Predicting House Prices using Linear Regression

Problem Statement

- Create and select the best regression model for predicting housing sales price.

Approach (Data Cleaning)

- Data Imputation
- Separating variable types to nominal, ordinal, discrete and continuous
- EDA filtered features with scatter plots and histograms
- Dropping outliers
- Label encoding for ordinal and one hot encoding for nominal variables
- Increase features by using polynomial feature for filtered continuous variables

Approach (Training and Evaluation)

- Perform train-test split
- Standard Scaling
- Fit data with Linear regression, LassoCV and RidgeCV
- Model Scoring and Selection
- Fitting test data to selected model
- Visualize top coefficients

Data Imputation

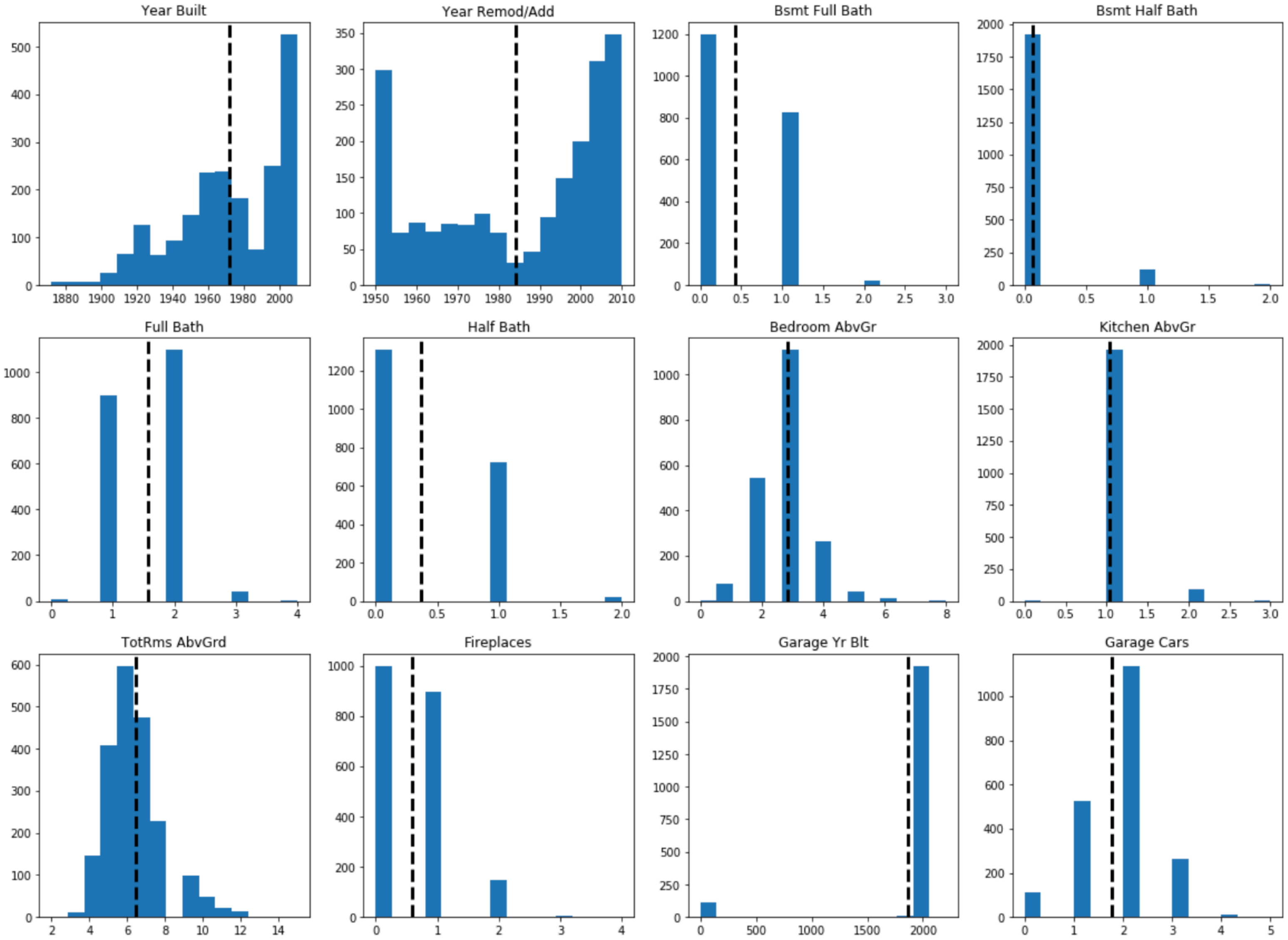
- Features with missing count and their missing percentage

