PROJECT 2

PREDICTING HOUSE PRICES IN AMES, IOWA

AGENDA

- Problem Statement
- Cleaning Data
- Exploring Data
- Creating the Model
- Test the Model
- Recommendations

PROBLEM STATEMENT

- What is the best features for you to maximize the price of your house?

- Perform analysis on Data collected from the area

- Create a model to predict prices

CLEANING DATA

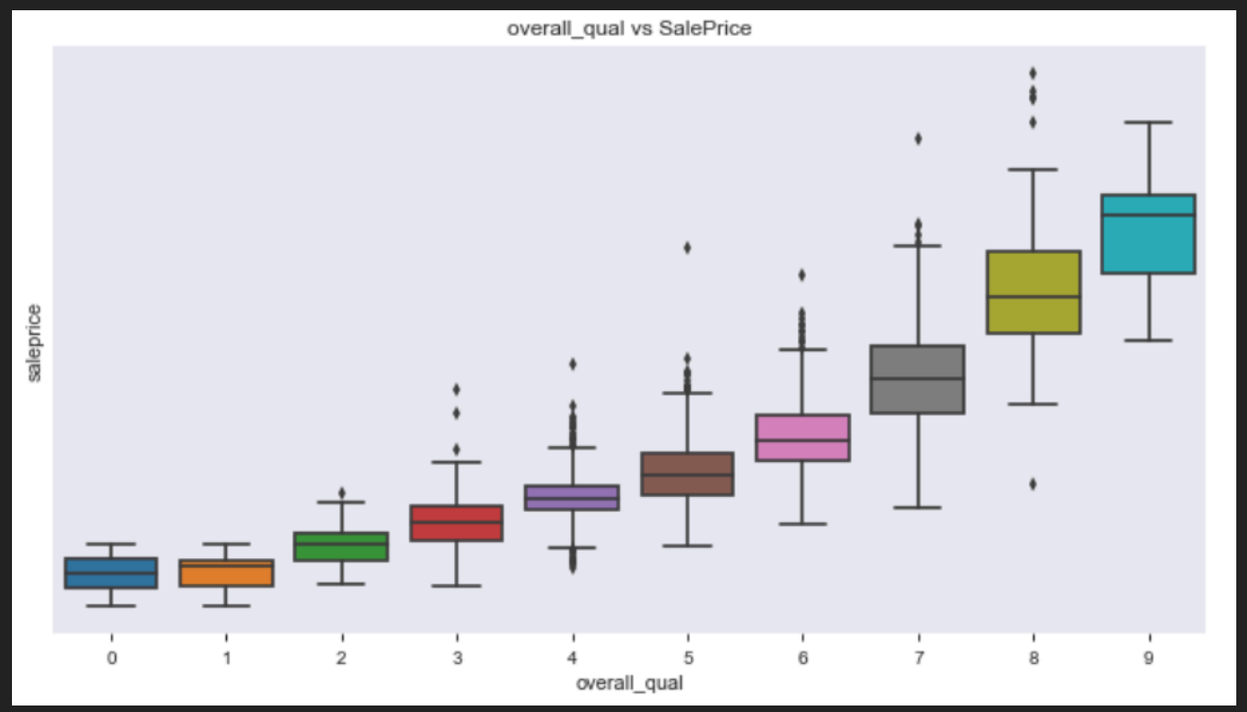
- Many Variable in the data had null values
- Dropped Features with too many null values
- Converted the rest according to description

