18.4.2024 Markus Lähde

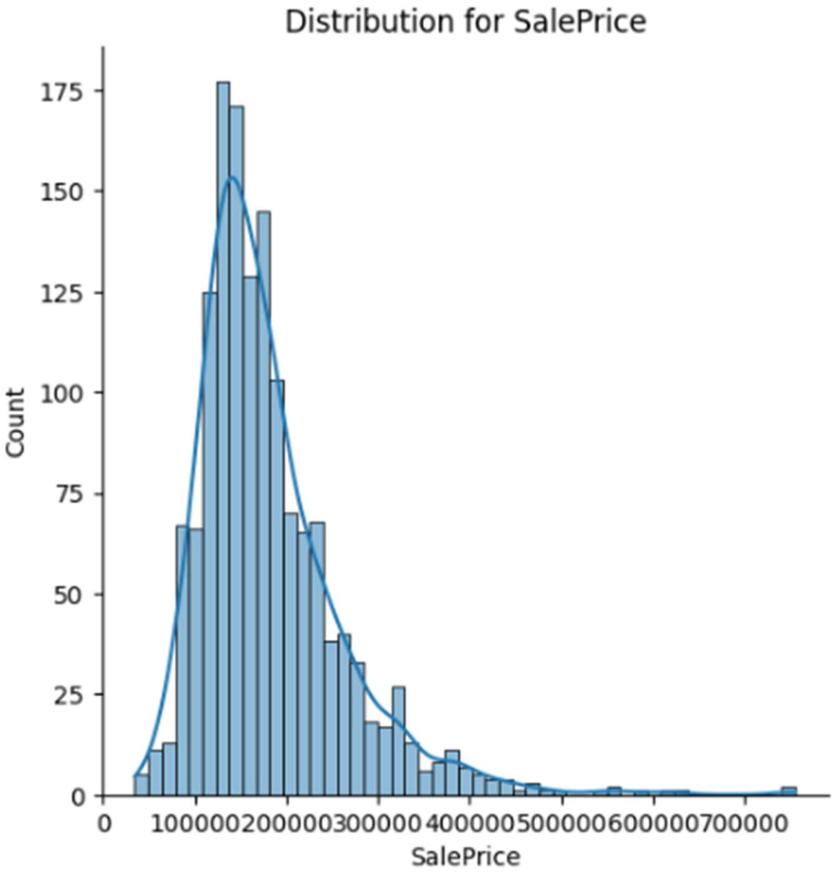
Exploratory Data Analysis on Housing Prices Dataset

https://www.kaggle.com/competitions/home-data-for-ml-course/overview

The Target

The target variable in the dataset is the SalePrice, which is a continuous feature.

SalePrice has a right skewed distribution with minimum value of **34900** and maximum value of **755000**.



Dividing data by type

Dataset came with a very good data dictionary. Based on dictionary and observations made, the dataset was divided into three buckets: Ordinal, Nominal, Discrete and Continuous.

Correlation

Correlation of variables with the SalePrice was examined.

Ordinal variables were encoded and correlation of each variable with the target was examined by calculating the Spearman's rank correlation coefficient and p-value.

Based on this analysis, these variables were considered for the model:

ExterQual - SalePrice: 0.684 (p-value: 5.605572202388652e-202)

KitchenQual - SalePrice: 0.673 (p-value: 4.400509397208977e-193)

HeatingQC - SalePrice: -0.471 (p-value: 1.2728804763417352e-81)

