Removing these columns as there are more than 80% missing values: (code cell 5)

'PoolQC', 'MiscFeature', 'Alley' , 'Fence'

Dropping these columns as there is a huge disparity in the category count: (code cell 7)

'Utilities', 'Street', 'Condition2', 'RoofMatl', 'Heating'

Based on domain knowledge these columns are most important: (code cell 8)

'OverallQual' ,'GrLivArea', 'GarageCars', 'TotalBsmtSF', '1stFlrSF' ,'YearBuilt', 'FullBath', 'TotRmsAbvGrd' ,'GarageArea', 'Foundation', 'ExterCond' ,'ExterQual' , 'Neighborhood', 'KitchenQual' ,'BsmtQual', 'BsmtCond', 'BsmtExposure', 'SalePrice','LotArea','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3Ssn Porch','ScreenPorch','MiscVal','PoolArea','SaleType','Functional'

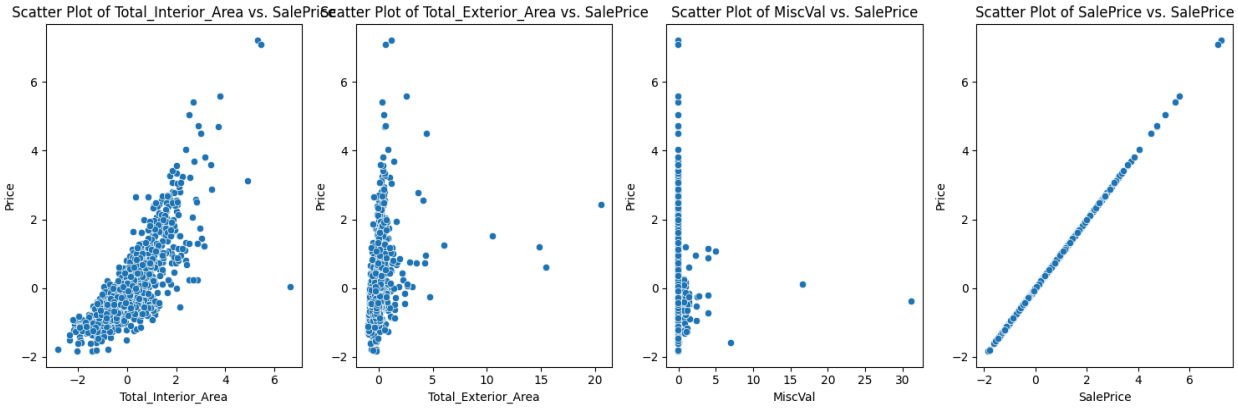
combining similar values into one column: (code cell 20)

```
df['Total_Interior_Area'] = df[['GrLivArea', 'TotalBsmtSF']].sum(axis=1)
df['Total_Exterior_Area'] = df[['GarageArea', 'LotArea','OpenPorchSF','EnclosedPorch','3SsnPorch','ScreenPorch','PoolArea']].sum(axis=1)
```

All these columns are absent at the same time, and non of these column have ‘NA” values which according to metadata means no basement. Thus i make a judgement call to impute all of these with ‘NA’ (code cell 17)

