Ames Housing Data

Understanding Features that Affect Sale Price

Problem Statement

In this project, we will understand the features that affect the Sale Price of housing in Ames, Iowa. We aim to find a production model that we can use to get a good score on Kaggle, as well as find business insights on what helps to increase Sale Price.

Data Cleaning

We remove 3 rows where only 1 or 2 values are missing for all rows of the feature.

We then check the Data Documentation from http://jse.amstat.org/ and impute the categorical features with ‘N_A’ as appropriate.

After the first round of cleaning, we have 2046 rows compared to the original 2051 rows.

Feature Selection

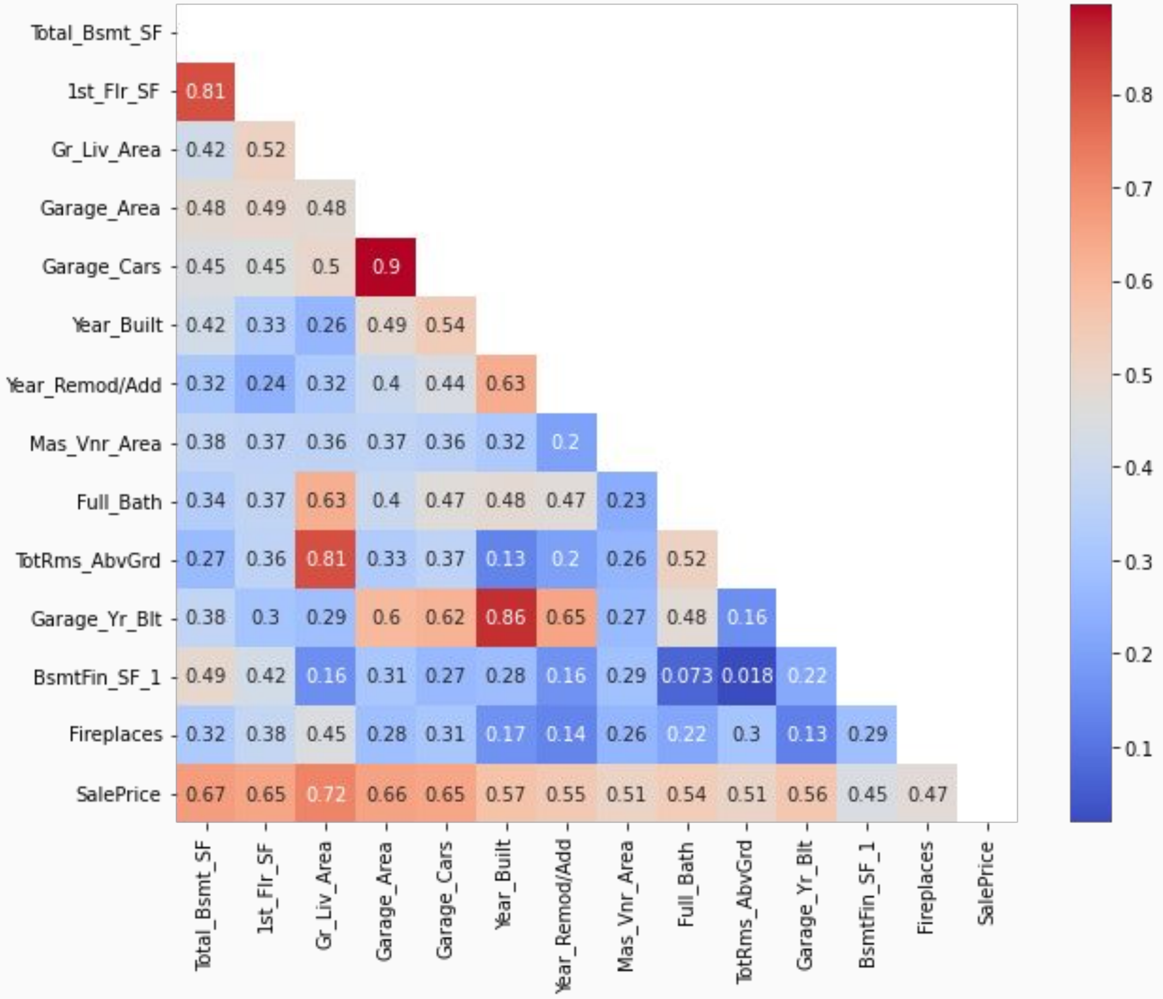
We divide the features into 3 types

- Values (numeric, either integer or float)
- Ordinal (categories with inherent order)
- Nominal (categories without inherent order)

Features Selection (Values)

We use Seaborn’s Heatmap to see which features have high correlation with Sale Price.

