Project 2:
Ames Housing Data

Le-Anne Lee

Scope

- Problem statement
- Analysis of Data
- Model Findings
- Conclusion

Problem Statement

How can we predict the sale price of a house in Ames for a potential seller?

What data we had?

- 2051 observations (train set)
- 879 observations (test set)

}

<u>80 Features</u>

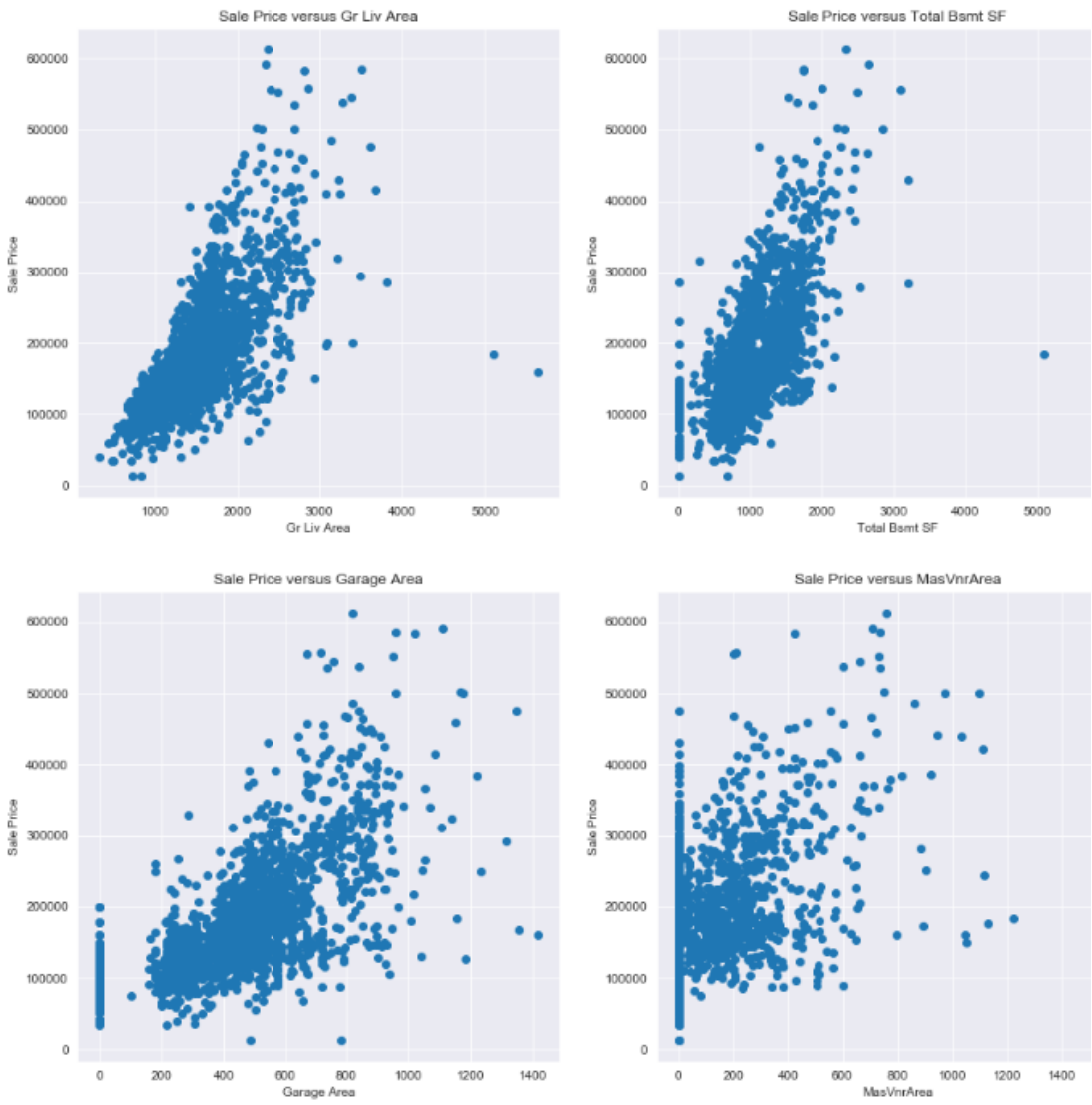
- 22 Nominal Features
- 14 Discrete Features
- 23 Ordinal Features
- 19 Continuous Features
- ID
- Sale Price