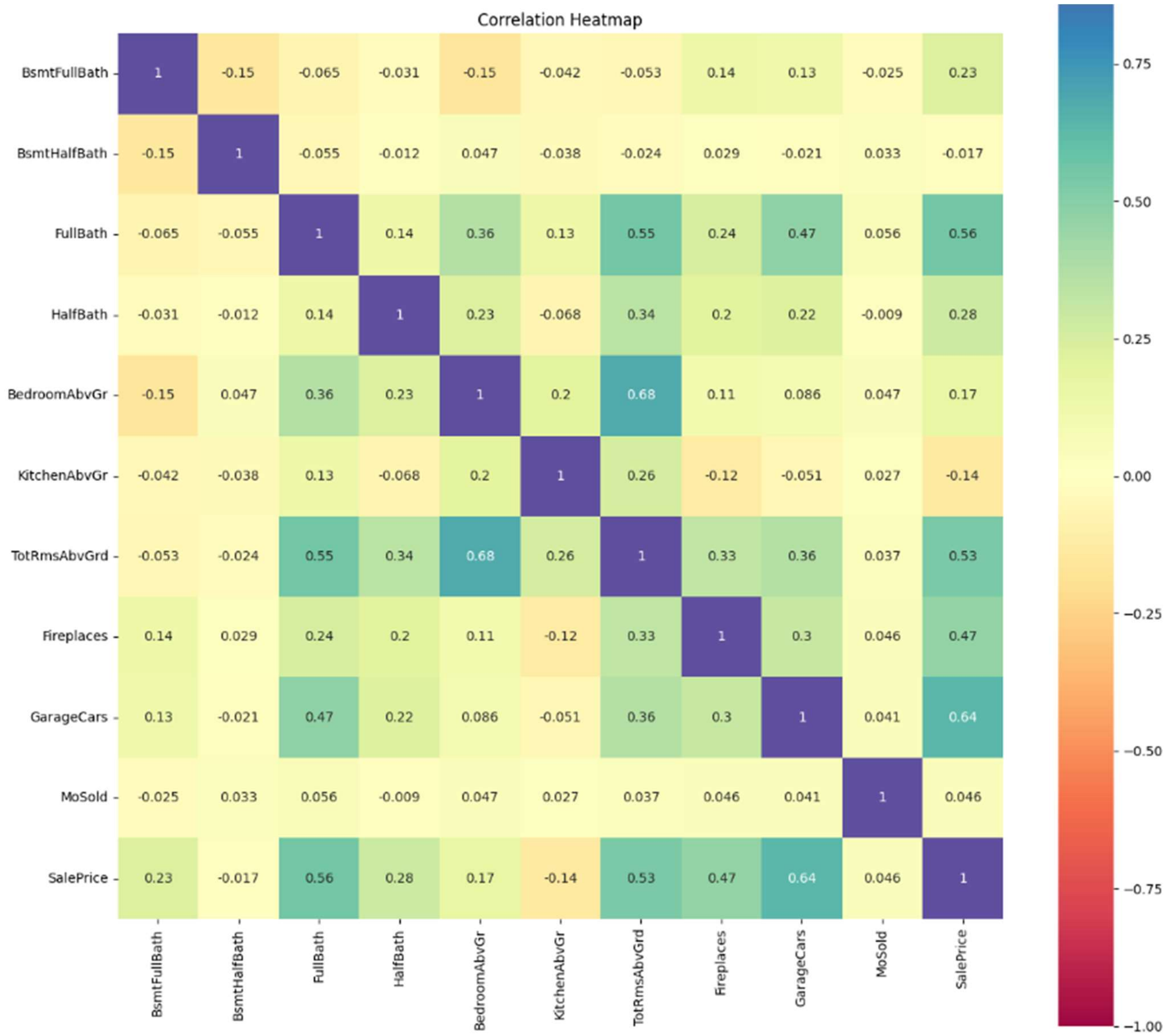
CentralAir - SalePrice: 0.313 (p-value: 1.3028329648335056e-34)

Nominal variables were encoded with the get_dummies -method and their correlation with the target was examined by calculating the Eta squared correlation test.

Based on this analysis, these variables were considered for the model:

Variable	Eta Squared Correlation
BsmtQual	2.805709e-01
Neighborhood	1.531942e-01
Foundation	1.379532e-01
BsmtFinType1	1.347997e-01
GarageFinish	1.336655e-01
SaleType	1.171327e-01
SaleCondition	1.133343e-01

Discrete variables' Pearson-correlation was checked with the target and a heatmap was generated to check multicollinearity between variables.



	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces	GarageCars	MoSold	SalePrice
BsmtFullBath	1	-0.15	-0.065	-0.031	-0.15	-0.042	-0.053	0.14	0.13	-0.025	0.23
BsmtHalfBath	-0.15	1	-0.055	-0.012	0.047	-0.038	-0.024	0.029	-0.021	0.033	-0.017
FullBath	-0.065	-0.055	1	0.14	0.36	0.13	0.55	0.24	0.47	0.056	0.56
HalfBath	-0.031	-0.012	0.14	1	0.23	-0.068	0.34	0.2	0.22	-0.009	0.28
BedroomAbvGr	-0.15	0.047	0.36	0.23	1	0.2	0.68	0.11	0.086	0.047	0.17
KitchenAbvGr	-0.042	-0.038	0.13	-0.068	0.2	1	0.26	-0.12	-0.051	0.027	-0.14
TotRmsAbvGrd	-0.053	-0.024	0.55	0.34	0.68	0.26	1	0.33	0.36	0.037	0.53
Fireplaces	0.14	0.029	0.24	0.2	0.11	-0.12	0.33	1	0.3	0.046	0.47
GarageCars	0.13	-0.021	0.47	0.22	0.086	-0.051	0.36	0.3	1	0.041	0.64
MoSold	-0.025	0.033	0.056	-0.009	0.047	0.027	0.037	0.046	0.041	1	0.046
SalePrice	0.23	-0.017	0.56	0.28	0.17	-0.14	0.53	0.47	0.64	0.046	1

Based on this analysis, these variables were considered for the model:

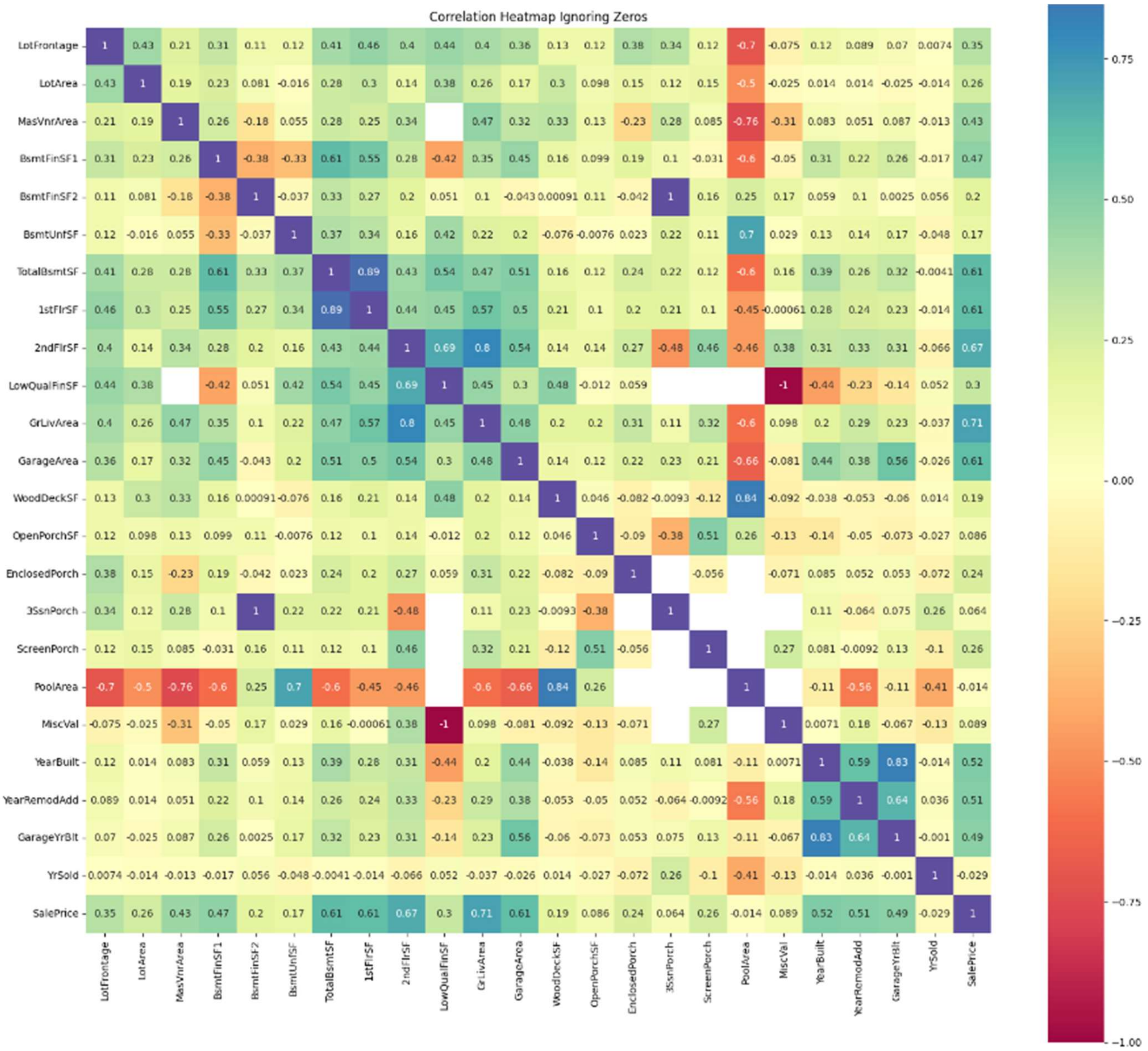
GarageCars: 0.640 correlation with the target

FullBath: 0.561 correlation with the target

TotRmsAbvGrd: 0.534 correlation with the target

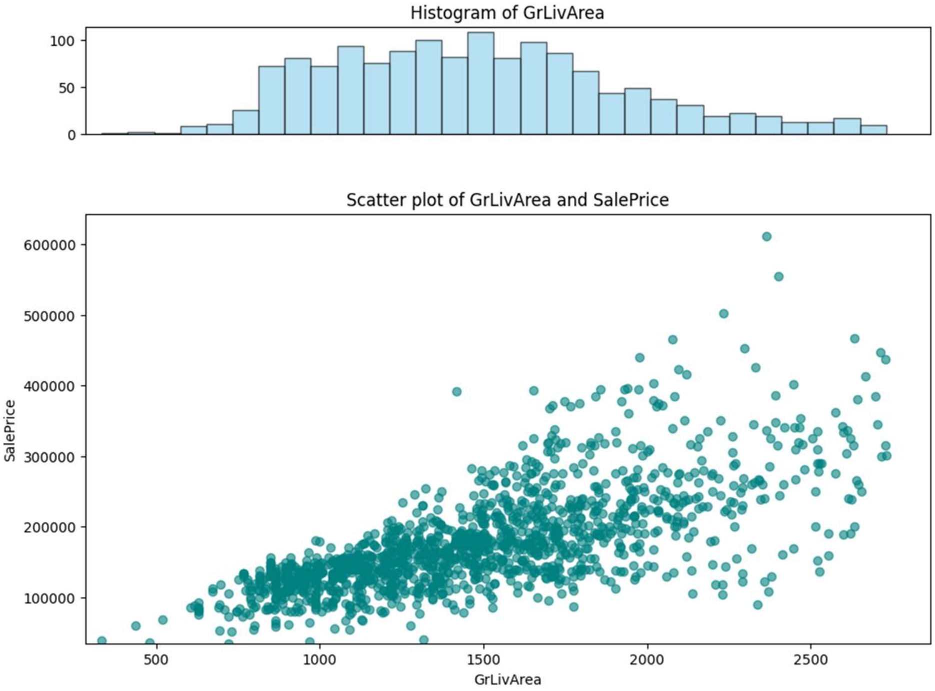
Fireplaces: 0.467 correlation with the target

