

Feature Selection and EDA Report

Null Values

Null Values by Columns/Features

Out of 80 total data columns, there are 19 columns with at least 1 null value.

Feature	Num of Nulls	Frac Null
PoolQC	1453	1.0
MiscFeature	1406	0.96
Alley	1369	0.94
Fence	1179	0.81
MasVnrType	872	0.6
FireplaceQu	690	0.47
LotFrontage	259	0.18
GarageType	81	0.06

Null Values by Rows/Data Samples

Out of 1460 total rows/data samples, 1460 rows have at least one null value.

The row with the most NULL values has 16 NULLs.

Numeric vs Non-Numeric Features and Unique Values Count

Out of 79 total feature columns, there are 36 numeric columns and 43 non-numeric columns.

Numeric Feature	Num Unique Values
BsmtHalfBath	3
HalfBath	3
BsmtFullBath	4
FullBath	4
KitchenAbvGr	4
Fireplaces	4
GarageCars	5
YrSold	5

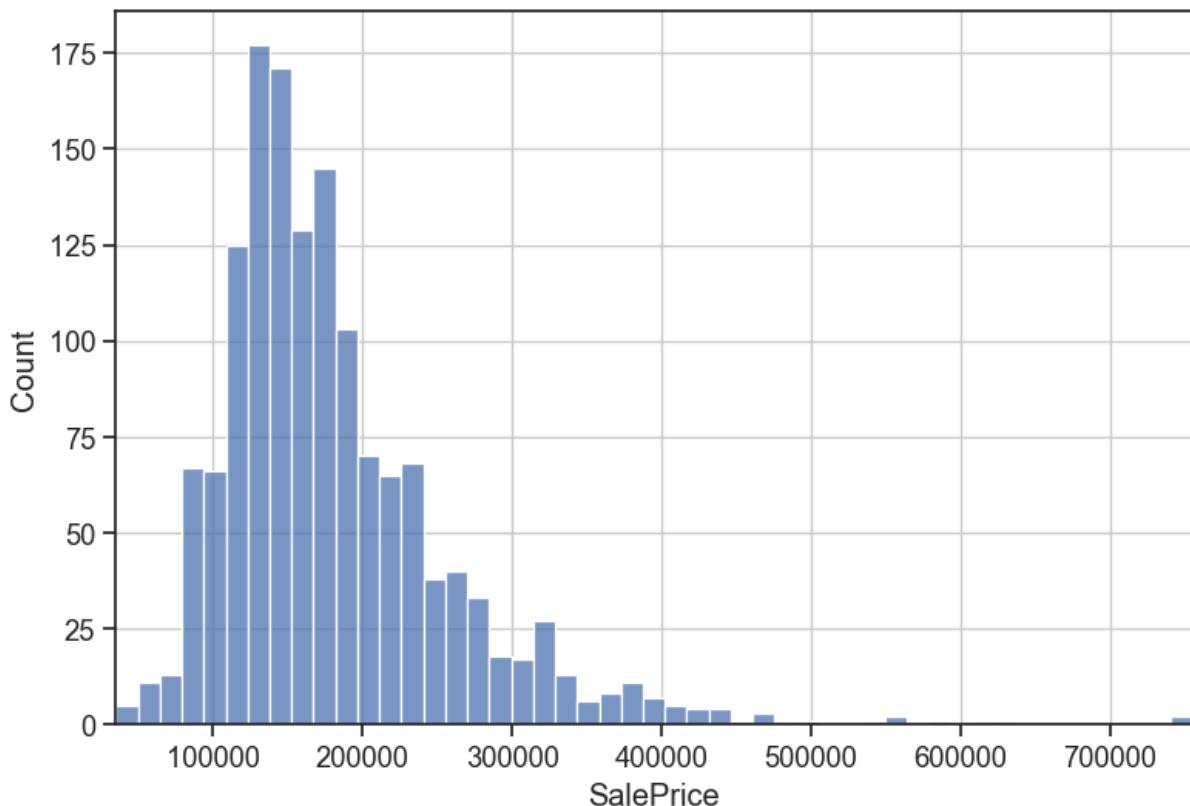
There are an additional 4 numeric feature columns with 10 or fewer unique values.

Non-Numeric Feature	Num Unique Values
Neighborhood	25
Exterior2nd	16
Exterior1st	15
Condition1	9
SaleType	9
Condition2	8
HouseStyle	8
RoofMatl	8

There are an additional 8 non-numeric feature columns with more than 5 unique values.

Target Column

For the chosen target column ('SalePrice'), this appears to be a regression problem. The target column has 0 null values and 663 unique values.

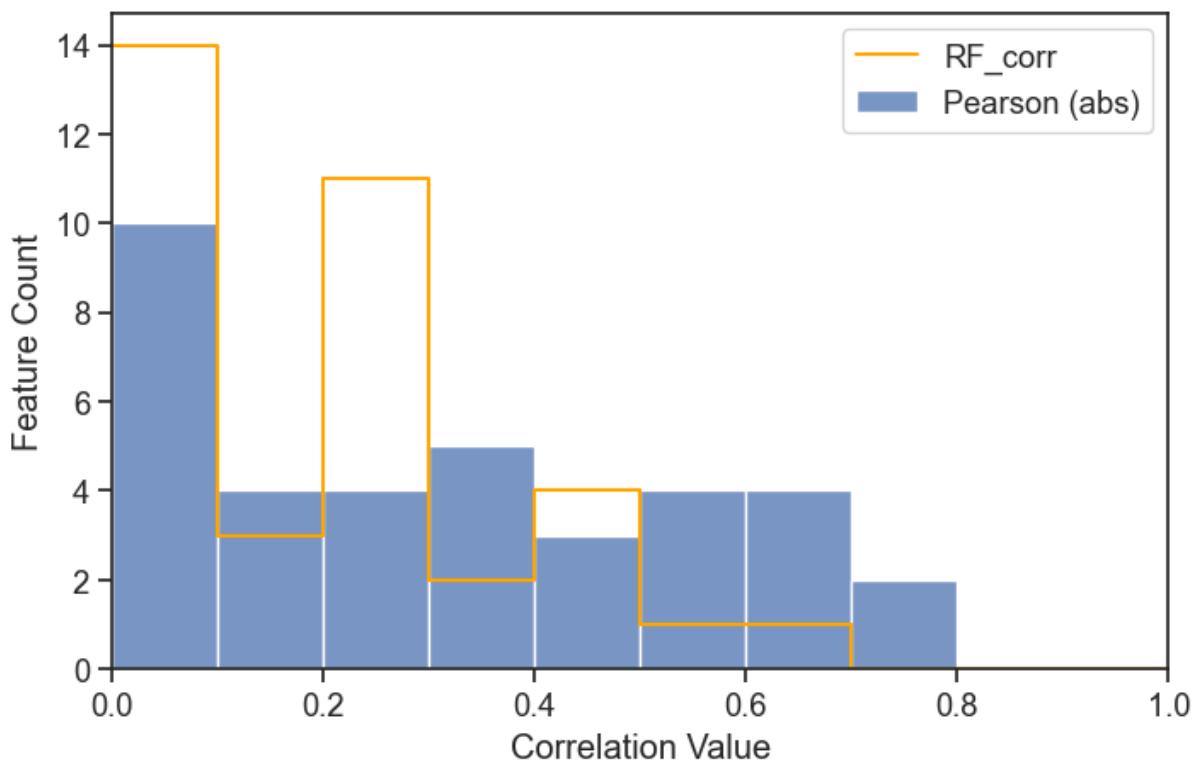


The above plot shows the distribution of values in the target column.

Feature Correlations

Correlations of Numeric Features with Target Variable

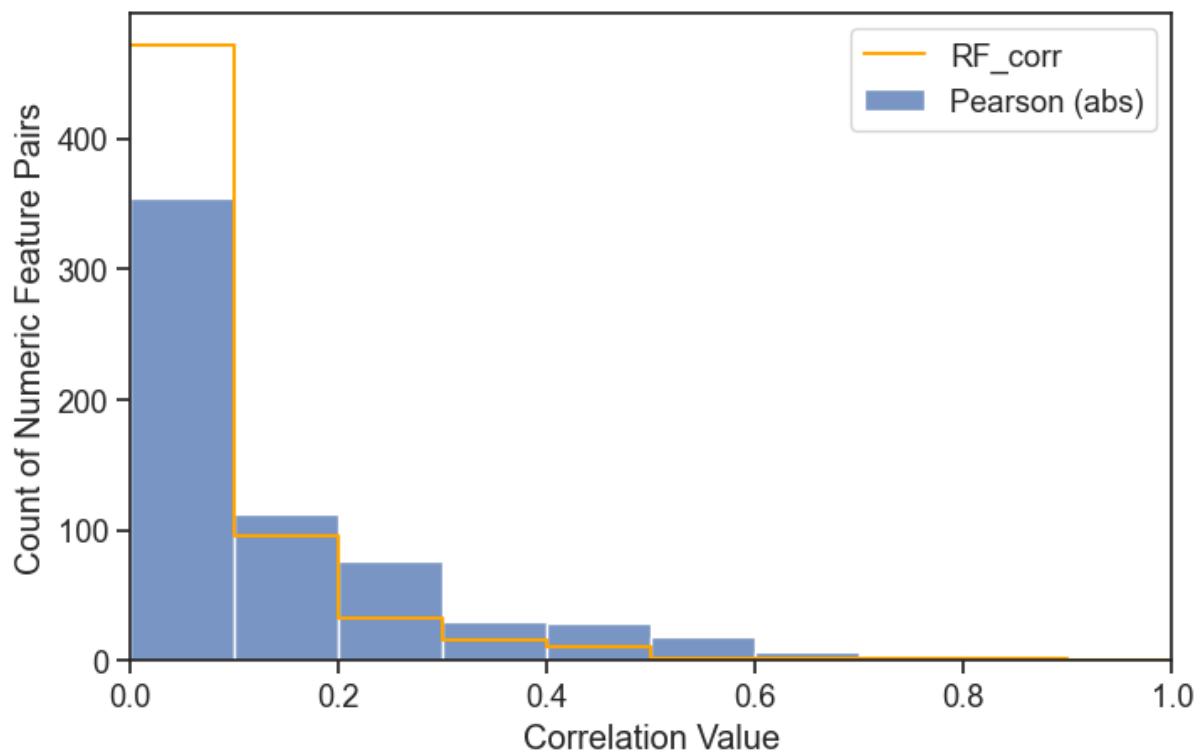
Numeric Feature	Count non-Null	Pearson Corr	RF Corr
OverallQual	1460	0.79	0.64
GrLivArea	1460	0.71	0.51
GarageCars	1460	0.64	0.48
GarageArea	1460	0.62	0.46
TotalBsmtSF	1460	0.61	0.46
1stFlrSF	1460	0.61	0.42
YearBuilt	1460	0.52	0.38
GarageYrBlt	1379	0.49	0.32



The above plot shows a histogram of all numeric features and their correlation value with the target variable.

Correlations between Numeric Features

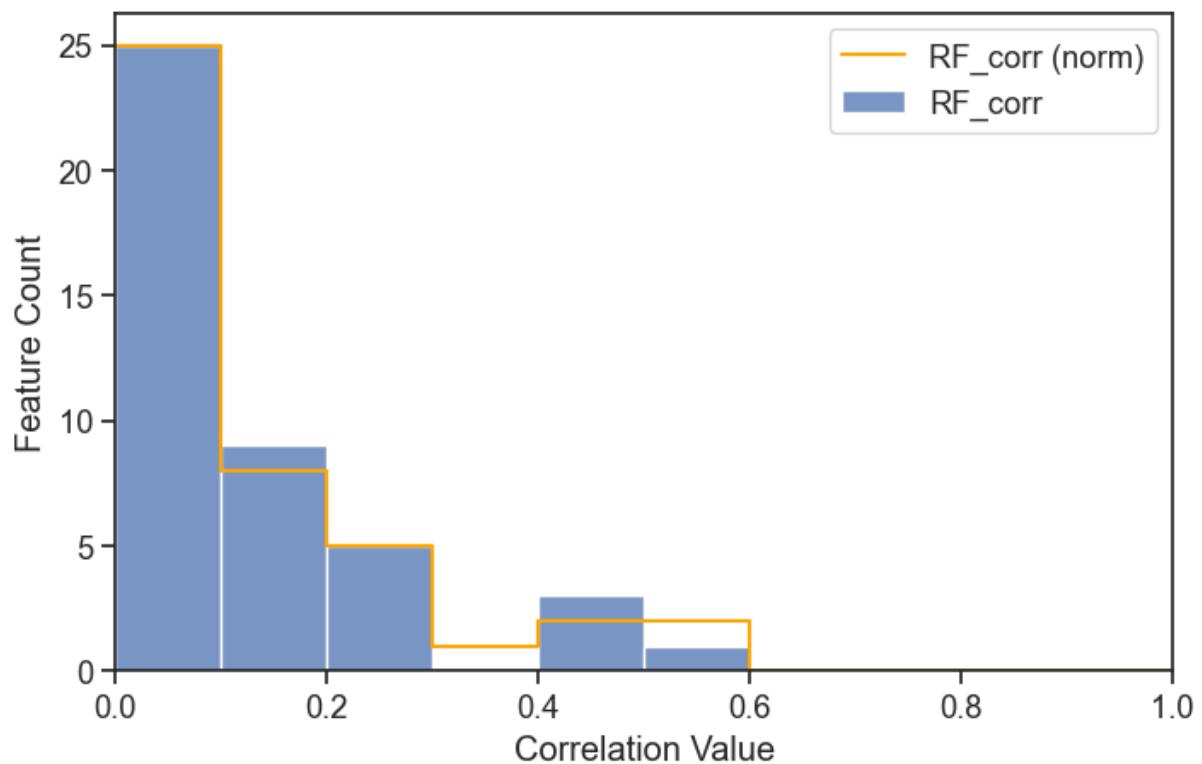
Numeric Feature	Avg Pearson Correlation	Avg RF Correlation	Feat with Max Pear Corr	Max Pear Corr	Feat with Max RF Corr	Max RF Corr
GarageArea	0.24	0.14	GarageCars	0.88	GarageCars	0.83
GarageCars	0.25	0.13	GarageArea	0.88	GarageArea	0.83
YearBuilt	0.22	0.16	GarageYrBlt	0.83	GarageYrBlt	0.73
GarageYrBlt	0.21	0.14	YearBuilt	0.83	YearBuilt	0.73
TotRmsAbvGrd	0.23	0.11	GrLivArea	0.83	GrLivArea	0.69
GrLivArea	0.28	0.16	TotRmsAbvGrd	0.83	TotRmsAbvGrd	0.69
TotalBsmtSF	0.25	0.13	1stFlrSF	0.82	1stFlrSF	0.68
1stFlrSF	0.25	0.13	TotalBsmtSF	0.82	TotalBsmtSF	0.68



The above plot shows a histogram of all unique pairs of numeric features and the correlation between the two features of each the pair.

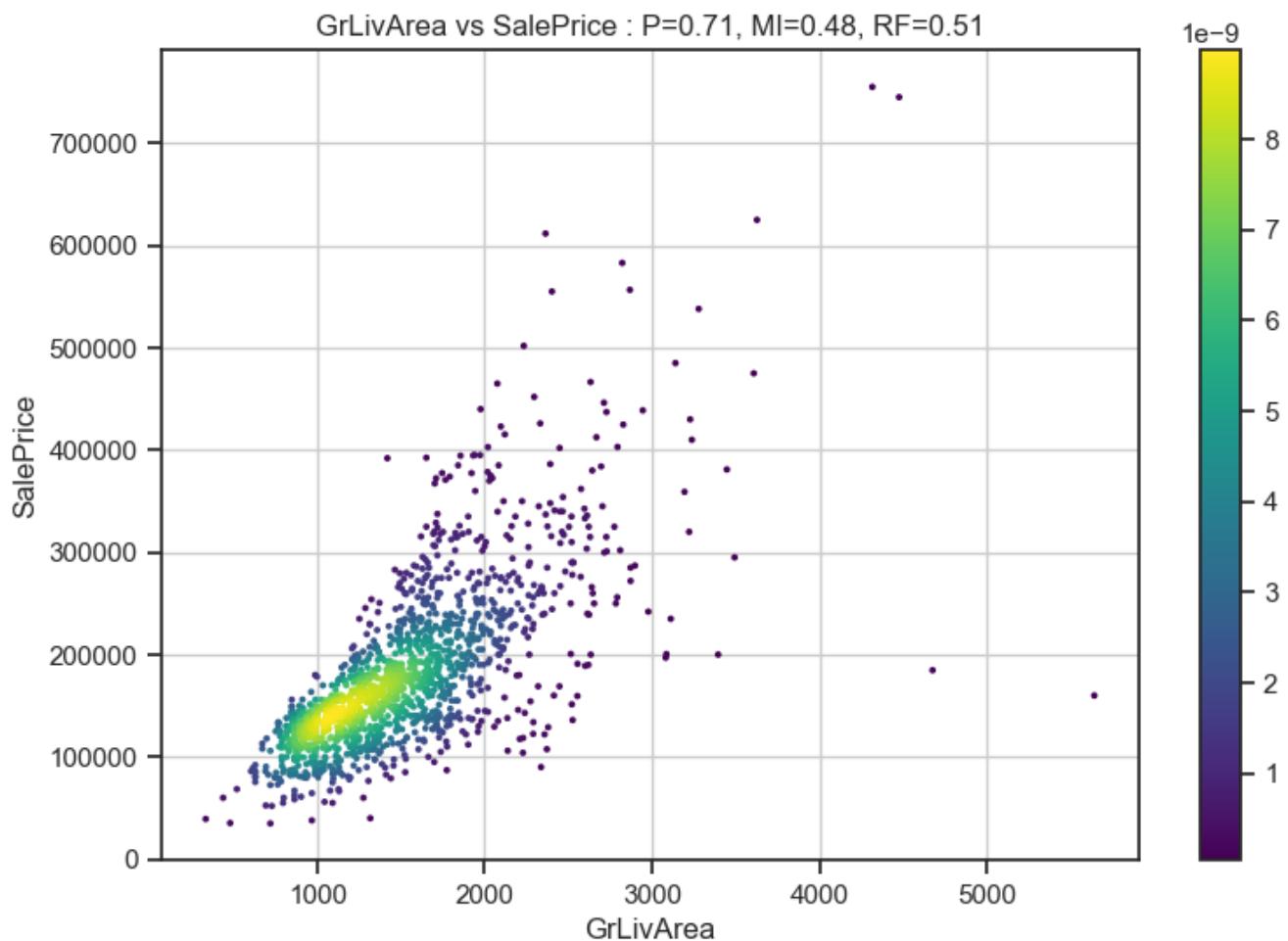
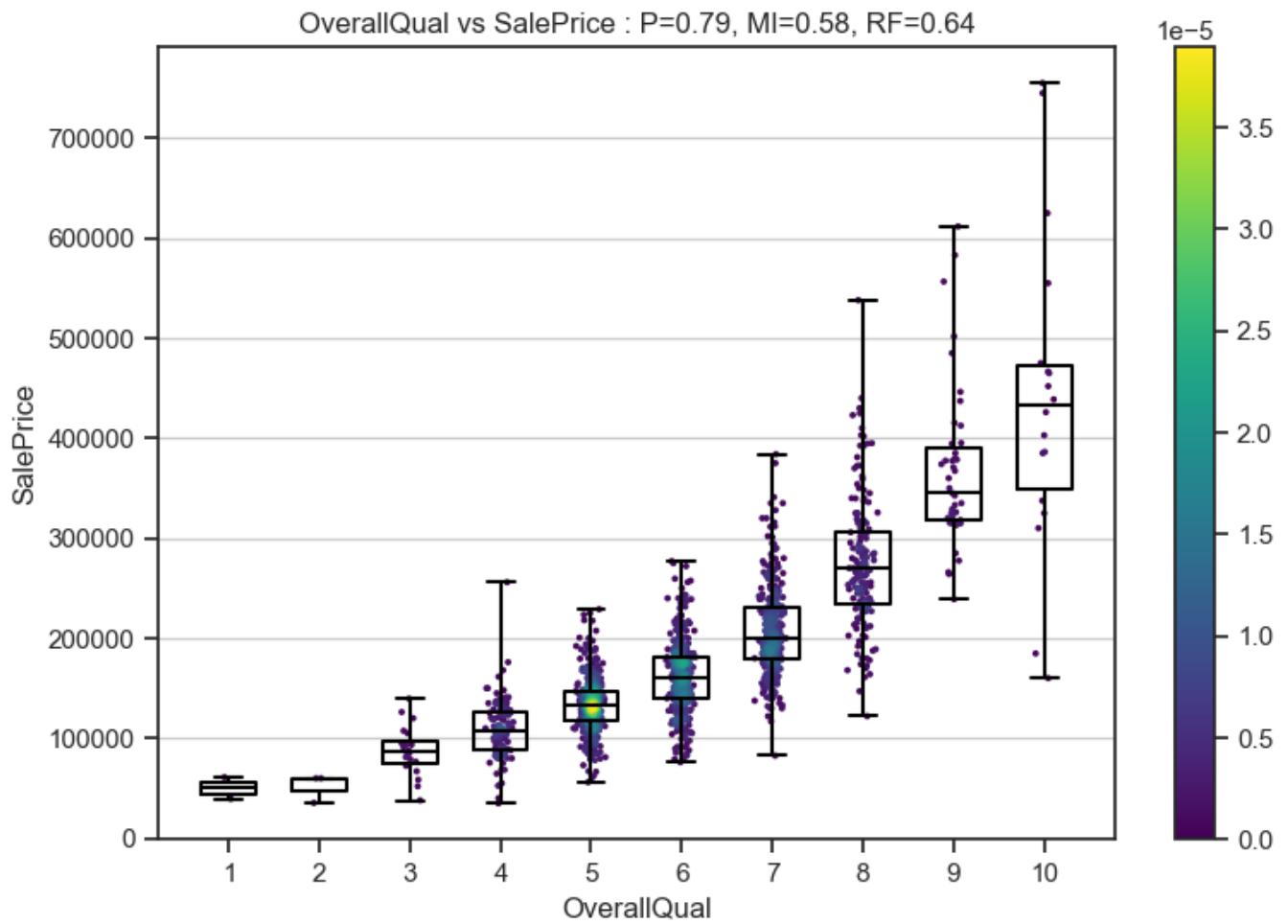
Correlations of Non-Numeric Features with Target Variable

