Housing Price Prediction

Submitted By :

ASHUTOSH CHAUDHARY

ACKNOWLEDGMENT

I would like to express my gratitude towards my internship mentor Ms. Srishti Maan for helping me in completion of the project.

BUSINESS PROBLEM

From my understanding the problem is about making prediction of house prices for the company which is about to buy that property & to help make decision that whether it would be profitable or not to buy it on the basis of other variables given so that we could make this venture more profitable for the company.

OBJECTIVE FOR PROBLEM UNDERTAKEN

We have to study every feature's behavious present in dataset & make observation from its behaviour about how every feature is giving the signs about that whether it would be profitable to buy the property or not & building predictive model on the basis of those features information to reduce the loss & increase the profit of the company.

ANALYTICAL PROBLEM FRAMING

- **Origin of dataset & data types of every features**

Dataset is provided by the company & we have to import it using various libraries necessary for the project to get completed. Also data types of features are both continuous & categorical.

- **Mathematical/Analytical modelling of the problem**

For visualization we only use four plots most of the times that were countplot, boxplot, distplot & scatterplot & for model building we use Linear Regression, Decision Tree Regression, Random Forest Regression & Bagging Regression models to opt best out of them to work on dataset.

- **Assumptions related to problem statement**

No assumptions were made while working on the dataset.

- **Libraries & Tools used**

We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warnings libraries for this task.

STEPS TAKEN FOR THE TASK

1. **Importing Libraries for the task**

   Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickles & warnings were imported for task to get completed.

2. **Importing Dataset using libraries**

   Imported the datasets using pandas library in jupyter notebook.

3. **Checking Dimension of dataset**

   By checking dimension of train dataset we get to know that it contains 1161 rows & 81 columns & for test dataset we get to know that it contains 292 rows & 80 columns.

4. **Checking Description of dataset**

   From description we find mean, min value, max value, etc of every column which contains continuous data in them

5. **Checking for presence of null values in dataset**

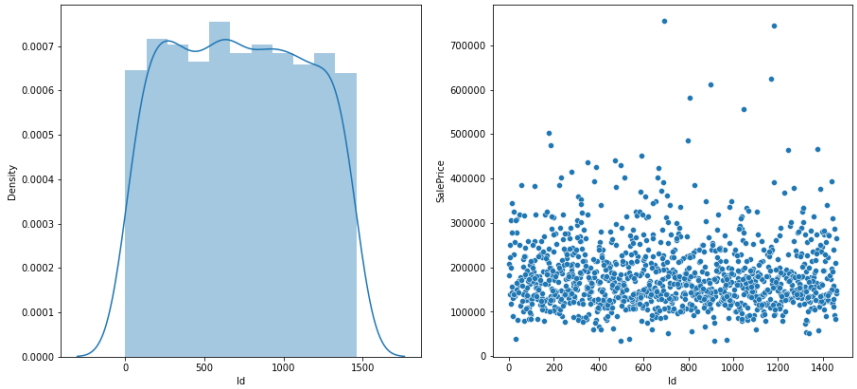
   We checked for the presence of null values in every column of dataset by doing it repeatedly for every column as doing it repeatedly for every column & null values were present. Then we filled them randomly to avoid data bias.

6. Identifying Target variable

By looking at dataset we identified target variable which is named SalePrice.

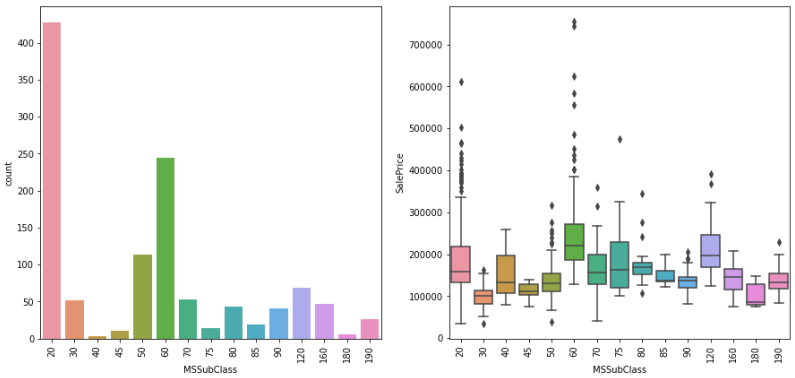
7. Performing EDA on whole Train dataset

Id



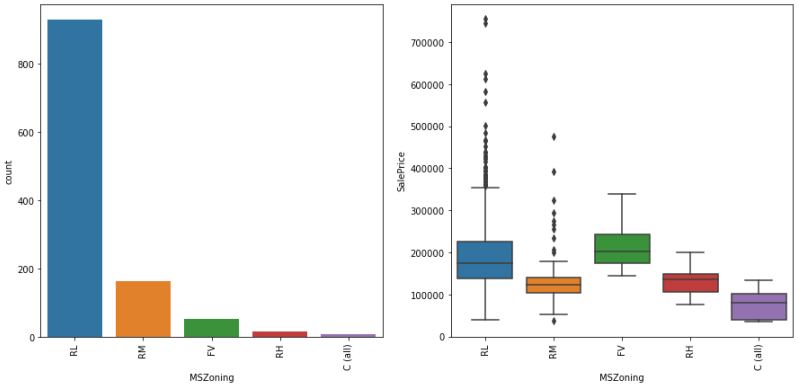
From visualization we conclude that could not determine the skewness of data with data having highest density at 600 & we could not determine the correlation of the data with target column (SalePrice) as data is scattered all over the place.

MSSubClass



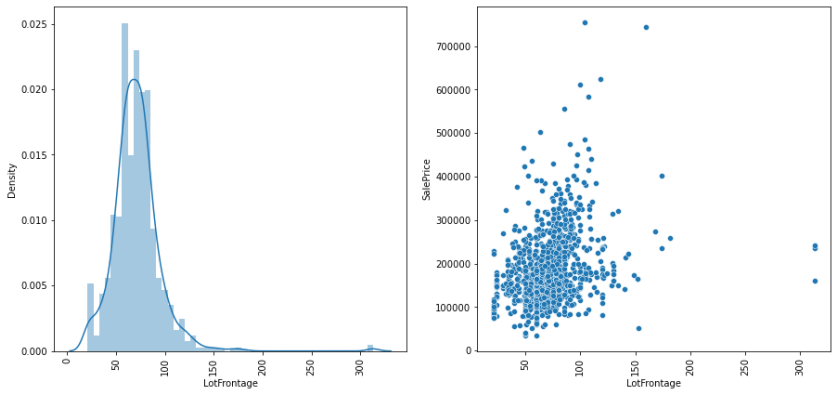
From visualization we concluded that most data is present in category 20 of MSSubClass & outliers are present in 10 categories of MSSubClass out of total 15 categories of MSSubclass when plotted against SalePrice.

MSZoning



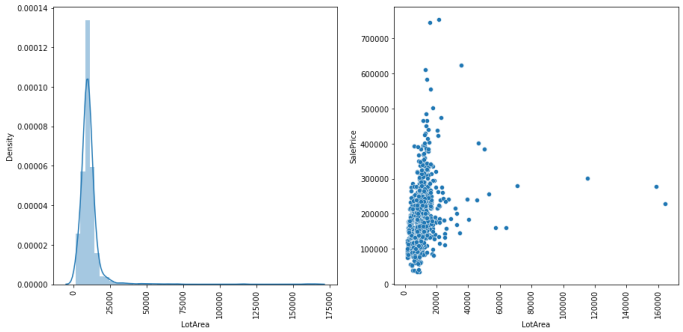
From visualization we concluded that most data is present in RL category & outliers of SalePrice are present in only 2 categories of MSZoning out of total 5 categories.

LotFrontage



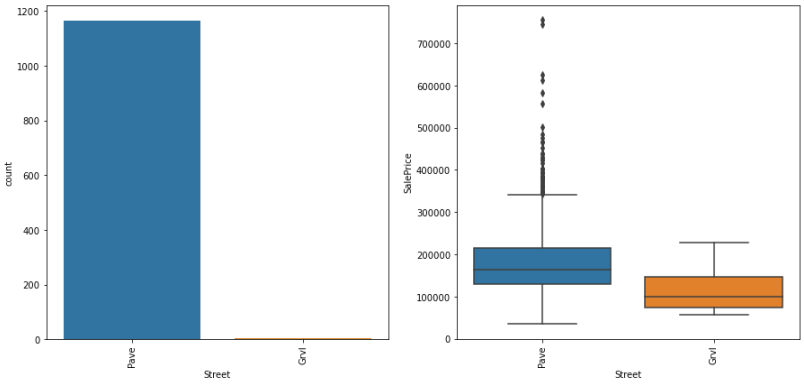
From visualization we concluded that data is positively skewed with data having highest density at around 60 & data is showing somewhat positive correlation with SalePrice with data being concentrated mostly under 150 range.

LotArea



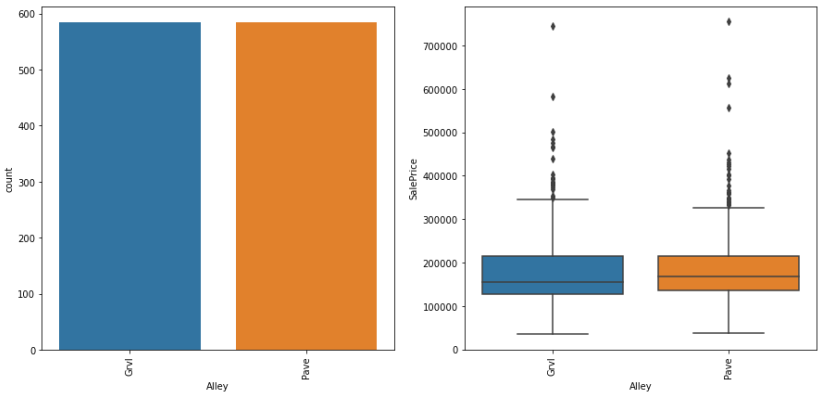
From visualization we concluded that data is showing positive skewness with data having highest density at around 2000 & data is showing somewhat positive correlation with SalePrice with data being concentrated mostly under range of 300000 of LotArea.

Street



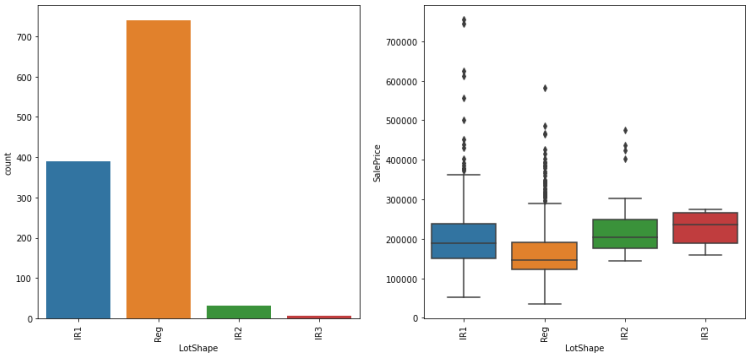
From visualization we concluded that most data is present in pave category & also outliers of SalePrice are present in only pave category of Street.

Alley



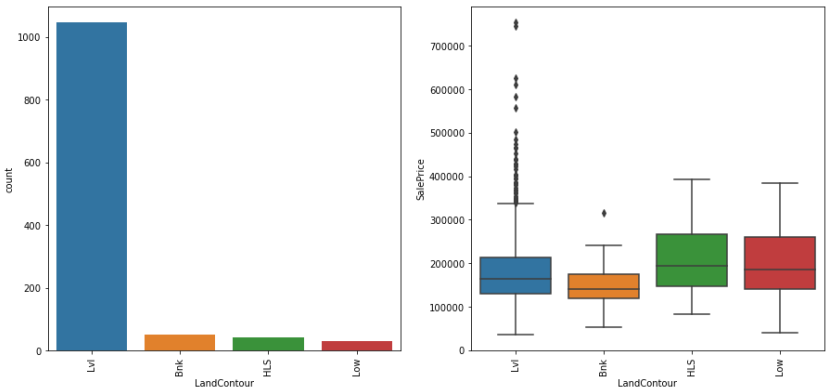
From visualization we concluded that data is present equally in both categories & outliers of SalePrice are present in both categories of Alley.

