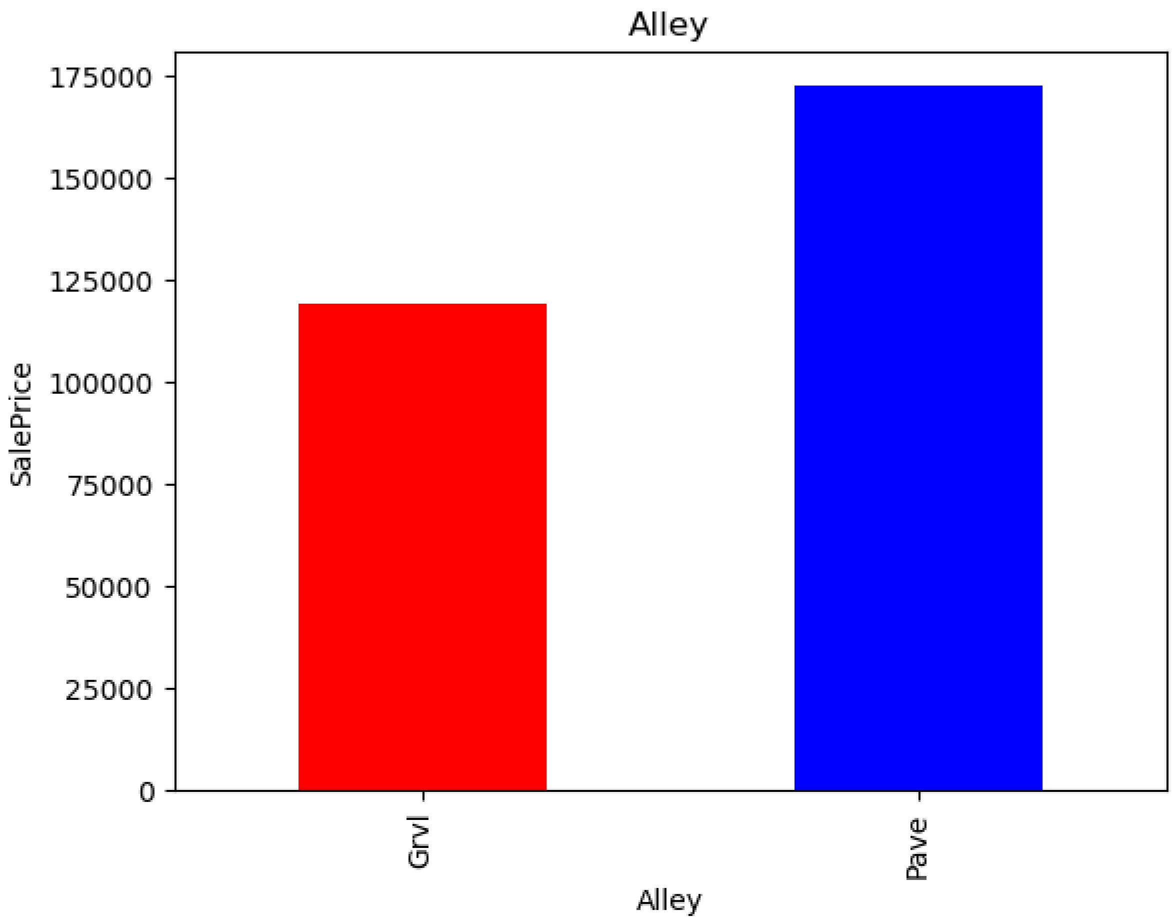
	MSSubClass	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	Lan
0	60	RL	Pave	NaN	Reg	Lvl	AllPub	Inside	
1	20	RL	Pave	NaN	Reg	Lvl	AllPub	FR2	
2	60	RL	Pave	NaN	IR1	Lvl	AllPub	Inside	
3	70	RL	Pave	NaN	IR1	Lvl	AllPub	Corner	
4	60	RL	Pave	NaN	IR1	Lvl	AllPub	FR2	

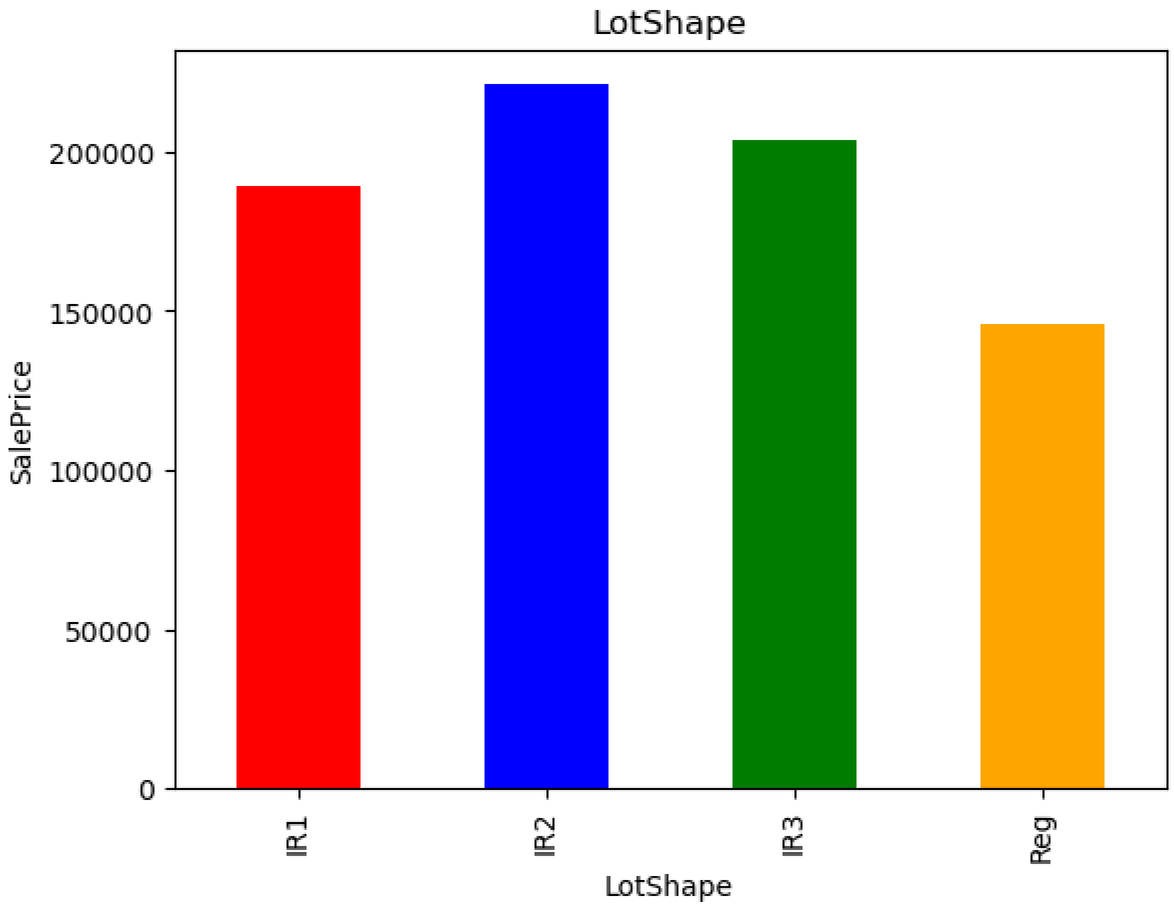
In [19]:
```
#Let's find the relationship between them and Sale Price

for feature in discrete_feature:
    data=ds.copy()
    data.groupby(feature)['SalePrice'].median().plot.bar(color=['red','blue','green
    plt.xlabel(feature)
    plt.ylabel('SalePrice')
    plt.title(feature)
    plt.show()
```

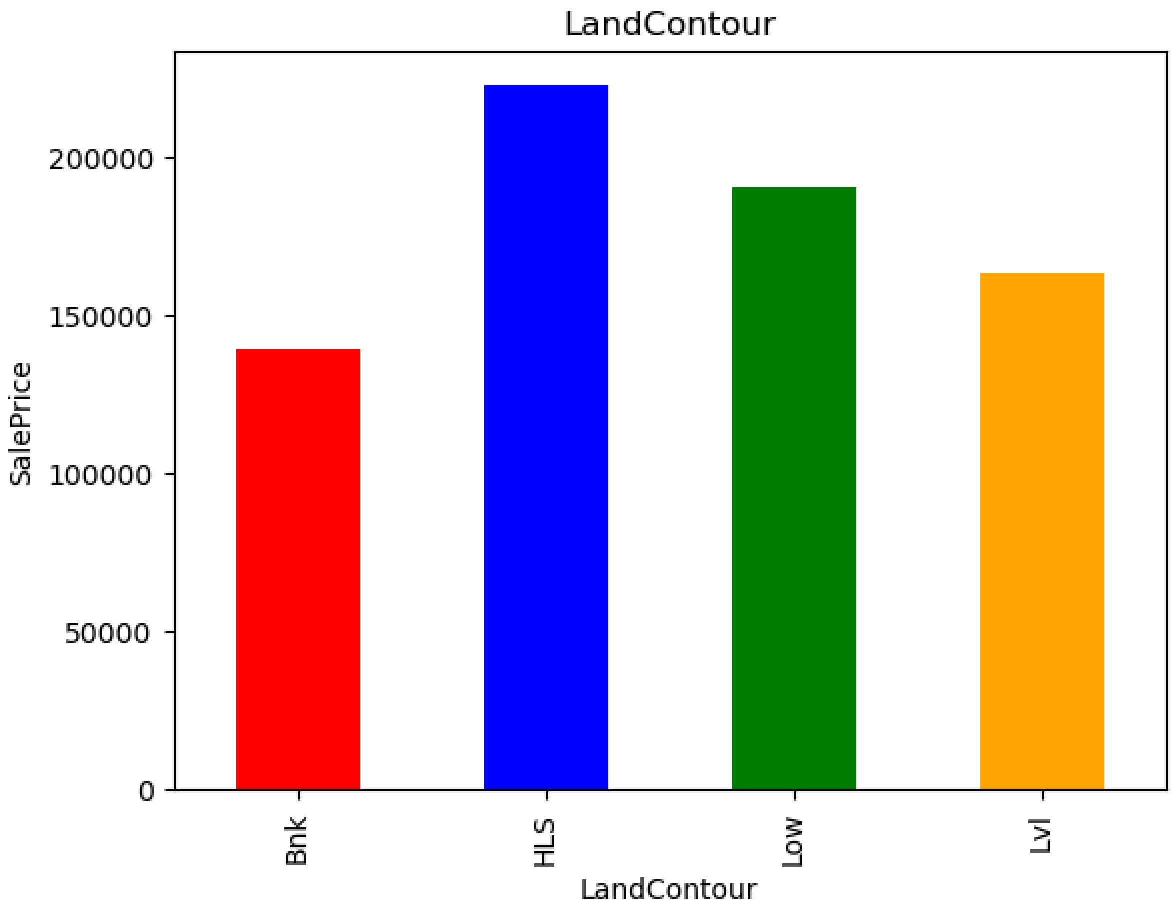


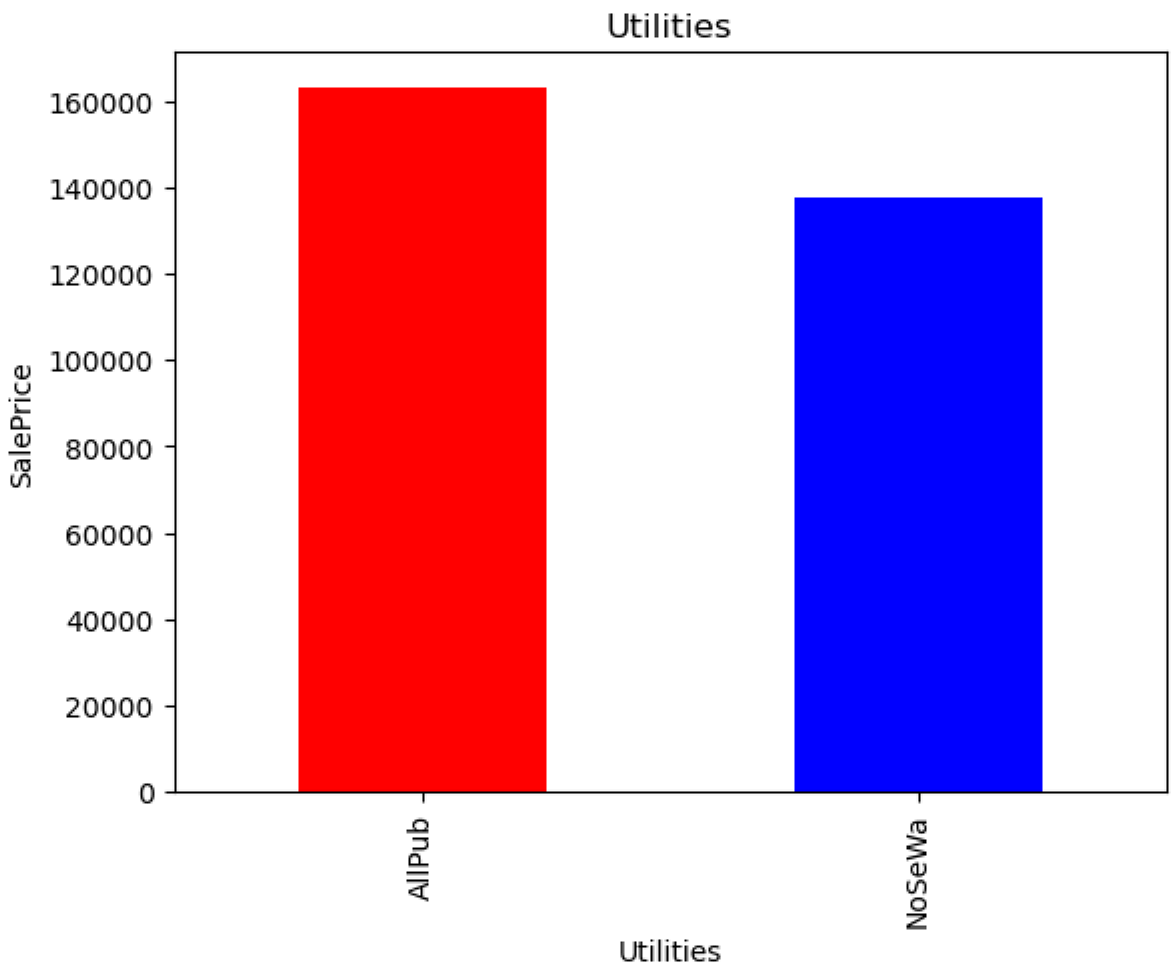

MSZoning
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
C (all)
FV
RH
RL
RM
MSZoning

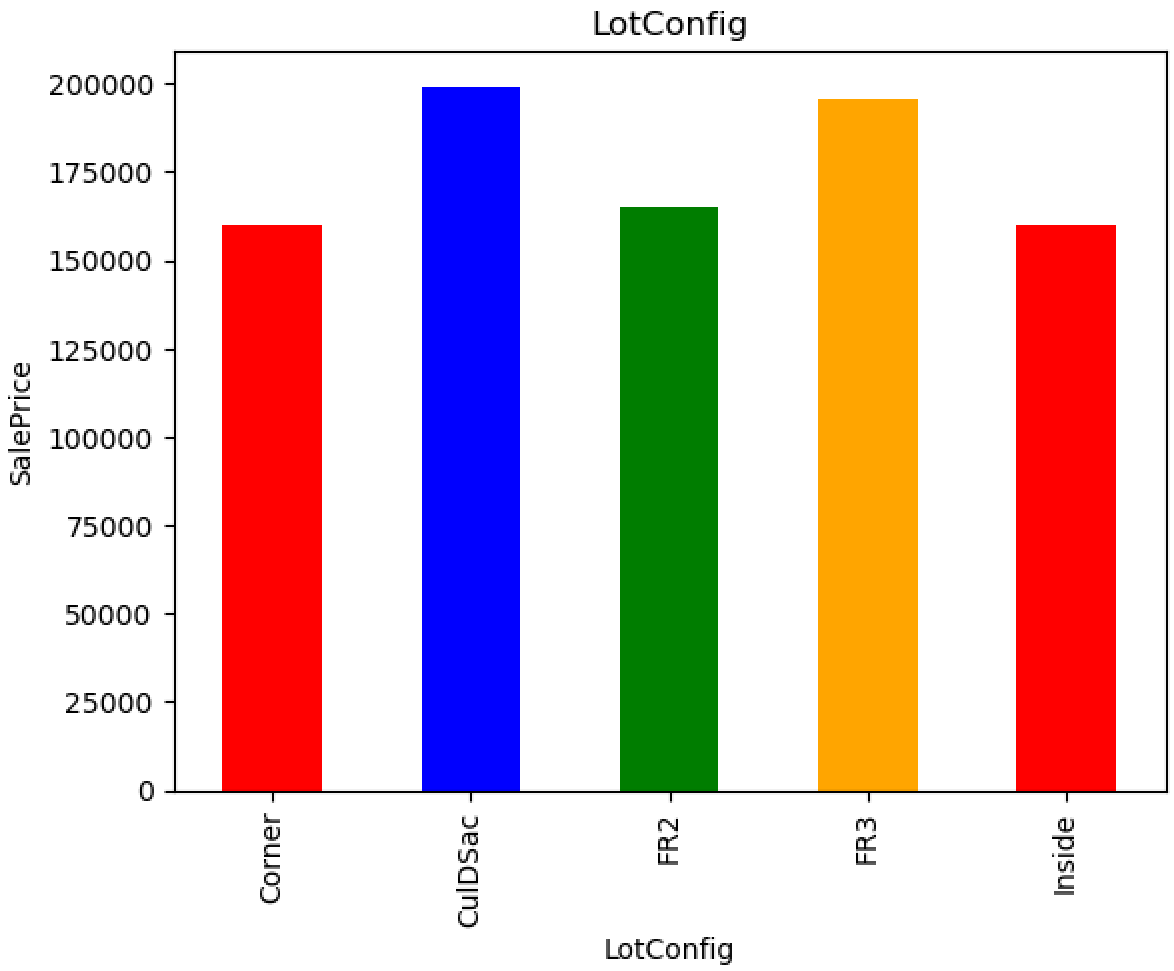

Street
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Grvl
Pave
Street

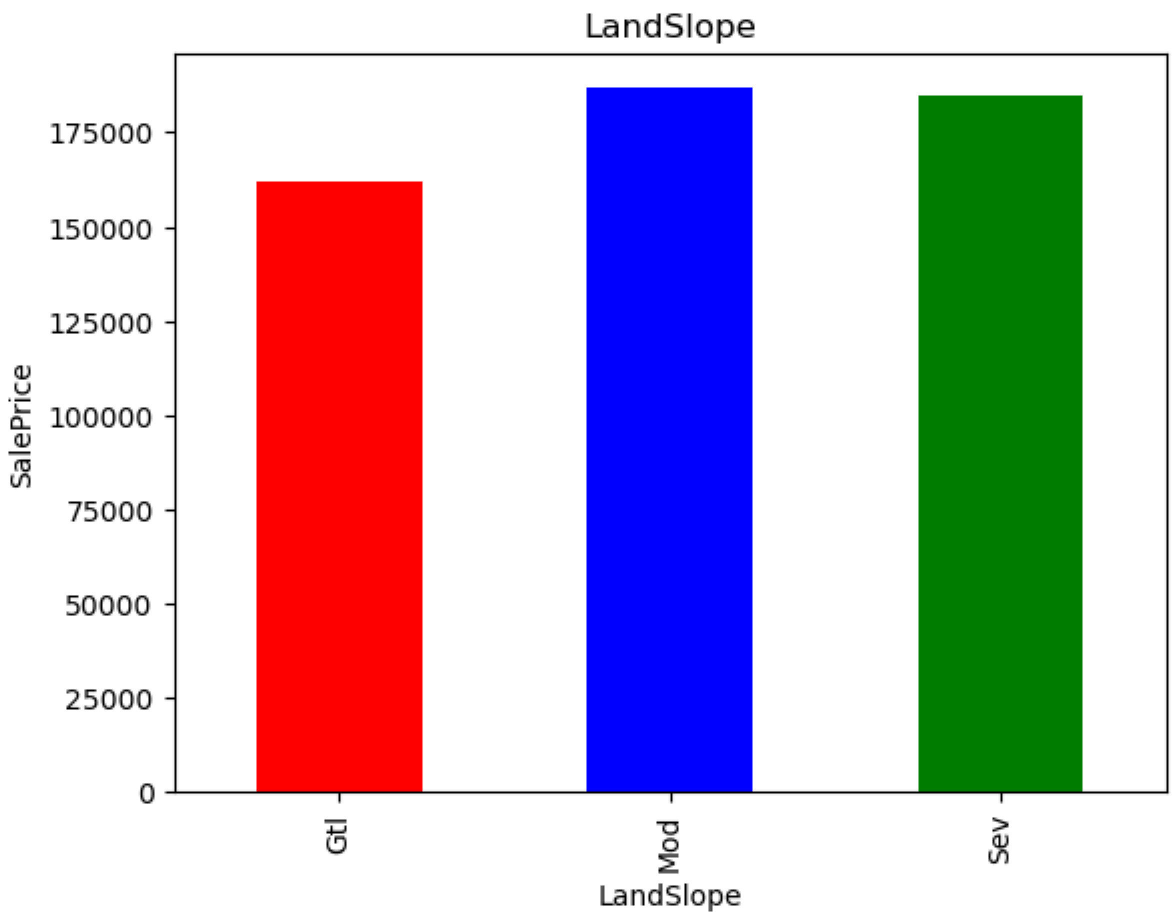

Alley
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
Grvl
Pave
Alley

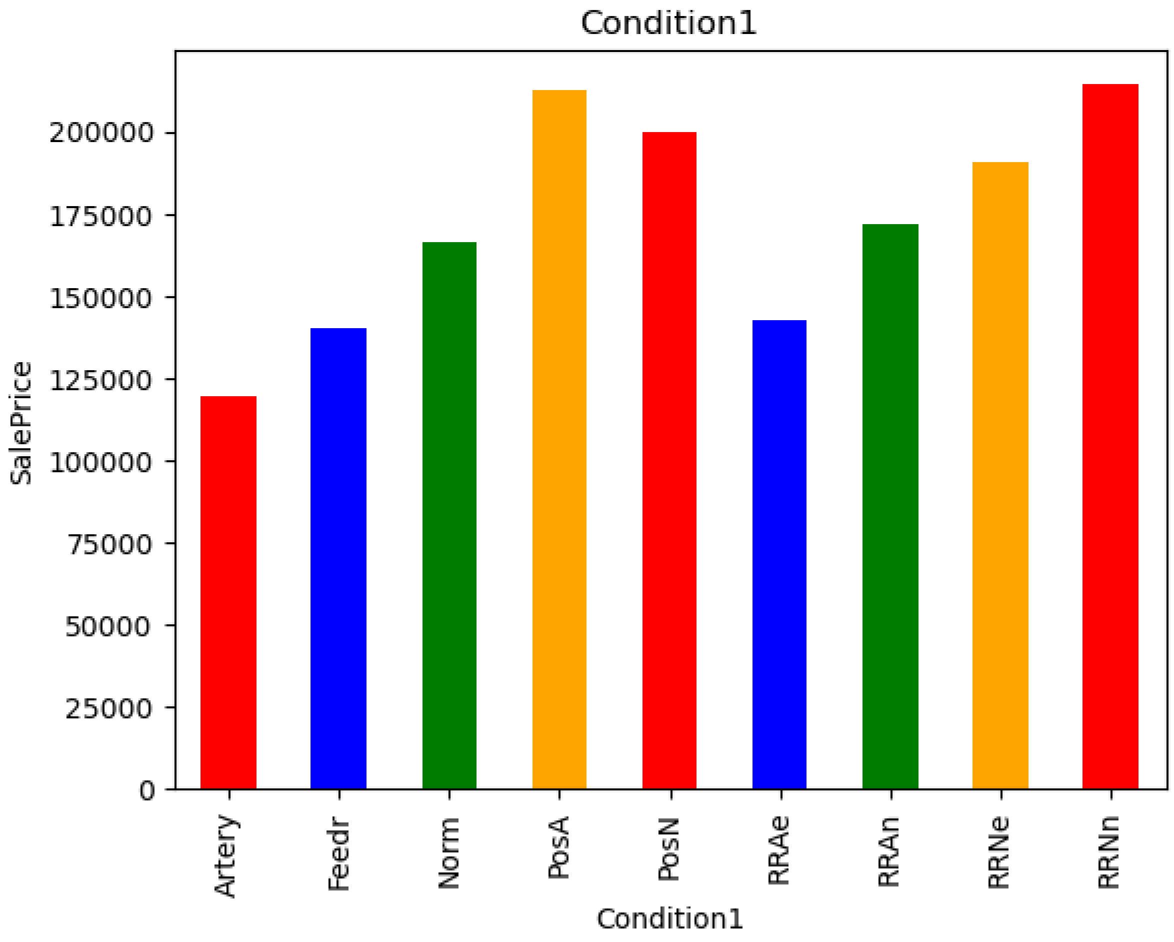

LotShape
SalePrice
200000
150000
100000
50000
0
IR1
IR2
IR3
Reg
LotShape

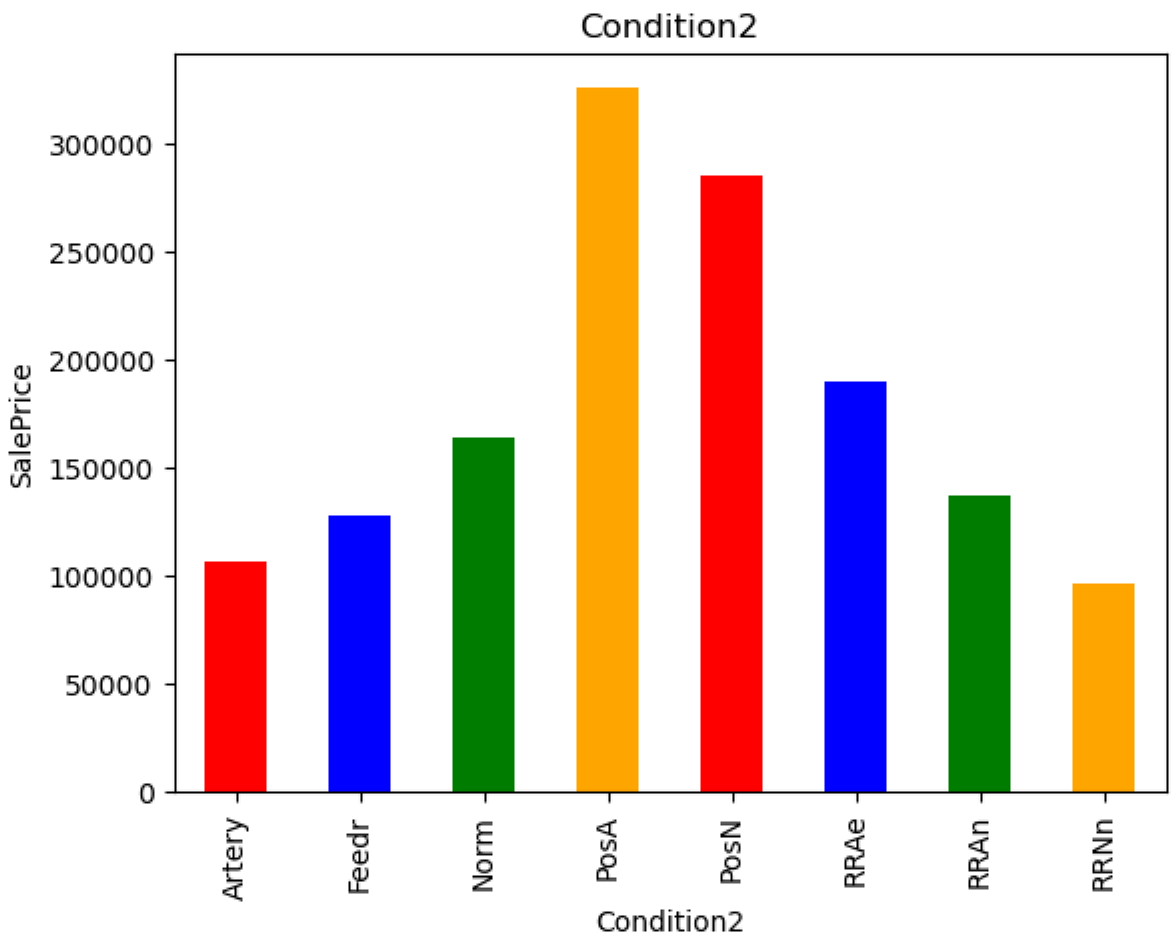

LandContour
SalePrice
200000
150000
100000
50000
0
Bnk
HLS
Low
Lvl
LandContour

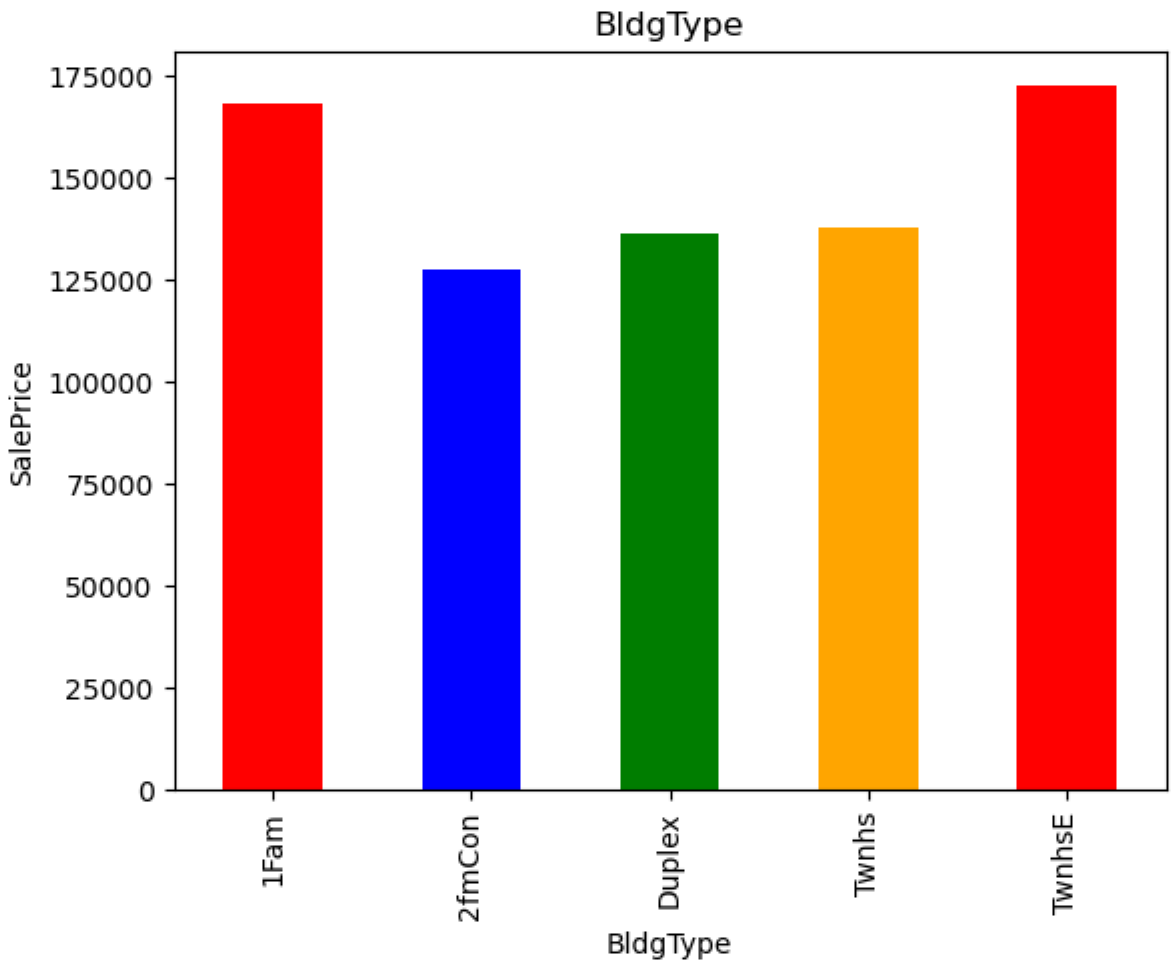

Utilities
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
AllPub
NoSeWa
Utilities

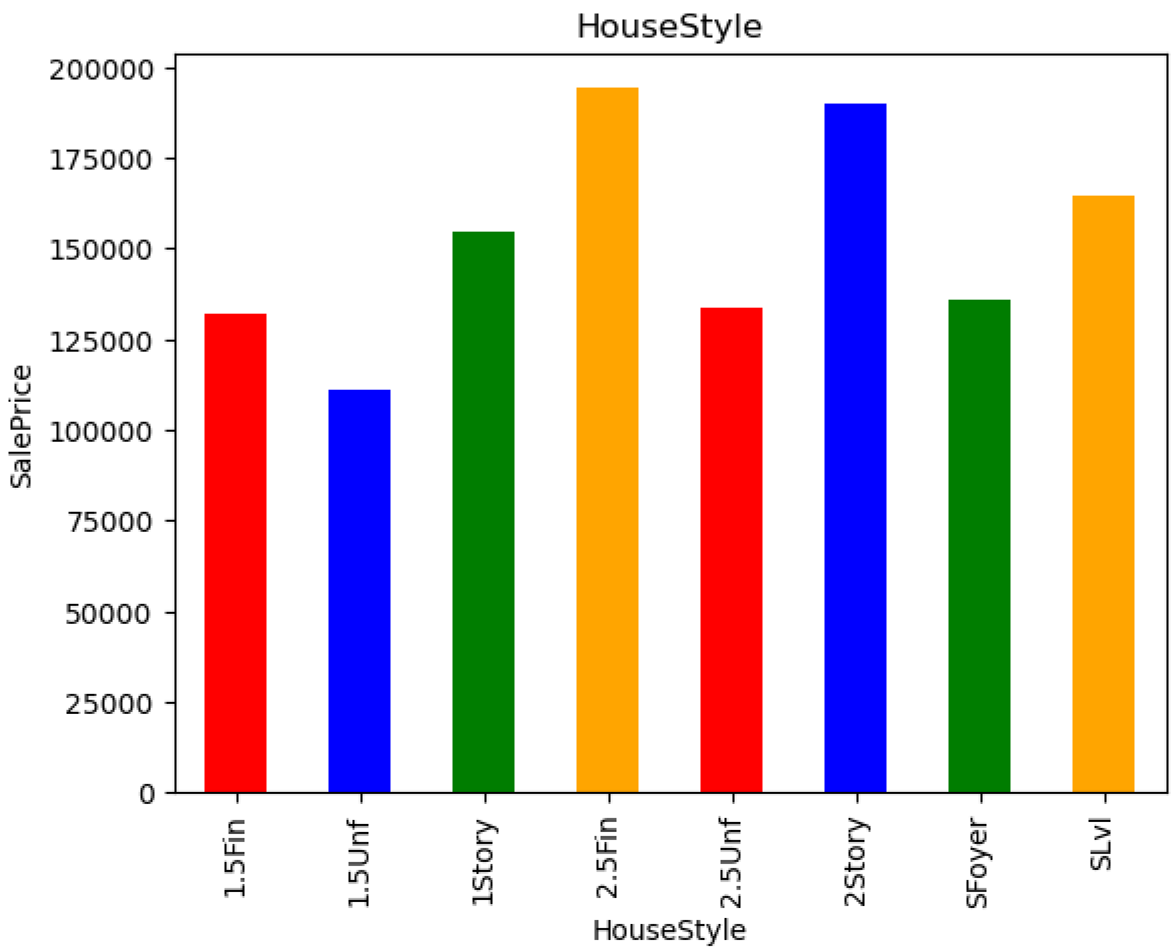

LotConfig
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Corner
CulDSac
FR2
FR3
Inside
LotConfig

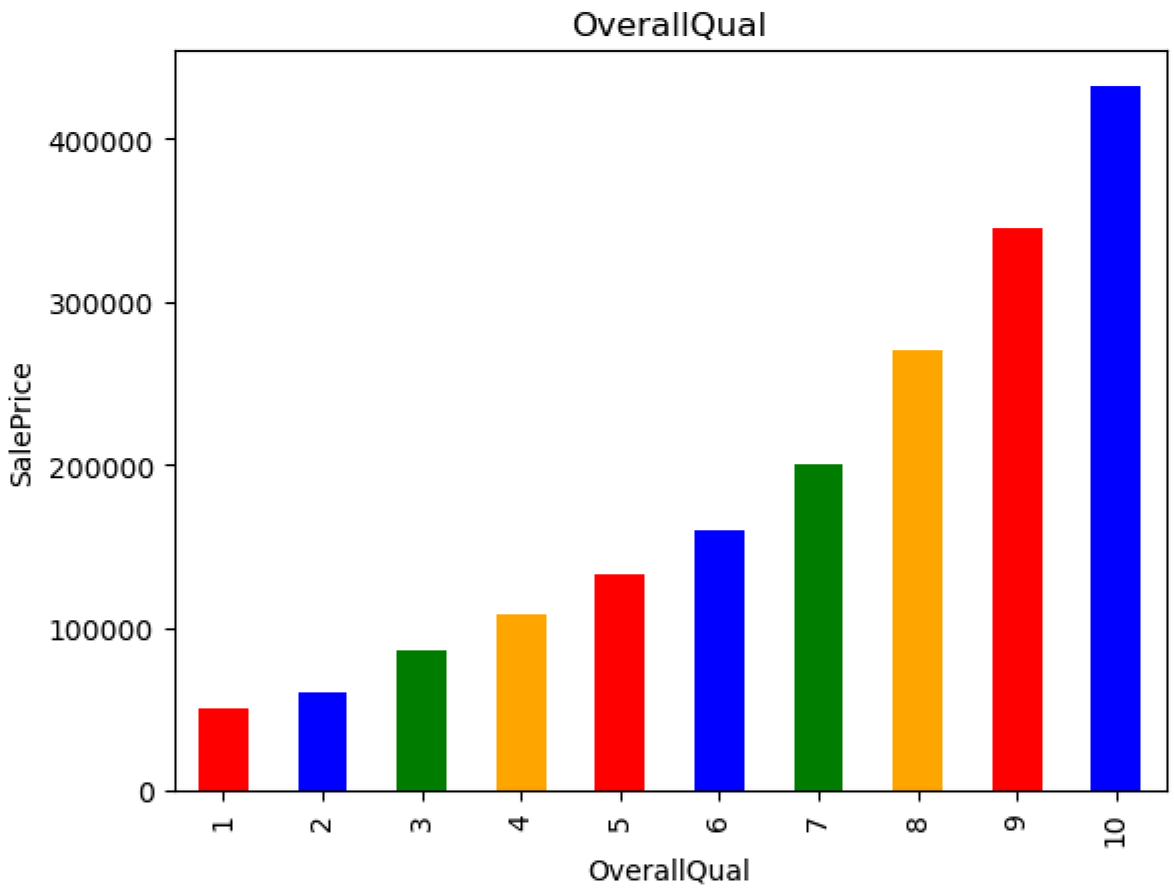

LandSlope
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
Gtl
Mod
Sev
LandSlope

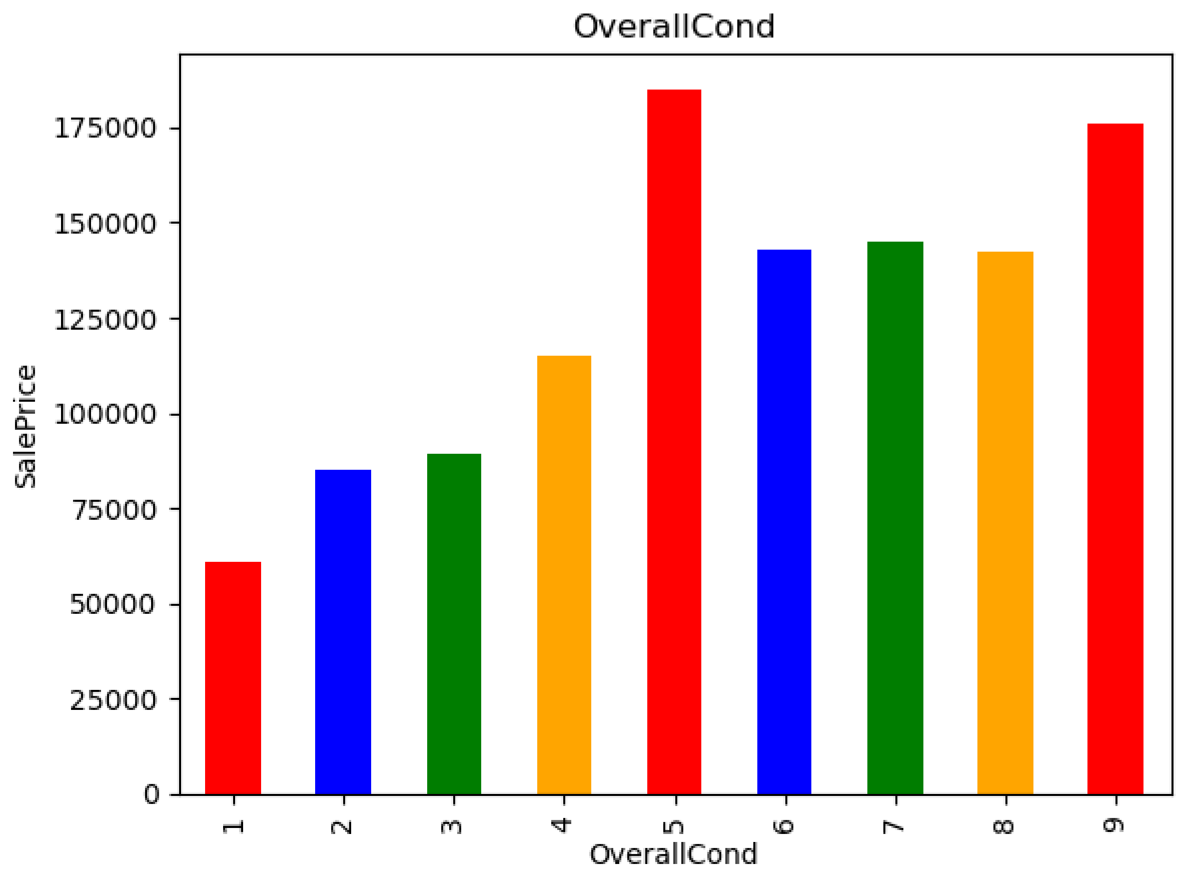

Condition1
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn
Condition1

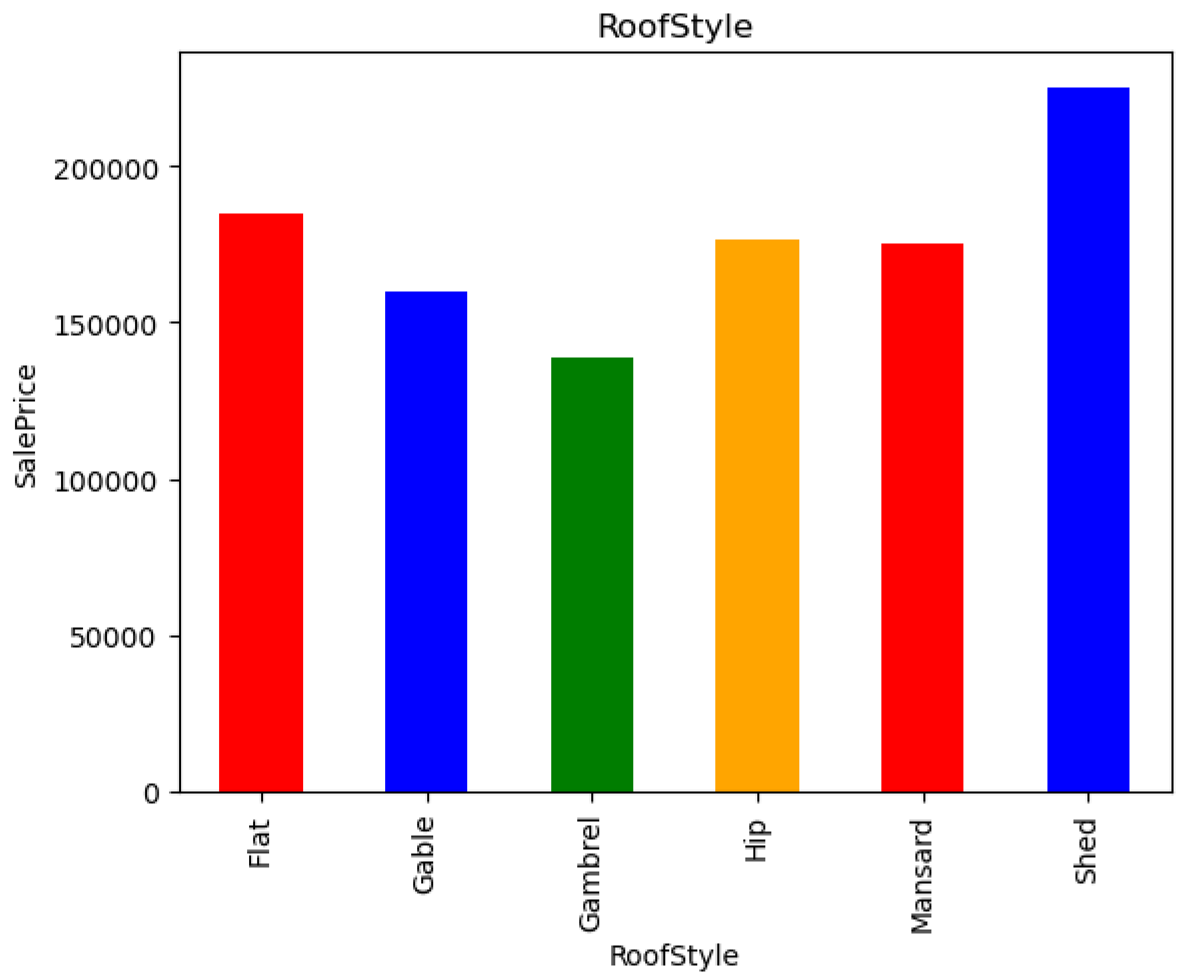

Condition2
SalePrice
300000
250000
200000
150000
100000
50000
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNn
Condition2

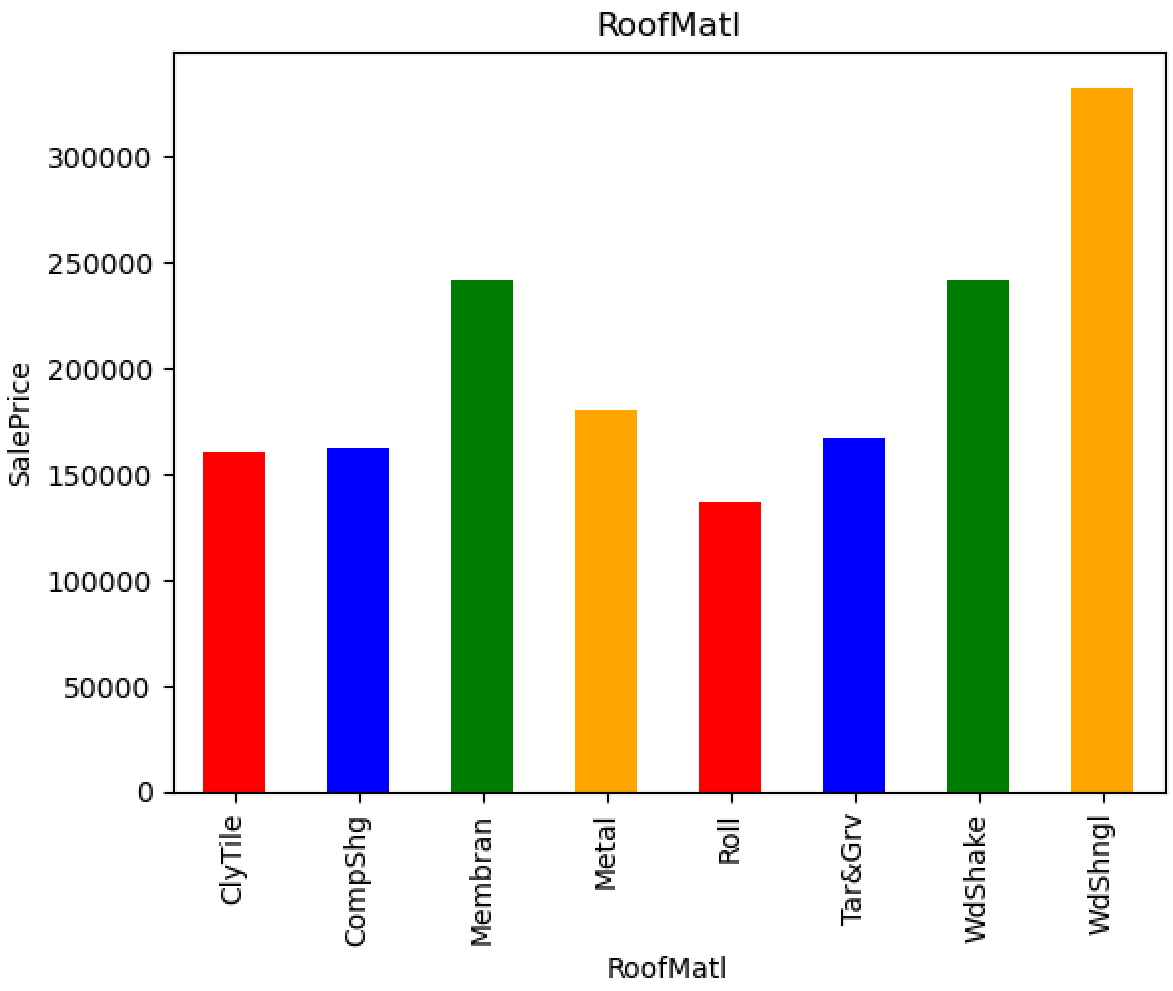

BldgType
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
BldgType

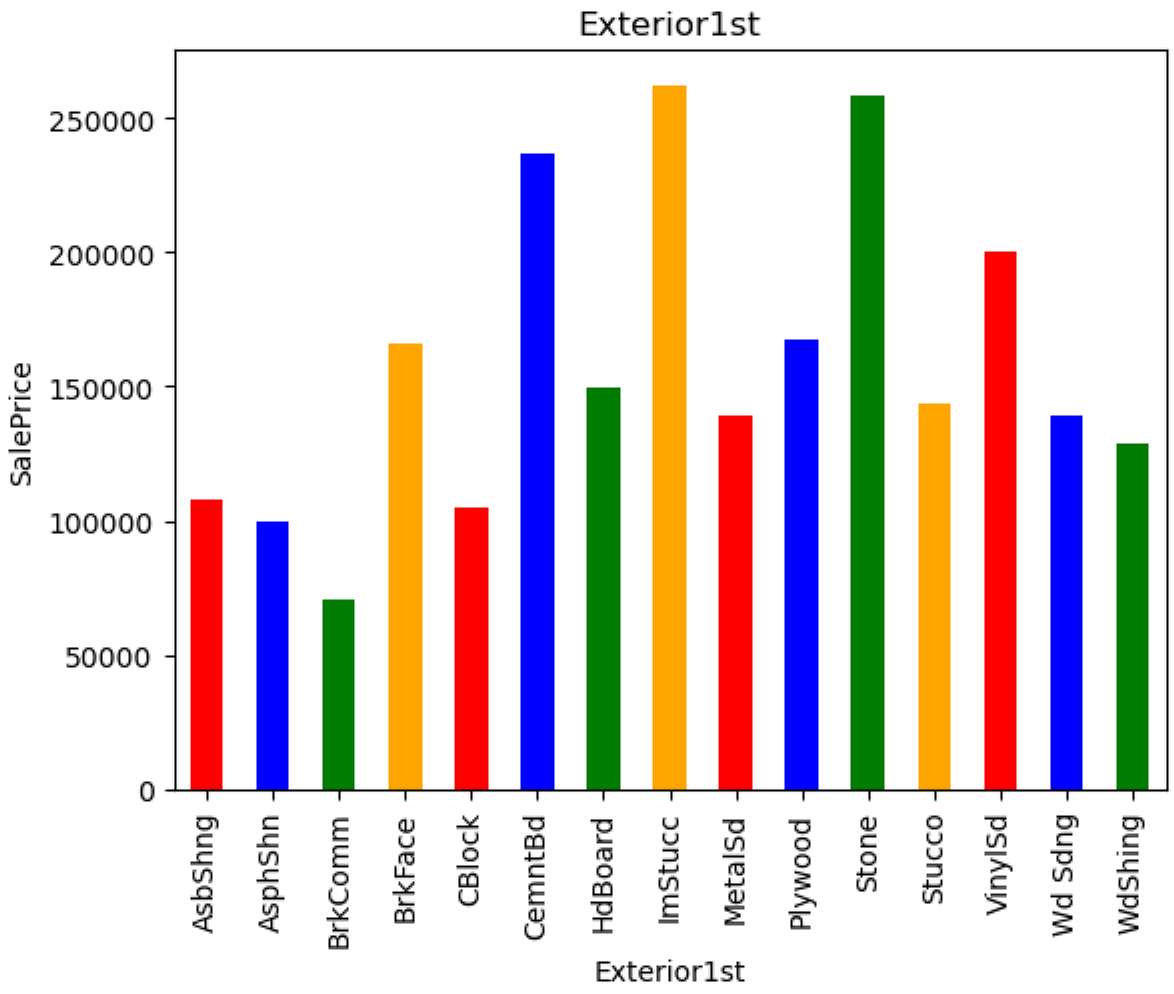

HouseStyle
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle

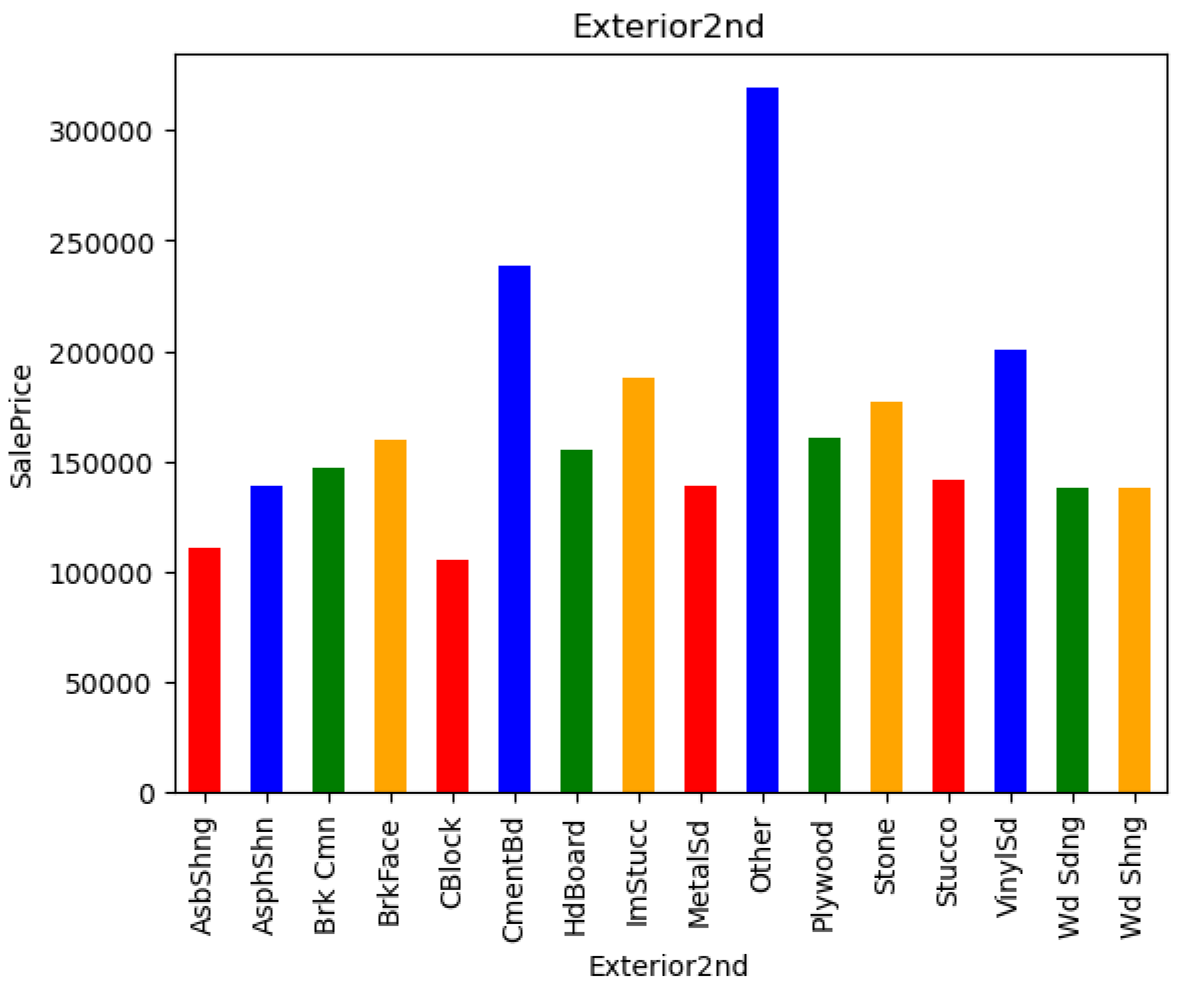

OverallQual
SalePrice
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
OverallQual