Non-Numeric Feature	Count non-Null	Num Unique	Mutual Info	RF Corr	RF Corr (norm)
Neighborhood	1460	25	0.53	0.55	0.55
ExterQual	1460	4	0.32	0.48	0.51
KitchenQual	1460	4	0.33	0.46	0.49
BsmtQual	1423	4	0.32	0.45	0.48
Alley	91	2	0.28	0.29	0.38
GarageFinish	1379	3	0.23	0.27	0.3
Foundation	1460	6	0.2	0.26	0.26
GarageType	1379	6	0.17	0.21	0.21

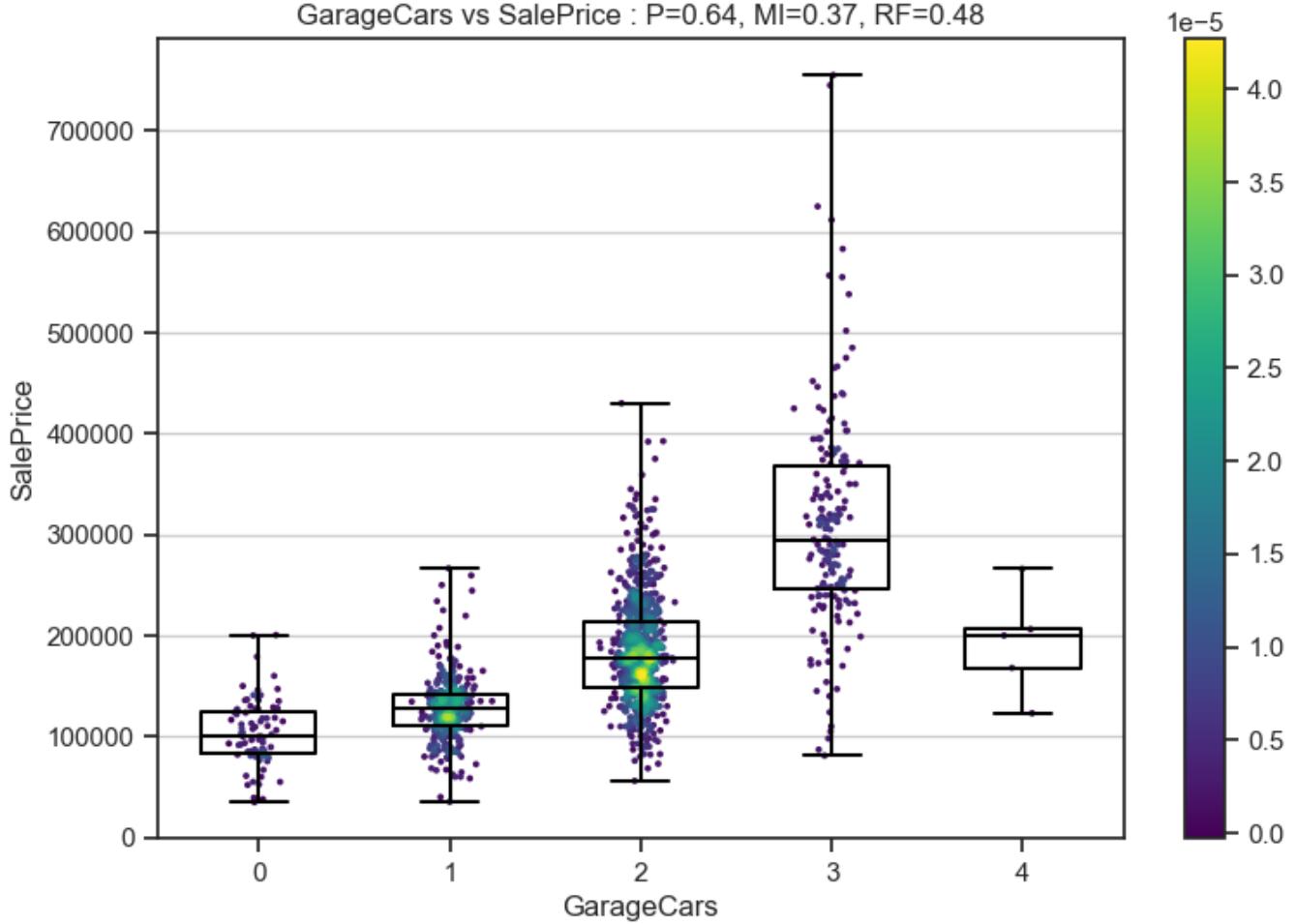


The above plot shows a histogram of all non-numeric features and their correlation value with the target variable.

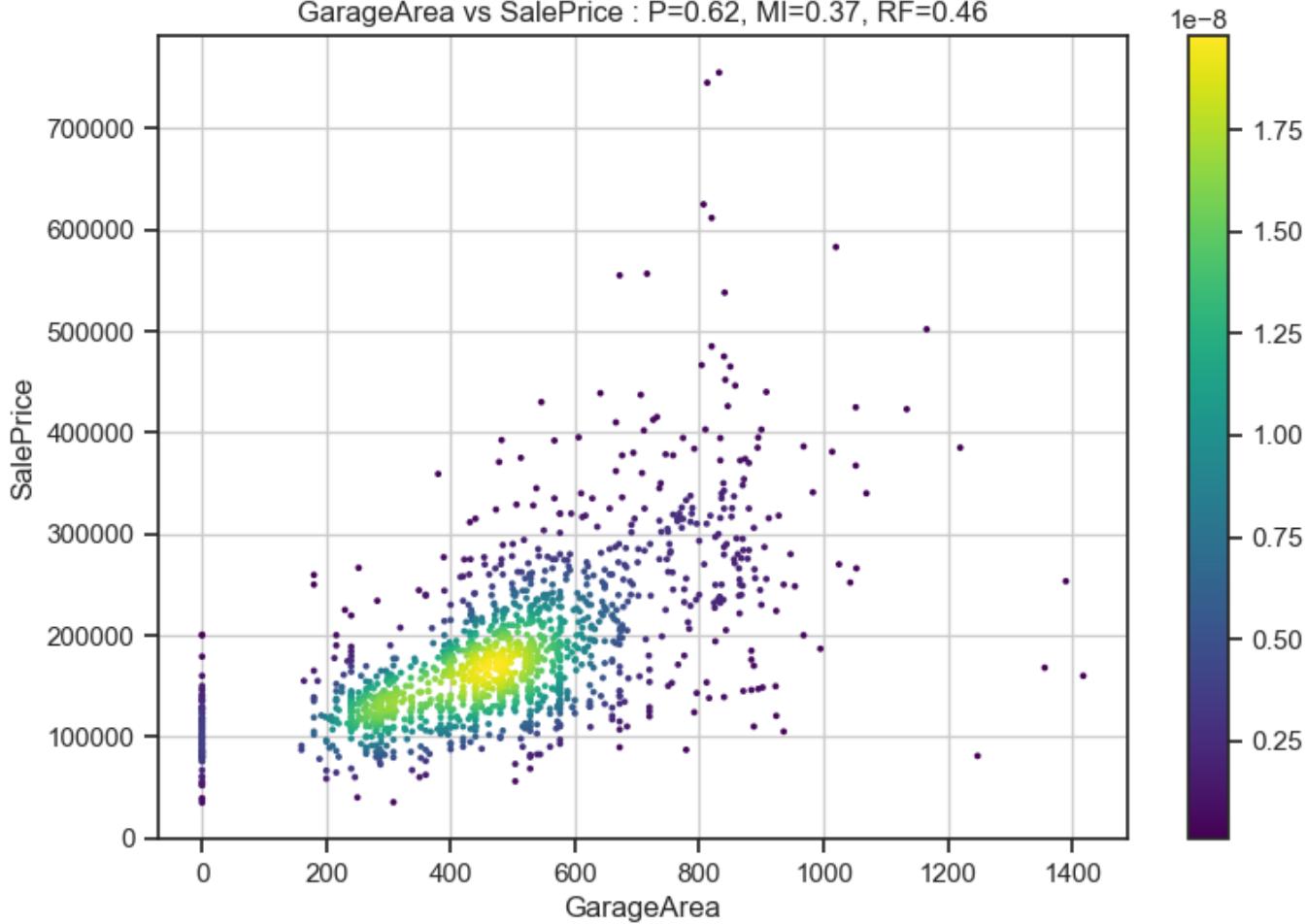
Plots of Numeric Columns versus the Target Variable



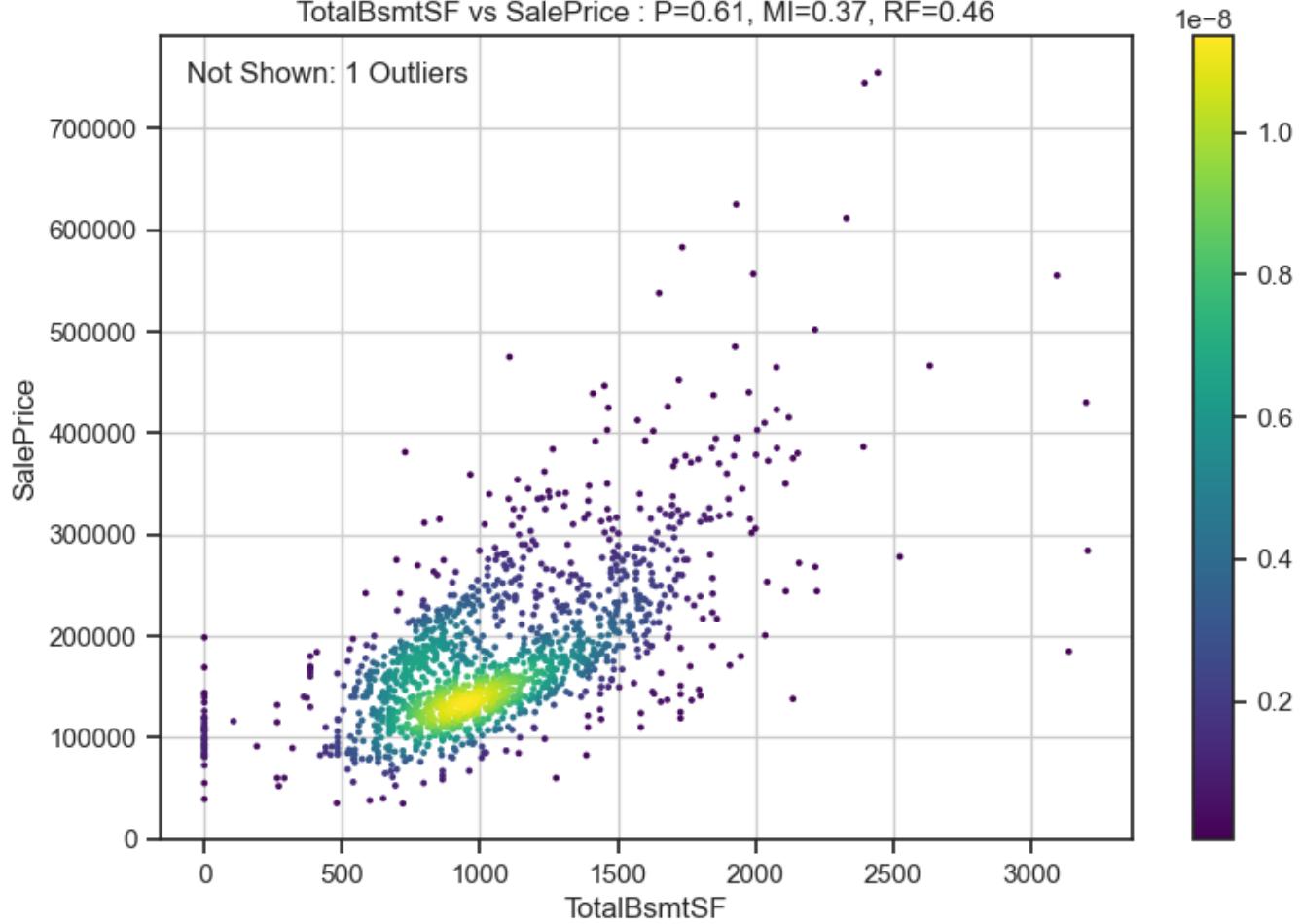
GarageCars vs SalePrice : P=0.64, MI=0.37, RF=0.48



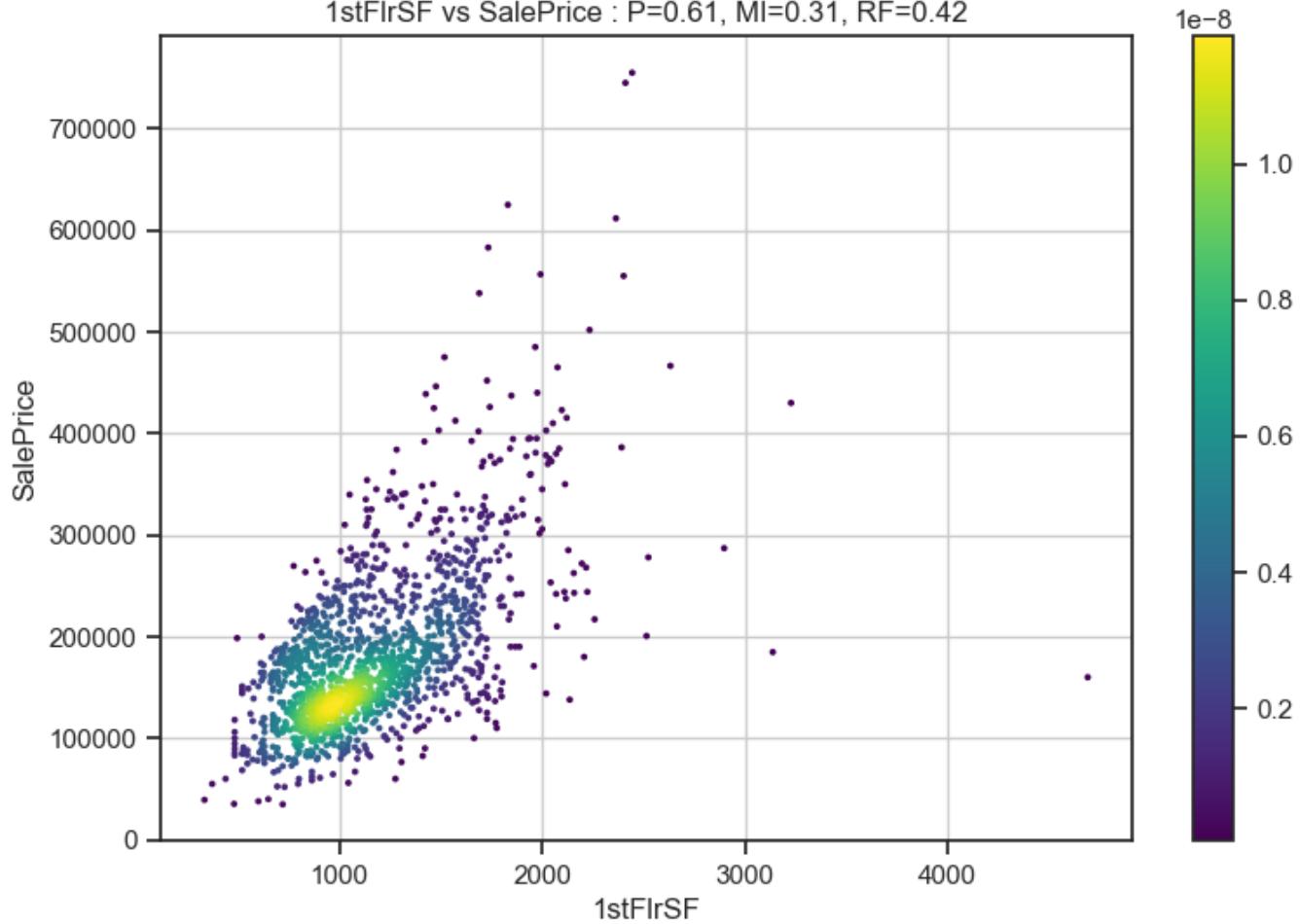
GarageArea vs SalePrice : P=0.62, MI=0.37, RF=0.46



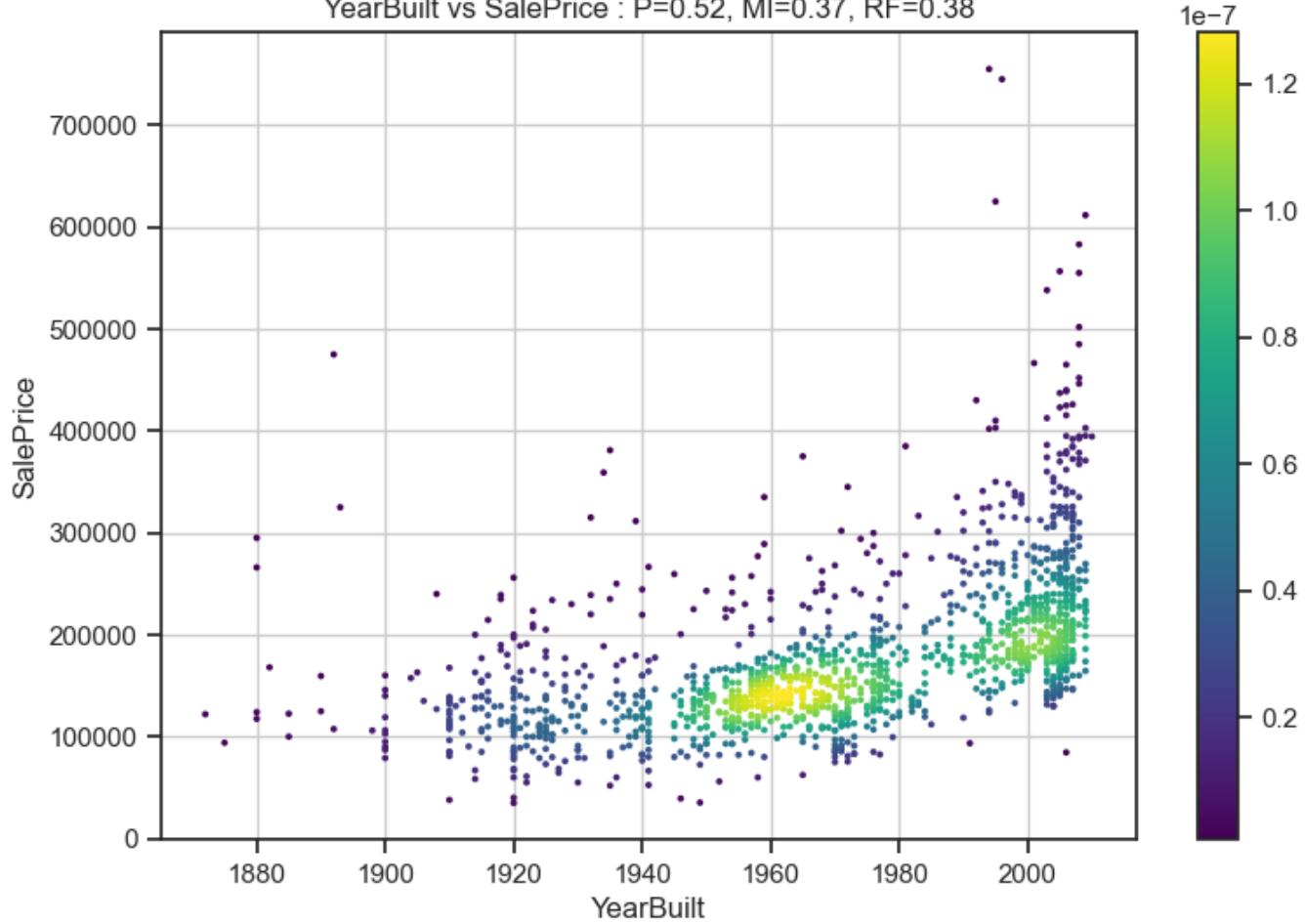
TotalBsmtSF vs SalePrice : P=0.61, MI=0.37, RF=0.46