- Dropped rows with missing percentage of less than 1

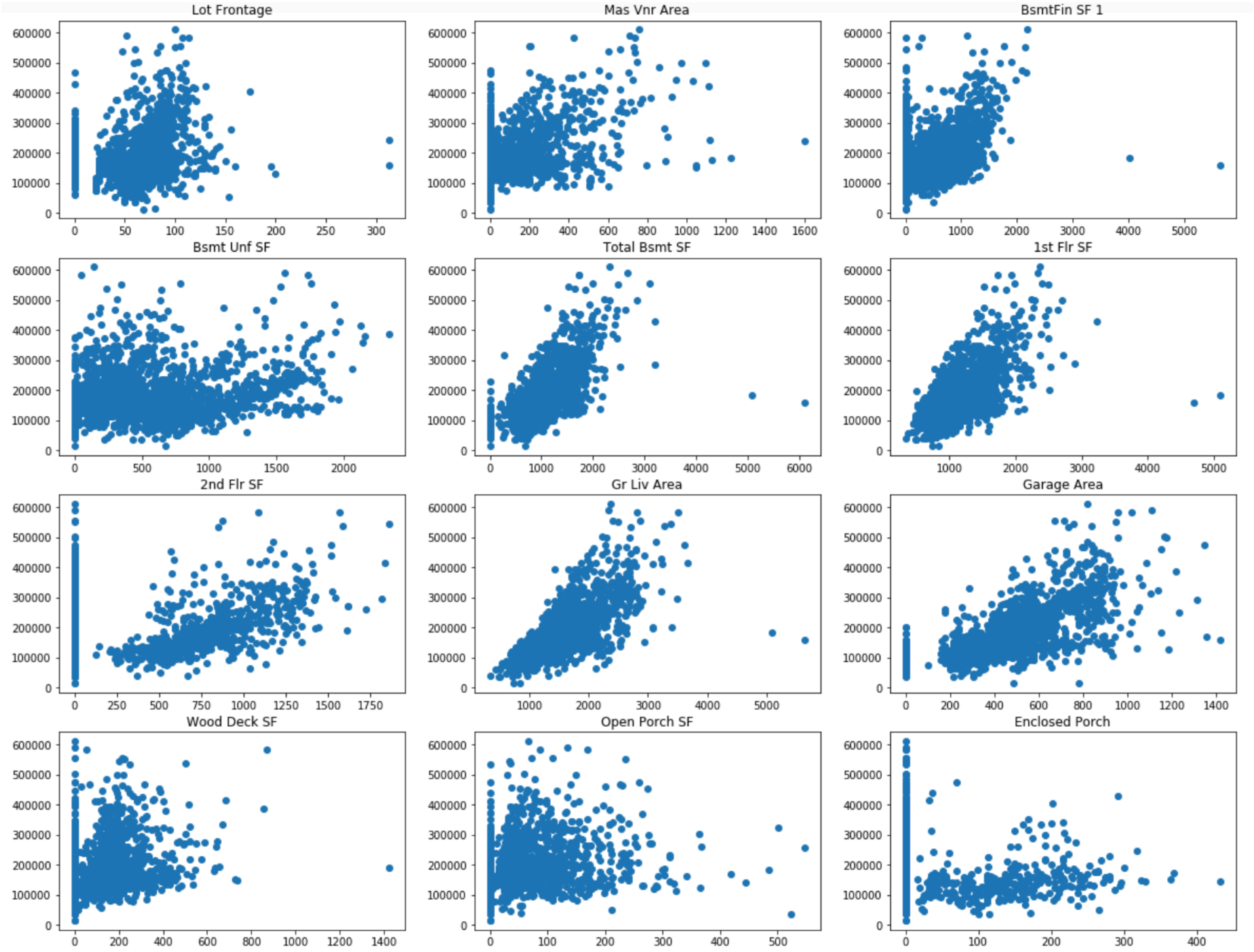
- Missing values means that house does not have these features
- Filled the rest with 0 if is numeric or 'none' if is string as