LotShape



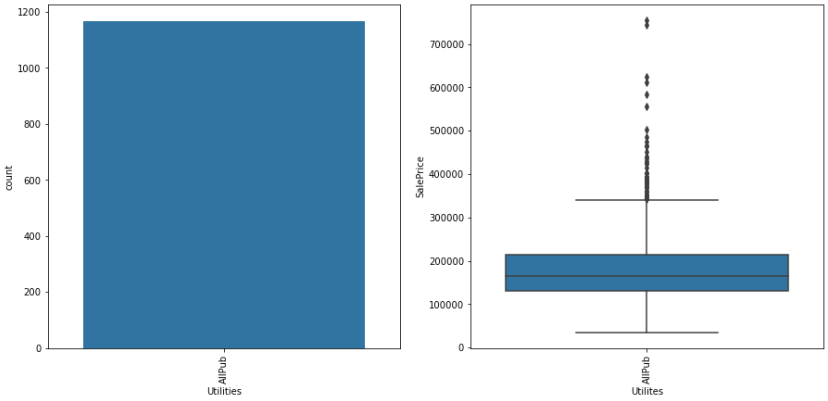
From visualization we concluded that most data is present in Reg category of LotShape & outliers of SalePrice are present in 3 categories of LotShape out of total 4 categories of LotShape.

LandContour



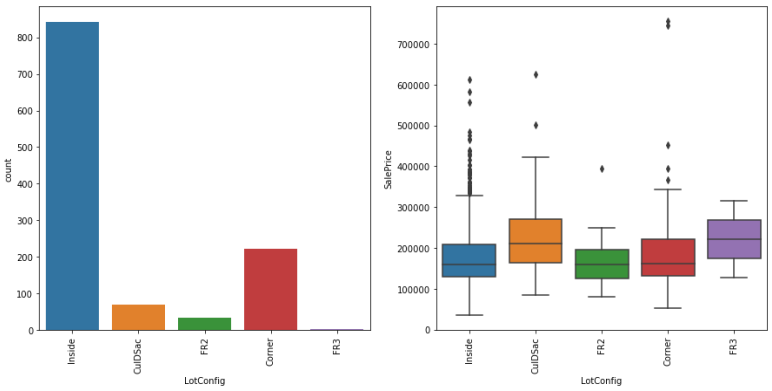
From visualization we concluded that most data is present in lvl category of LandContour & outliers of SalePrice are present in 2 categories of LandContour out of total 4 categories of LandContour.

Utilities



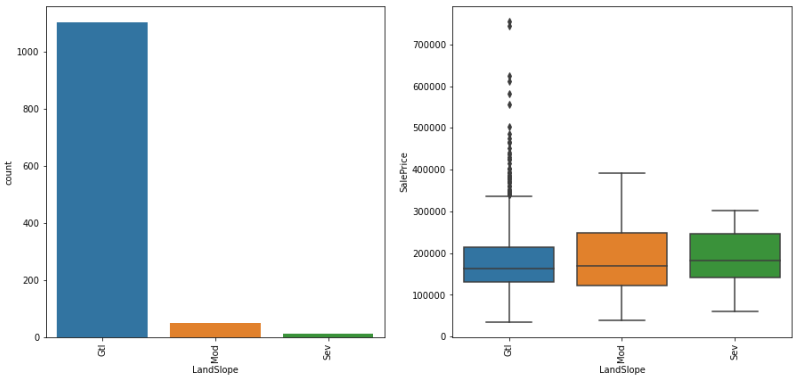
From visualization we concluded that only one values is present in every row of whole column & outliers of SalePrice are present in it.

LotConfig



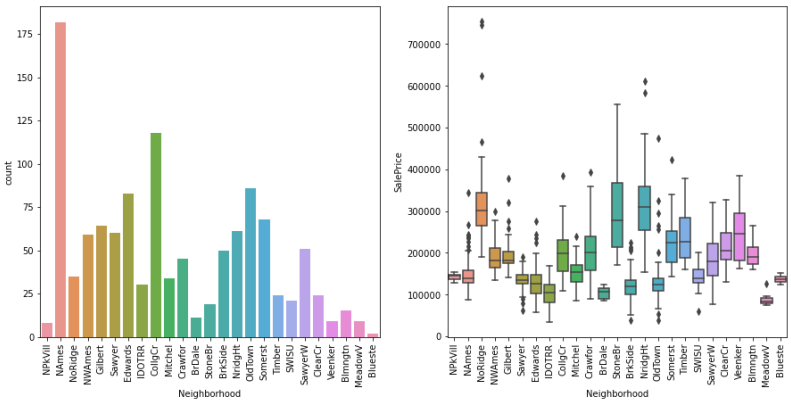
From visualization we concluded that most data is present in Inside category of LotConfig & outliers of SalePrice are present in 4 categories of LotConfig out of total 5 categories of LotConfig.

LandSlope



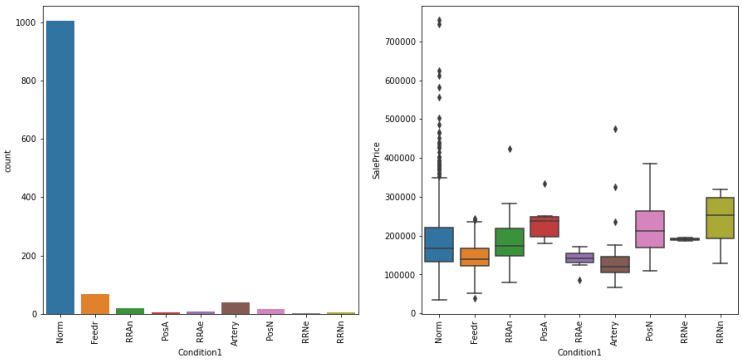
From visualization we concluded that most data is present in Gtl category of LandSlope & outliers of SalePrice is present in only Gtl category of LandSlope out of total 3 categories of LandSlope.

Neighborhood



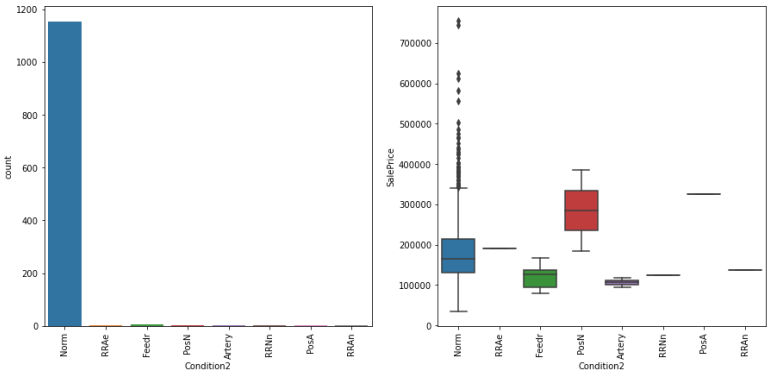
From visualization we concluded that most data is present in NAmes category of Neighborhood & outliers of SalePrice are present in 14 categories of Neighborhood out of total 25 categories of Neighborhood.

Condition1



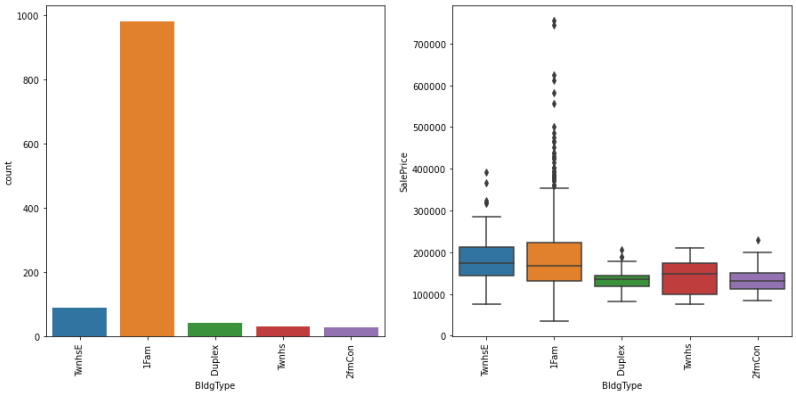
From visualization we concluded that most data is present in Norm category of Condition1 & Outliers of SalePrice are present in every category of Condition1 except 1 category.

Condition2



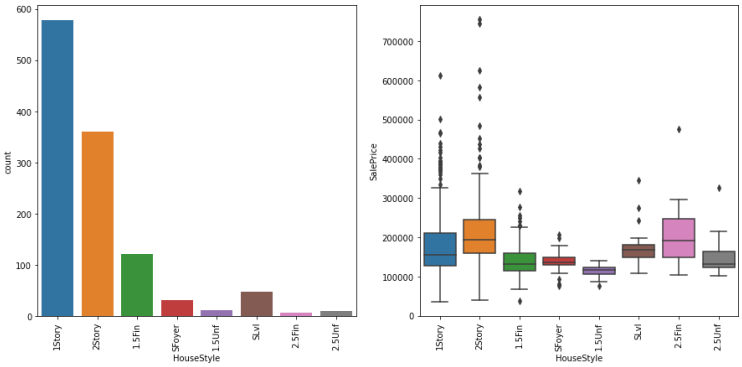
From visualization we concluded that most data is present in Norm category of Condition2 & outliers of SalePrice are present in only Norm Category of Condition2 out of total 8 categories of Condition2.

BldgType



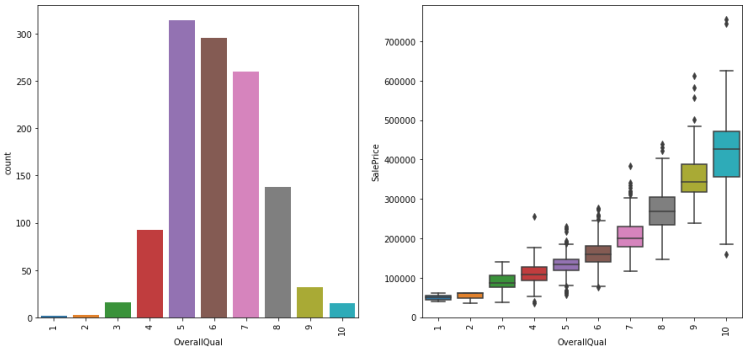
From visualization we concluded that most data is present in 1Fam category of BldgType & Outliers of SalePrice are present in every category of BldgType except 1 category which is Twnhs.

HouseStyle



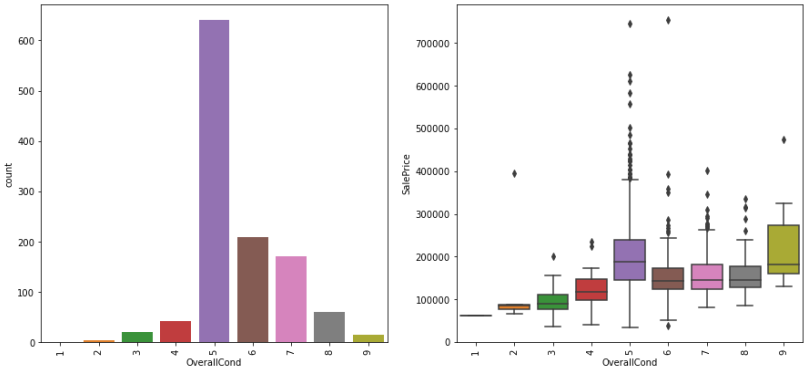
From visualization we concluded that most data is present in 1Story category of houseStyle & Outliers of SalePrice are present in every category of HouseStyle.

OverallQual



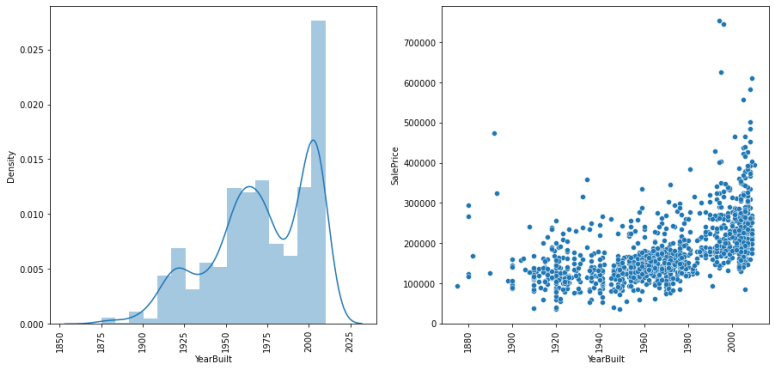
From visualization we concluded that most data is present in category 5 of OverallQual & outliers of SalePrice are present in 7 categories of OverallQual out of total 10 categories of OverallQual.