Continuous variables’ Pearson-correlation with the target was checked ignoring 0, nan and outlier values, correlation with other variables was also checked with a heatmap.

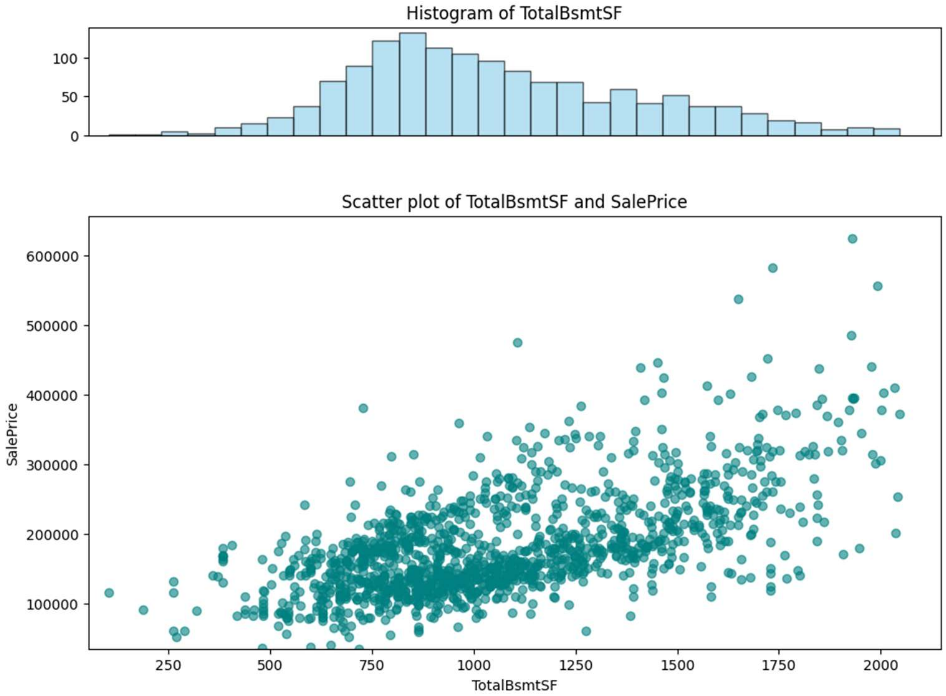


Based on this analysis, these variables were considered for the model:

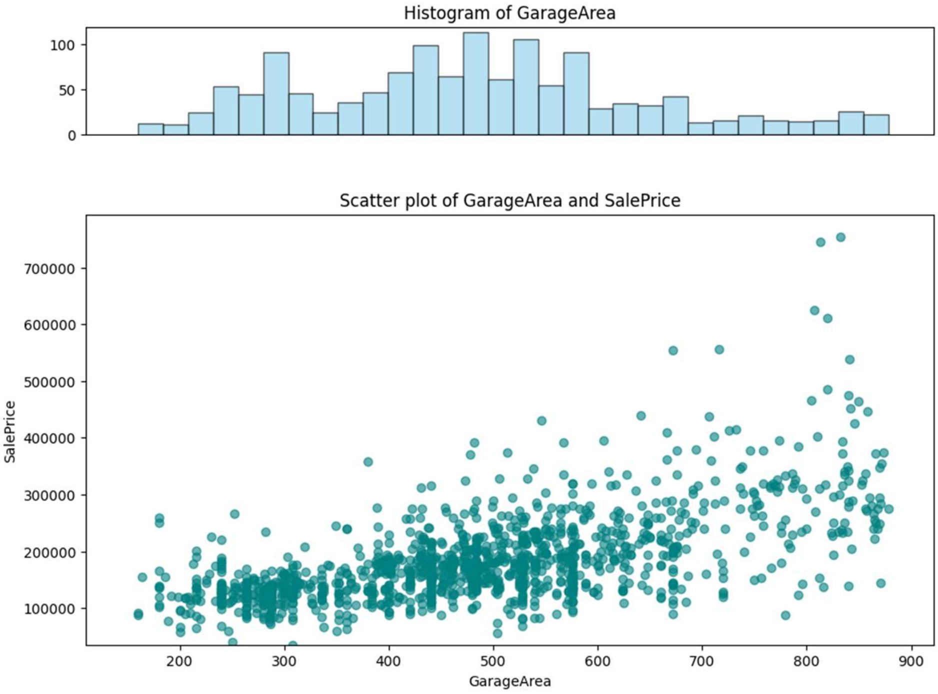
GrLivArea: correlation 0.709



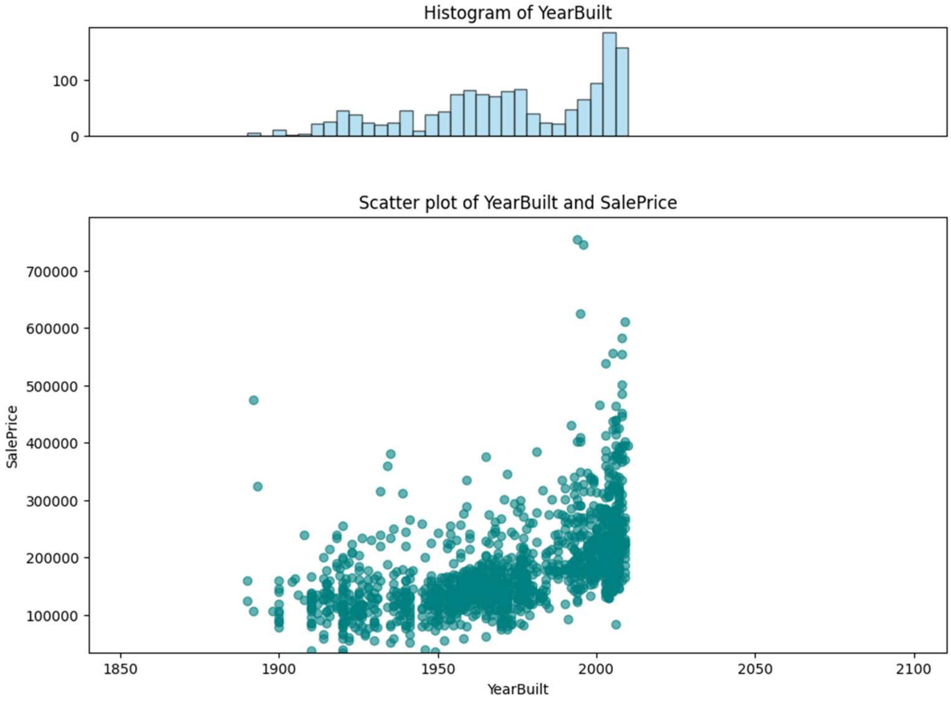
TotalBsmtSF: correlation 0.610



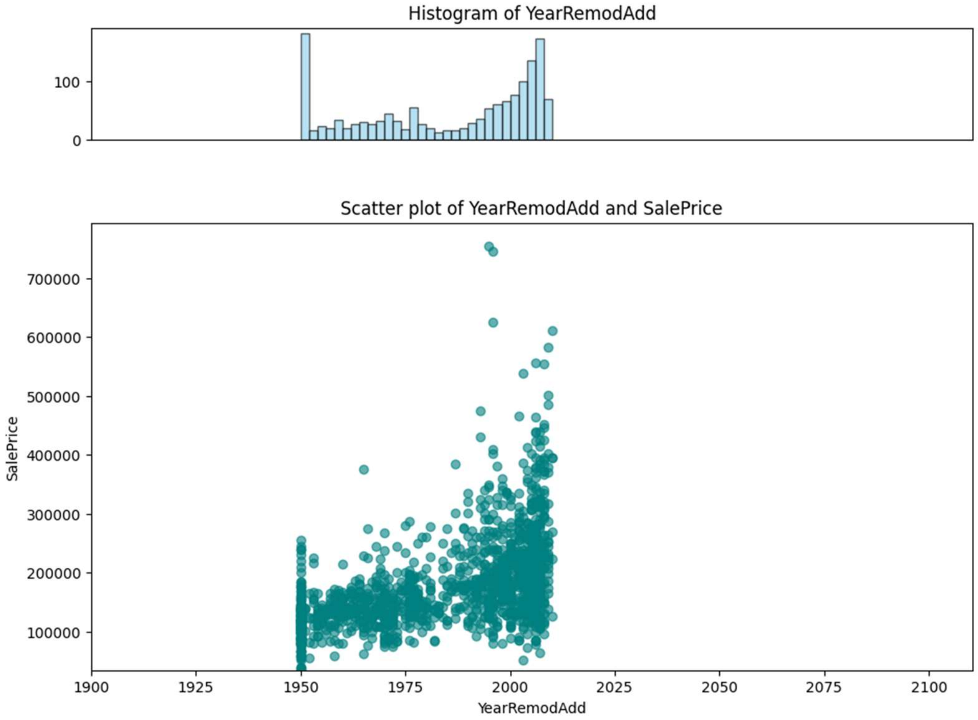
GarageArea: correlation 0.608



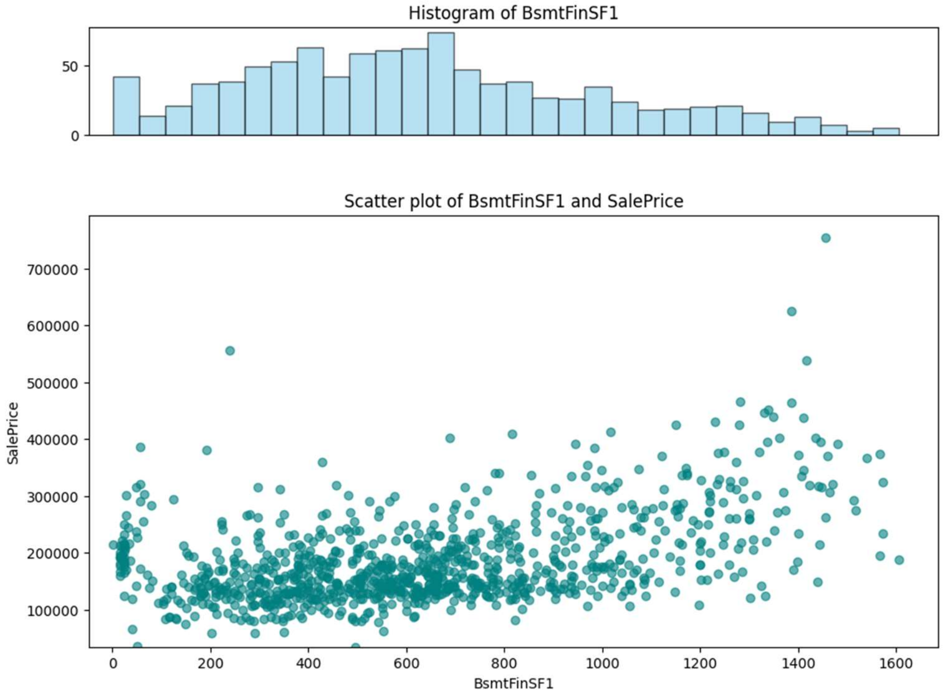
YearBuilt: correlation 0.523



YearRemodAdd: correlation 0.507



