Horse price prediction - Kaggel data set

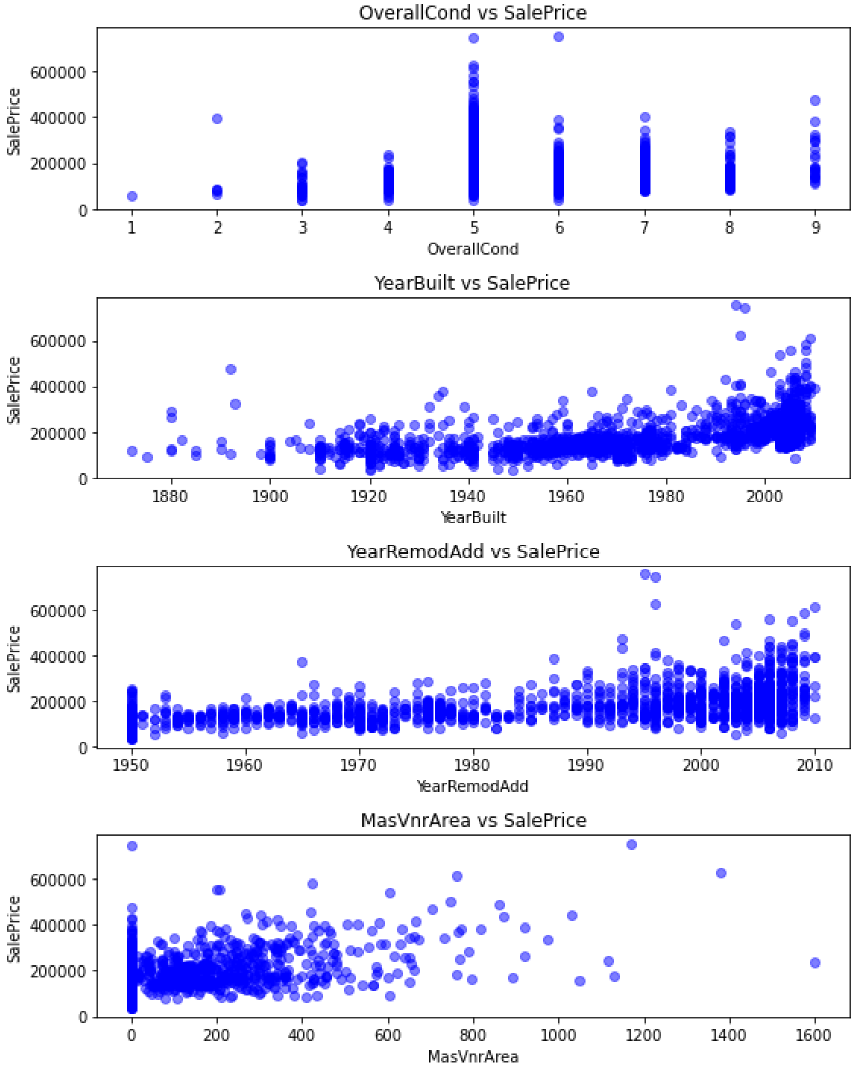
The purpose of this project is to build regression models to predict the price of the house based on the provided data.

Data Overview:

Data has 80 columns and 1460 rows. The target variable is SalePrice. 37 columns are numeric and 43 are categorical.