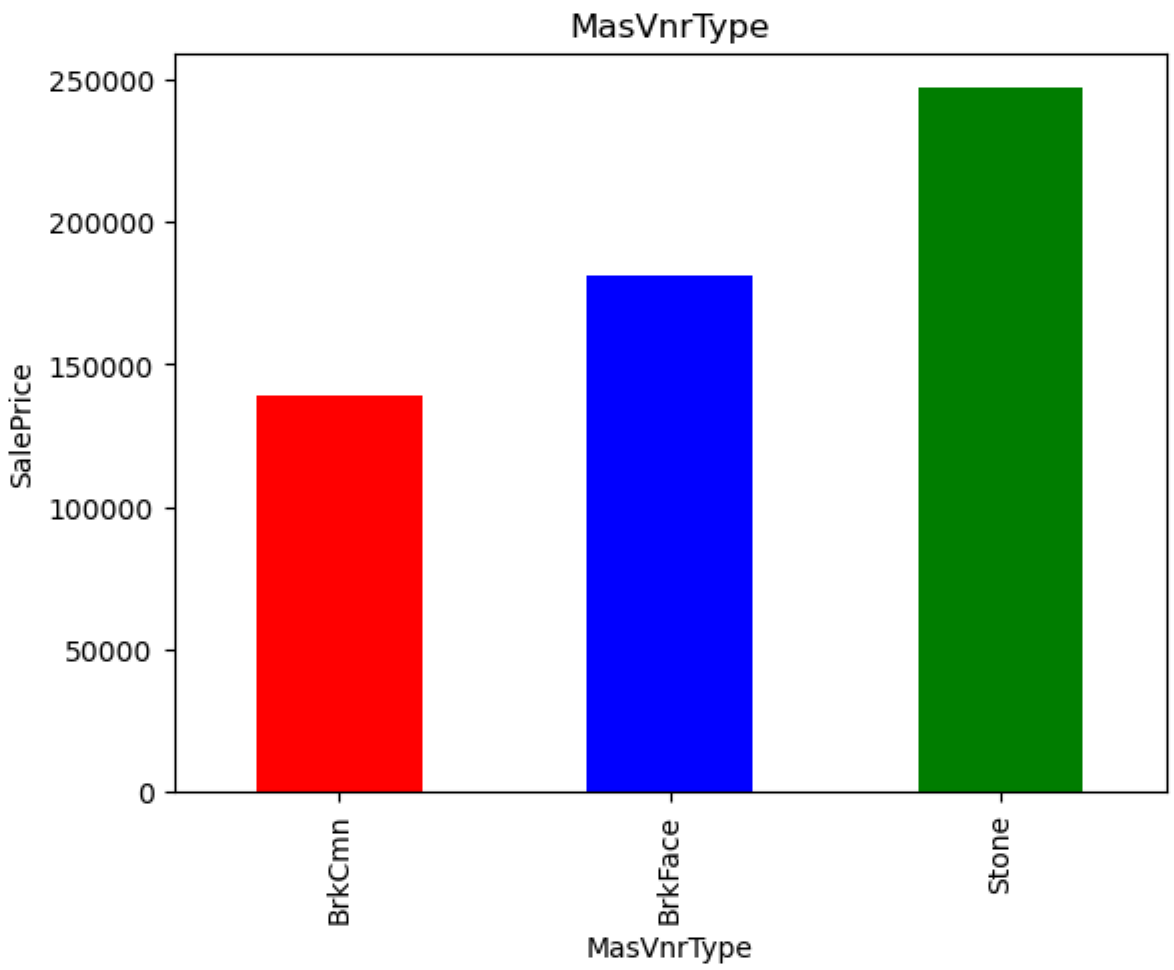

OverallCond
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
1
2
3
4
5
6
7
8
9
OverallCond

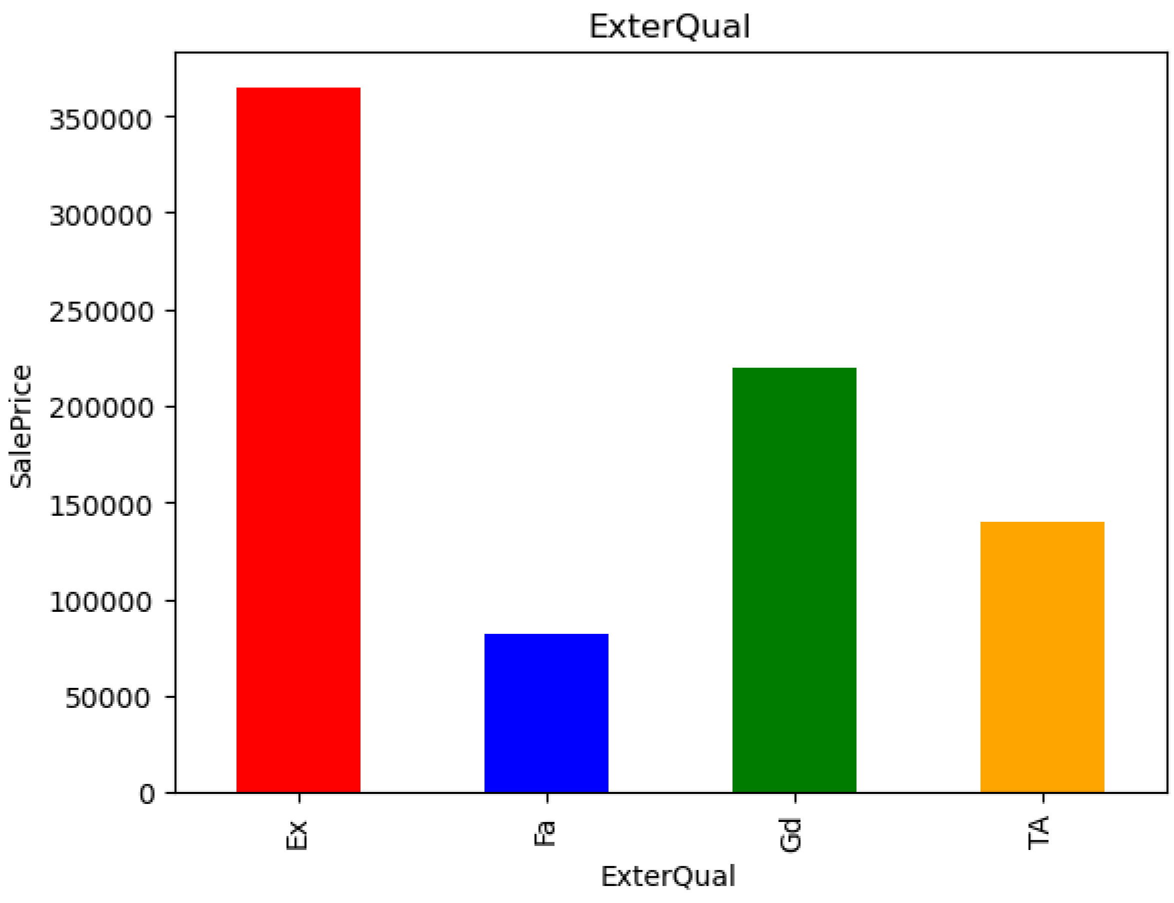

RoofStyle
SalePrice
0
50000
100000
150000
200000
Flat
Gable
Gambrel
Hip
Mansard
Shed
RoofStyle

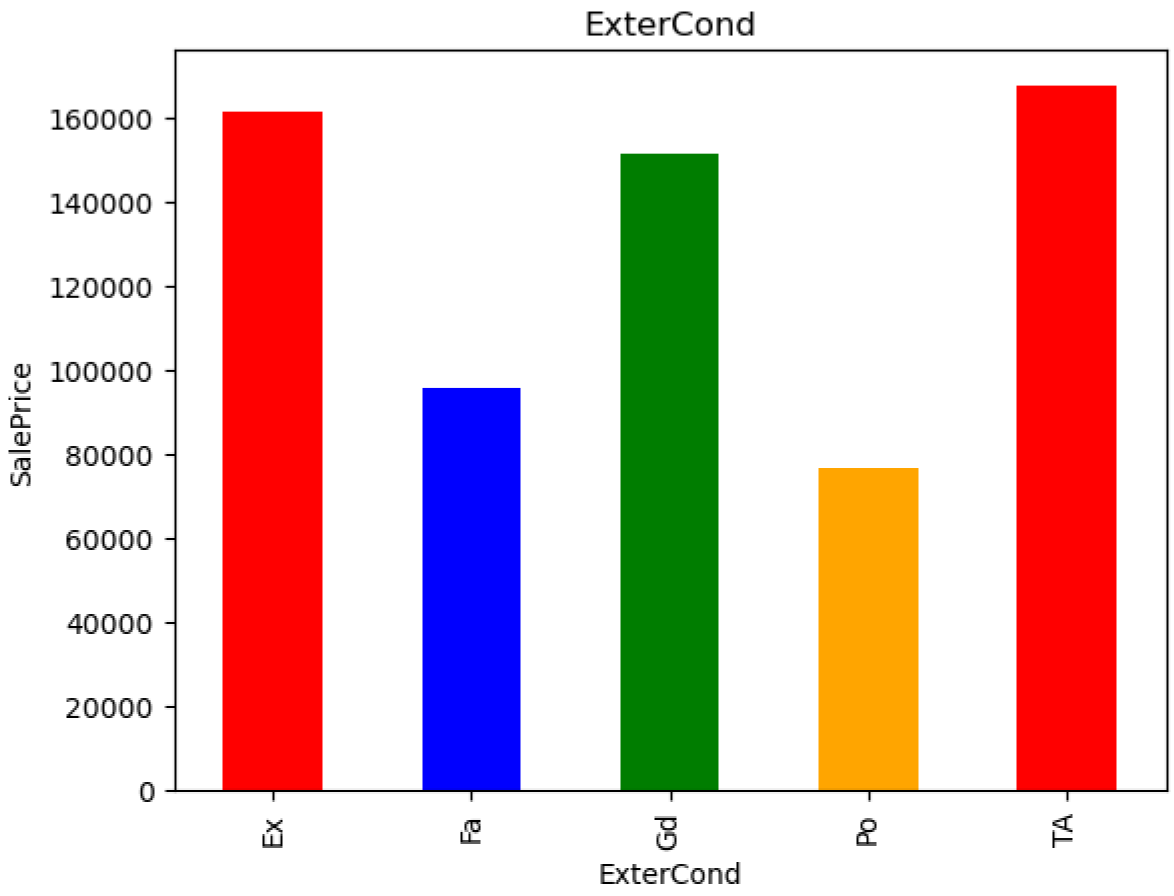

RoofMatl
SalePrice
300000
250000
200000
150000
100000
50000
0
ClyTile
CompShg
Membran
Metal
Roll
Tar&Grv
WdShake
WdShngl
RoofMatl

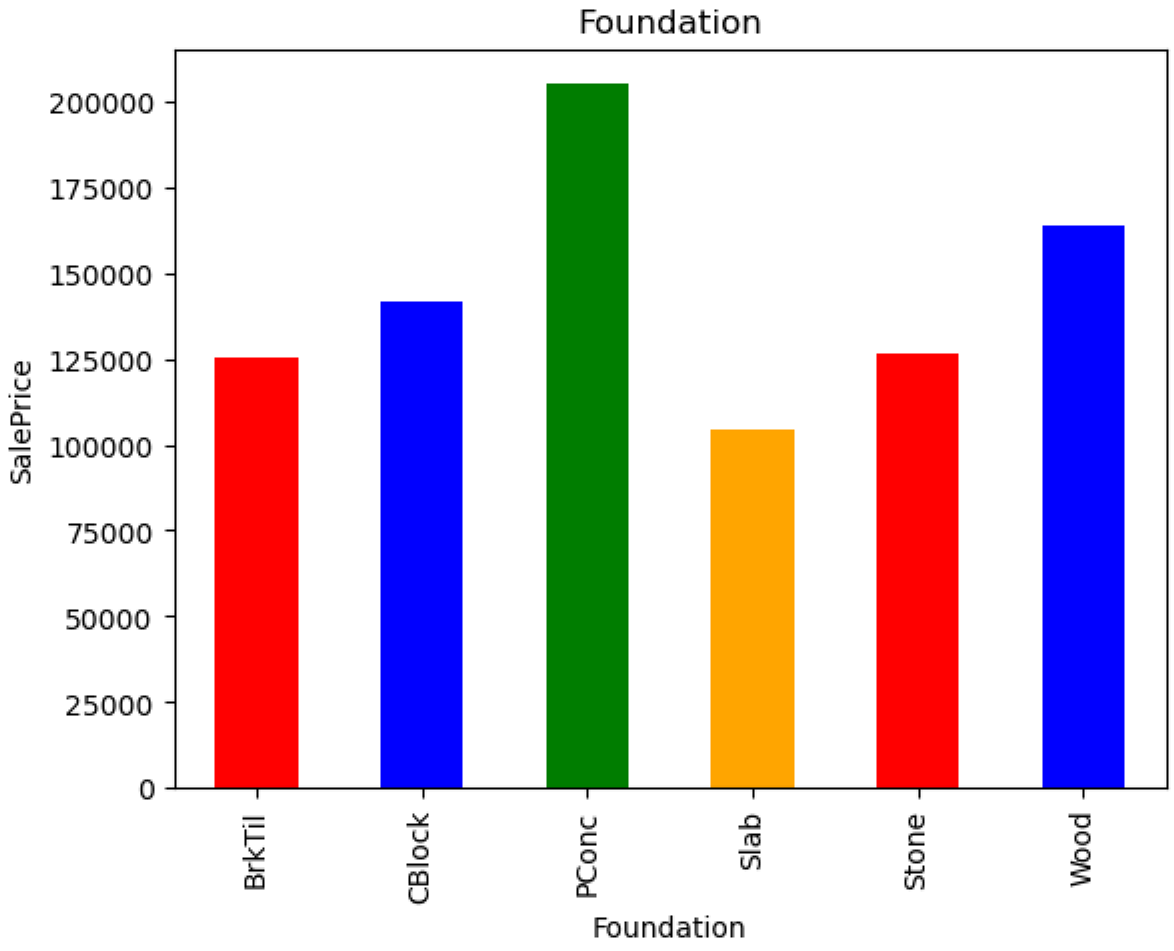

Exterior1st
SalePrice
250000
200000
150000
100000
50000
0
AsbShng
AsphShn
BrkComm
BrkFace
CBlock
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
WdShing
Exterior1st

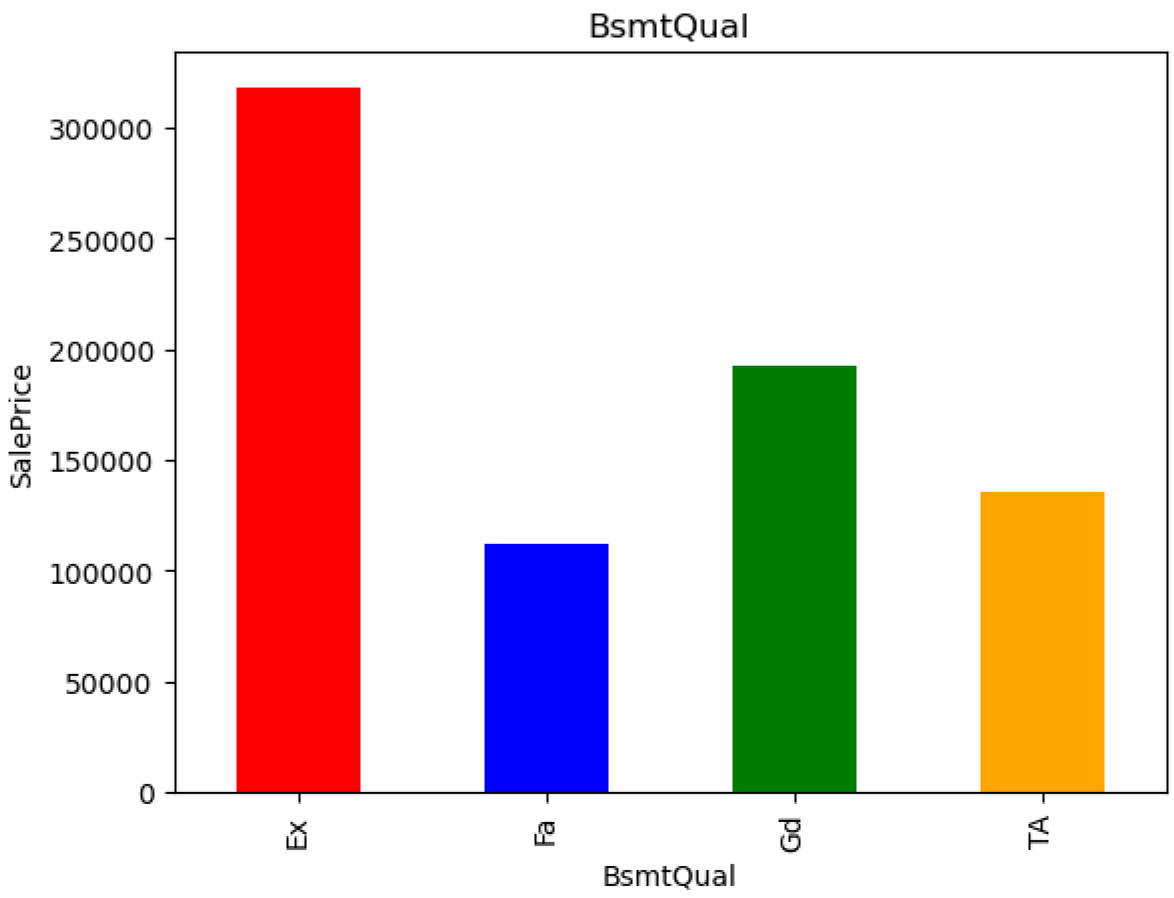

Exterior2nd
SalePrice
0
50000
100000
150000
200000
250000
300000
AsbShng
AsphShn
Brk Cmn
BrkFace
CBlock
CmentBd
HdBoard
ImStucc
MetalSd
Other
Plywood
Stone
Stucco
VinylSd
Wd Sdng
Wd Shng
Exterior2nd

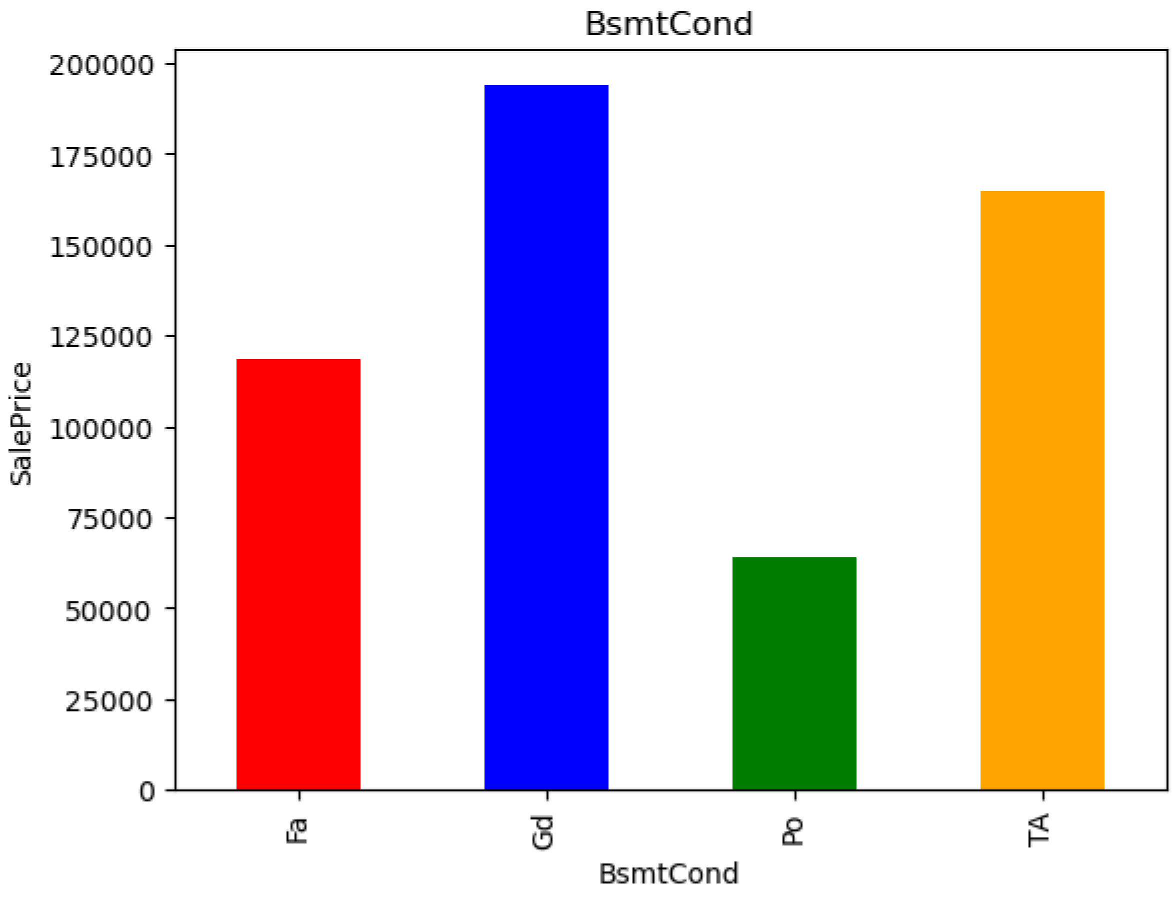

MasVnrType
SalePrice
250000
200000
150000
100000
50000
0
BrkCmn
BrkFace
Stone
MasVnrType

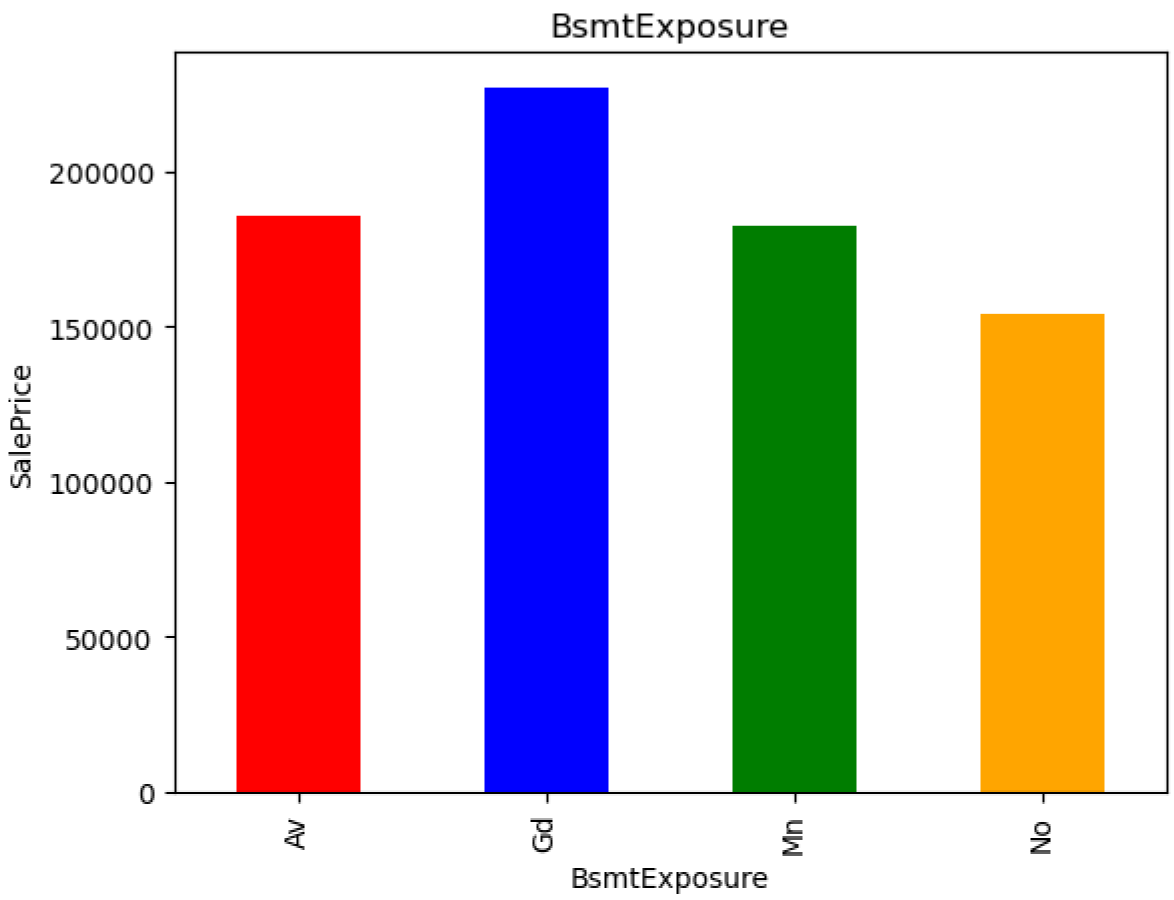

ExterQual
SalePrice
350000
300000
250000
200000
150000
100000
50000
0
Ex
Fa
Gd
TA
ExterQual

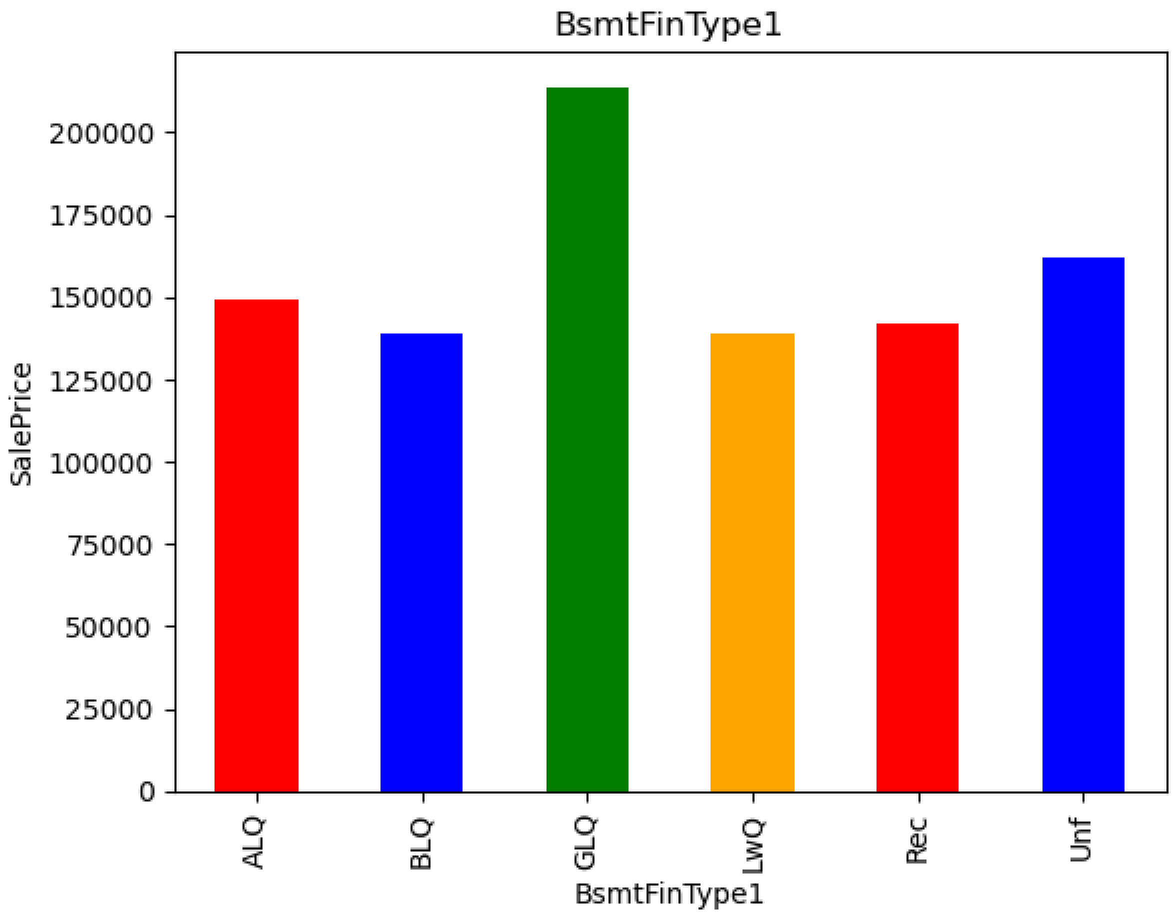

ExterCond
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Ex
Fa
Gd
Po
TA
ExterCond

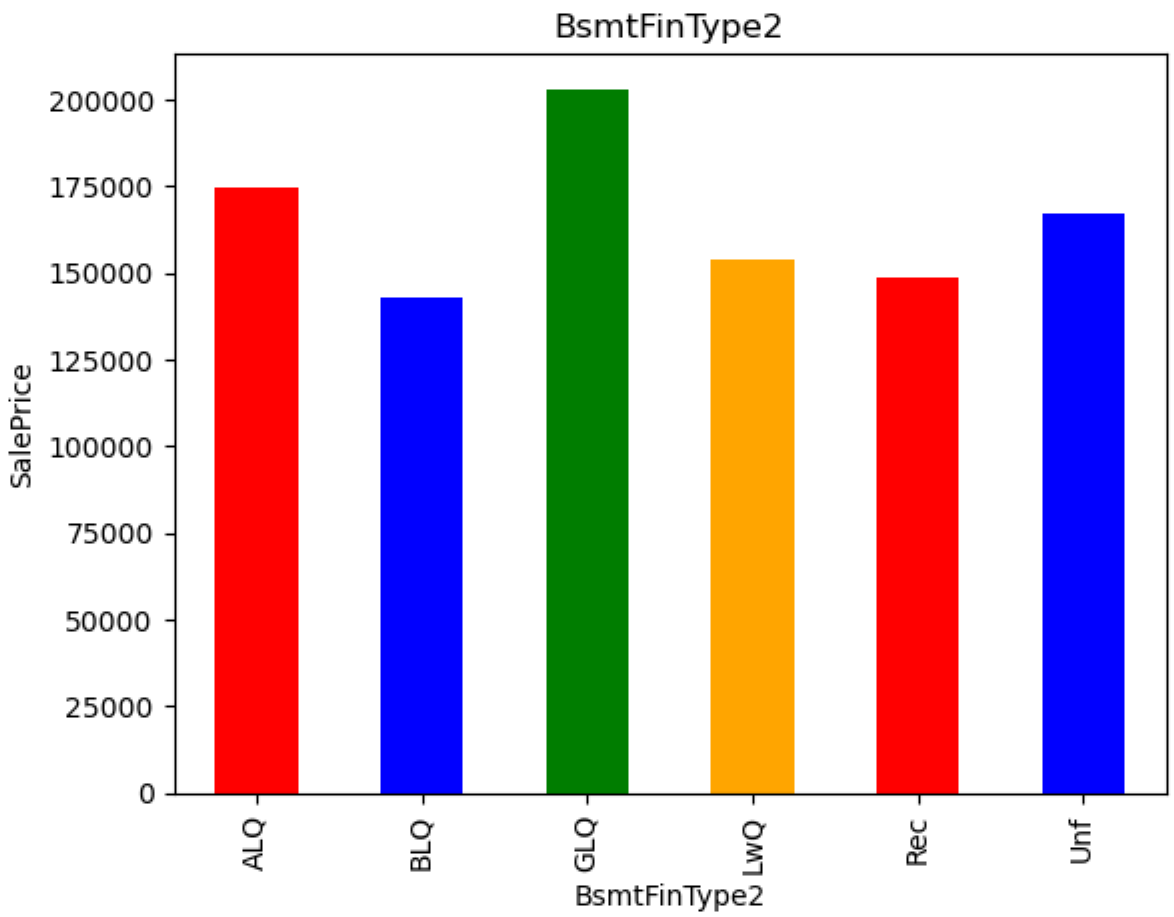

Foundation
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
BrkTil
CBlock
PConc
Slab
Stone
Wood
Foundation

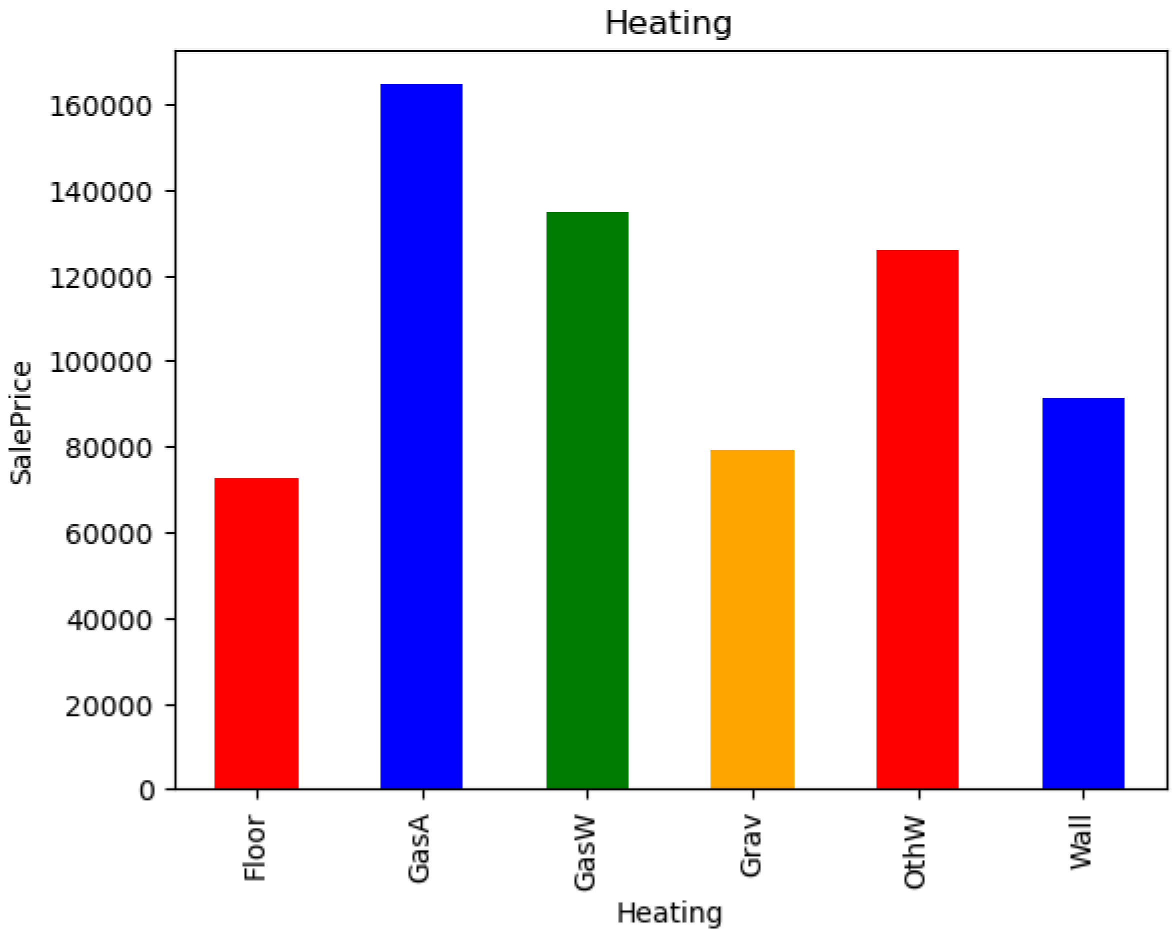

BsmtQual
SalePrice
300000
250000
200000
150000
100000
50000
0
Ex
Fa
Gd
TA
BsmtQual

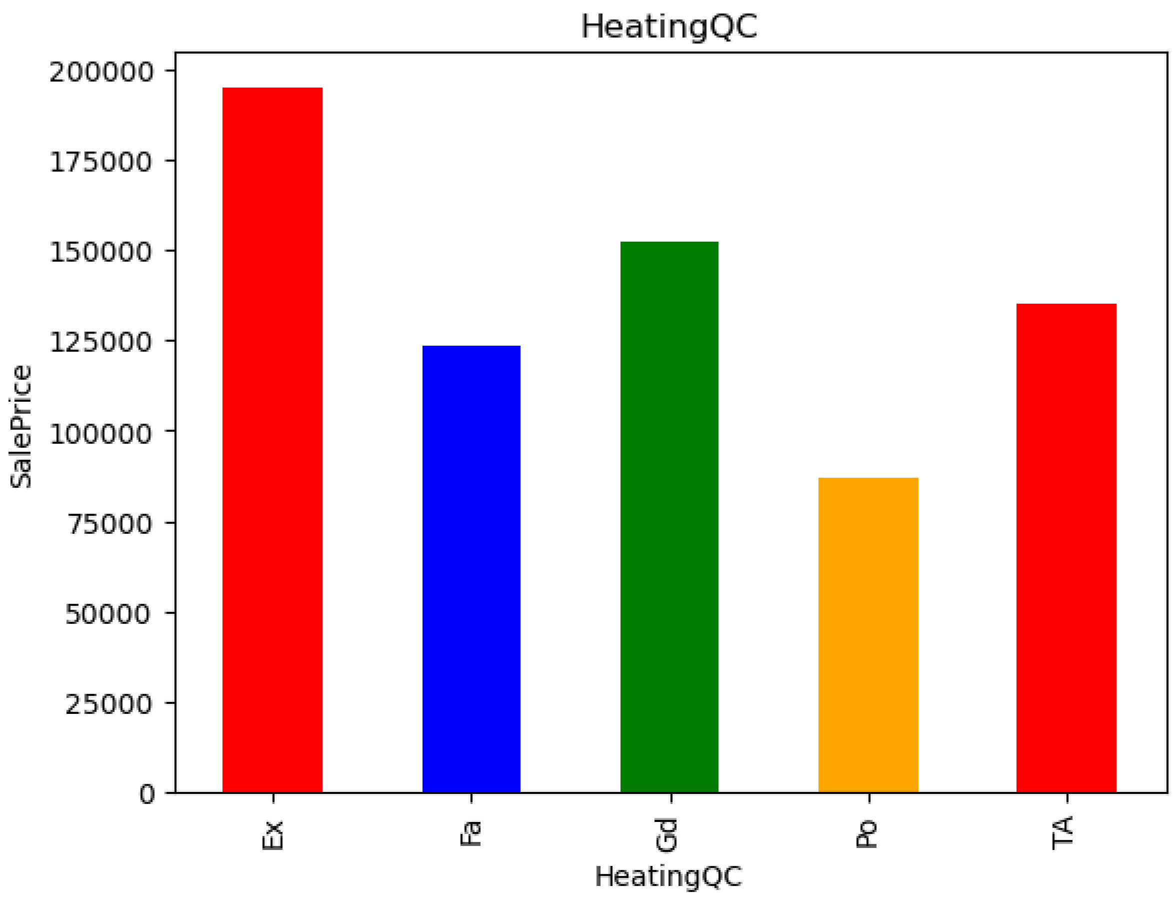

BsmtCond
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Fa
Gd
Po
TA
BsmtCond

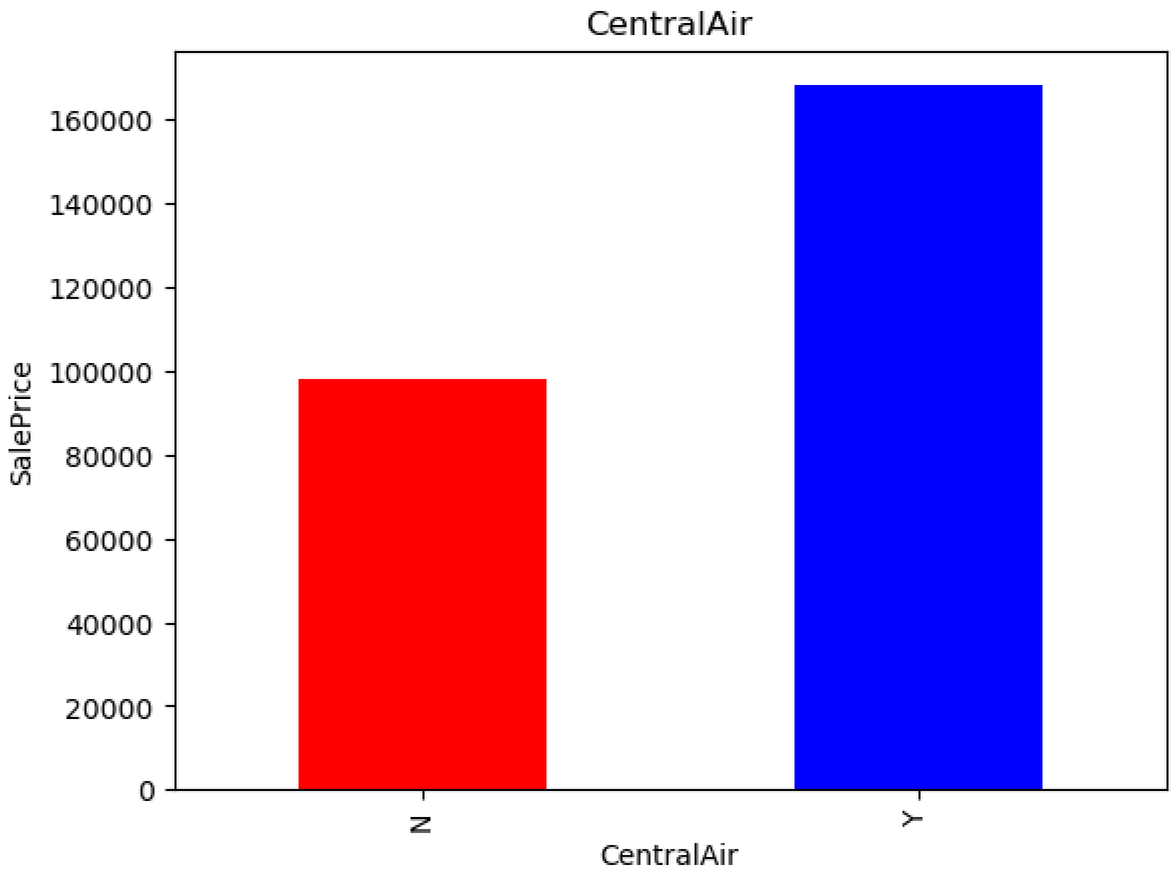

BsmtExposure
SalePrice
200000
150000
100000
50000
0
Av
Gd
Mn
No
BsmtExposure

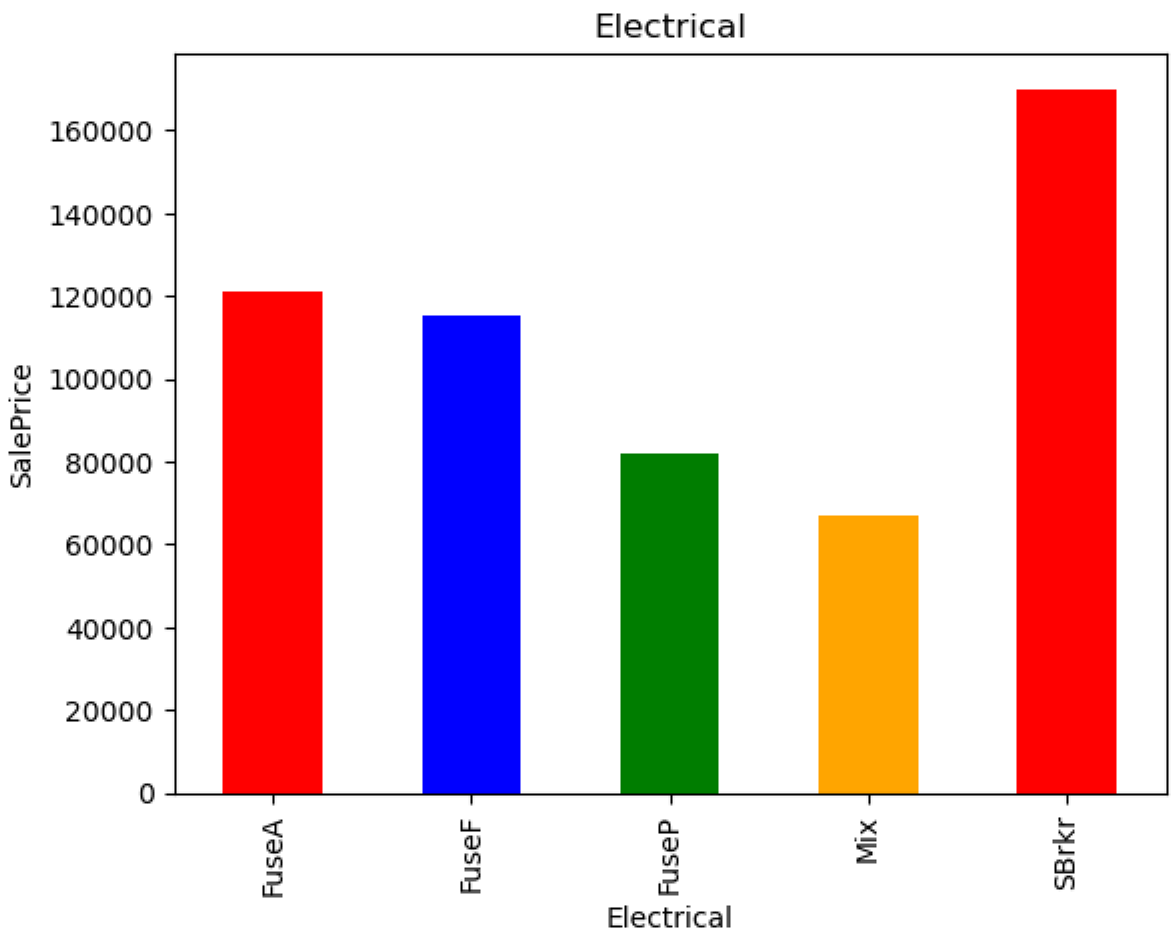

BsmtFinType1
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinType1

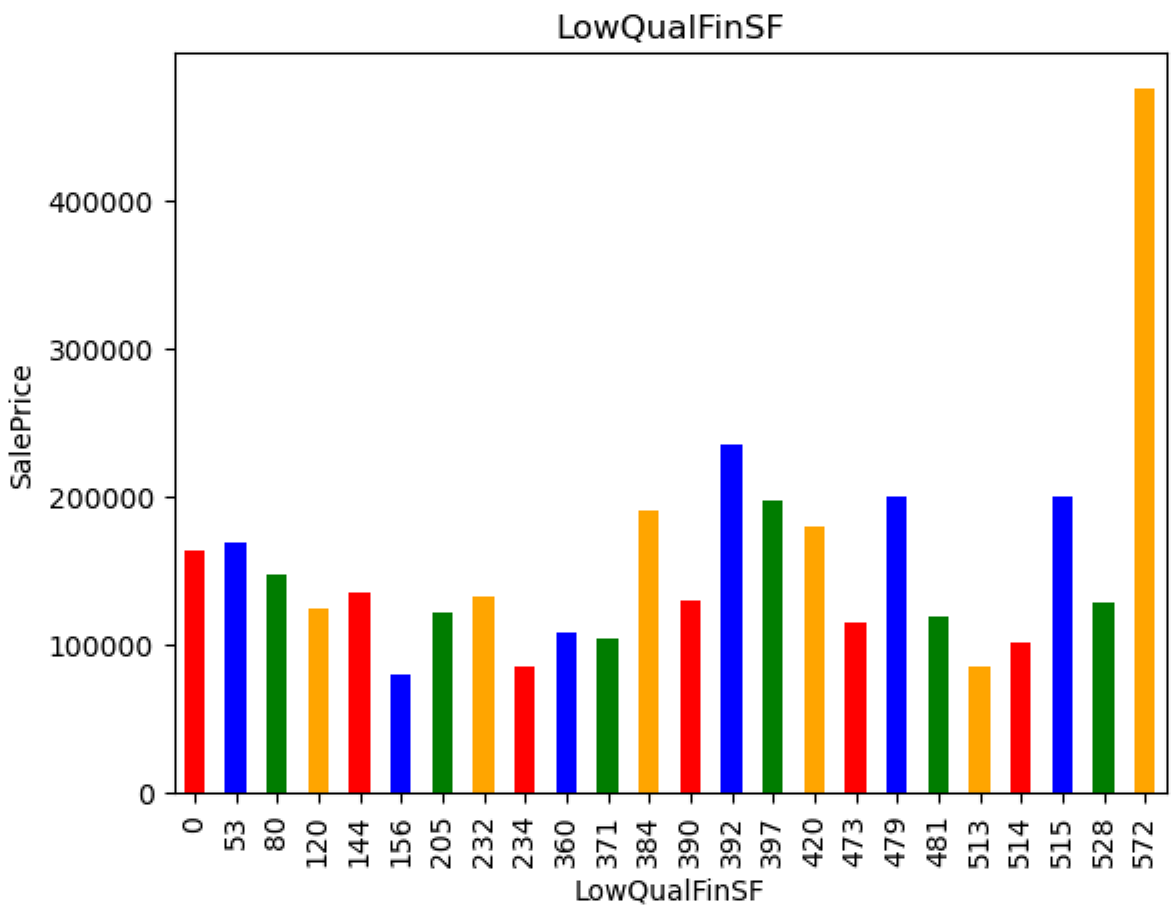

BsmtFinType2
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinType2

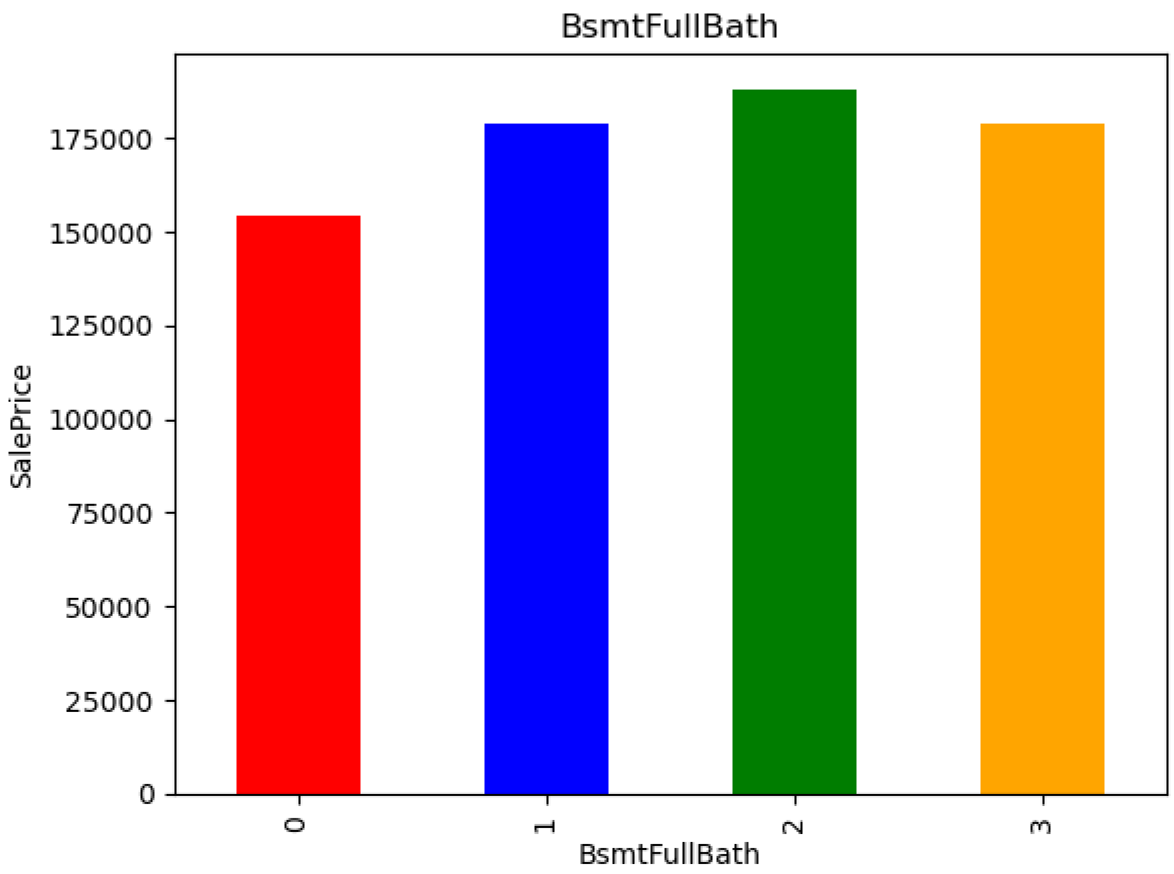

Heating
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
Floor
GasA
GasW
Grav
OthW
Wall
Heating