	missing_count	missing_percentage
Pool QC	2042	99.561190
Misc Feature	1986	96.830814
Alley	1911	93.174061
Fence	1651	80.497318
Fireplace Qu	1000	48.756704
Lot Frontage	330	16.089712
Garage Qual	114	5.558264
Garage Cond	114	5.558264
Garage Yr Blt	114	5.558264
Garage Finish	114	5.558264
Garage Type	113	5.509508
Bsmt Exposure	58	2.827889
BsmtFin Type 2	56	2.730375
Bsmt Qual	55	2.681619
BsmtFin Type 1	55	2.681619
Bsmt Cond	55	2.681619
Mas Vnr Type	22	1.072647
Mas Vnr Area	22	1.072647
Bsmt Half Bath	2	0.097513
Bsmt Full Bath	2	0.097513
Bsmt Unf SF	1	0.048757
Total Bsmt SF	1	0.048757
BsmtFin SF 1	1	0.048757
BsmtFin SF 2	1	0.048757
Garage Cars	1	0.048757
Garage Area	1	0.048757

Histograms (Discrete variables)

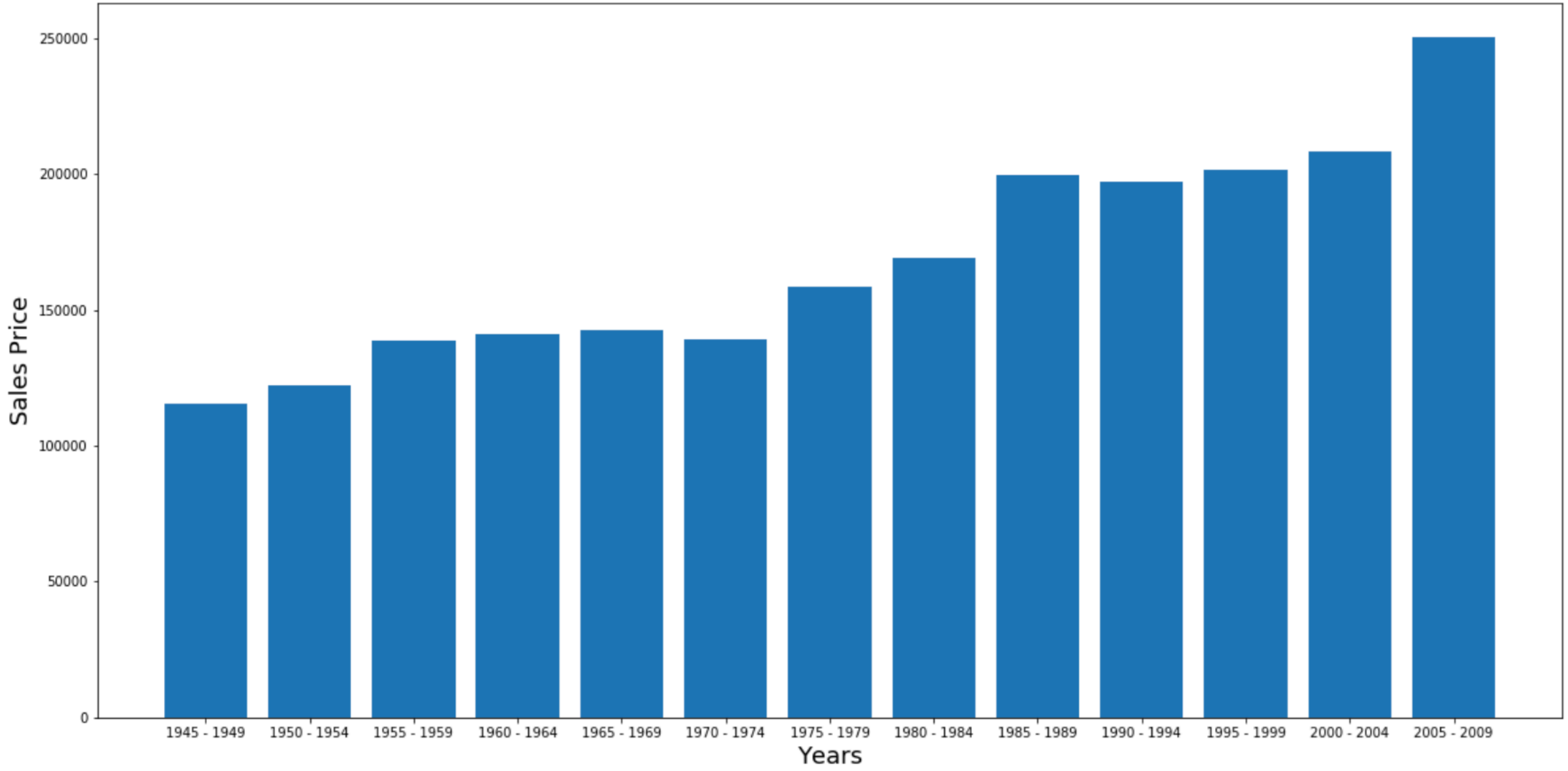


Scatter plots (Continuous variables)