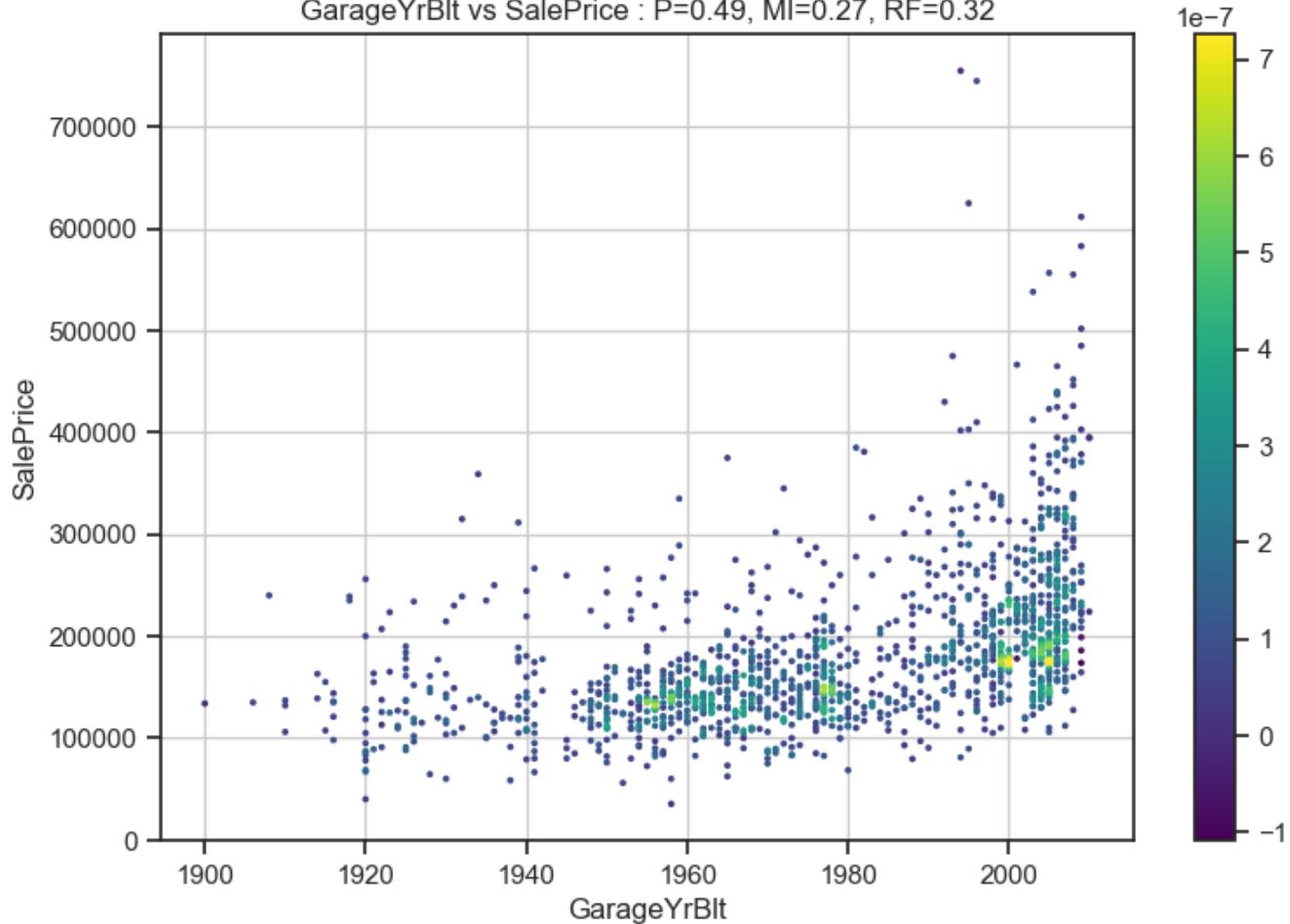
1stFlrSF vs SalePrice : P=0.61, MI=0.31, RF=0.42



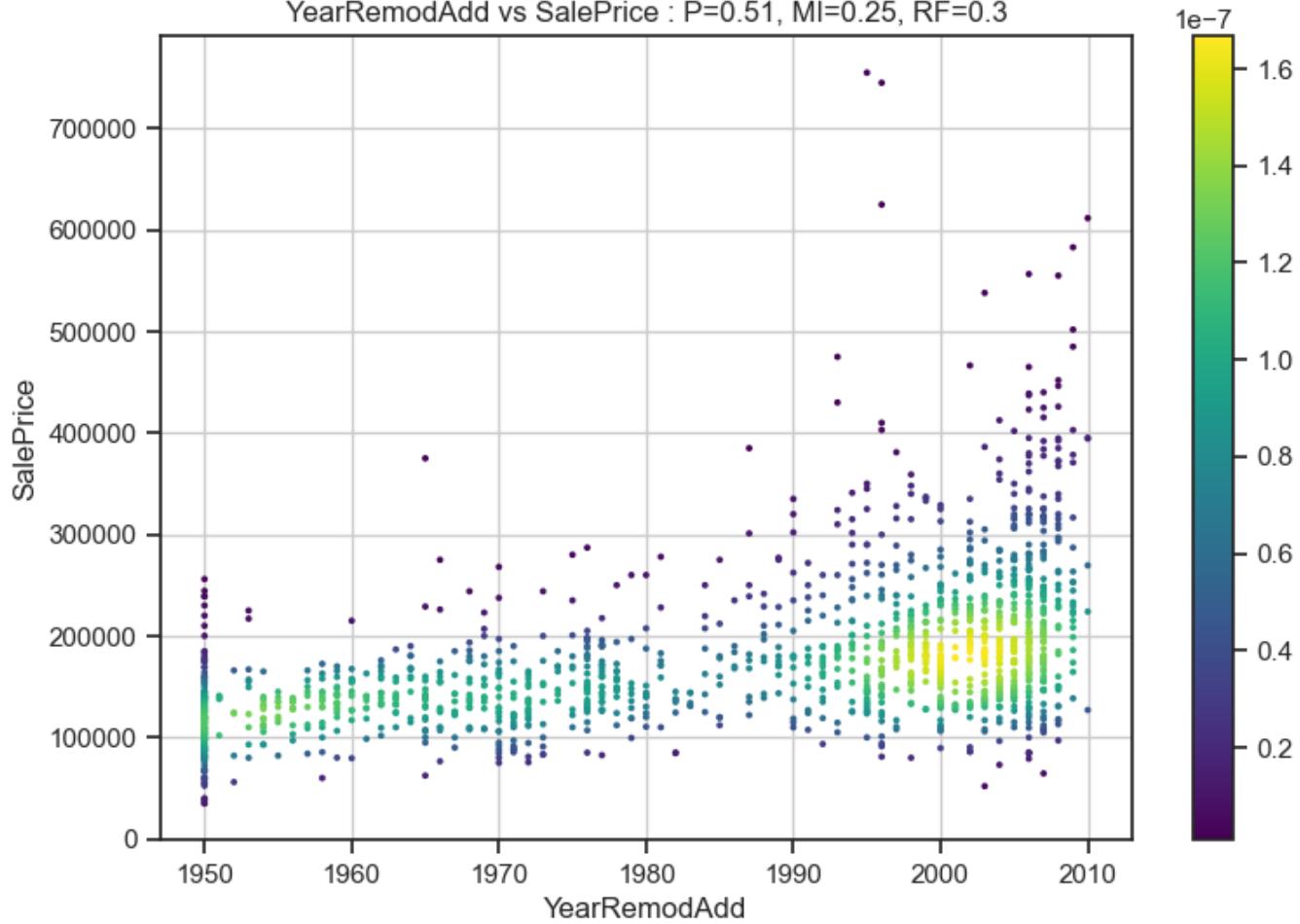
YearBuilt vs SalePrice : P=0.52, MI=0.37, RF=0.38



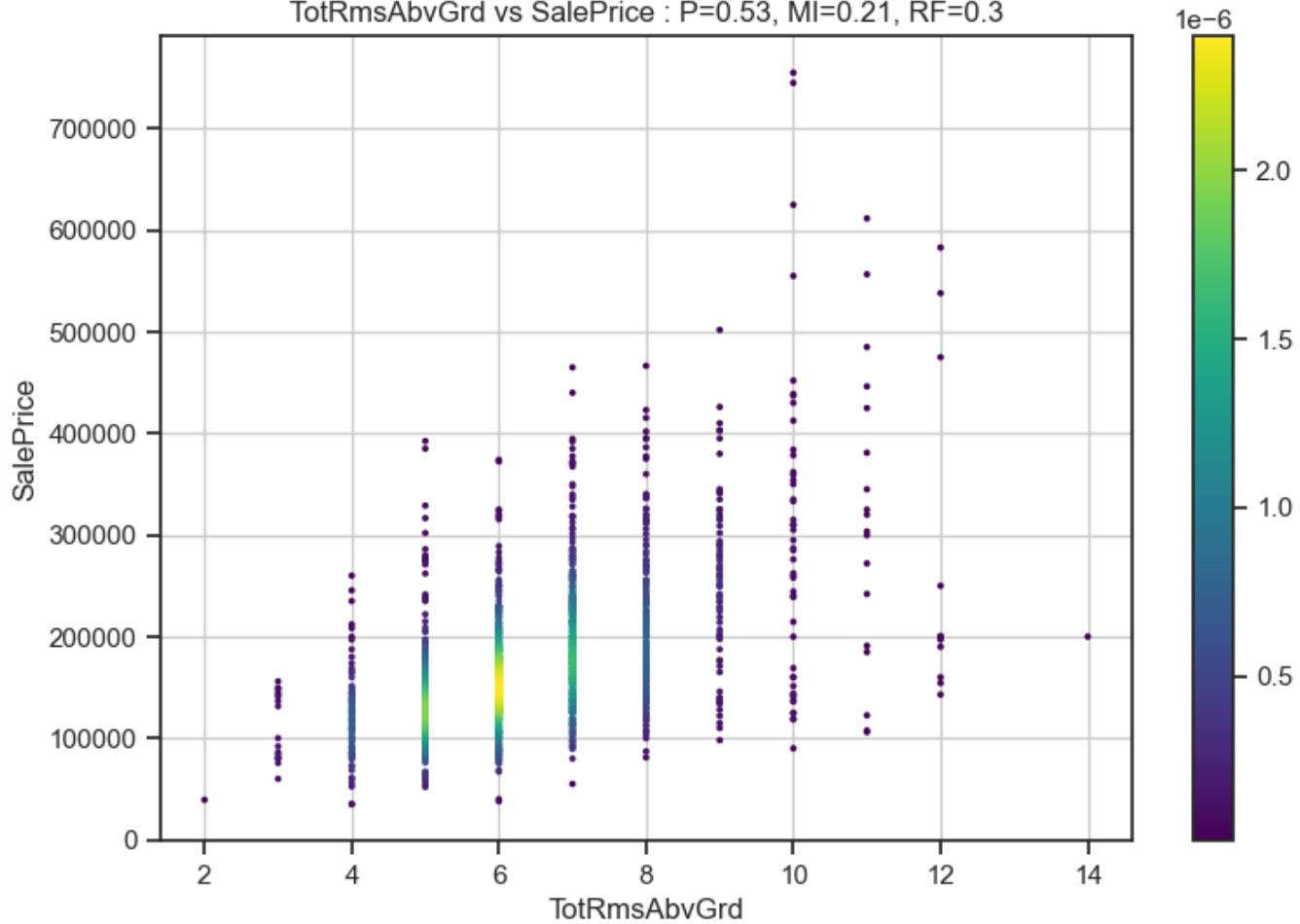
GarageYrBlt vs SalePrice : P=0.49, MI=0.27, RF=0.32



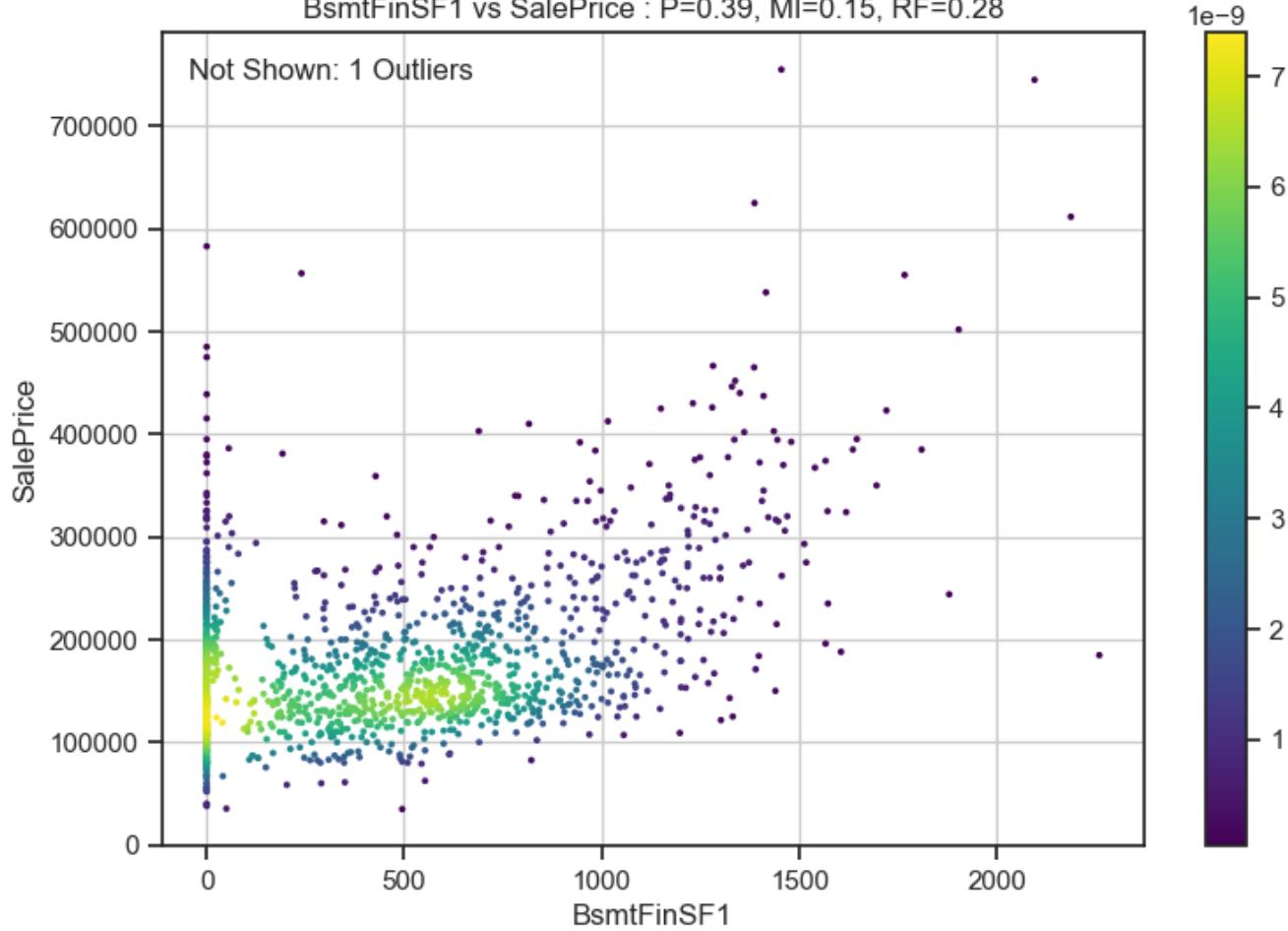
YearRemodAdd vs SalePrice : P=0.51, MI=0.25, RF=0.3



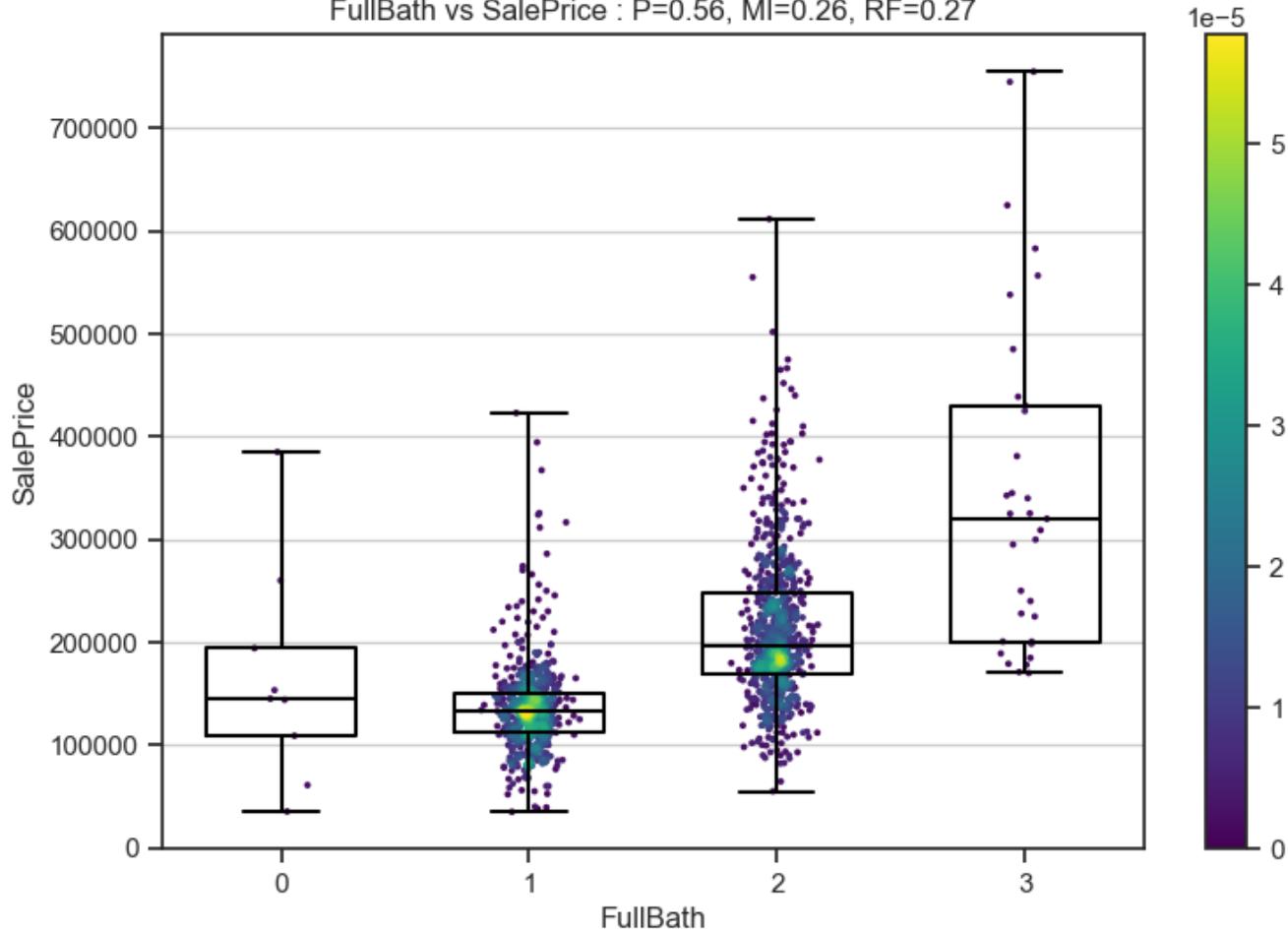
TotRmsAbvGrd vs SalePrice : P=0.53, MI=0.21, RF=0.3