Null Values Data Dictionary

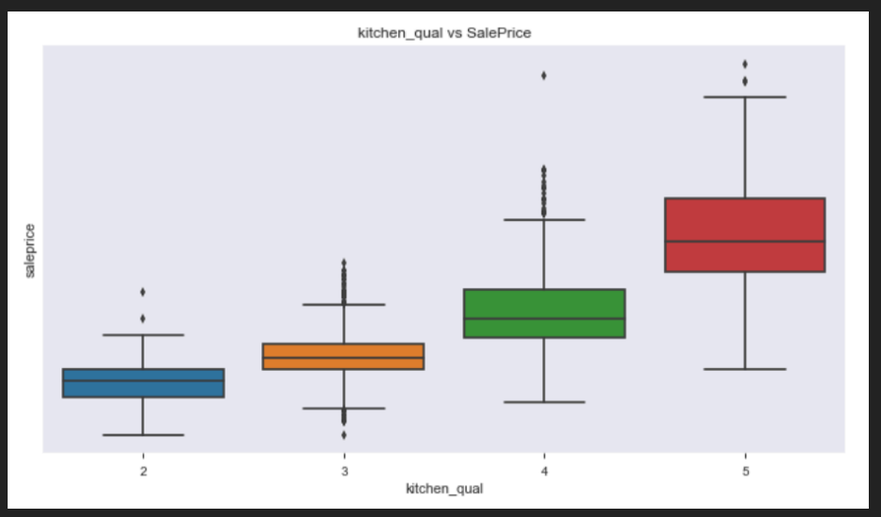
Feature	Type	Number of Null	Description
Pool QC	Object	2042	Null Value should be NA(Ordinal)
Misc Feature	Object	1986	Null Value should be NA(Ordinal)
Alley	Object	1911	Null Value should be NA(Nominal)
Fence	Object	1651	Null Value should be NA(Ordinal)
Fireplace Qu	Object	1000	Null Value should be NA(Ordinal)
Lot Frontage	Int	330	Null Value should be Null/0(Continous)
Garage Finish	Object	114	Null Value should be NA(Ordinal)
Garage Cond	Object	114	Null Value should be NA(Ordinal)
Garage Qual	Object	114	Null Value should be NA(Ordinal)
Garage Yr Blt	Int	114	Null Value remain Null(Discete)
Garage Type	Object	113	Null Value should be NA(Nomaial)
Bsmt Exposure	Object	58	Null Values should be NA(Ordinal)
BsmtFin Type 2	Object	56	Null Values should be NA(Ordinal)
BsmtFin Type 1	Object	55	Null Values should be NA(Ordinal)
Bsmt Cond	Object	55	Null Values should be NA(Ordinal)
Bsmt Qual	Object	55	Null Values should be NA(Ordinal)
Mas Vnr Type	Object	22	Null Values should be None(Nominal)
Mas Vnr Area	Int	22	Null Values should be 0(Continuous)
Bsmt Half Bath	Int	2	Null Values should be 0(Discrete)
Bsmt Full Bath	Int	2	Null Values should be 0(Discrete)
Garage Cars	Int	1	Null Values should be 0(Discrete)
Garage Area	Int	1	Null Values should be 0(Continous)
Bsmt Unf SF	Int	1	Null Values should be 0(Continous)
BsmtFin SF 2	Int	1	Null Values should be 0
Total Bsmt SF	Int	1	Null Values should be 0
BsmtFin SF 1	Int	1	Null Values should be 0

EXPLORING THE DATA

- Split Data into 3 sections (Continuous, Nominal, Ordinal)
- Ranked the Ordinal variables accordingly