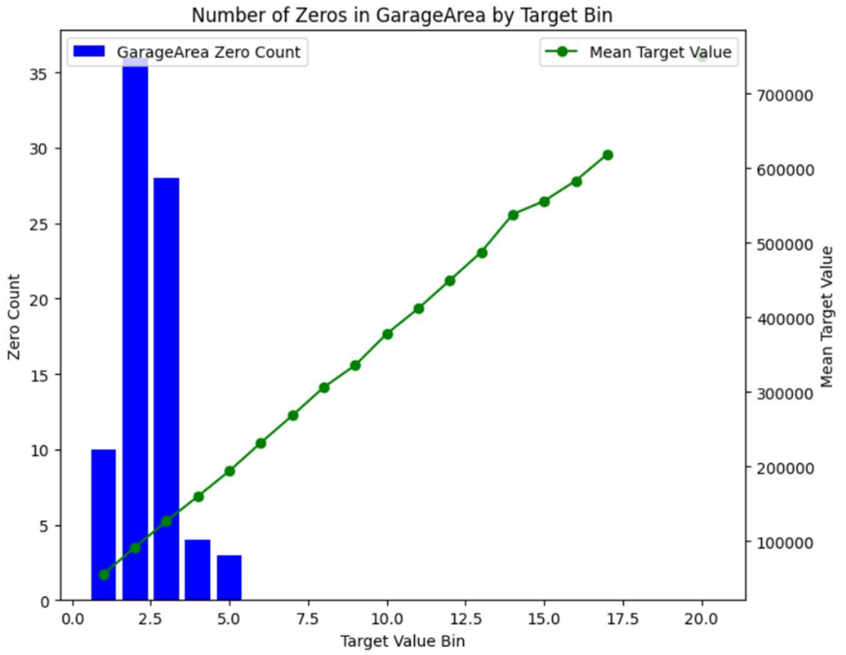
BsmtFinSF1: correlation 0.472

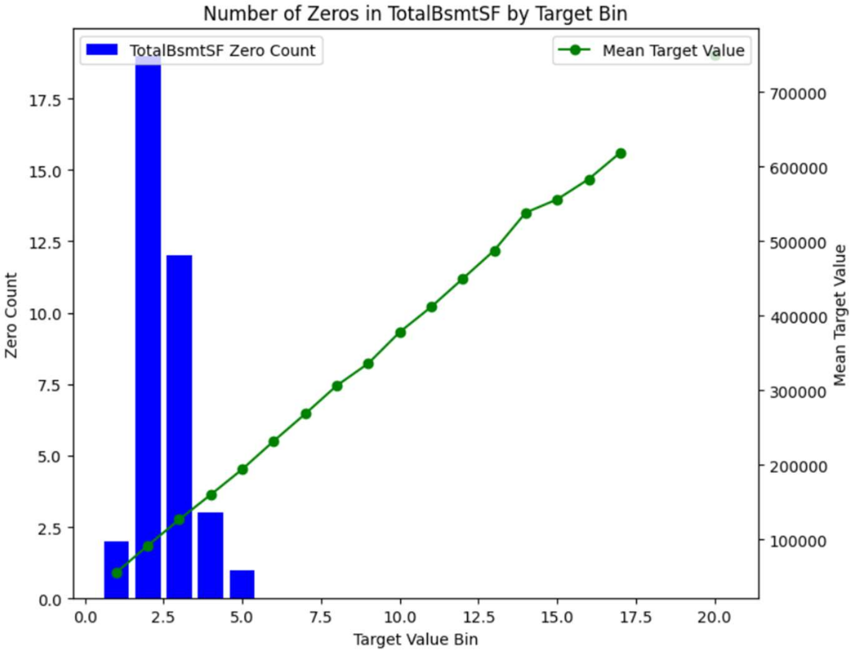


Variables 2ndFlrSF, 1stFlrSF and GarageYrBlt were discarded based on their high multicollinearity with variables that had higher correlation with the target.

Zero value assessment