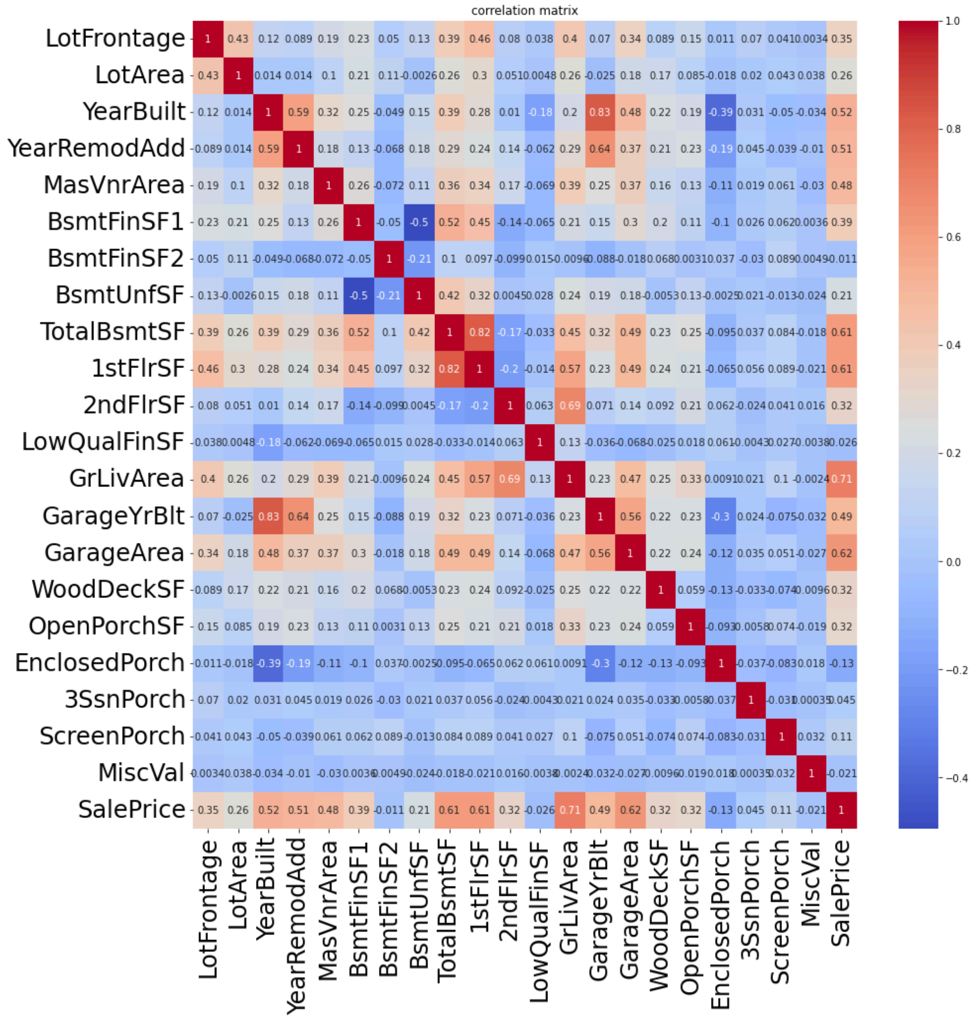
Scatter plot:

Below show the scatter plot of SalePrice plotted as a function of other numeric variables:



Collinearity:

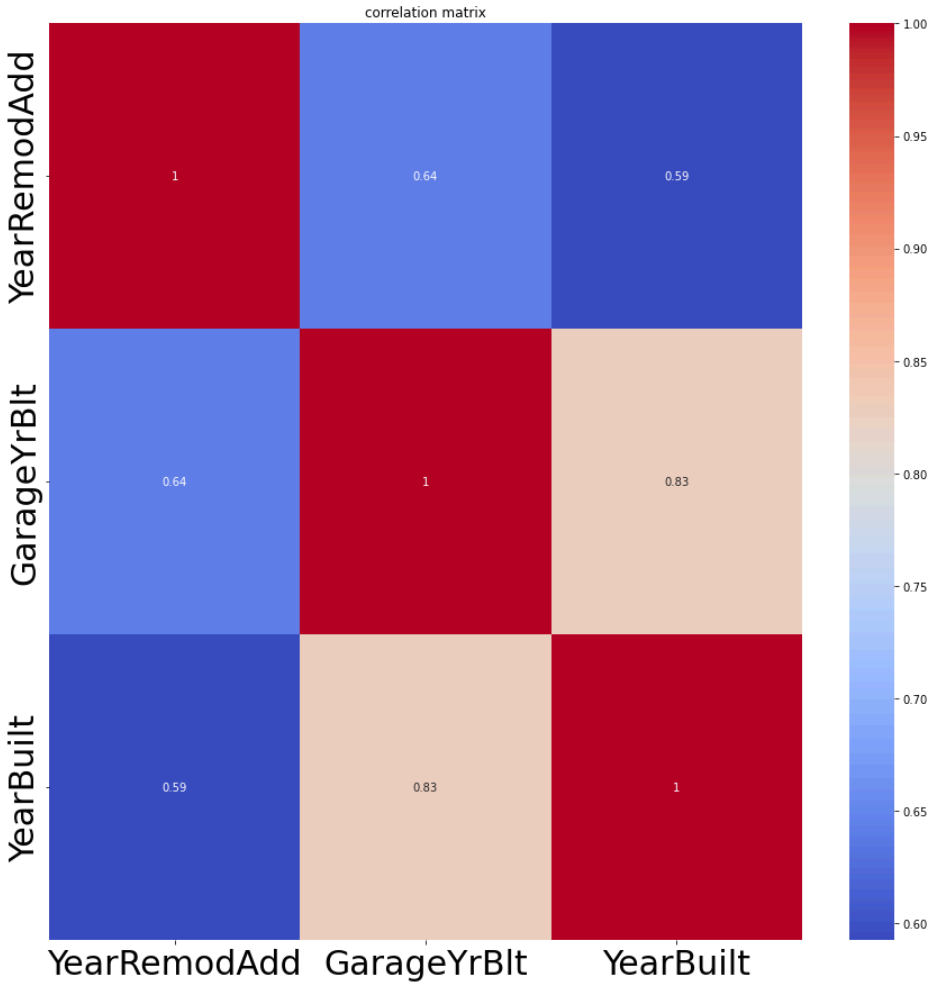
Before implementing any model we should examine the collinearity of data. The following is a correlation matrix, showing how correlated different features are:



	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	GarageYrBlt	GarageArea	WoodDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	MiscVal	SalePrice
LotFrontage	1	0.43	0.12	0.089	0.19	0.23	0.05	0.13	0.39	0.46	0.08	0.038	0.4	0.07	0.34	0.089	0.15	0.011	0.07	0.041	0.0034	0.35
LotArea	0.43	1	0.014	0.014	0.1	0.21	0.11	-0.0026	0.26	0.3	0.051	0.0048	0.26	-0.025	0.18	0.17	0.085	-0.018	0.02	0.043	0.038	0.26
YearBuilt	0.12	0.014	1	0.59	0.32	0.25	-0.049	0.15	0.39	0.28	0.01	-0.18	0.2	0.83	0.48	0.22	0.19	-0.39	0.031	-0.05	-0.034	0.52
YearRemodAdd	0.089	0.014	0.59	1	0.18	0.13	-0.068	0.18	0.29	0.24	0.14	-0.062	0.29	0.64	0.37	0.21	0.23	-0.19	0.045	-0.039	-0.01	0.51
MasVnrArea	0.19	0.1	0.32	0.18	1	0.26	-0.072	0.11	0.36	0.34	0.17	-0.069	0.39	0.25	0.37	0.16	0.13	-0.11	0.019	0.061	-0.03	0.48
BsmtFinSF1	0.23	0.21	0.25	0.13	0.26	1	-0.05	-0.5	0.52	0.45	-0.14	-0.065	0.21	0.15	0.3	0.2	0.11	-0.1	0.026	0.062	0.0036	0.39
BsmtFinSF2	0.05	0.11	-0.049	-0.068	-0.072	-0.05	1	-0.21	0.1	0.097	-0.099	0.015	-0.0096	-0.088	-0.018	0.068	0.0031	0.037	-0.03	0.089	0.0049	-0.011
BsmtUnfSF	0.13	-0.0026	0.15	0.18	0.11	-0.5	-0.21	1	0.42	0.32	0.0045	0.028	0.24	0.19	0.18	-0.0053	0.13	-0.0025	0.021	-0.013	-0.024	0.21
TotalBsmtSF	0.39	0.26	0.39	0.29	0.36	0.52	0.1	0.42	1	0.82	-0.17	-0.033	0.45	0.32	0.49	0.23	0.25	-0.095	0.037	0.084	-0.018	0.61
1stFlrSF	0.46	0.3	0.28	0.24	0.34	0.45	0.097	0.32	0.82	1	-0.2	-0.014	0.57	0.23	0.49	0.24	0.21	-0.065	0.056	0.089	-0.021	0.61
2ndFlrSF	0.08	0.051	0.01	0.14	0.17	-0.14	-0.099	0.0045	-0.17	-0.2	1	0.063	0.69	0.071	0.14	0.092	0.21	0.062	-0.024	0.041	0.016	0.32
LowQualFinSF	0.038	0.0048	-0.18	-0.062	-0.069	-0.065	0.015	0.028	-0.033	-0.014	0.063	1	0.13	-0.036	-0.068	-0.025	0.018	0.061	-0.0043	0.027	-0.0038	-0.026
GrLivArea	0.4	0.26	0.2	0.29	0.39	0.21	-0.0096	0.24	0.45	0.57	0.69	0.13	1	0.23	0.47	0.25	0.33	0.0091	0.021	0.1	-0.0024	0.71
GarageYrBlt	0.07	-0.025	0.83	0.64	0.25	0.15	-0.088	0.19	0.32	0.23	0.071	-0.036	0.23	1	0.56	0.22	0.23	-0.3	0.024	-0.075	-0.032	0.49
GarageArea	0.34	0.18	0.48	0.37	0.37	0.3	-0.018	0.18	0.49	0.49	0.14	-0.068	0.47	0.56	1	0.22	0.24	-0.12	0.035	0.051	-0.027	0.62
WoodDeckSF	0.089	0.17	0.22	0.21	0.16	0.2	0.068	-0.0053	0.23	0.24	0.092	-0.025	0.25	0.22	0.22	1	0.059	-0.13	-0.033	-0.074	-0.0096	0.32
OpenPorchSF	0.15	0.085	0.19	0.23	0.13	0.11	0.0031	0.13	0.25	0.21	0.21	0.018	0.33	0.23	0.24	0.059	1	-0.093	-0.0058	0.074	-0.019	0.32
EnclosedPorch	0.011	-0.018	-0.39	-0.19	-0.11	-0.1	0.037	-0.0025	-0.095	-0.065	0.062	0.061	0.0091	-0.3	-0.12	-0.13	-0.093	1	-0.037	-0.083	0.018	-0.13
3SsnPorch	0.07	0.02	0.031	0.045	0.019	0.026	-0.03	0.021	0.037	0.056	-0.024	-0.0043	0.021	0.024	0.035	-0.033	-0.0058	-0.037	1	-0.031	0.00035	0.045
ScreenPorch	0.041	0.043	-0.05	-0.039	0.061	0.062	0.089	-0.013	0.084	0.089	0.041	0.027	0.1	-0.075	0.051	-0.074	0.074	-0.083	-0.031	1	0.032	0.11
MiscVal	0.0034	0.038	-0.034	-0.01	-0.03	0.0036	0.0049	-0.024	-0.018	-0.021	0.016	-0.0038	-0.0024	-0.032	-0.027	-0.0096	-0.019	0.018	0.00035	0.032	1	-0.021
SalePrice	0.35	0.26	0.52	0.51	0.48	0.39	-0.011	0.21	0.61	0.61	0.32	-0.026	0.71	0.49	0.62	0.32	0.32	-0.13	0.045	0.11	-0.021	1