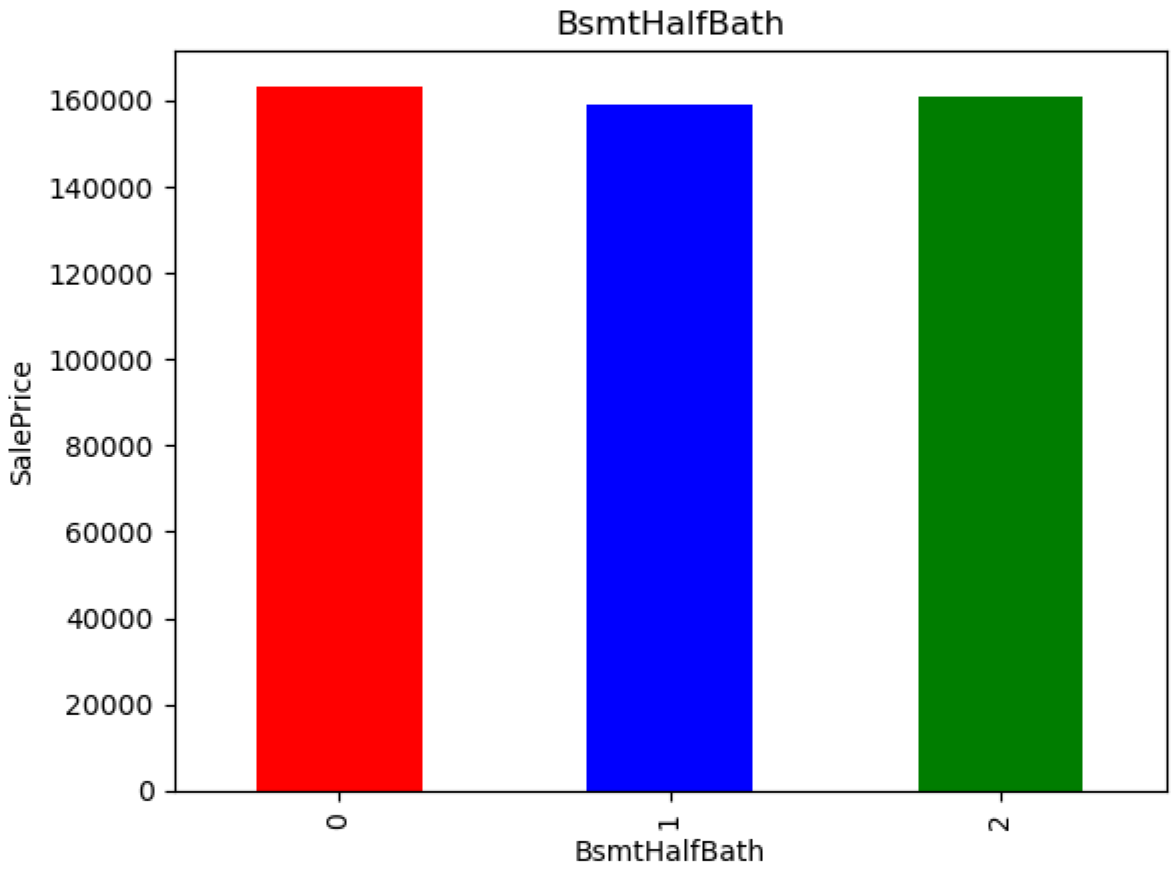

HeatingQC
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Ex
Fa
Gd
Po
TA
HeatingQC

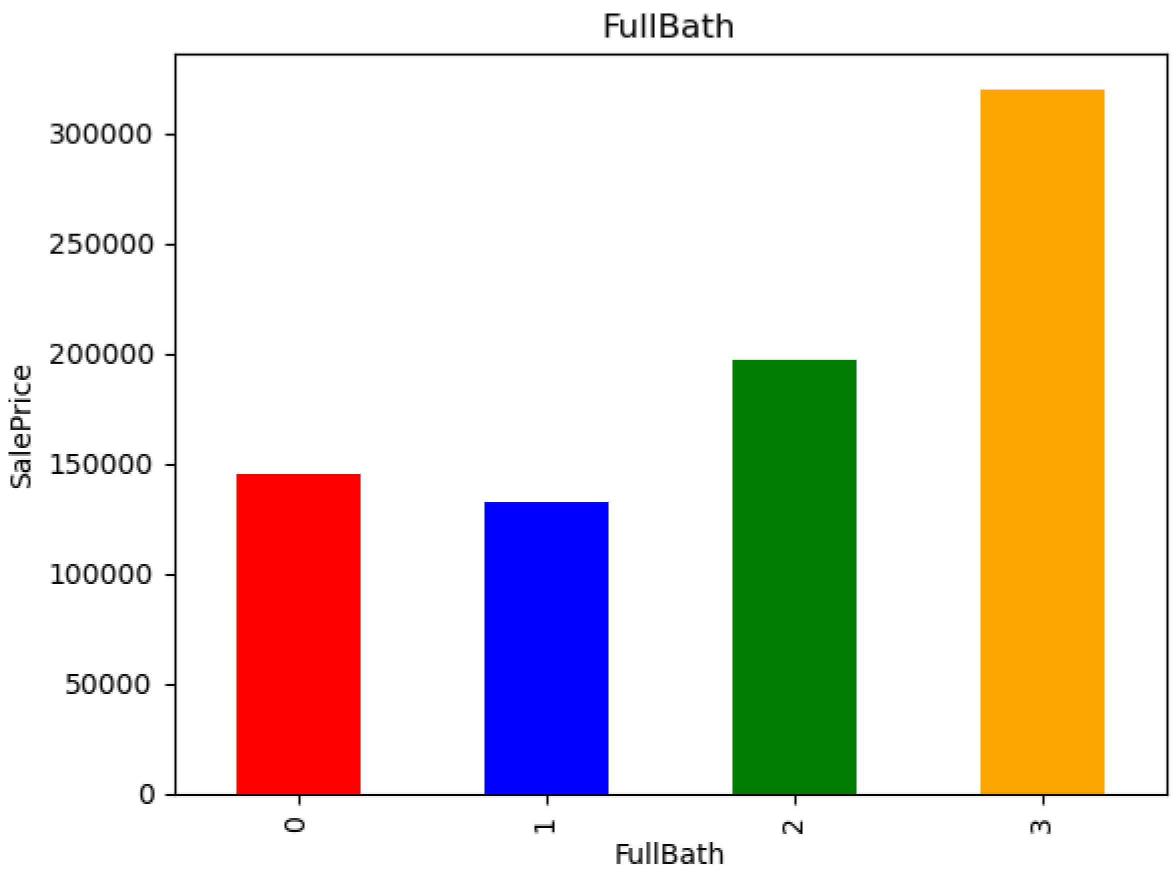

CentralAir
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
N
Y
CentralAir

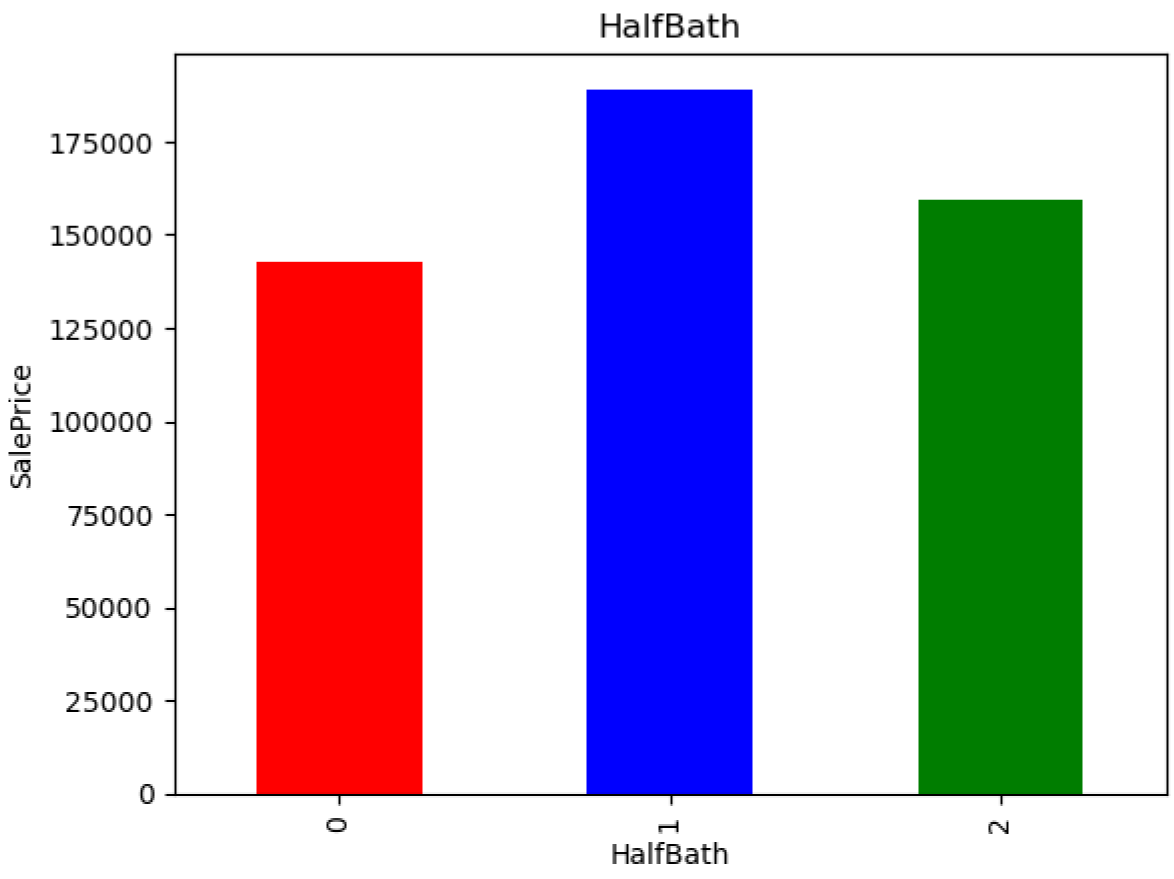

Electrical
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
FuseA
FuseF
FuseP
Mix
SBrkr
Electrical

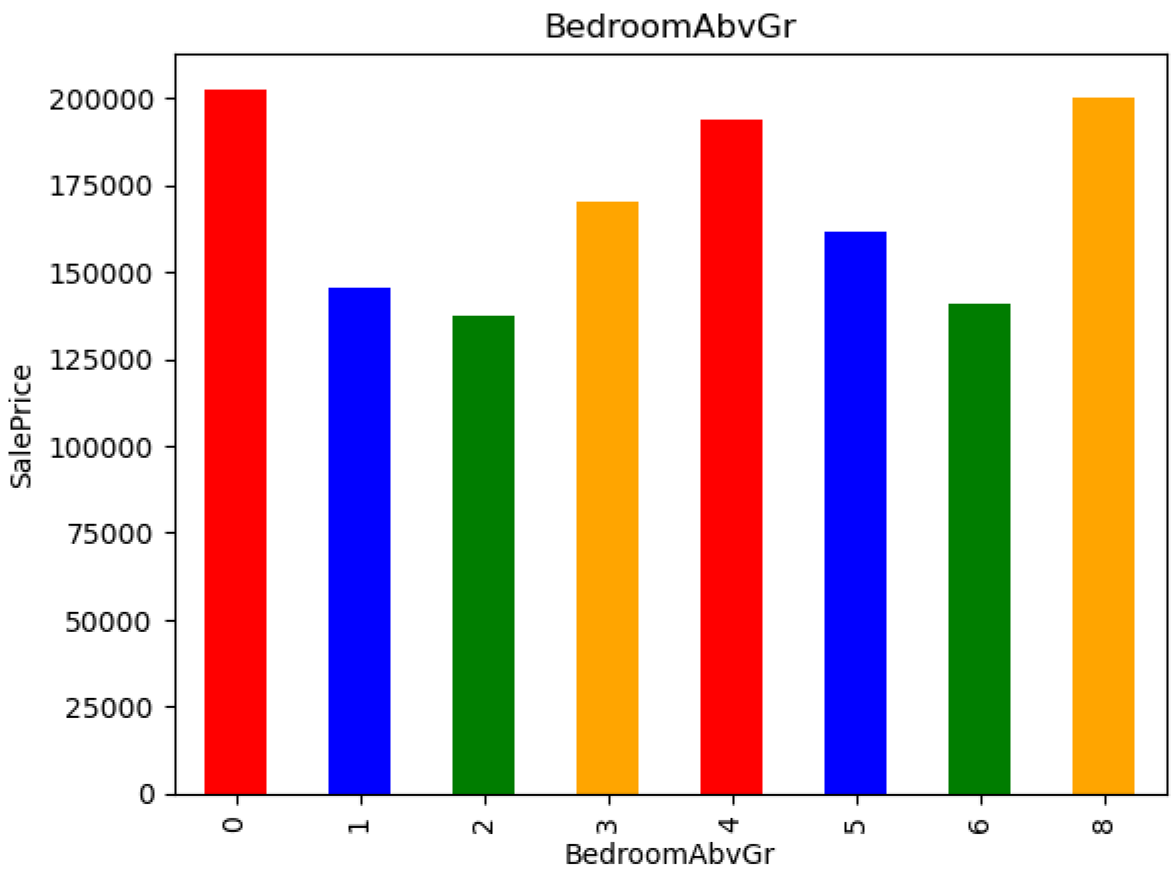

LowQualFinSF
SalePrice
0
100000
200000
300000
400000
0
53
80
120
144
156
205
232
234
360
371
384
390
392
397
420
473
479
481
513
514
515
528
572
LowQualFinSF

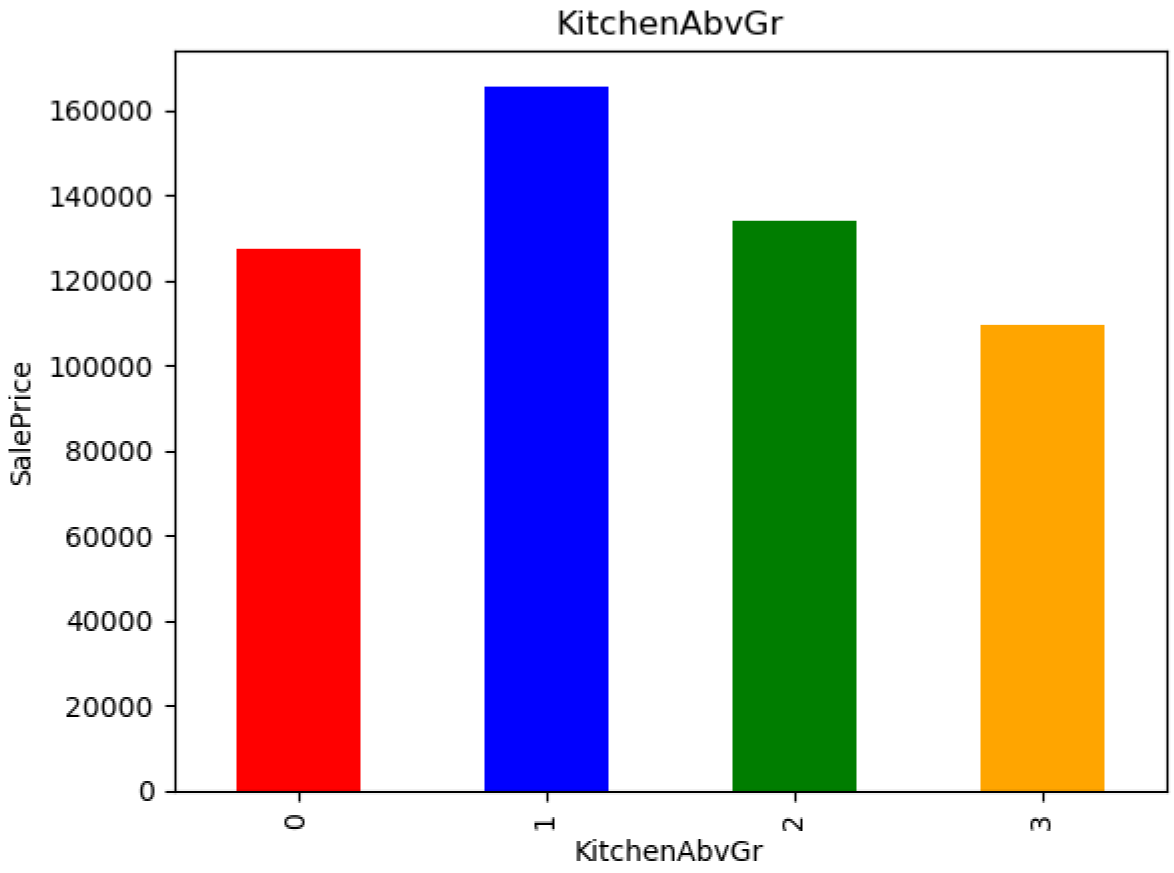

BsmtFullBath
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
BsmtFullBath

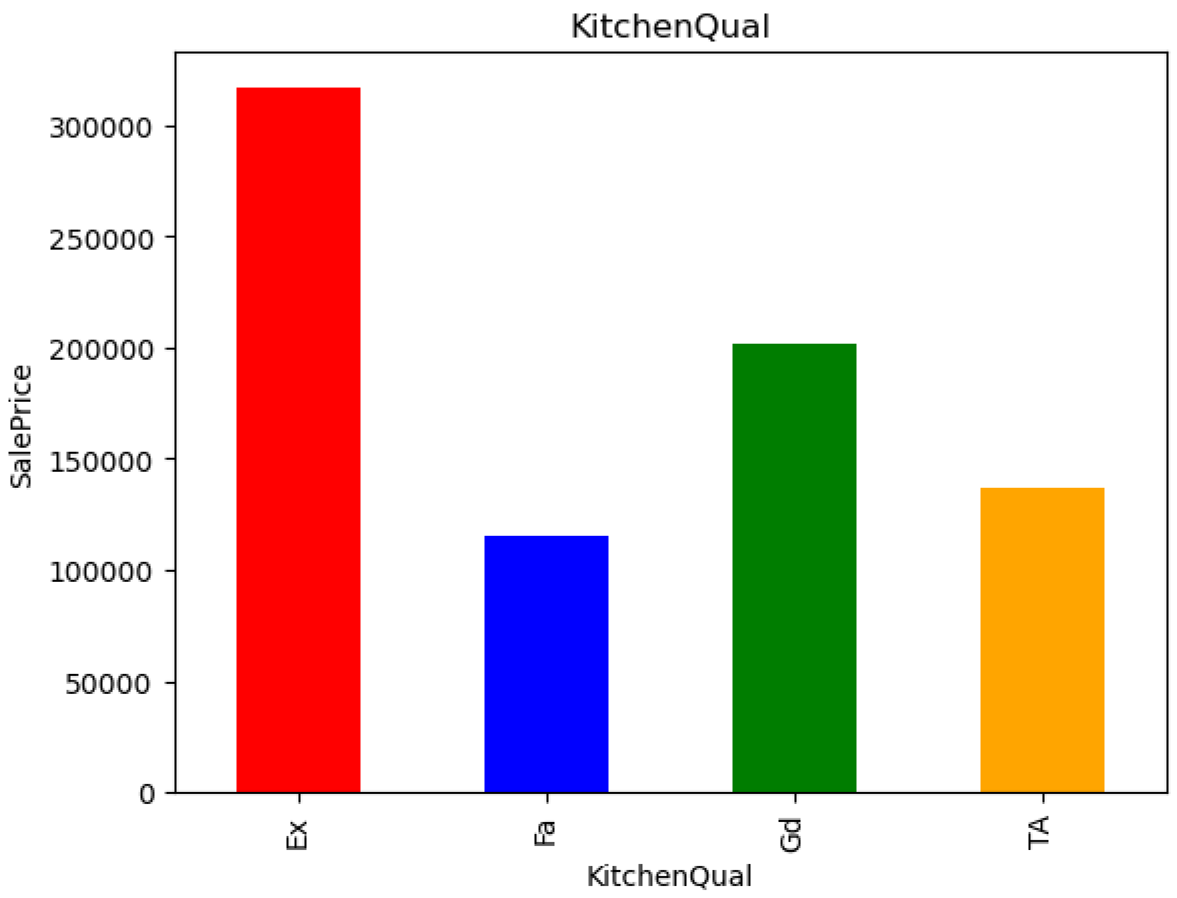

BsmtHalfBath
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
BsmtHalfBath

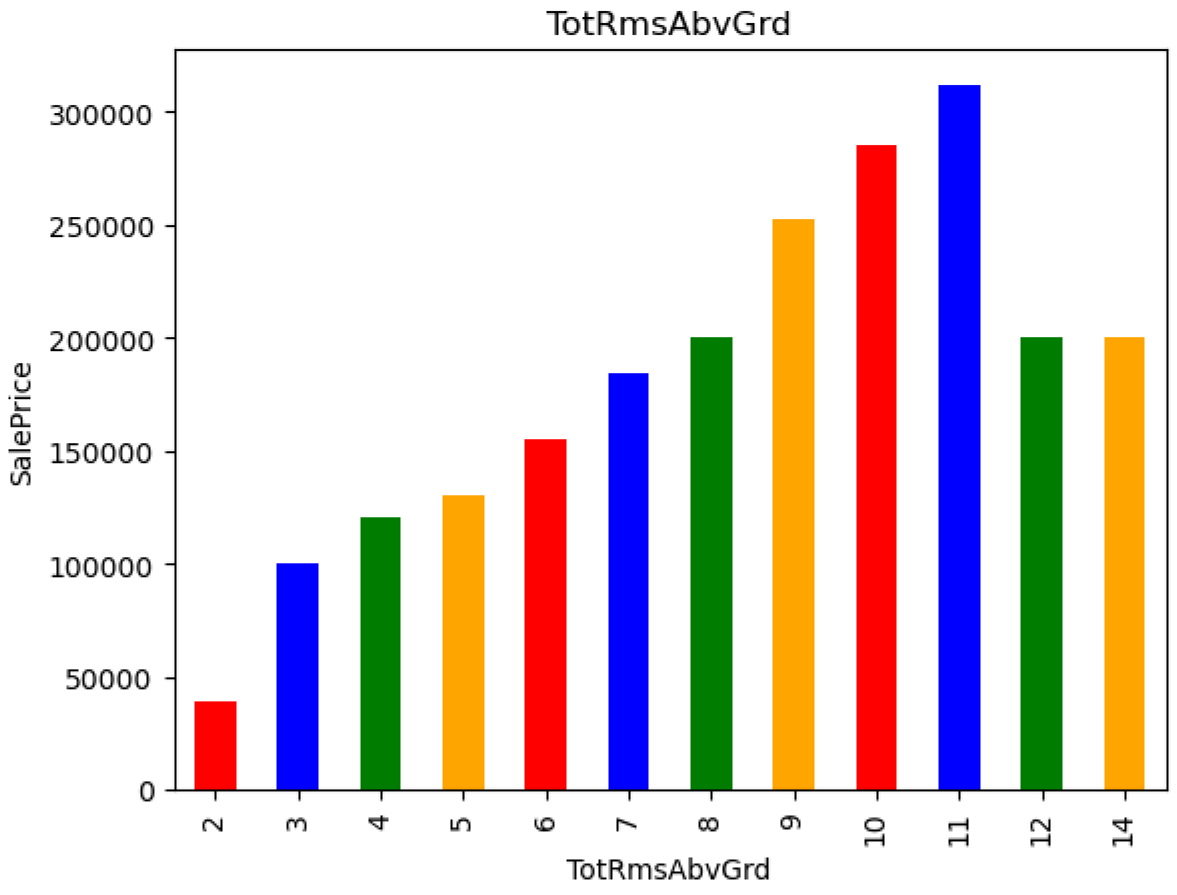

FullBath
SalePrice
300000
250000
200000
150000
100000
50000
0
0
1
2
3
FullBath

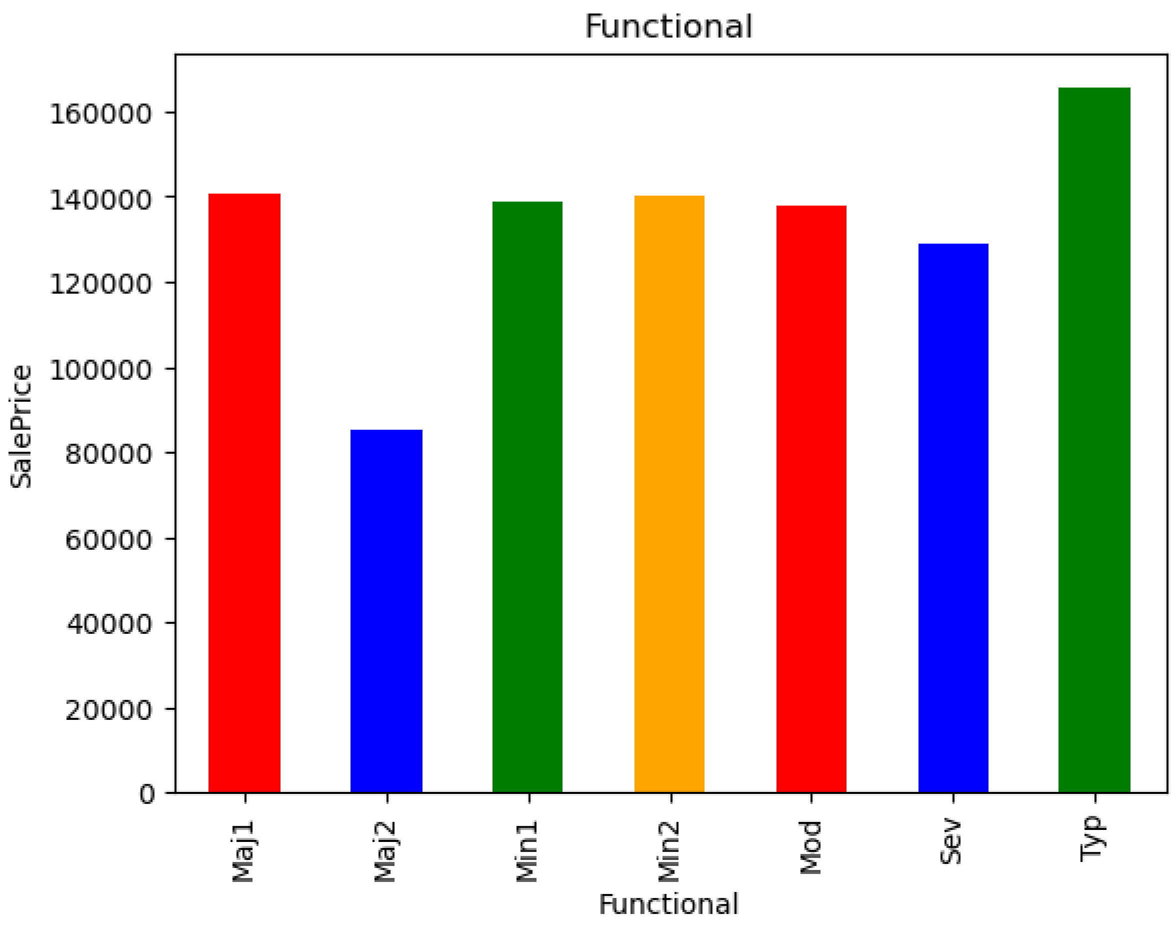

HalfBath
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
HalfBath

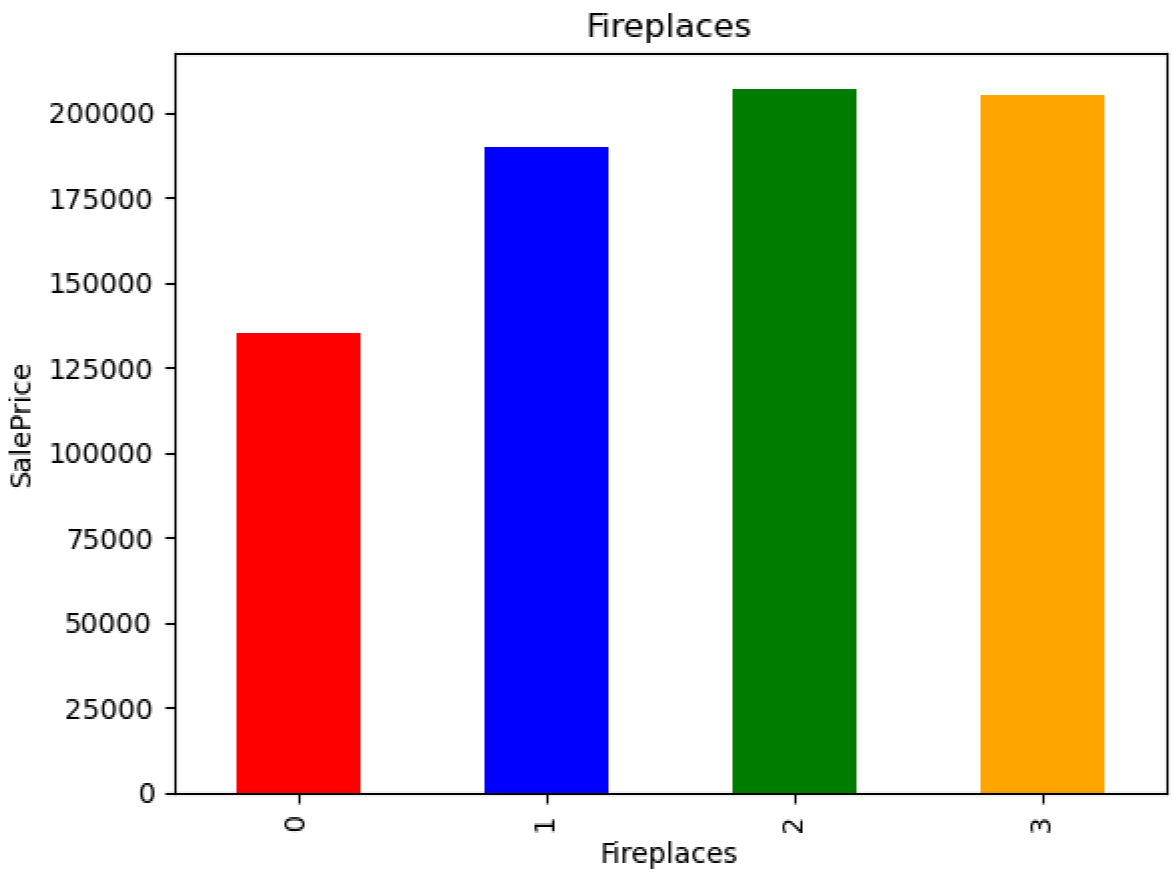

BedroomAbvGr
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
3
4
5
6
8
BedroomAbvGr

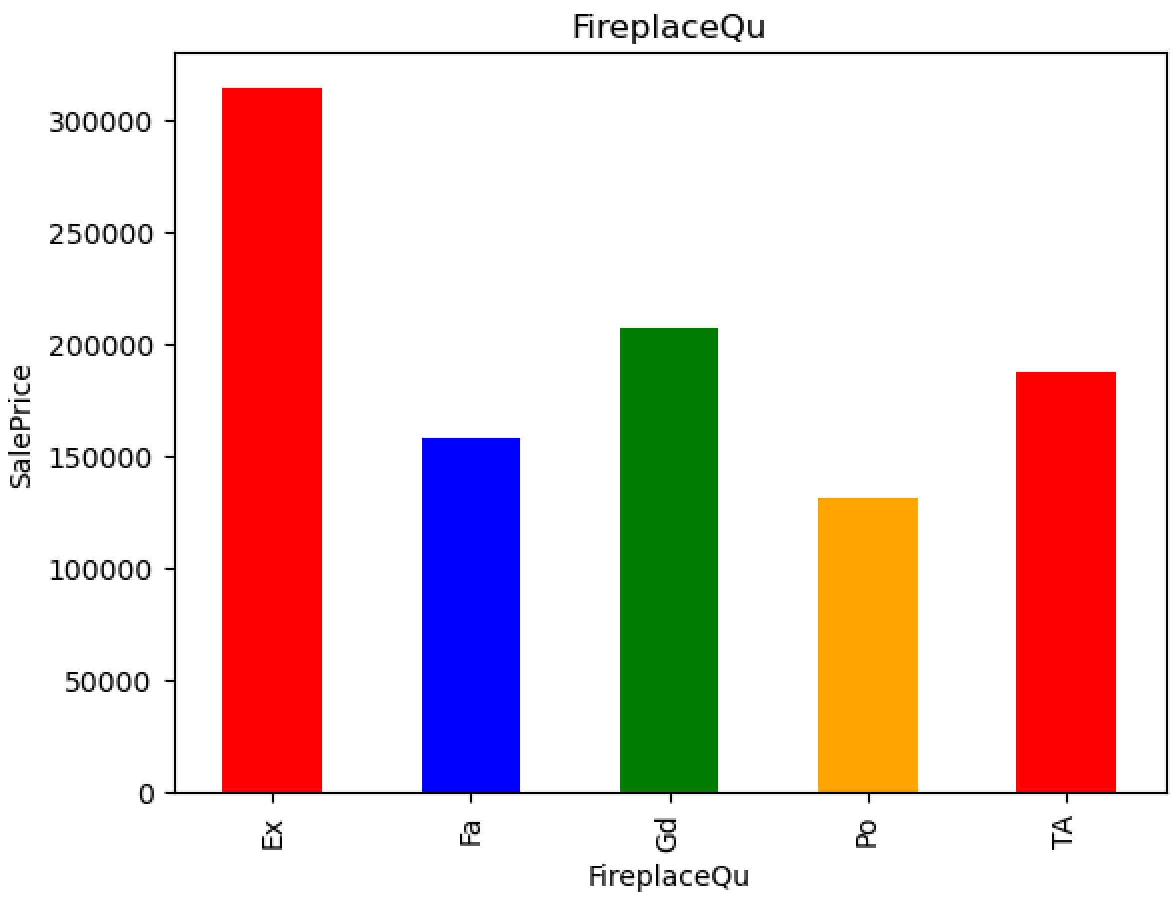

KitchenAbvGr
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
KitchenAbvGr

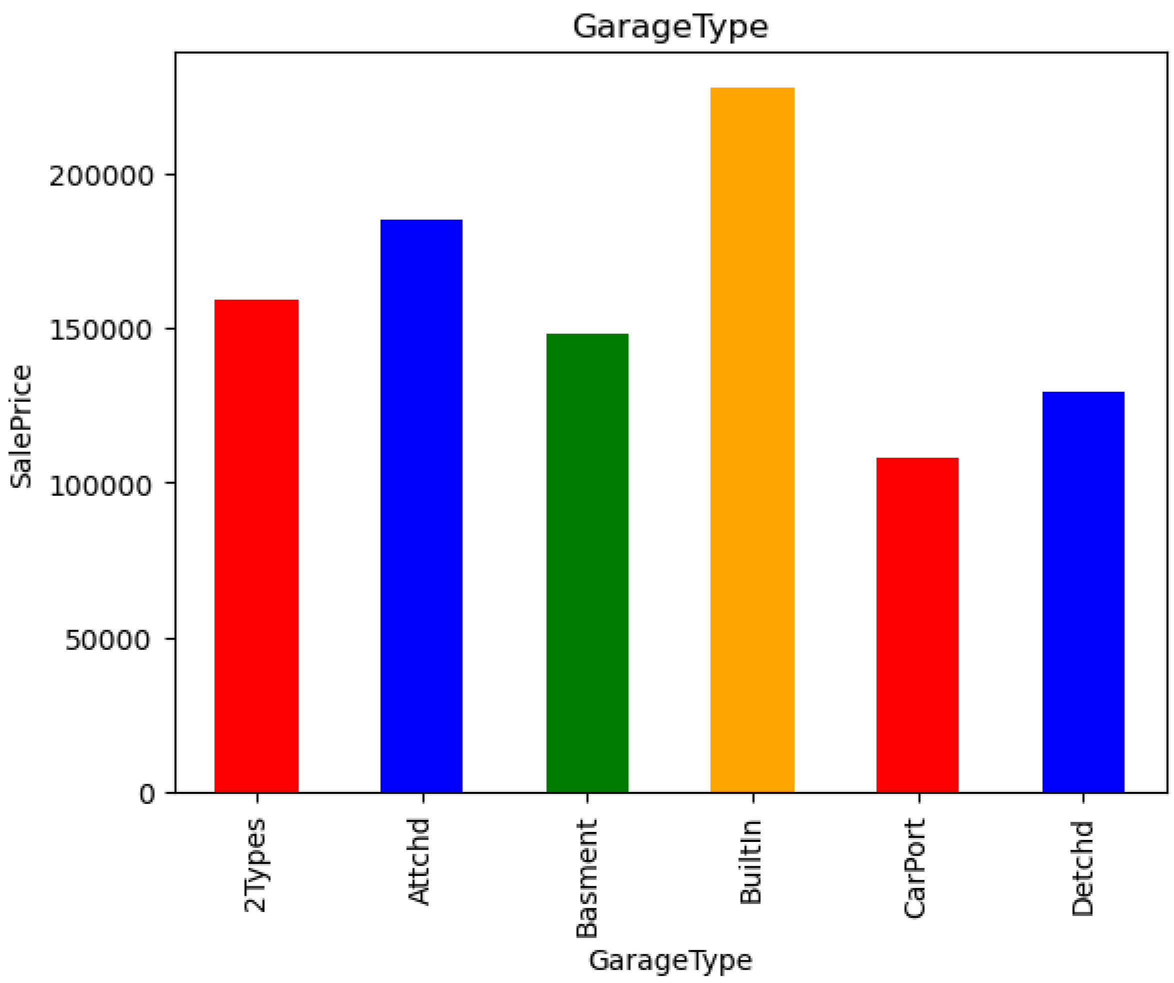

KitchenQual
SalePrice
0
50000
100000
150000
200000
250000
300000
Ex
Fa
Gd
TA
KitchenQual

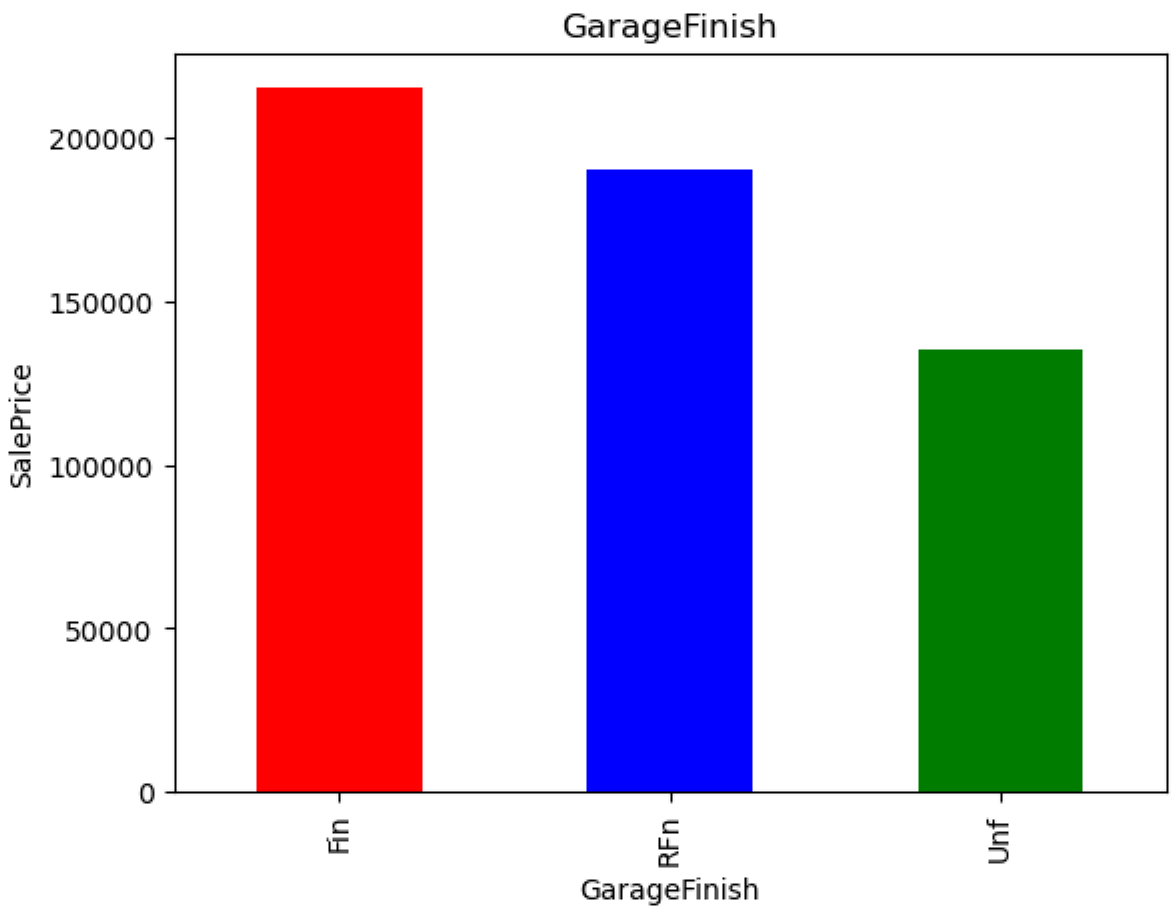

TotRmsAbvGrd
SalePrice
0
50000
100000
150000
200000
250000
300000
2
3
4
5
6
7
8
9
10
11
12
14
TotRmsAbvGrd

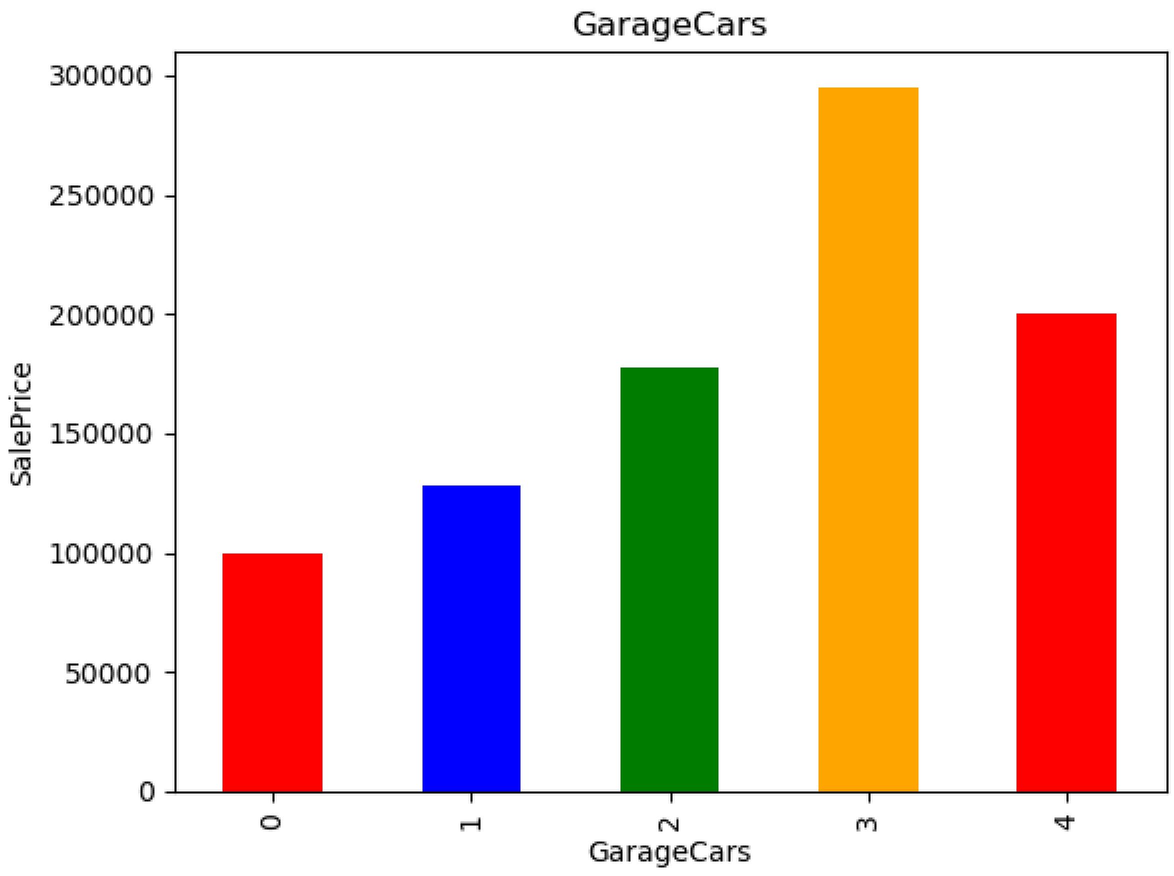

Functional
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ
Functional

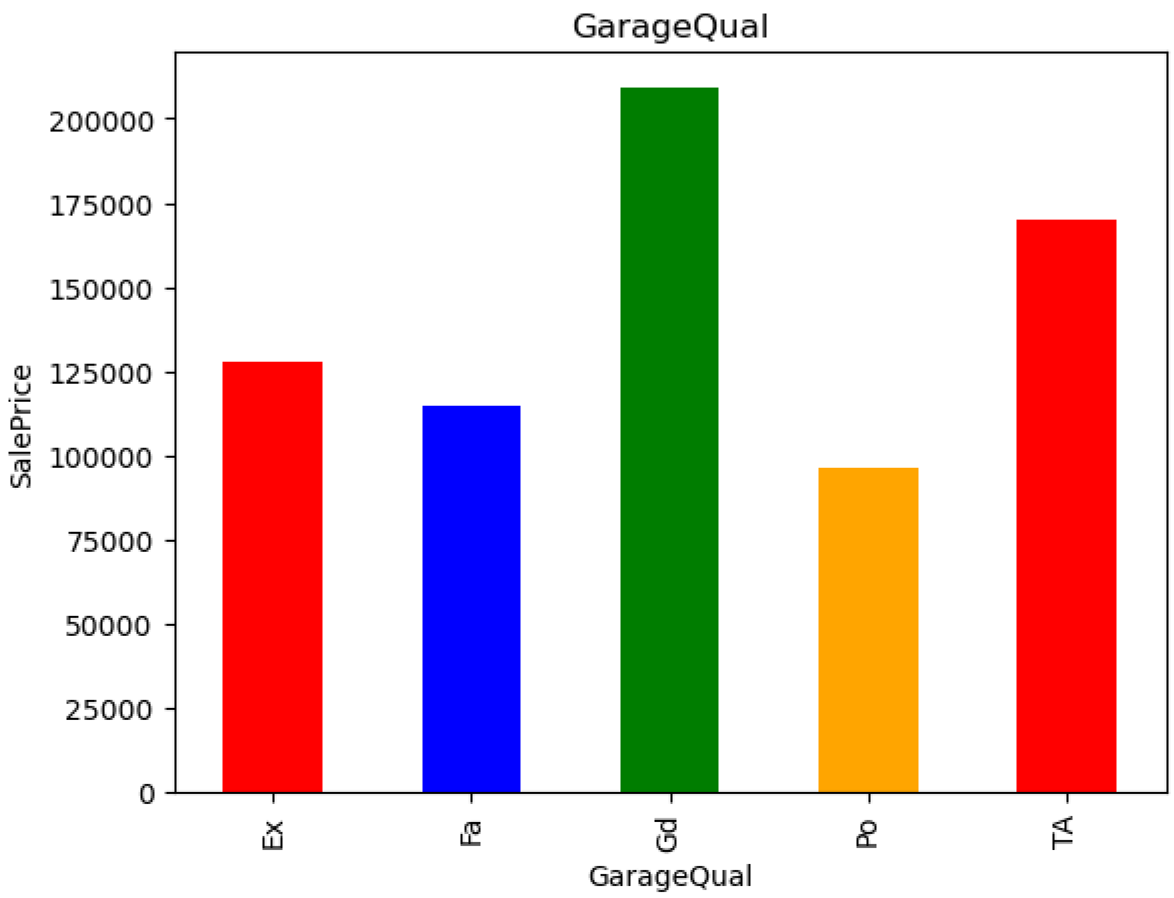

Fireplaces
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
0
1
2
3
Fireplaces

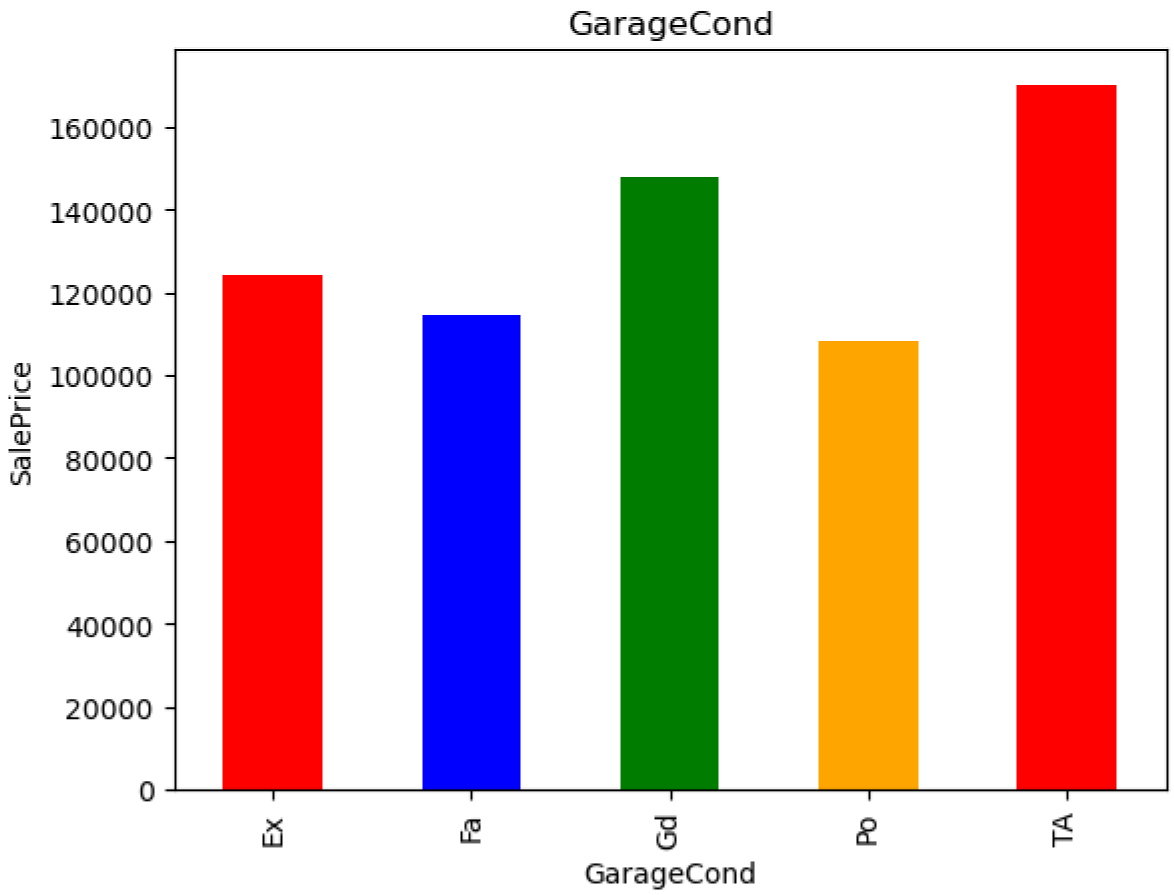

FireplaceQu
SalePrice
300000
250000
200000
150000
100000
50000
0
Ex
Fa
Gd
Po
TA
FireplaceQu

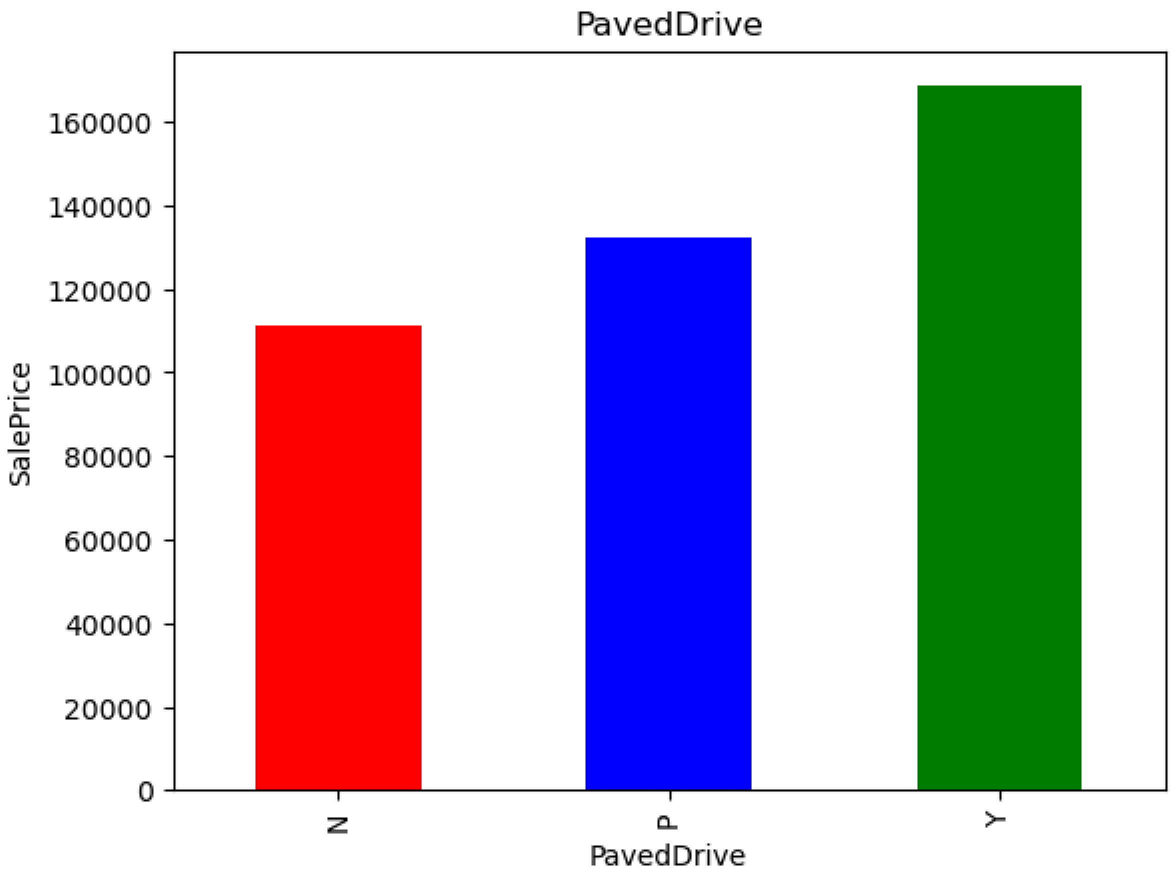

GarageType
SalePrice
0
50000
100000
150000
200000
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
GarageType