Able to spot outliers from the graph and Correlation with sales price

Bar graph of year built and sales price by binning to 5 years

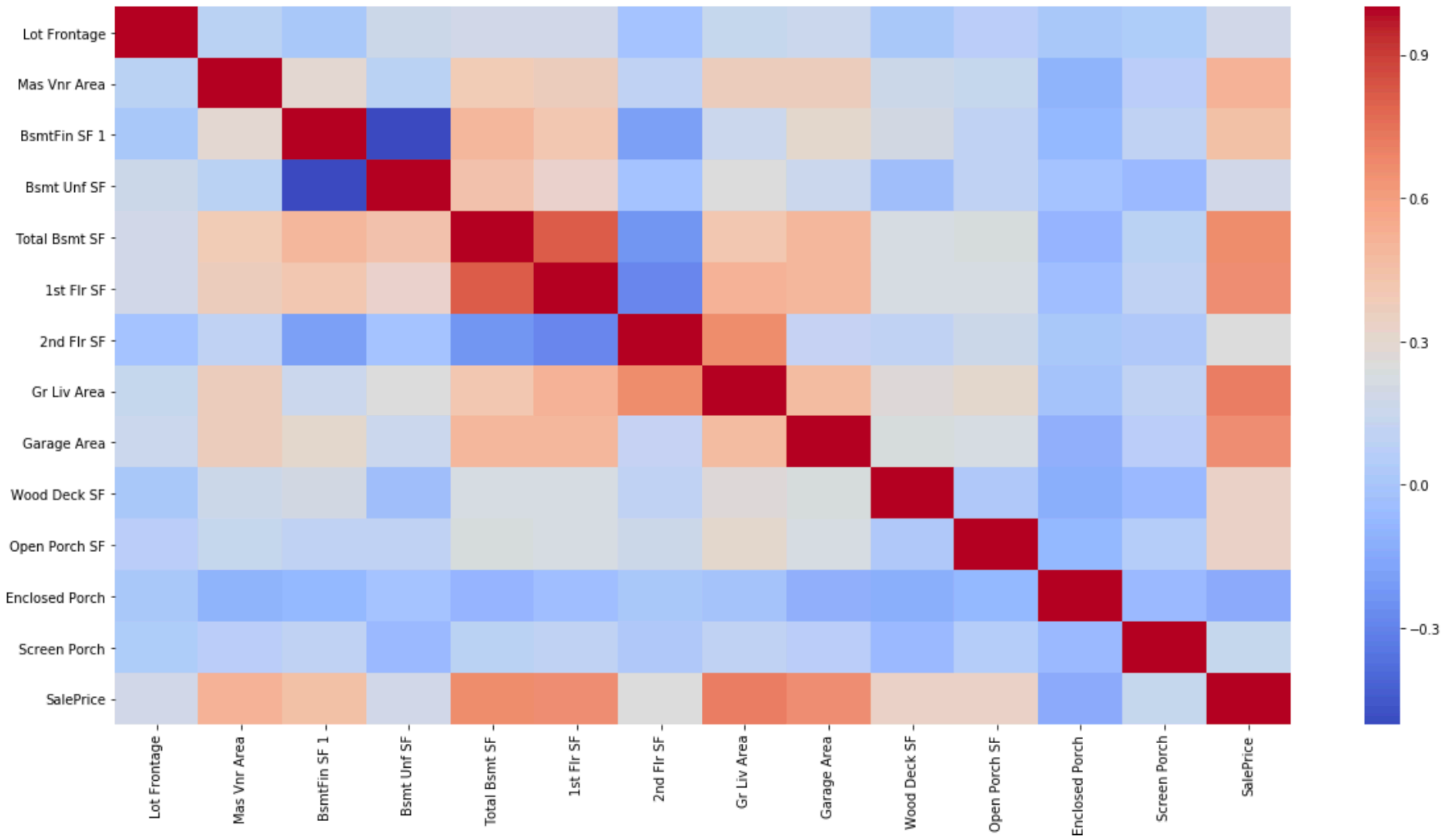


Label Encoding for Ordinal Categories

	Lot Shape	Overall Qual	Overall Cond	Exter Qual	Exter Cond	Bsmt Qual	Bsmt Cond	Bsmt Exposure	BsmtFin Type 1	BsmtFin Type 2	Heating QC	Kitchen Qual	Fireplace Qu	Garage Finish	Garage Qual	Fence
0	IR1	6	8	Gd	TA	TA	TA	No	GLQ	Unf	Ex	Gd	None	RFn	TA	None
1	IR1	7	5	Gd	TA	Gd	TA	No	GLQ	Unf	Ex	Gd	TA	RFn	TA	None
2	Reg	5	7	TA	Gd	TA	TA	No	GLQ	Unf	TA	Gd	None	Unf	TA	None
3	Reg	5	5	TA	TA	Gd	TA	No	Unf	Unf	Gd	TA	None	Fin	TA	None
4	IR1	6	8	TA	TA	Fa	Gd	No	Unf	Unf	TA	TA	None	Unf	TA	None