	Id	PID	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood
0	109	533352170	60	RL	NaN	13517	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	Sawyer
1	544	531379050	60	RL	43.0	11492	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	SawyerW
2	153	535304180	20	RL	68.0	7922	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	NAmes
3	318	916386060	60	RL	73.0	9802	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Timber
4	255	906425045	50	RL	82.0	14235	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	SawyerW

Processing the data…

Continuous Features

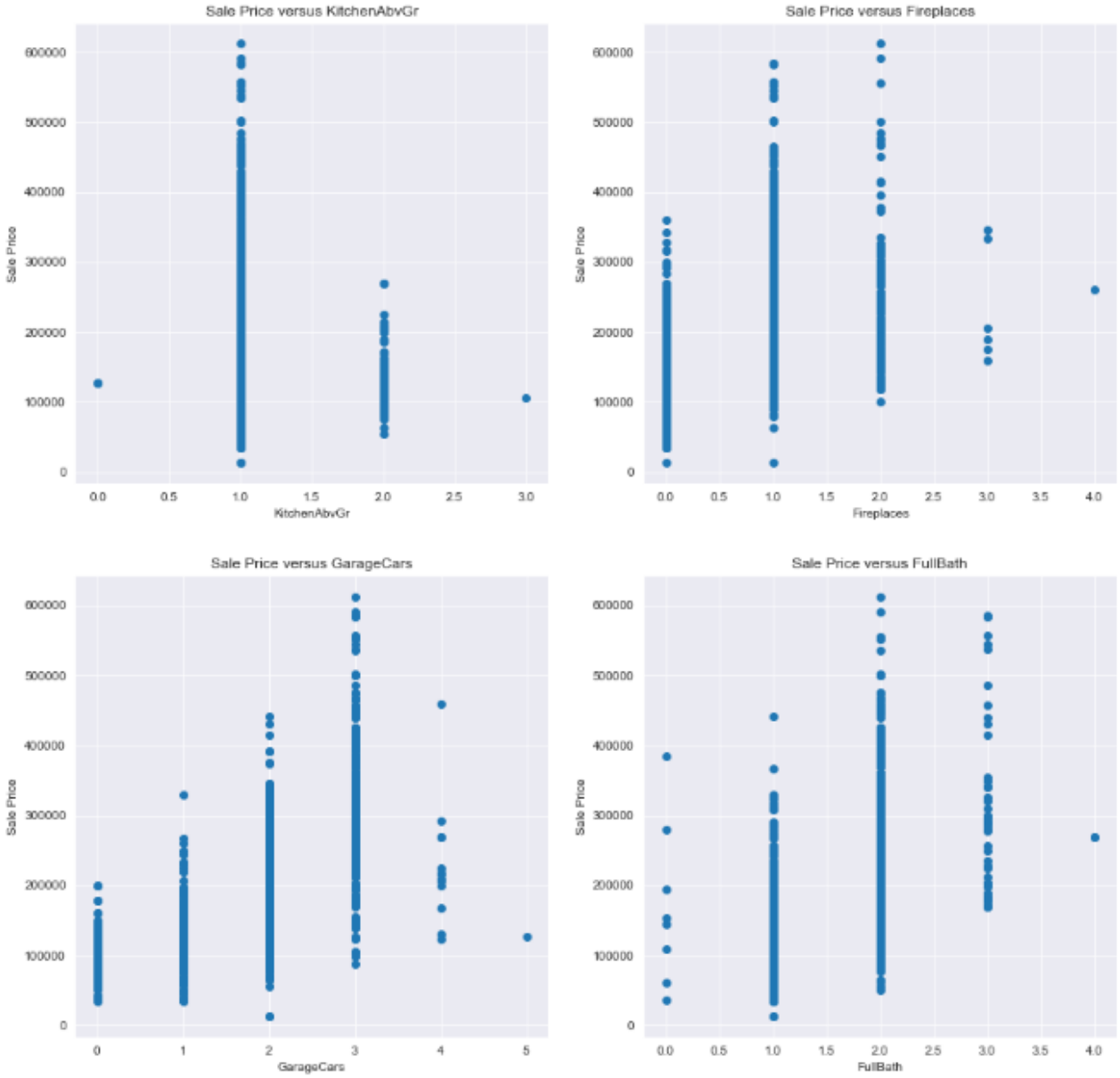
- High correlation between sale price and Gr Liv Area, Tot Bsmt SqFt and Garage Area
- Positive correlation
- Zeros…