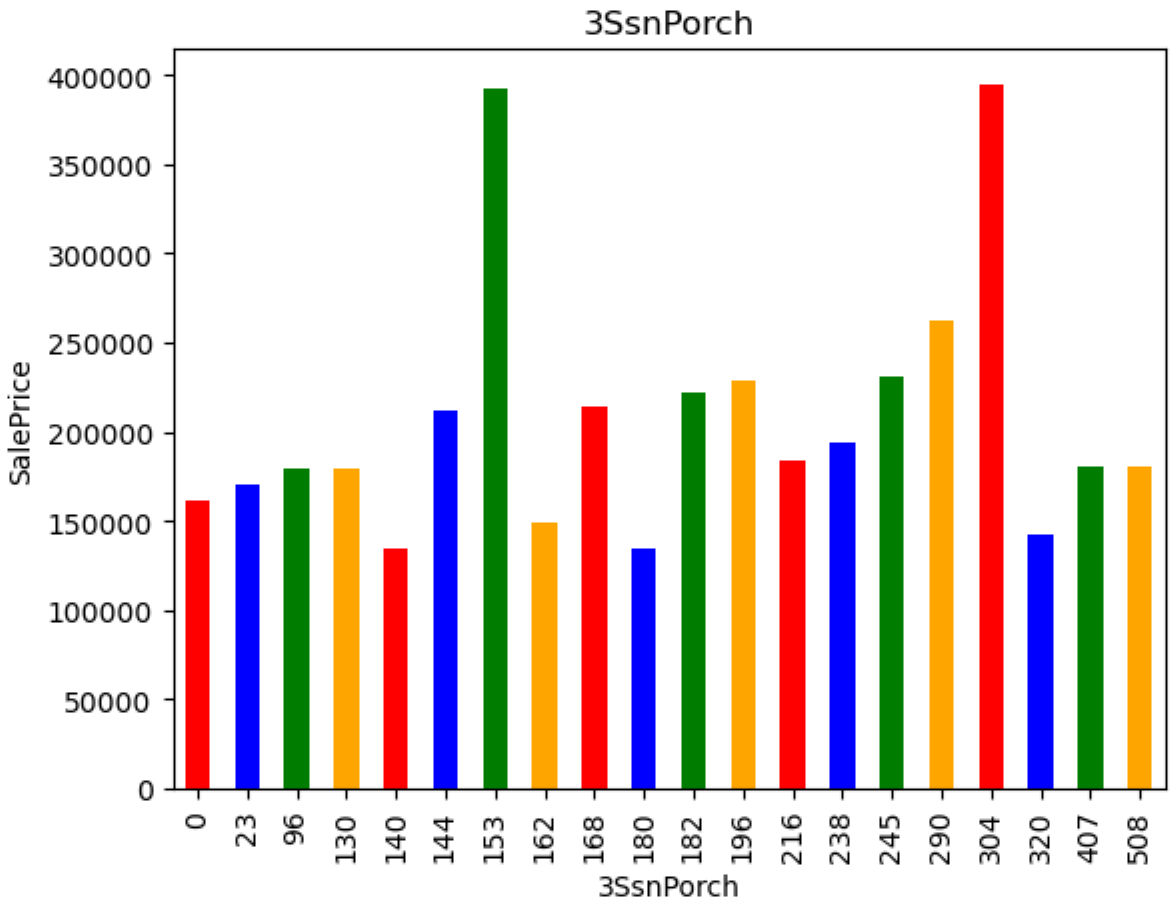

GarageFinish
SalePrice
200000
150000
100000
50000
0
Fin
RFn
Unf
GarageFinish

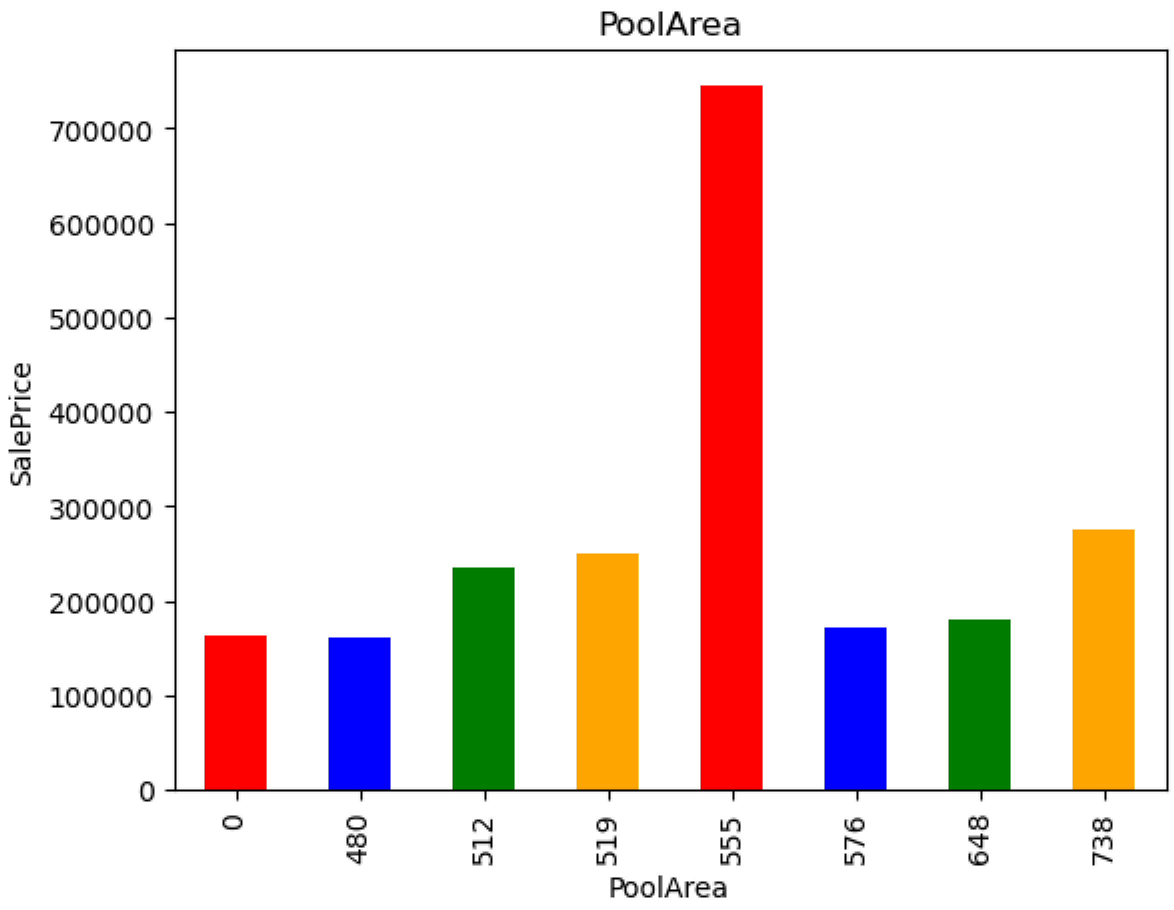

GarageCars
SalePrice
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
GarageCars

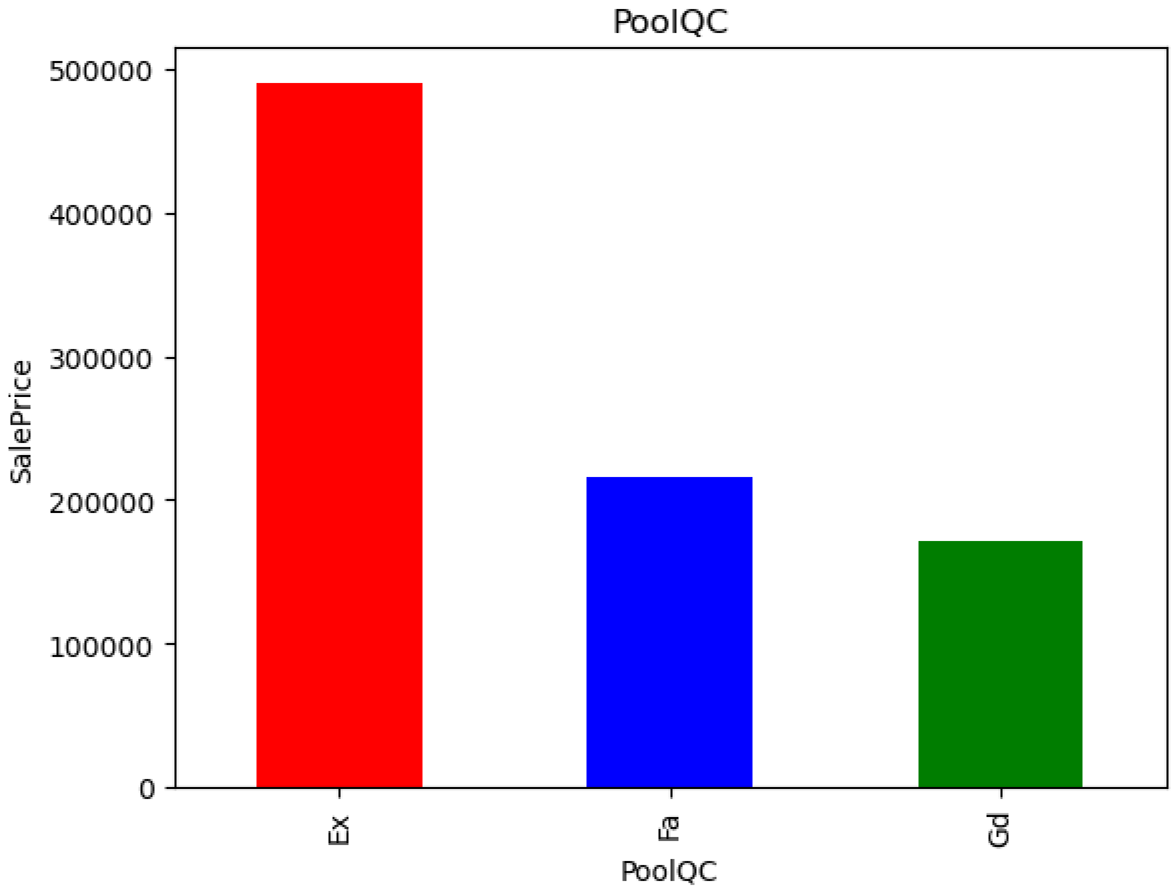

GarageQual
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
Ex
Fa
Gd
Po
TA
GarageQual

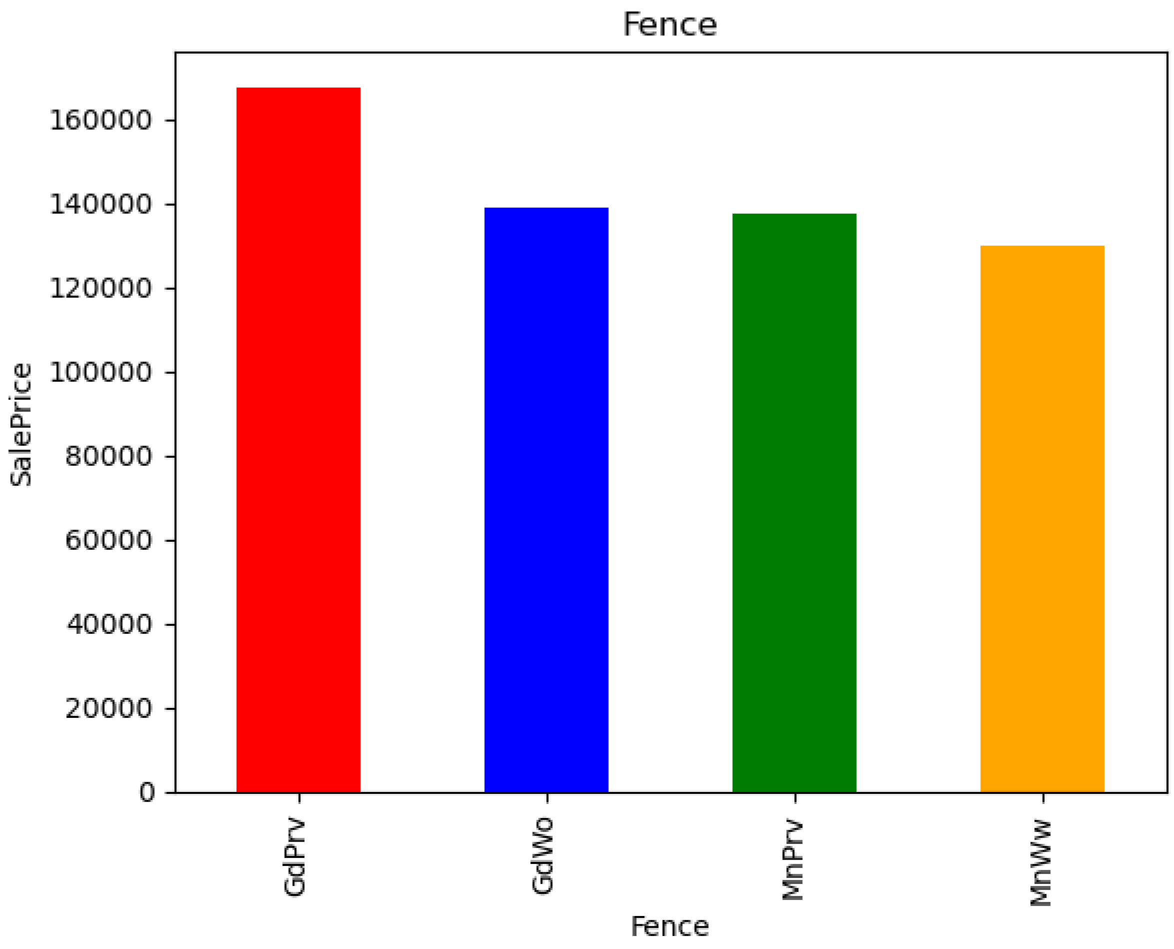

GarageCond
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Ex
Fa
Gd
Po
TA
GarageCond

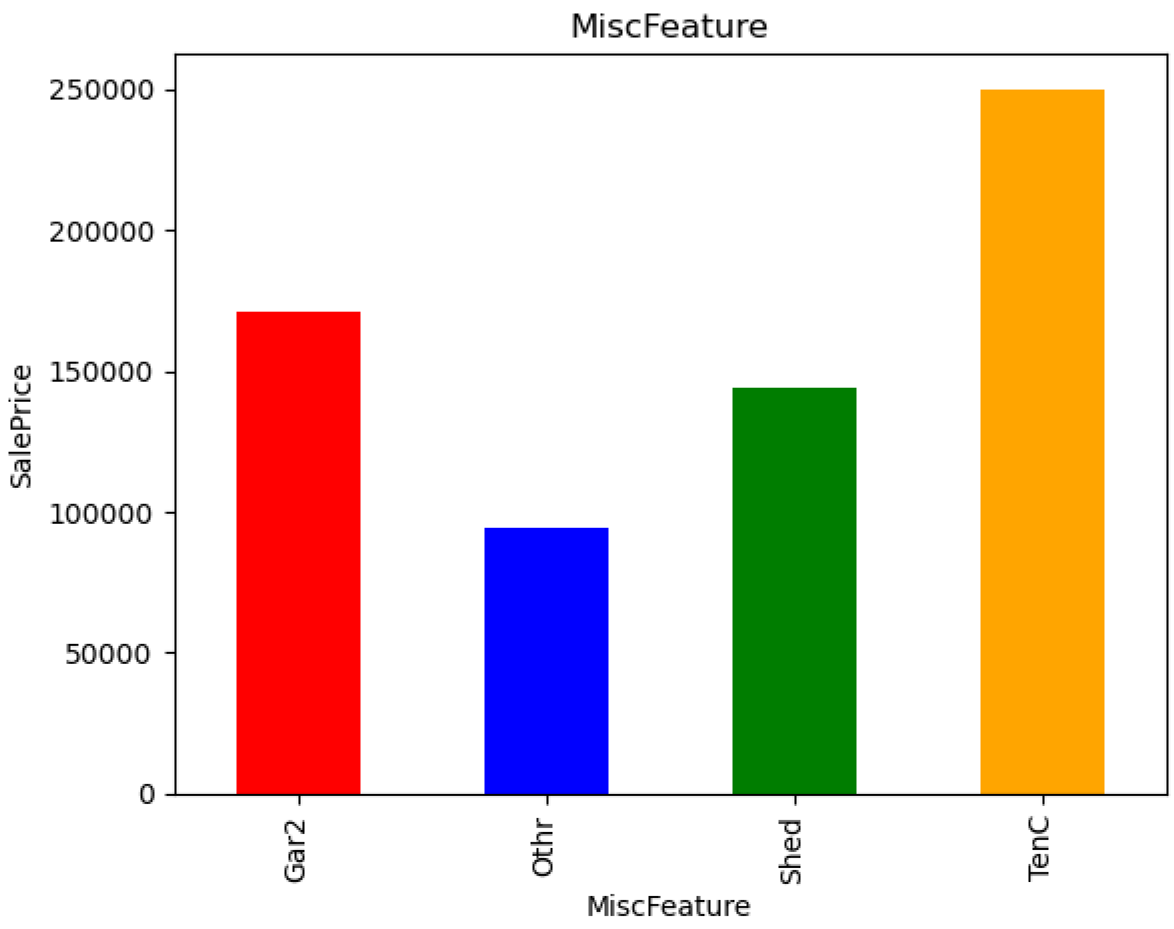

PavedDrive
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
N
P
Y
PavedDrive

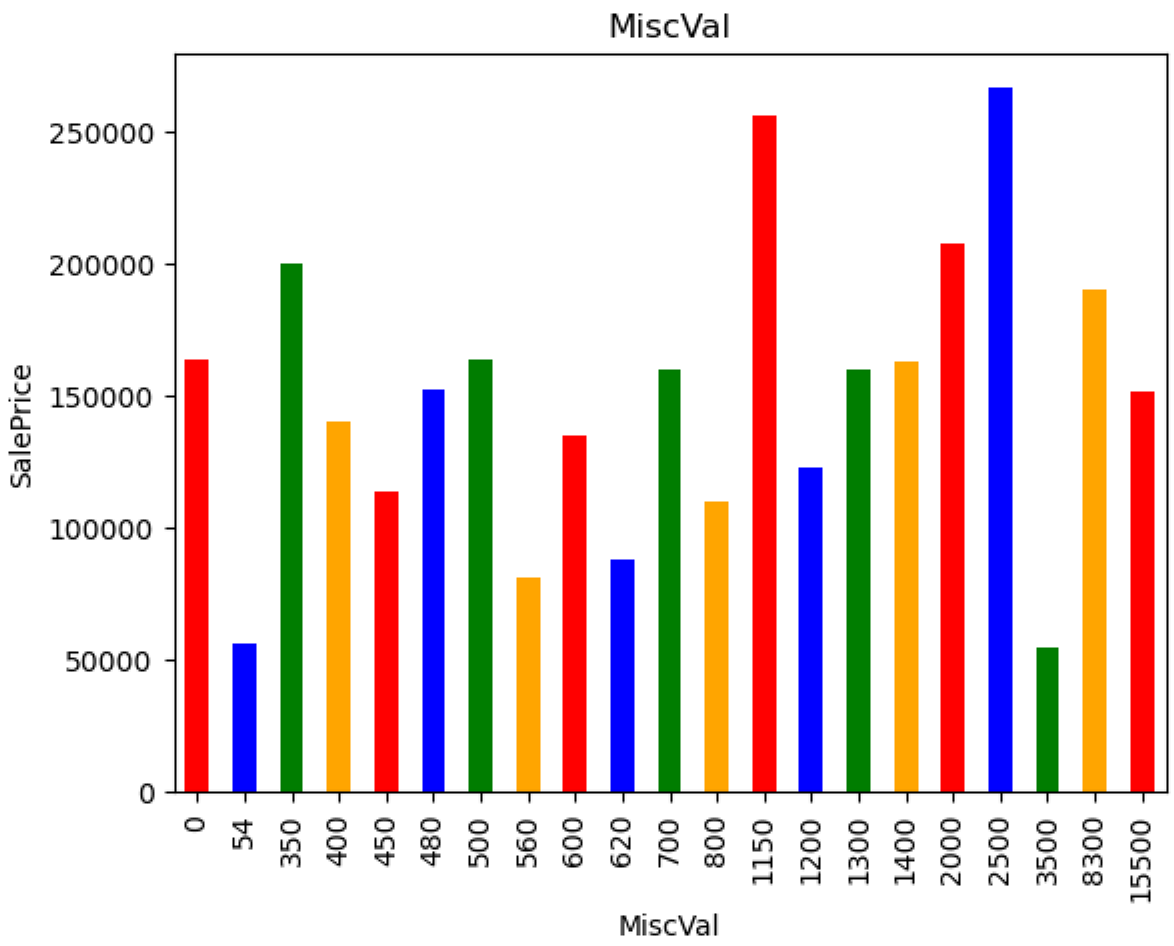

3SsnPorch
SalePrice
0
50000
100000
150000
200000
250000
300000
350000
400000
0
23
96
130
140
144
153
162
168
180
182
196
216
238
245
290
304
320
407
508
3SsnPorch

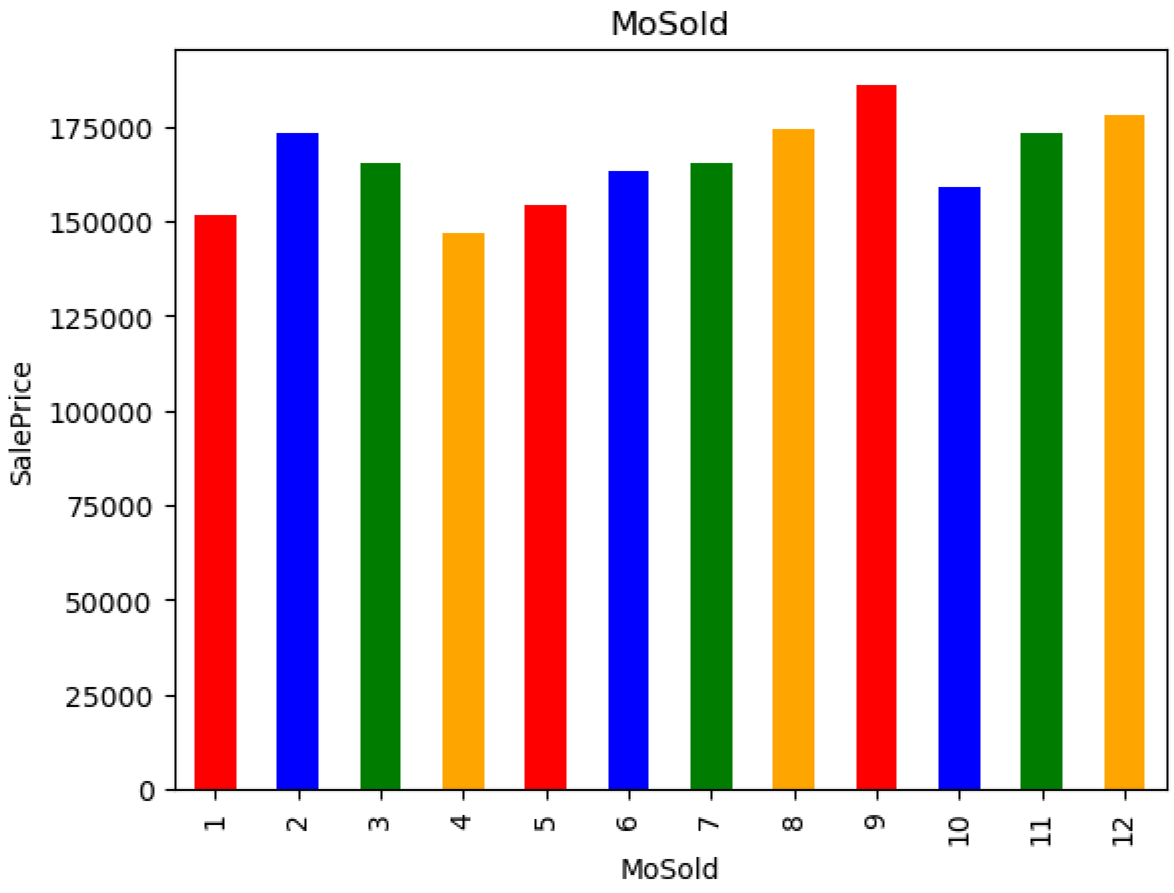

PoolArea
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
0
480
512
519
555
576
648
738
PoolArea

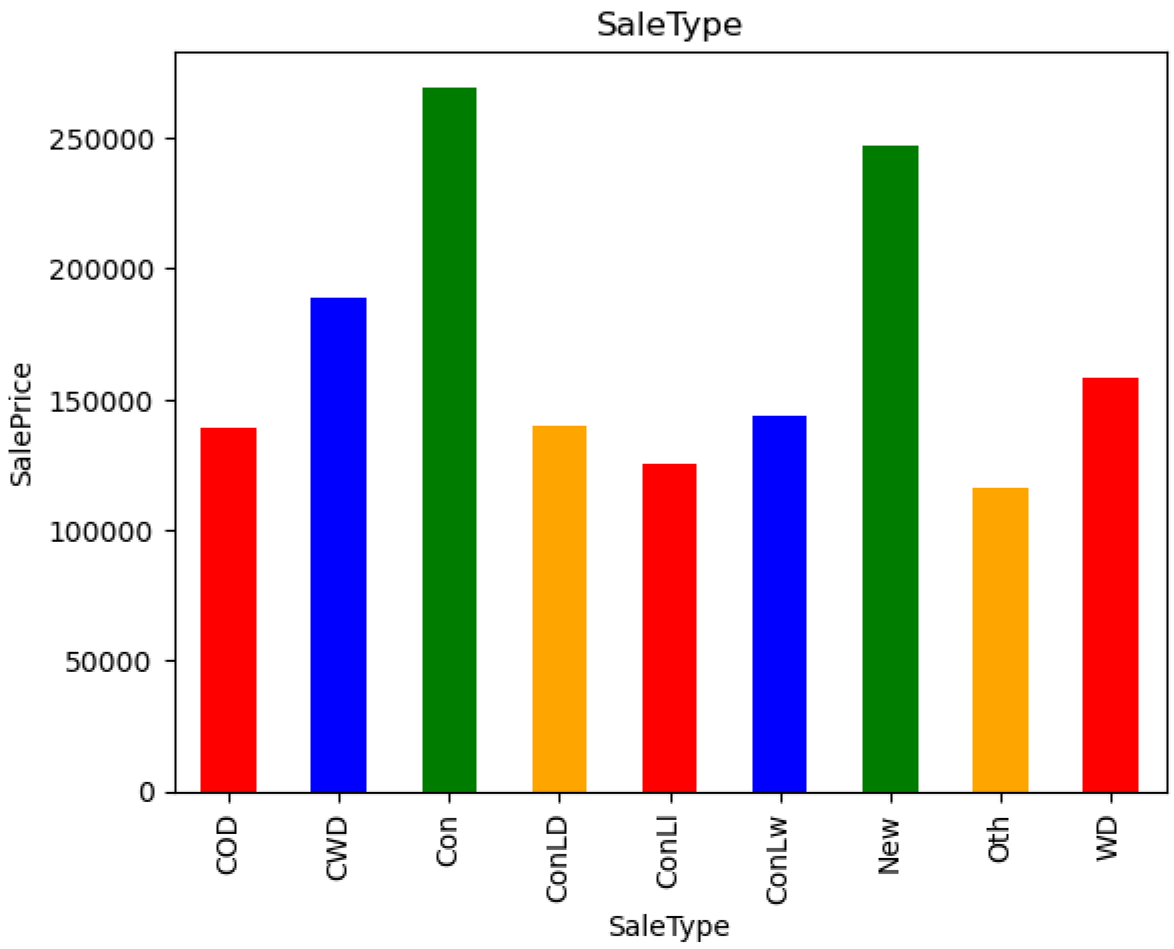

PoolQC
SalePrice
0
100000
200000
300000
400000
500000
Ex
Fa
Gd
PoolQC

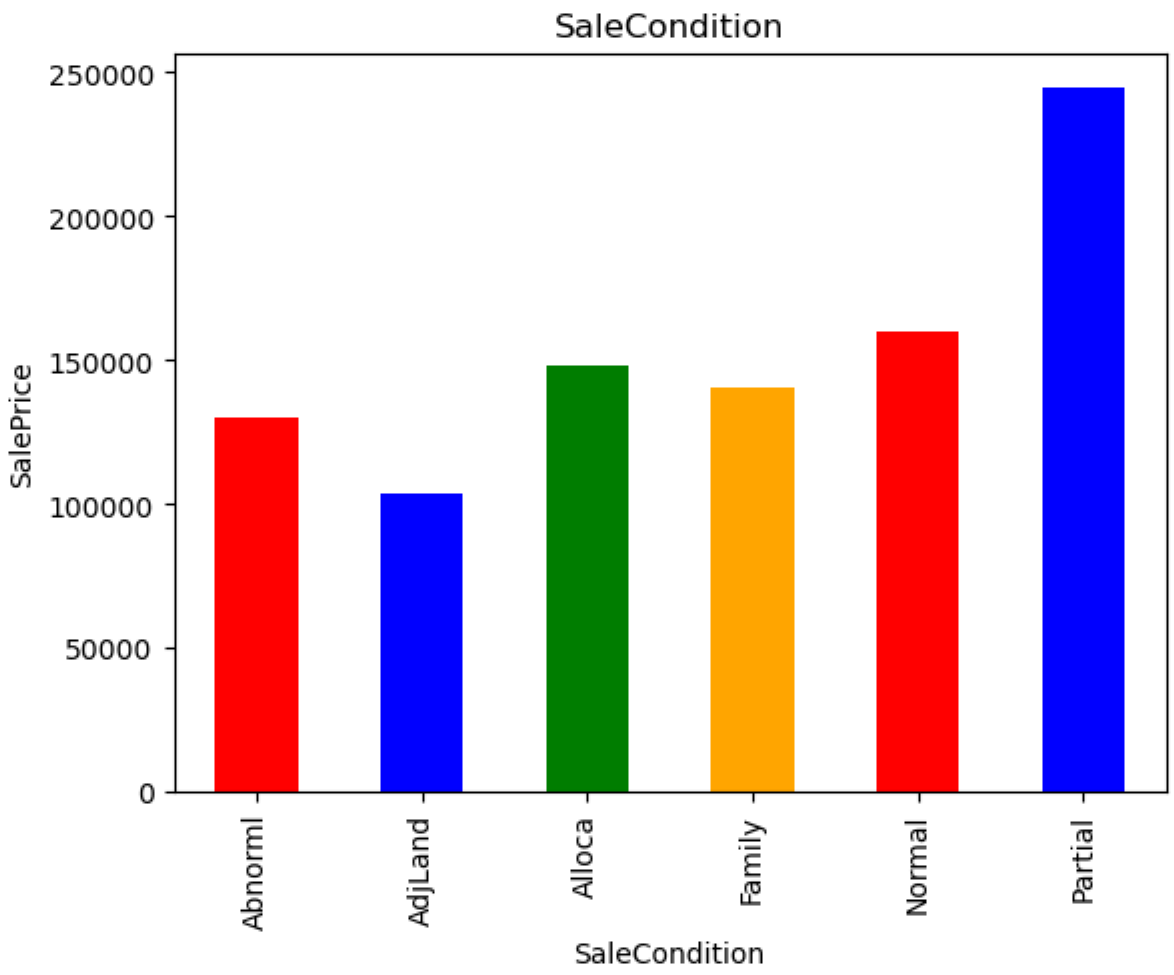

Fence
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
GdPrv
GdWo
MnPrv
MnWw
Fence


MiscFeature
SalePrice
0
50000
100000
150000
200000
250000
Gar2
Othr
Shed
TenC
MiscFeature


MiscVal
SalePrice
0
50000
100000
150000
200000
250000
0
54
350
400
450
480
500
560
600
620
700
800
1150
1200
1300
1400
2000
2500
3500
8300
15500
MiscVal


MoSold
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
1
2
3
4
5
6
7
8
9
10
11
12
MoSold


SaleType
SalePrice
0
50000
100000
150000
200000
250000
COD
CWD
Con
ConLD
ConLI
ConLw
New
Oth
WD
SaleType



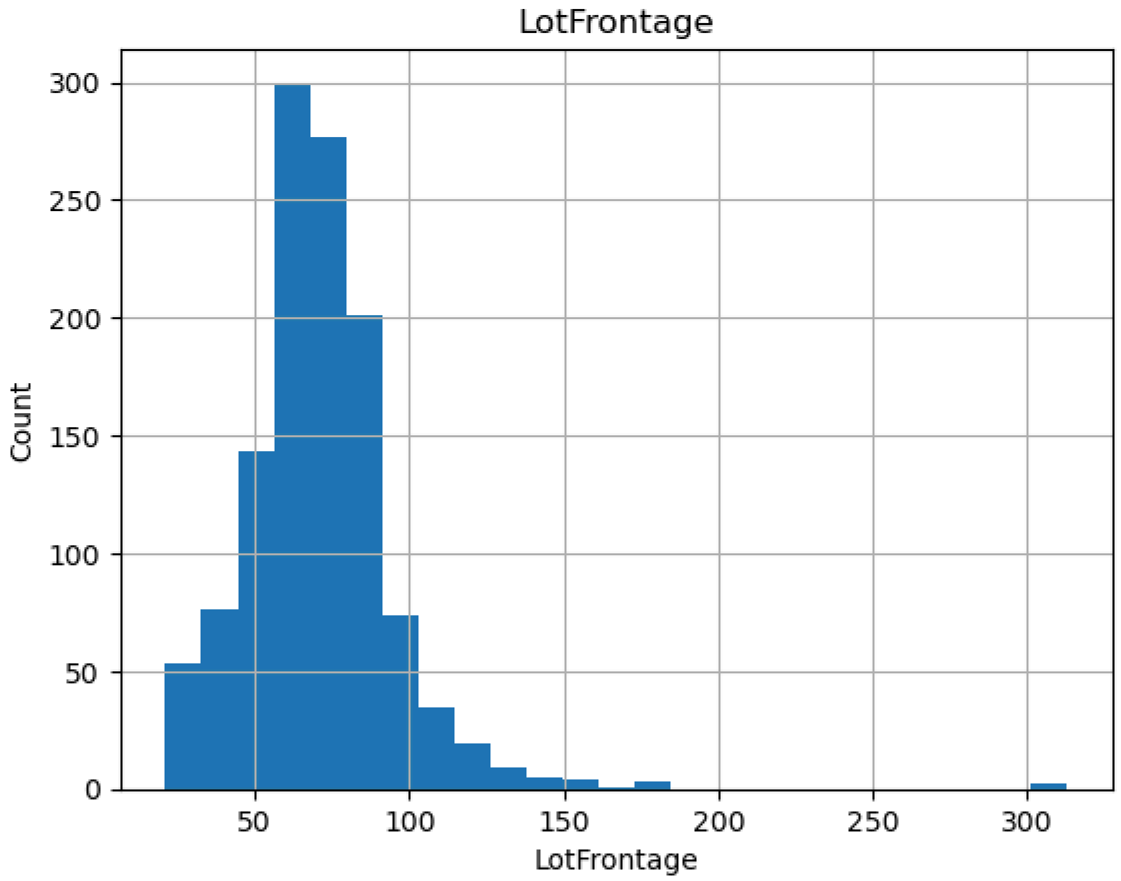
In [20]:
```
## There is a relationship between variable number and Saleprice
```

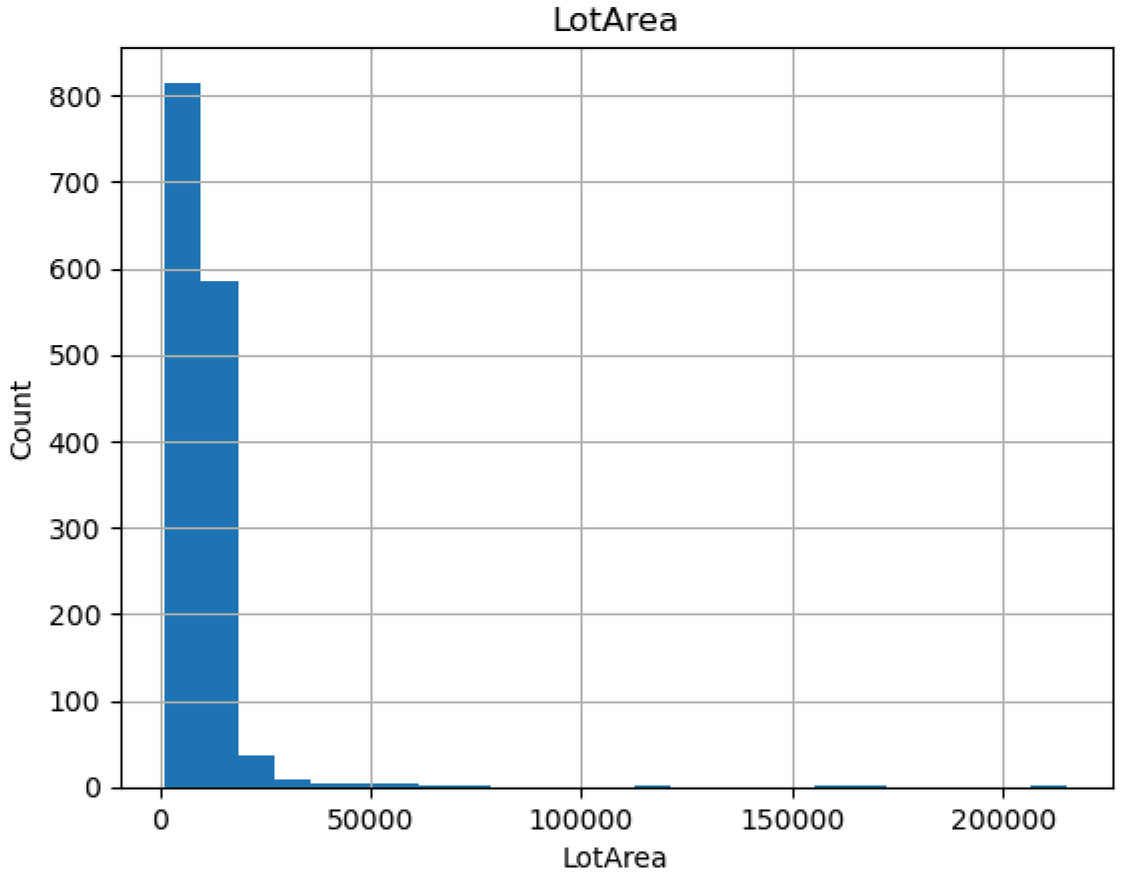
Continous Variable

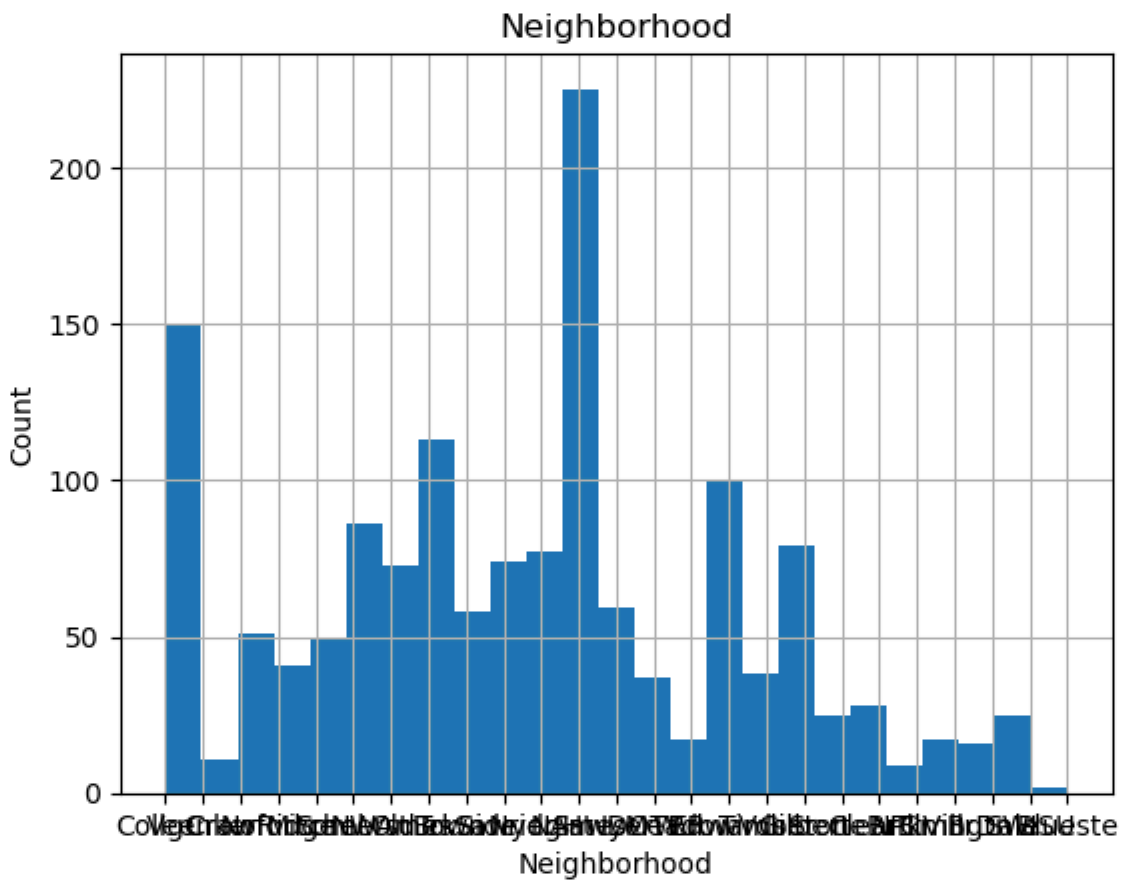
In [21]:
```
continous_feature=[feature for feature in numerical_features if feature not in disc
print("Continous feature count {}".format(len(continous_feature)))
```

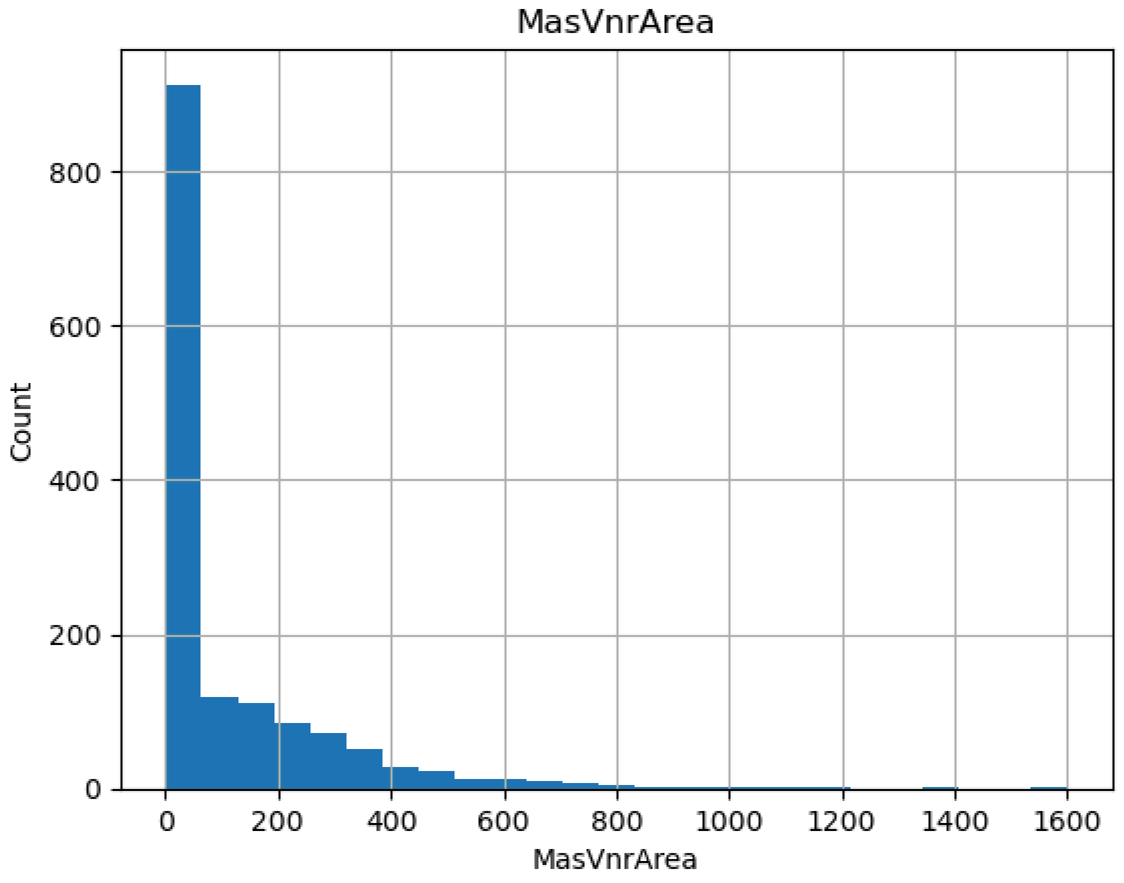
```
Continous feature count 17
```

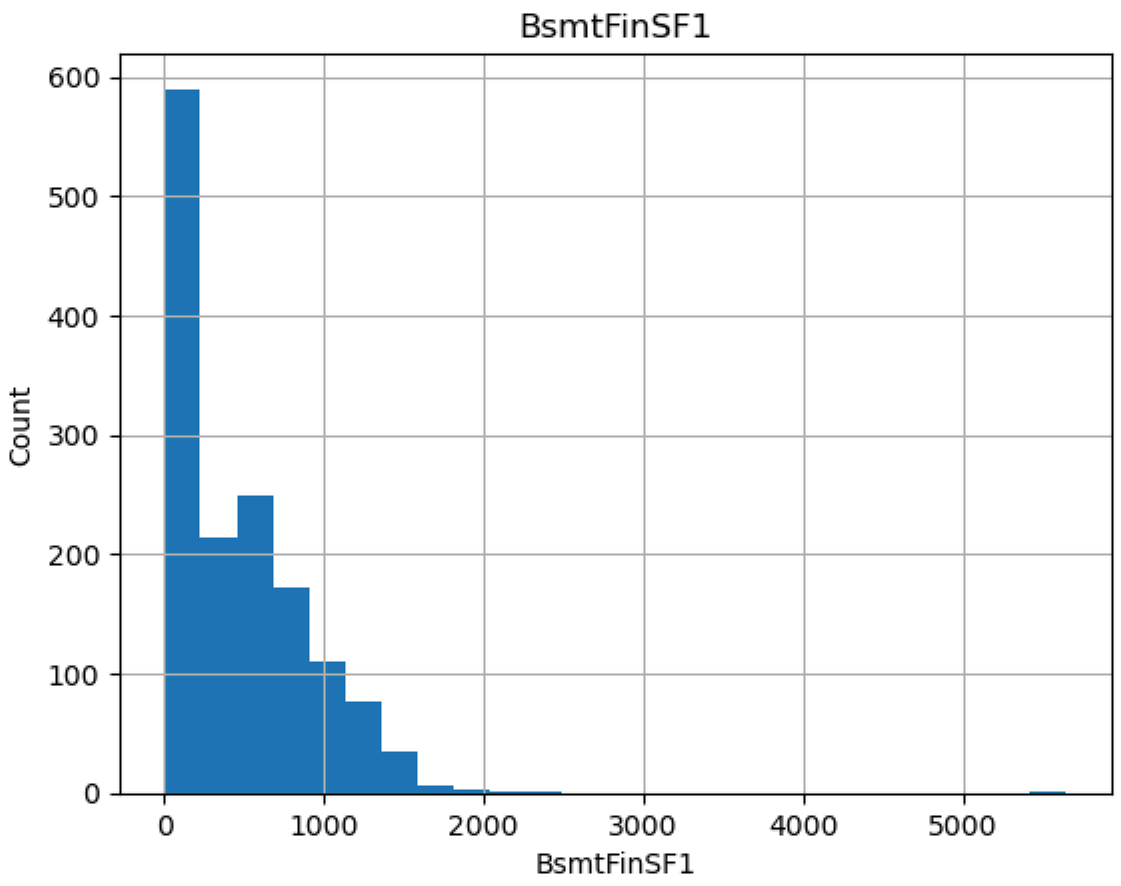
In [22]:
```
#Lets analyze the continous values by creating histograms to understand

for feature in continous_feature:
    data=ds.copy()
    data[feature].hist(bins=25)
    plt.xlabel(feature)
    plt.ylabel("Count")
    plt.title(feature)
    plt.show()
```

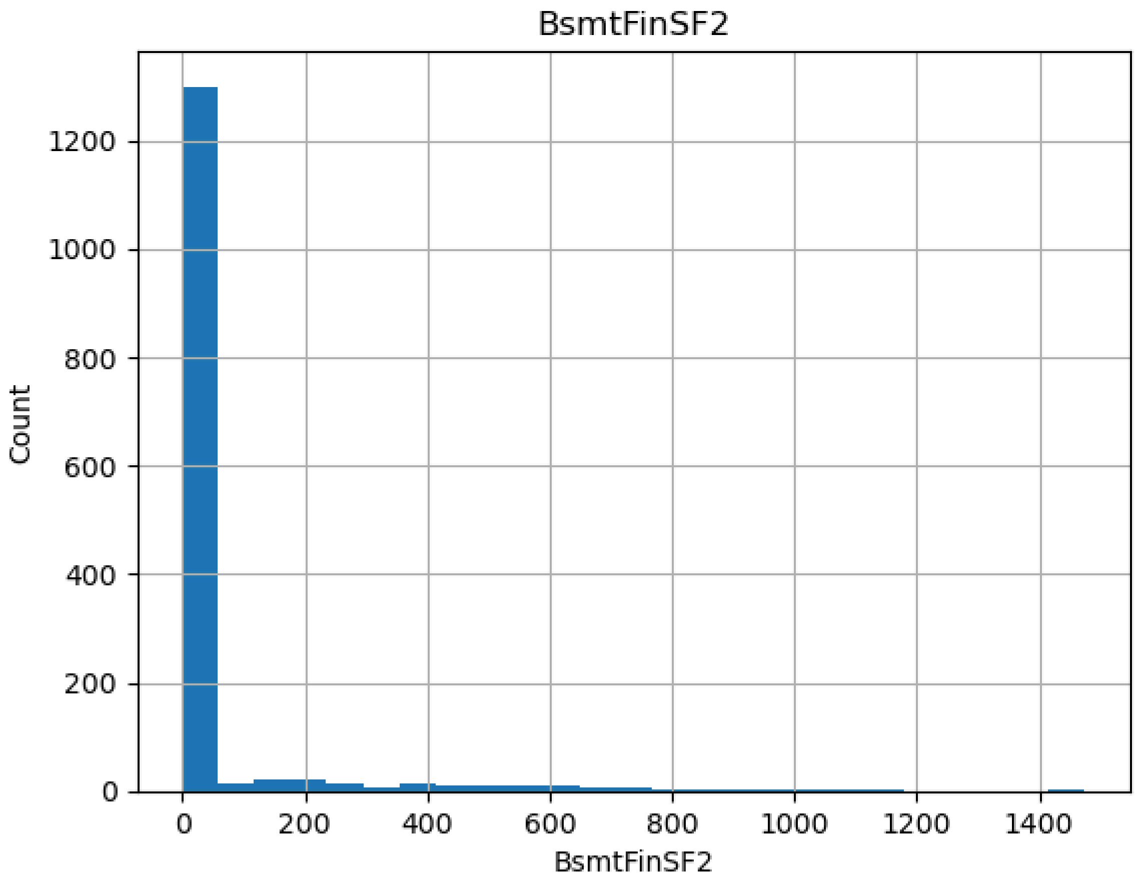

LotFrontage
Count
0
50
100
150
200
250
300
50
100
150
200
250
300
LotFrontage

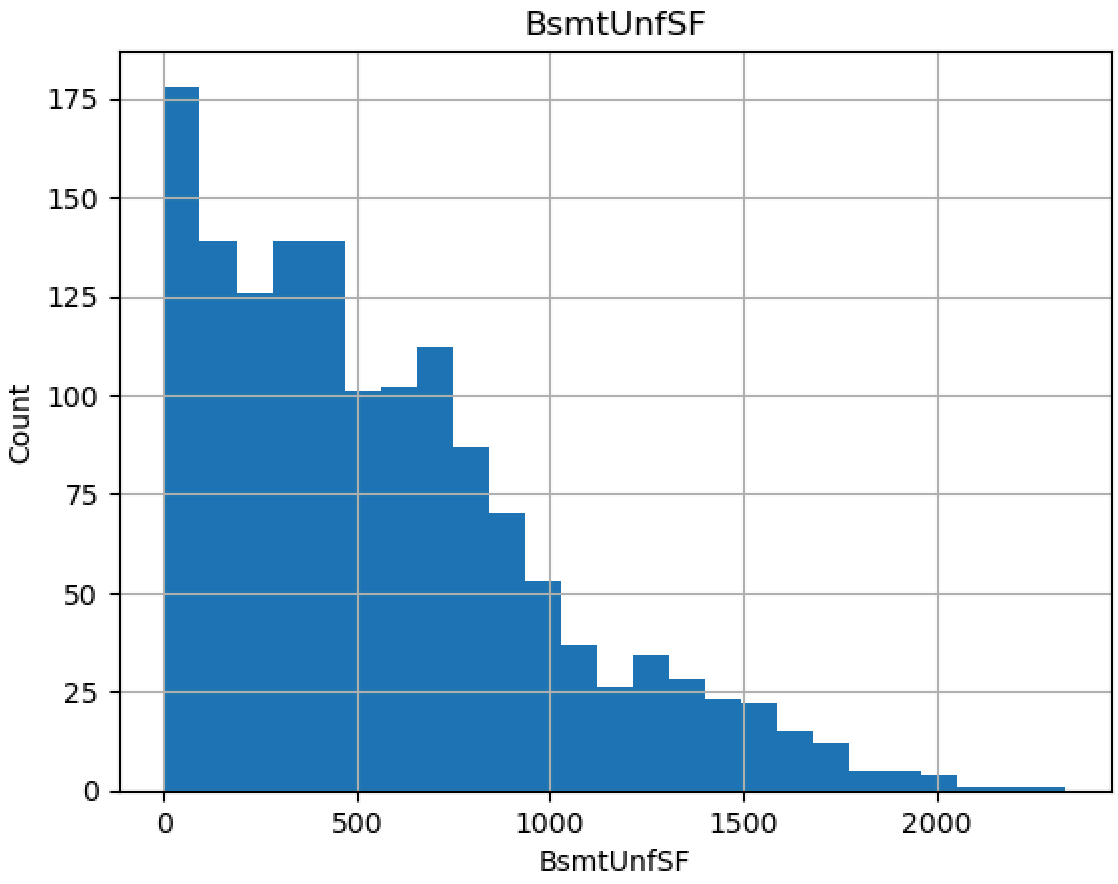

LotArea
Count
0
100
200
300
400
500
600
700
800
0
50000
100000
150000
200000
LotArea