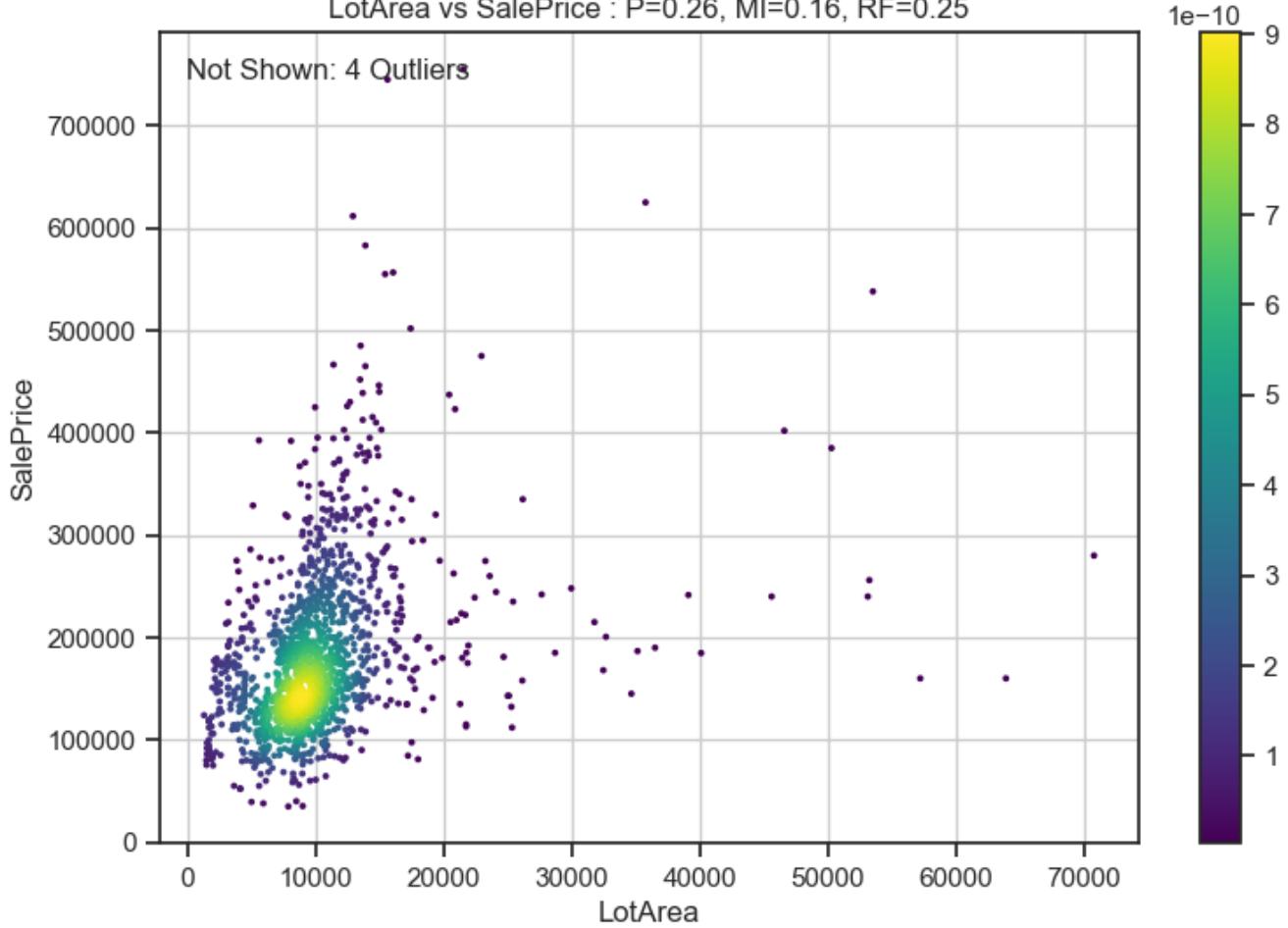
BsmtFinSF1 vs SalePrice : P=0.39, MI=0.15, RF=0.28



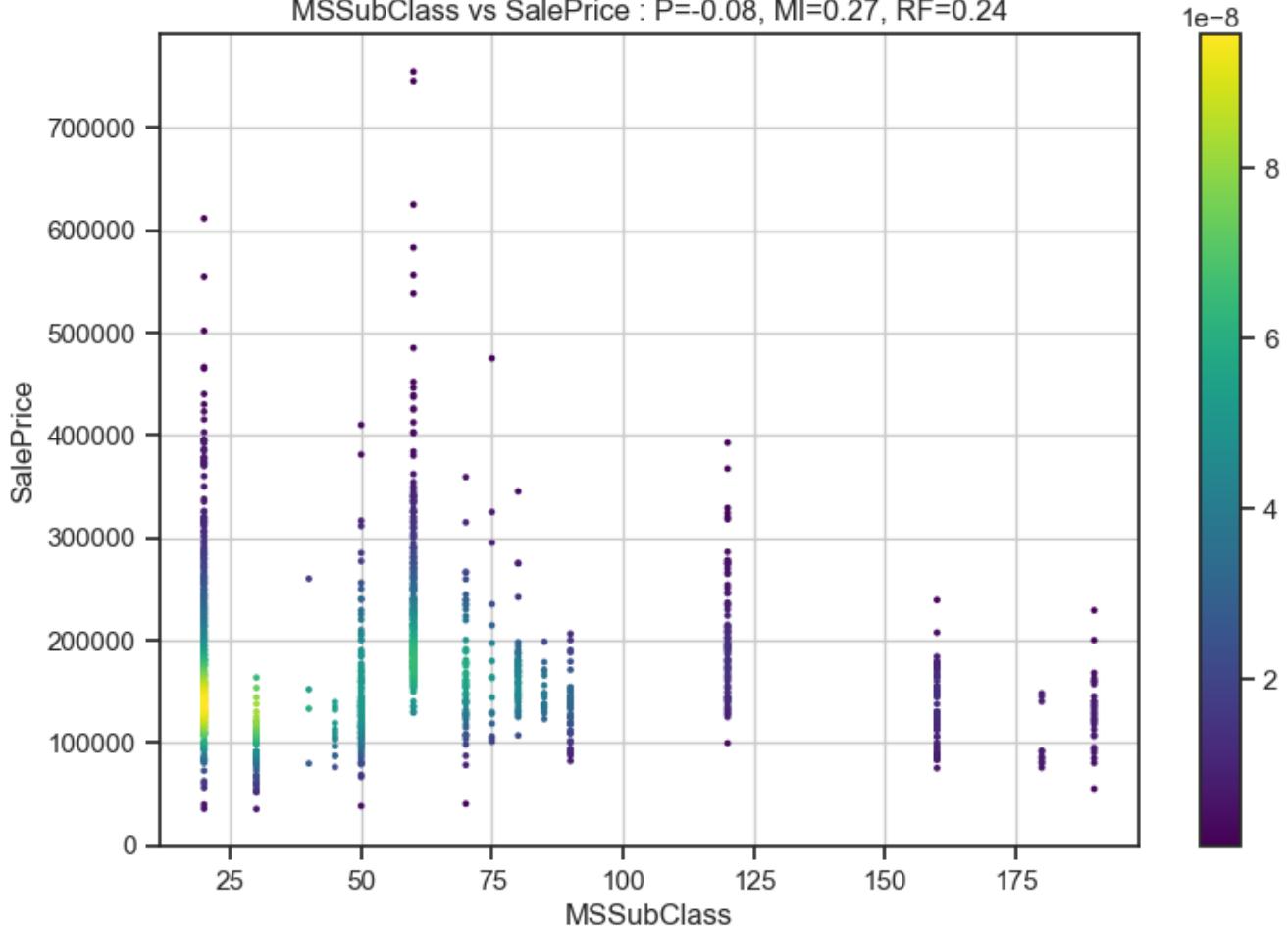
FullBath vs SalePrice : P=0.56, MI=0.26, RF=0.27

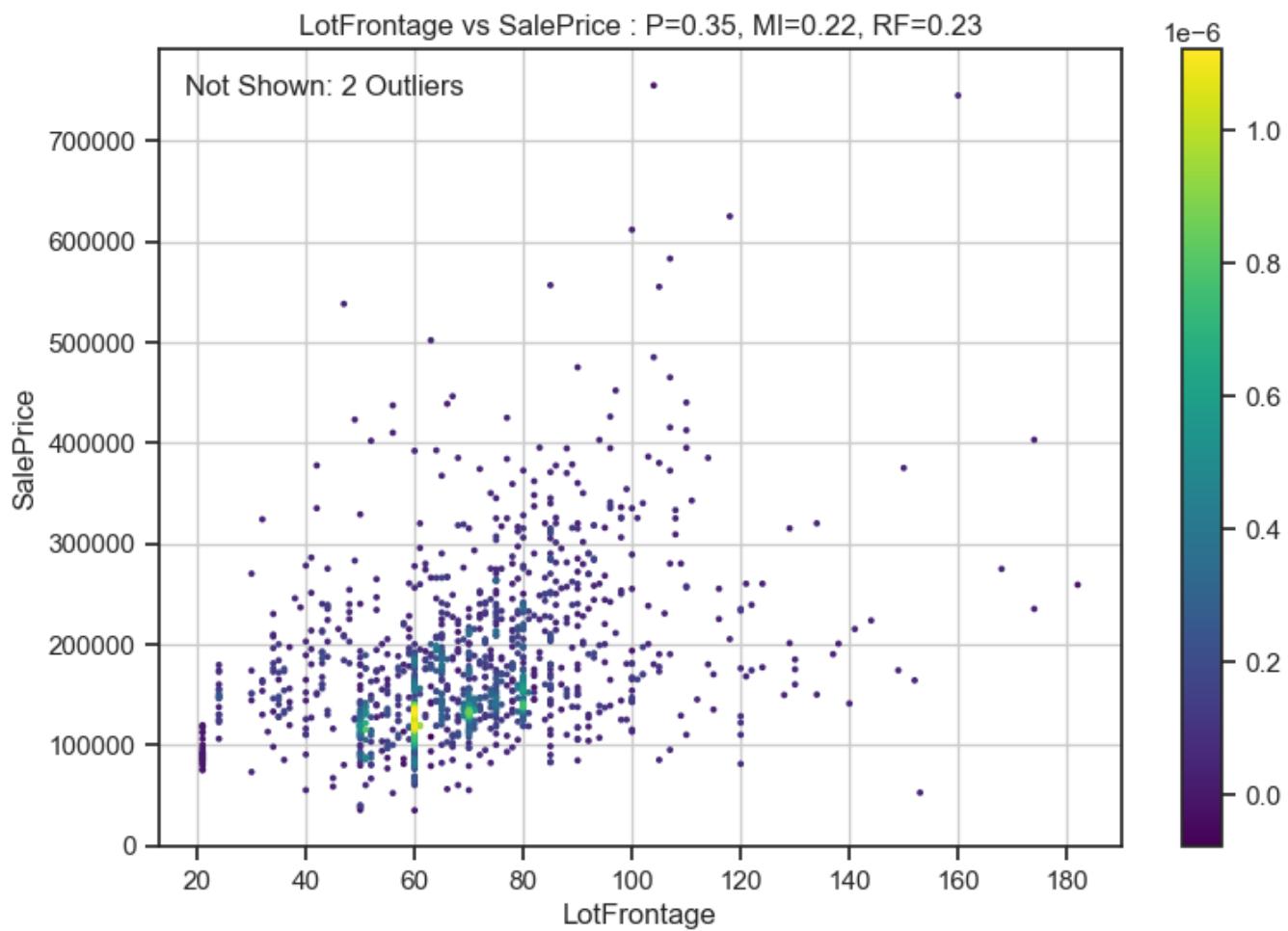
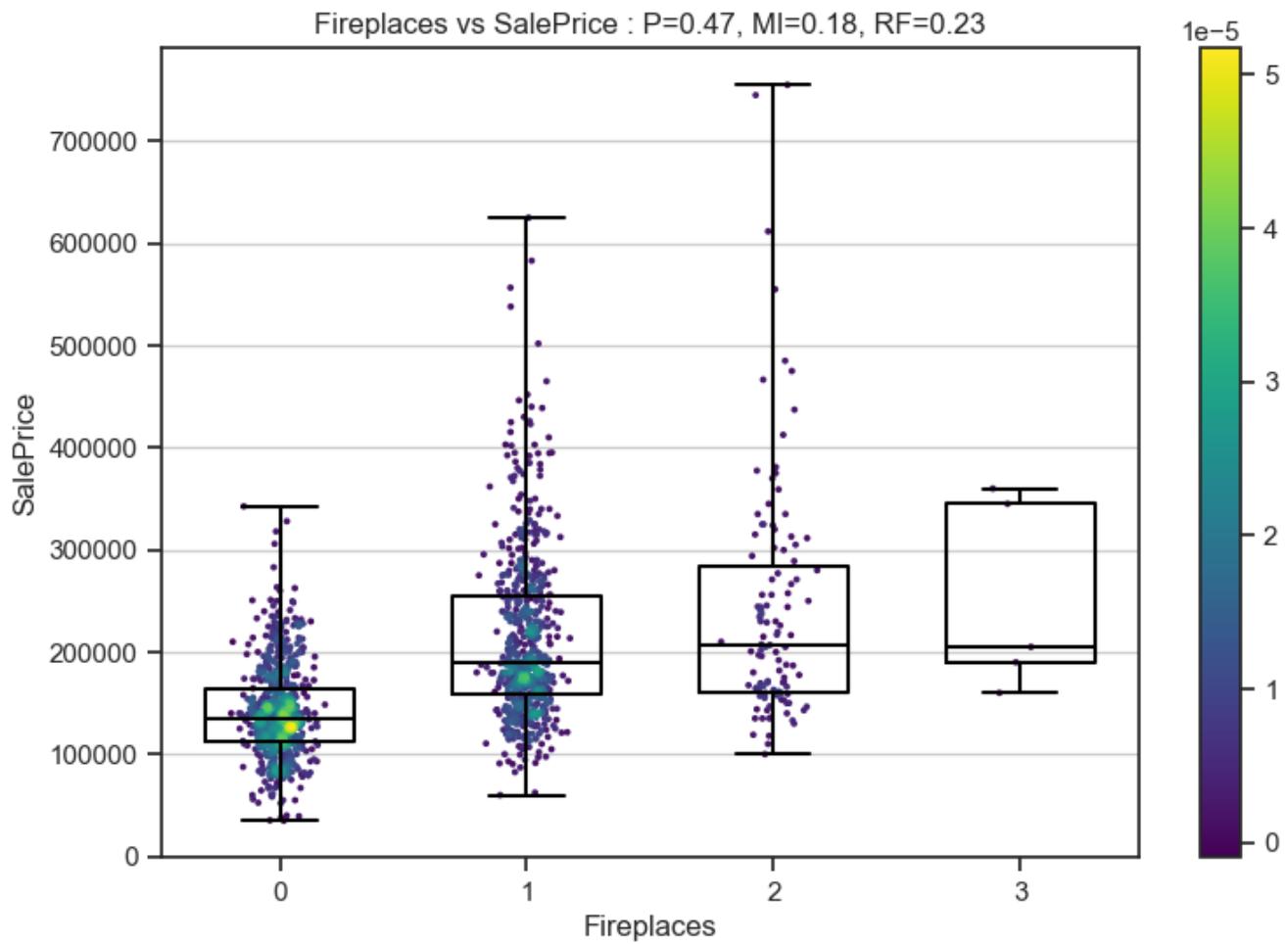