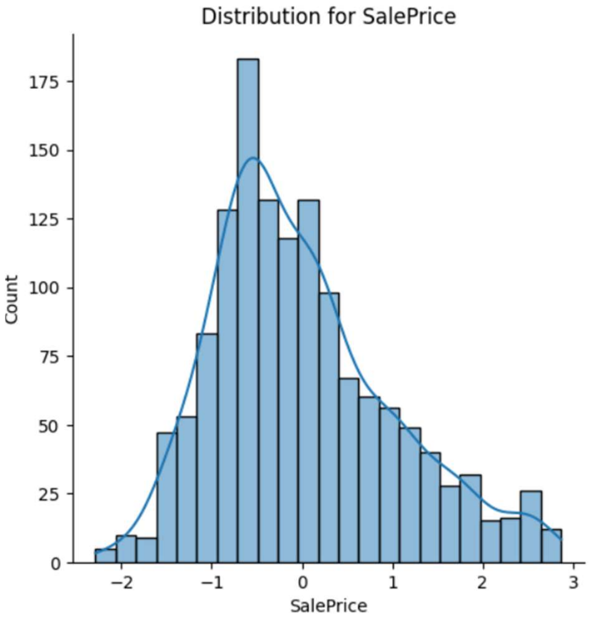
Placement of 0 values of continuous variables in relation to the target variable was examined by visualizing them with binned and sorted target values. In general, 0 values seem to be corresponding with lower house prices (below 200 000). Fit of a model around this area should be well planned.