These 13 features have correlation above 0.4, but 4 features were removed as they were inter-linked.



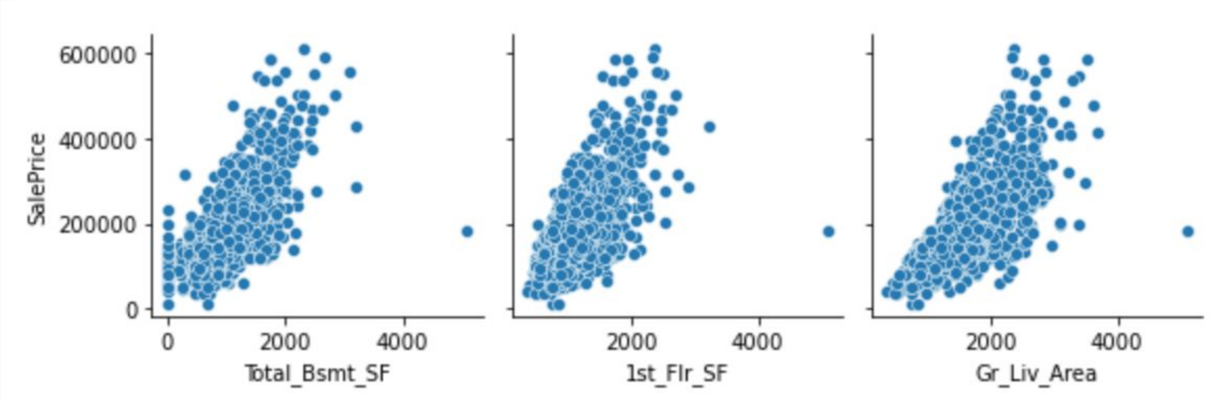
	Total_Bsmt_SF	1st_Flr_SF	Gr_Liv_Area	Garage_Area	Garage_Cars	Year_Built	Year_Remod/Add	Mas_Vnr_Area	Full_Bath	TotRms_AbvGrd	Garage_Yr_Blt	BsmtFin_SF_1	Fireplaces	SalePrice
Total_Bsmt_SF														
1st_Flr_SF	0.81													
Gr_Liv_Area	0.42	0.52												
Garage_Area	0.48	0.49	0.48											
Garage_Cars	0.45	0.45	0.5	0.9										
Year_Built	0.42	0.33	0.26	0.49	0.54									
Year_Remod/Add	0.32	0.24	0.32	0.4	0.44	0.63								
Mas_Vnr_Area	0.38	0.37	0.36	0.37	0.36	0.32	0.2							
Full_Bath	0.34	0.37	0.63	0.4	0.47	0.48	0.47	0.23						
TotRms_AbvGrd	0.27	0.36	0.81	0.33	0.37	0.13	0.2	0.26	0.52					
Garage_Yr_Blt	0.38	0.3	0.29	0.6	0.62	0.86	0.65	0.27	0.48	0.16				
BsmtFin_SF_1	0.49	0.42	0.16	0.31	0.27	0.28	0.16	0.29	0.073	0.018	0.22			
Fireplaces	0.32	0.38	0.45	0.28	0.31	0.17	0.14	0.26	0.22	0.3	0.13	0.29		
SalePrice	0.67	0.65	0.72	0.66	0.65	0.57	0.55	0.51	0.54	0.51	0.56	0.45	0.47	