LotArea vs SalePrice : P=0.26, MI=0.16, RF=0.25

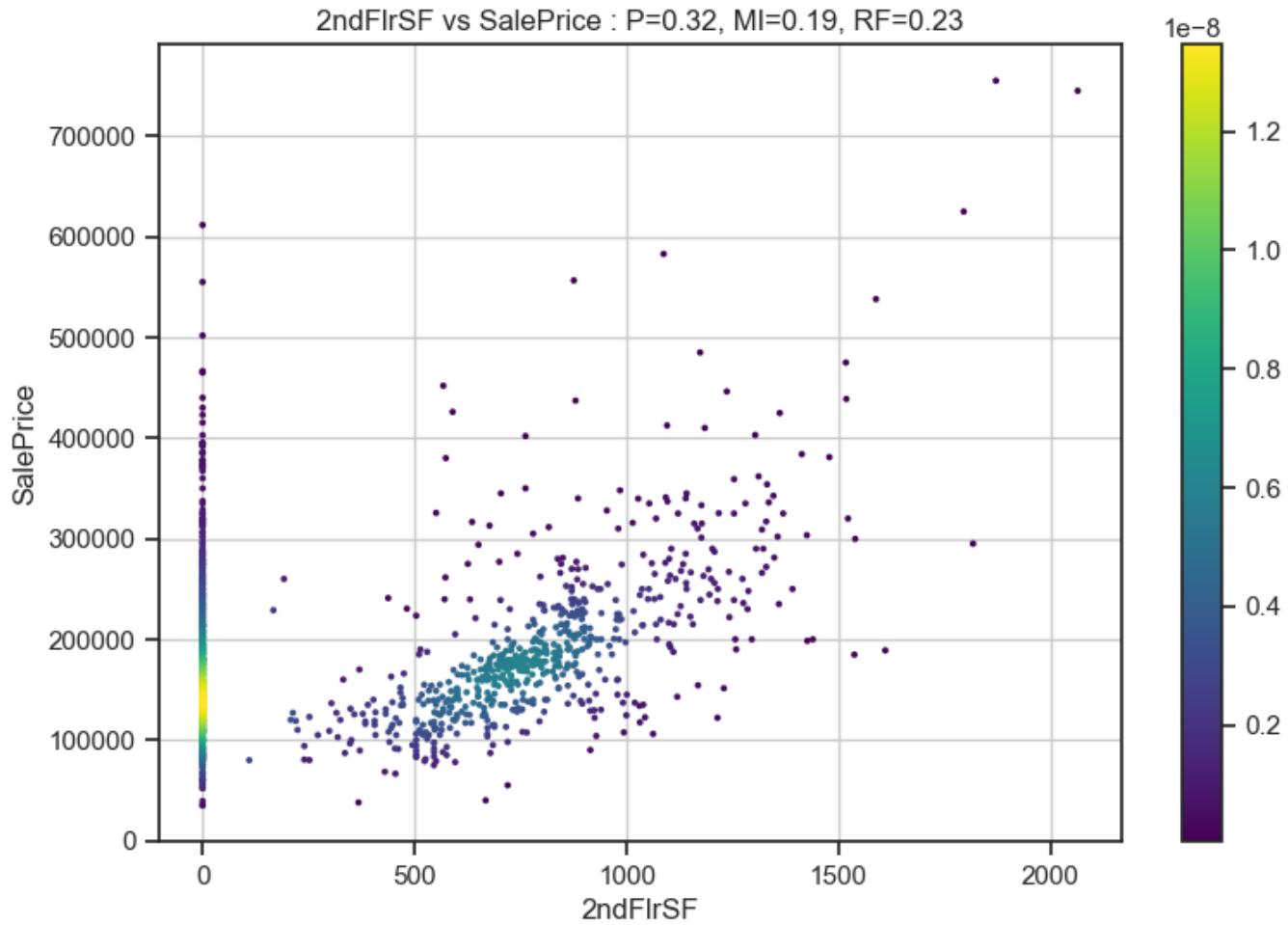


MSSubClass vs SalePrice : P=-0.08, MI=0.27, RF=0.24

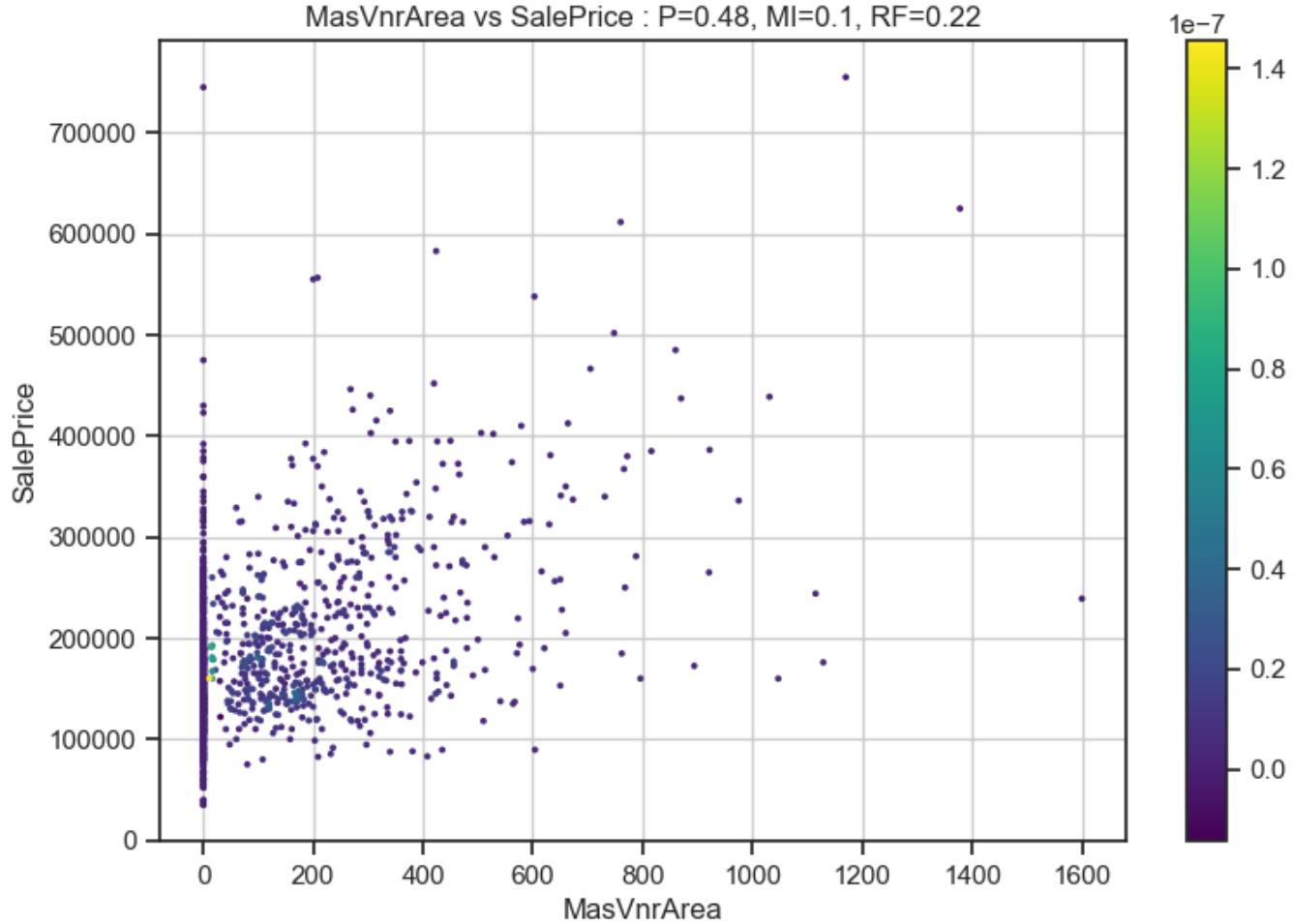




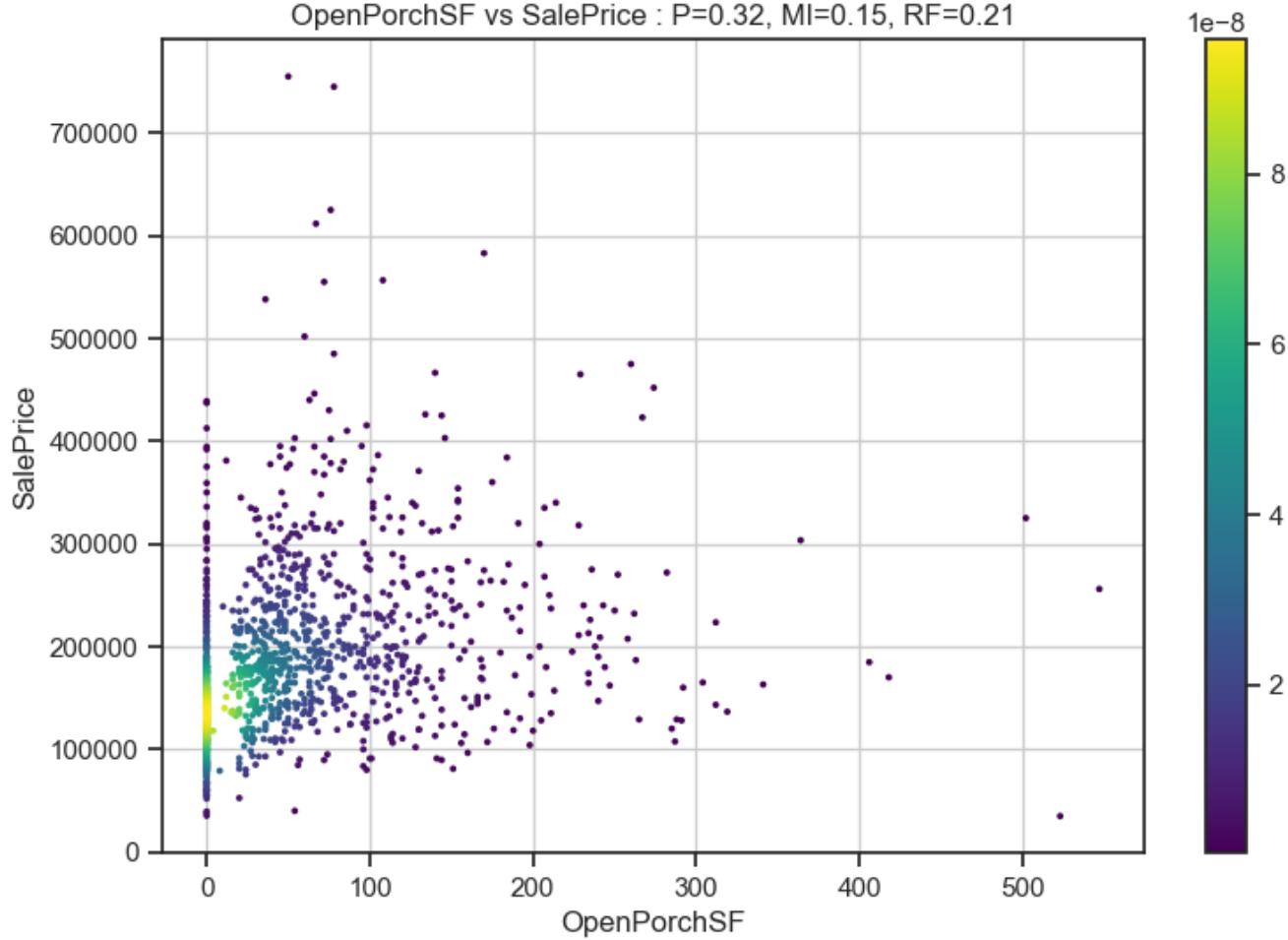
2ndFlrSF vs SalePrice : P=0.32, MI=0.19, RF=0.23



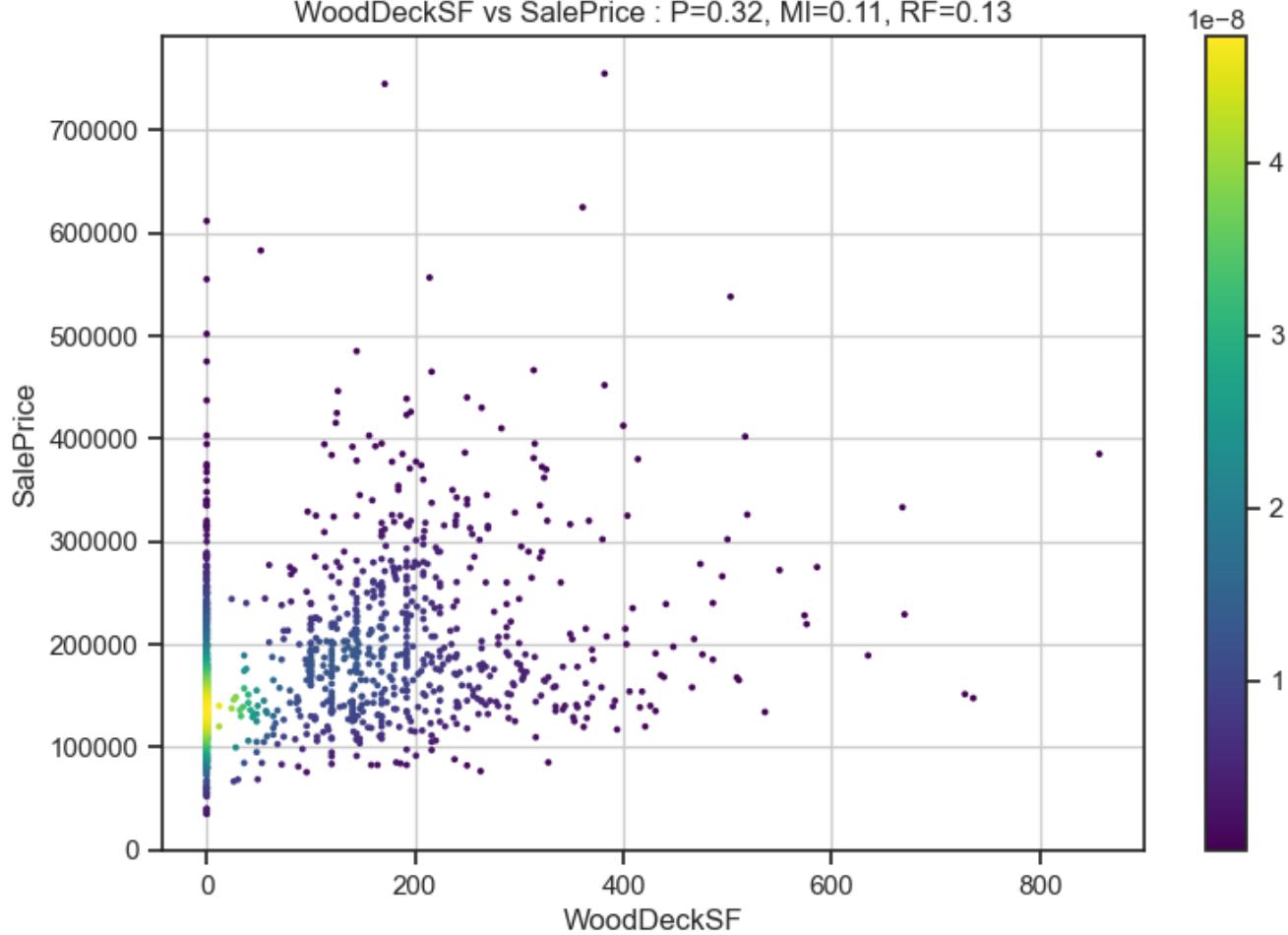
MasVnrArea vs SalePrice : P=0.48, MI=0.1, RF=0.22

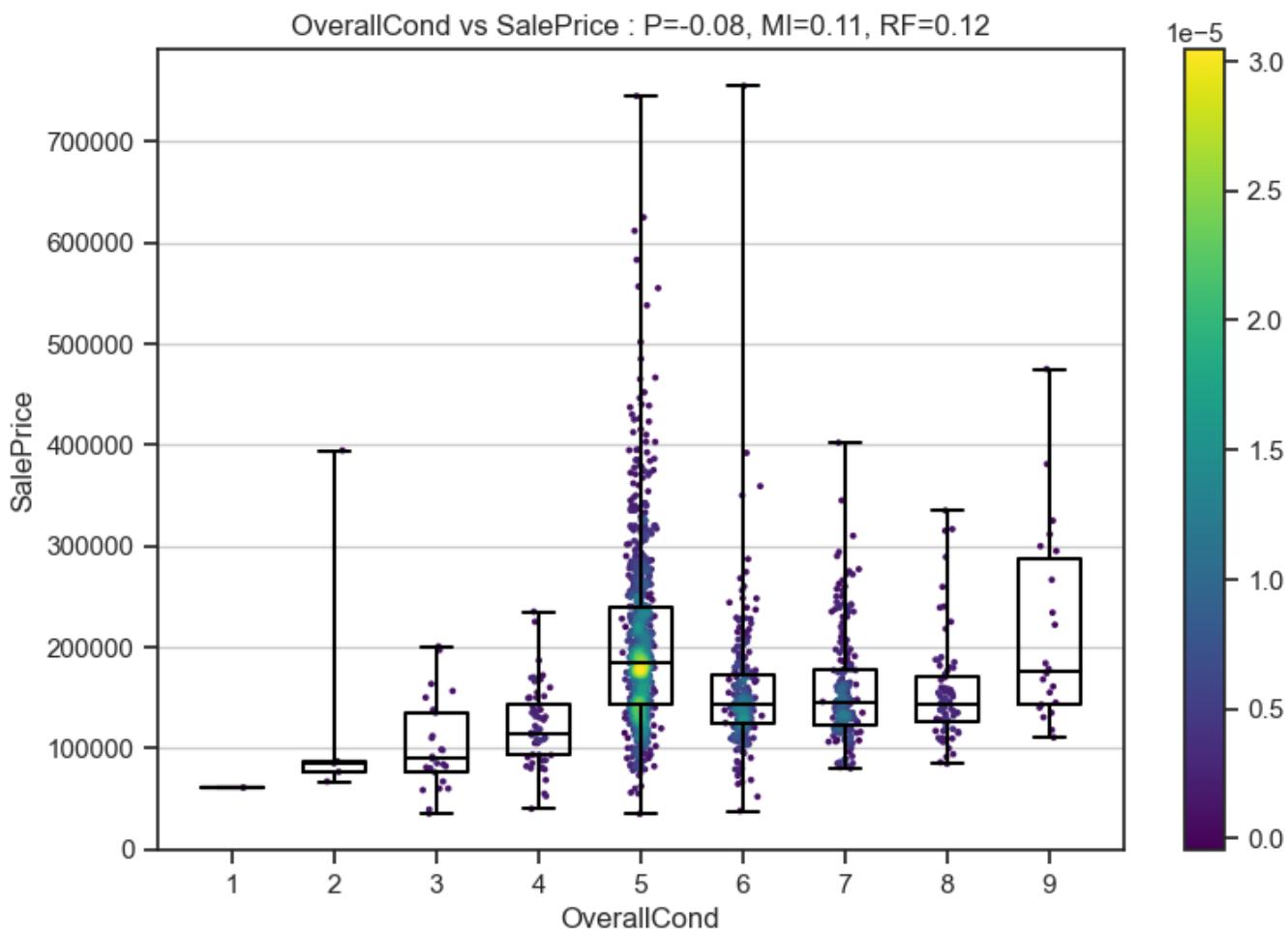
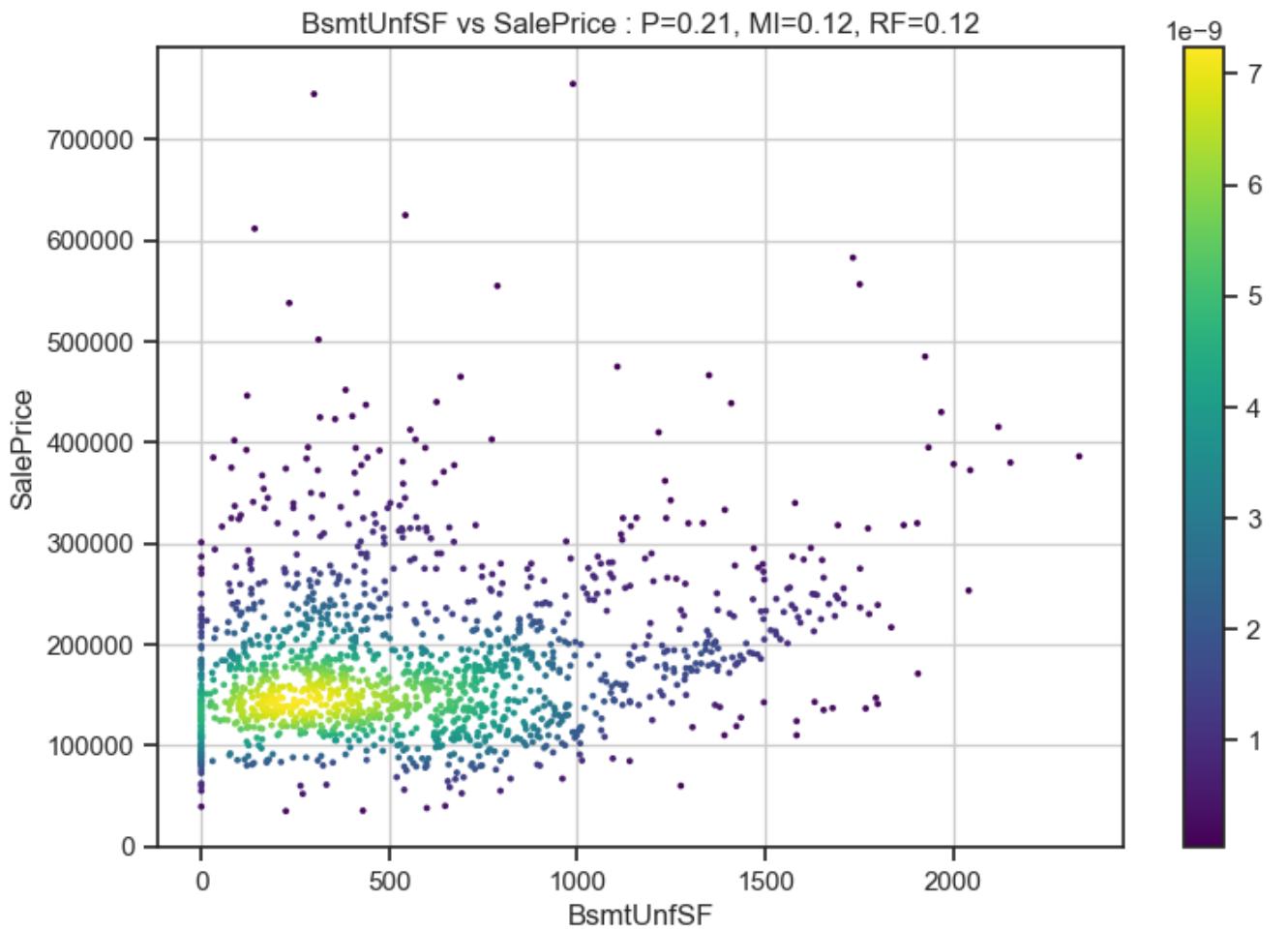