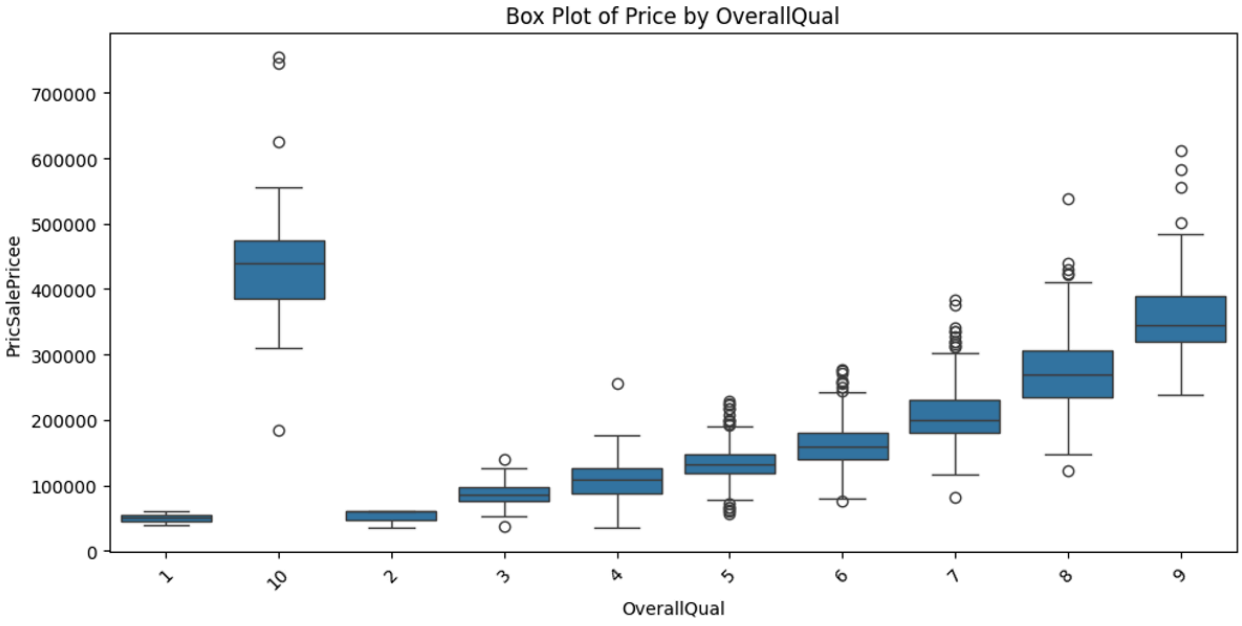
	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinType2
17	NaN	NaN	NaN	NaN	NaN
39	NaN	NaN	NaN	NaN	NaN
90	NaN	NaN	NaN	NaN	NaN
102	NaN	NaN	NaN	NaN	NaN
156	NaN	NaN	NaN	NaN	NaN
182	NaN	NaN	NaN	NaN	NaN
259	NaN	NaN	NaN	NaN	NaN
342	NaN	NaN	NaN	NaN	NaN
362	NaN	NaN	NaN	NaN	NaN
371	NaN	NaN	NaN	NaN	NaN
392	NaN	NaN	NaN	NaN	NaN
520	NaN	NaN	NaN	NaN	NaN
532	NaN	NaN	NaN	NaN	NaN
533	NaN	NaN	NaN	NaN	NaN
553	NaN	NaN	NaN	NaN	NaN

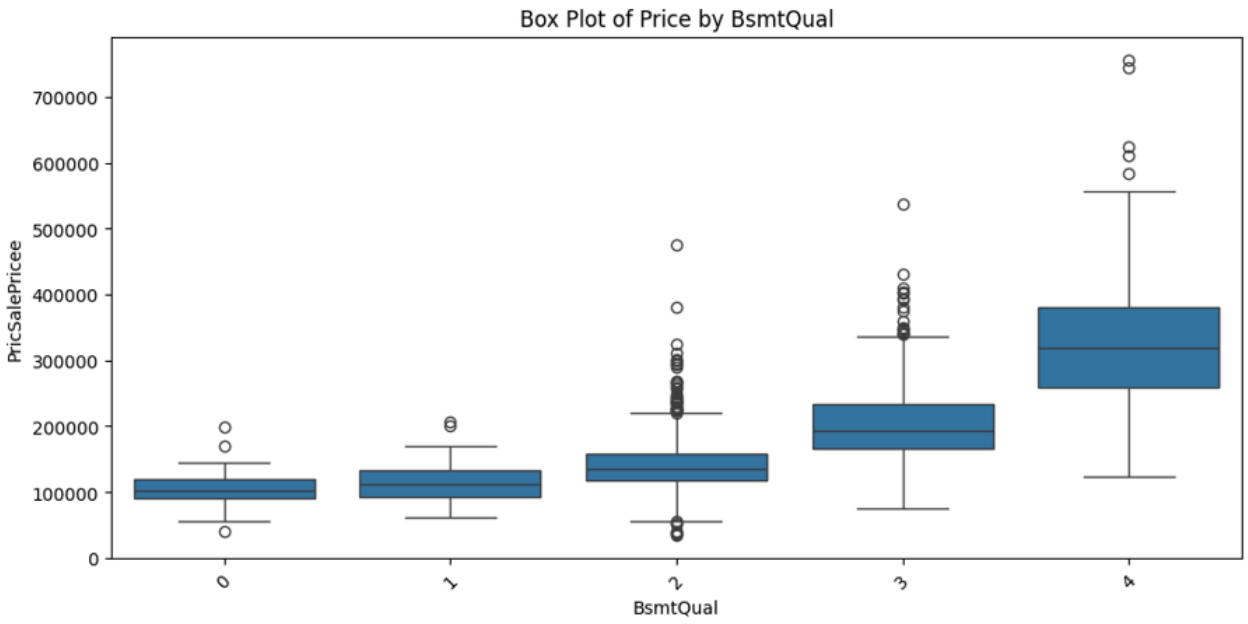
There is a high corelation of these columns [total interior area] with price. (code cell 30)