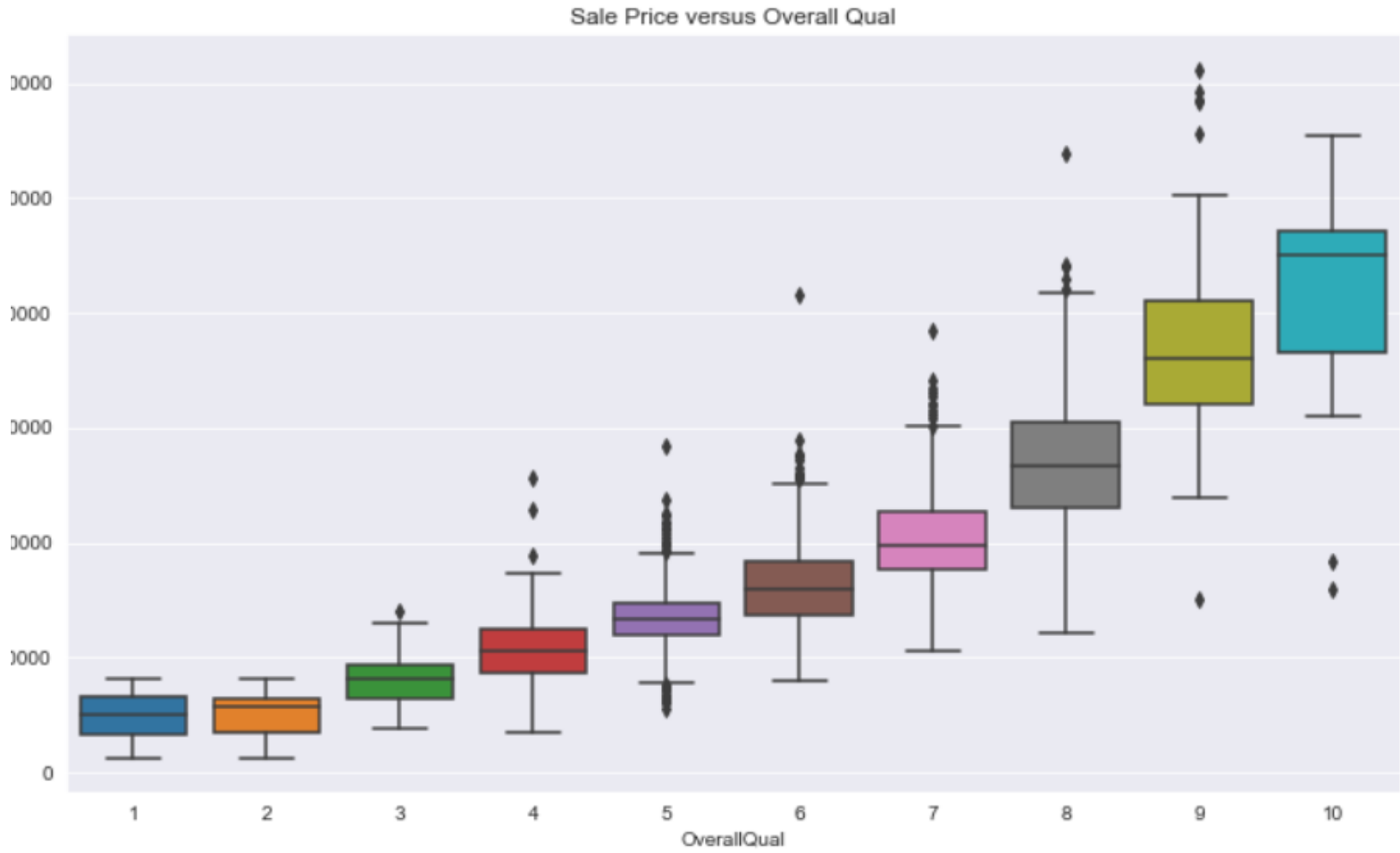
More is more…

Discrete Features

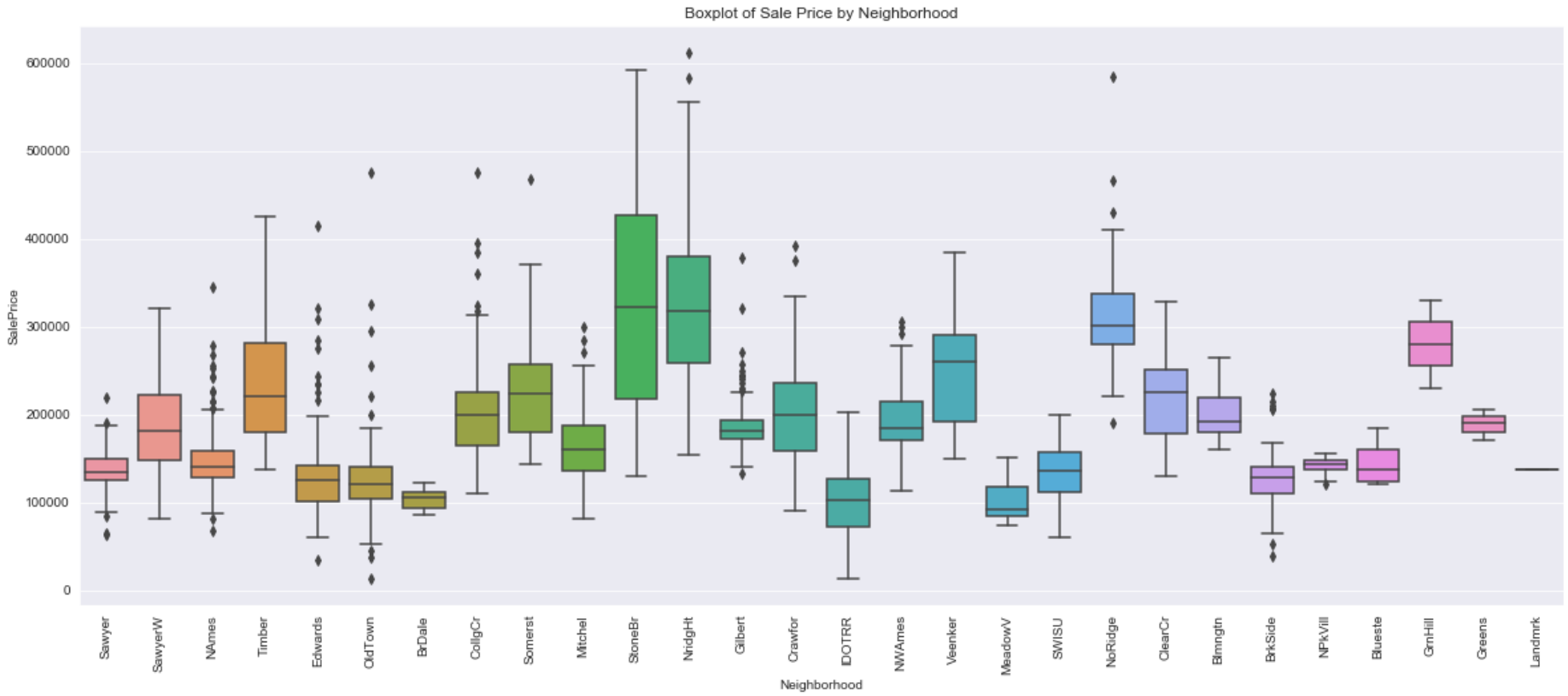
- General trend that sale price increases when the feature increases
- Kitchen Abv Gr being an exception