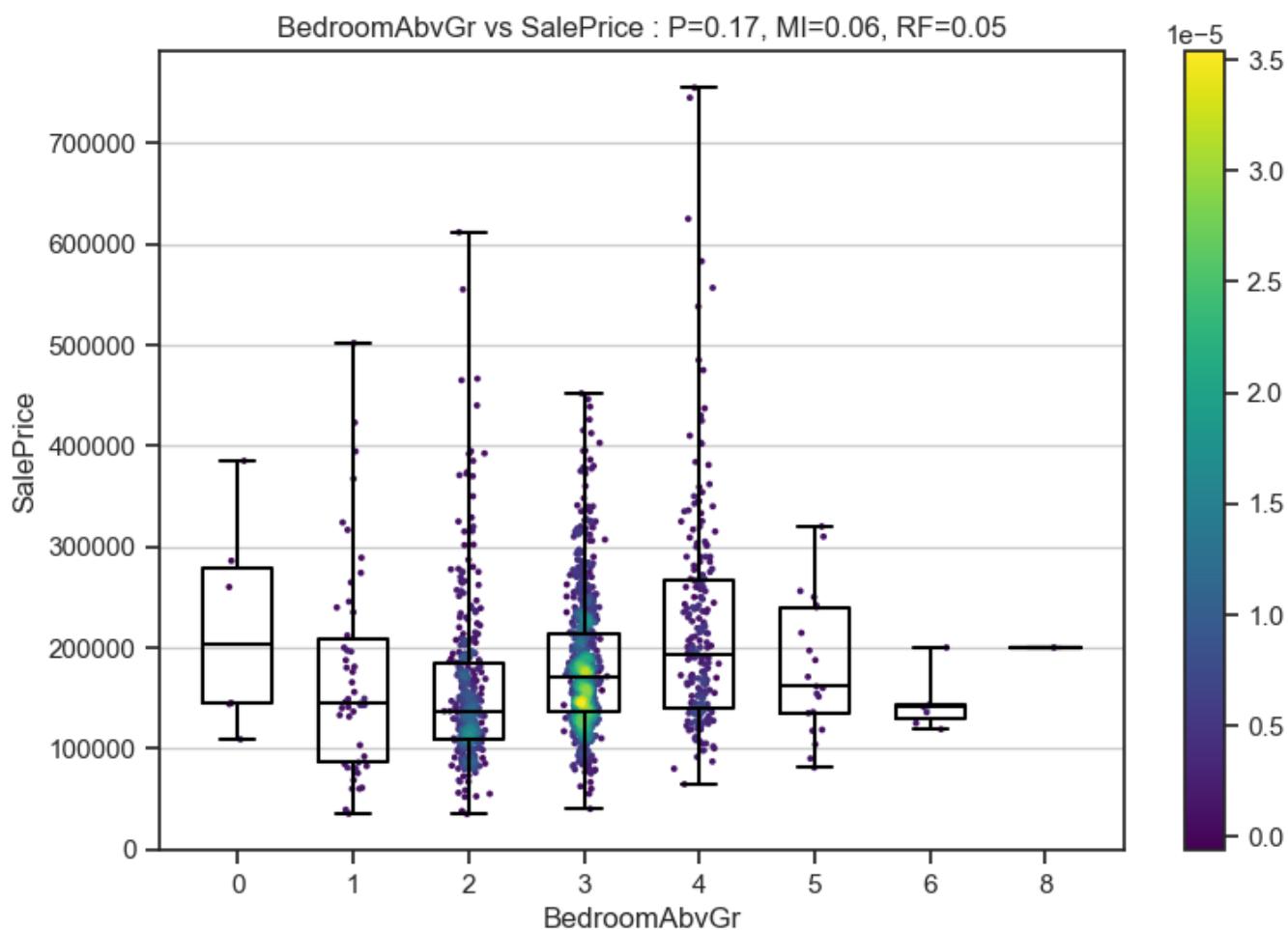
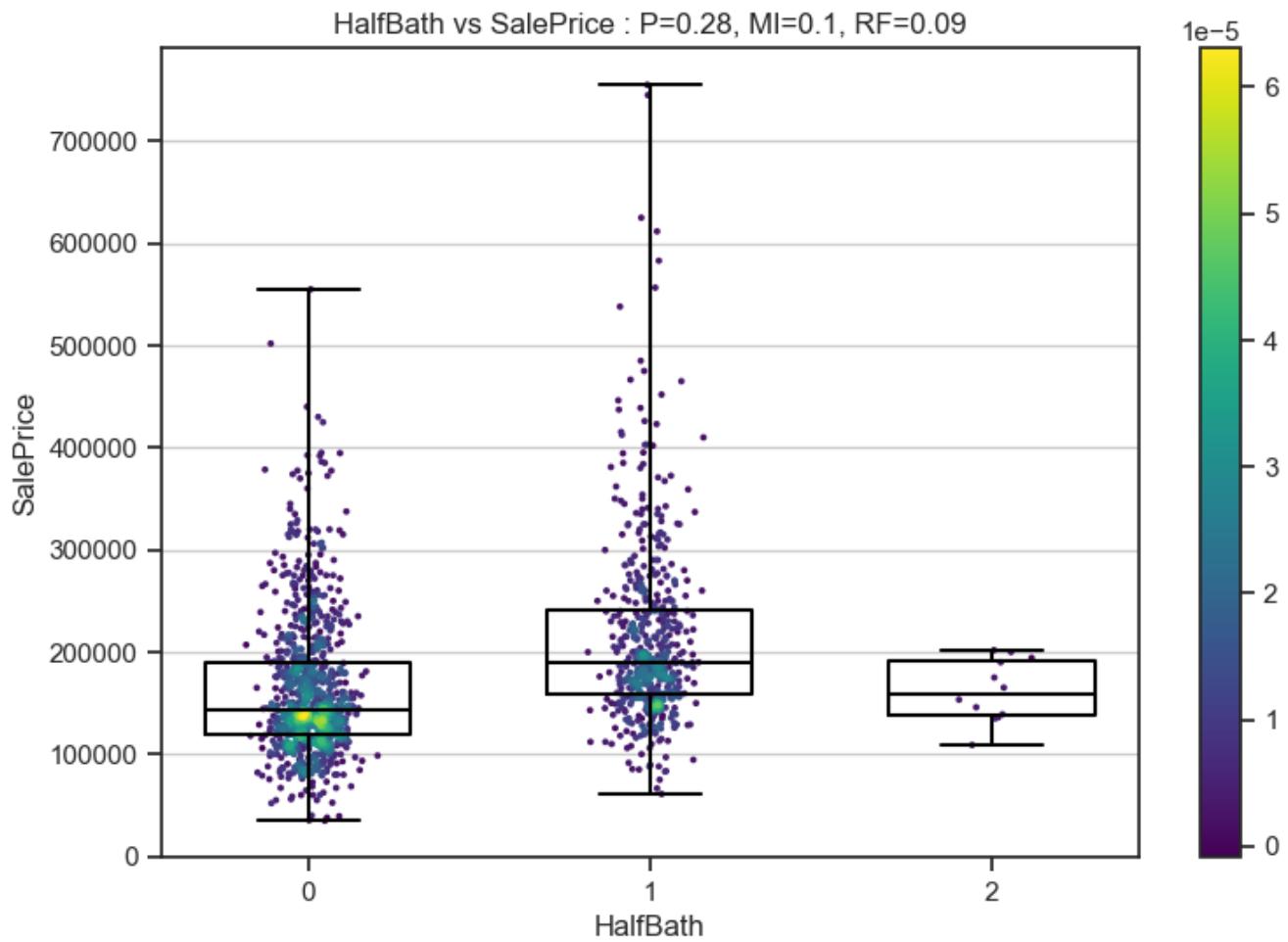
OpenPorchSF vs SalePrice : P=0.32, MI=0.15, RF=0.21

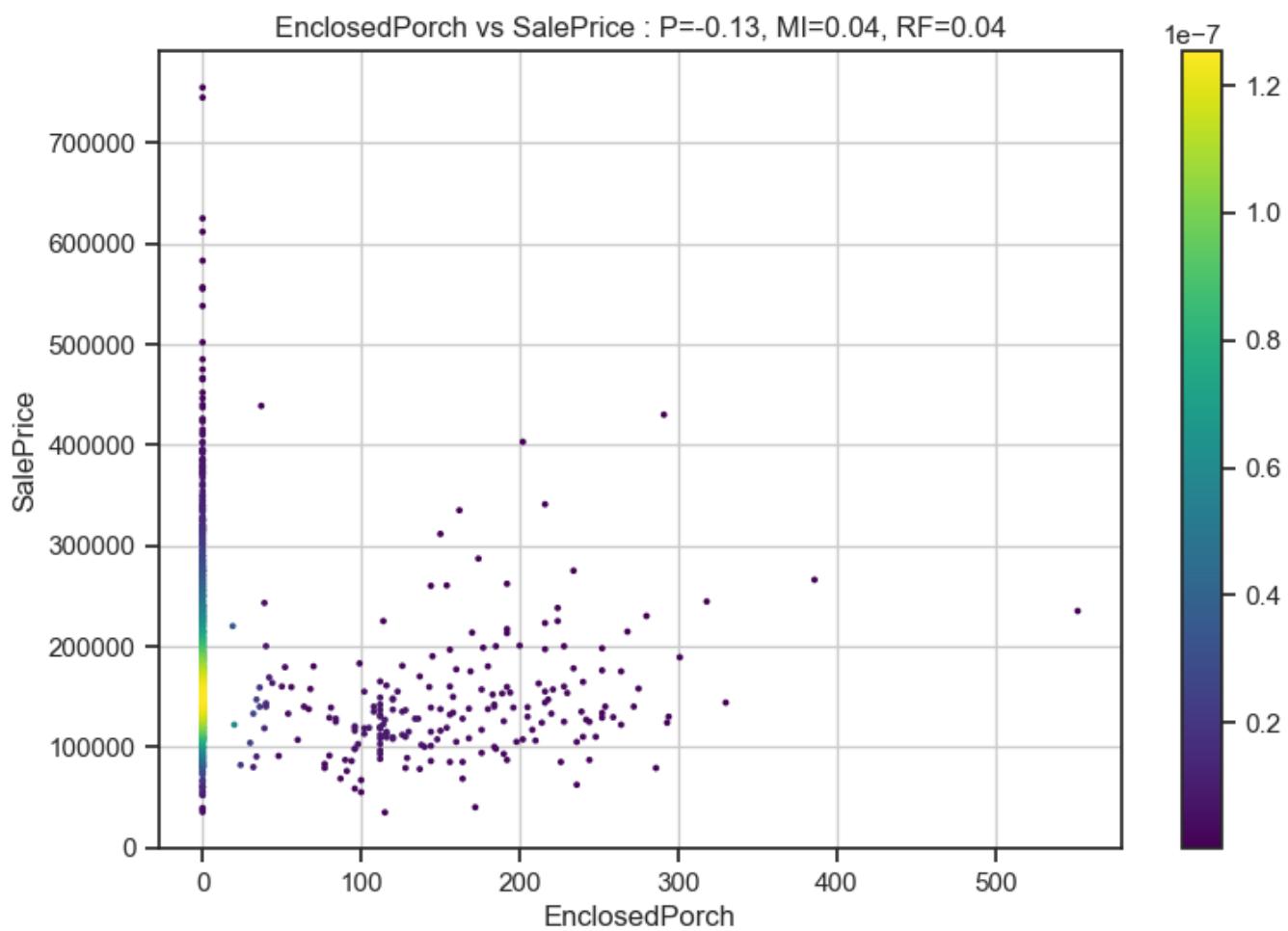
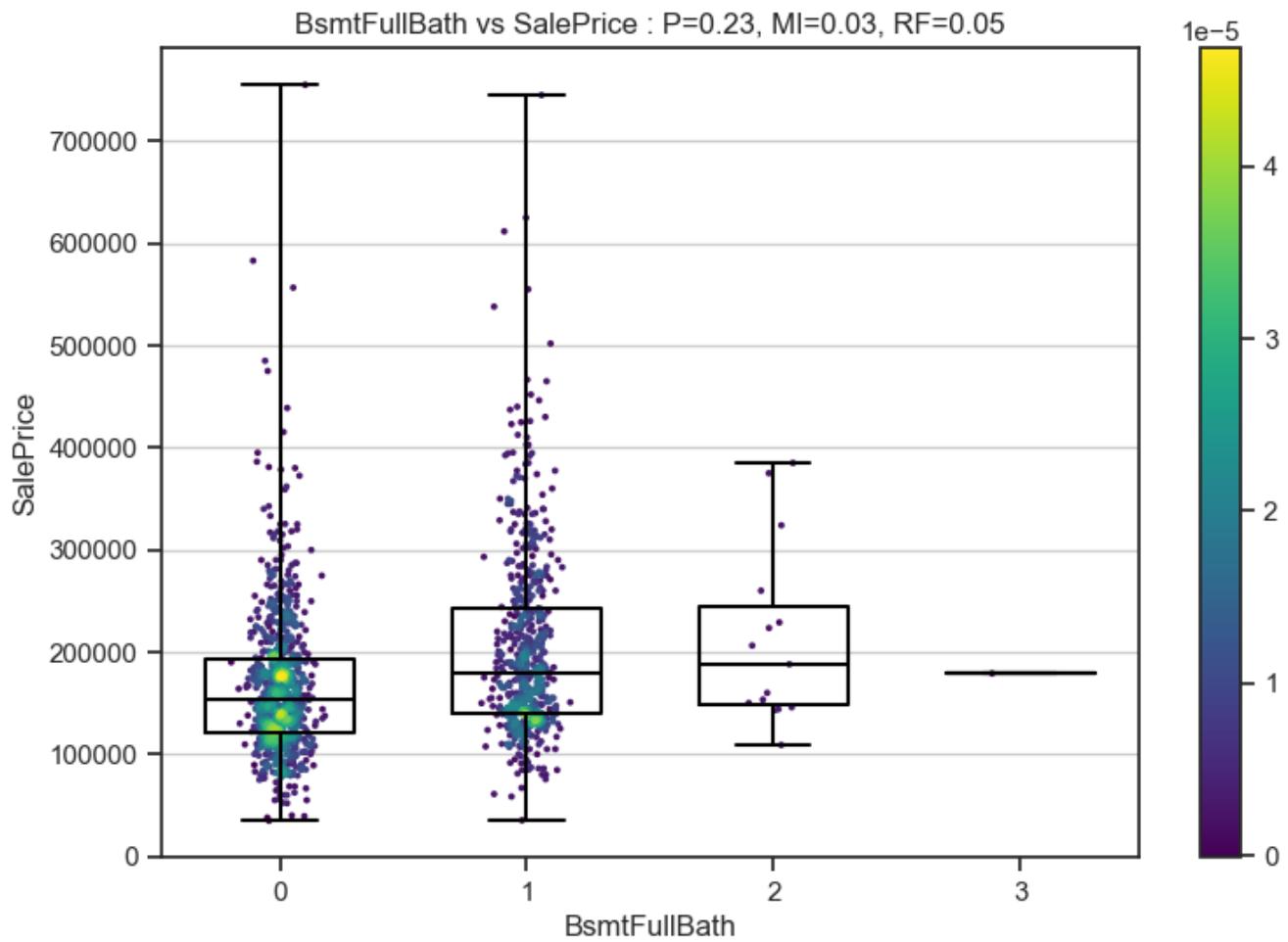


WoodDeckSF vs SalePrice : P=0.32, MI=0.11, RF=0.13

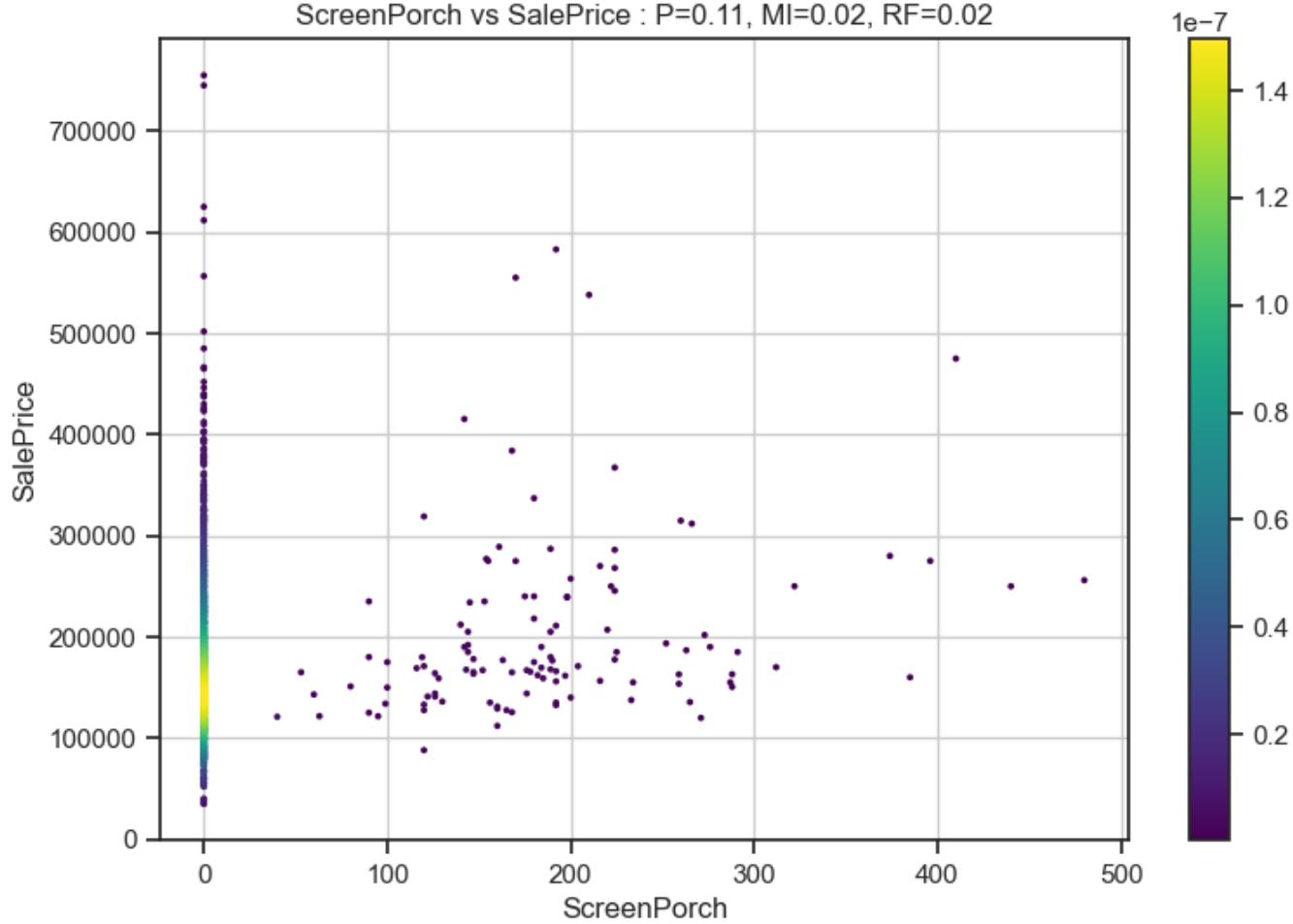




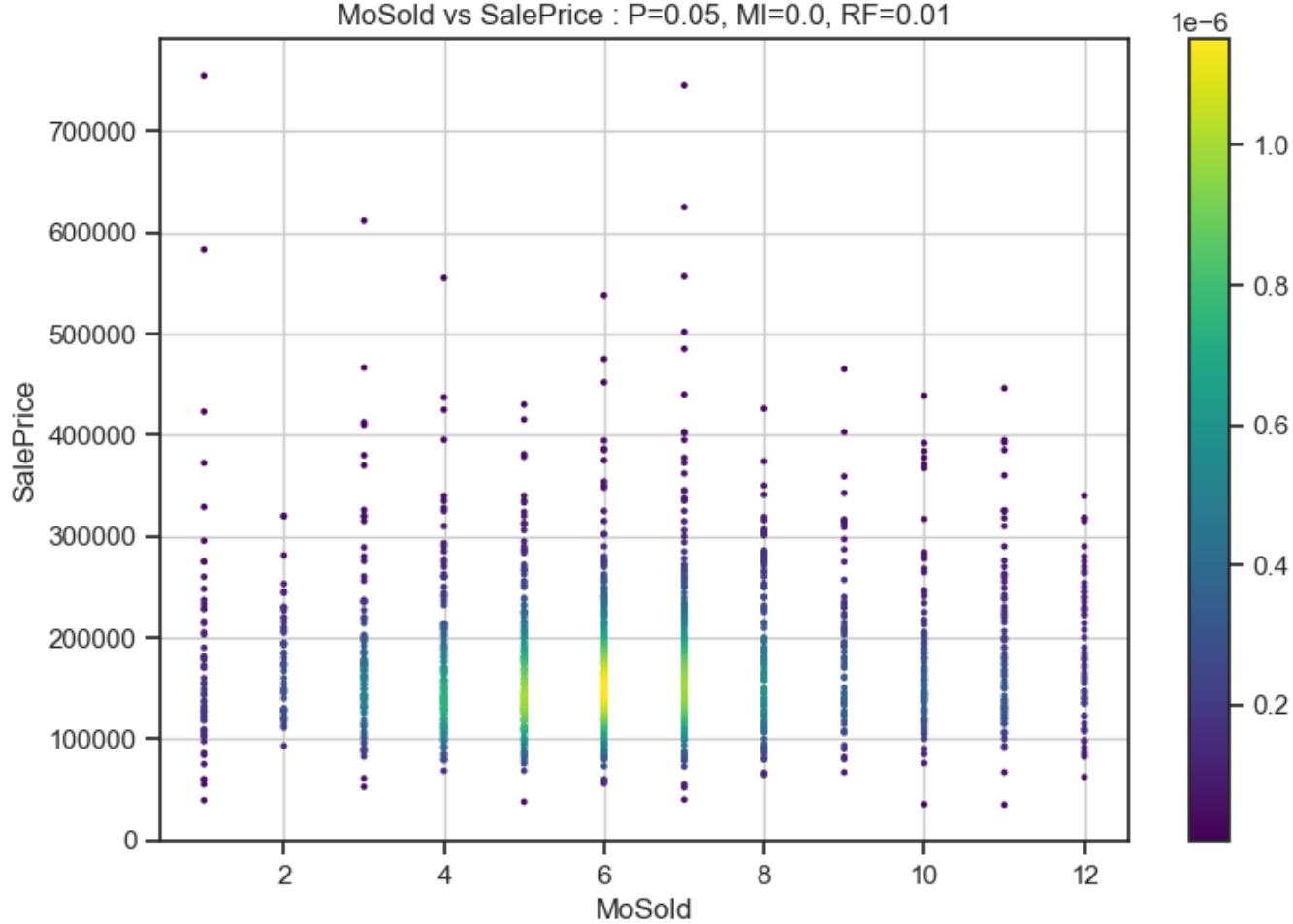




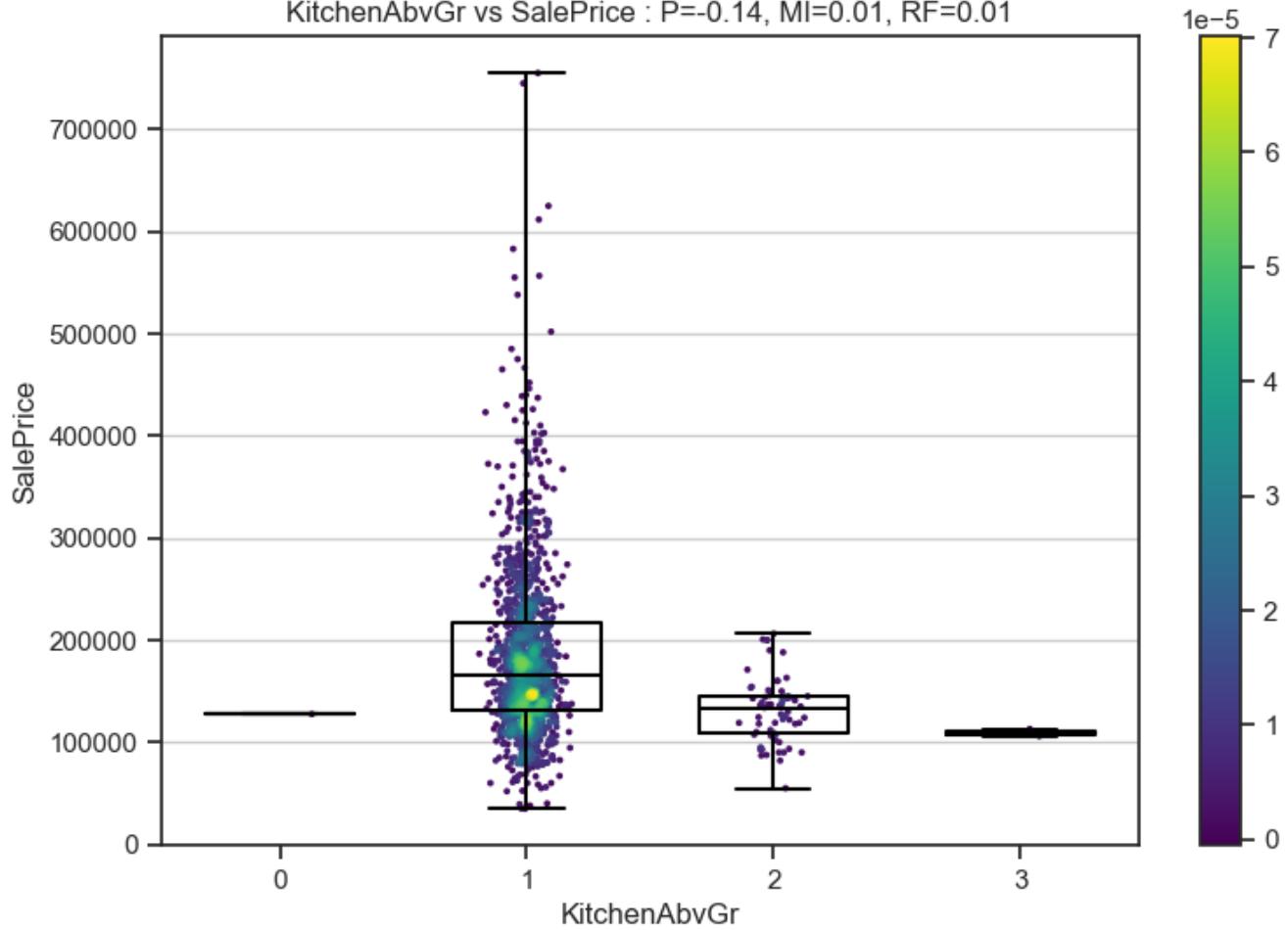
ScreenPorch vs SalePrice : P=0.11, MI=0.02, RF=0.02