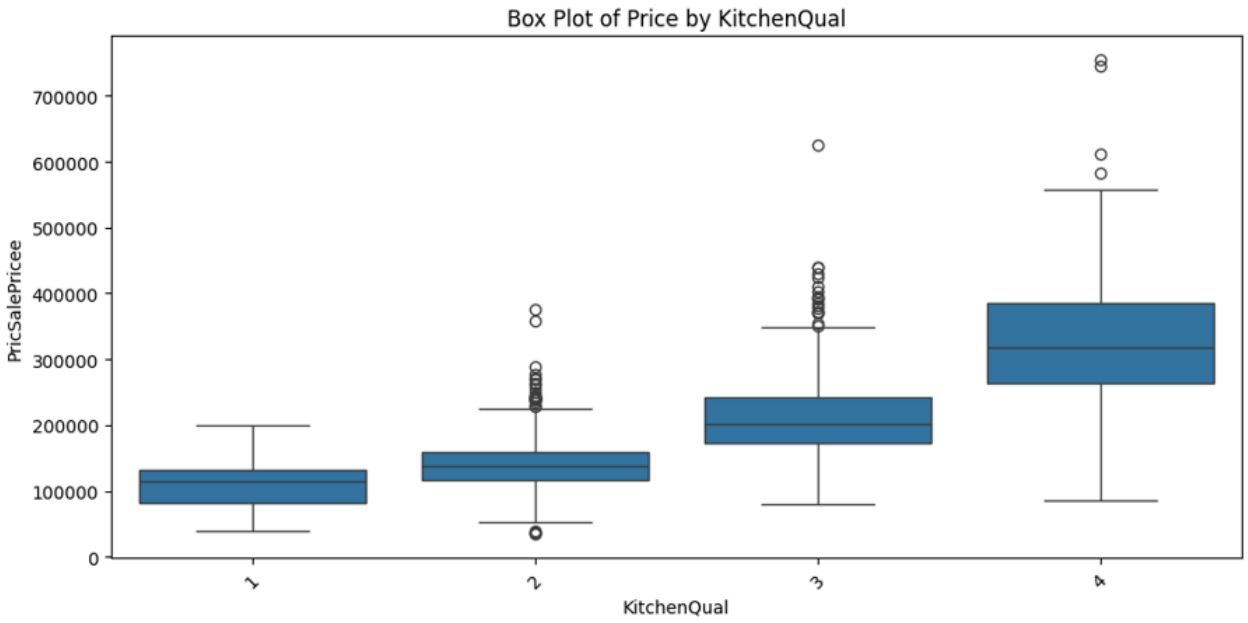


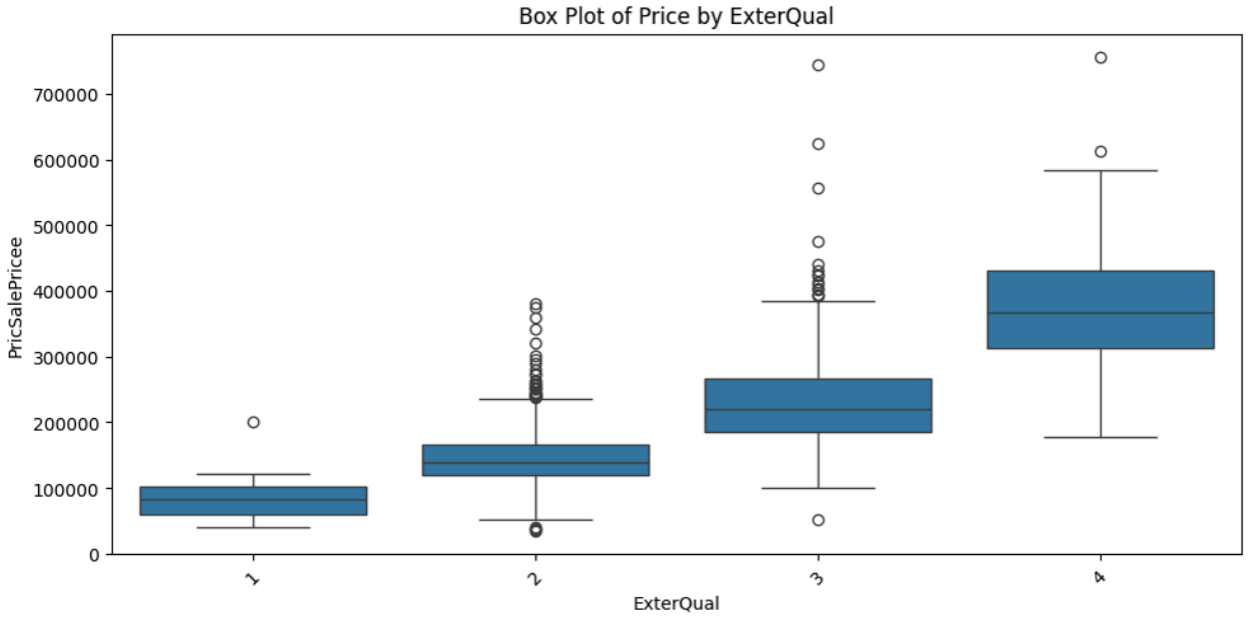
This makes sense as total interior area is defenetly gonna increase the price.
But its counter intuitive that total exterior area and micVal is not corealtied with price