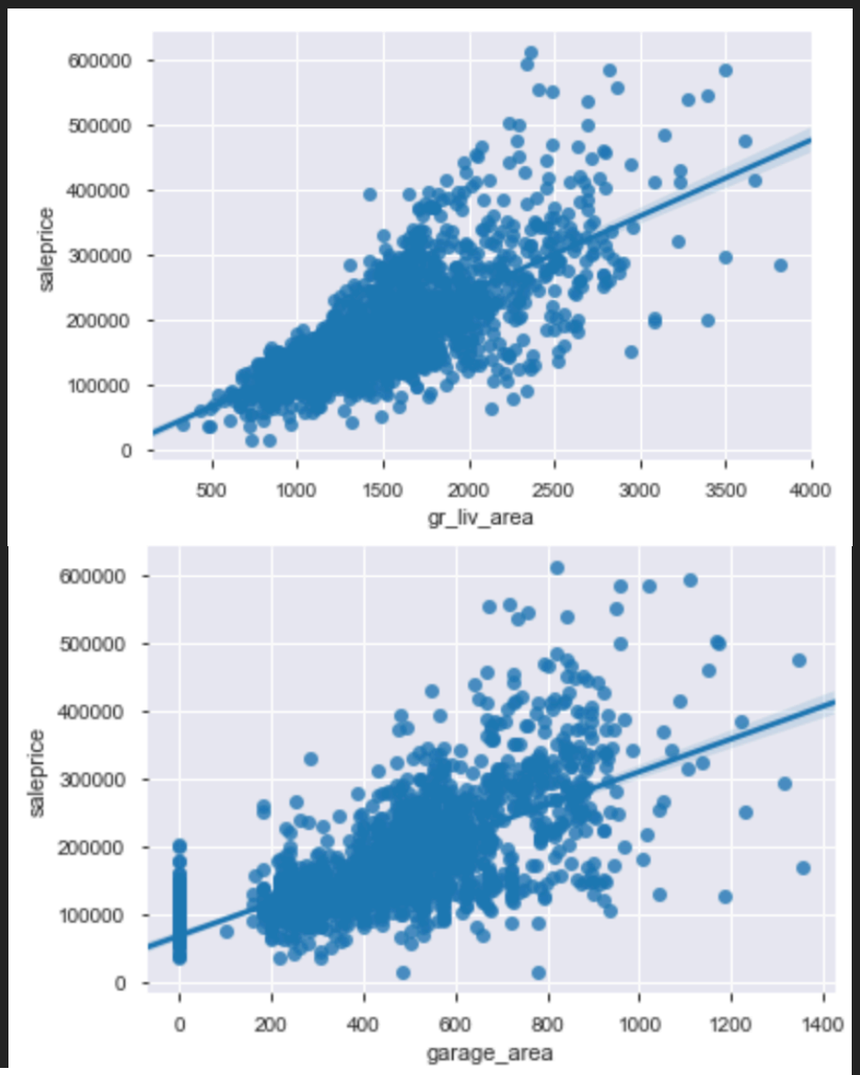


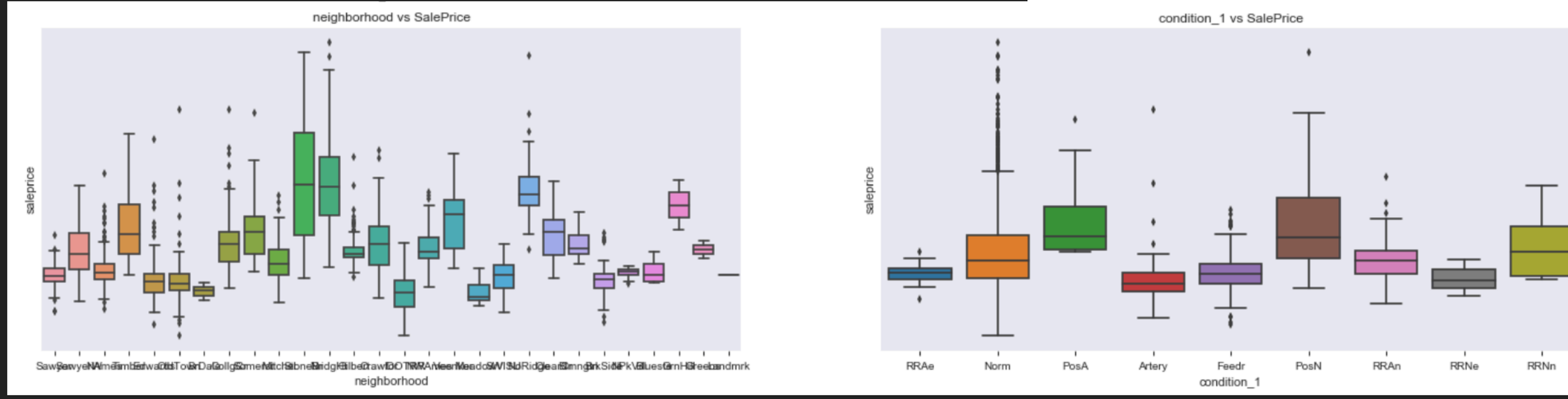
Feature	Type	Ranking Representation
lot_shape	Object	{'IR3':0,'IR2':1,'IR1':2,'Reg':3}
overall_qual	Object	{'1':0,'2':1,'3':2,'4':3,"5":4,"6":5,'7':6,'8':7,'9':8,'10':9}
overall_cond	Object	{'1':0,'2':1,'3':2,'4':3,"5":4,"6":5,'7':6,'8':7,'9':8}
exter_qual	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
exter_cond	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
bsmt_qual	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
bsmt_cond	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
bsmt_exposure	Object	{'NA':0,'No':1,'Mn':2,'Av':3,'Gd':4}
bsmtfin_type_1	Object	{'NA':0,"Unf":1,'LwQ':2,'Rec':3,'BLQ':4,"ALQ":5,'GLQ':6}
bsmtfin_type_2	Object	{'NA':0,"Unf":1,'LwQ':2,'Rec':3,'BLQ':4,"ALQ":5,'GLQ':6}
functional	Object	{'Sal':0,'Sev':1,'Maj2':2,'Maj1':3,'Mod':4,'Min2':5,'Min1':6,'Typ':7}
garage_finish	Object	{'NA':0, 'Unf':1, 'RFn':2,'Fin':3}
garage_qual	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
garage_cond	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
paved_drive	Object	{'N':0,'P':1,'Y':2}
heating_qc	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
kitchen_qual	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}
fireplace_qu	Object	{'NA':0,"Po":1, 'Fa':2, 'TA': 3,'Gd': 4,"Ex":5}