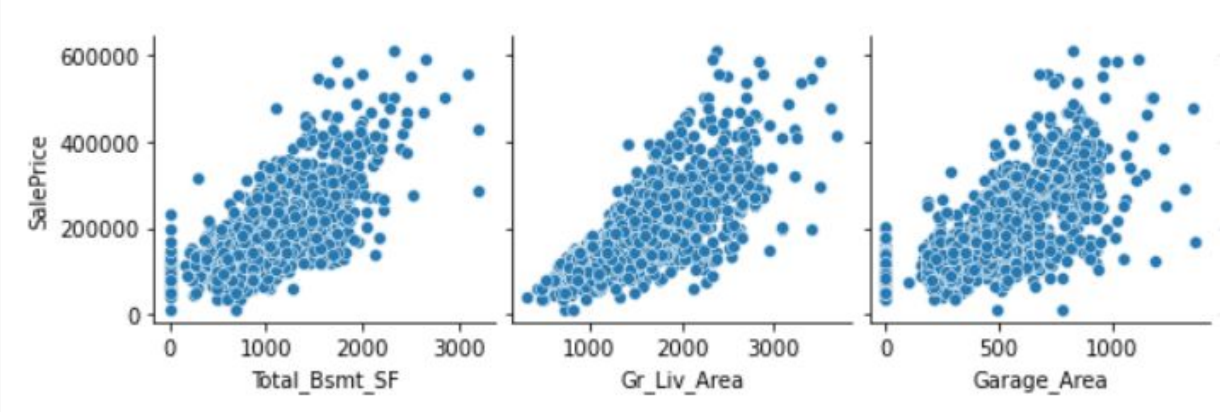
Features Selection (Values)

We use Seaborn's Pairplot and found an outlier, which we removed.

We also see from the plots that all 9 features are largely linearly correlated with sale price and decide to use Linear models later.

We end up with 9 features at the end of this step.

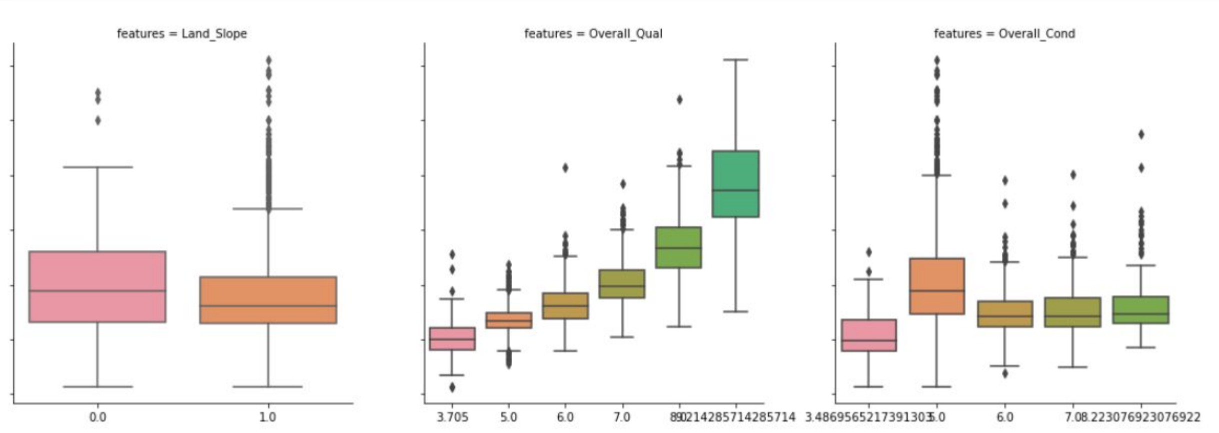




Features Selection (Ordinal)

We create a function to encode the features such that higher numbers should correspond to higher sale price.

We then use boxplots to visualize the data to shortlist the features we wish to further investigate.