here is a clear increase in mean price for the quality columns Overalqual, kitchenqual, basementqual, exterqual. (code cell 34)




Box Plot of Price by BsmtQual
PricSalePricee
0
100000
200000
300000
400000
500000
600000
700000
0
1
2
3
4
BsmtQual

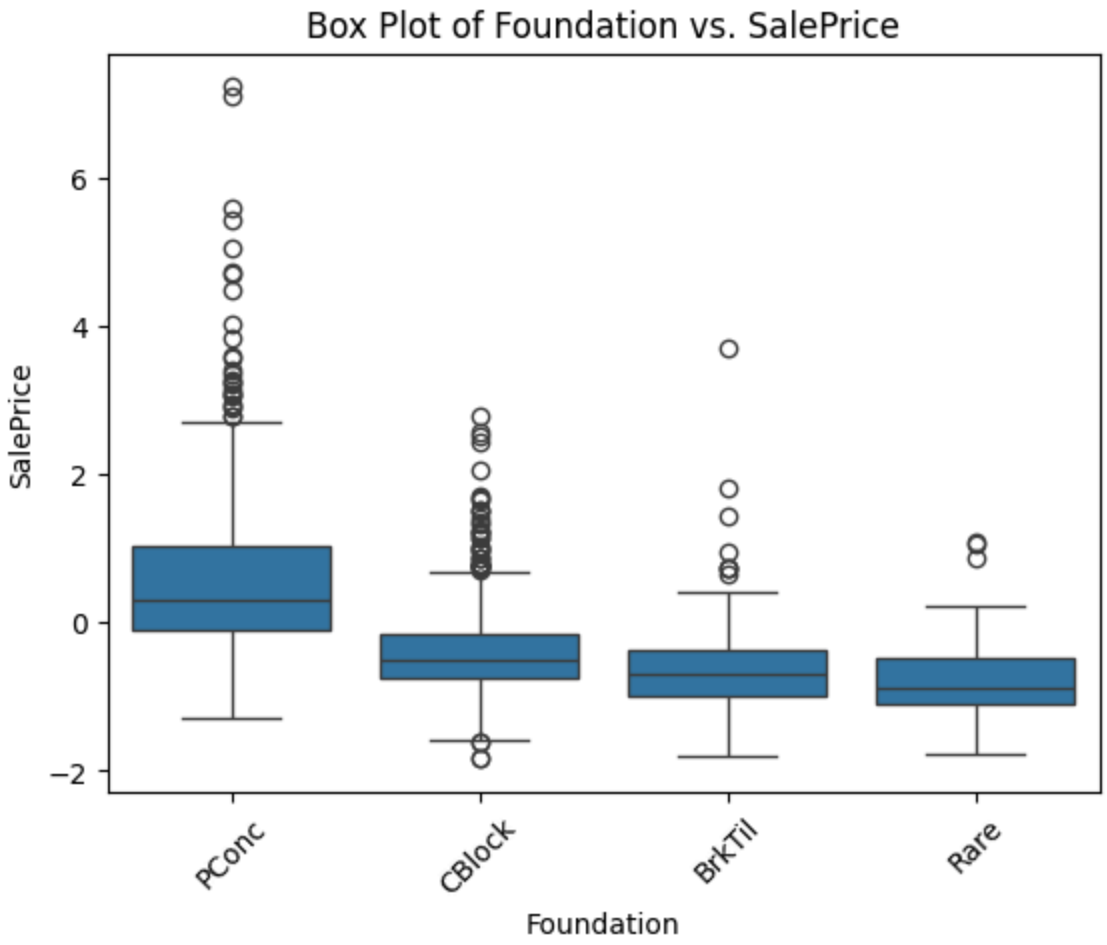

Box Plot of Price by KitchenQual
PricSalePricee
0
100000
200000
300000
400000
500000
600000
700000
1
2
3
4
KitchenQual



This finding makes sense as quality of stuff directly influences the price.