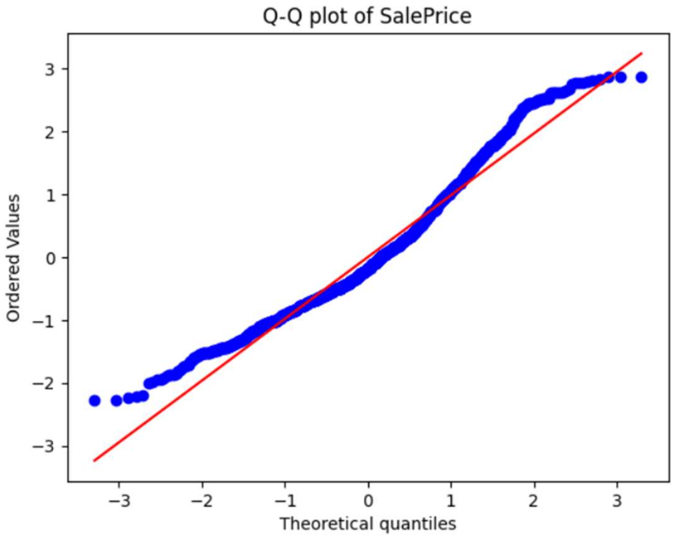




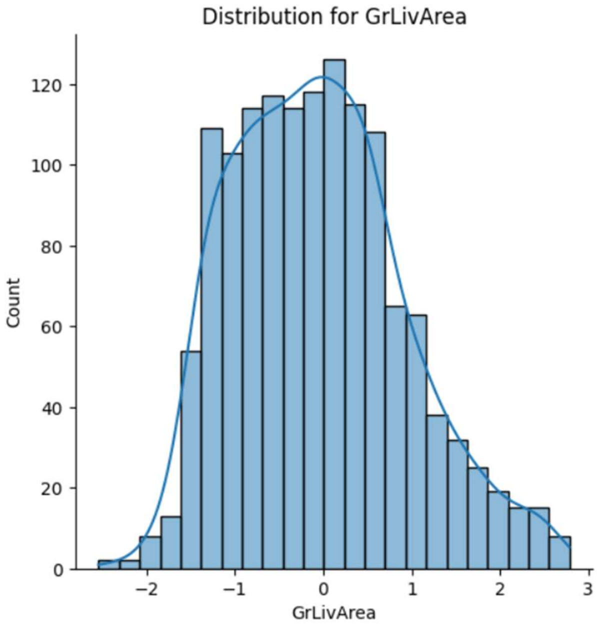
Normality and distribution assessment

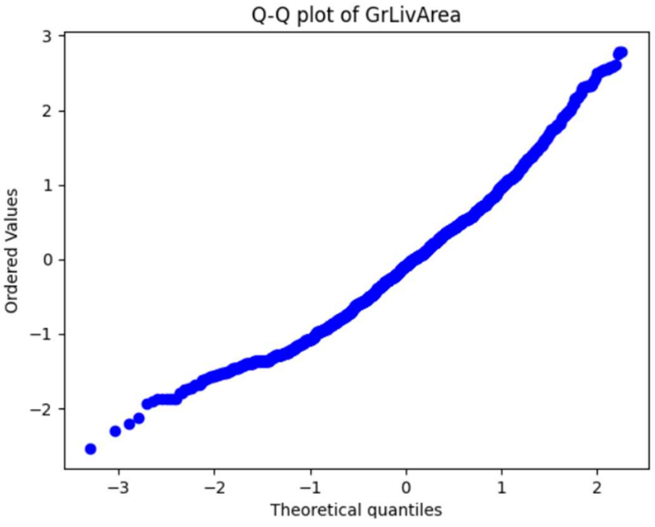
To use parametric methods an assumption of linearity between continuous variables and the target is made. Although not strictly necessary, it is a good practice to transform continuous variables with skewed distribution to achieve a better linear relationship with the target. To assess need for transformations, normality of variables was measured by counting their Shapiro-Wilk test p-value and Q-Q plotting distribution. This assessment was made to variables without their outliers and 0 values. Data was also scaled with Standard Scaler before doing the distribution analysis.



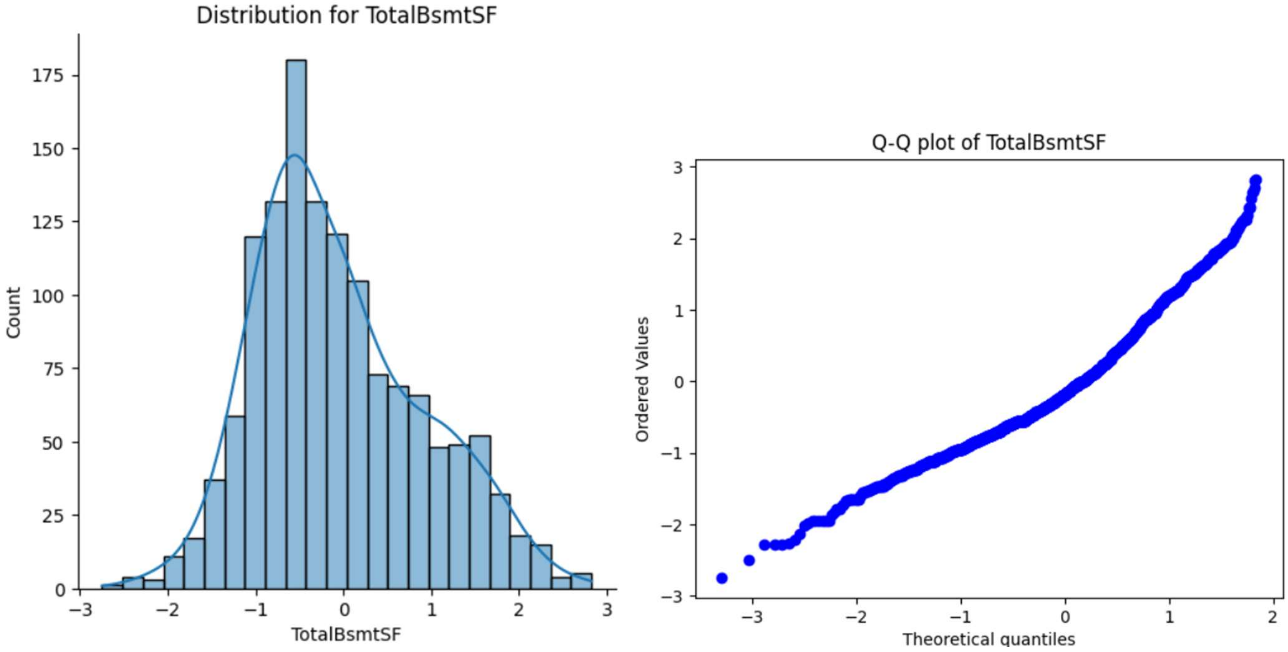


SalePrice - Shapiro Test p-value: 1.5383312564147468e-18





GrLivArea - Shapiro Test p-value: 6.986455684809822e-13



TotalBsmtSF - Shapiro Test p-value: 1.949321136636168e-14

Conclusions

With this information given, building predictive model should be easier. Datasets features don’t generally follow normal distribution, but I believe sufficient linear correlation between well chosen variables can be achieved. This correlation can be helped with scaling and possibly transforming some of the features.