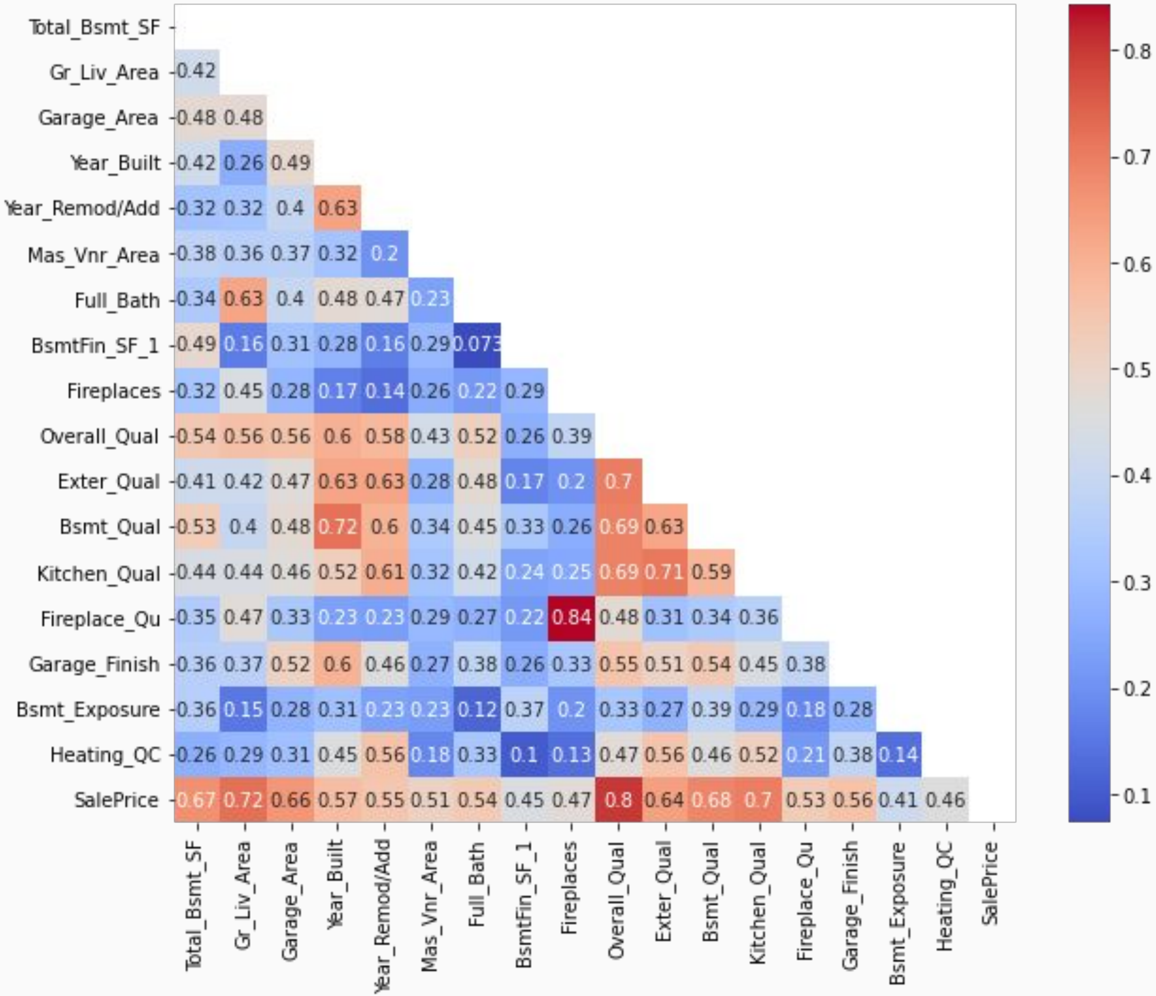


Features Selection (Ordinal)

We use Seaborn's Heatmap to see which features have high correlation with Sale Price, to select 8 ordinal features with correlation above 0.4.

We then plot another Heatmap with the 9 features with values and find one pair with high correlation, so we remove 1 feature.

We end up with 16 features at the end of this step.