OverallCond



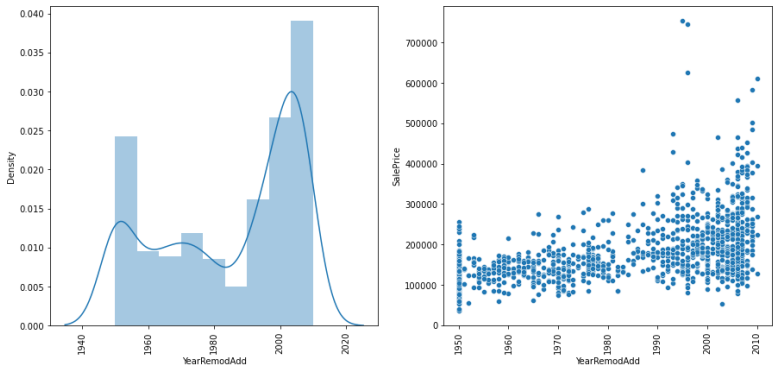
From visualization we concluded that most data is present in category 5 of OverallCond & outliers are present in 8 categories of OverallCond out of total 9 categories of OverallCond.

YearBuilt



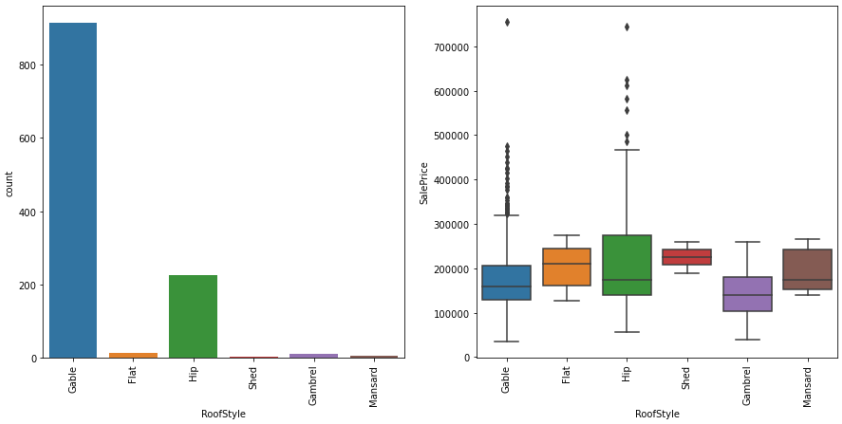
From visualization we concluded that data is negatively skewed with having highest density at around 2015 & data is somewhat positively correlated with SalePrice.

YearRemodAdd



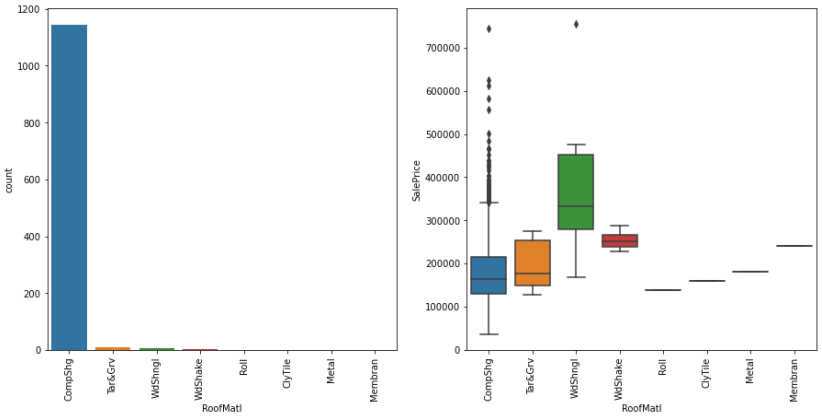
From visualization we concluded that we cannot determine skewness of data with having highest density at around 2010 & data being spread all over the plot equally so correlation cannot be determine with SalePrice.

RoofStyle



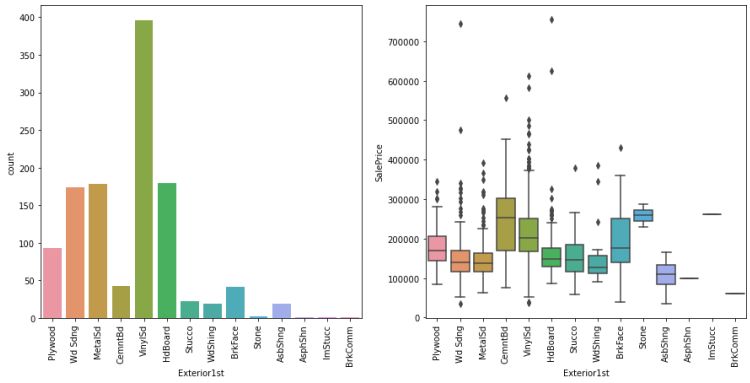
From visualization we concluded that most data is present in Gable category of RoofStyle & outliers of SalePrice are present in only 2 categories of RoofStyle out of total 6 categories of RoofStyle.

RoofMatl



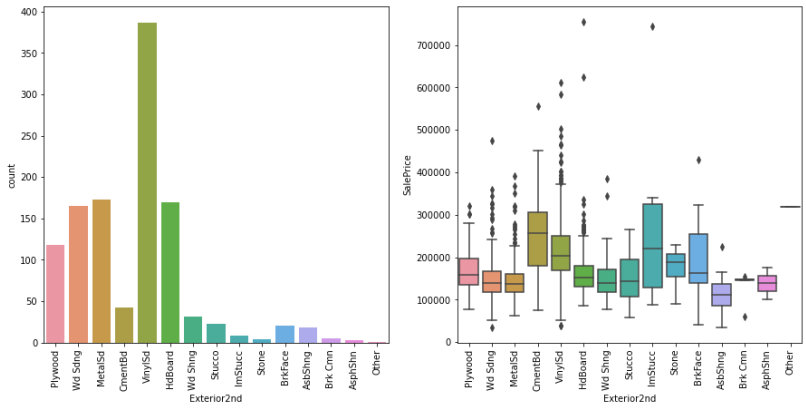
From visualization we concluded that most data are present in CompShg category of RoofMatl & outliers of SalePrice are present in only 2 categories out of total 8 categories of RoofMatl.

Exterior1st



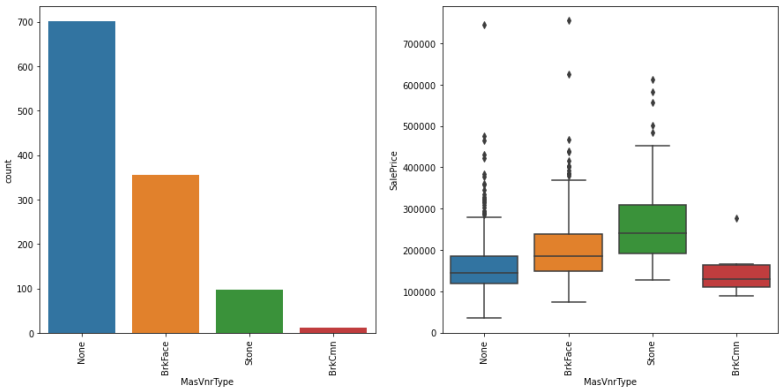
From visualization we concluded that most data is present in VinylSd category of Exterior1st & outliers of SalePrice are present in 9 categories out of total 14 categories of Exterior1st.

Exterior2nd



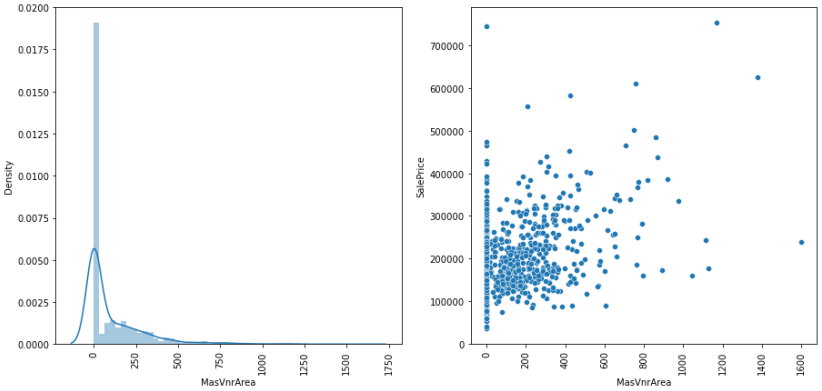
From visualization we concluded that most data is present in VinylSd category of Exterior2nd & outliers of SalePrice are present in 11 categories out of total 15 categories of Exterior2nd.

MasVnrType



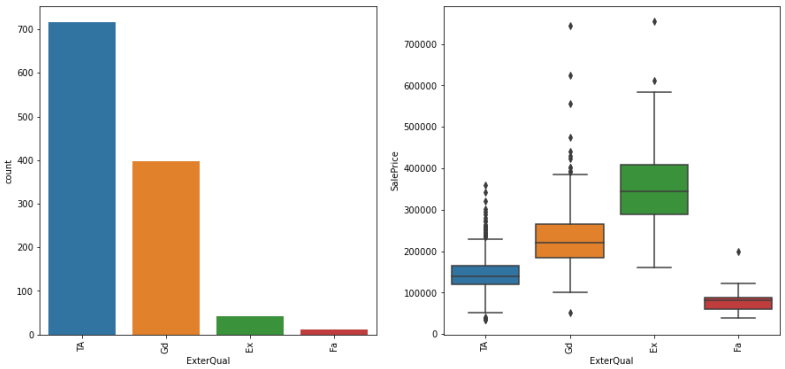
From Visualization we concluded that most data is present in None category of MasVnrType & outliers of SalePrice are present in every category of MasVnrType.

MasVnrArea



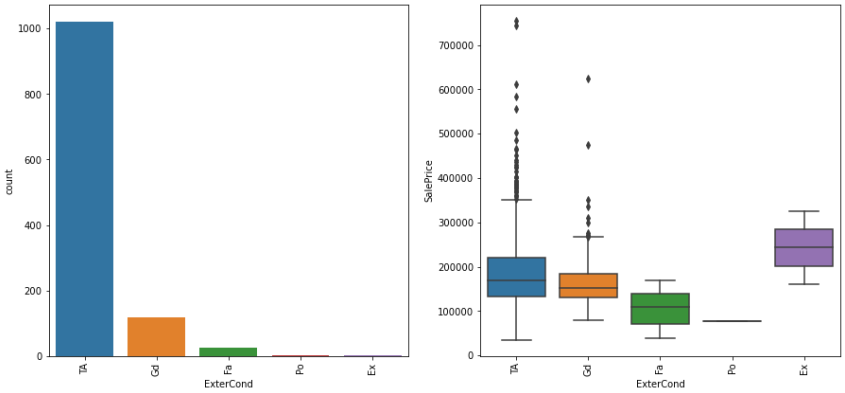
From visualization we concluded that data is positively skewed with having highest density at around 0 & data is having positive correlation with SalePrice.

ExterQual



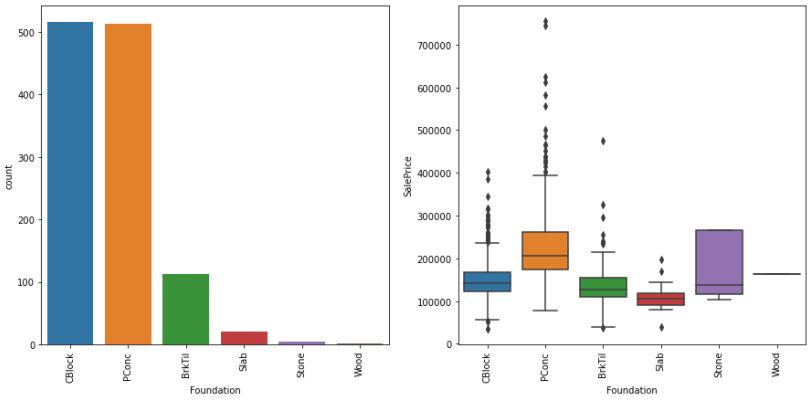
From visualization we concluded that most data is present in TA category of ExterQual & outliers of SalePrice are present in every category of ExterQual.

ExterCond



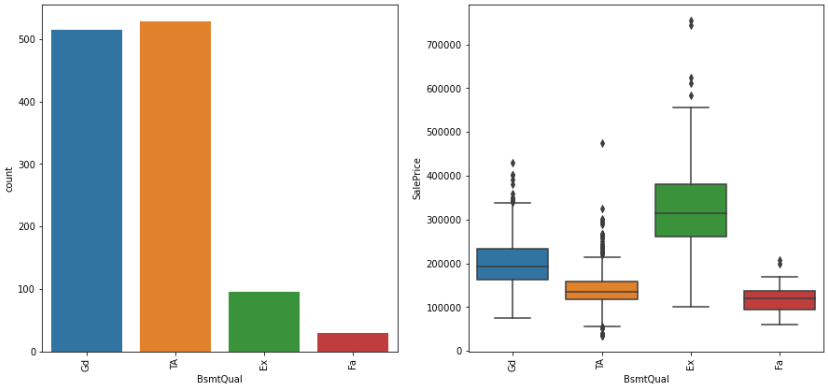
From visualization we concluded that most data is present in TA category of ExterCond & outliers of SalePrice are present in only 2 categories out of total 5 categories of ExterCond.