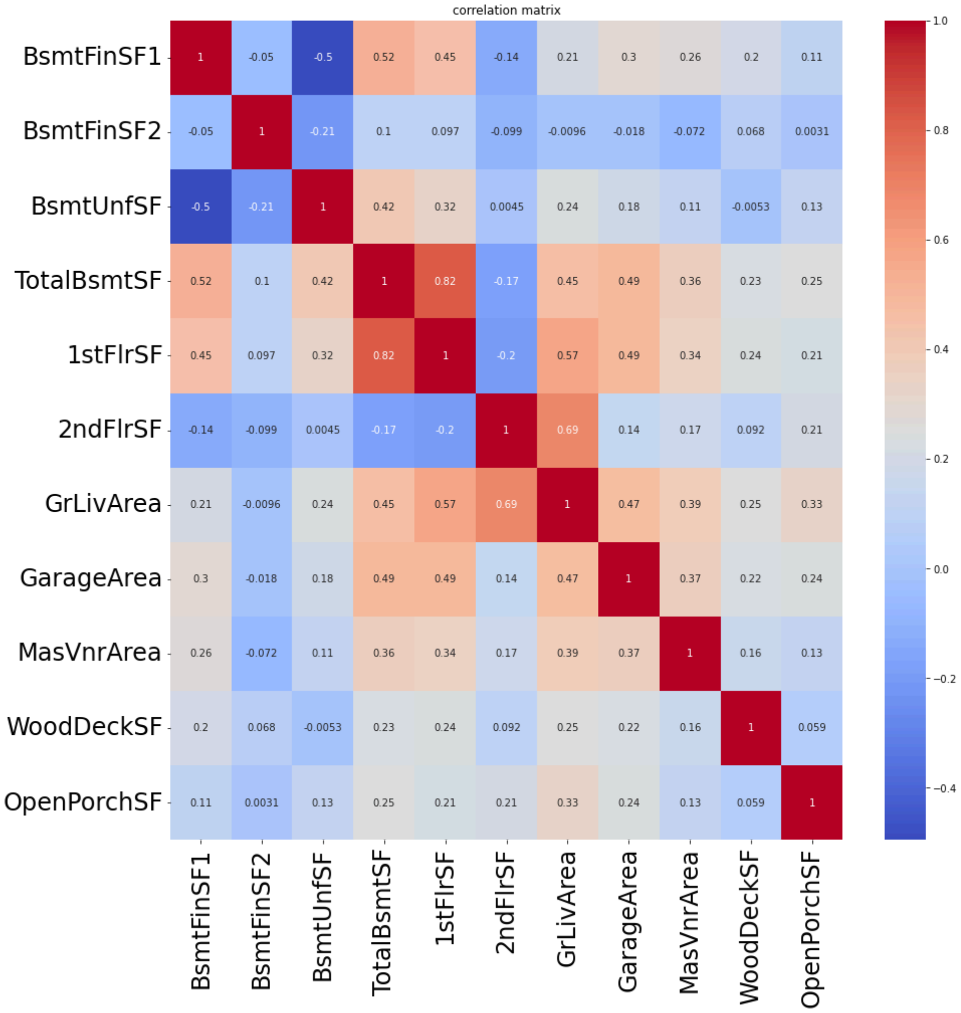
There are three groups of feature that have high correlation values:

1. Features that include year: ['YearRemodAdd', 'GarageYrBlt', "YearBuilt"]



The above image shows their correlation matrix. The decision was to drop two of these columns and keep only YearBuilt

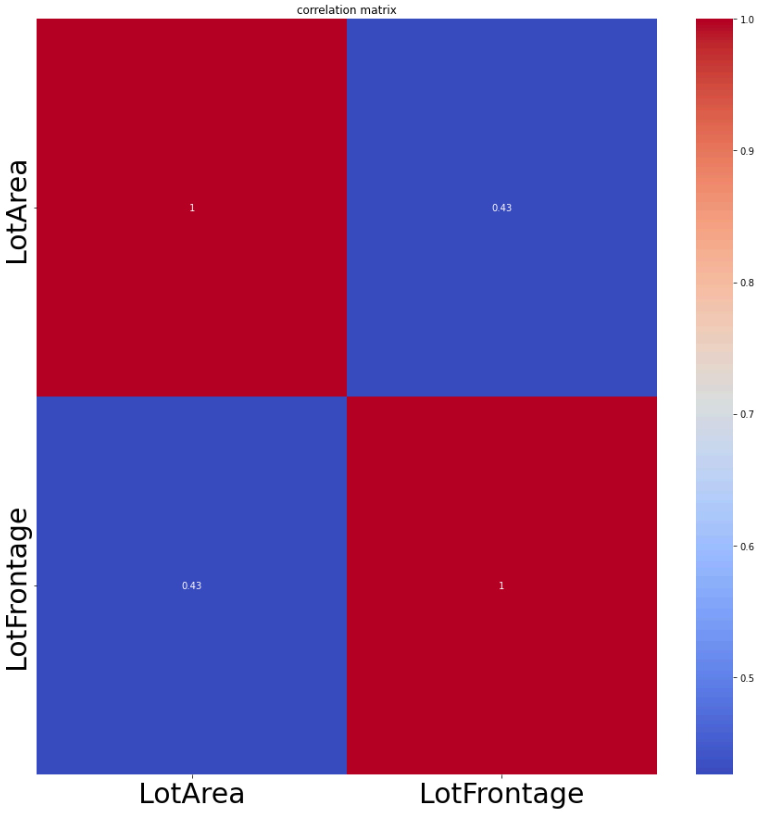
2. Correlation between features that have units of area. The image below show the correlation matrix for these columns:



	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	1stFlrSF	2ndFlrSF	GrLivArea	GarageArea	MasVnrArea	WoodDeckSF	OpenPorchSF
BsmtFinSF1	1	-0.05	-0.5	0.52	0.45	-0.14	0.21	0.3	0.26	0.2	0.11
BsmtFinSF2	-0.05	1	-0.21	0.1	0.097	-0.099	-0.0096	-0.018	-0.072	0.068	0.0031
BsmtUnfSF	-0.5	-0.21	1	0.42	0.32	0.0045	0.24	0.18	0.11	-0.0053	0.13
TotalBsmtSF	0.52	0.1	0.42	1	0.82	-0.17	0.45	0.49	0.36	0.23	0.25
1stFlrSF	0.45	0.097	0.32	0.82	1	-0.2	0.57	0.49	0.34	0.24	0.21
2ndFlrSF	-0.14	-0.099	0.0045	-0.17	-0.2	1	0.69	0.14	0.17	0.092	0.21
GrLivArea	0.21	-0.0096	0.24	0.45	0.57	0.69	1	0.47	0.39	0.25	0.33
GarageArea	0.3	-0.018	0.18	0.49	0.49	0.14	0.47	1	0.37	0.22	0.24
MasVnrArea	0.26	-0.072	0.11	0.36	0.34	0.17	0.39	0.37	1	0.16	0.13
WoodDeckSF	0.2	0.068	-0.0053	0.23	0.24	0.092	0.25	0.22	0.16	1	0.059
OpenPorchSF	0.11	0.0031	0.13	0.25	0.21	0.21	0.33	0.24	0.13	0.059	1

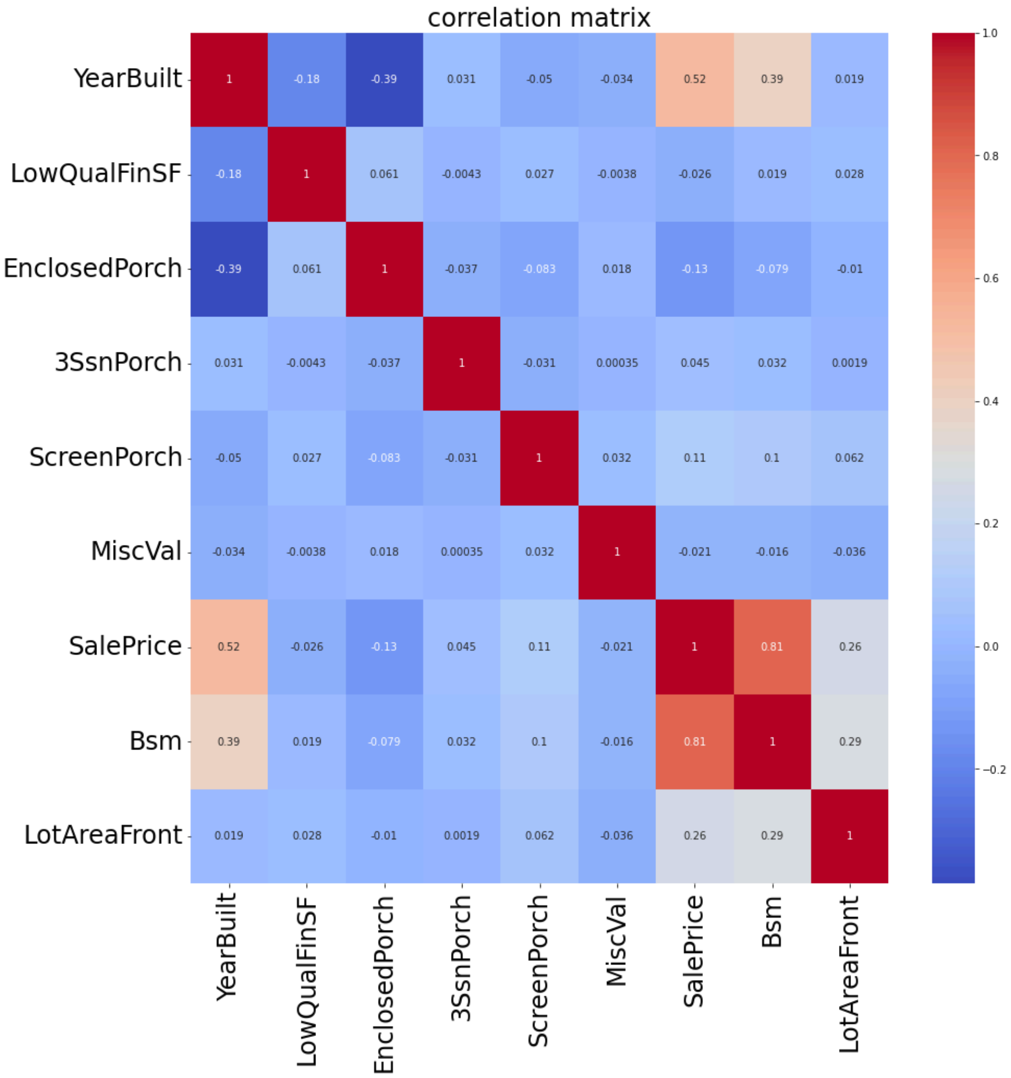
Since these features all have the same unite, a new feature was created by adding those features and followed by dropping the original columns.