	Lot Shape	Overall Qual	Overall Cond	Exter Qual	Exter Cond	Bsmt Qual	Bsmt Cond	Bsmt Exposure	BsmtFin Type 1	BsmtFin Type 2	Heating QC	Kitchen Qual	Fireplace Qu	Garage Finish	Garage Qual	Fence
0	3	6	8	4	3	4	4	2	7	2	5	4	1	3	4	1
1	3	7	5	4	3	5	4	2	7	2	5	4	4	3	4	1
2	4	5	7	3	4	4	4	2	7	2	3	4	1	2	4	1
3	4	5	5	3	3	5	4	2	2	2	4	3	1	4	4	1
4	3	6	8	3	3	3	5	2	2	2	3	3	1	2	4	1

Heatmap of continuous variable coefficients



To visualize and select feature for polynomial fit

Model Scoring and Lasso Coefficients

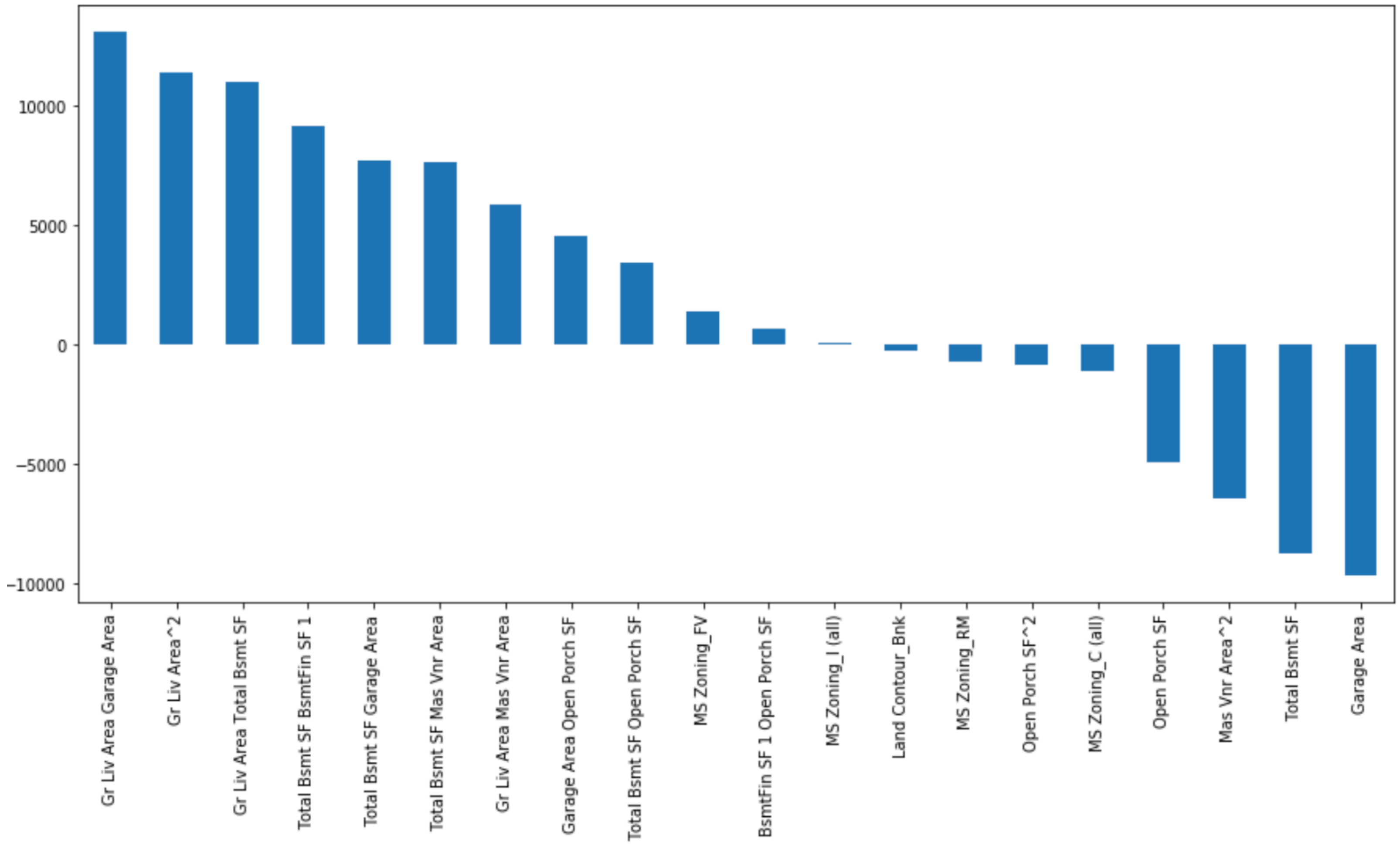
```
lr_rmse: 23891.142701776815
lasso_rmse: 22097.805793542546
ridge_rmse: 23324.57635448145

lr_adj_r2 score: 0.9057774284266177
lasso_adj_r2 score: 0.9193917647460038
ridge_adj_r2 score: 0.9101933209149107
```