Foundation



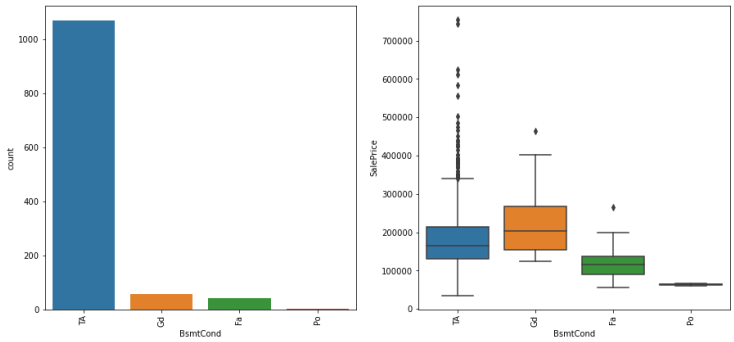
From visualization we concluded that most data is present in CBlock category of Foundation & outliers of SalePrice are present in 4 categories out of total 6 categories of Foundation.

BsmtQual



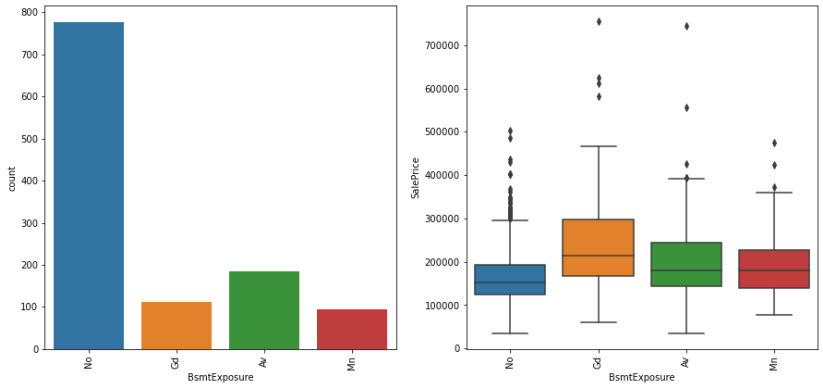
From visualization we concluded that most data is present TA category of BsmtQual & outliers of SalePrice are present in every category of BsmtQual.

BsmtCond



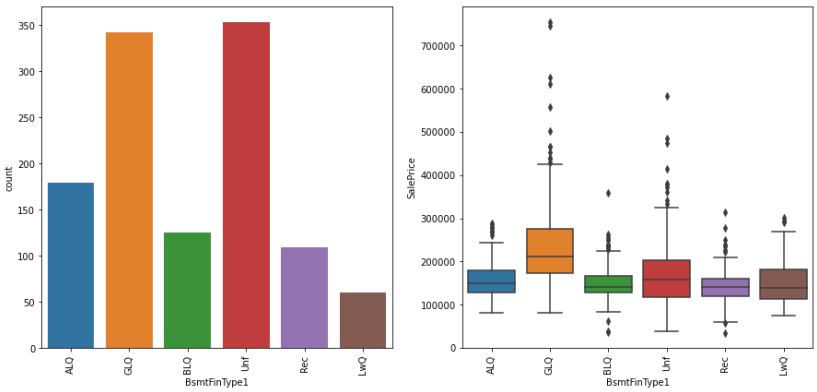
From visualization we concluded that most data is present in TA category of BsmtCond & outliers of SalePrice are present in 3 categories out of total 4 categories of BsmtCond.

BsmtExposure



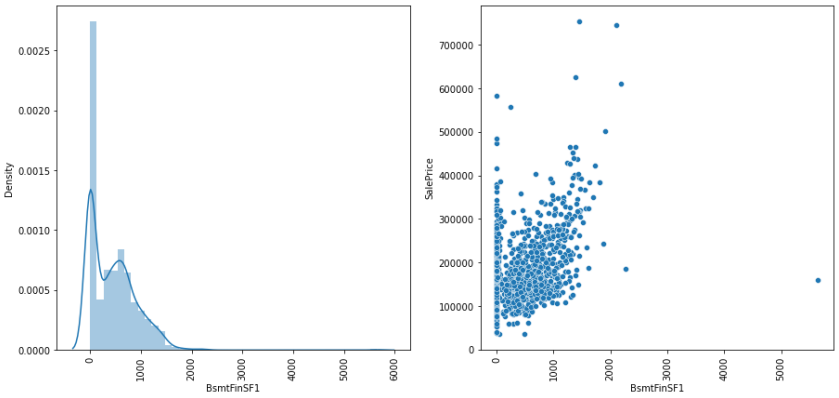
From visualization we conclude that most data is present in No category of BsmtExposure & outliers of SalePrice are present in every category of BsmtExposure.

BsmtFinType1



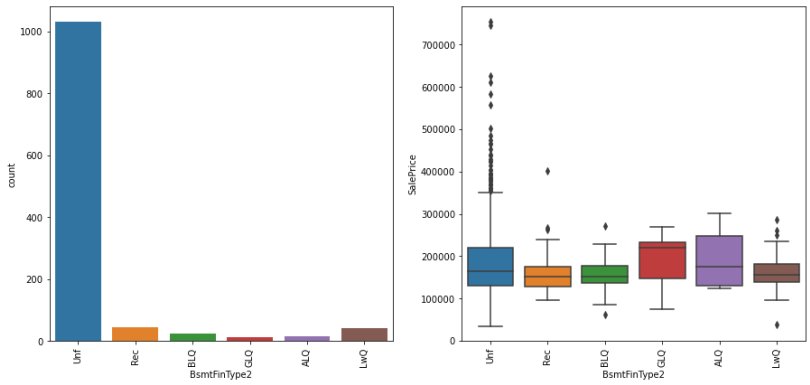
From visualization we concluded that most data is present in Unf category of BsmtFinType1 & outliers are present in every category BsmtFinType1.

BsmtFinSF1



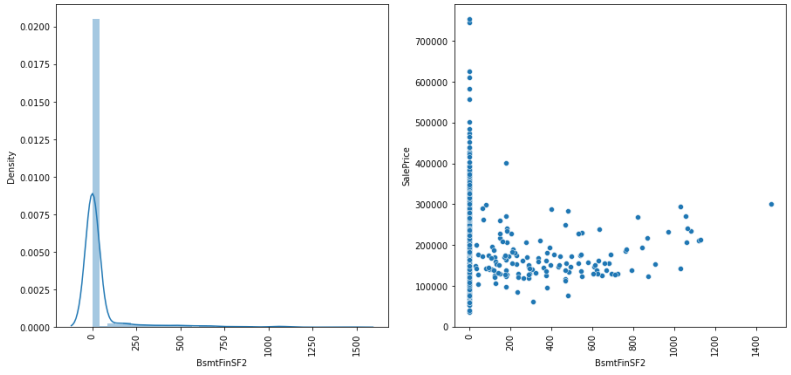
From visualization we concluded that data is positively skewed with having highest density at around 0 & it is showing positive correlation with SalePrice on plot.

BsmtFinType2



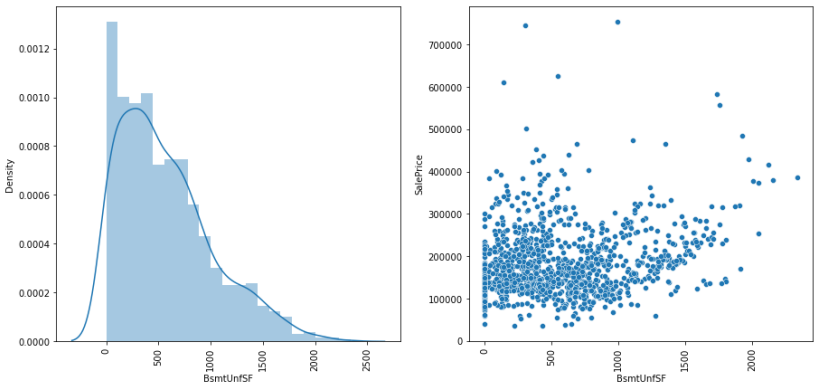
From visualization we concluded that most data is present Unf category of BsmtFinType2 & outliers of SalePrice are present in 4 categories out of total 6 categories of BsmtFinType2.

BsmtFinSF2



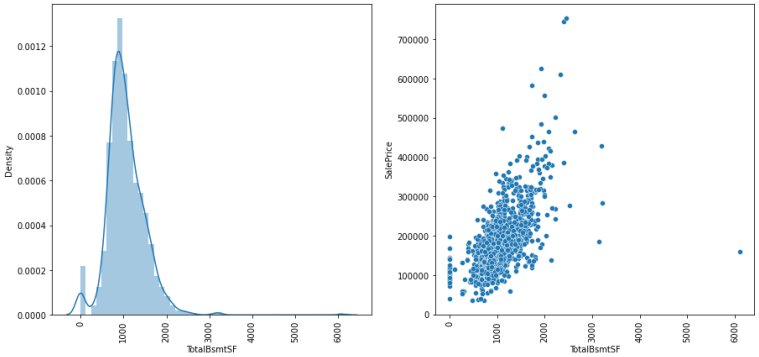
From visualization we concluded that data is showing somewhat positive skewness with data having highest density at around 0 & we cannot determine correlation of data with SalePrice on plot.

BsmtUnfSF



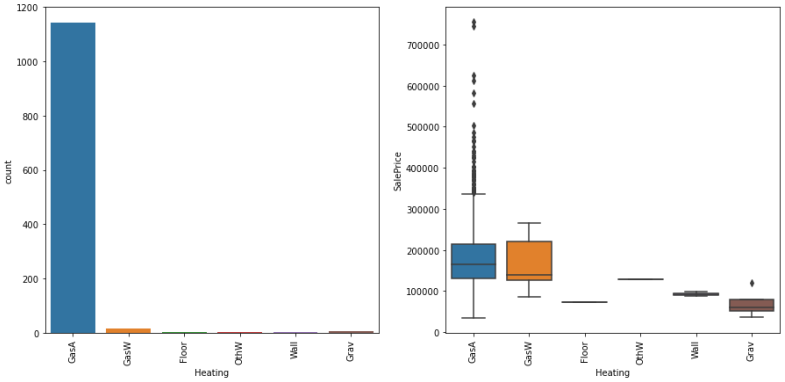
From visualization we conclude that data is showing positive skewness with having highest density at around 0 & data is having somewhat positive correlation with SalePrice on plot.

TotalBsmtSF



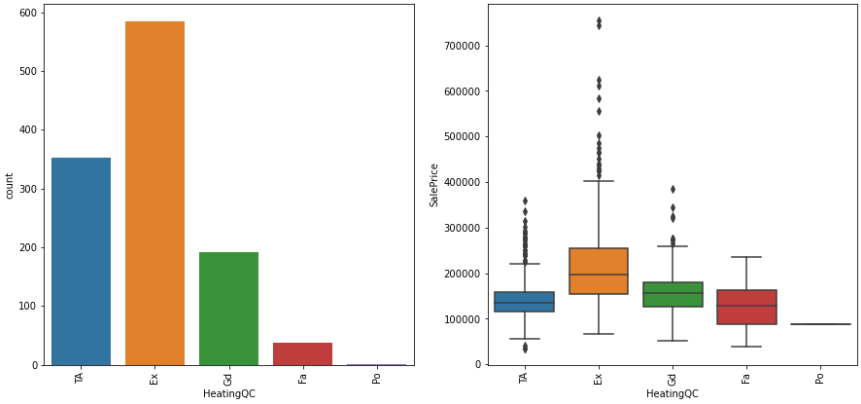
From visualization we conclude that data is showing positive skewness with having highest density at around 1000 & it is showing positive correlation with SalePrice on plot.