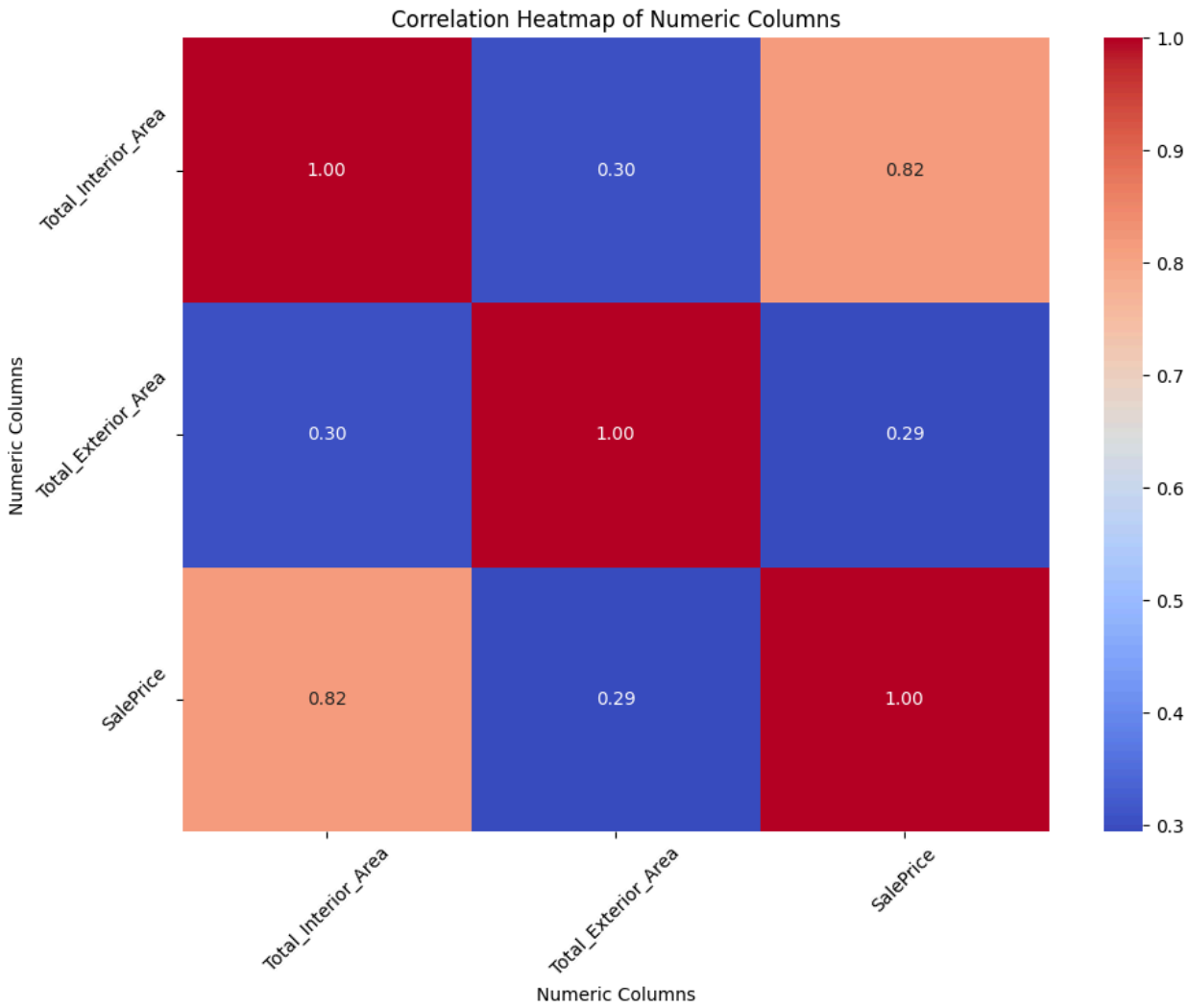
Boxplot of price and foundation: (code cell 30)

there is a pattern in the median.



It makes sense that poured concrete is costlier than brick and tiles.
Its counter intuitive that rare foundations (like wood stone) are low in price, as rare foundations are supposed to be in highly priced houses.

Corelation heatmap (code cell 37)



There is a high corelation between total interior area and sale price.

Feature selection

Based on eda.
['Total_Interior_Area', 'SalePrice', 'Foundation',
'SaleType', 'OverallQual', 'BsmtQual', 'KitchenQual', 'ExterQual'],

This makes sense on domain knowledge as total area must be correlated with price,\
sale type helps show that new house are costlier
All quality columns must be related with rice.