Quality pays...

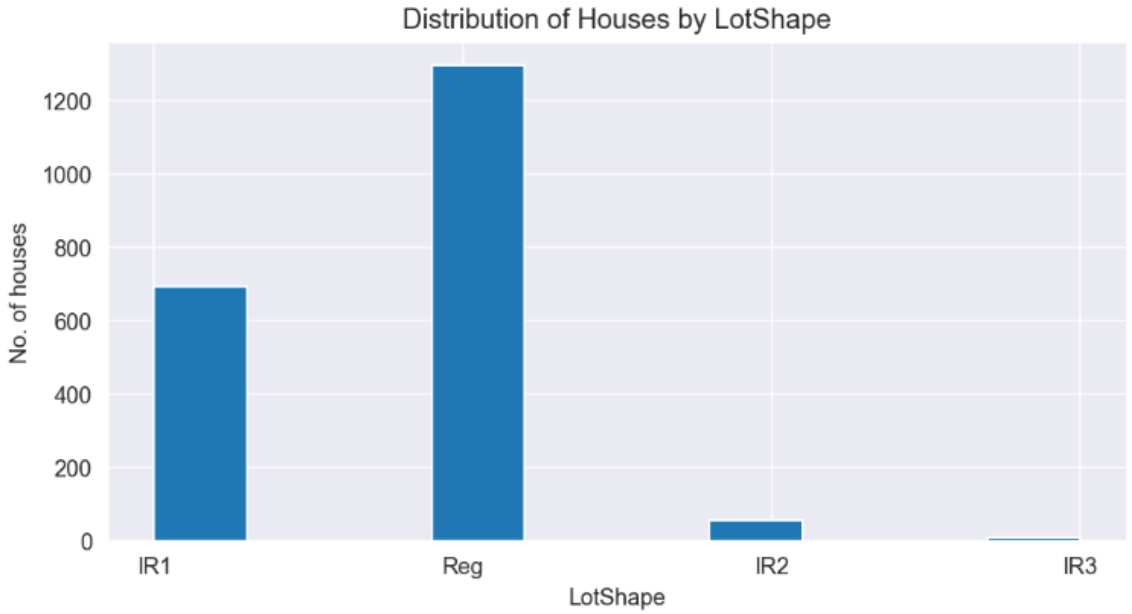


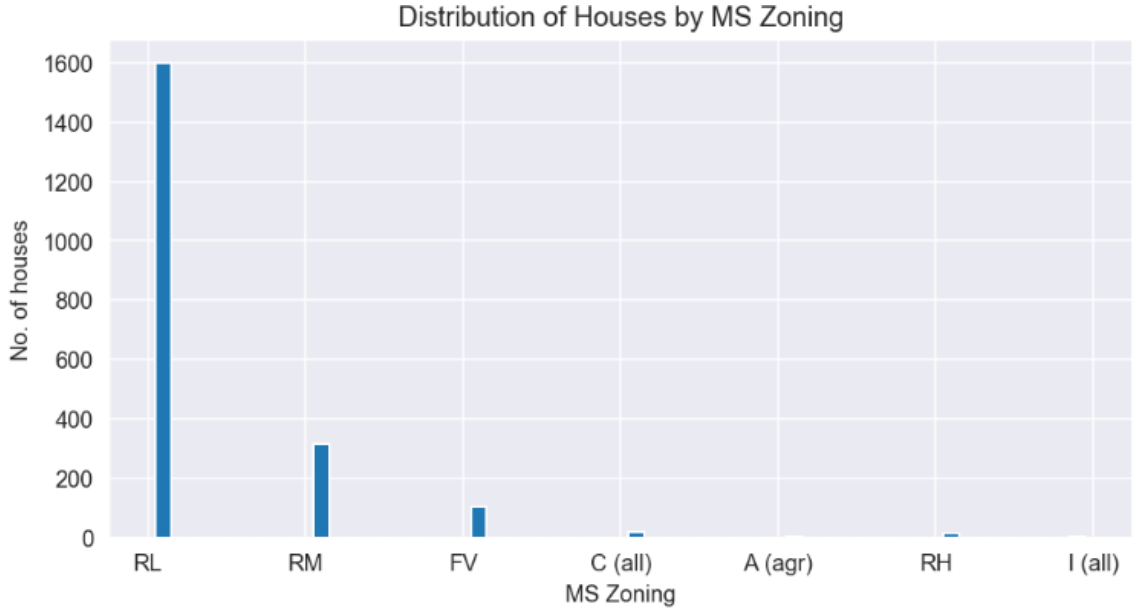
My neighbourhood is better than yours…



Other Observations

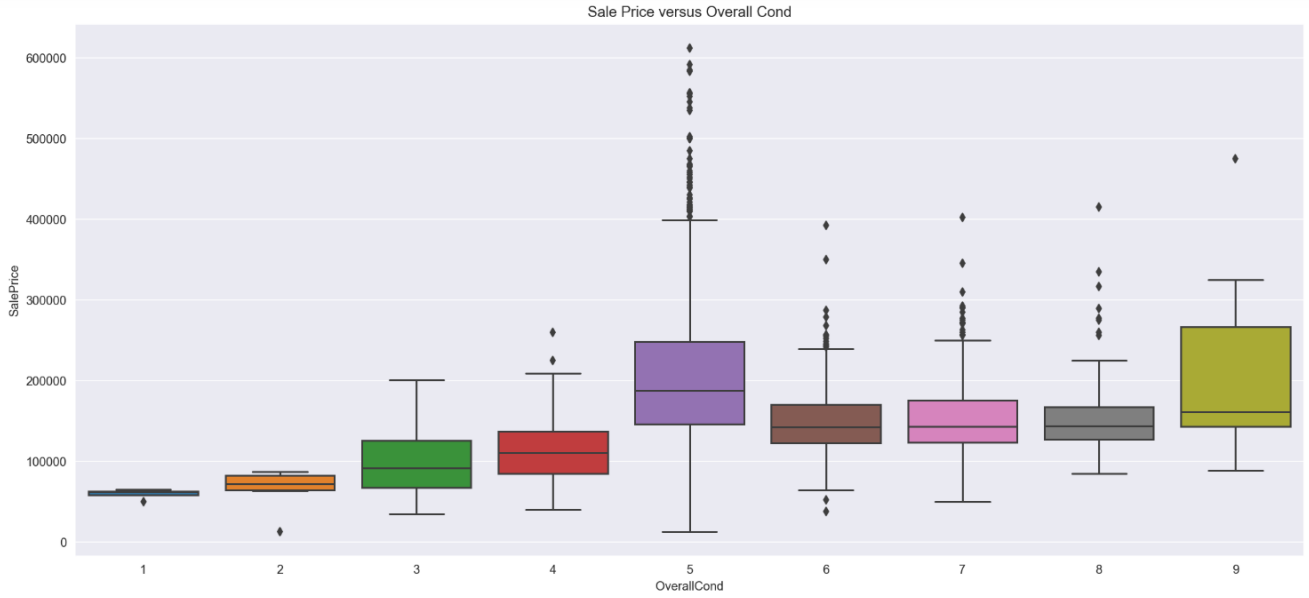
- Highly Skewed Distributions





Other Observations

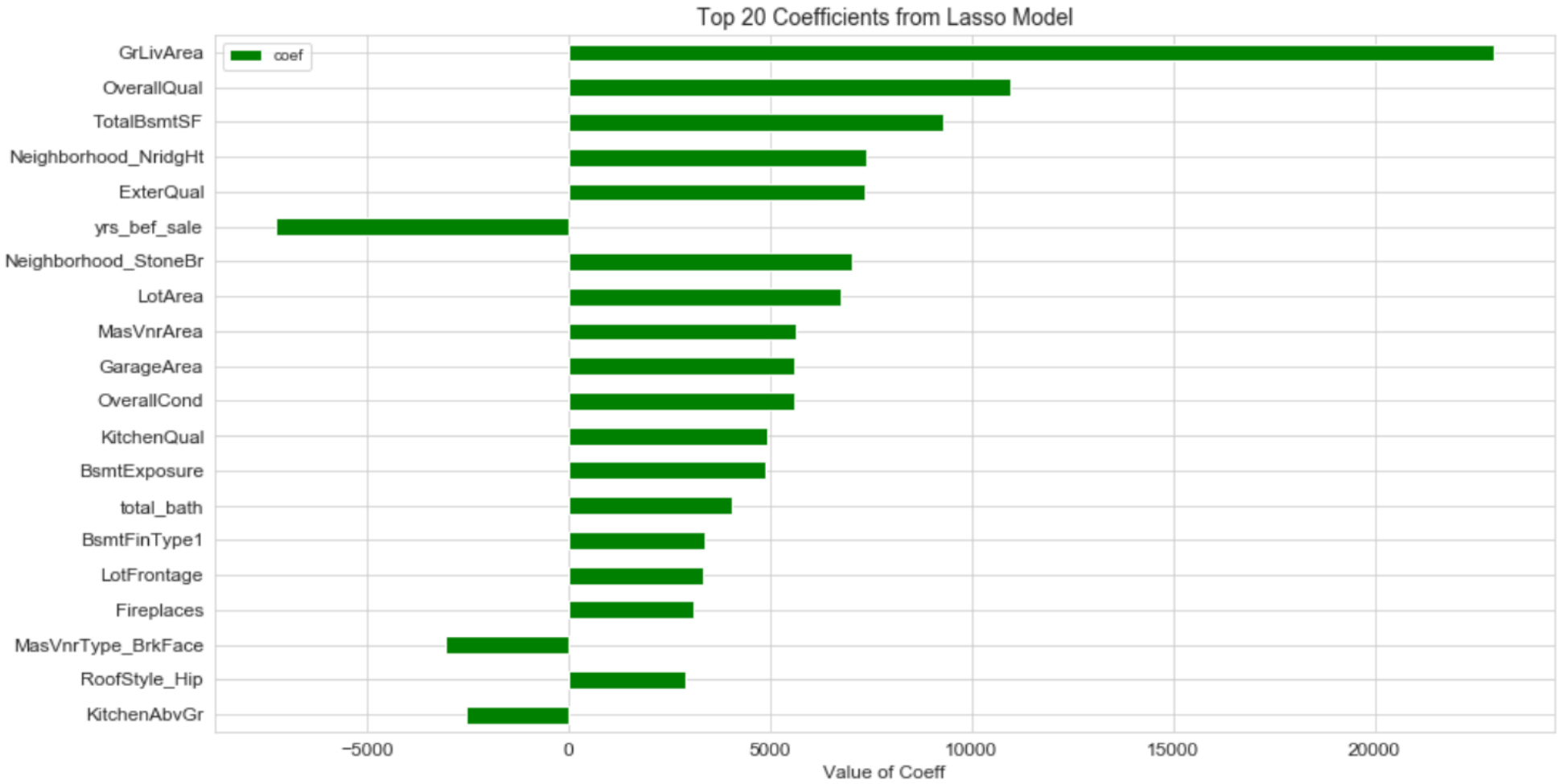