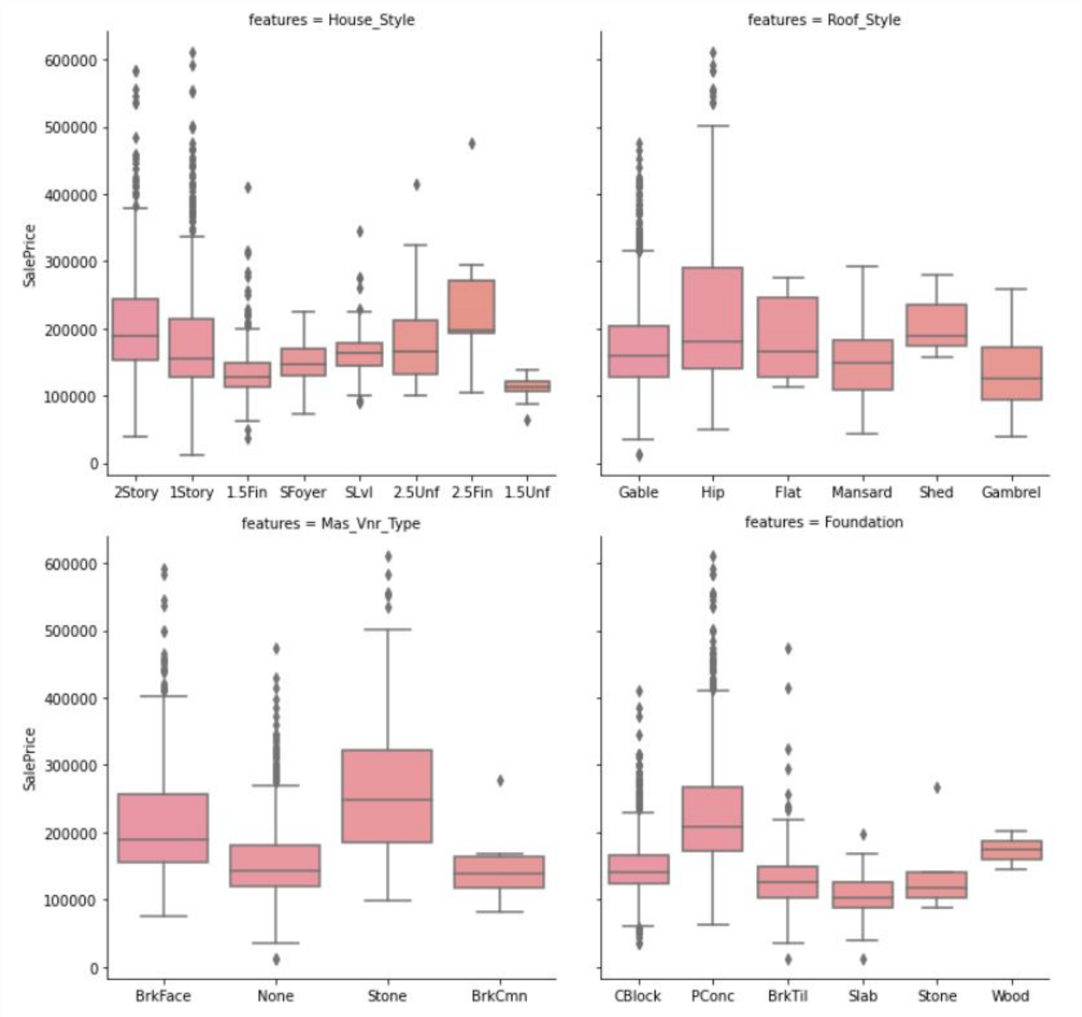


	Total_Bsmt_SF	Gr_Liv_Area	Garage_Area	Year_Built	Year_Remod/Add	Mas_Vnr_Area	Full_Bath	BsmtFin_SF_1	Fireplaces	Overall_Qual	Exter_Qual	Bsmt_Qual	Kitchen_Qual	Fireplace_Qu	Garage_Finish	Bsmt_Exposure	Heating_QC	SalePrice
Total_Bsmt_SF																		
Gr_Liv_Area	0.42																	
Garage_Area	0.48	0.48																
Year_Built	0.42	0.26	0.49															
Year_Remod/Add	0.32	0.32	0.4	0.63														
Mas_Vnr_Area	0.38	0.36	0.37	0.32	0.2													
Full_Bath	0.34	0.63	0.4	0.48	0.47	0.23												
BsmtFin_SF_1	0.49	0.16	0.31	0.28	0.16	0.29	0.073											
Fireplaces	0.32	0.45	0.28	0.17	0.14	0.26	0.22	0.29										
Overall_Qual	0.54	0.56	0.56	0.6	0.58	0.43	0.52	0.26	0.39									
Exter_Qual	0.41	0.42	0.47	0.63	0.63	0.28	0.48	0.17	0.2	0.7								
Bsmt_Qual	0.53	0.4	0.48	0.72	0.6	0.34	0.45	0.33	0.26	0.69	0.63							
Kitchen_Qual	0.44	0.44	0.46	0.52	0.61	0.32	0.42	0.24	0.25	0.69	0.71	0.59						
Fireplace_Qu	0.35	0.47	0.33	0.23	0.23	0.29	0.27	0.22	0.84	0.48	0.31	0.34	0.36					
Garage_Finish	0.36	0.37	0.52	0.6	0.46	0.27	0.38	0.26	0.33	0.55	0.51	0.54	0.45	0.38				
Bsmt_Exposure	0.36	0.15	0.28	0.31	0.23	0.23	0.12	0.37	0.2	0.33	0.27	0.39	0.29	0.18	0.28			
Heating_QC	0.26	0.29	0.31	0.45	0.56	0.18	0.33	0.1	0.13	0.47	0.56	0.46	0.52	0.21	0.38	0.14		
SalePrice	0.67	0.72	0.66	0.57	0.55	0.51	0.54	0.45	0.47	0.8	0.64	0.68	0.7	0.53	0.56	0.41	0.46	

Features Selection (Nominal)

We create boxplots to see if there is a reasonable way to split the categories of a nominal feature into 2 groups with distinct Sale Price.