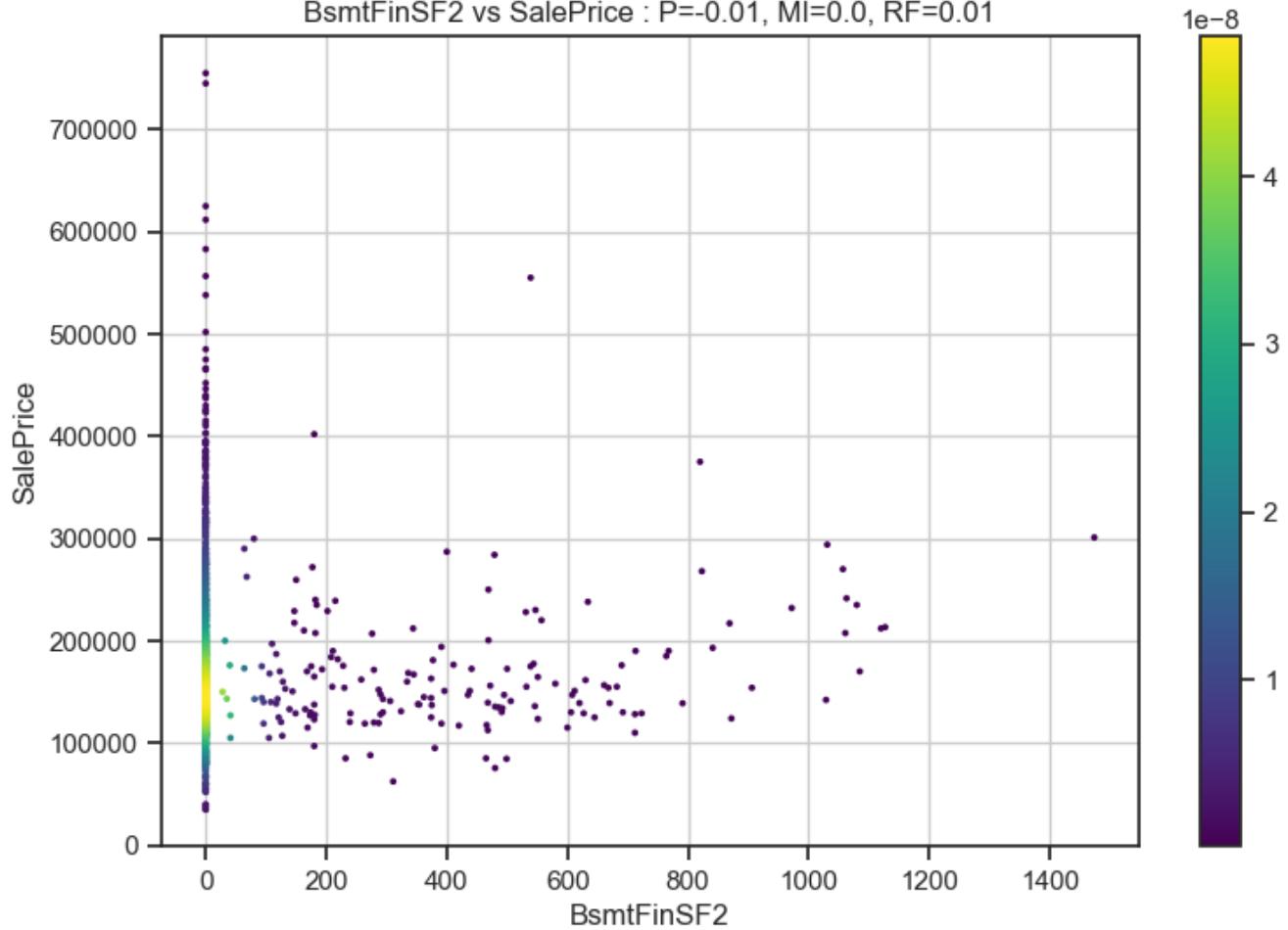
MoSold vs SalePrice : P=0.05, MI=0.0, RF=0.01



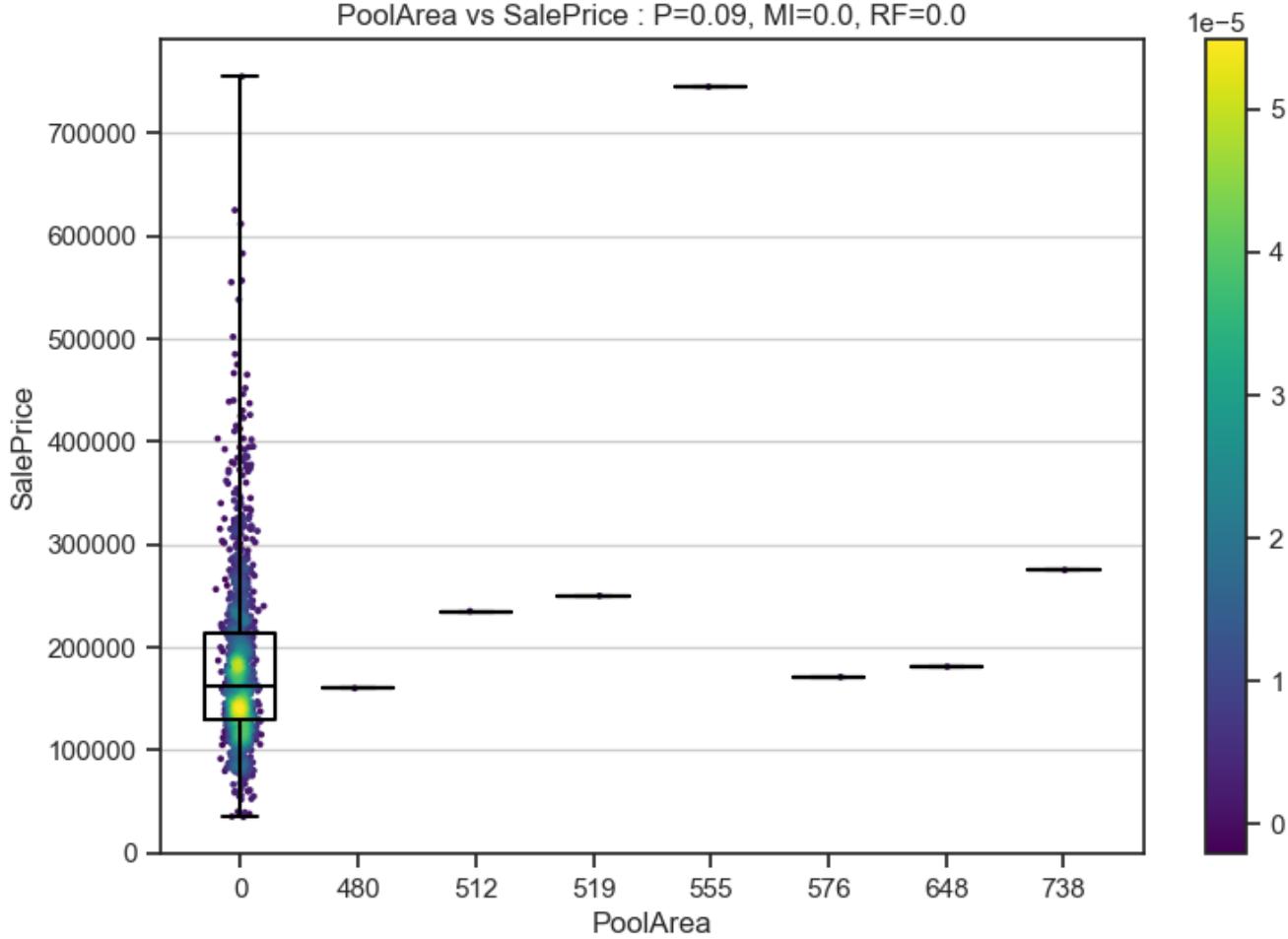
KitchenAbvGr vs SalePrice : P=-0.14, MI=0.01, RF=0.01



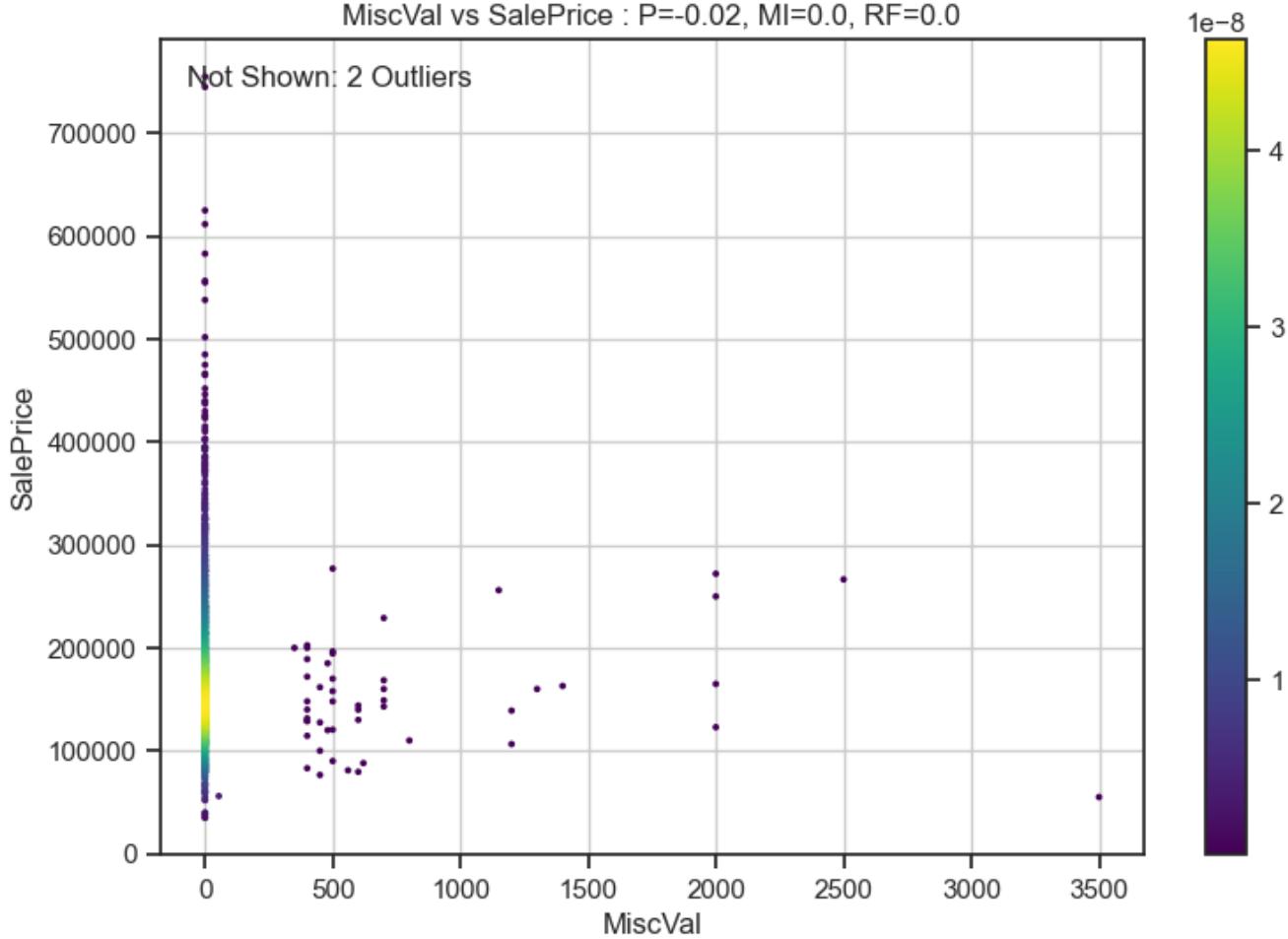
BsmtFinSF2 vs SalePrice : P=-0.01, MI=0.0, RF=0.01



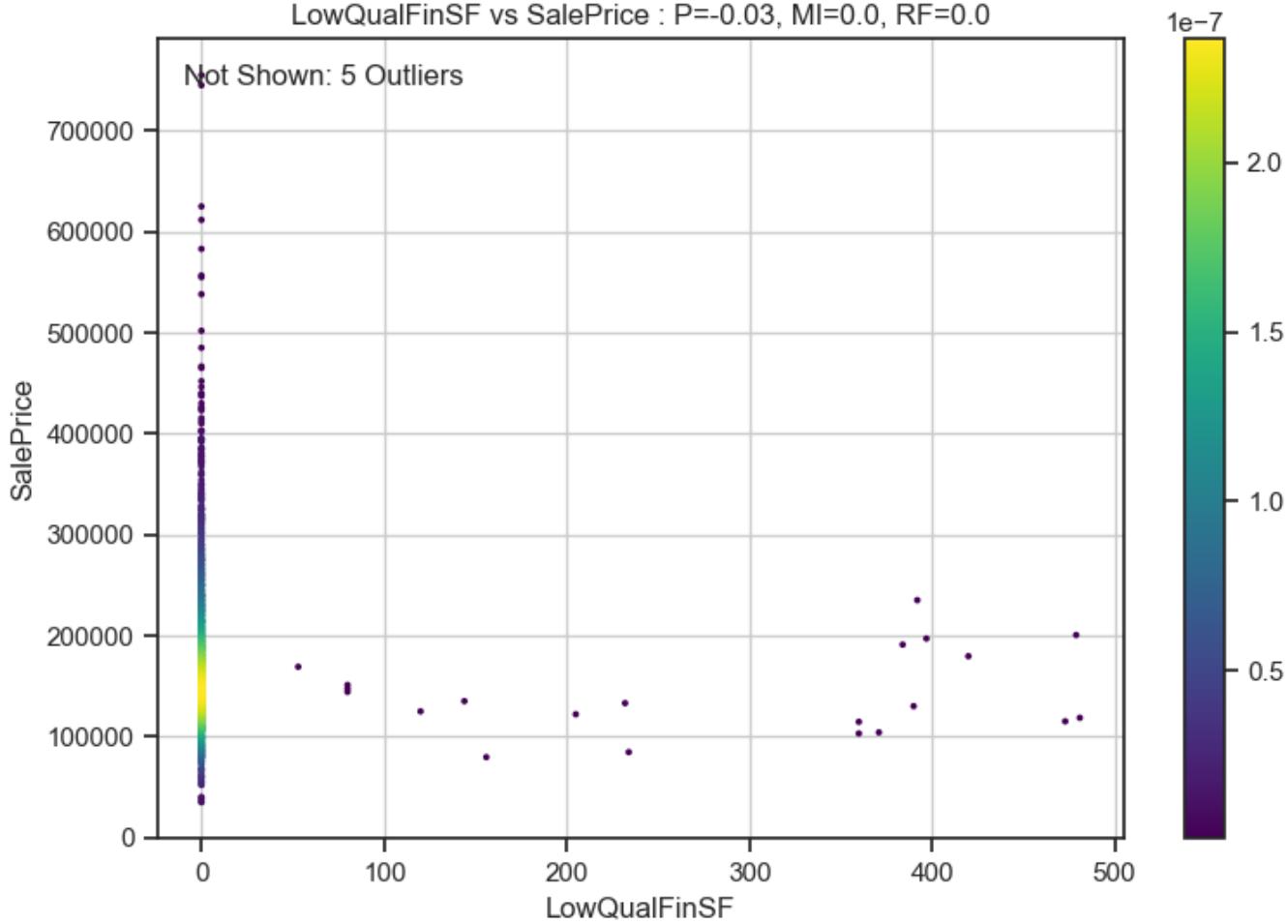
PoolArea vs SalePrice : P=0.09, MI=0.0, RF=0.0



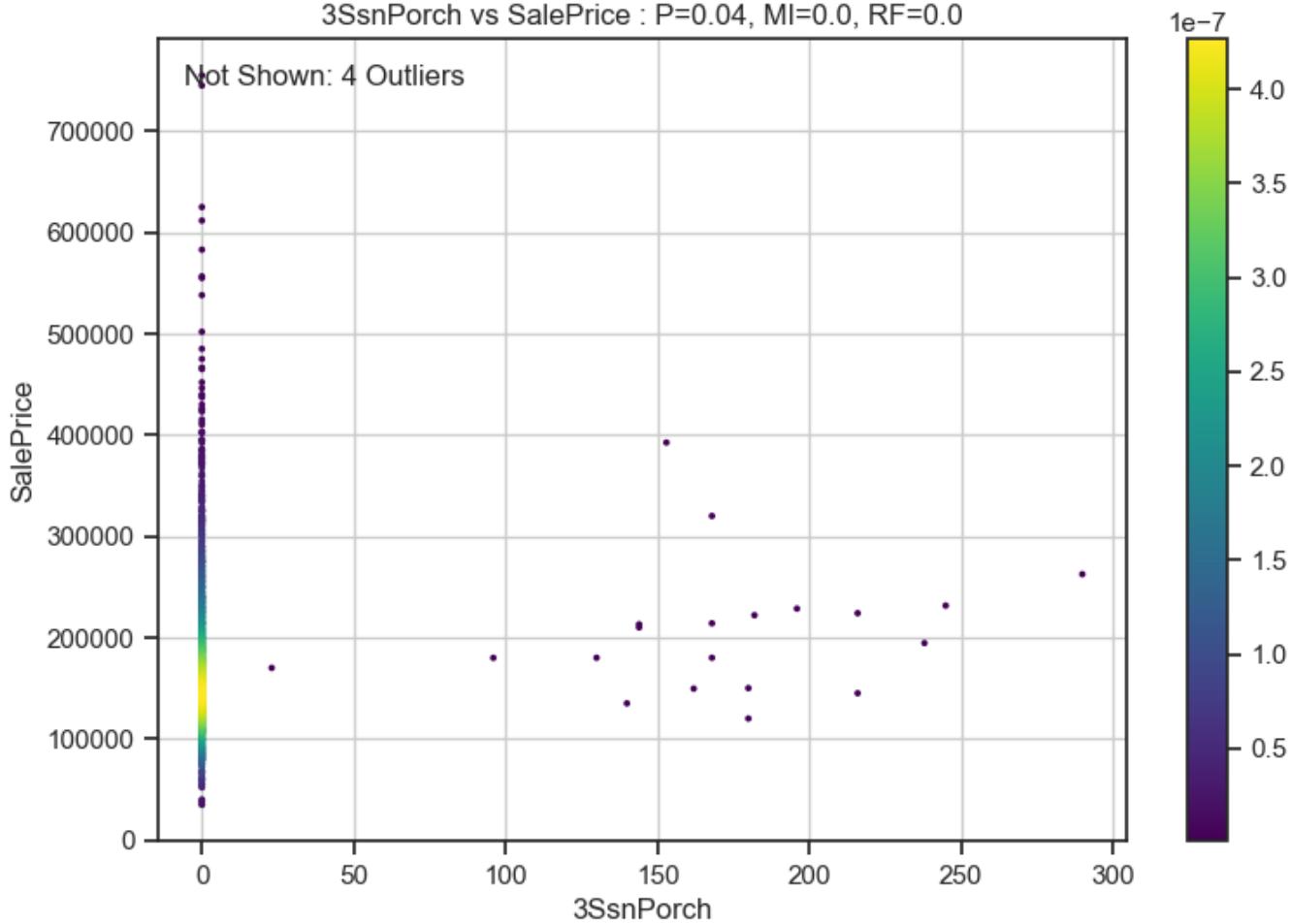
MiscVal vs SalePrice : P=-0.02, MI=0.0, RF=0.0