Heating



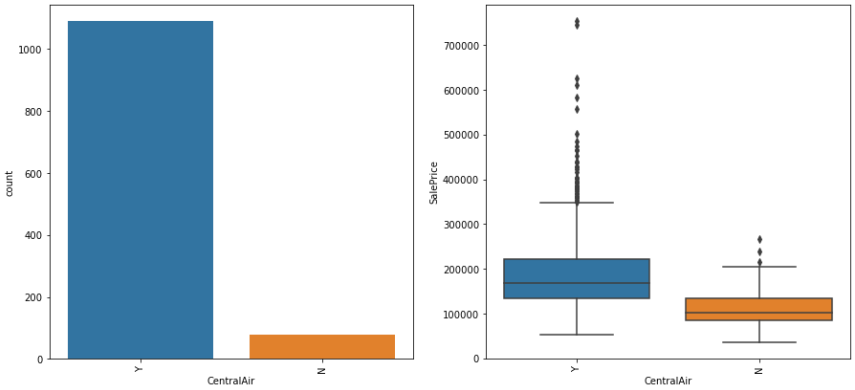
From visualization we conclude that most data is present in GasA category of Heating & outliers of SalePrice are present in only 2 categories out of total 6 categories of Heating.

HeatingQC



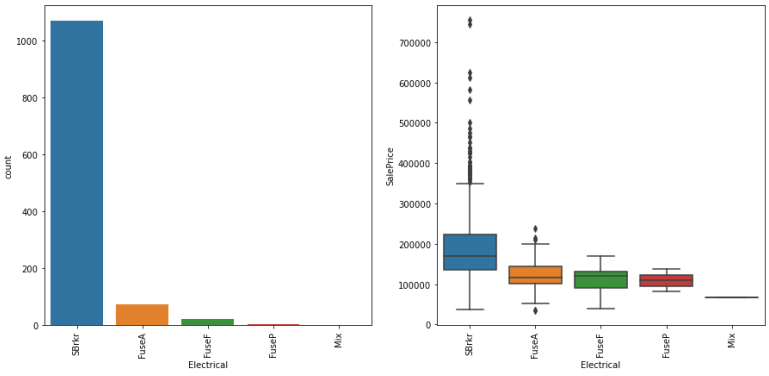
From visualization we concluded that most data is present in Ex category of HeatingQC & outliers of SalePrice are present in 3 categories out of total 5 categories of HeatingQC.

CentralAir



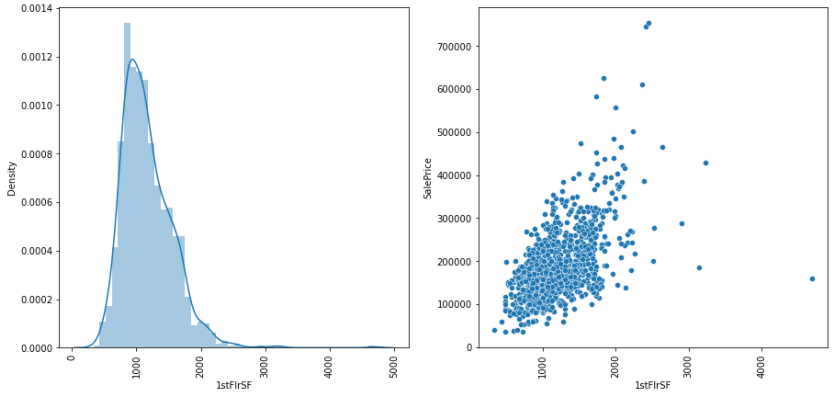
From visualization we concluded that most data is present in Y category of CentralAir & outliers of SalePrice are present in both categories of CentralAir.

Electrical



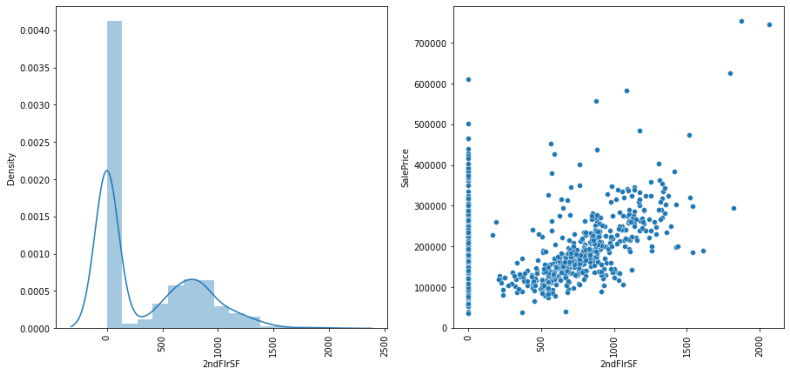
From visualization we concluded that most data is present in SBrkr category of Electrical & outliers of SalePrice are present in 2 categories out of total 5 categories of Electrical.

1stFlrSF



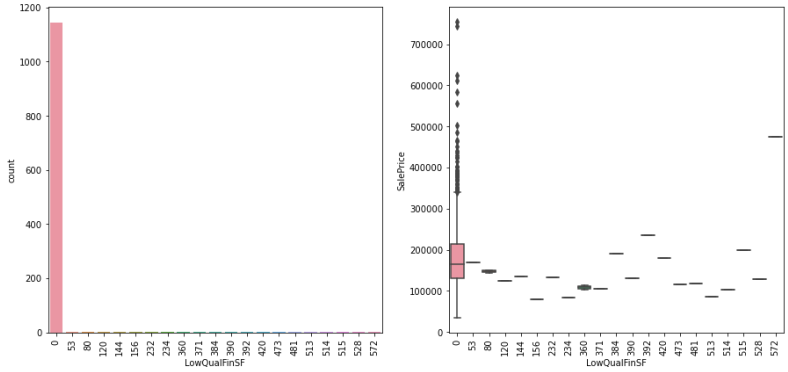
From visualization we concluded that data is positively skewed with having highest density at around 600 & data is showing positive correlation with SalePrice on plot.

2ndFlrSF



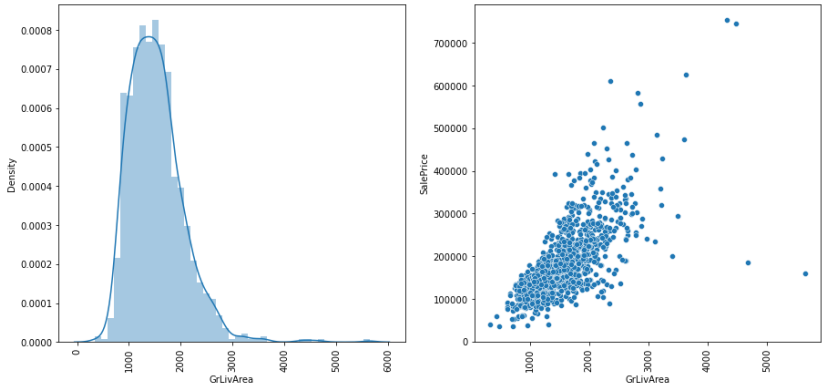
From visualization we concluded that data is showing somewhat positive skewness with multiple peaks & data having highest density at around 0. Also data is positively correlated with SalePrice on plot.

LowQualFinSF



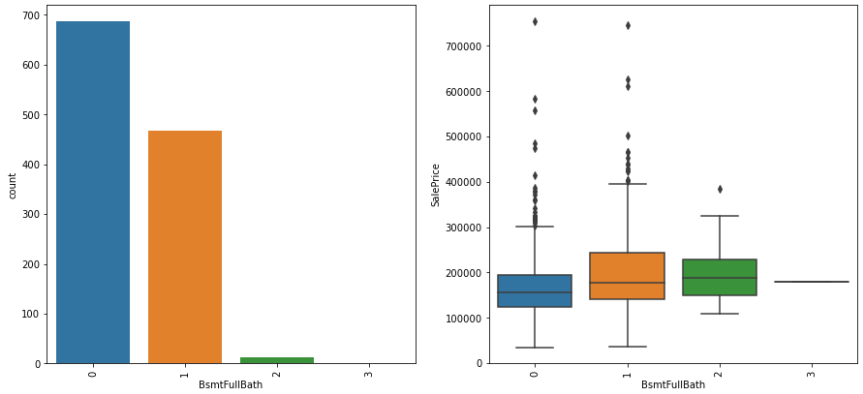
From visualization we concluded that most data is present in 0 category of LowQualFinSF & outliers of SalePrice are present in only 1 category out of total 21 categories of LowQualFinSF.

GrLivArea



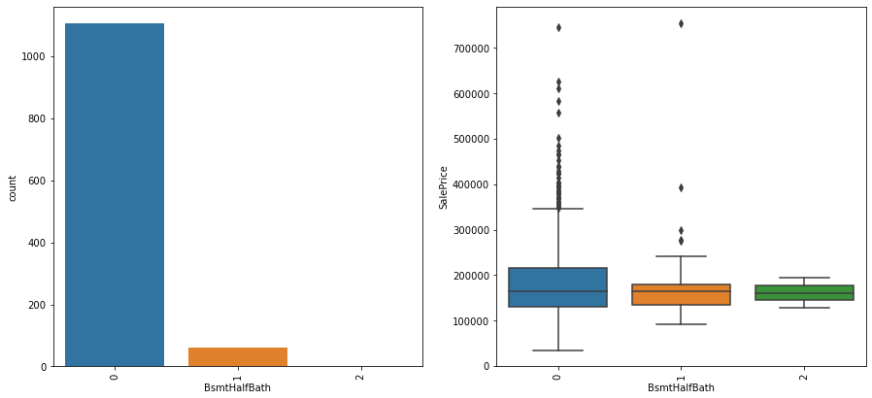
From visualization we conclude that data is positively skewed with having highest density at around 1800 & it is showing positive correlation with SalePrice on plot.

BsmtFullBath



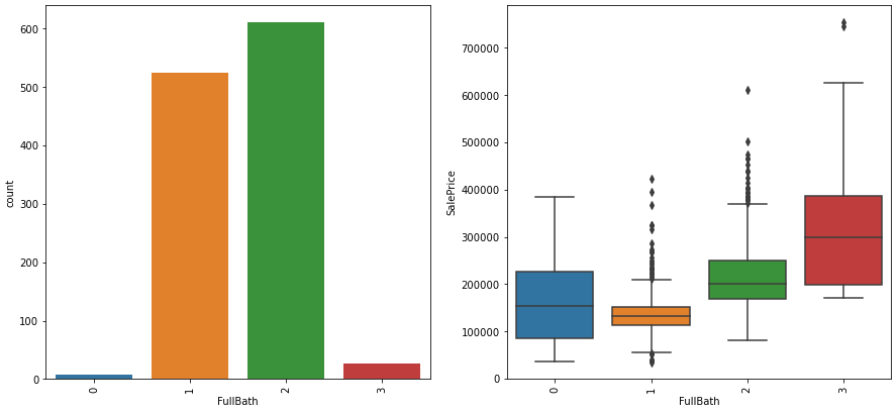
From visualization we concluded that most data is present in 0 category of BsmtFullBath & outliers of SalePrice are present in 3 categories out of total 4 categories of BsmtFullBath.

BsmtHalfBath



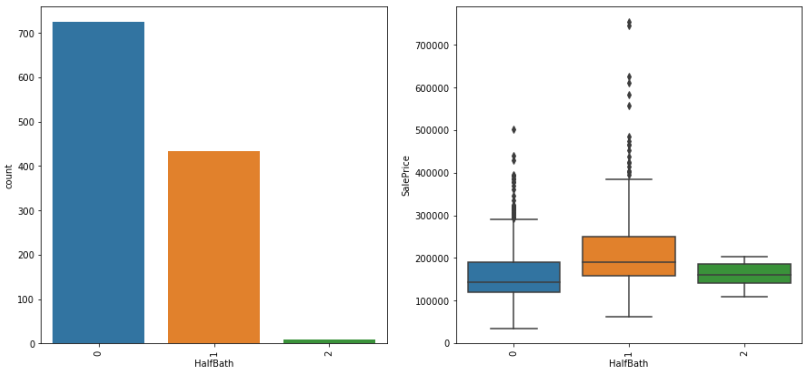
From visualization we conclude that most data is present in 0 category of BsmtHalfBath & outliers of SalePrice are present in 2 categories out of total 3 categories of BsmtHalfBath.