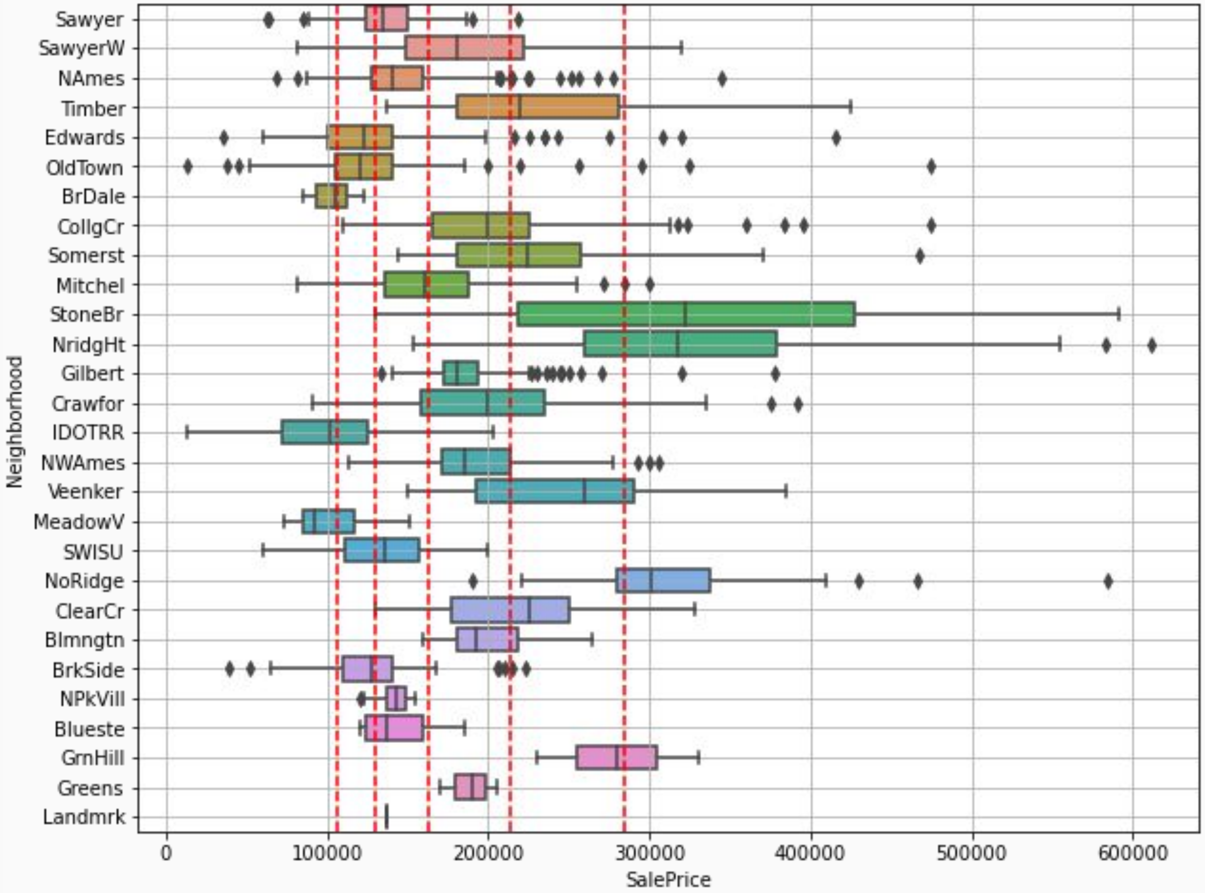
For example, for the feature Foundation, PConc has higher Sale Price and will be encoded with 1, while the remaining categories will be encoded with 0.



Features Selection (Nominal)

For Neighborhood, we use the percentile lines at 10%, 25%, 50%, 75% and 90% to try to split the data.

We decide to encode into Neighborhood_High the categories where median Sale Price exceeds the 90% line and Neighborhood_Low the categories where median Sale Price is below the 25% line.