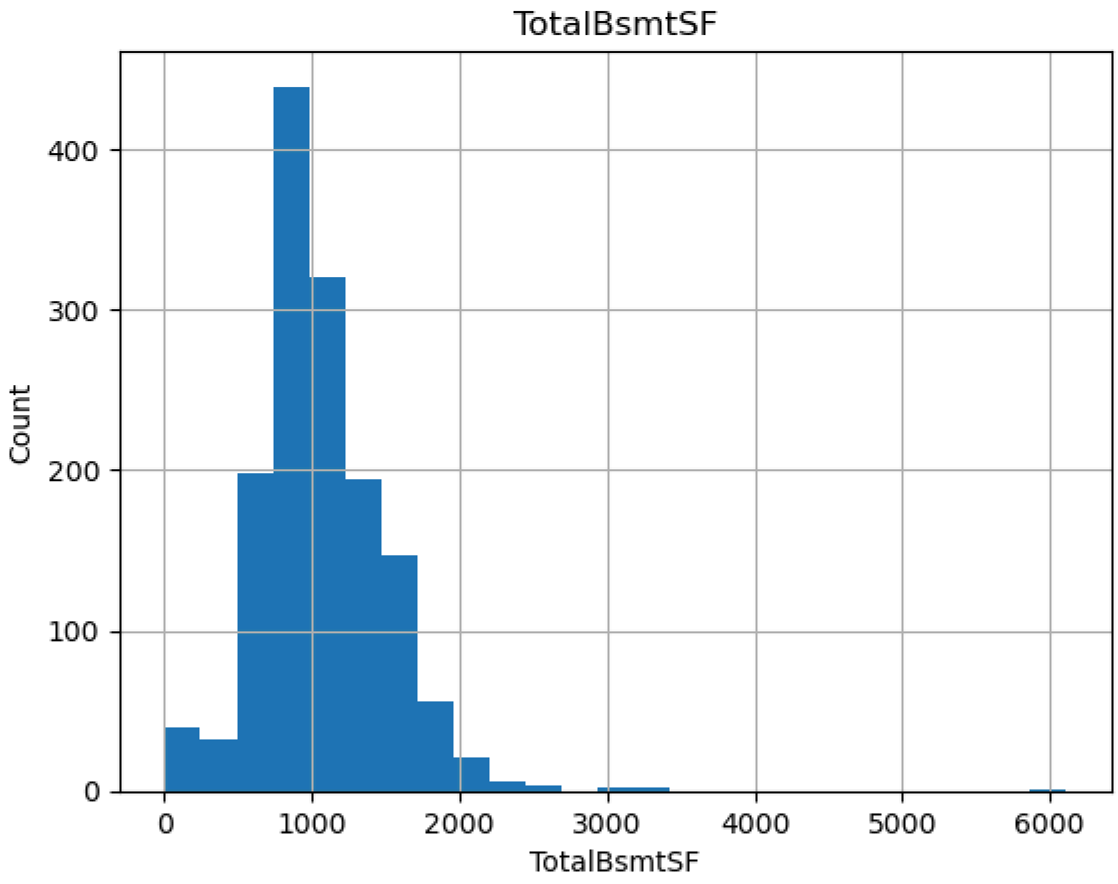

Neighborhood
Count
Neighborhood
0
50
100
150
200

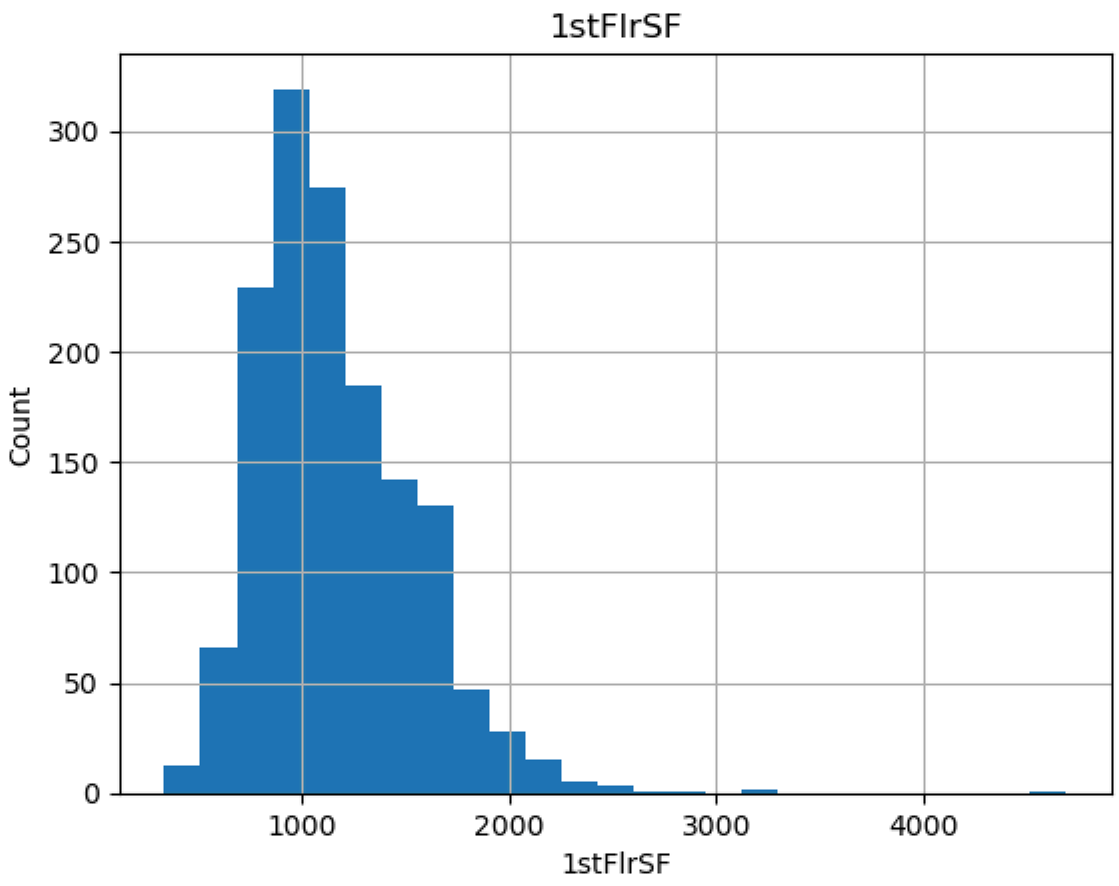

MasVnrArea
Count
MasVnrArea
0
200
400
600
800
1000
1200
1400
1600


BsmtFinSF1
Count
600
500
400
300
200
100
0
0
1000
2000
3000
4000
5000
BsmtFinSF1

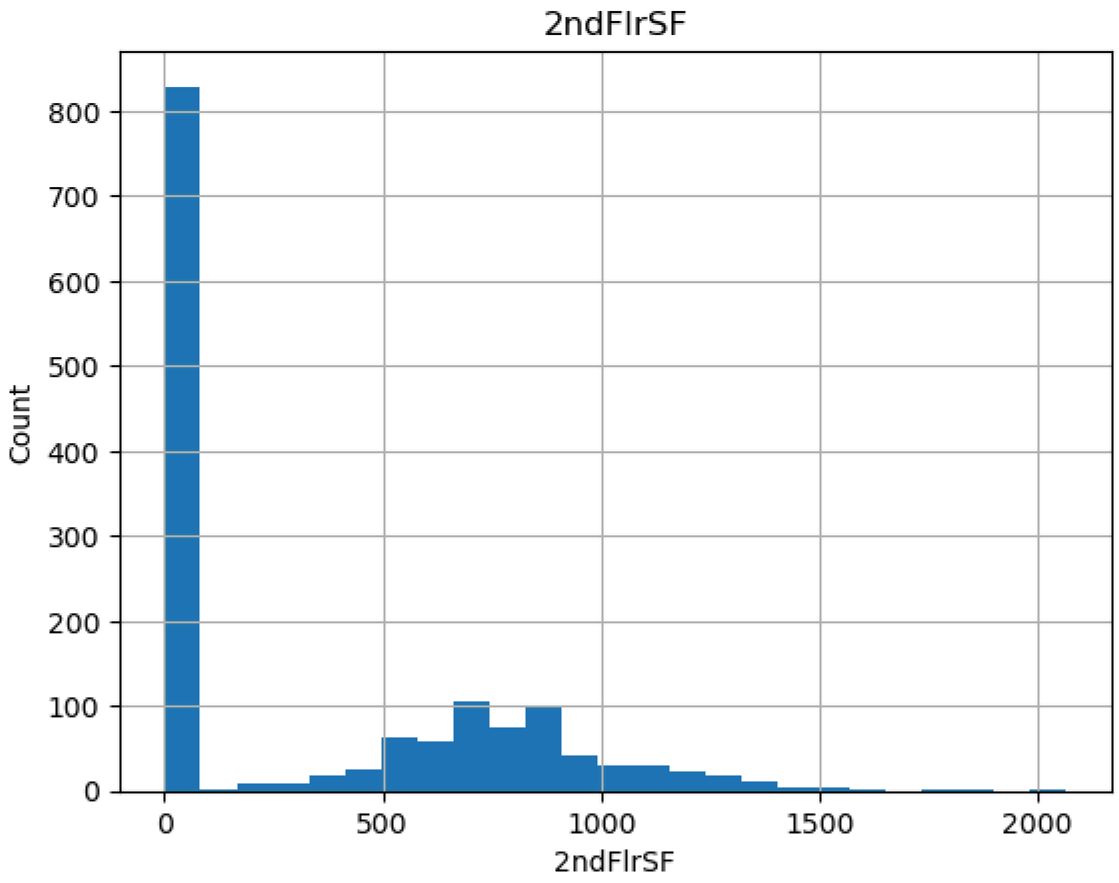

BsmtFinSF2
Count
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
BsmtFinSF2


BsmtUnfSF
Count
175
150
125
100
75
50
25
0
0
500
1000
1500
2000
BsmtUnfSF

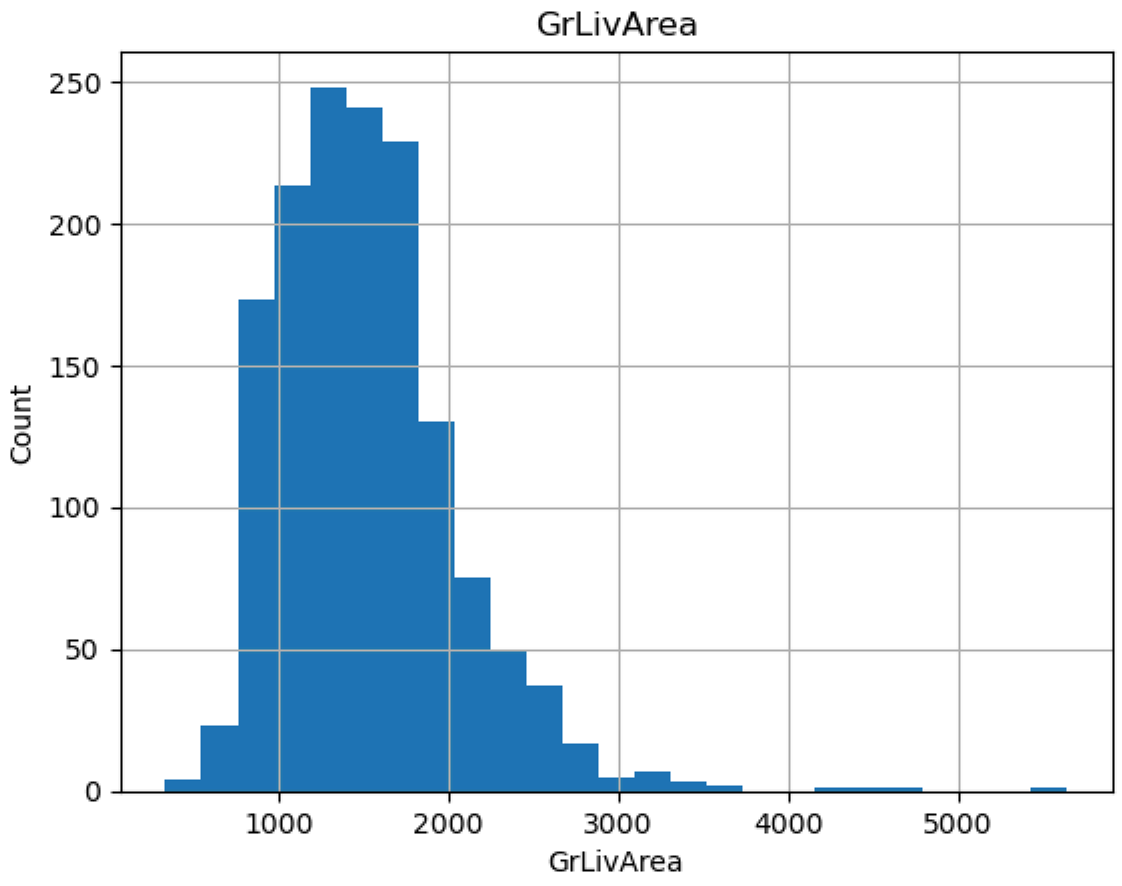

TotalBsmtSF
Count
400
300
200
100
0
0
1000
2000
3000
4000
5000
6000
TotalBsmtSF

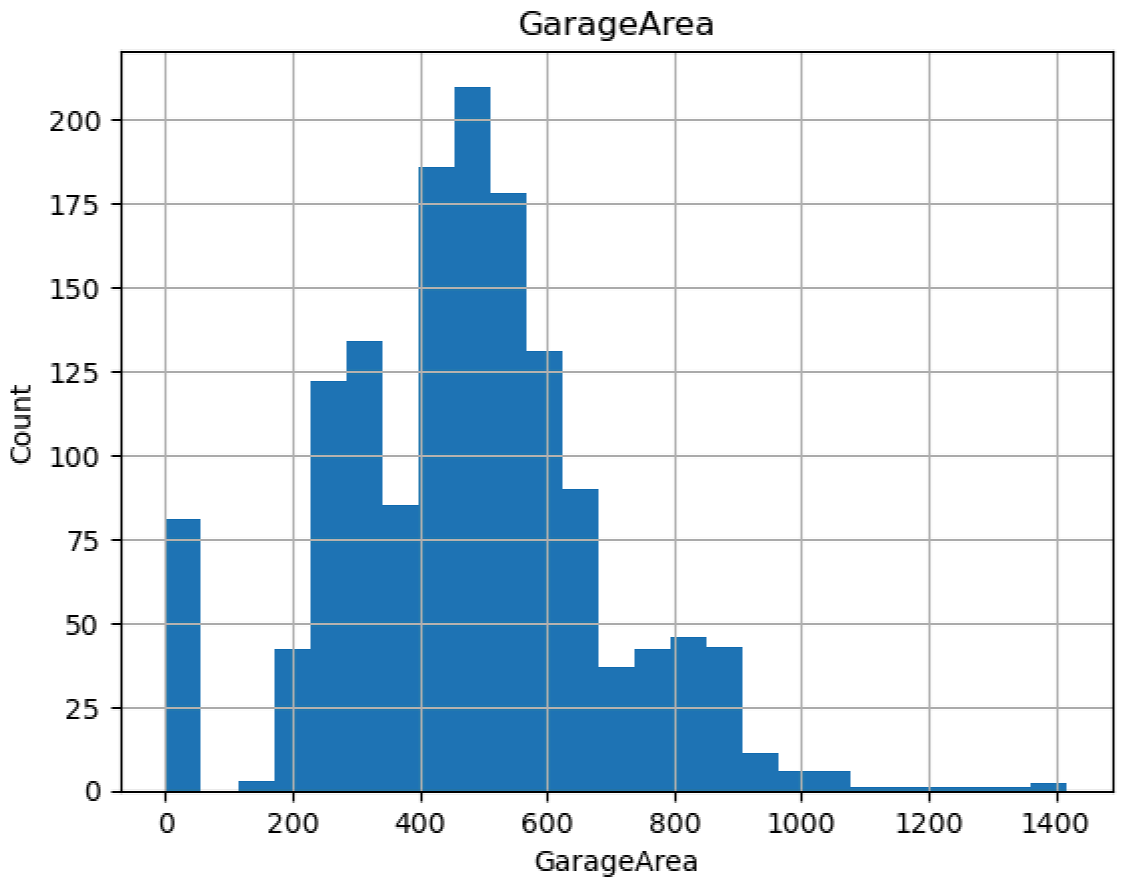


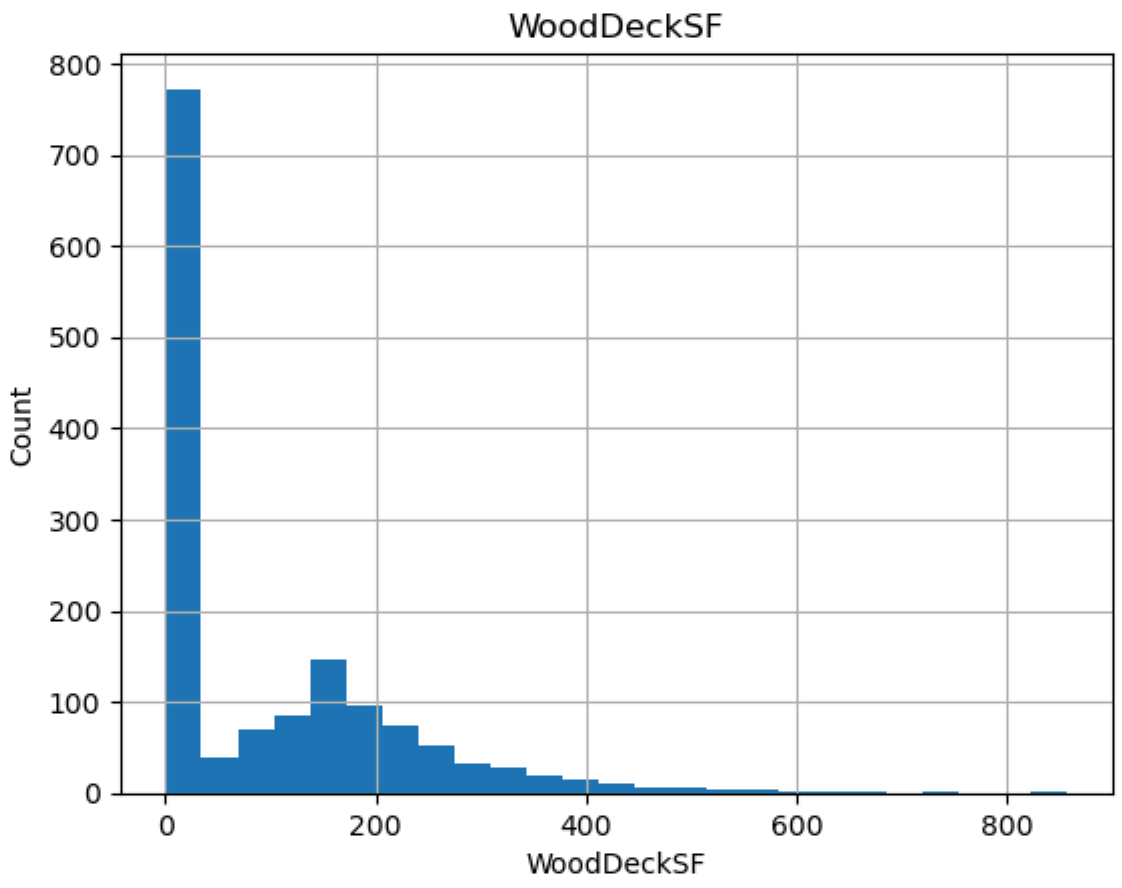
1stFlrSF
Count
0
50
100
150
200
250
300
1000
2000
3000
4000
1stFlrSF

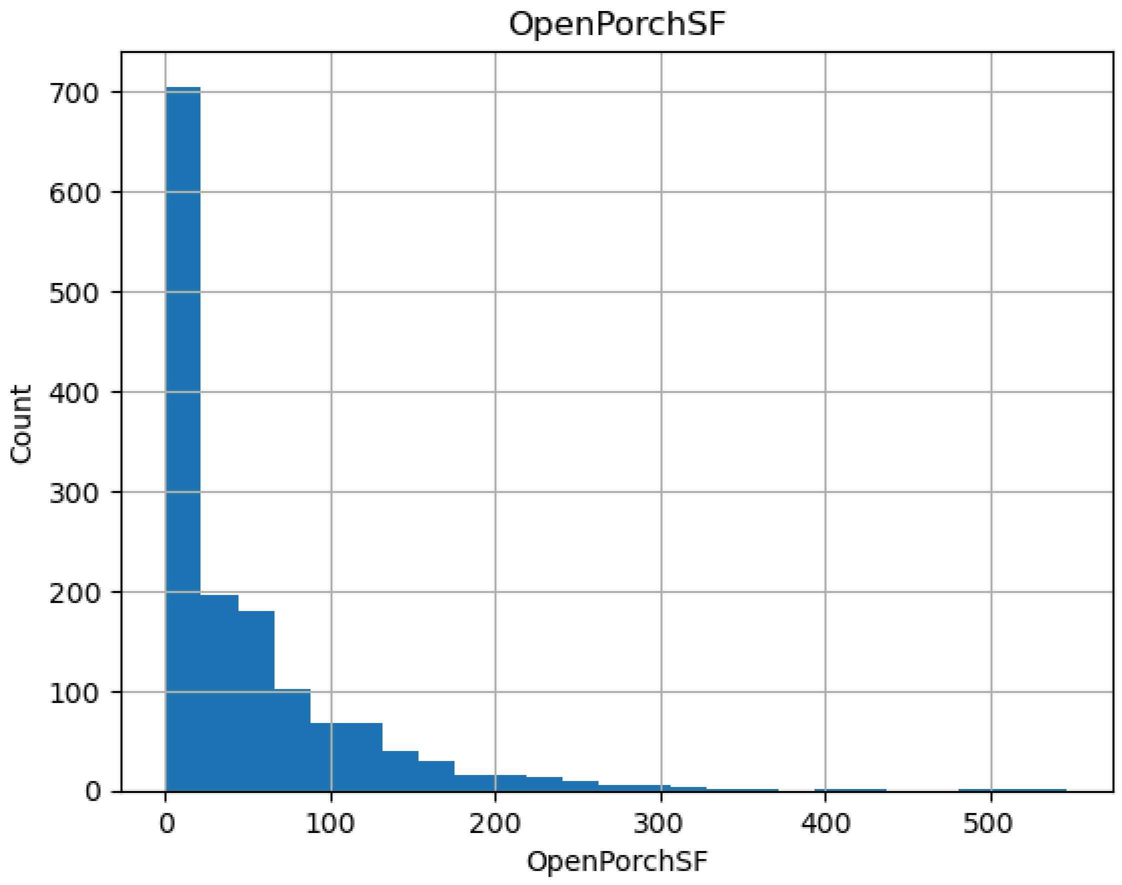


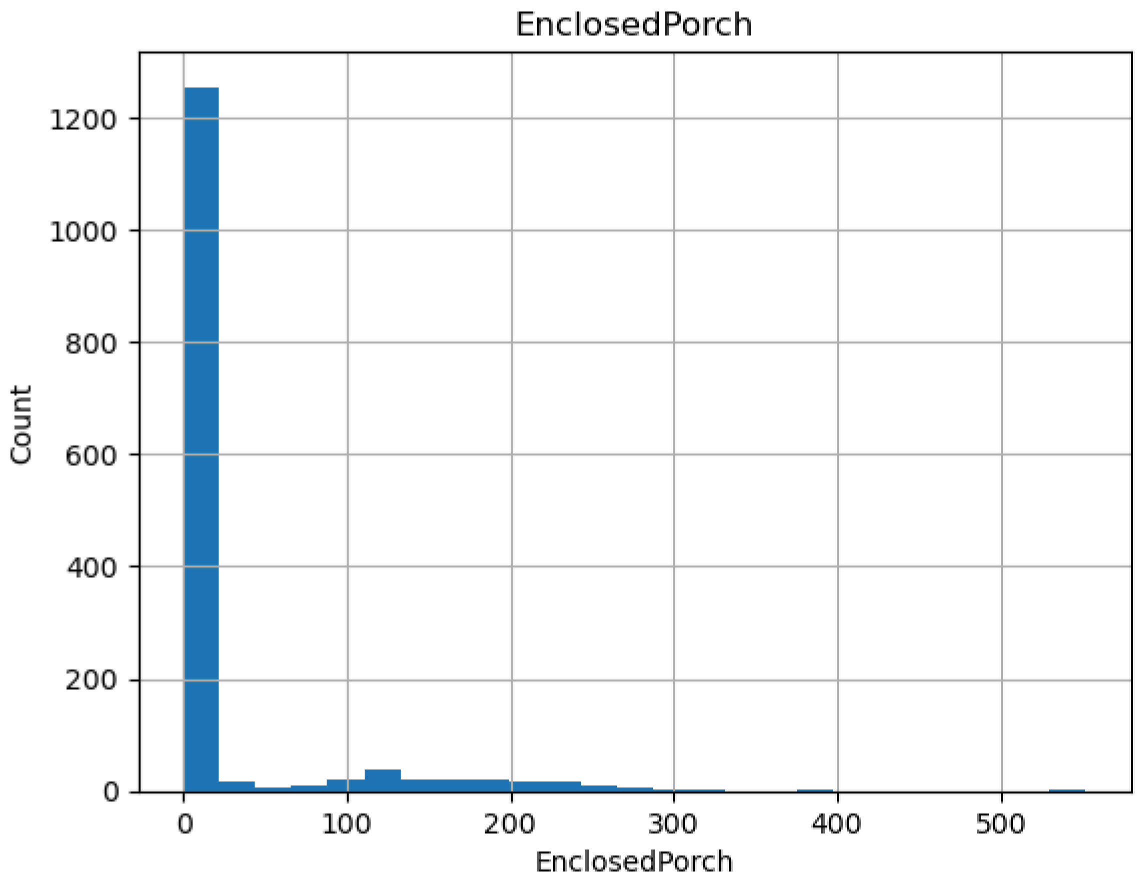
2ndFlrSF
Count
0
100
200
300
400
500
600
700
800
0
500
1000
1500
2000
2ndFlrSF

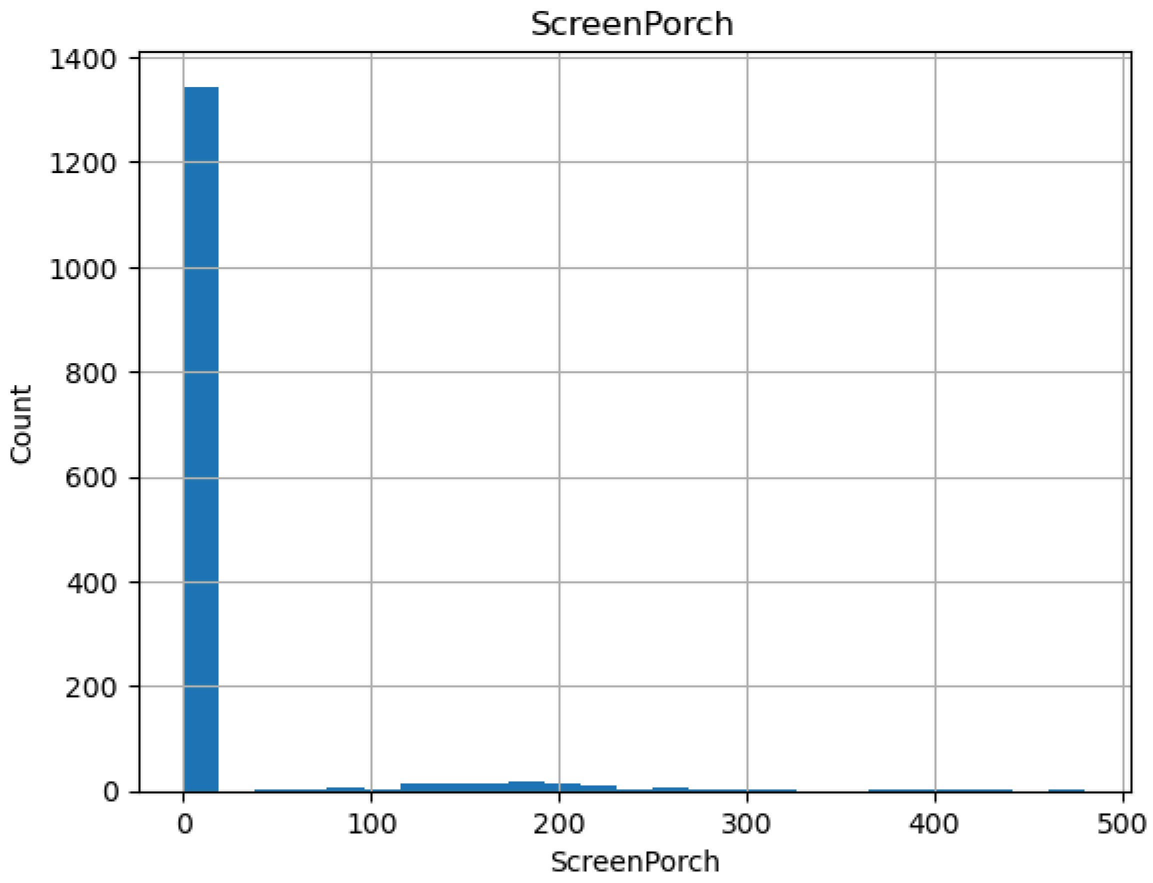

GrLivArea
Count
250
200
150
100
50
0
1000
2000
3000
4000
5000
GrLivArea

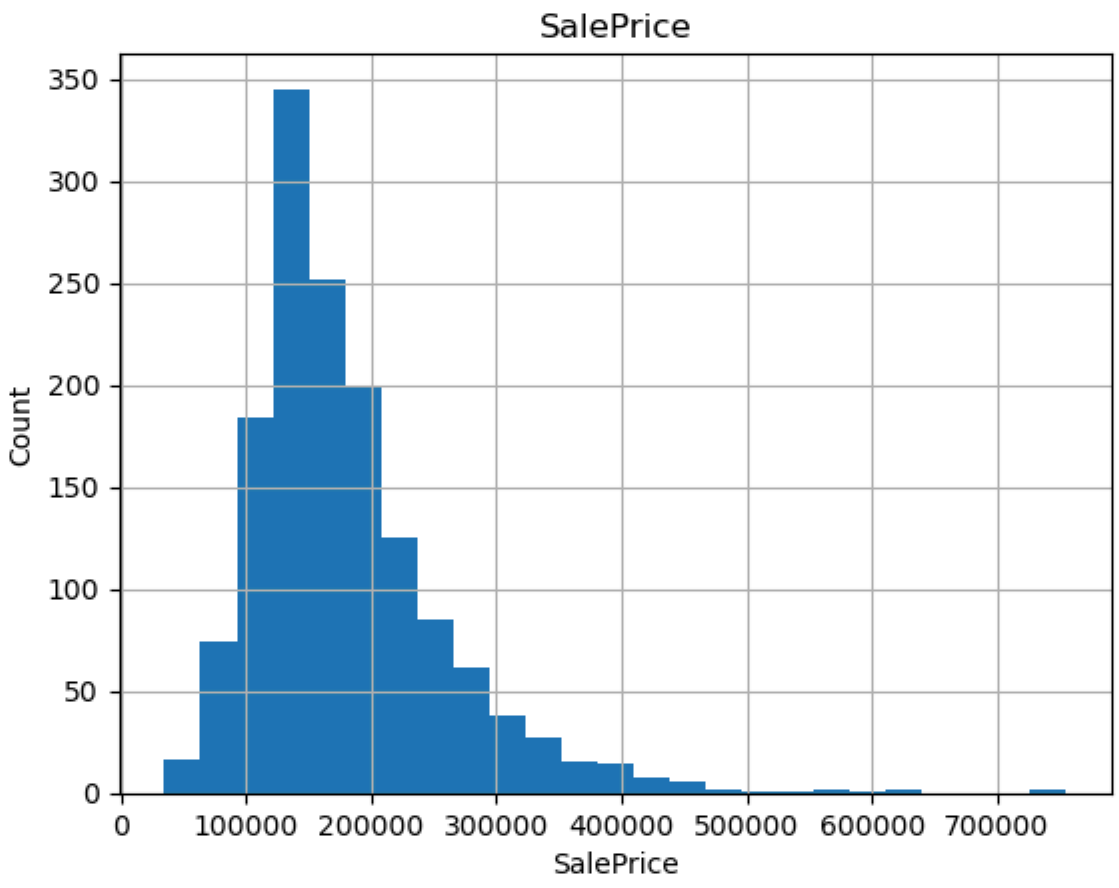

GarageArea
Count
200
175
150
125
100
75
50
25
0
0
200
400
600
800
1000
1200
1400
GarageArea


WoodDeckSF
Count
800
700
600
500
400
300
200
100
0
0
200
400
600
800
WoodDeckSF


OpenPorchSF
Count
700
600
500
400
300
200
100
0
0
100
200
300
400
500
OpenPorchSF

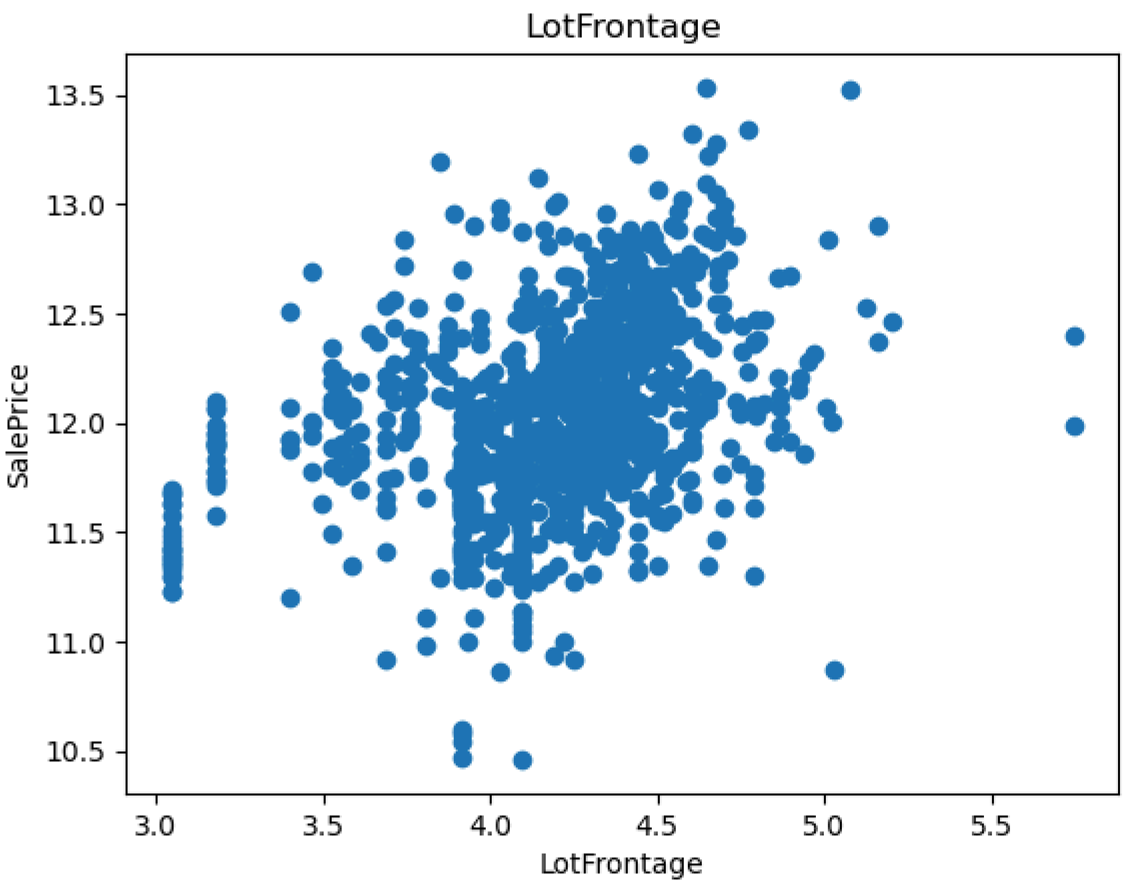

EnclosedPorch
Count
1200
1000
800
600
400
200
0
0
100
200
300
400
500
EnclosedPorch

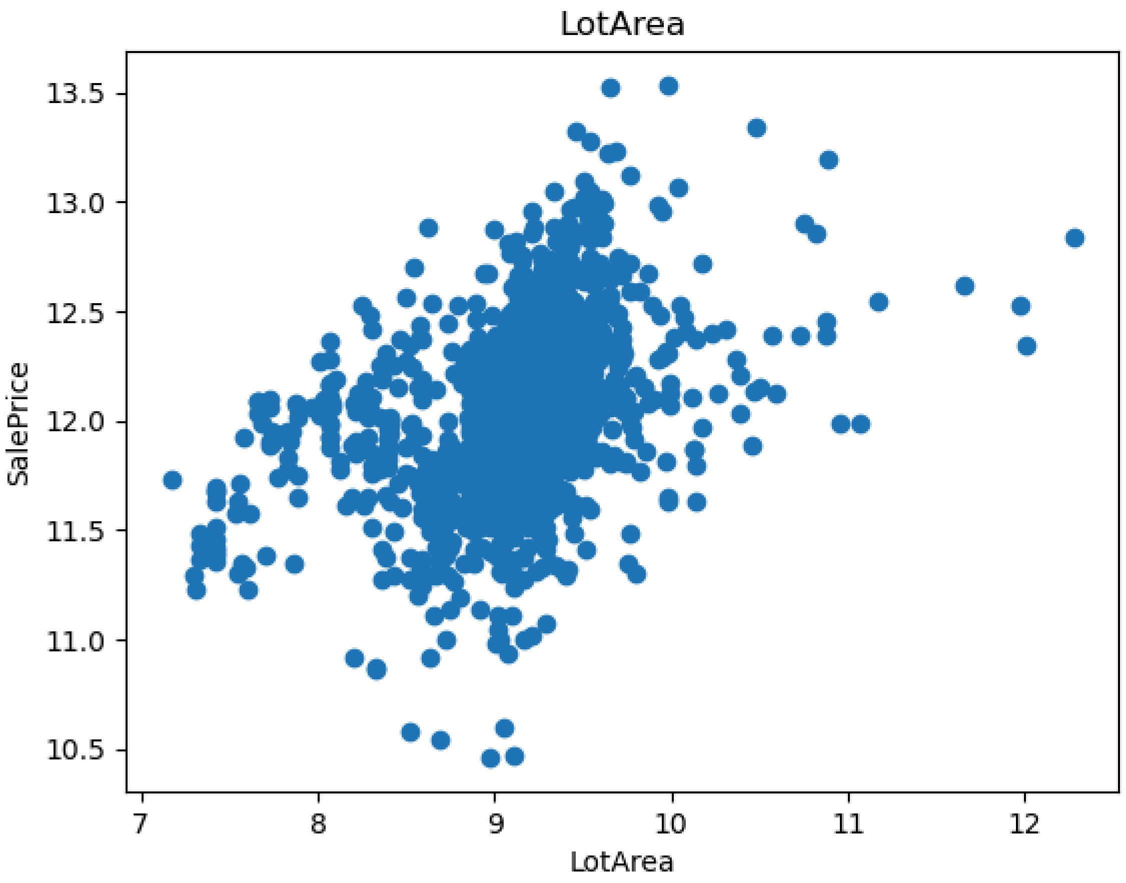

ScreenPorch
Count
1400
1200
1000
800
600
400
200
0
0
100
200
300
400
500
ScreenPorch

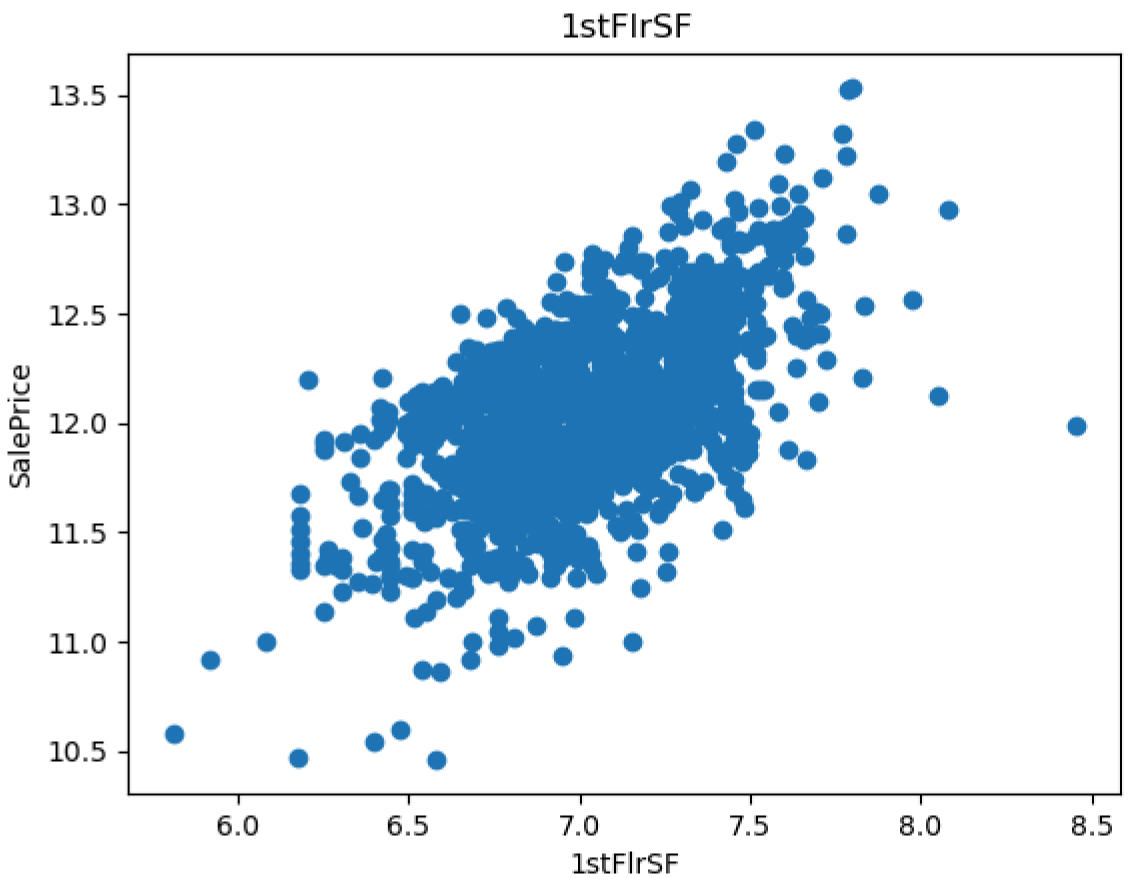


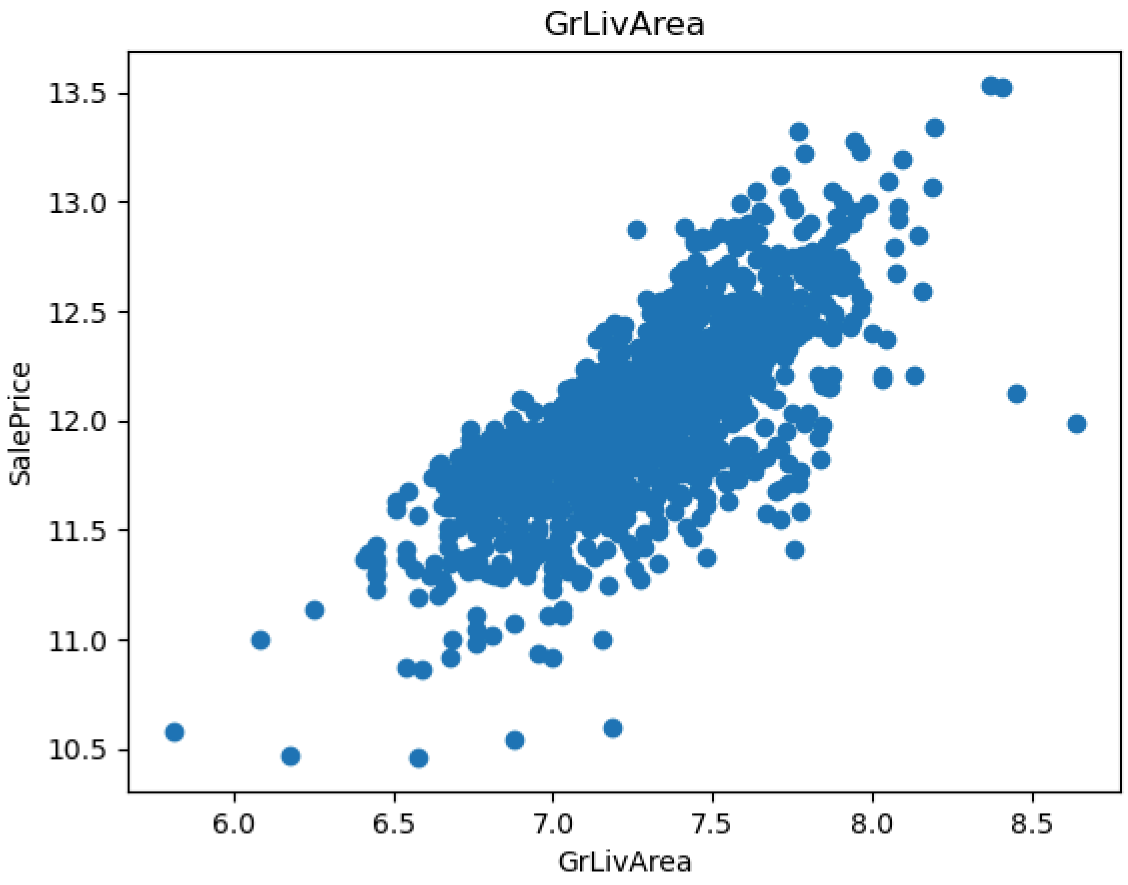
Exploratory Data Analiysis

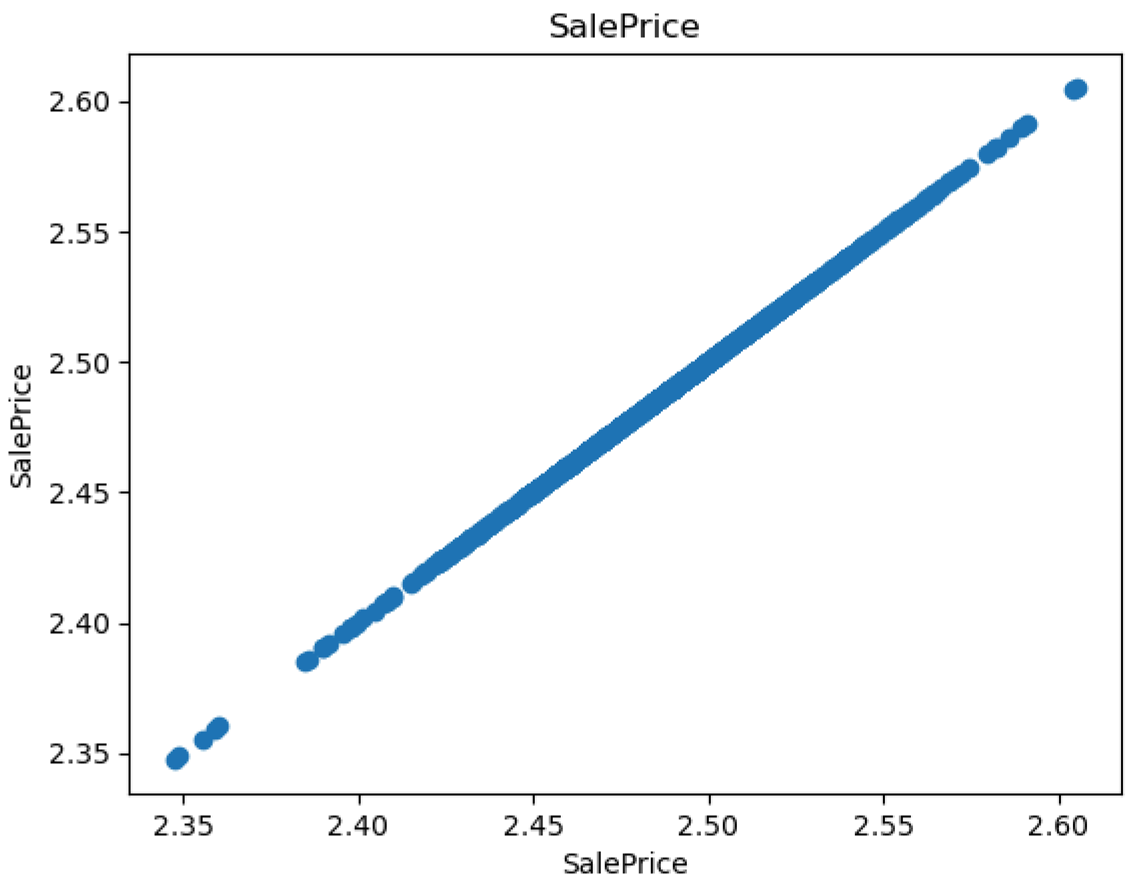
In [23]:
```
## we will be using logarithmic transformation

for feature in continous_feature:
    data=ds.copy()

    # skip if column is not numeric
    if data[feature].dtype == 'object':
        continue

    # skip if column has 0 values
    if 0 in data[feature].values:
        continue


    else:
        data[feature]=np.log(data[feature])
        data['SalePrice']=np.log(data['SalePrice'])
        plt.scatter(data[feature],data['SalePrice'])
        plt.xlabel(feature)
        plt.ylabel('SalePrice')
        plt.title(feature)
        plt.show()
```


LotFrontage
SalePrice
LotFrontage
13.5
13.0
12.5
12.0
11.5
11.0
10.5
3.0
3.5
4.0
4.5
5.0
5.5


LotArea
SalePrice
LotArea
13.5
13.0
12.5
12.0
11.5
11.0
10.5
7
8
9
10
11
12

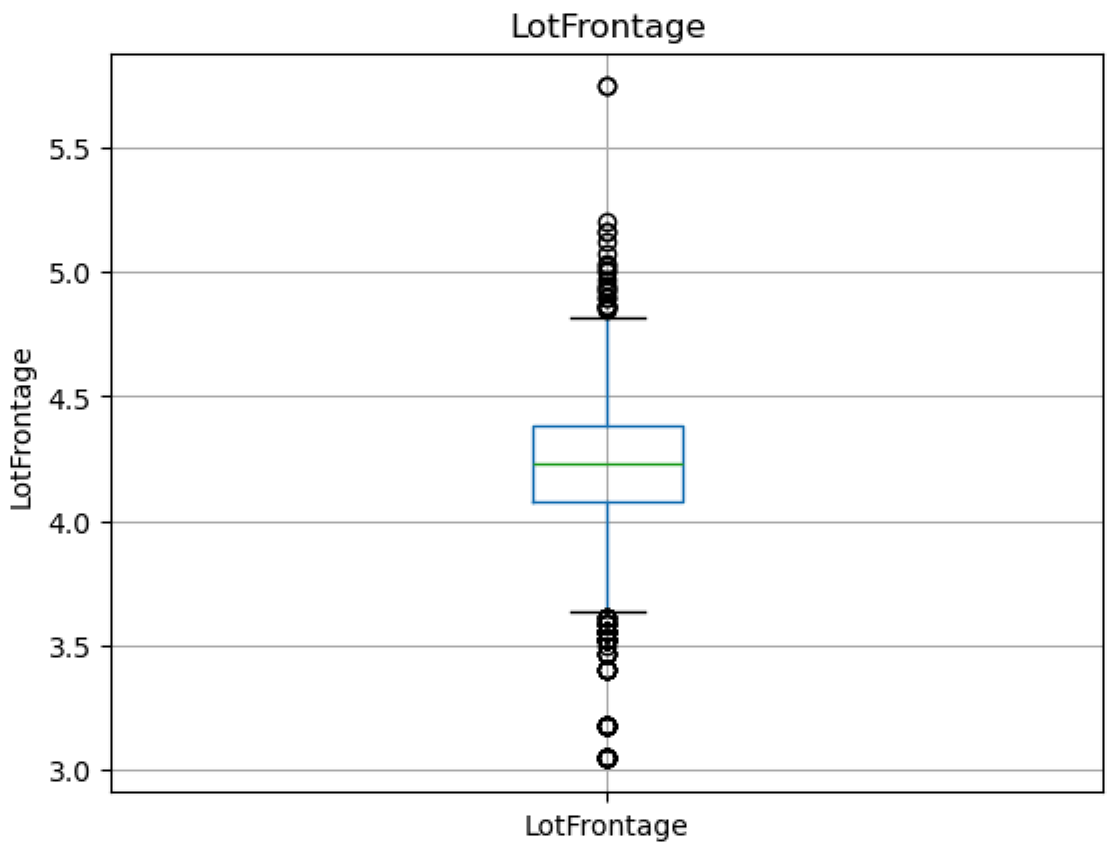

1stFlrSF
SalePrice
1stFlrSF
13.5
13.0
12.5
12.0
11.5
11.0
10.5
6.0
6.5
7.0
7.5
8.0
8.5

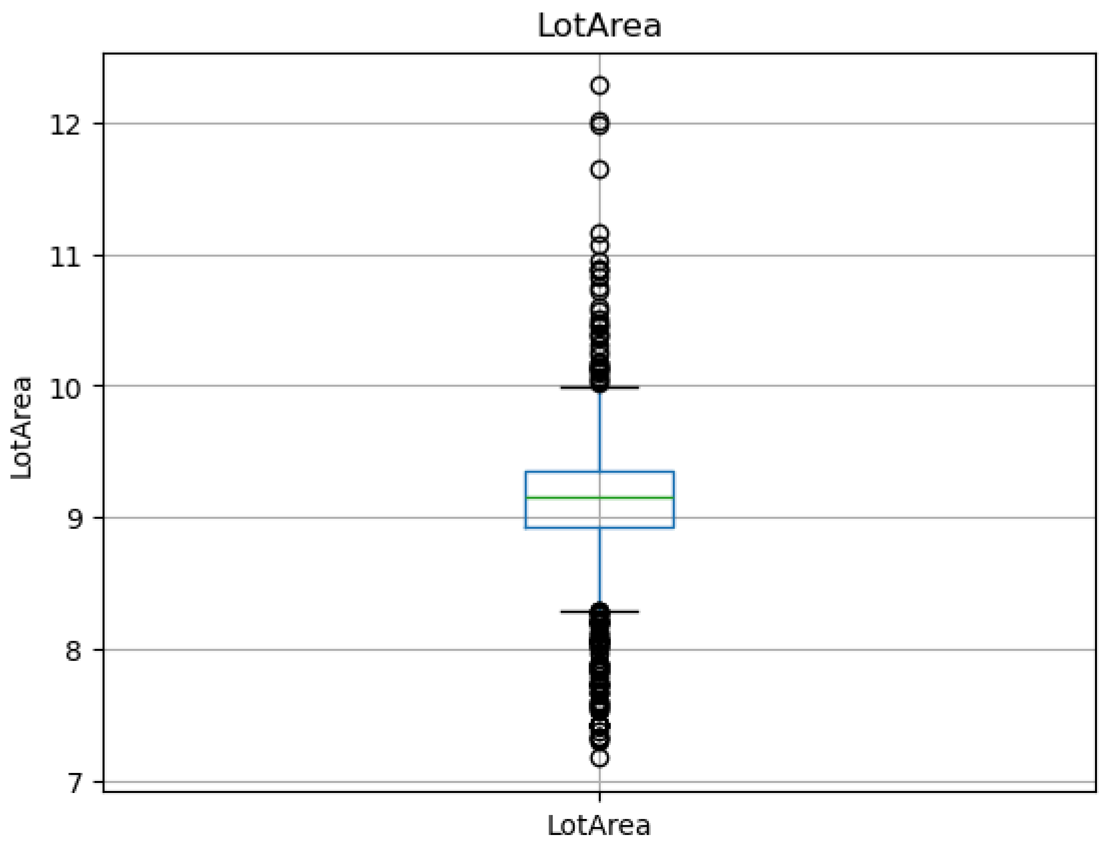

GrLivArea
SalePrice
GrLivArea
13.5
13.0
12.5
12.0
11.5
11.0
10.5
6.0
6.5
7.0
7.5
8.0
8.5