FullBath



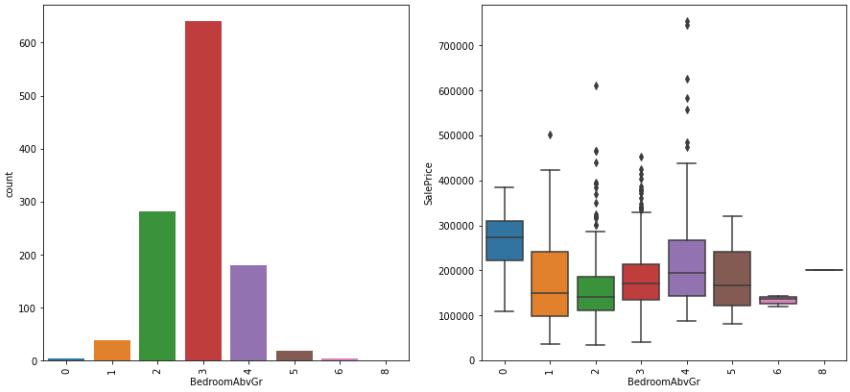
From visualization we concluded that most data is present in category 2 of FullBath & outliers of SalePrice are present in 3 categories out of total 4 categories of FullBath.

HalfBath



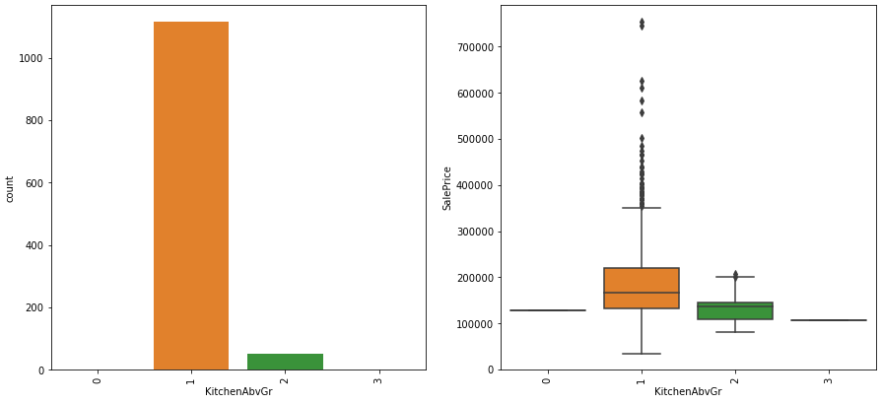
From visualization we concluded that most data is present in category 0 of HalfBath & outliers of SalePrice are present in 2 categories out of total 3 categories of HalfBath.

BedroomAbvGr



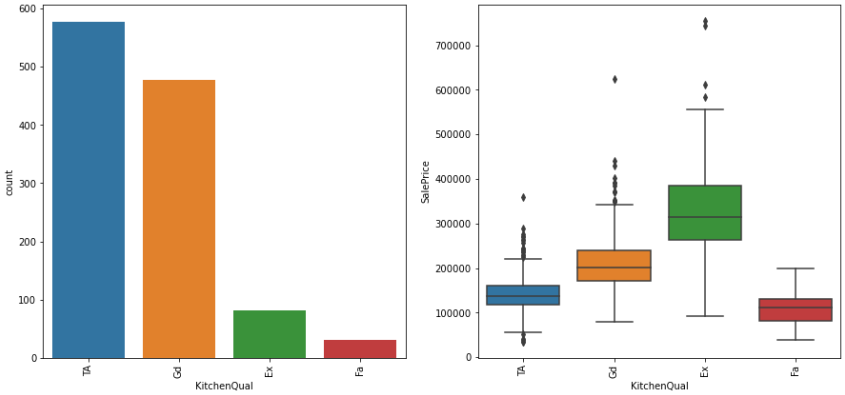
From visualization we concluded that most data is present in category 3 of BedroomAbvGr & outliers of SalePrice are present in 4 categories out of total 8 categories of BedroomAbvGr.

KitchenAbvGr



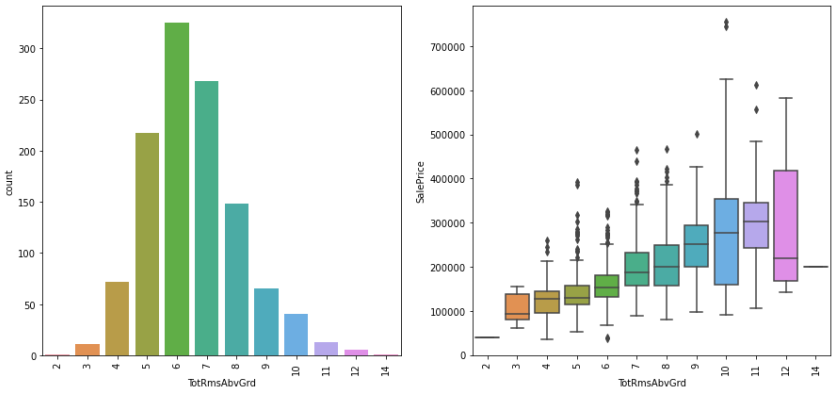
From visualization we concluded that most data is present in category 1 of KitchenAbvGr & outliers of SalePrice are present in 2 categories out of total 4 categories of KitchenAbvGr.

KitchenQual



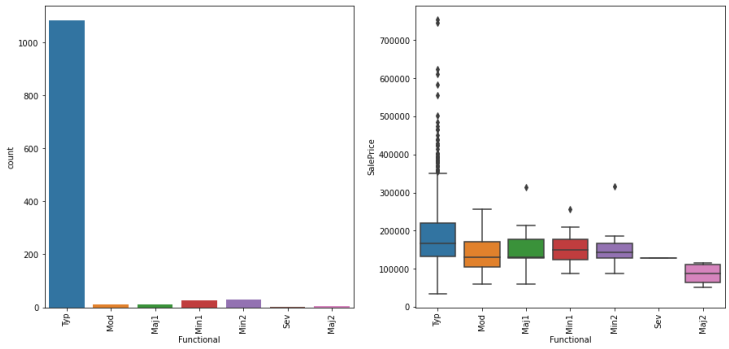
From visualization we conclude that most data is present in TA category of KitchenQual & outliers of SalePrice are present in 3 categories out of total 4 categories of KitchenQual

TotRmsAbvGr



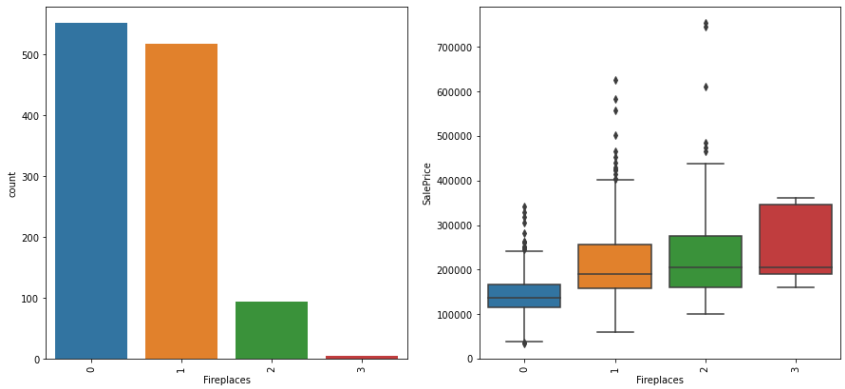
From visualization we concluded that most data is present in category 6 of TotRmsAbvGr & outliers of SalePrice are present in 8 categories out of total 12 categories of TotRmsAbvGr.

Functional



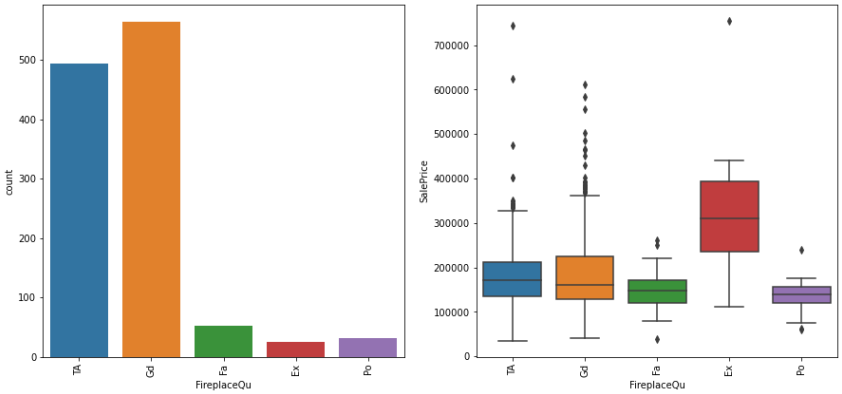
From visualization we conclude that most data is present in category Typ of Functional & outliers of SalePrice are present in 4 categories out of total 7 categories of Functional.

Fireplaces



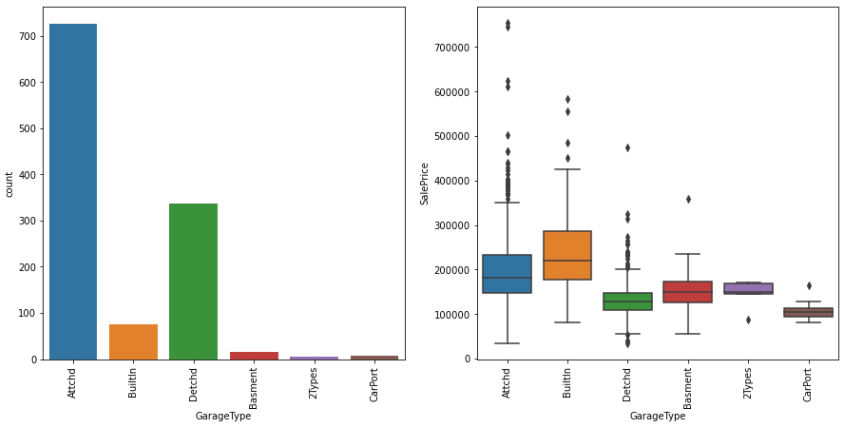
From visualization we conclude that most data is present in category 0 of Fireplaces & outliers of SalePrice are present in 3 categories out of total 4 categories of Fireplaces

FireplaceQu



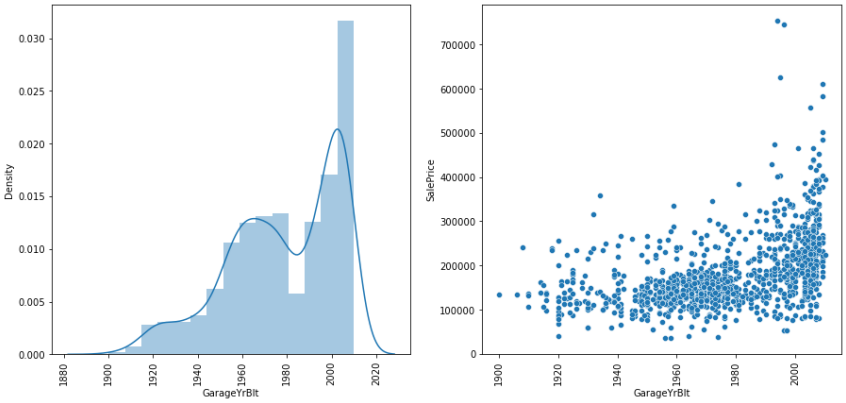
From visualization we conclude that most data is present Gd category of FireplaceQu & outliers of SalePrice are present in every category of FireplaceQu.

GarageType



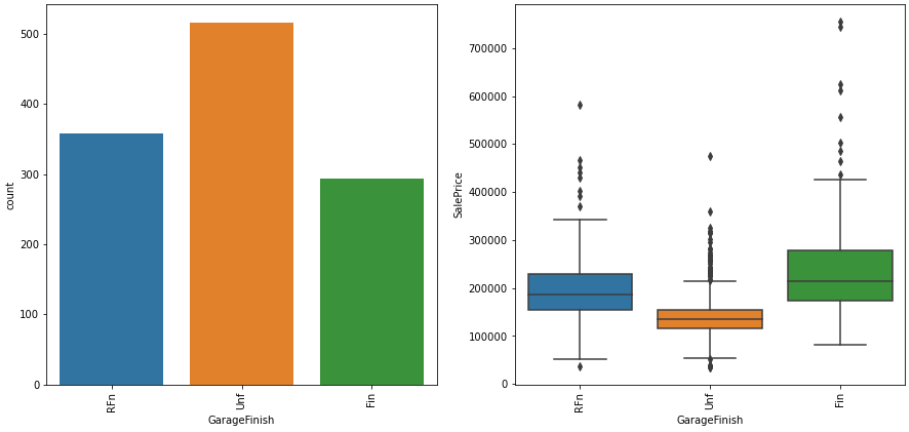
From visualization we conclude that data is mostly present in Attchd category of GarageType & outliers of SalePrice are present in every category of GarageType.