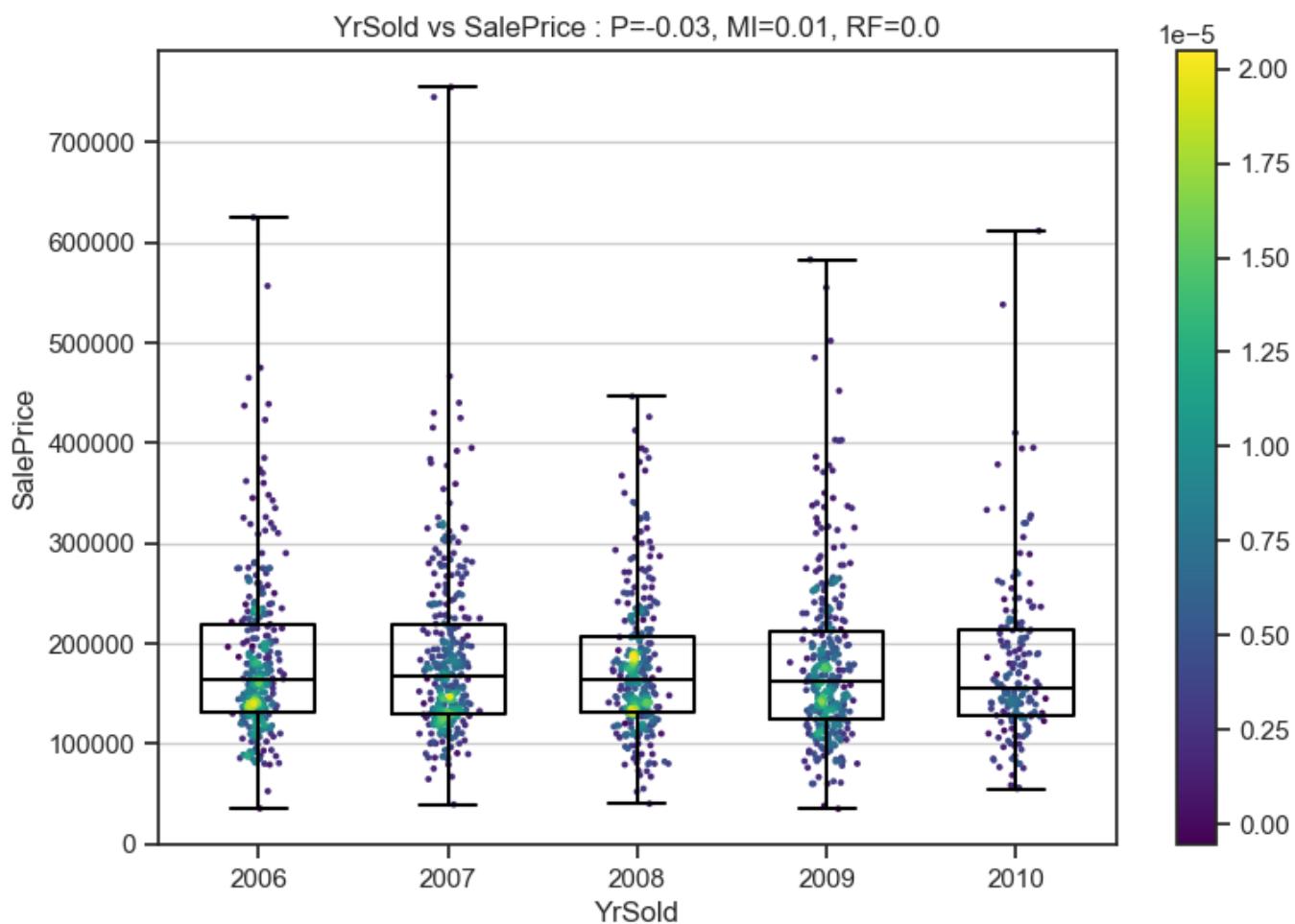
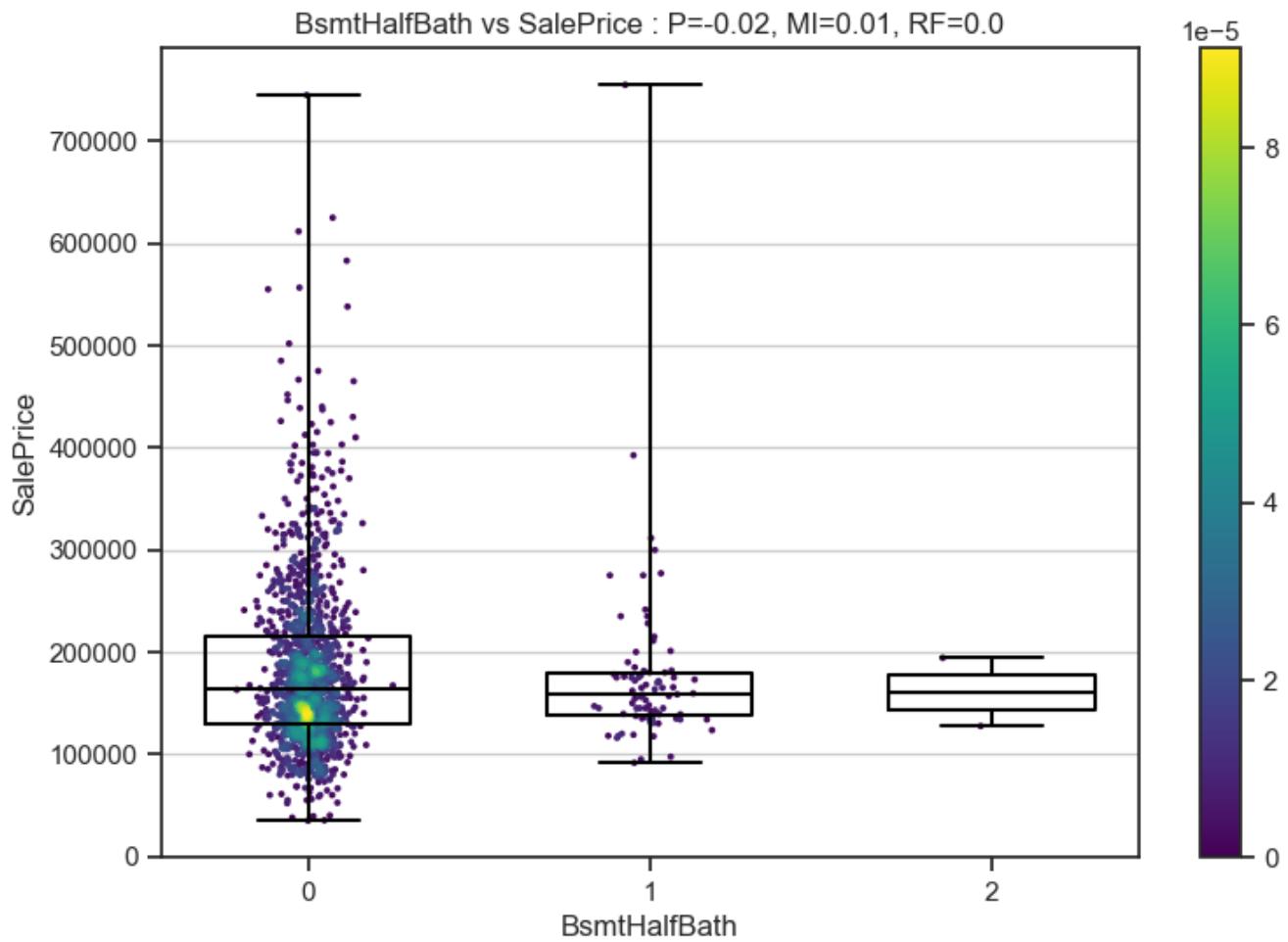


LowQualFinSF vs SalePrice : P=-0.03, MI=0.0, RF=0.0

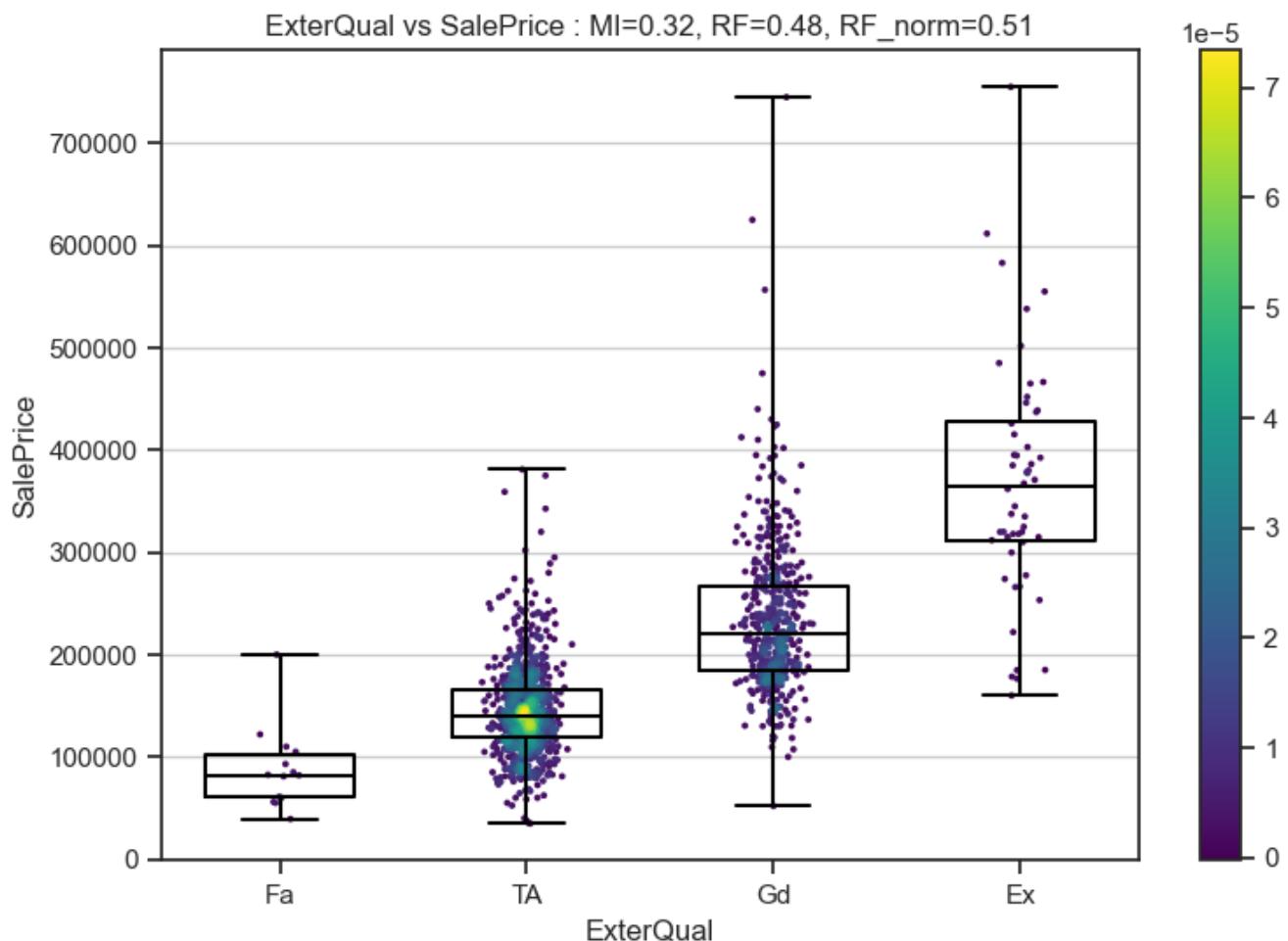
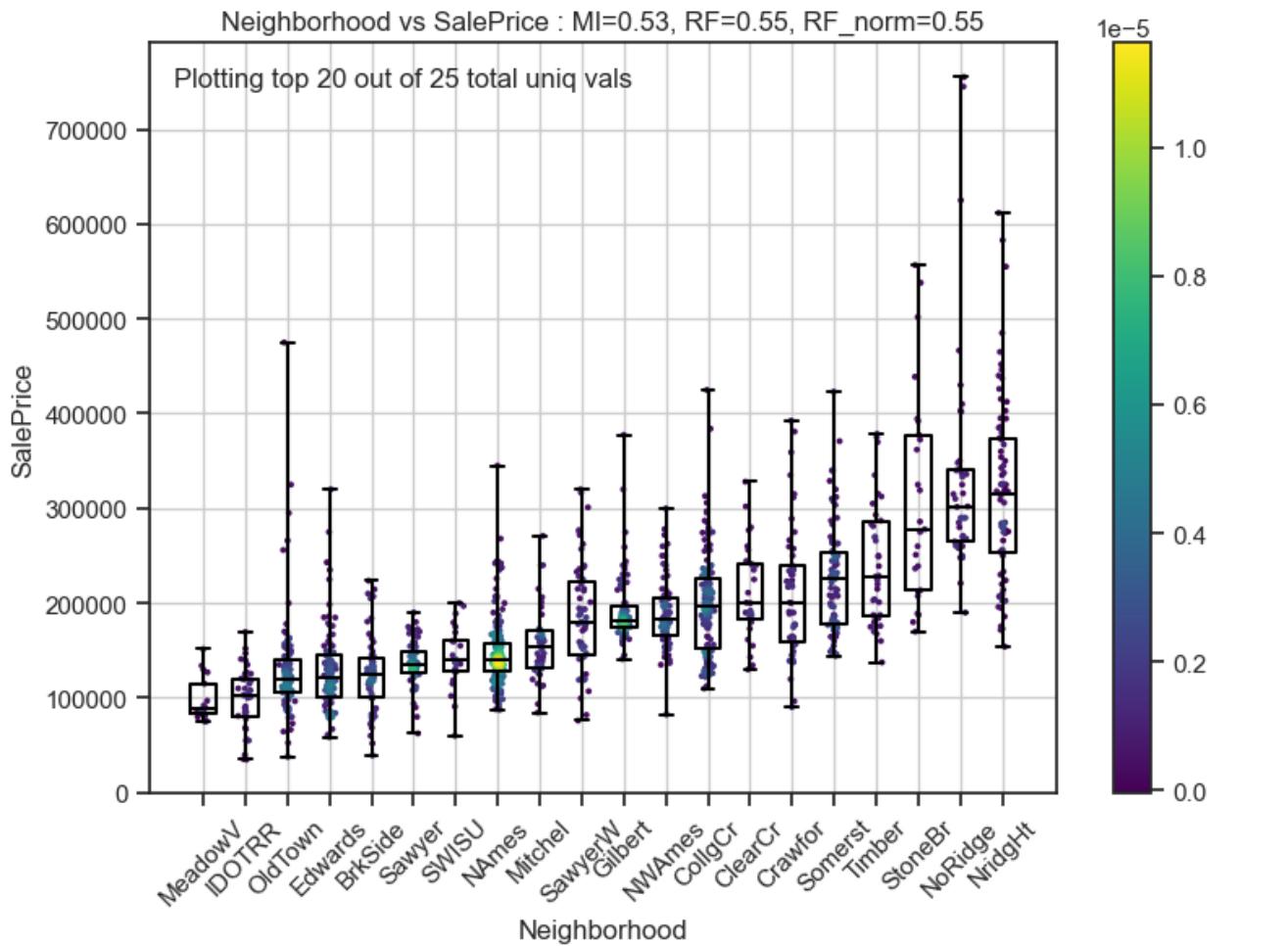


3SsnPorch vs SalePrice : P=0.04, MI=0.0, RF=0.0

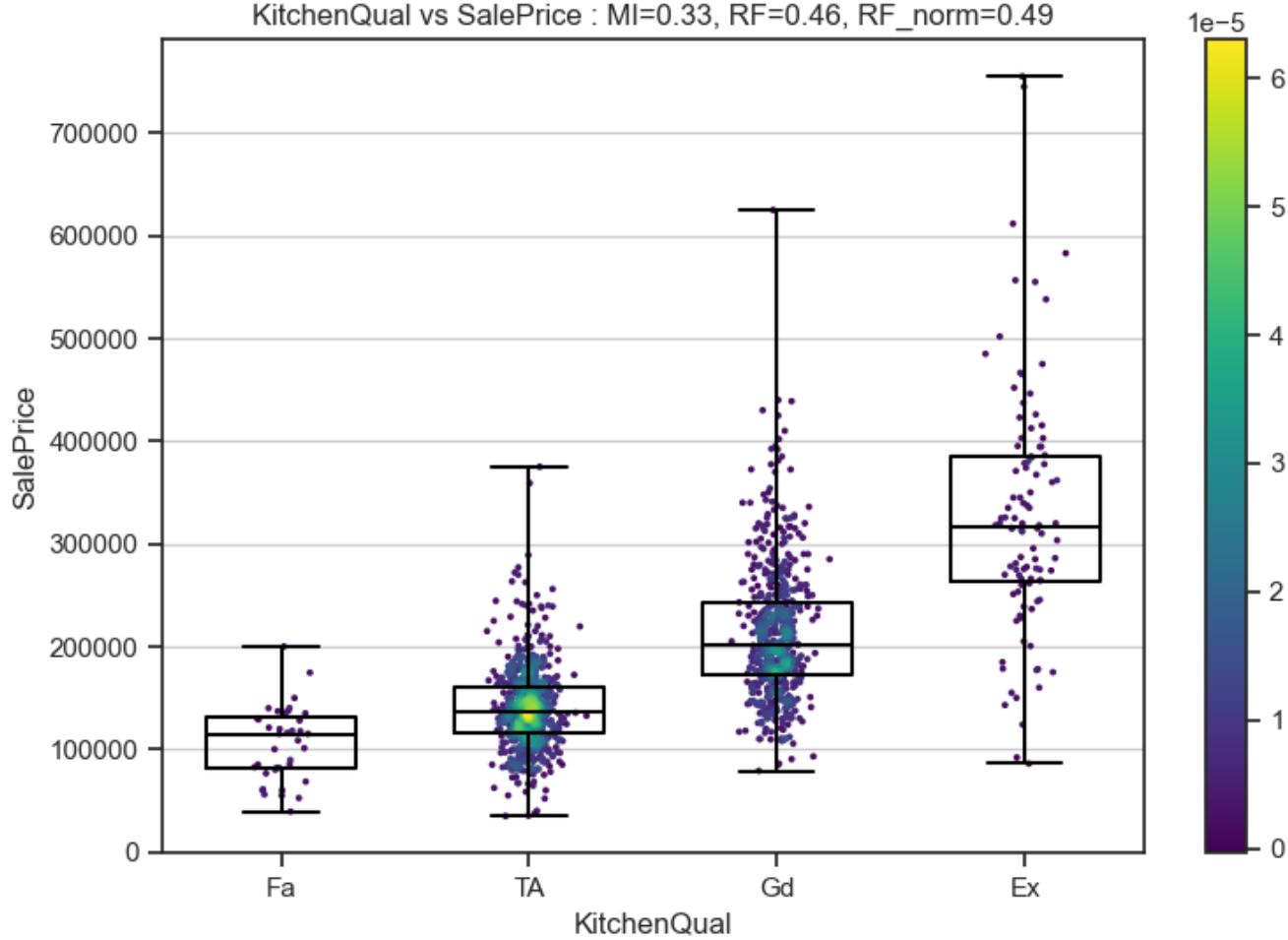




Plots of Non-Numeric Columns versus the Target Variable



KitchenQual vs SalePrice : MI=0.33, RF=0.46, RF_norm=0.49



BsmtQual vs SalePrice : MI=0.32, RF=0.45, RF_norm=0.48