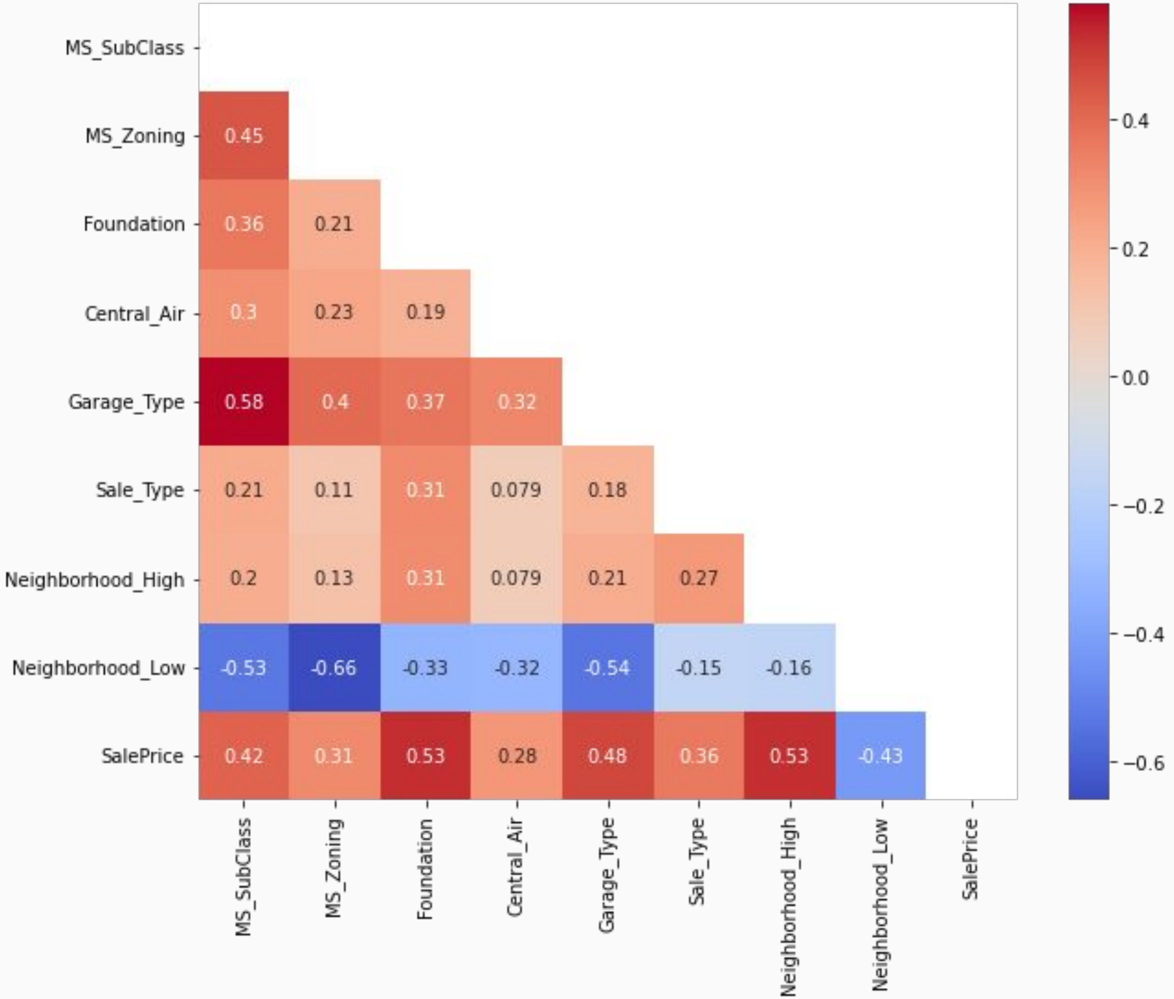


Features Selection (Nominal)

We shortlist the 5 features with correlation above 0.4.

Double checking using Heatmap with the earlier 16 features, no pairs with high correlation were found.