GarageYrBlt



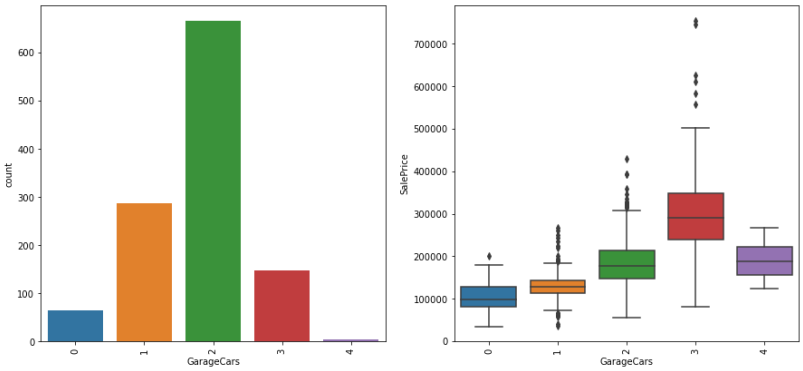
From visualization we conclude that data is negatively skewed with having highest density at around 2010 & data is showing somewhat positive correlation with SalePrice & data is dispersed all over the place on plot.

GarageFinish



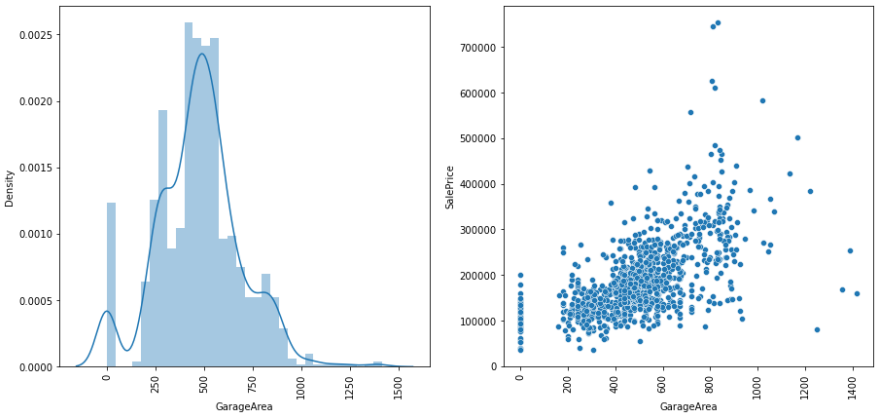
From visualization we conclude that most data is present in Unf category of GarageFinish & outliers of SalePrice are present in every category of GarageFinish.

GarageCars



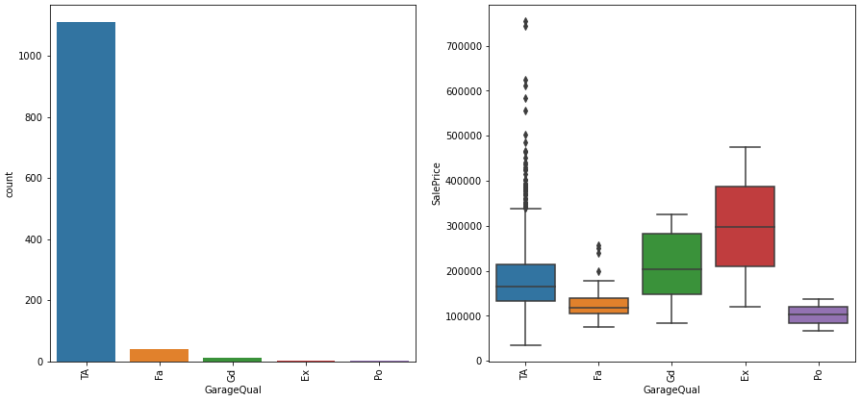
From visualization we conclude that most data is present in category 2 of GarageCars & outliers of SalePrice are present in 4 categories out of total 5 categories of GarageCars.

GarageArea



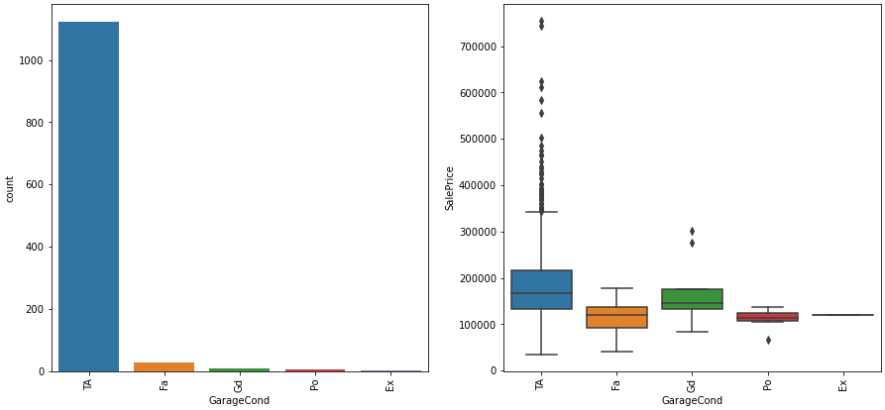
From visualization we conclude that data is positively skewed with having highest density around 350 & data is showing positive correlation with SalePrice in plot.

GarageQual



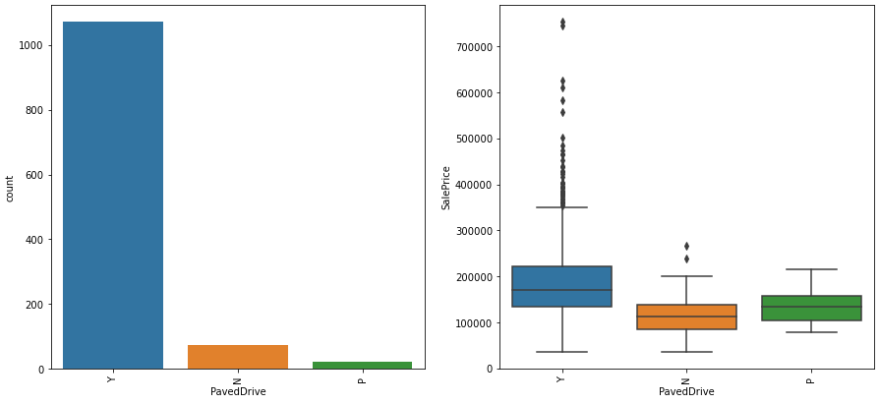
From visualization we conclude that most data is present in category TA of GarageQual & outliers of SalePrice are present in only 2 categories out of total 5 categories of GarageQual.

GarageCond



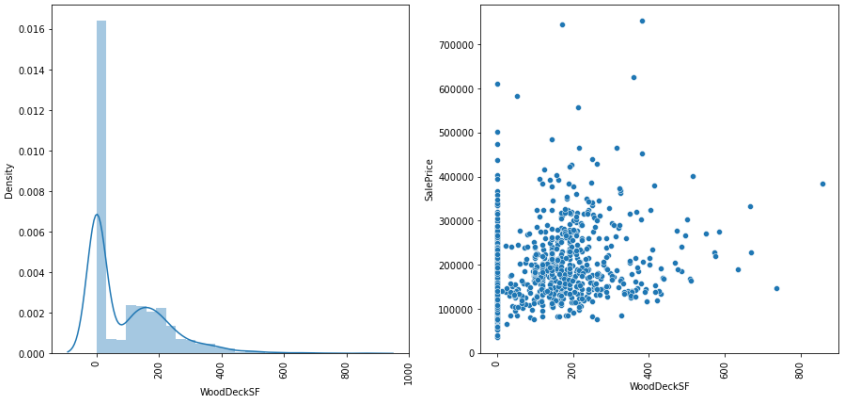
From visualization we conclude that most data is present in category TA of GarageCond & outliers of SalePrice are present in 3 categories out of total 5 categories of GarageCond.

PavedDrive



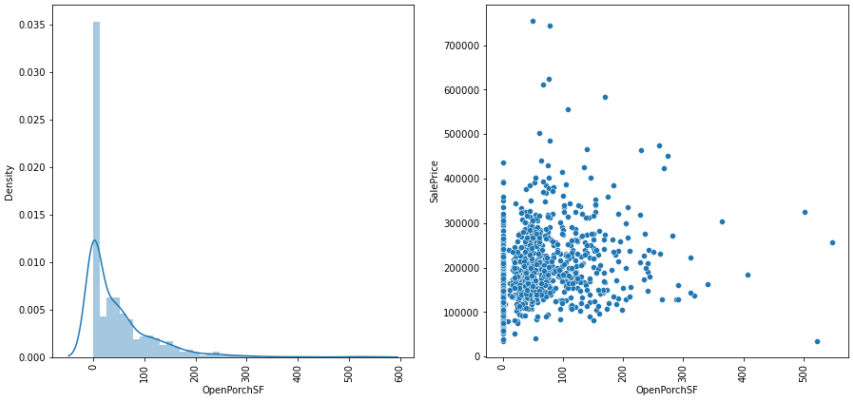
From visualization we conclude that most data is present in category Y of PavedDrive & outliers of SalePrice are present in 2 categories out of total 3 categories of PavedDrive.

WoodDeckSF



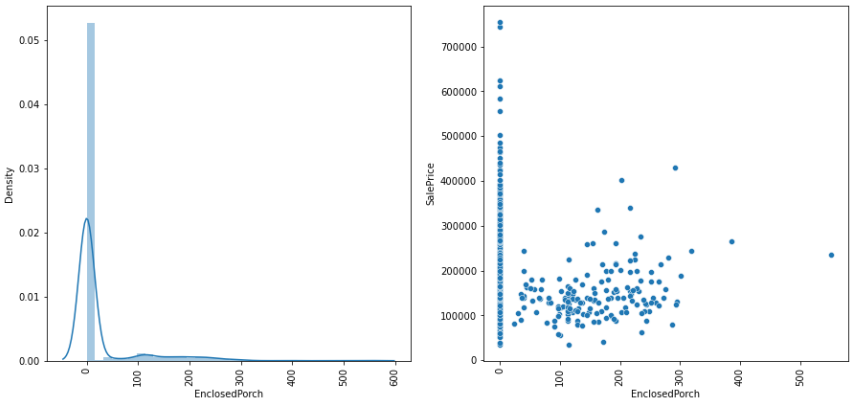
From visualization we conclude that data is positively skewed with multiple peaks & having highest density at around 0. Also data is having positive correlation with Saleprice in plot.

OpenPorchSF



From visualization we conclude that data is positively skewed with having highest density at around 0 & data having positive correlation with SalePrice in plot.

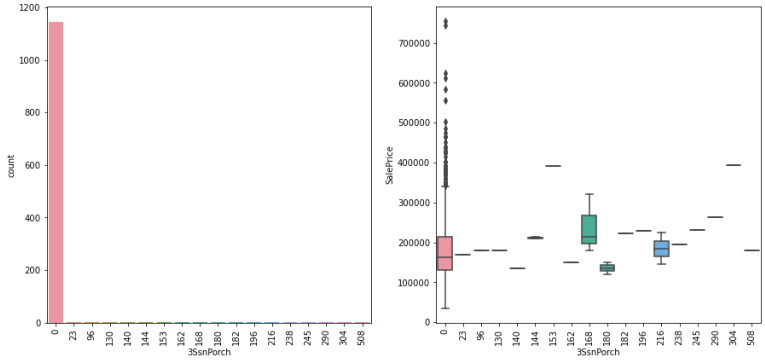
EnclosedPorch



From visualization we conclude that data is somewhat positively skewed with having highest density at around 0 & data is having positive correlation with SalePrice on plot.