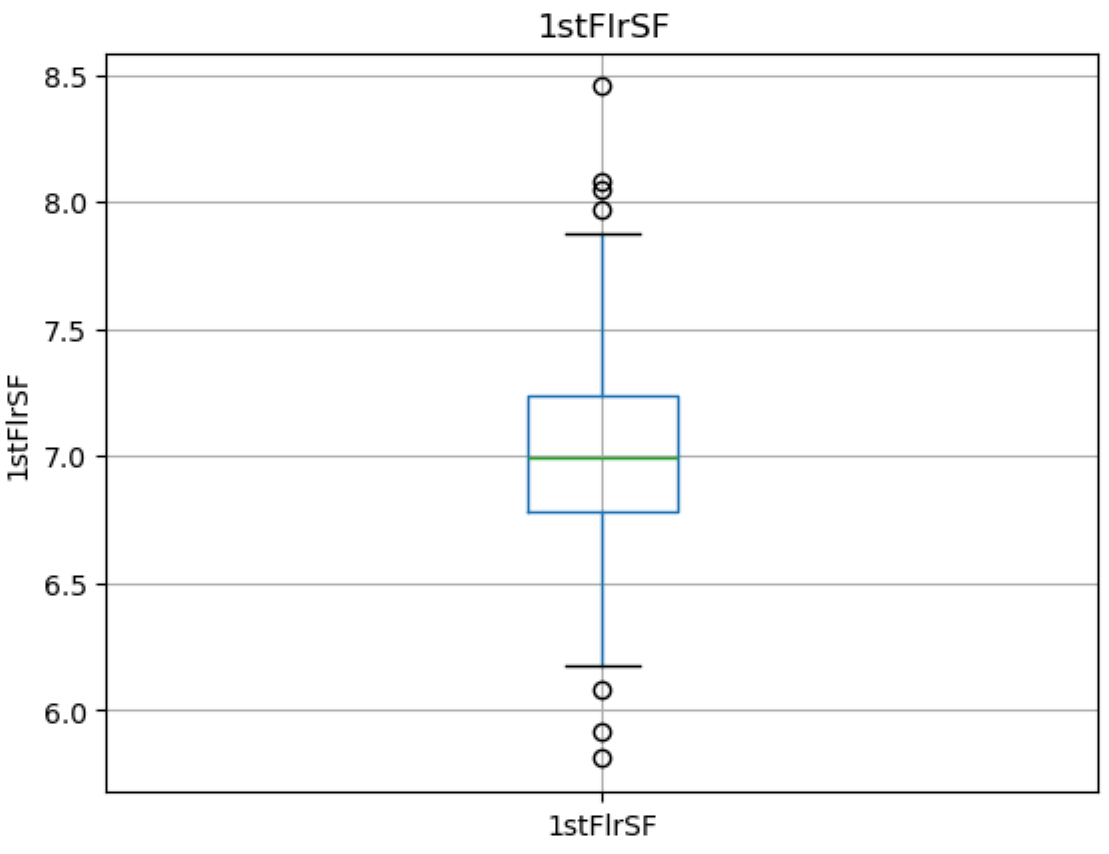


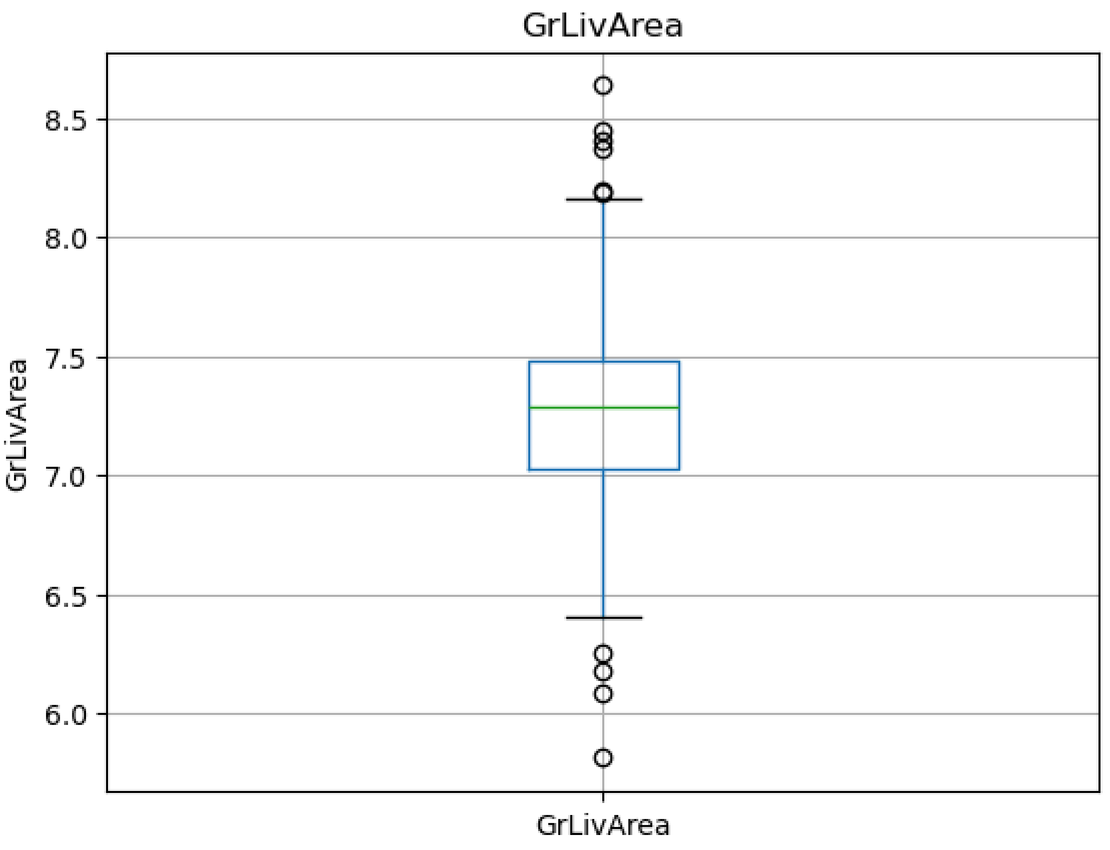
Outliers

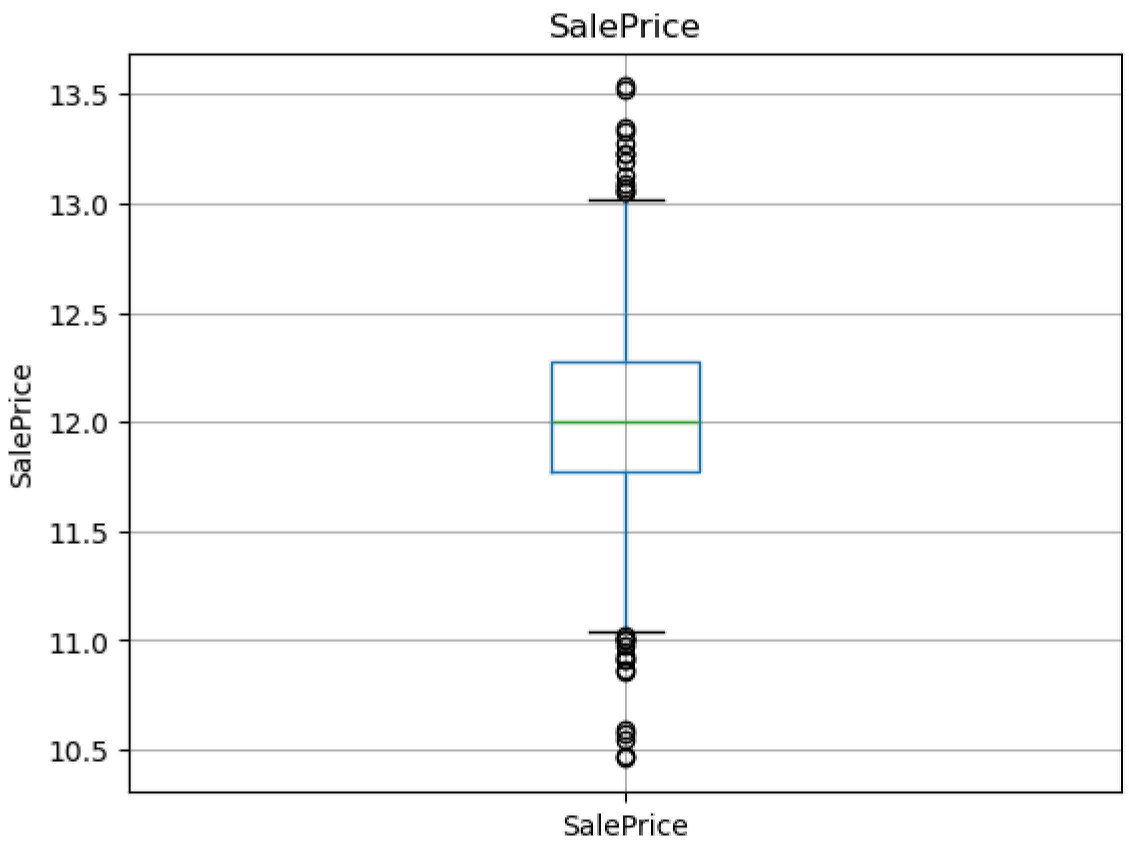
In [24]:
```
for feature in continous_feature:
    data=ds.copy()

    # skip if column is not numeric
    if data[feature].dtype == 'object':
        continue

    if 0 in data[feature].values:
        continue


    else:
        data[feature]=np.log(data[feature])
        data.boxplot(column=feature)
        plt.ylabel(feature)
        plt.title(feature)
        plt.show()
```


LotFrontage
5.5
5.0
4.5
4.0
3.5
3.0
LotFrontage
LotFrontage


LotArea
12
11
10
9
8
7
LotArea
LotArea


1stFlrSF
8.5
8.0
7.5
7.0
6.5
6.0
1stFlrSF
1stFlrSF


GrLivArea
8.5
8.0
7.5
7.0
6.5
6.0
GrLivArea
GrLivArea



Categorical Variables

In [25]:
```
categorical_features=[feature for feature in ds.columns if data[feature].dtype=='O'
categorical_features
```

```
Out[25]: ['MSZoning',
          'Street',
          'Alley',
          'LotShape',
          'LandContour',
          'Utilities',
          'LotConfig',
          'LandSlope',
          'Neighborhood',
          'Condition1',
          'Condition2',
          'BldgType',
          'HouseStyle',
          'RoofStyle',
          'RoofMatl',
          'Exterior1st',
          'Exterior2nd',
          'MasVnrType',
          'ExterQual',
          'ExterCond',
          'Foundation',
          'BsmtQual',
          'BsmtCond',
          'BsmtExposure',
          'BsmtFinType1',
          'BsmtFinType2',
          'Heating',
          'HeatingQC',
          'CentralAir',
          'Electrical',
          'KitchenQual',
          'Functional',
          'FireplaceQu',
          'GarageType',
          'GarageFinish',
          'GarageQual',
          'GarageCond',
          'PavedDrive',
          'PoolQC',
          'Fence',
          'MiscFeature',
          'SaleType',
          'SaleCondition']
```

```
In [26]: ds[categorical_features].head()
```

Out[26]:

	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neig
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	

In [27]:

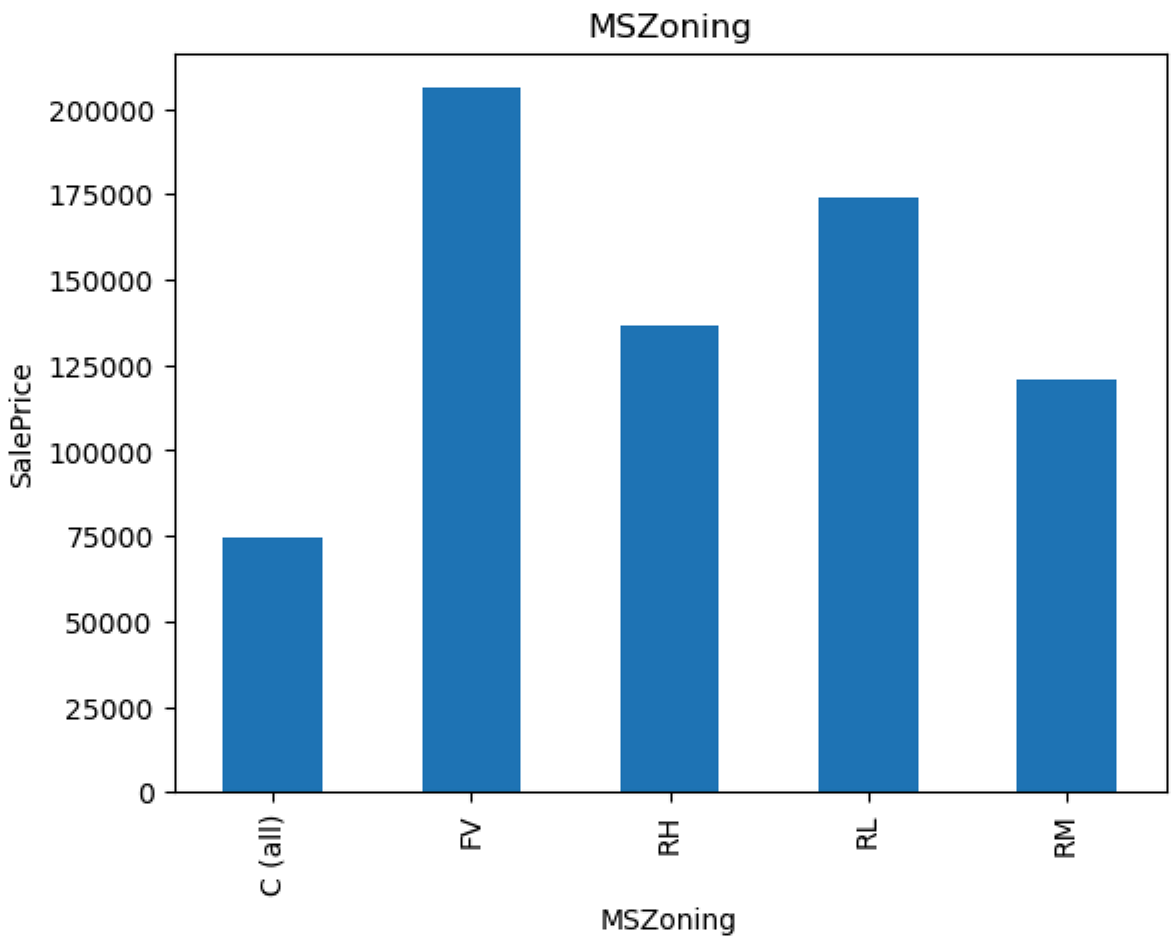
```
for feature in categorical_features:
    print('The feature is {} and number of categories are {}'.format(feature,len(ds
The feature is MSZoning and number of categories are 5
The feature is Street and number of categories are 2
The feature is Alley and number of categories are 3
The feature is LotShape and number of categories are 4
The feature is LandContour and number of categories are 4
The feature is Utilities and number of categories are 2
The feature is LotConfig and number of categories are 5
The feature is LandSlope and number of categories are 3
The feature is Neighborhood and number of categories are 25
The feature is Condition1 and number of categories are 9
The feature is Condition2 and number of categories are 8
The feature is BldgType and number of categories are 5
The feature is HouseStyle and number of categories are 8
The feature is RoofStyle and number of categories are 6
The feature is RoofMatl and number of categories are 8
The feature is Exterior1st and number of categories are 15
The feature is Exterior2nd and number of categories are 16
The feature is MasVnrType and number of categories are 4
The feature is ExterQual and number of categories are 4
The feature is ExterCond and number of categories are 5
The feature is Foundation and number of categories are 6
The feature is BsmtQual and number of categories are 5
The feature is BsmtCond and number of categories are 5
The feature is BsmtExposure and number of categories are 5
The feature is BsmtFinType1 and number of categories are 7
The feature is BsmtFinType2 and number of categories are 7
The feature is Heating and number of categories are 6
The feature is HeatingQC and number of categories are 5
The feature is CentralAir and number of categories are 2
The feature is Electrical and number of categories are 6
The feature is KitchenQual and number of categories are 4
The feature is Functional and number of categories are 7
The feature is FireplaceQu and number of categories are 6
The feature is GarageType and number of categories are 7
The feature is GarageFinish and number of categories are 4
The feature is GarageQual and number of categories are 6
The feature is GarageCond and number of categories are 6
The feature is PavedDrive and number of categories are 3
The feature is PoolQC and number of categories are 4
The feature is Fence and number of categories are 5
The feature is MiscFeature and number of categories are 5
The feature is SaleType and number of categories are 9
The feature is SaleCondition and number of categories are 6
```

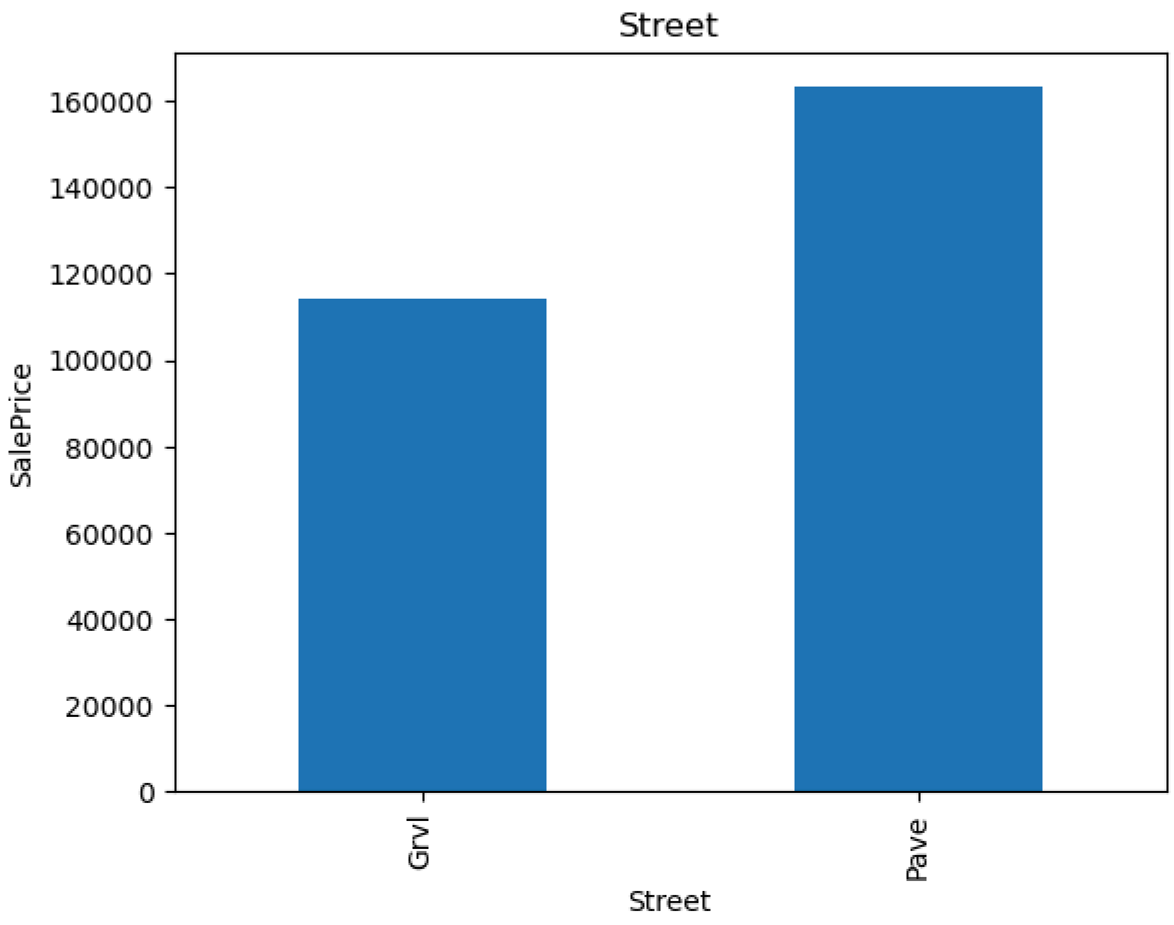
In [28]:

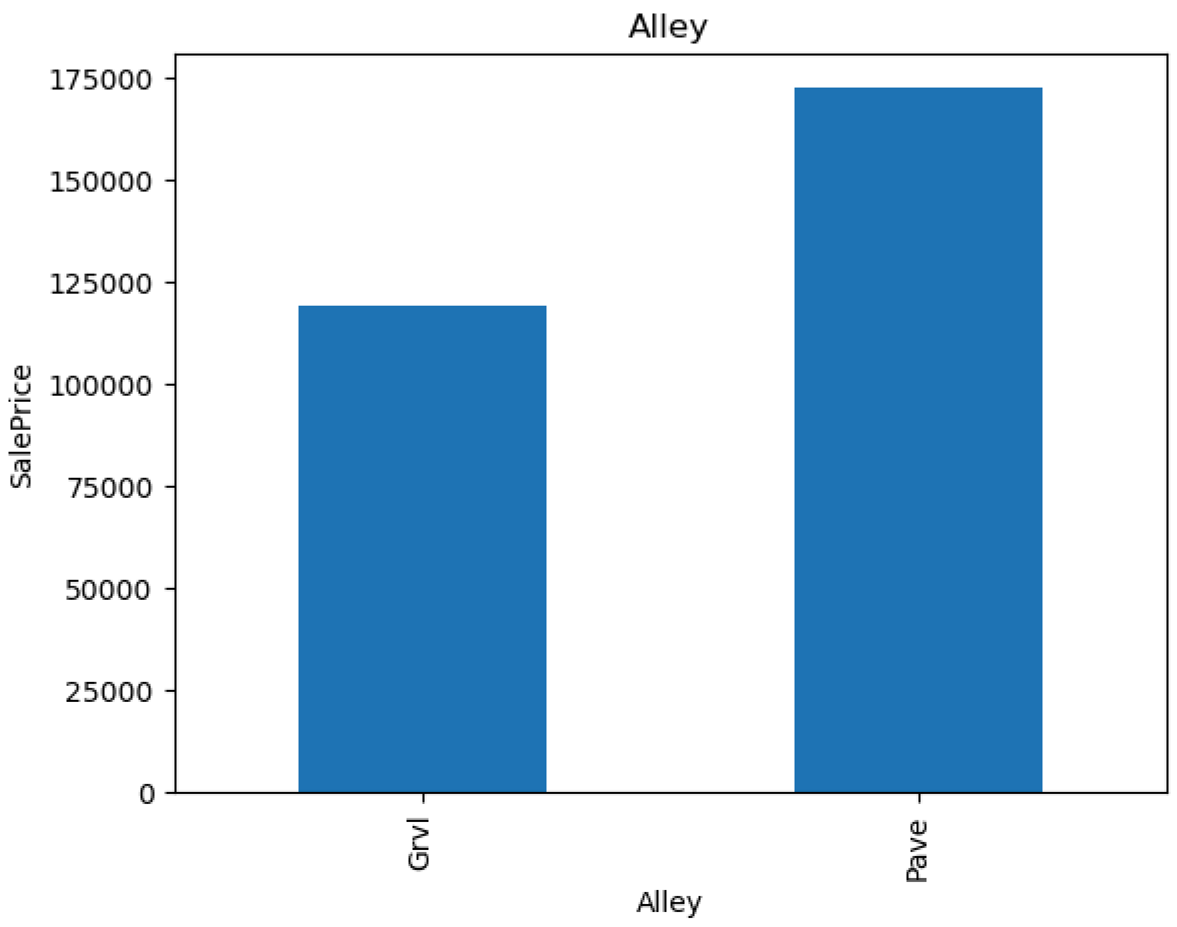
```
## Findout the relationship between categorical variables and dependent feature i.
```

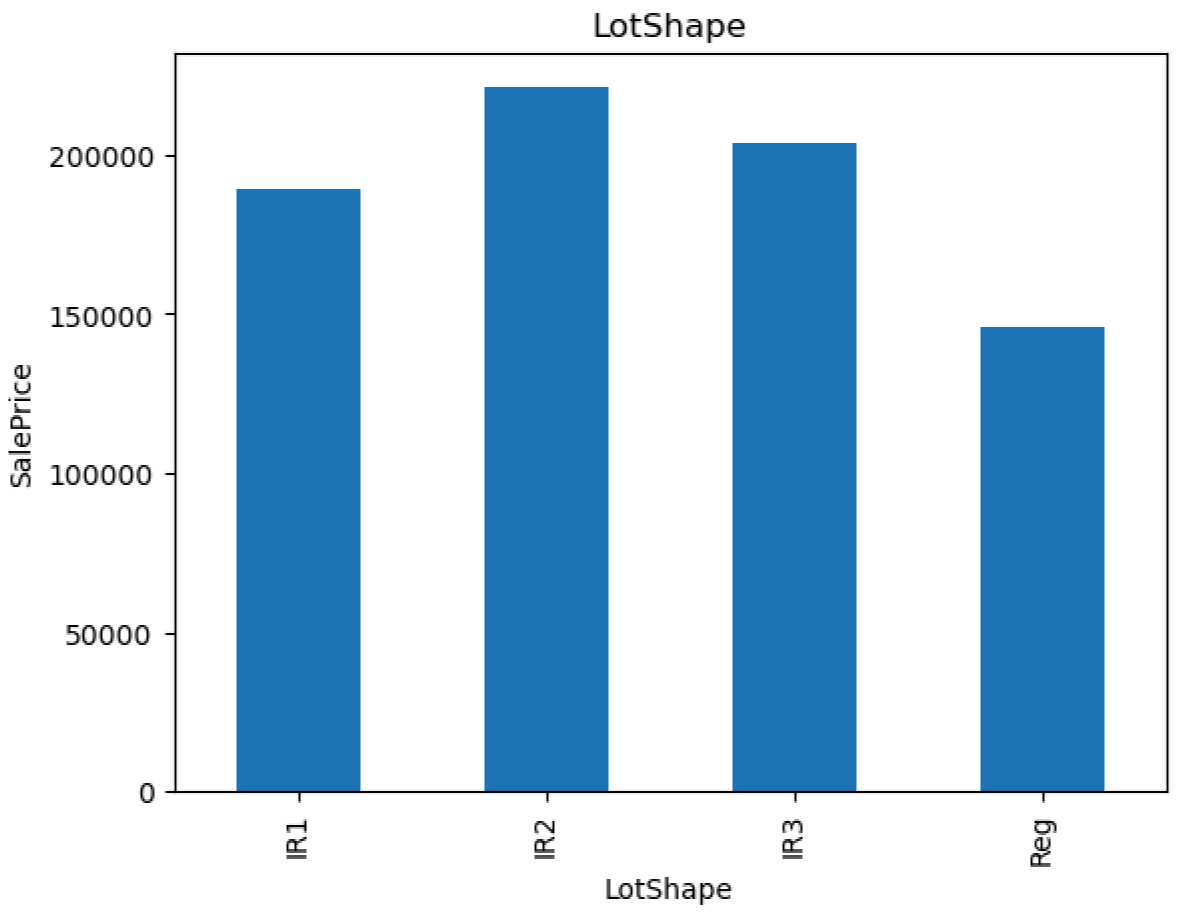
In [29]:

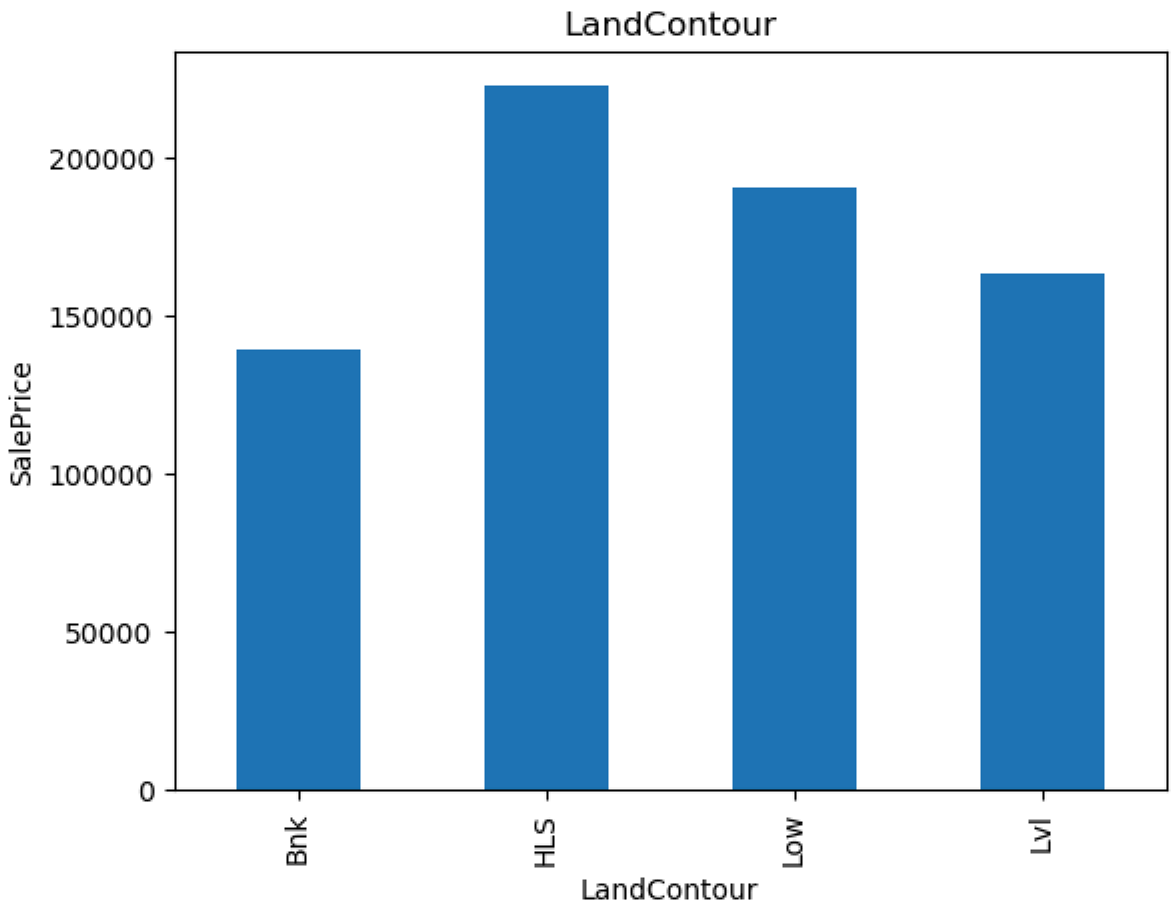
```
for feature in categorical_features:
    data=ds.copy()
    data.groupby(feature)['SalePrice'].median().plot.bar()
    plt.xlabel(feature)
    plt.ylabel('SalePrice')
    plt.title(feature)
    plt.show()
```

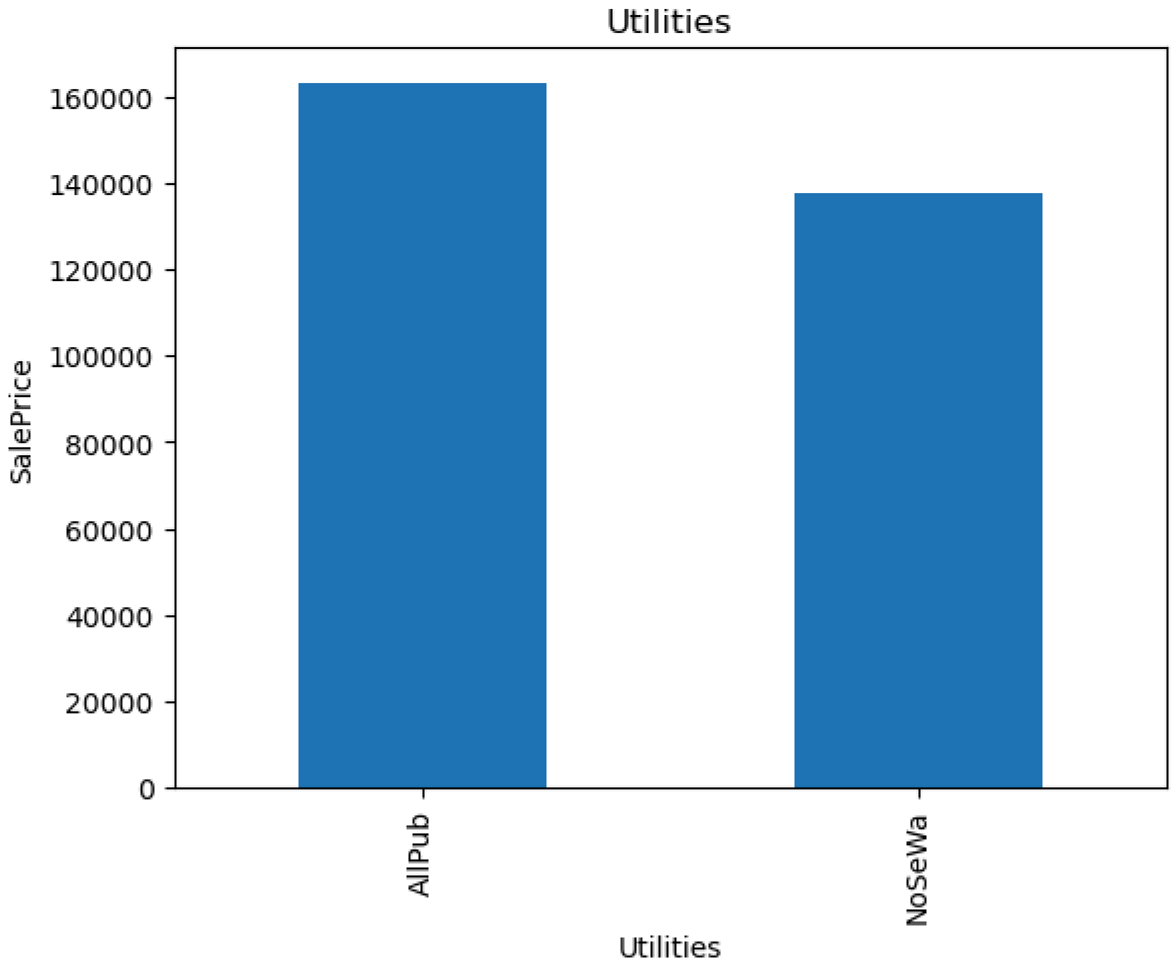

MSZoning
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
C (all)
FV
RH
RL
RM
MSZoning

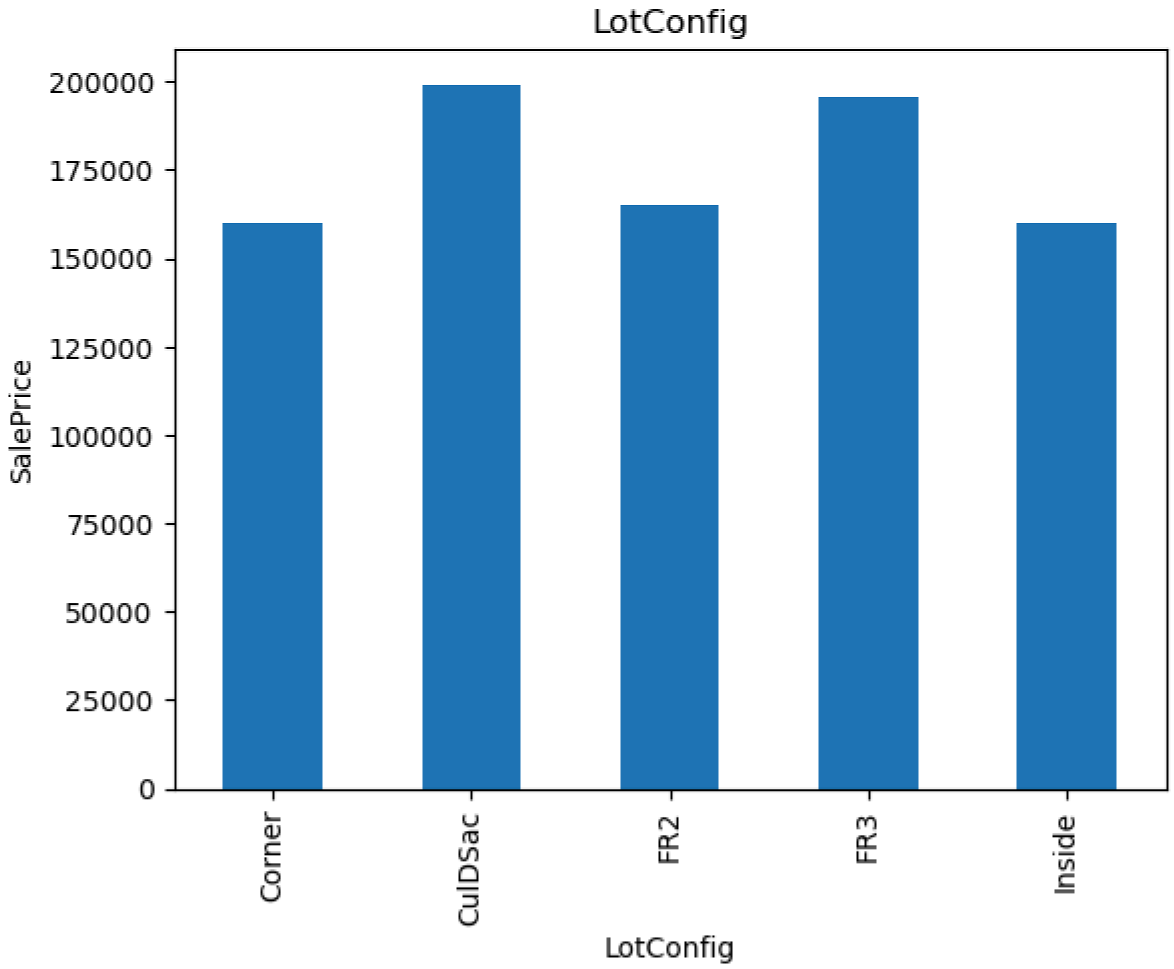

Street
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Grvl
Pave
Street

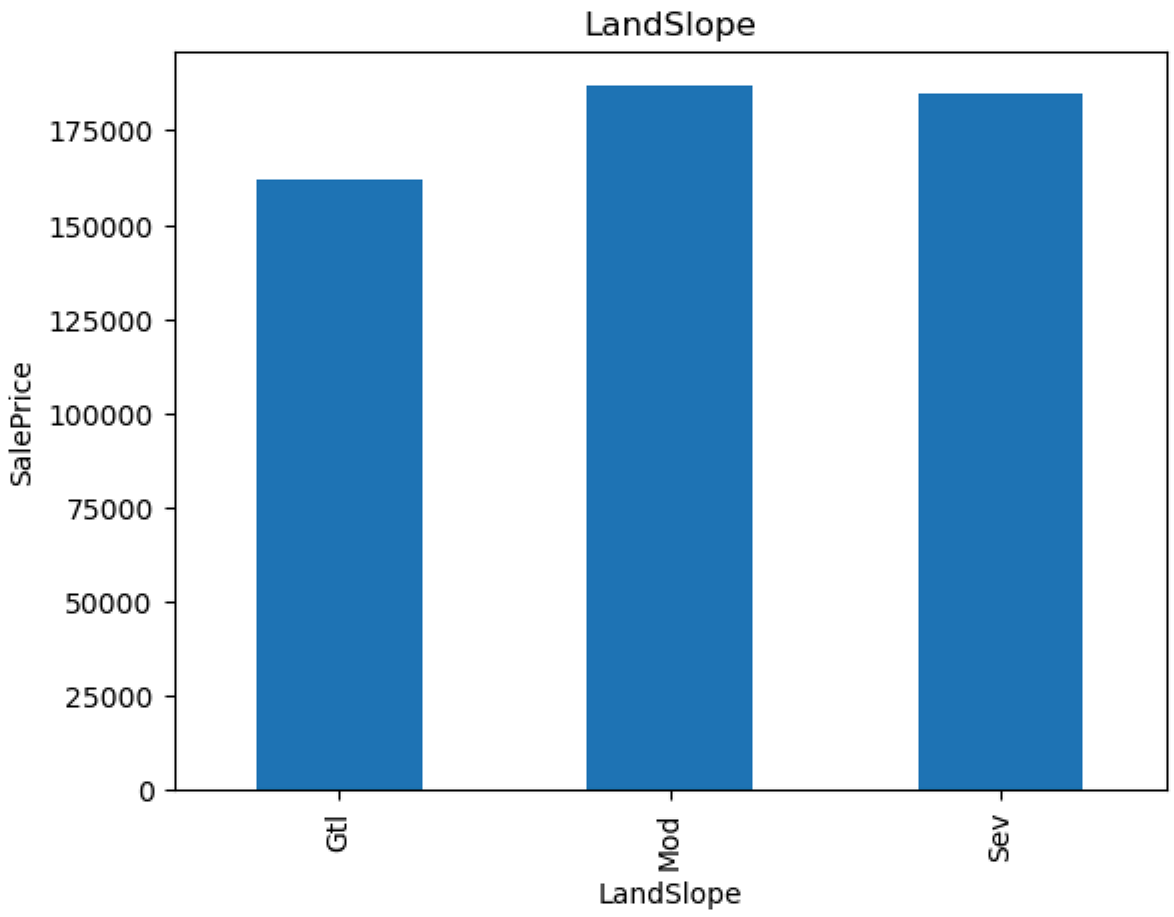

Alley
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
Grvl
Pave
Alley


LotShape
SalePrice
200000
150000
100000
50000
0
IR1
IR2
IR3
Reg
LotShape

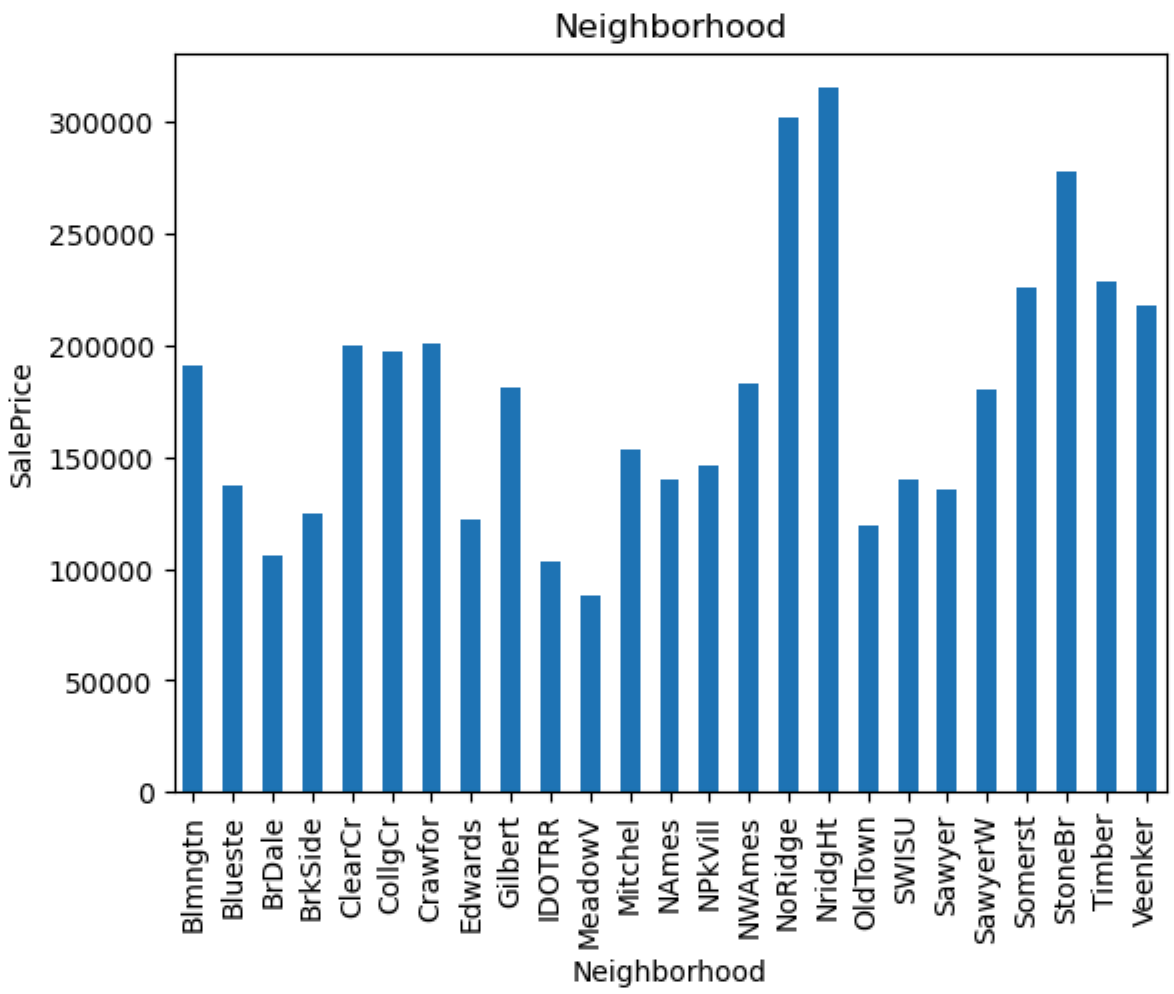

LandContour
SalePrice
200000
150000
100000
50000
0
Bnk
HLS
Low
Lvl
LandContour


Utilities
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
AllPub
NoSeWa
Utilities

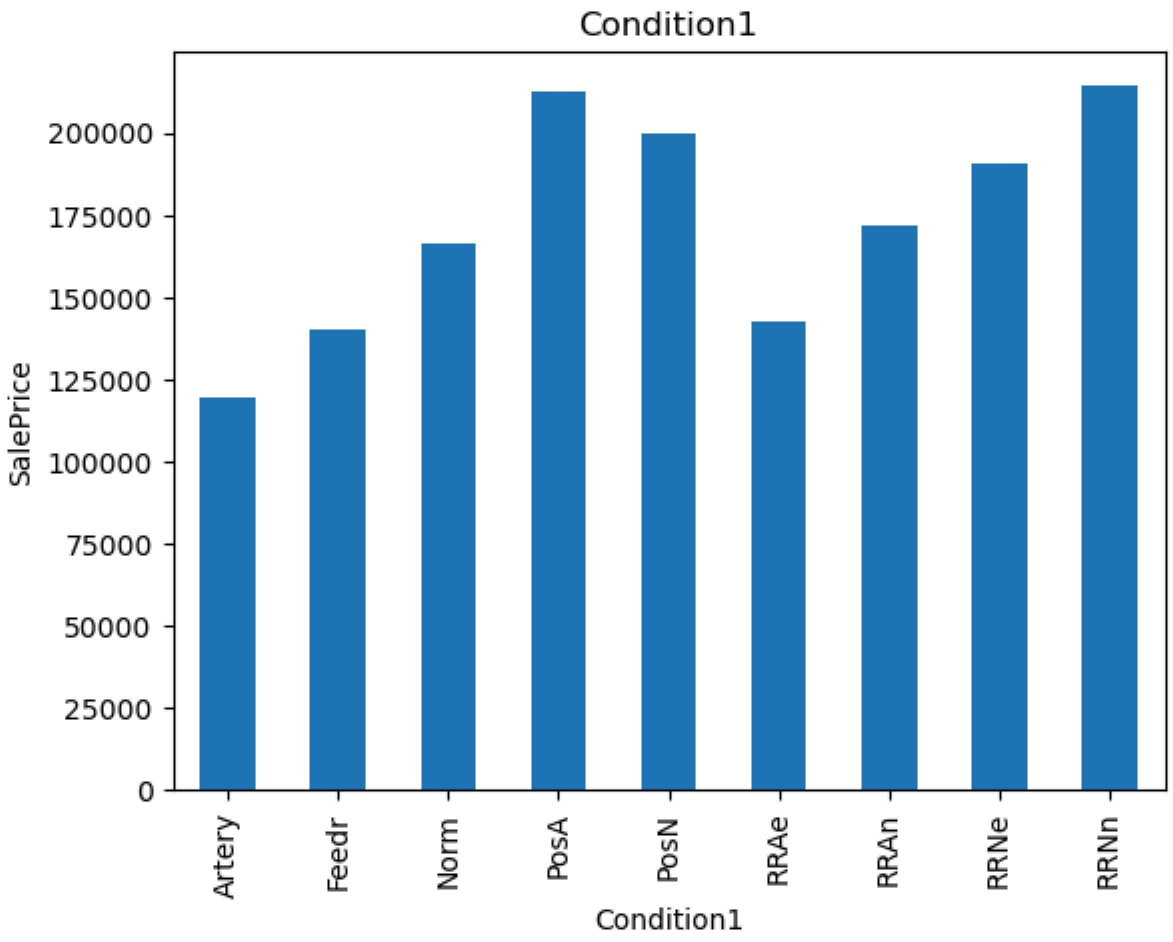

LotConfig
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Corner
CulDSac
FR2
FR3
Inside
LotConfig

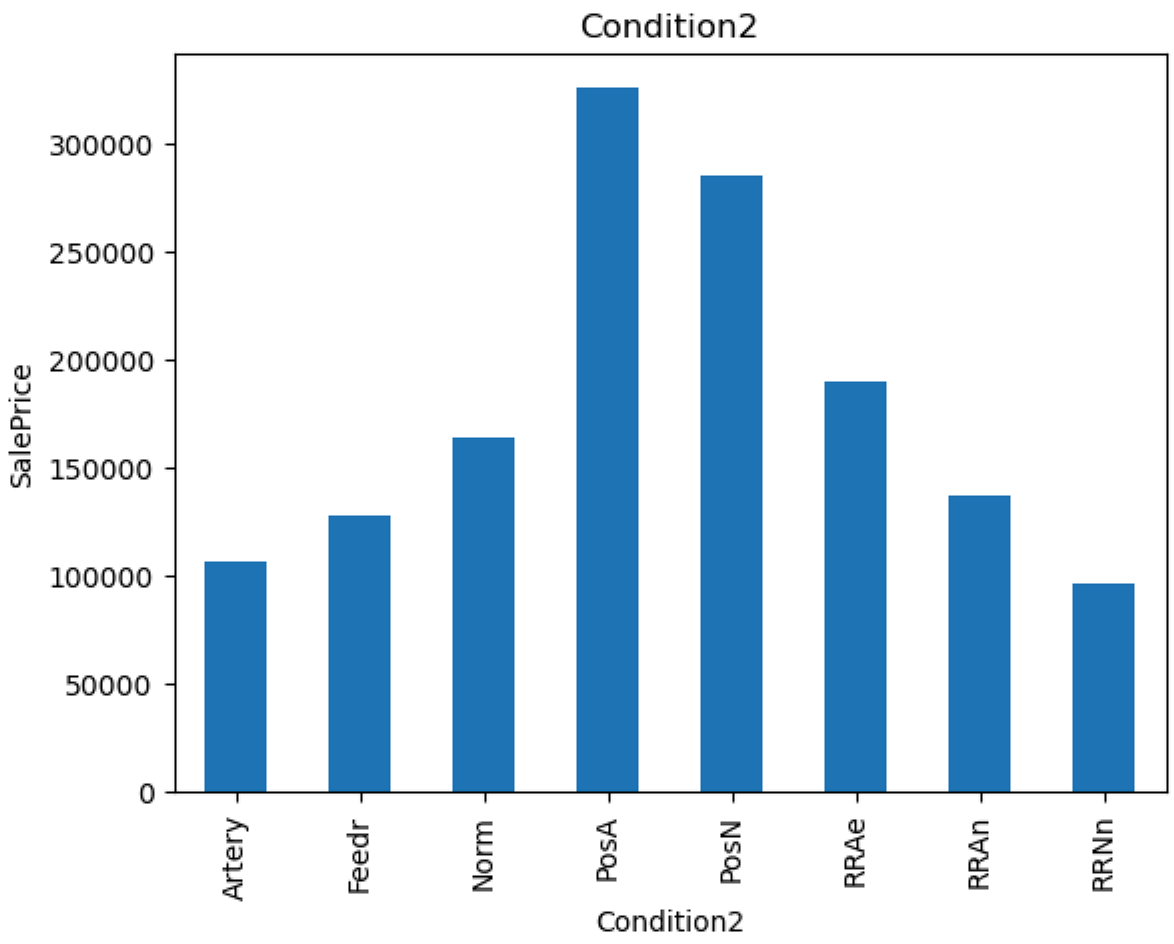


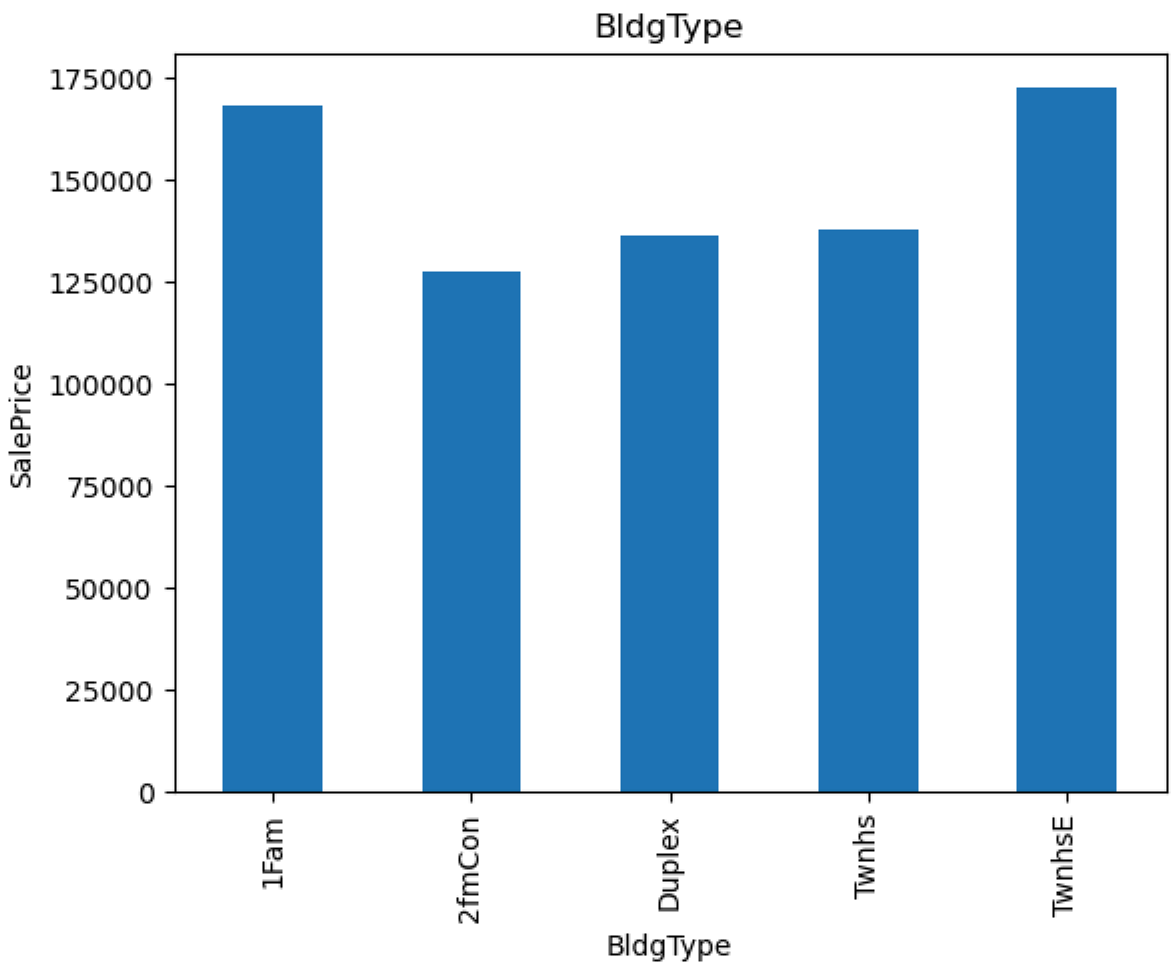
LandSlope
SalePrice
175000
150000
125000
100000
75000
50000
25000
0
Gtl
Mod
Sev
LandSlope