3. Correlation between features of type Lot:

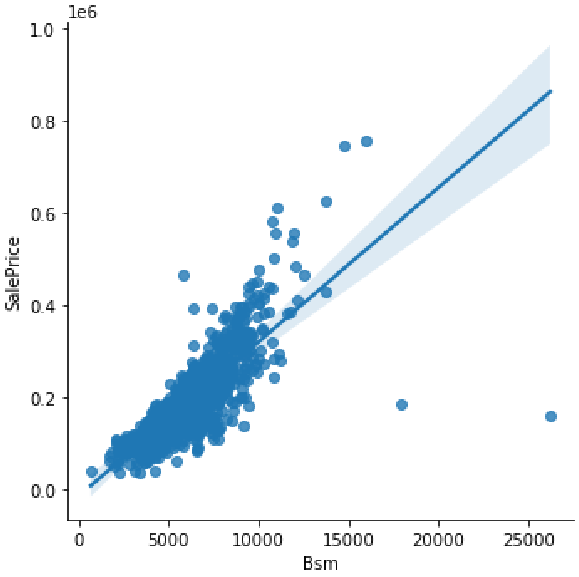


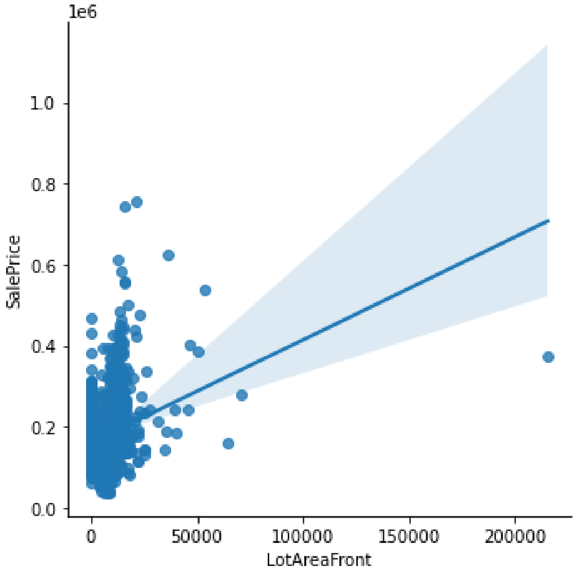
Again the decision was to make a new feature by adding them up followed by dropping the original columns.

Finally the correlation function after feature engineering is plotted in the figure below:



The following two graphs show how the new feature are correlated with the target variable:





Dealing with outliers:

In order the improve the model outliers were removed according to the following criteria:
multiplier = 2.5
Q1 = df[true_numeric_columns].quantile(0.25)
Q3 = df[true_numeric_columns].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - multiplier * IQR
upper_bound = Q3 + multiplier * IQR

Data splitting to test and train:
The data was divided to test and train in order to be able to measure the performance of each model.

Models:
Different regression model were implemented and table below compares all of them:

Model	RMSE
Mean	62834.4
Median	62113.1
Linear Regression	10283858416.3
Decision Tree Regression	31200.8
Decision Tree Regression **Hyperparameter tuning**	28947.3
Random Forest Regression (RFR)	22059.9
RFR hyperparameter tuning	22245.3
Gradient Boosting Regression (GBR)	20177.0
GBR hyperparameter tuning	19536.5
XGBoost hyperparameter tuning	19971.9

According to our results GBR and XGBoost are among the best model.