- Highly Skewed Distributions



Modeling

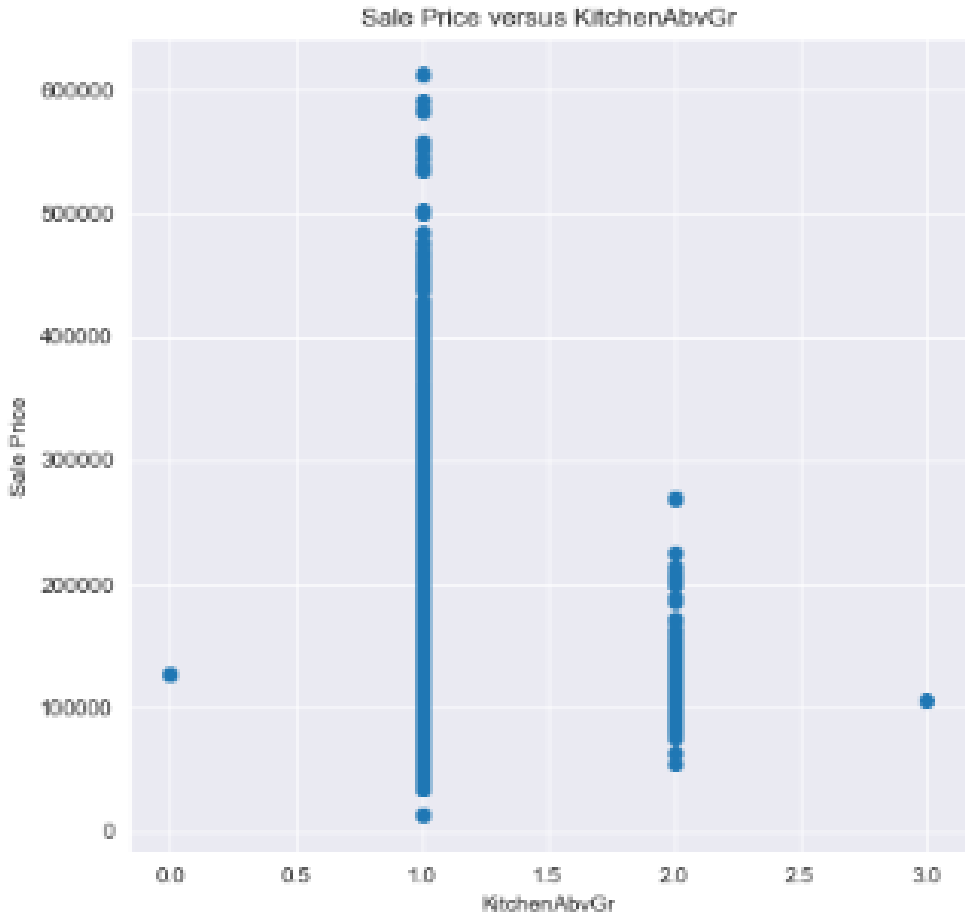
- Lasso and Ridge performed better at the onset where there were too many features (100 over) (Adj R^2 =0.866, RMSE=26,603)
- Top 20 features from Lasso model were selected but improvement in R^2 score was marginal (Adj R^2 = 0.8722, RMSE=28,912)
- Top 30 features yielded slightly better scores (Adj R^2 = 0.884, RMSE=27,169)
- Top 5 features were squared to explore polynomial relationships (Adj R^2 = 0.899, RMSE=25,741)

Top 20 Features



Tell me why?

- Highly Skewed Distributions



Conclusion & Recommendations

- Gr Liv Area stood out as the feature with the greatest weightage on sale price
- There may be a polynomial relationship with some of the features
- Effect of using features with low variance on the model
- Neighbourhood seems to have a significant effect on sale price and should be explored further
- To explore combining features further