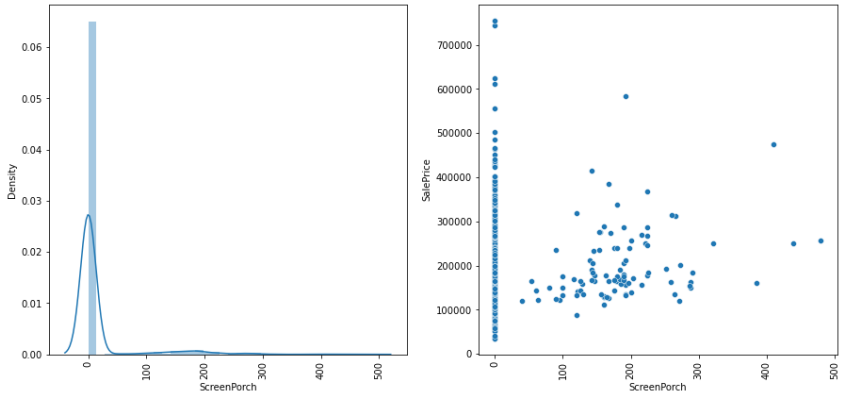
s

3SsnPorch



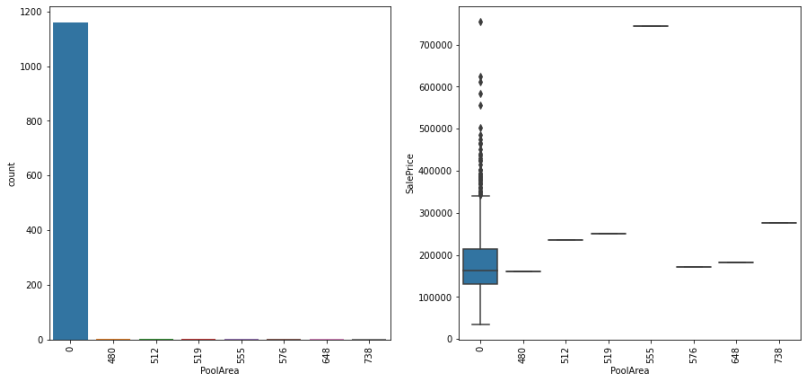
From visualization we conclude that most data is present in category 0 of 3SsnPorch & outliers of SalePrice are present in 1 category out of total 18 categories of 3SsnPorch.

ScreenPorch



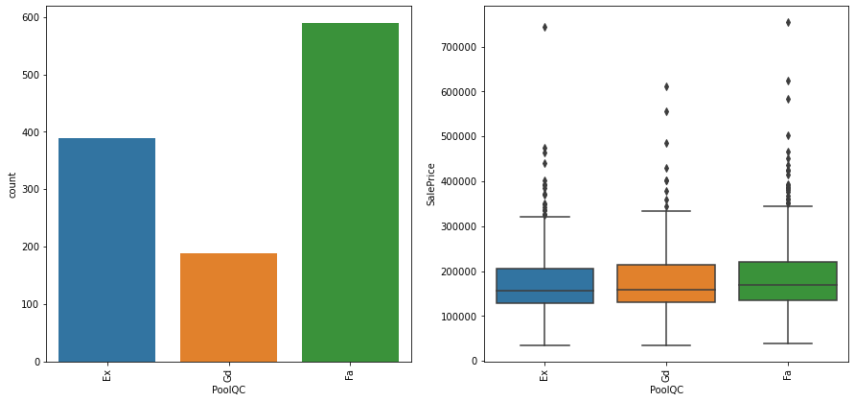
From visualization we conclude that data is somewhat positively skewed with having highest density at around 0 & data is having positive correlation with SalePrice on plot.

PoolArea



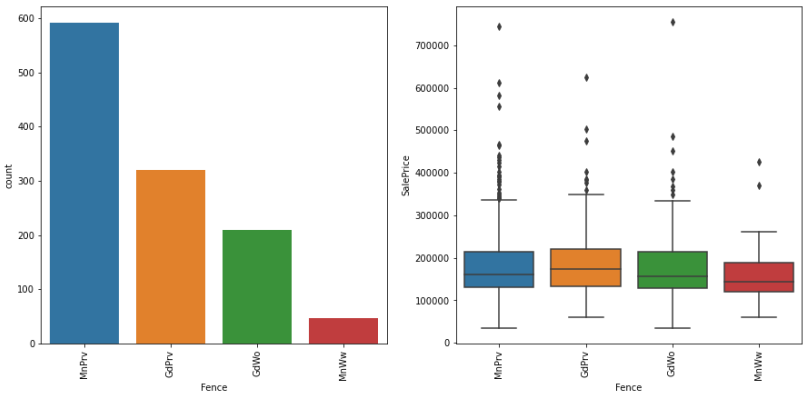
From visualization we conclude that most data is present in category 0 of PoolArea & outliers of SalePrice are present in only 1 category out of total 8 categories of PoolArea.

PoolQC



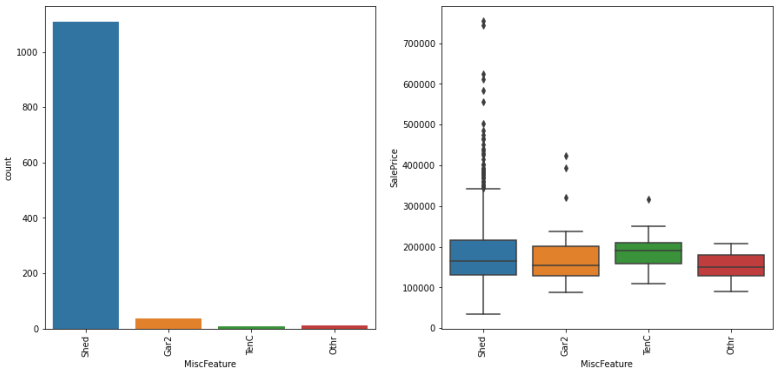
From visualization we conclude that most data is present in category Fa of PoolQC & outliers of SalePrice are present in every category of PoolQC

Fence



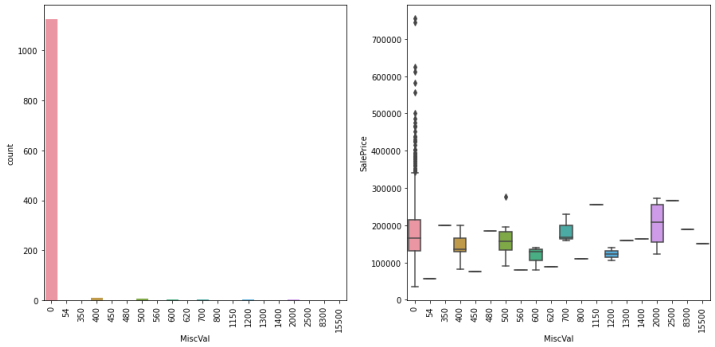
From visualization we conclude that most data is present in category MnPrv of Fence & outliers of SalePrice are present in every category of Fence.

MiscFeature



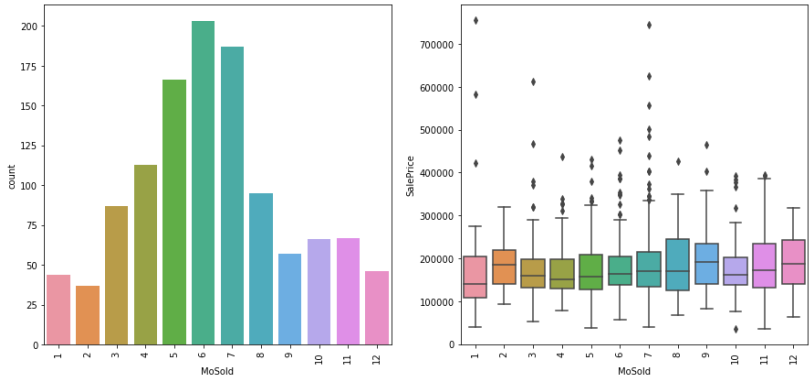
From visualization we conclude that most data is present in category Shed of MiscFeature & outliers of SalePrice are present in 3 categories out of total 4 categories of MiscFeature.

MiscVal



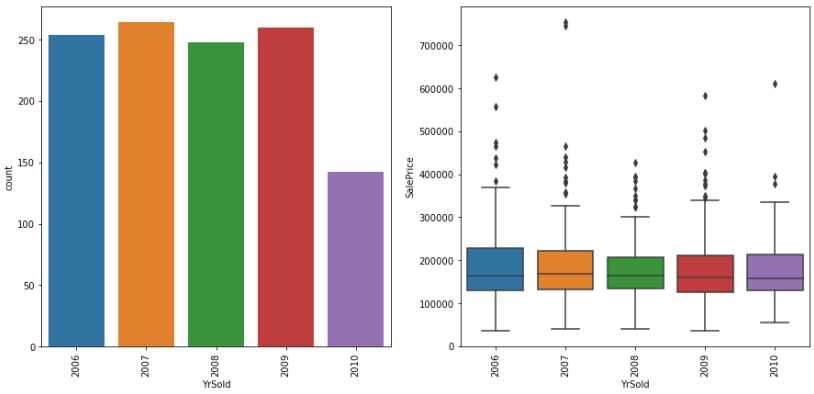
From visualization we conclude that most data is present in category 0 of MiscVal & outliers of SalePrice are present in 2 categories out of total 20 categories of MiscVal.

MoSold



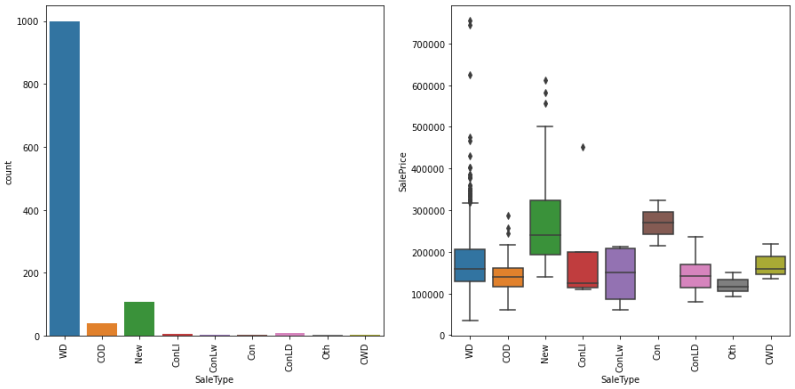
From visualization we conclude that most data is present in category 6 of MoSold & outliers of SalePrice are present in 10 categories out of total 12 categories of MoSold.

YrSold



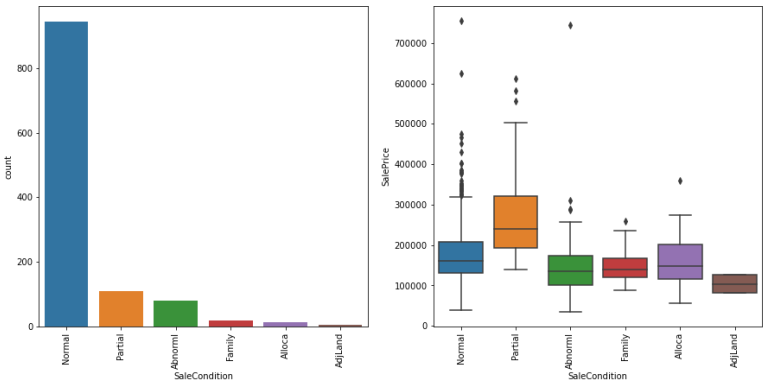
From visualization we conclude that most data is present in category 2007 of YrSold & outliers of SalePrice are present in every category of YrSold.

SaleType



From visualization we conclude that most data is present in category WD of SaleType & outliers of SalePrice are present in 4 categories out of total 9 categories of SaleType.

SaleCondition



From visualization we conclude that most data is present in Normal category of SaleCondition & outliers of SalePrice are present in 5 categories out of total 6 categories of SaleCondition.

Changed data type

After visualization we changed data type of every column of train dataset to int or float data type.

8. Plotting Heatmap

We plotted heatmap of train dataset to check correlation of every column with target column & we found that many columns having high positive correlation with target column. Also we dropped following (Utilities, SaleCondition, Alley, GarageQual) columns to reduce multicollinearity.

9. Preparing Test dataset

After plotting heatmap for train dataset, we filled the null values in test dataset & changed the data types of every column to int or float data type present in test dataset. Also target column was not present in test dataset.

10. Scaling the data

After preparing both train & test dataset we scaled them to remove outliers from them.

11. Applying GridSearchCV

We chose 4 regression models (Linear Regression, Decision Tree Regression, Random Forest Regression, Bagging Regression) to check performance of them on the dataset to find out best suited model for the dataset. Then one by one we applied GridSearchCV on each model to find best hyper parameter tuning while model working on dataset. After applying GridSearchCV on every model we concluded that Random Forest Regression model is giving best train accuracy which was 0.9110 with hyper parameter combination (max_depth : 5, min_samples_leaf : 1, min_samples_split : 2)

12. Applying Model on dataset

Chose Random Forest Regression Model & made it ready by tuning hyper parameter. After tuning hyper parameter we applied model on dataset to make prediction. After making prediction we made a dataframe which contained predicted value as test dataset does not have the target column. So we could’nt check the performance of the model.

13. Saving Model

After making prediction we saved the model using pickle library.

CONCLUSION

In the end we conclude that model is quite accurate in predicting the price of houses & price is getting affected positively because of the following variables (OverallQual, YearBuilt, YearRemodAdd, ExterQual, GrLivArea).