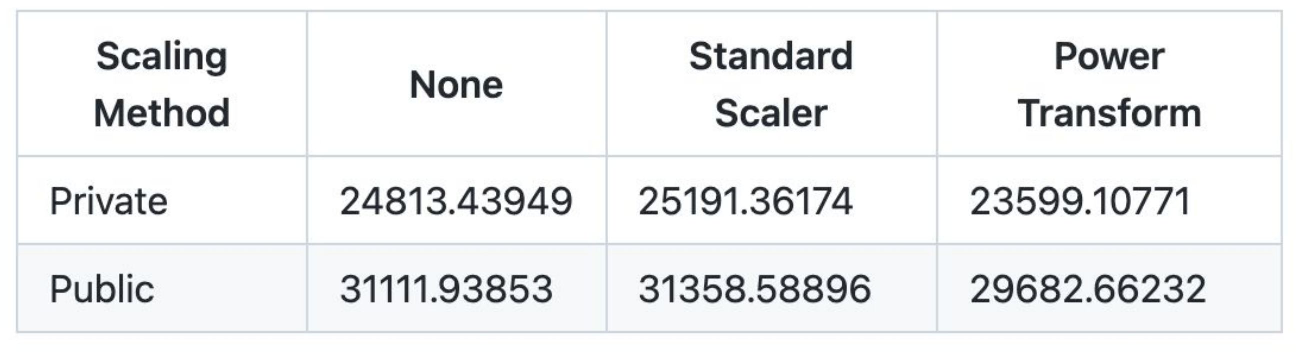
We proceed to model using these 21 features.



Model Selection

We use Linear Regression, LassoCV and RidgeCV with the 21 selected features with no scaling, using Standard Scaler and using Power Transform.

We found that the Linear Regression with Power Transform gave the best R2 and RMSE scores and use Linear Regression to submit for our Kaggle Scores.



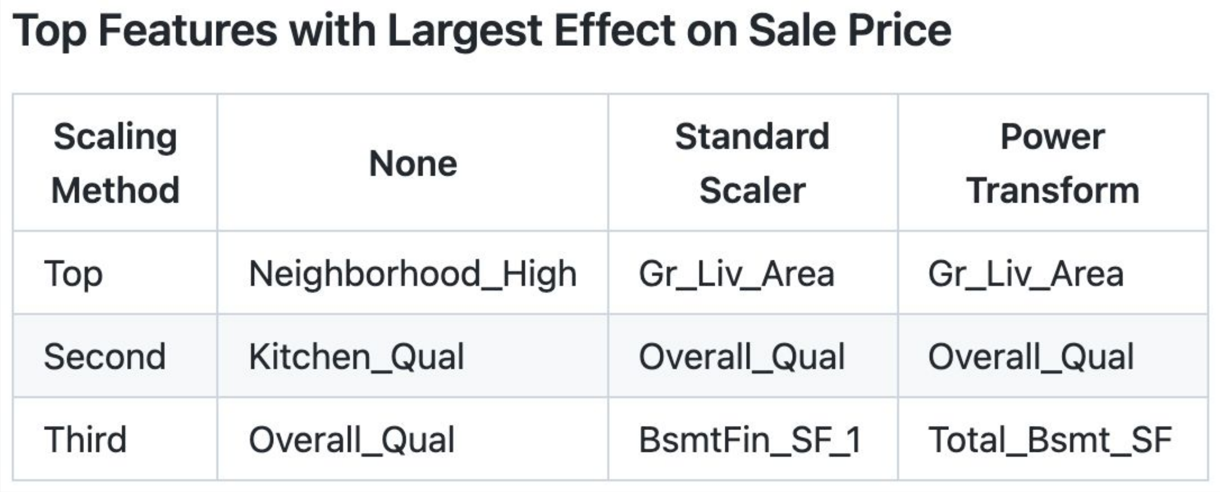
Scaling Method	None	Standard Scaler	Power Transform
Private	24813.43949	25191.36174	23599.10771
Public	31111.93853	31358.58896	29682.66232

Interpretation

We ignore the results from no scaling as it is inappropriate for our dataset with large variation in magnitudes of the raw values of different features.

The top 2 features are Gr_Liv_Area and Overall_Qual.

- For every 1 square foot increase in Gr_Liv_Area, SalePrice is expected to increase by around $13.65
- For every 1 point increase in Overall_Qual, SalePrice is expected to increase by around $1925.19



Top Features with Largest Effect on Sale Price

Scaling Method	None	Standard Scaler	Power Transform
Top	Neighborhood_High	Gr_Liv_Area	Gr_Liv_Area
Second	Kitchen_Qual	Overall_Qual	Overall_Qual
Third	Overall_Qual	BsmtFin_SF_1	Total_Bsmt_SF

Recommendations

To improve the chances of commanding a higher sale price, a house owner may wish to

- increase to living area above ground
- get a higher score in overall quality by using better material and having better finishing
- work on the basement to increase the area and exposure
- increase the garage area

For a house-hunter looking for a better deal, he might wish to avoid the 3 neighborhoods of Brookside, Northridge and Northridge Heights which command higher sale prices than the other neighborhoods in Ames.