Lasso performs the best

	variable	coef	abs_coef
8	Gr Liv Area Total Bsmt SF	13203.101473	13203.101473
7	Gr Liv Area^2	12866.575935	12866.575935
138	Overall Qual	11982.731448	11982.731448
9	Gr Liv Area Garage Area	8985.172512	8985.172512
18	Total Bsmt SF BsmtFin SF 1	8734.311691	8734.311691
1	Total Bsmt SF	-6676.548438	6676.548438
83	Bldg Type_1Fam	6625.739631	6625.739631
15	Total Bsmt SF Garage Area	6524.715362	6524.715362
71	Neighborhood_StoneBr	6377.002398	6377.002398
65	Neighborhood_NridgHt	5765.496508	5765.496508
153	Year Built	5739.688246	5739.688246
139	Overall Cond	5589.802803	5589.802803
17	Total Bsmt SF Mas Vnr Area	4615.313843	4615.313843
144	Bsmt Exposure	4131.457485	4131.457485
134	Sale Type_New	4072.044164	4072.044164

Top 20 lasso coefficients



Different iterations with different parameters

Data & Model	Score difference	Kaggle Score (RMSE)
Using top 5 continuous variables that are the most correlated with sales price without any cleaning	First Submission	43013
After cleaning most of the data mentioned from point 1 to 12 and performing RidgeCV with binning years	Improved	29537
Same as point 2 but using LassoCV and not binning years	Not much changes	29410
Same as point 3, and in addition dropping low variance nominal features which includes (Lot Config Roof Style, MS Sub Class and Exterior 2nd) with LassoCV	Improved	28066
Same as point 5 and dropping features with 0 lasso coefficient from training model	Deproved	23272
Same as point 4 and applying polynomial features with LassoCV	Improved	22542

Conclusion

- Final model selected based on highest adjusted r squared score on test and lowest RMSE score on Kaggle.
- The final selected model is to use lasso regression to reduce model complexity and uses about 113 features including dummies variables to achieve 22542 RMSE score.
  - It might however be overfitted due to the addition of polynomial features as we can see a sharp decrease in bias (low rmse score) and an increase between the difference the public and private score. (~8k difference)
- To have a more generalized model, it is better to not use polynomial features for training as it has low difference between public and private score. (less than 1k difference)

Conclusion

- To have a more generalized model, it is better to not use polynomial features for training as it has low difference between public and private score. (less than 1k difference)

- Using top 30 lasso coefficient features increases the RMSE score as compared to 100 features (Bias and Variance trade-off)