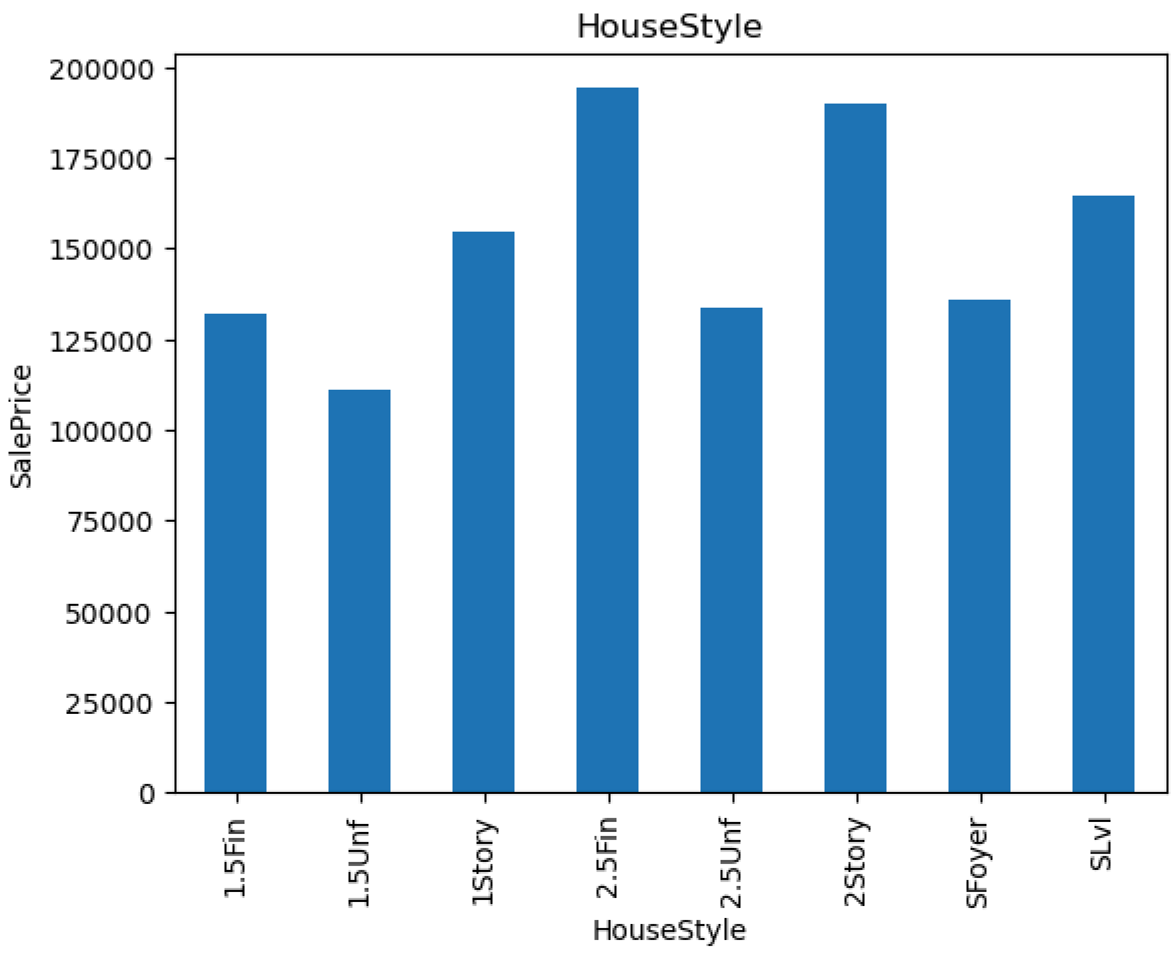


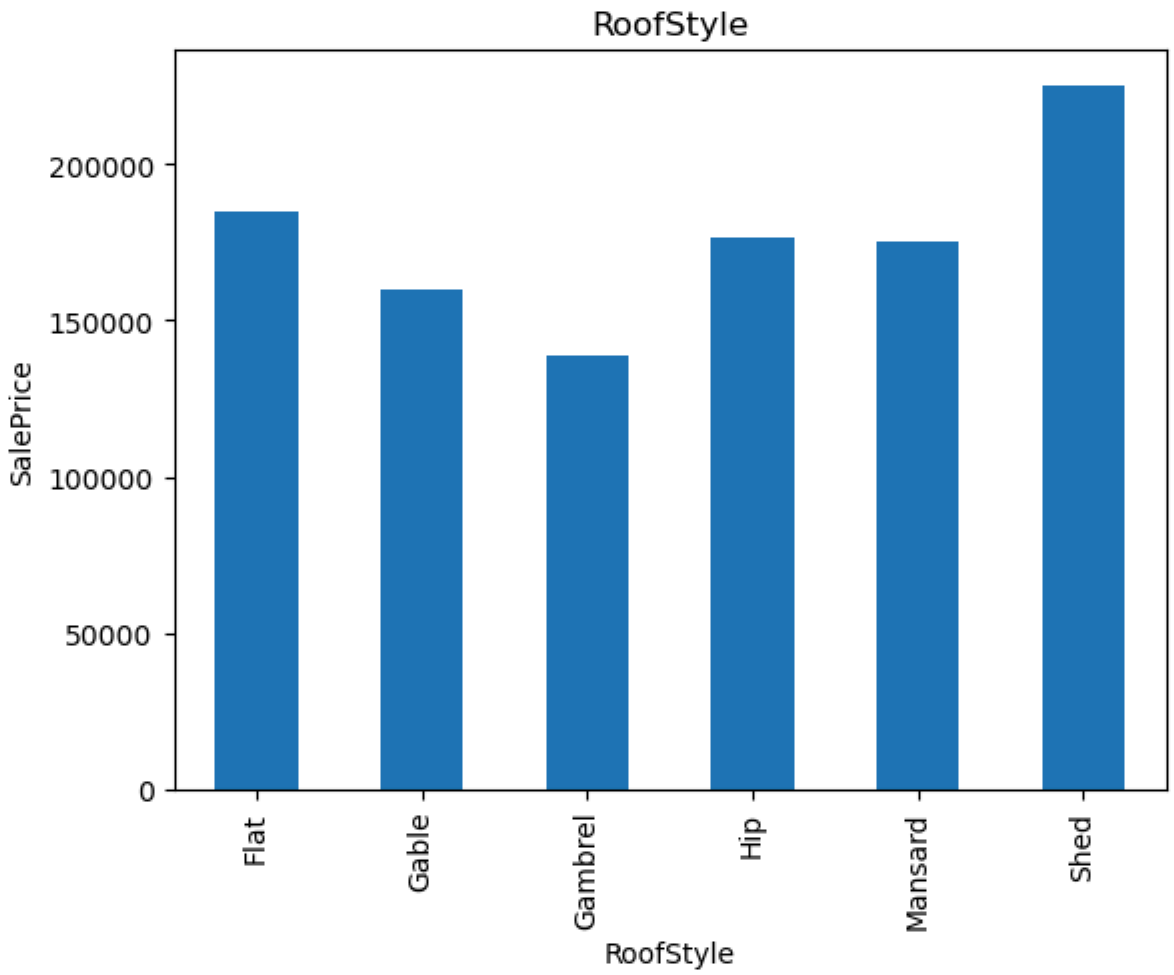
Neighborhood
SalePrice
0
50000
100000
150000
200000
250000
300000
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NPkVill
NWAmes
NoRidge
NridgHt
OldTown
SWISU
Sawyer
SawyerW
Somerst
StoneBr
Timber
Veenker
Neighborhood

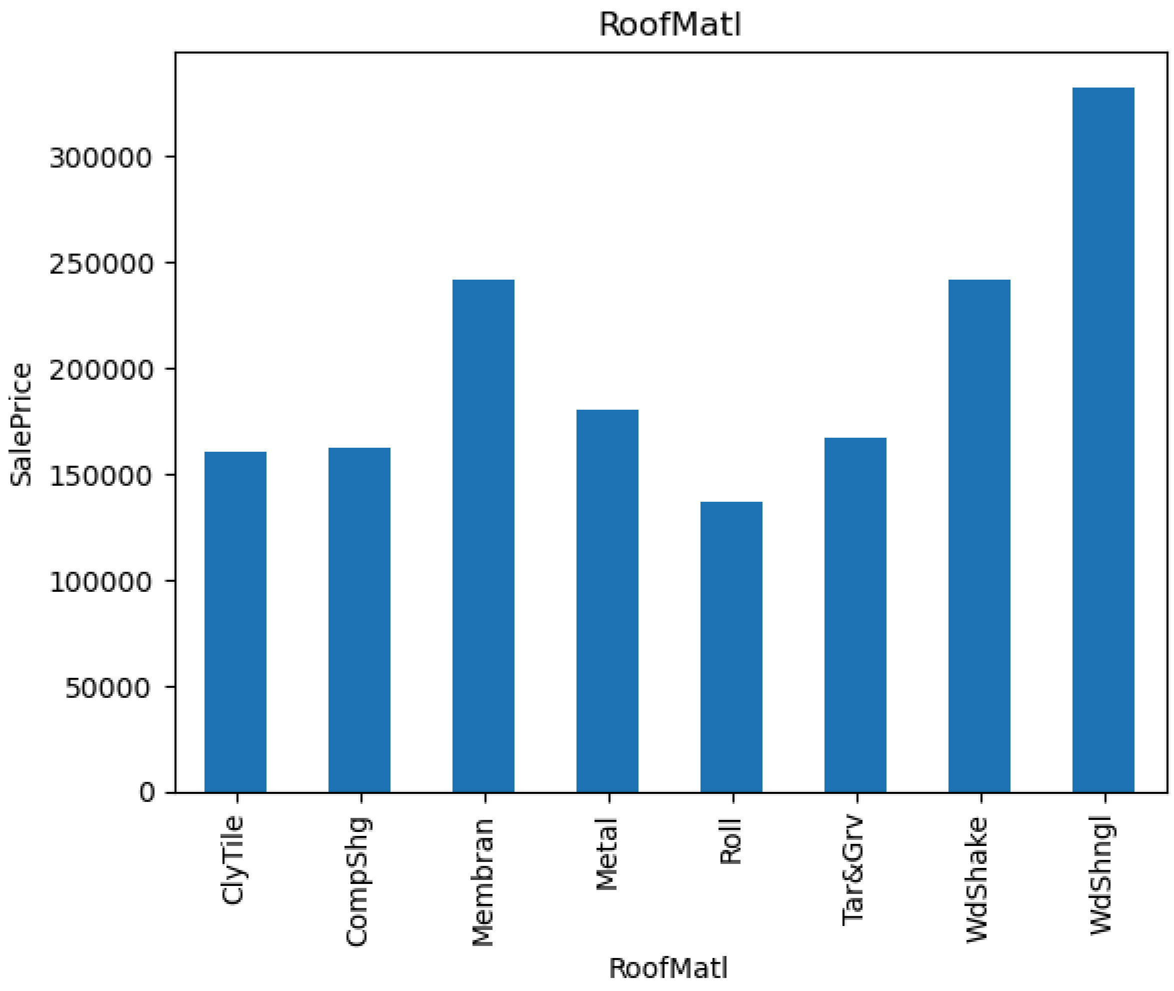

Condition1
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn
Condition1

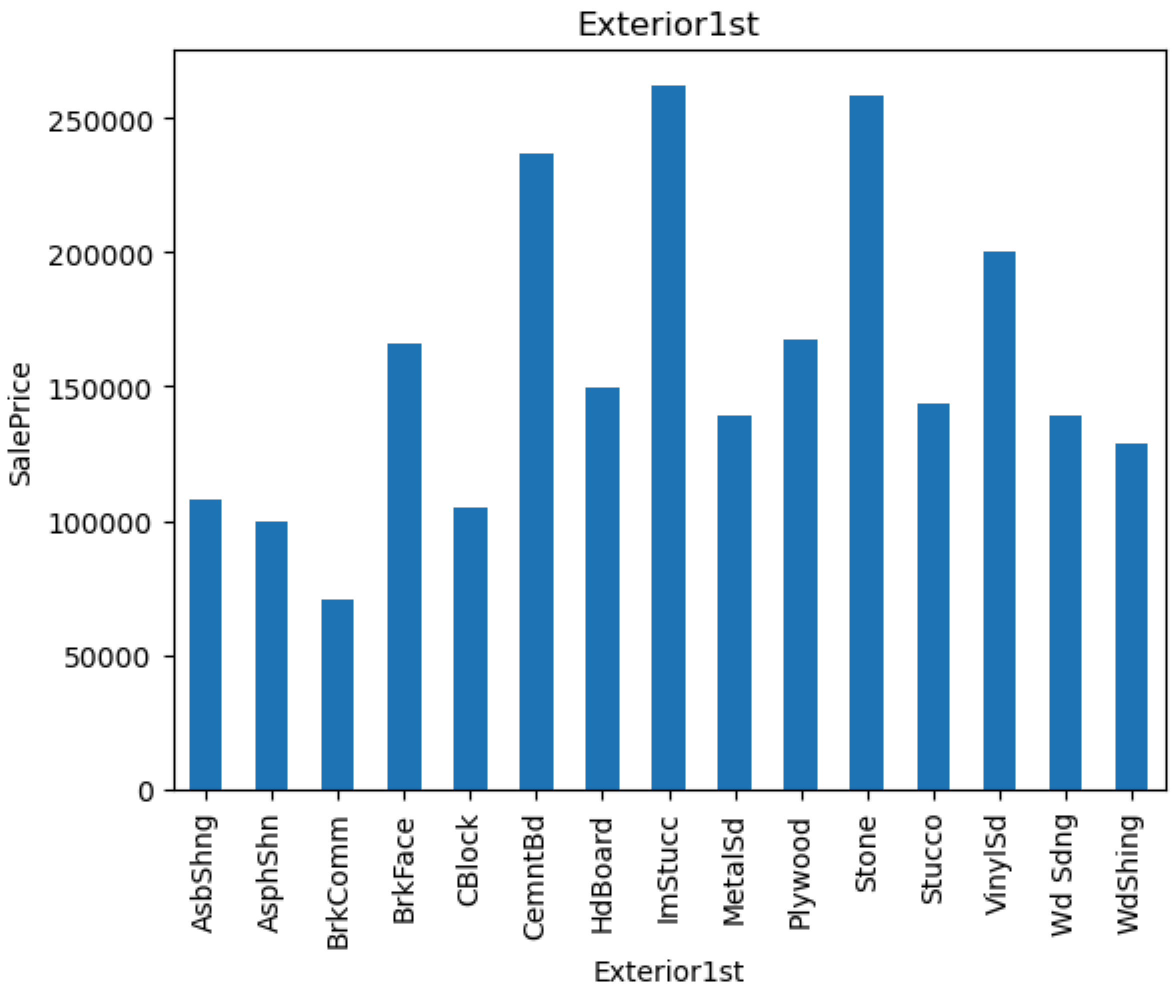

Condition2
SalePrice
300000
250000
200000
150000
100000
50000
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNn
Condition2

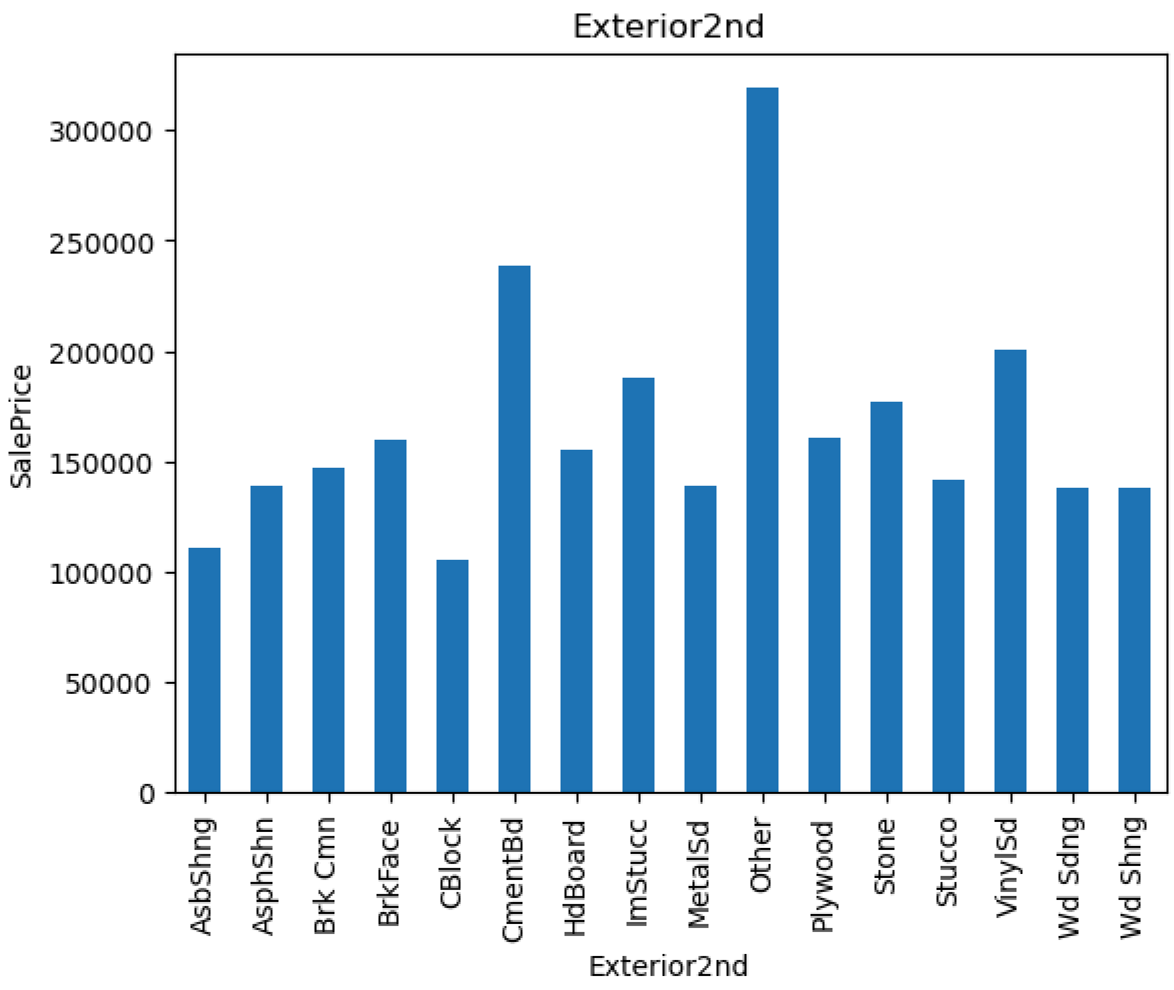

BldgType
175000
150000
125000
100000
75000
50000
25000
0
SalePrice
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
BldgType

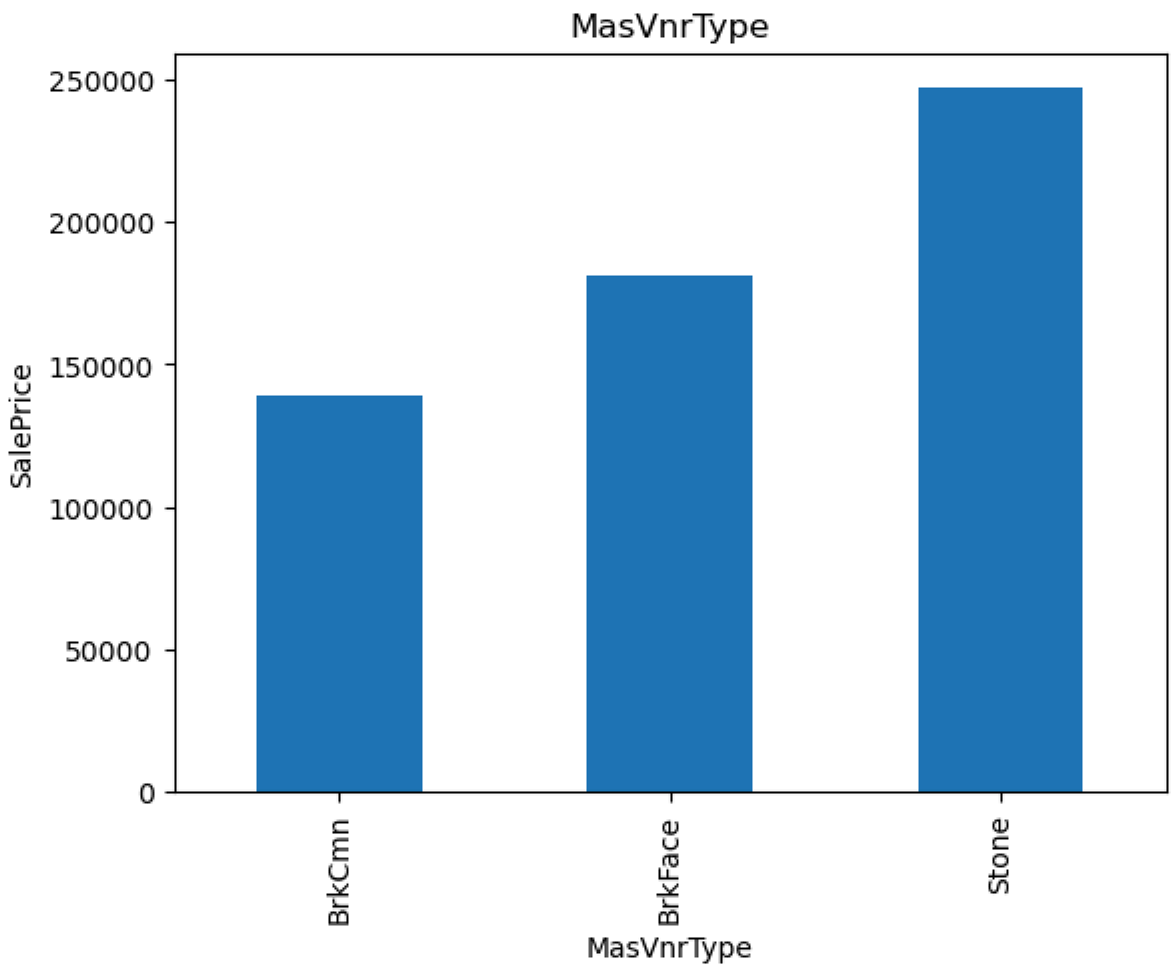

HouseStyle
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle

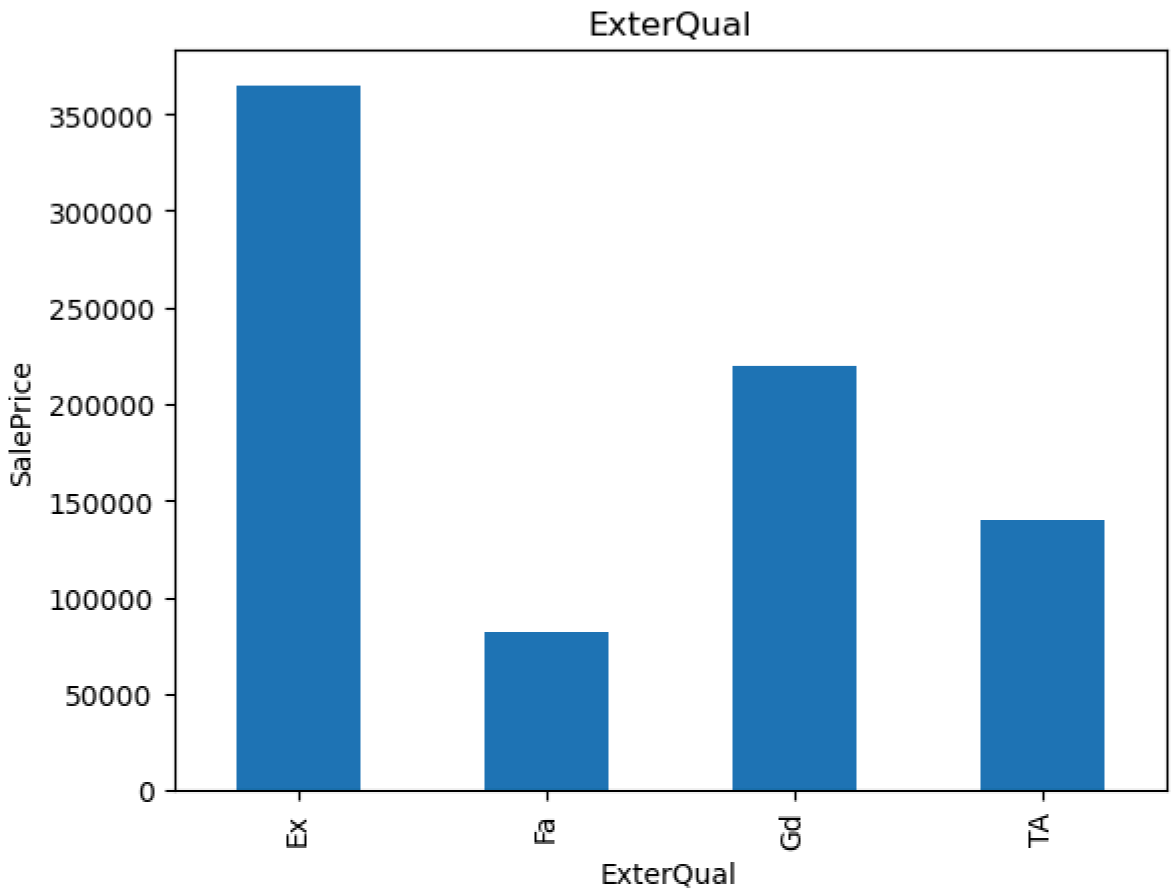

RoofStyle
SalePrice
0
50000
100000
150000
200000
Flat
Gable
Gambrel
Hip
Mansard
Shed
RoofStyle

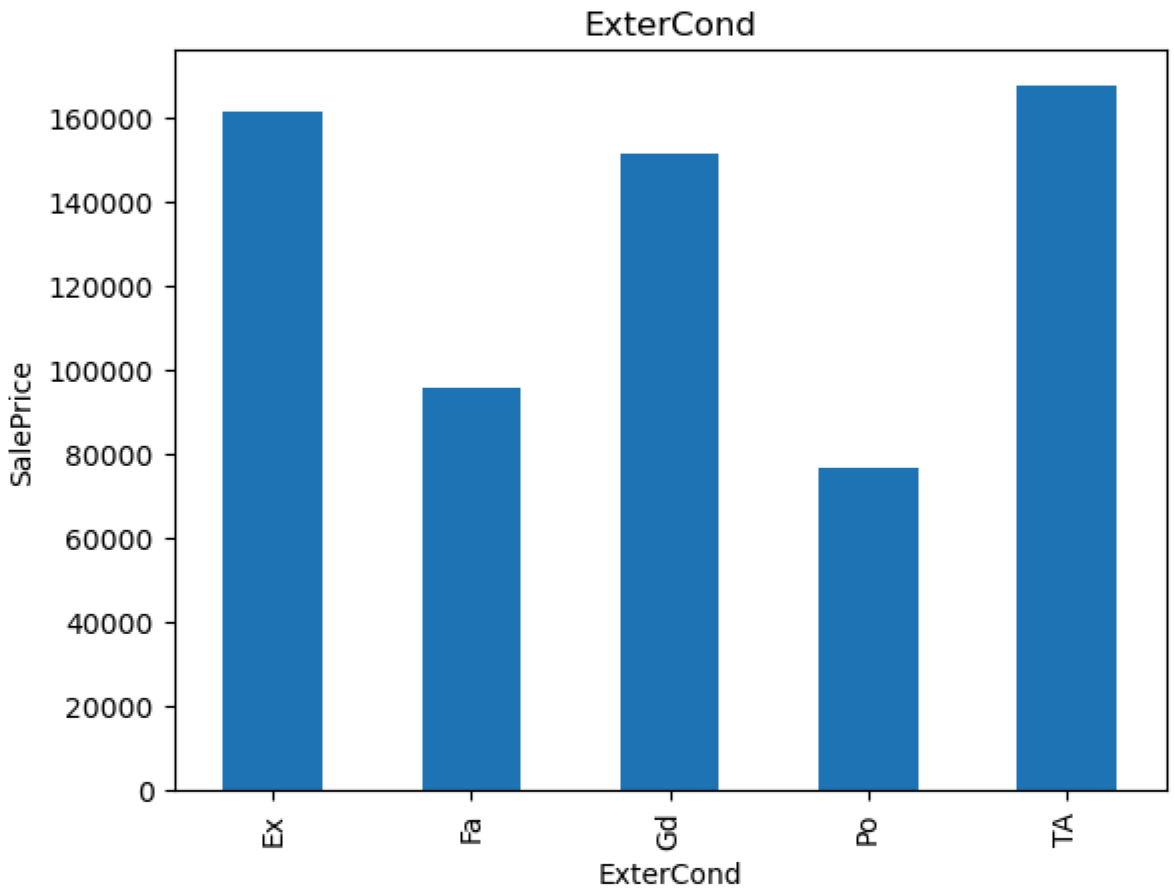

RoofMatl
SalePrice
300000
250000
200000
150000
100000
50000
0
ClyTile
CompShg
Membran
Metal
Roll
Tar&Grv
WdShake
WdShngl
RoofMatl

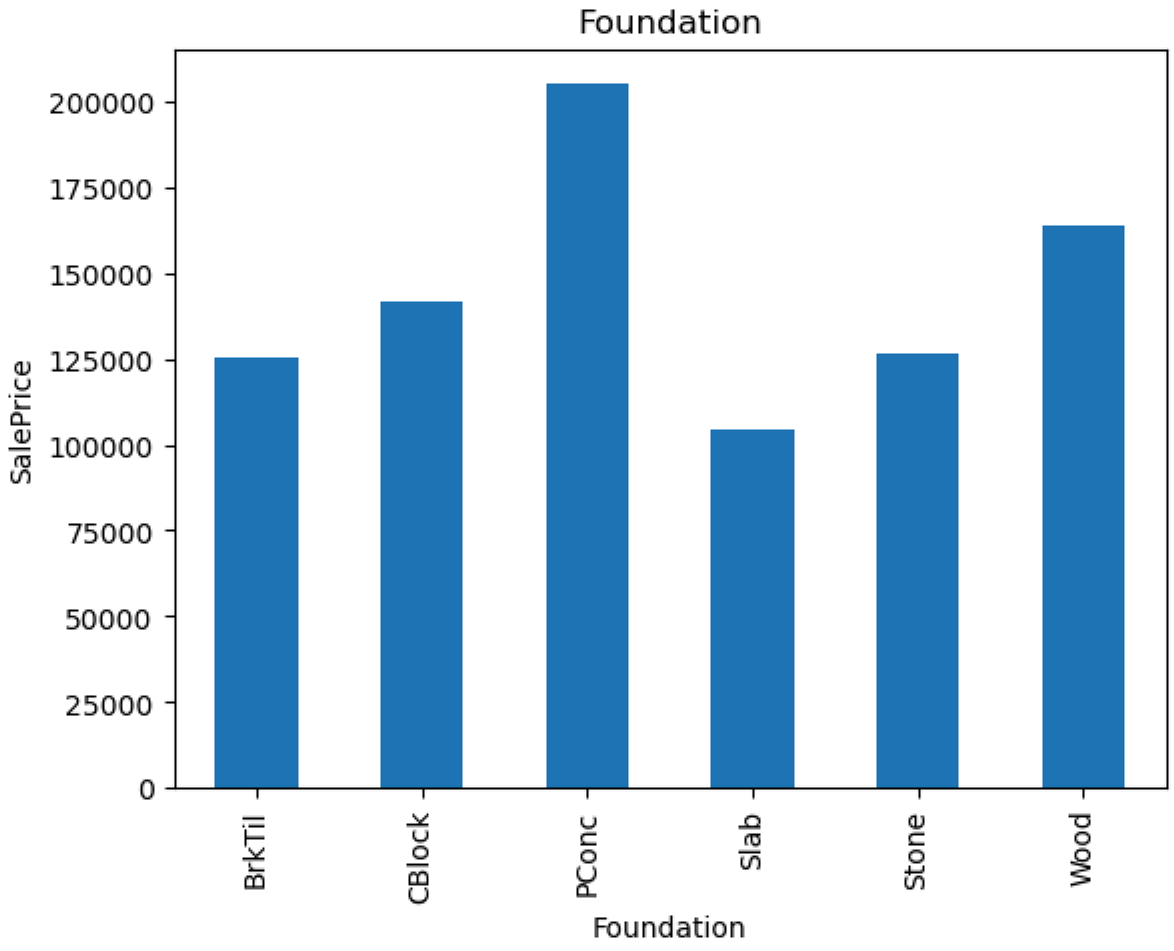

Exterior1st
SalePrice
0
50000
100000
150000
200000
250000
AsbShng
AsphShn
BrkComm
BrkFace
CBlock
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
WdShing
Exterior1st

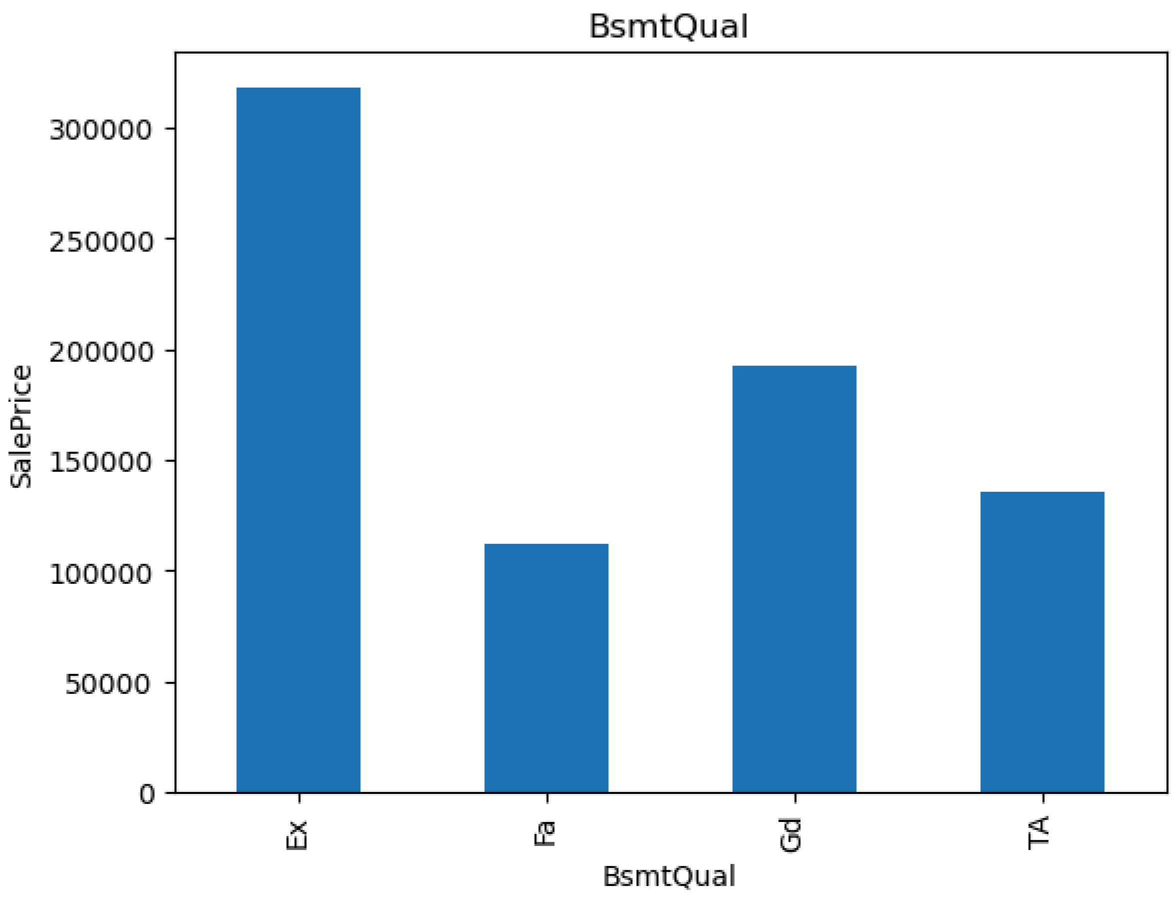

Exterior2nd
SalePrice
300000
250000
200000
150000
100000
50000
0
AsbShng
AsphShn
Brk Cmn
BrkFace
CBlock
CmentBd
HdBoard
ImStucc
MetalSd
Other
Plywood
Stone
Stucco
VinylSd
Wd Sdng
Wd Shng
Exterior2nd

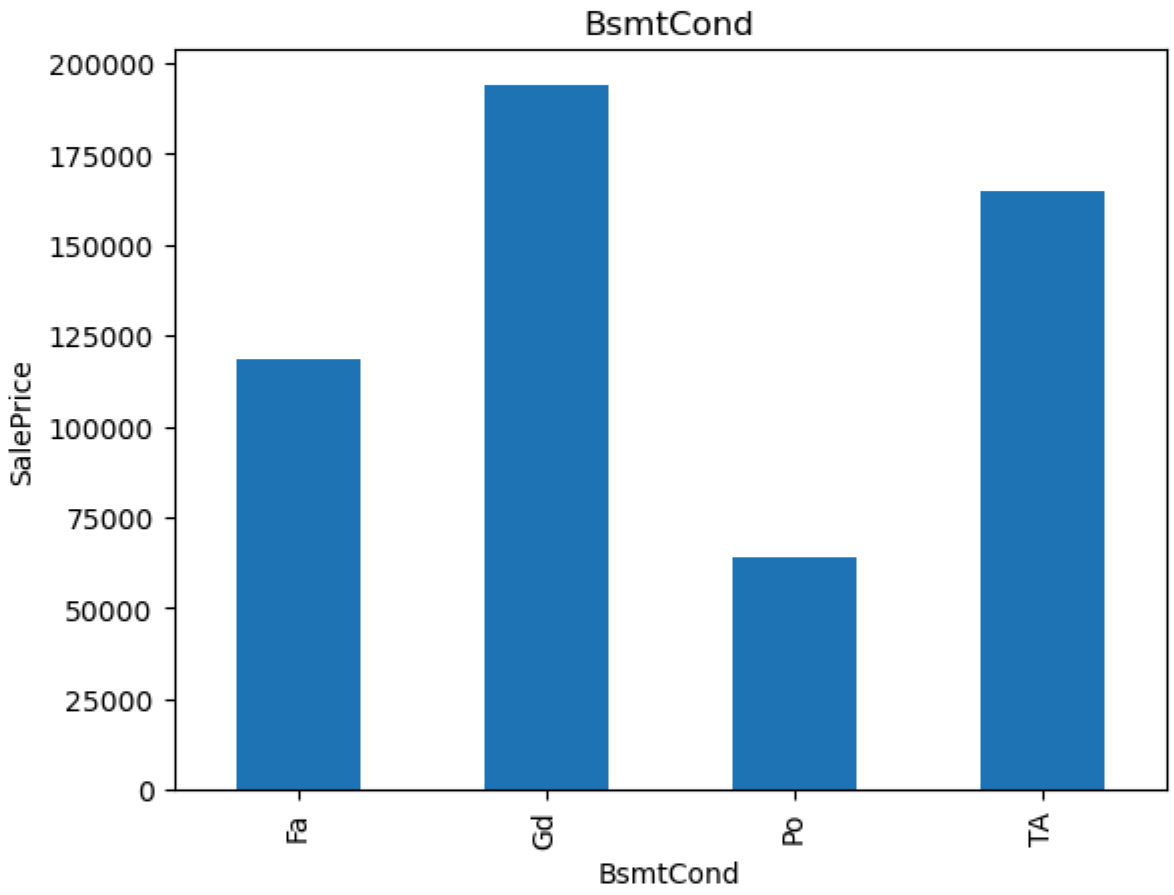

MasVnrType
SalePrice
0
50000
100000
150000
200000
250000
BrkCmn
BrkFace
Stone
MasVnrType

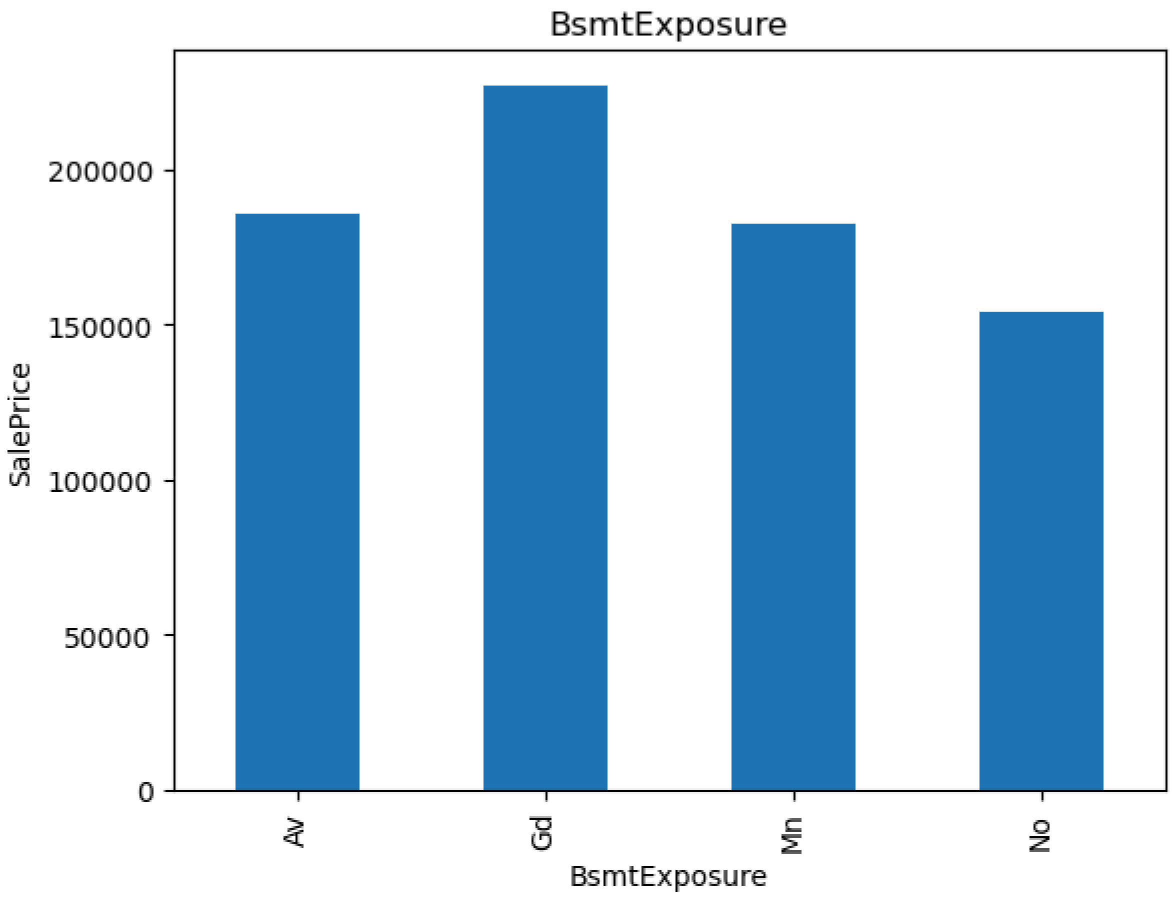

ExterQual
SalePrice
0
50000
100000
150000
200000
250000
300000
350000
Ex
Fa
Gd
TA
ExterQual

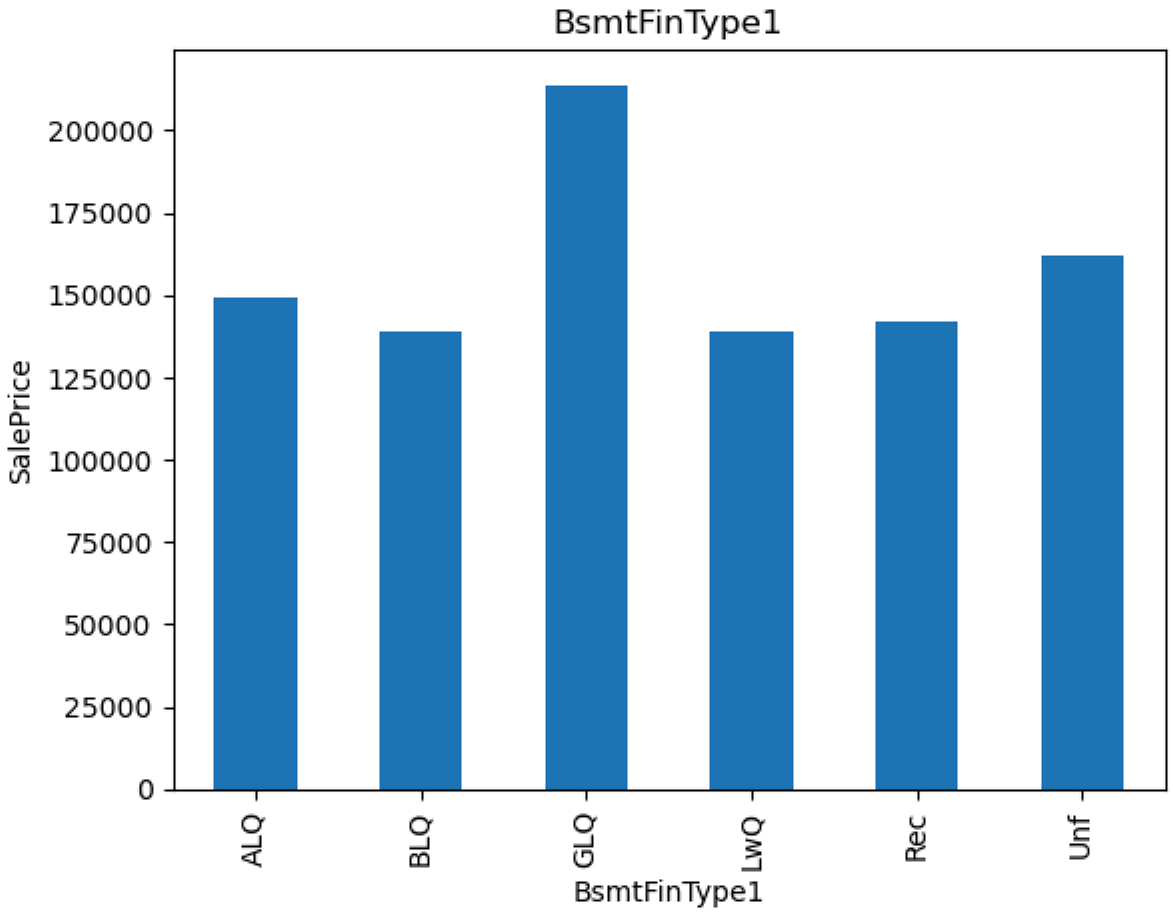

ExterCond
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Ex
Fa
Gd
Po
TA
ExterCond

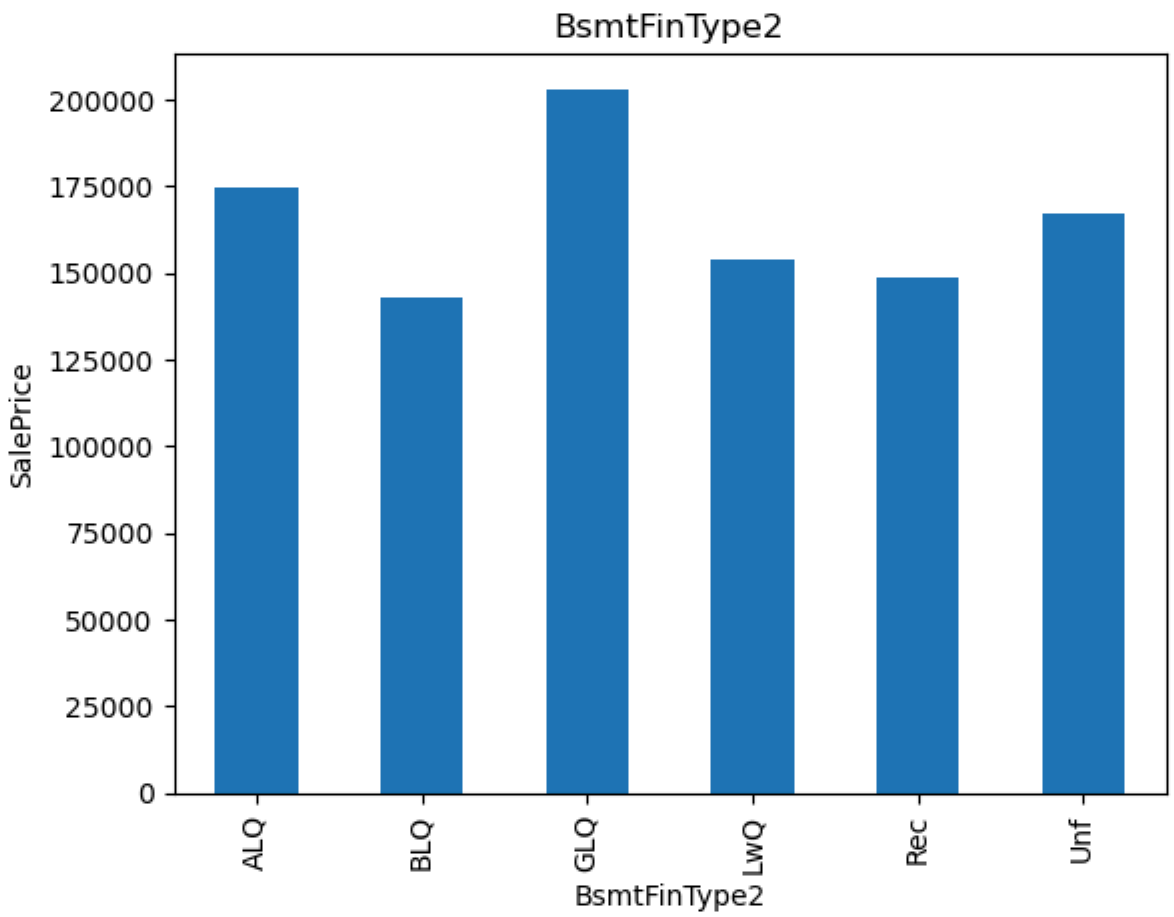

Foundation
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
BrkTil
CBlock
PConc
Slab
Stone
Wood
Foundation

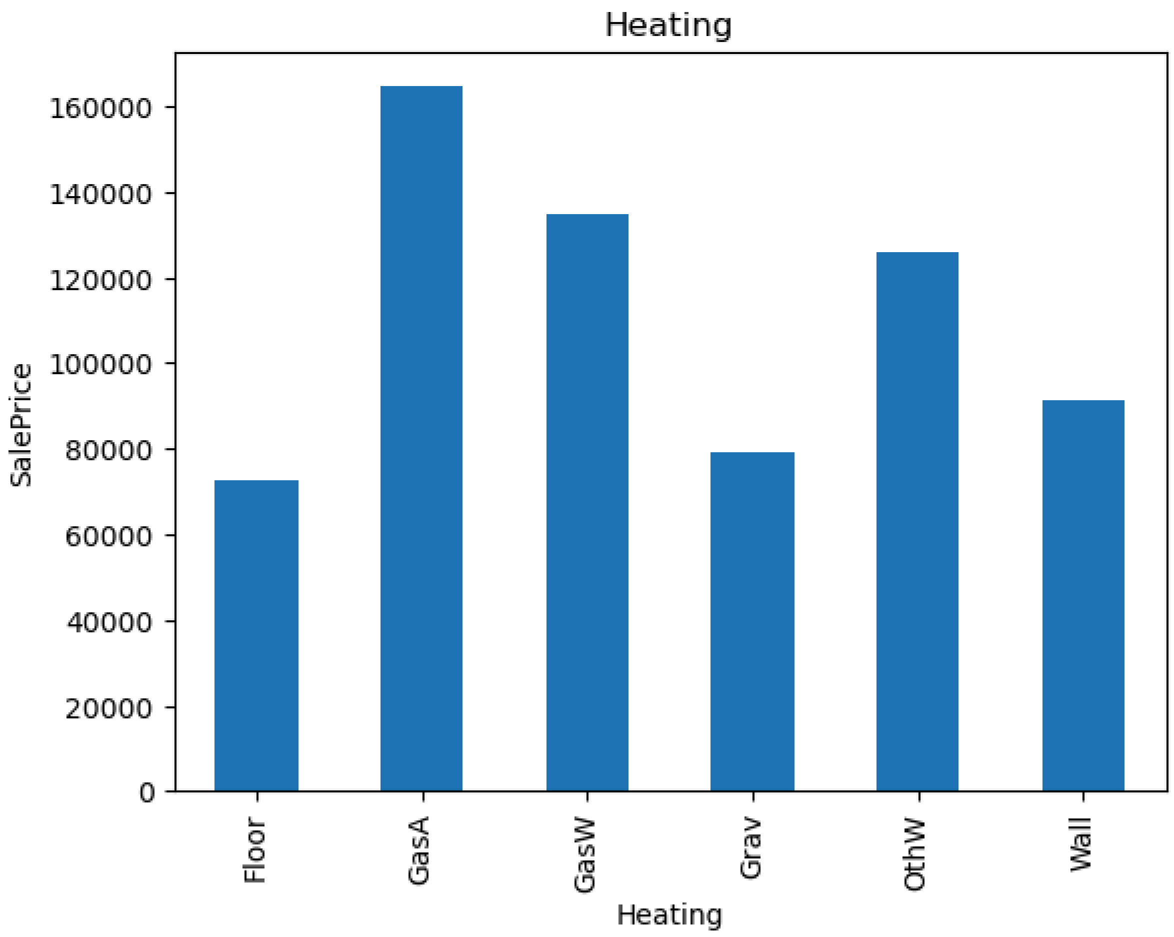

BsmtQual
SalePrice
300000
250000
200000
150000
100000
50000
0
Ex
Fa
Gd
TA
BsmtQual

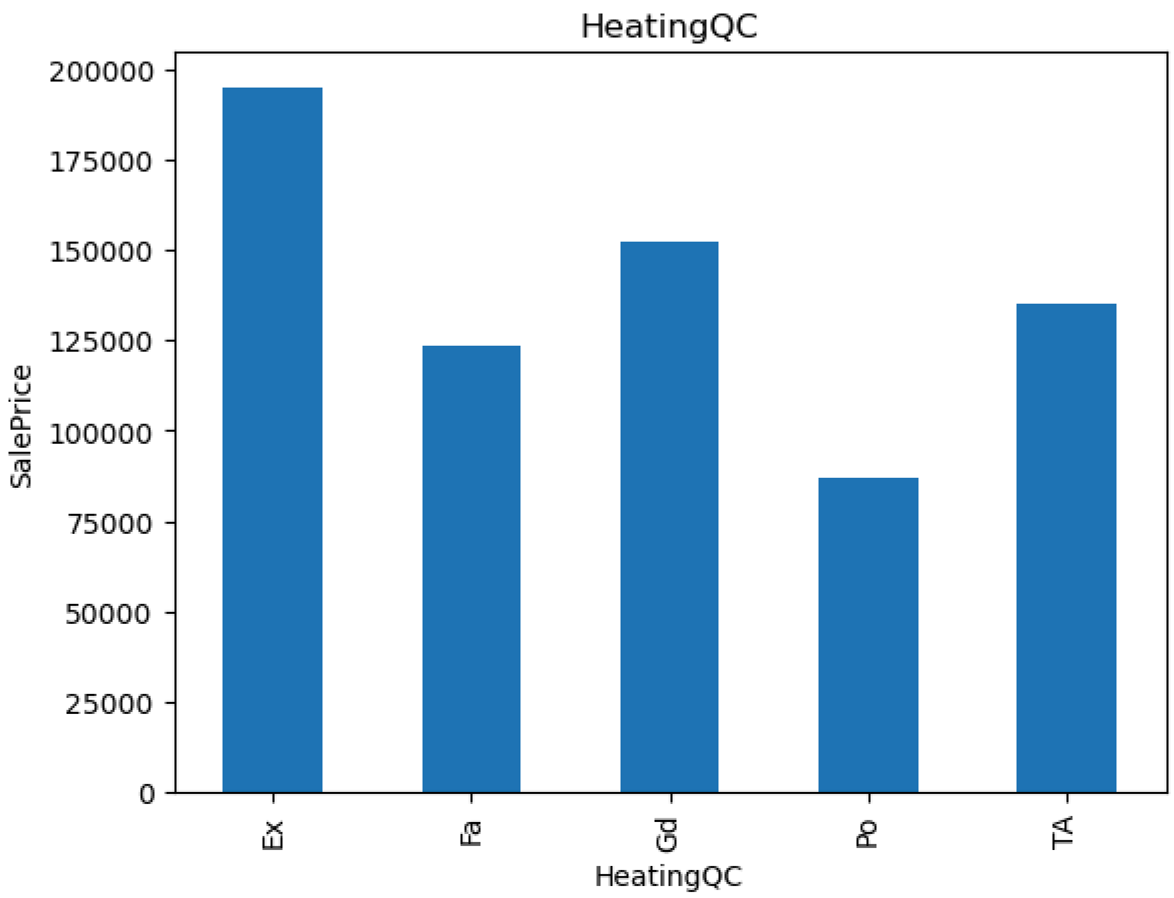

BsmtCond
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Fa
Gd
Po
TA
BsmtCond

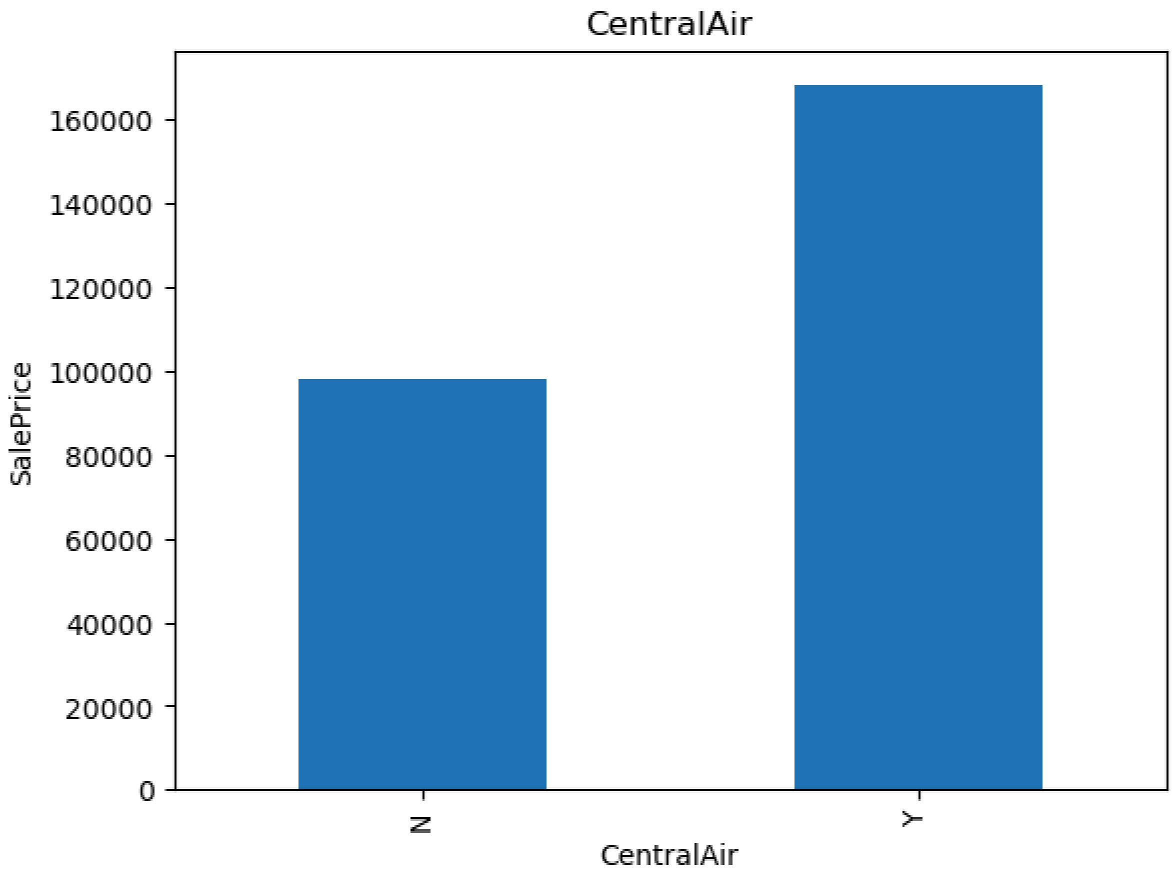

BsmtExposure
SalePrice
0
50000
100000
150000
200000
Av
Gd
Mn
No
BsmtExposure

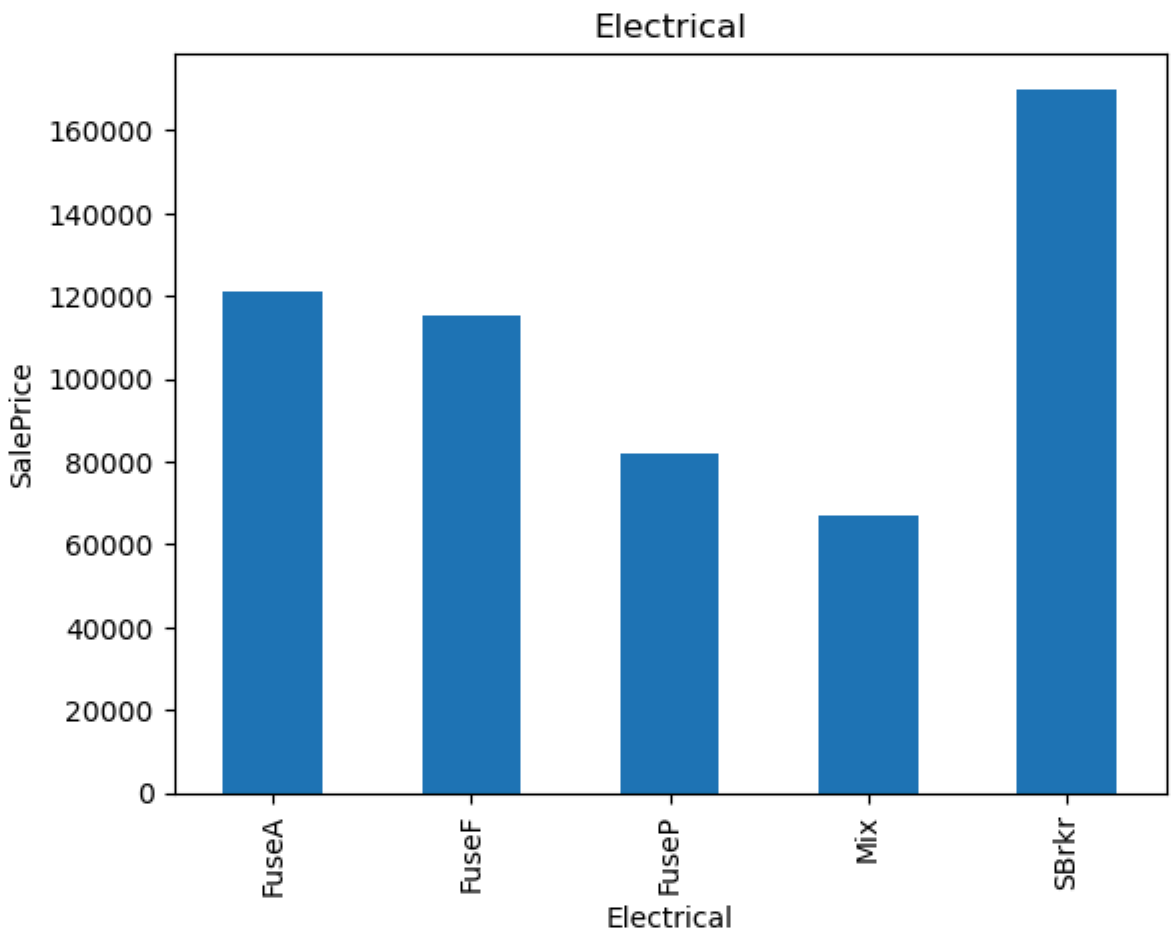

BsmtFinType1
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinType1

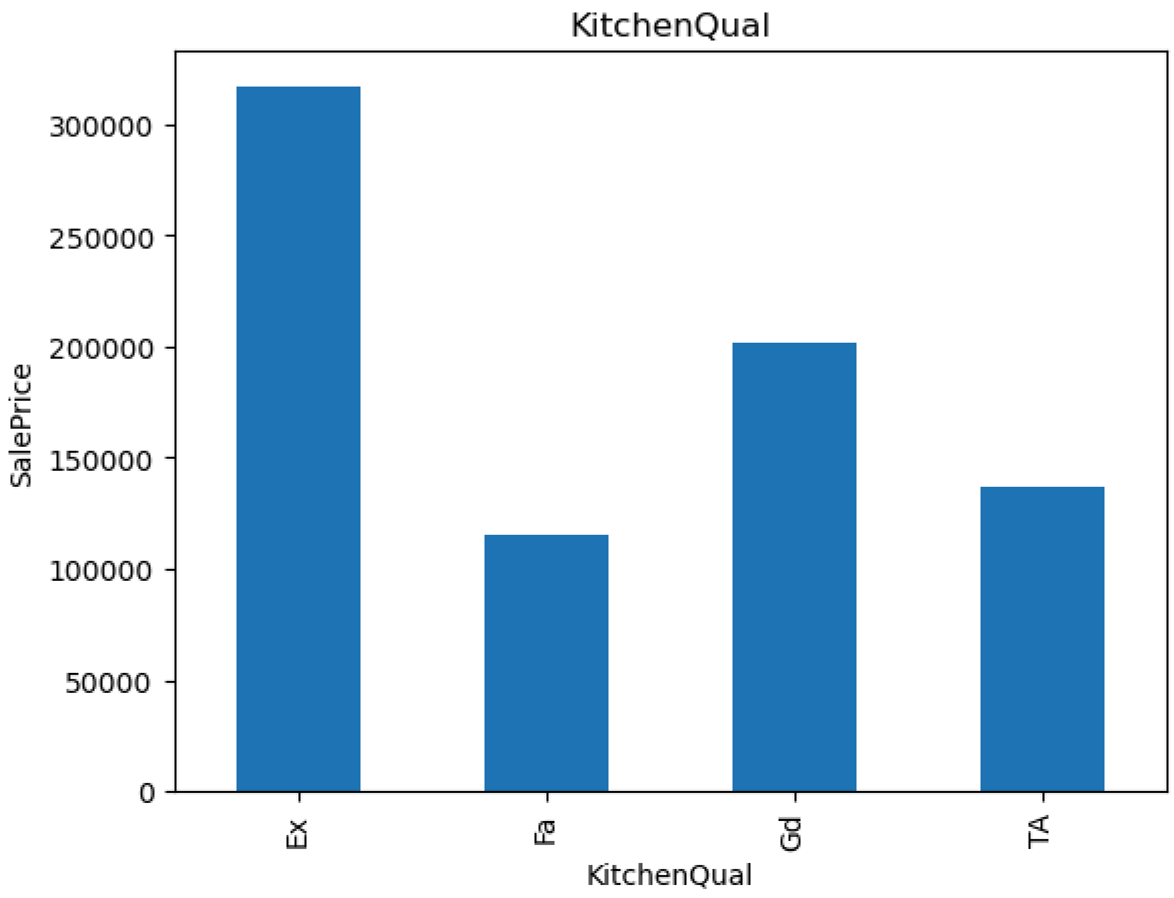

BsmtFinType2
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinType2

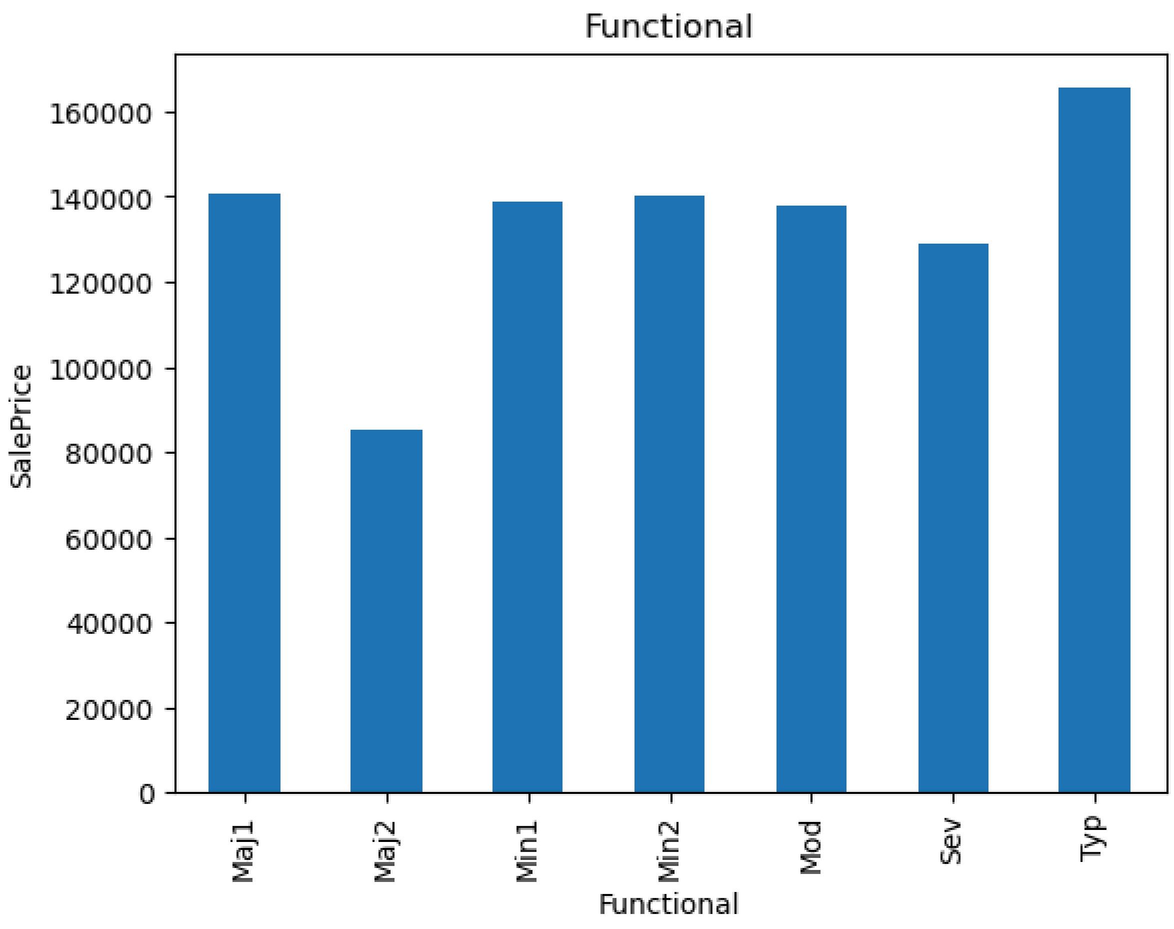

Heating
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
Floor
GasA
GasW
Grav
OthW
Wall
Heating

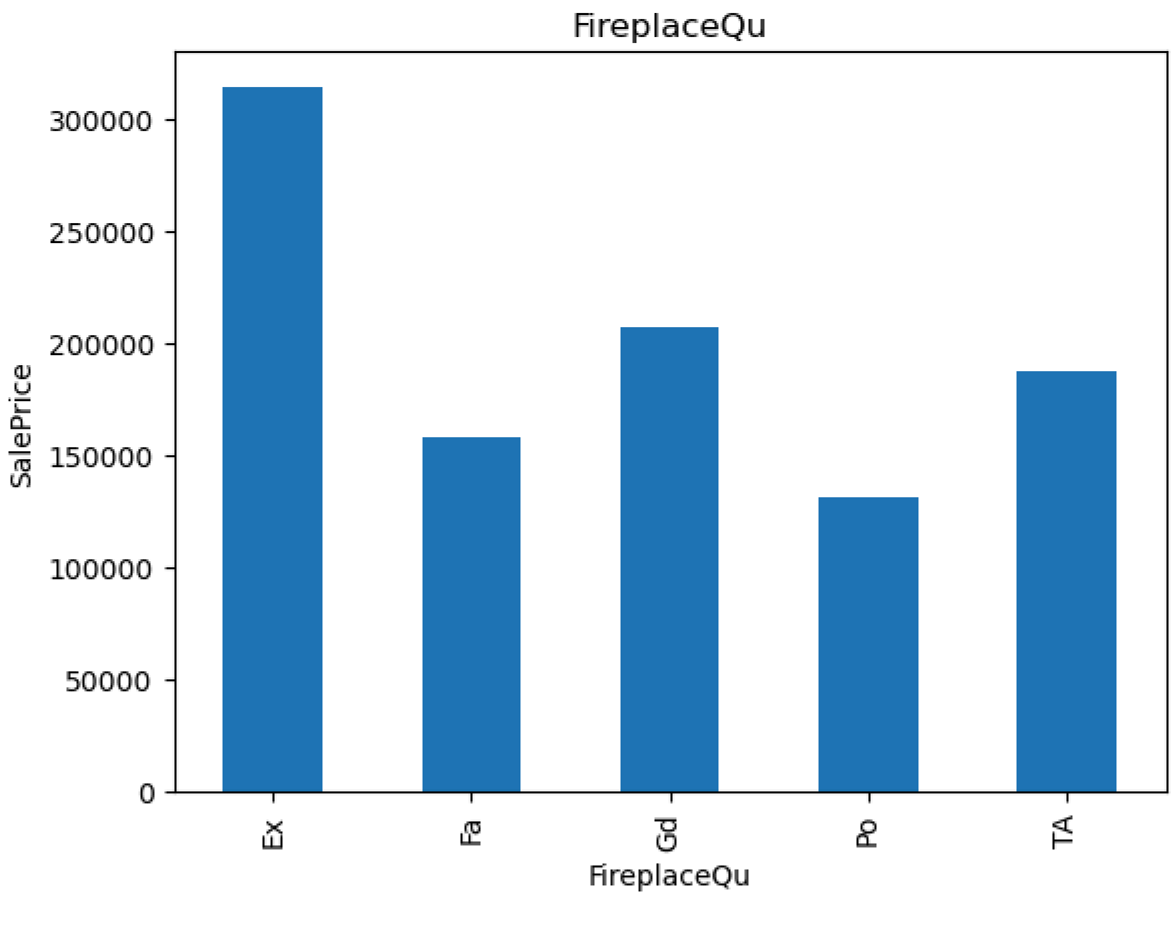

HeatingQC
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Ex
Fa
Gd
Po
TA
HeatingQC

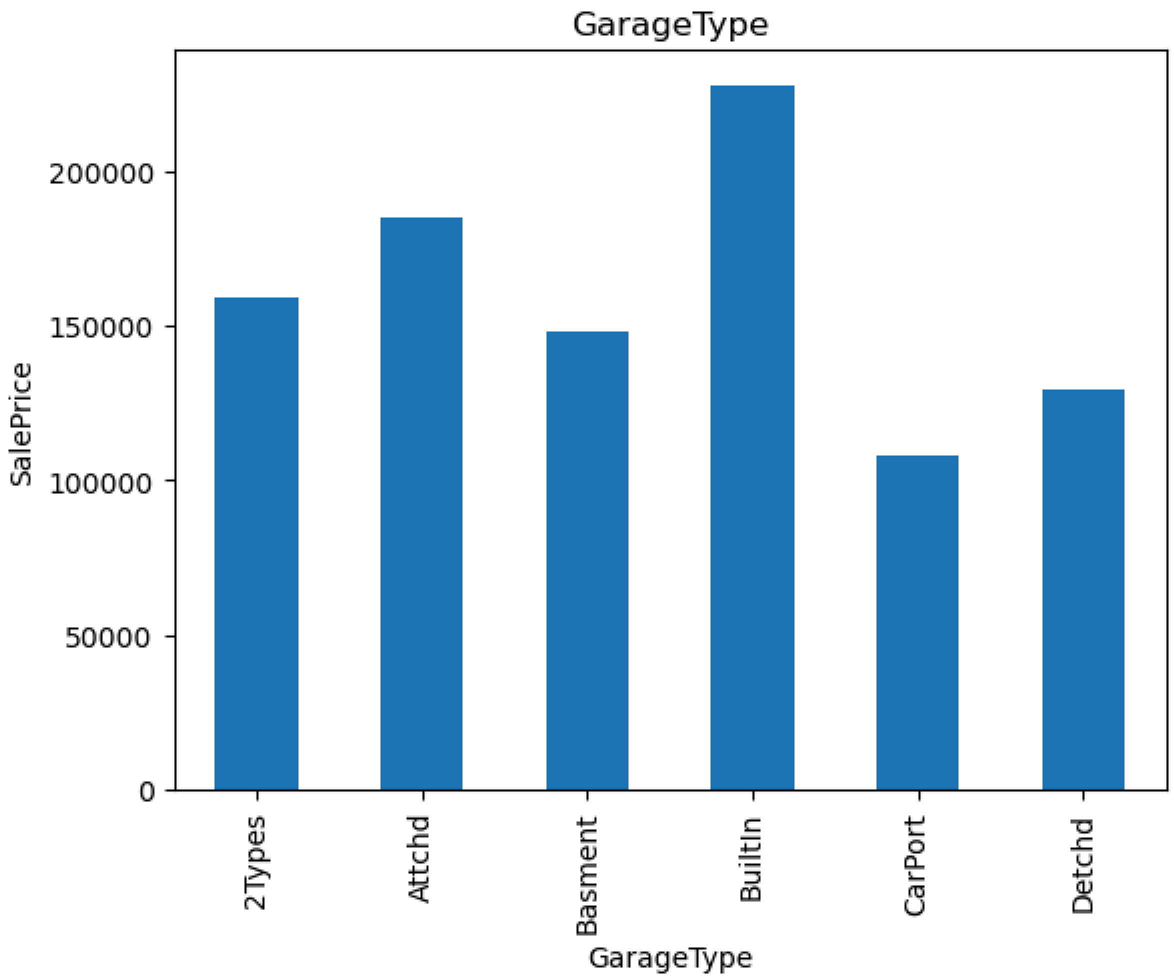

CentralAir
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
N
Y
CentralAir

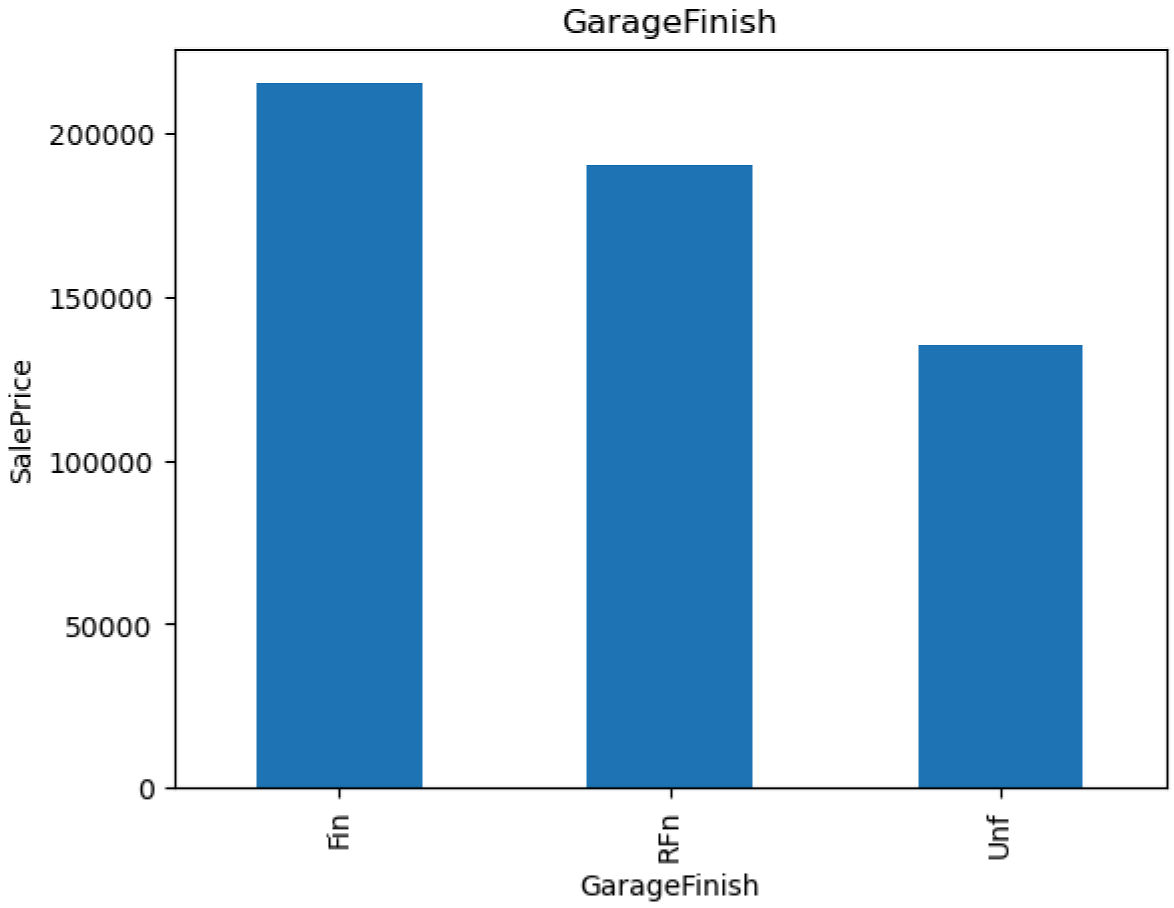

Electrical
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
FuseA
FuseF
FuseP
Mix
SBrkr
Electrical

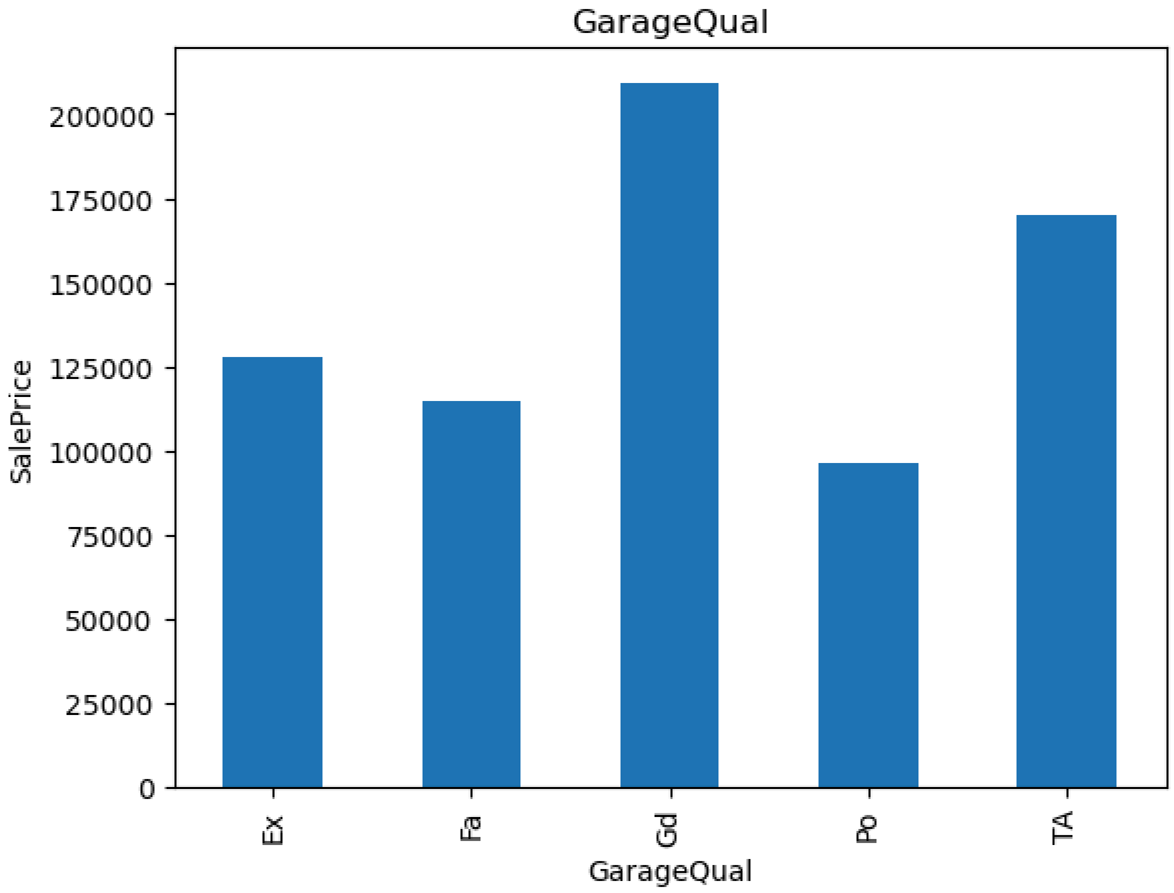

KitchenQual
SalePrice
300000
250000
200000
150000
100000
50000
0
Ex
Fa
Gd
TA
KitchenQual

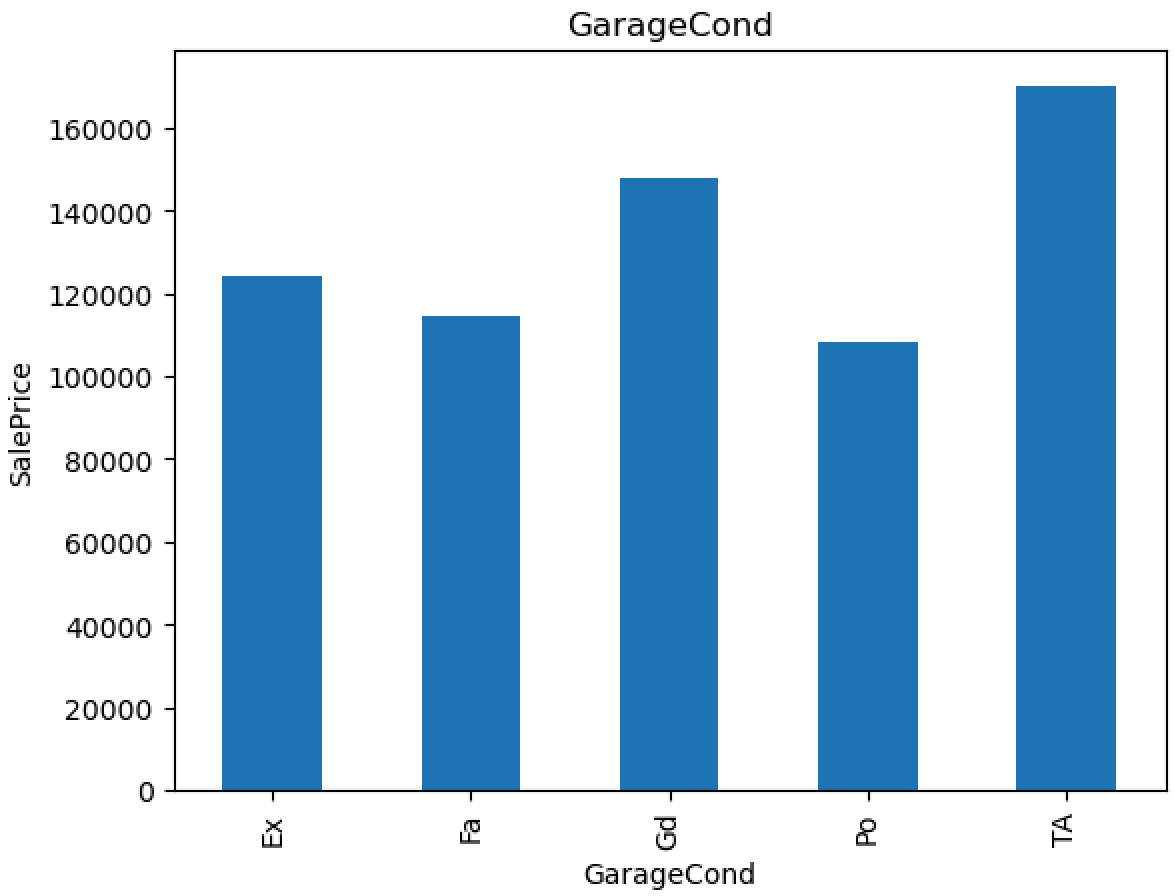
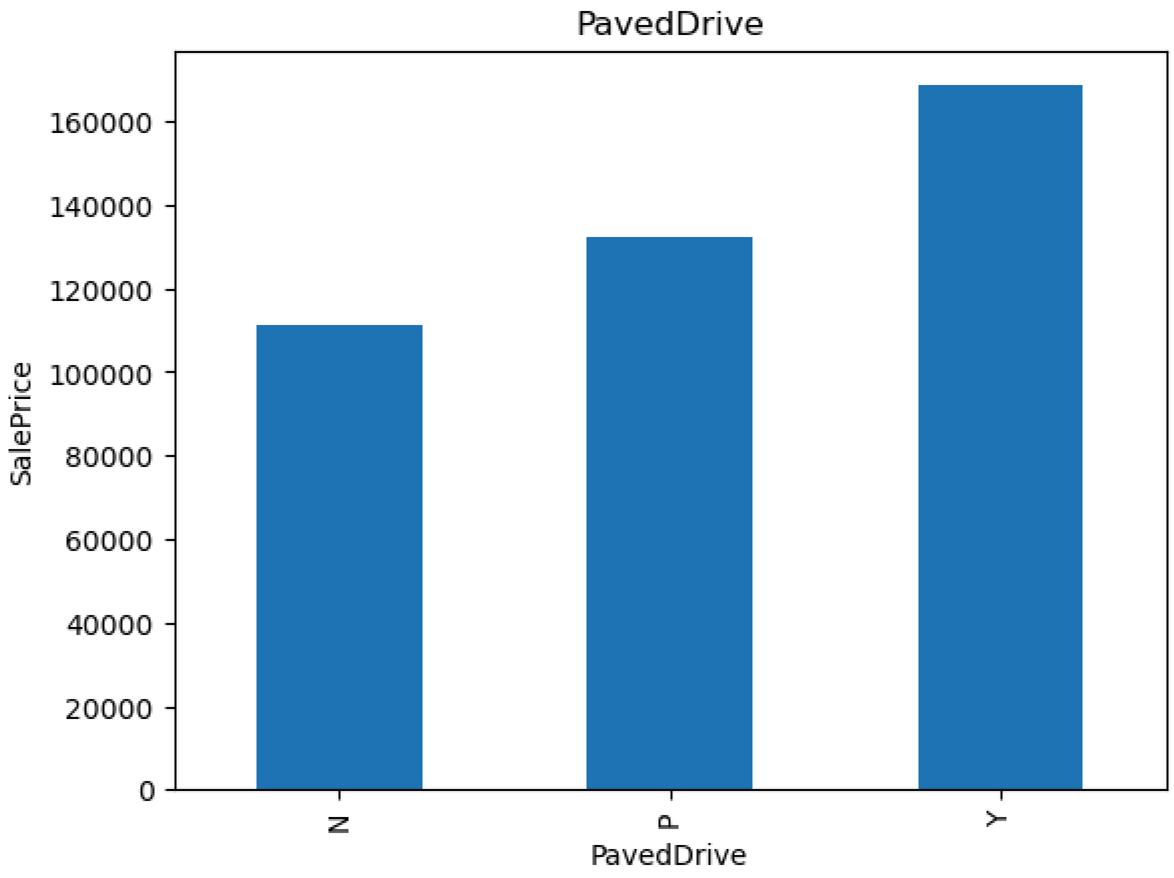

Functional
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ
Functional

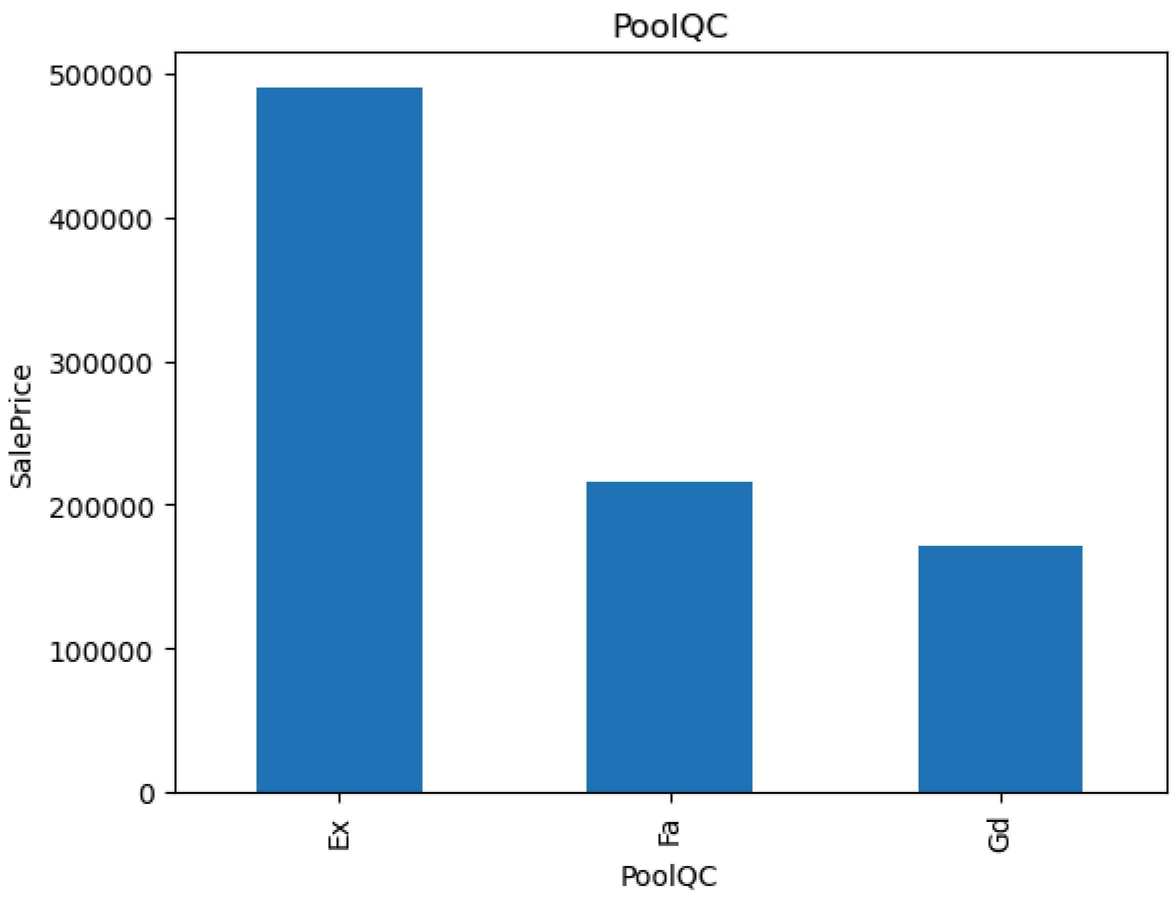

FireplaceQu
SalePrice
0
50000
100000
150000
200000
250000
300000
Ex
Fa
Gd
Po
TA
FireplaceQu

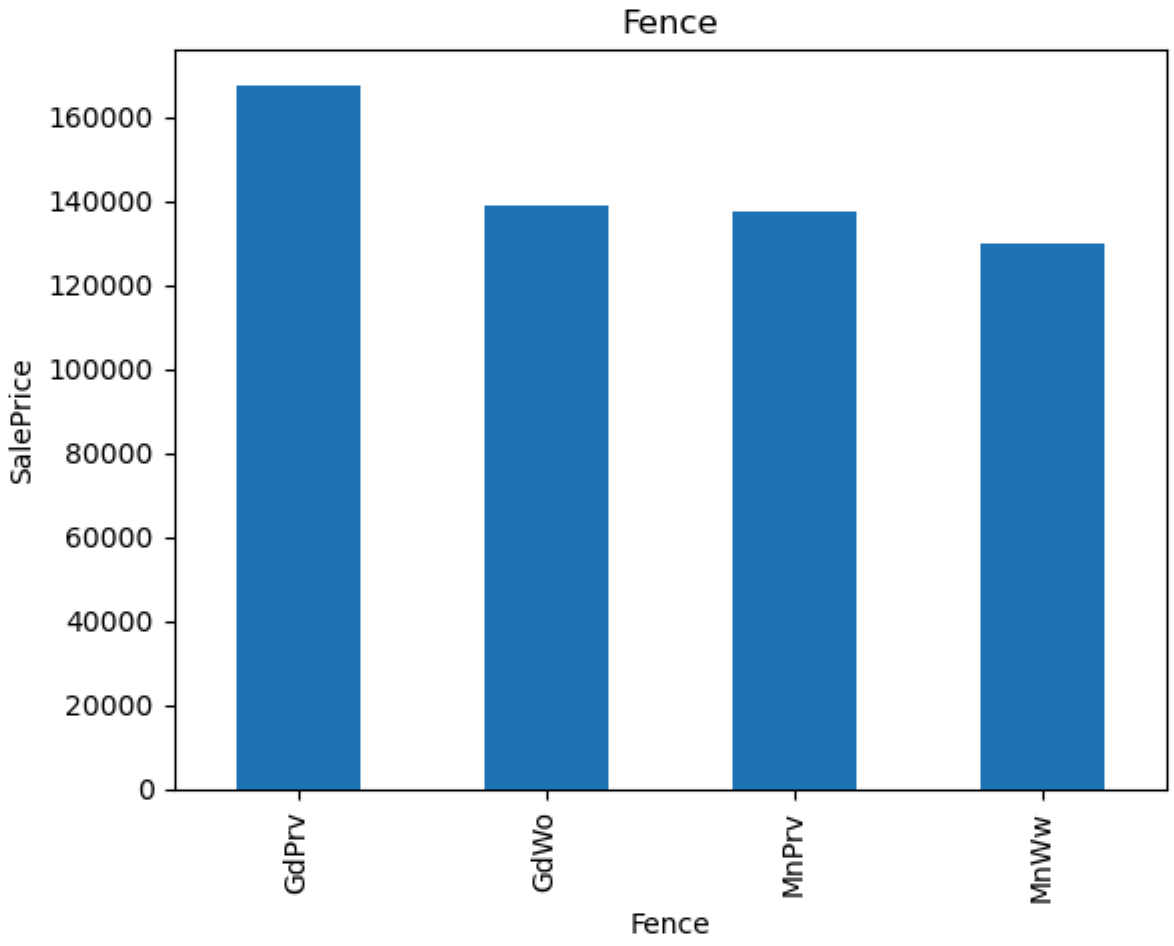

GarageType
SalePrice
200000
150000
100000
50000
0
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
GarageType

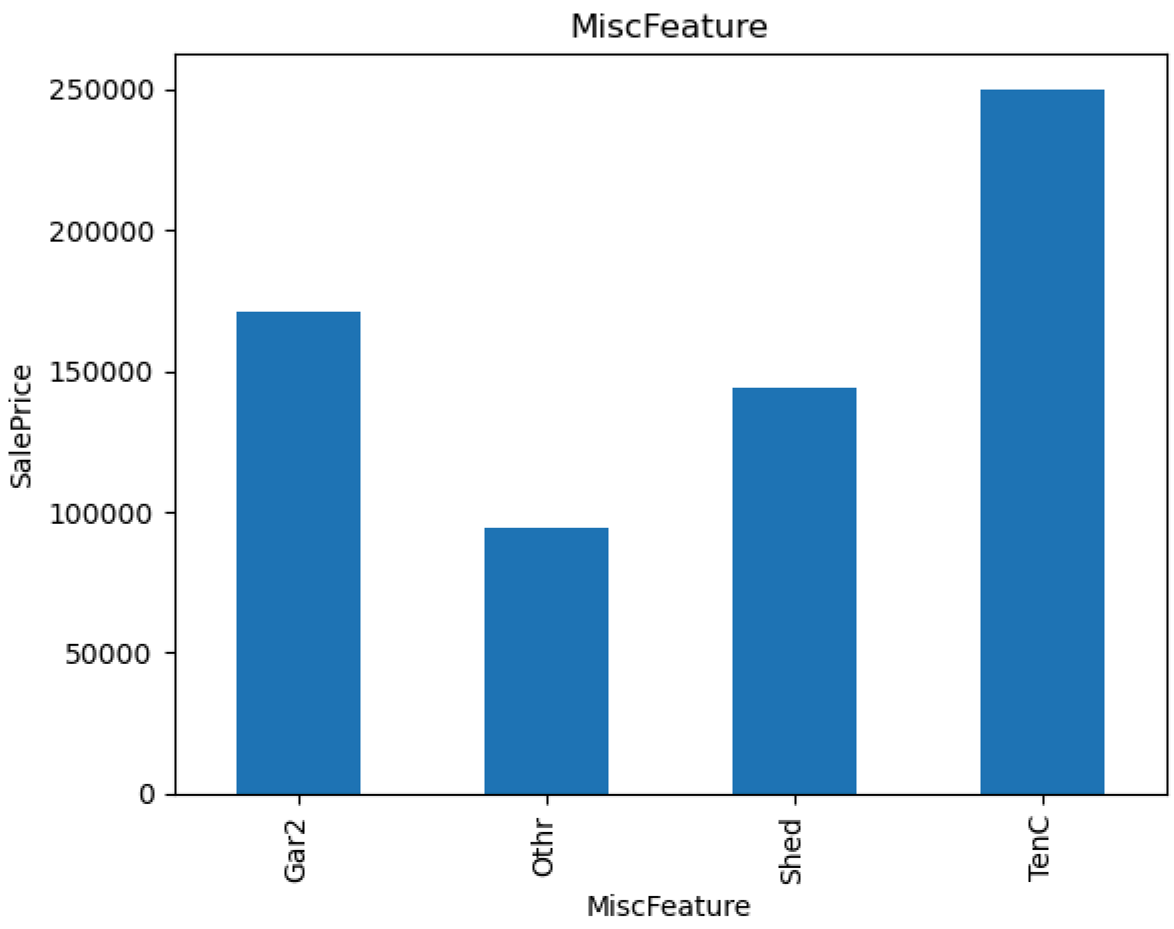

GarageFinish
SalePrice
200000
150000
100000
50000
0
Fin
RFn
Unf
GarageFinish

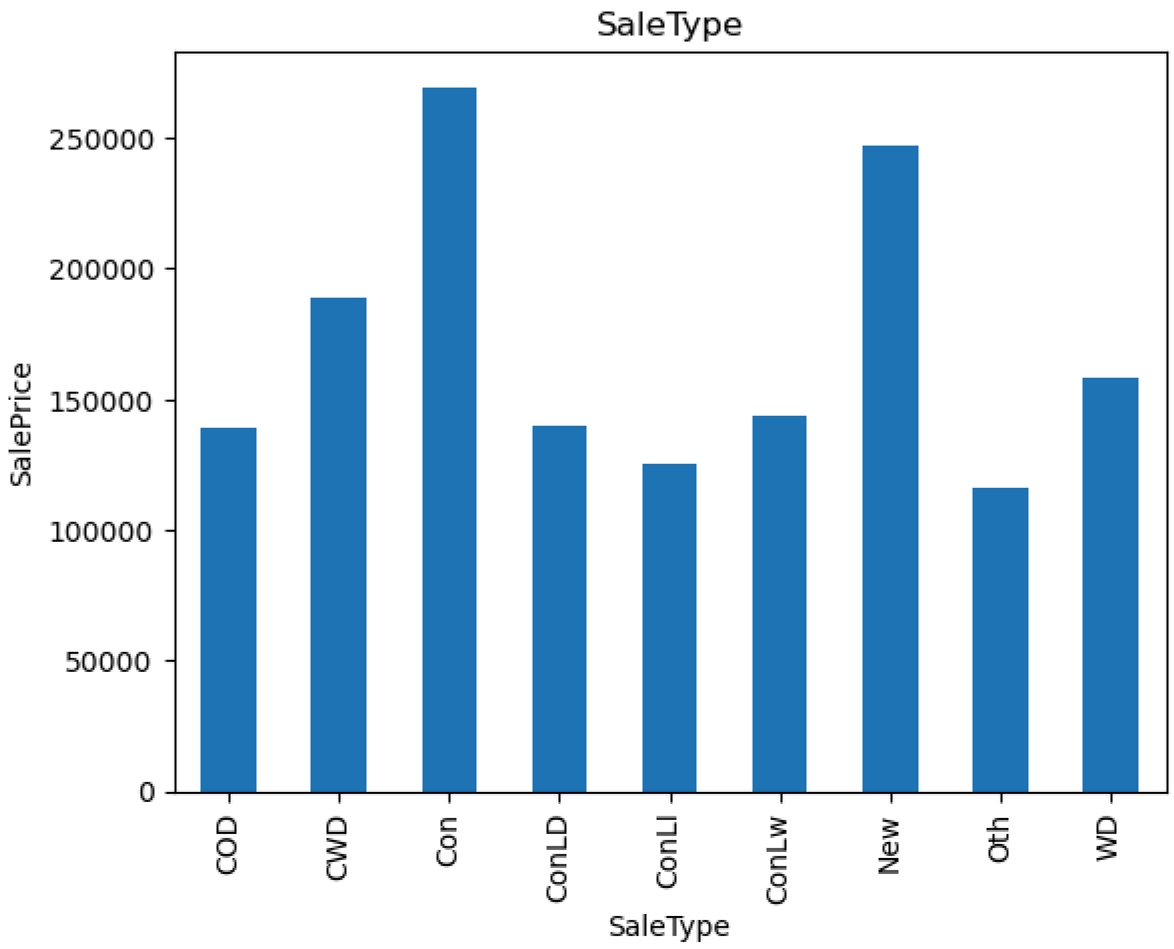

GarageQual
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
Ex
Fa
Gd
Po
TA
GarageQual

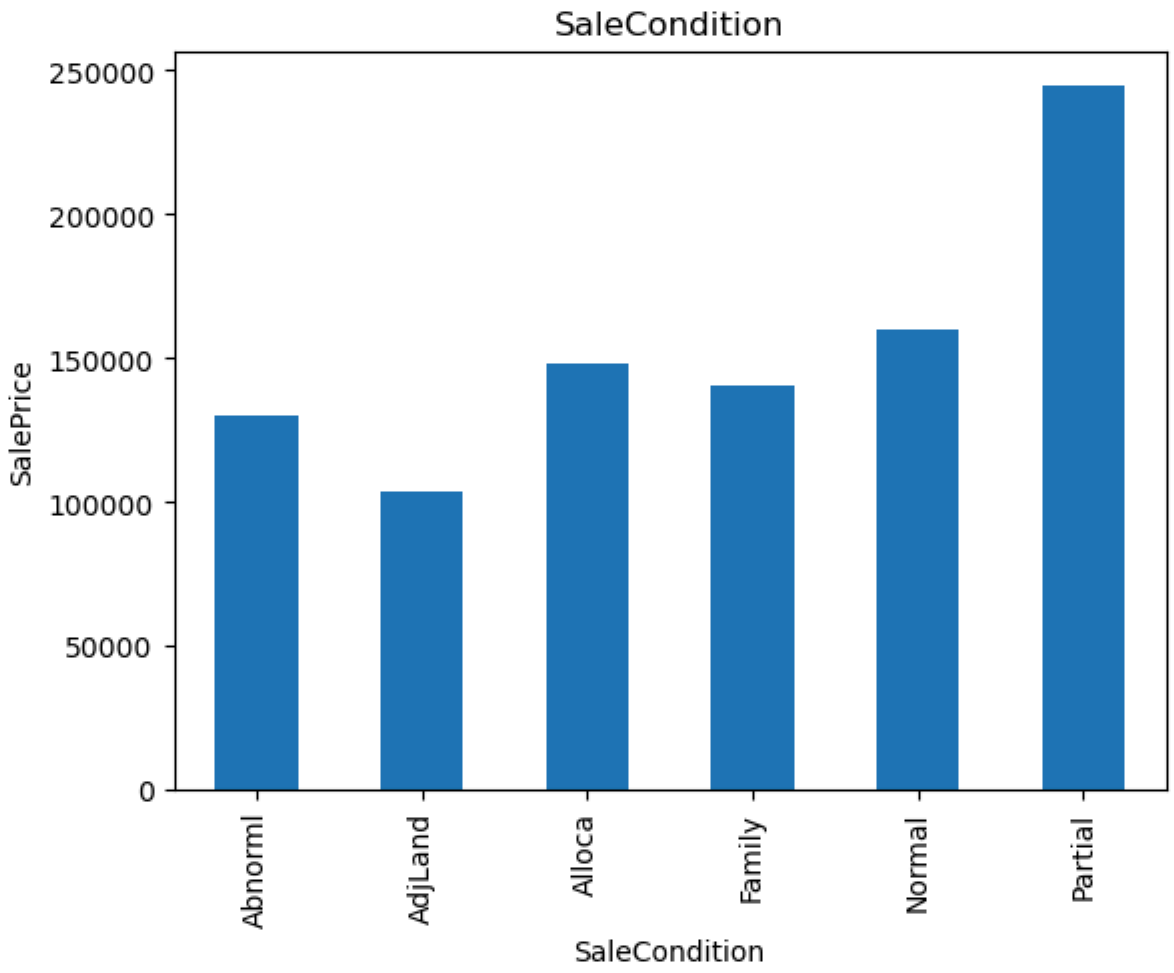

GarageCond
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
Ex
Fa
Gd
Po
TA
GarageCond

PavedDrive
SalePrice
0
20000
40000
60000
80000
100000
120000
140000
160000
N
P
Y
PavedDrive


PoolQC
SalePrice
500000
400000
300000
200000
100000
0
Ex
Fa
Gd
PoolQC


Fence
SalePrice
160000
140000
120000
100000
80000
60000
40000
20000
0
GdPrv
GdWo
MnPrv
MnWw
Fence


MiscFeature
SalePrice
250000
200000
150000
100000
50000
0
Gar2
Othr
Shed
TenC
MiscFeature


SaleType
SalePrice
0
50000
100000
150000
200000
250000
COD
CWD
Con
ConLD
ConLI
ConLw
New
Oth
WD
SaleType


SaleCondition
250000
200000
150000
100000
50000
0
SalePrice
Abnorml
AdjLand
Alloca
Family
Normal
Partial
SaleCondition

```
In [ ]:
```