EXPLORING THE DATA

- Looked at correlation between continuous variables against Sale Price
- Boxplot to compare nominal against Sale Price





CREATING THE MODEL

- Tested 3 different models
- Goal: Get the model with the lowest RMSE
- ↓ RMSE = ↑ Model Fit = ↑ Predicting Capabilities

- Model 1: RMSE = 31730.63840
- Model 2: RMSE = 28669.97609 ← Best Model
- Model 3: RMSE = 33611.97639

CREATING THE MODEL

- Combine the separated data frames
- Run Lasso Regression on the whole data (Eliminate useless variables)
- Picked 26 variables with the highest coefficient to Sales Price
- Checked for Multicollinearity (Threshold corr > 0.7)

```
multicollinearity(X_2)
```

```
overall_qual and exter_qual corr = 0.7383695193658895
exter_qual and overall_qual corr = 0.7383695193658895
exter_qual and kitchen_qual corr = 0.7290477397813664
kitchen_qual and exter_qual corr = 0.7290477397813664
```

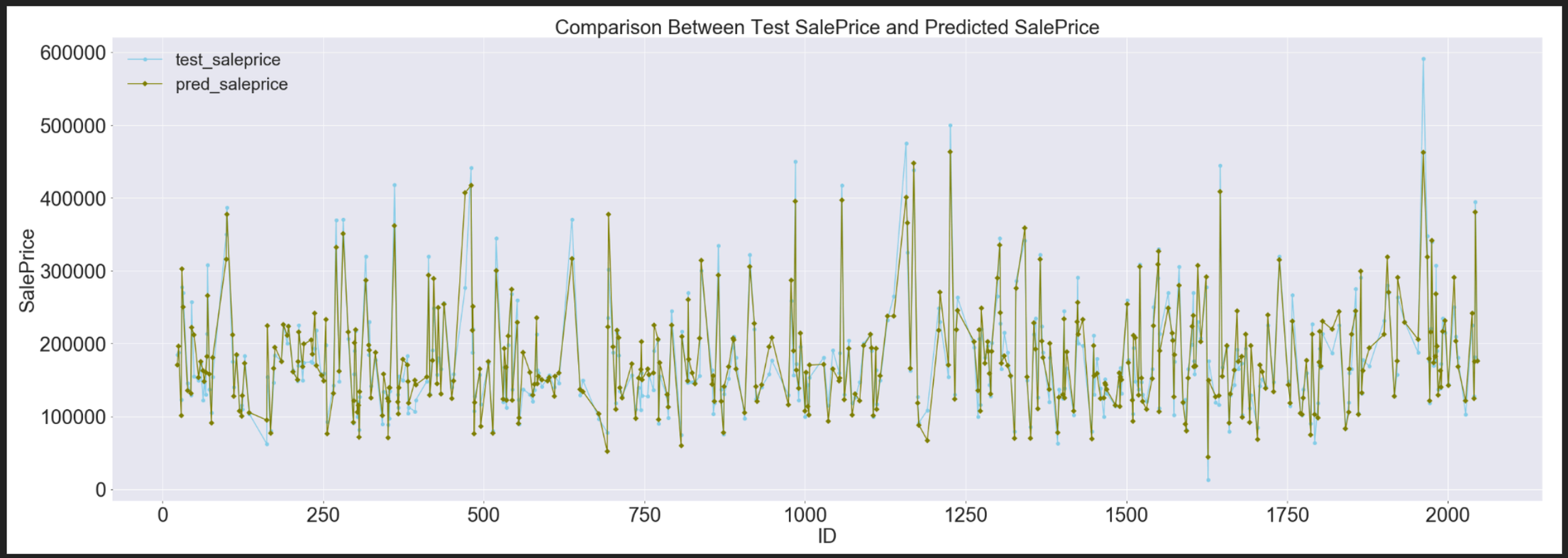
TEST THE MODEL

- Split the data into training and testing

```
X_train_2, X_test_2, y_train_2, y_test_2 = train_test_split(X_2, y_2, test_size=0.2, random_state = 42)
```

```
Lasso Regression Score :0.8987507849405587
Linear Regression Score :0.8706272821600907
Ridge Regression Score :0.898771692106255
```

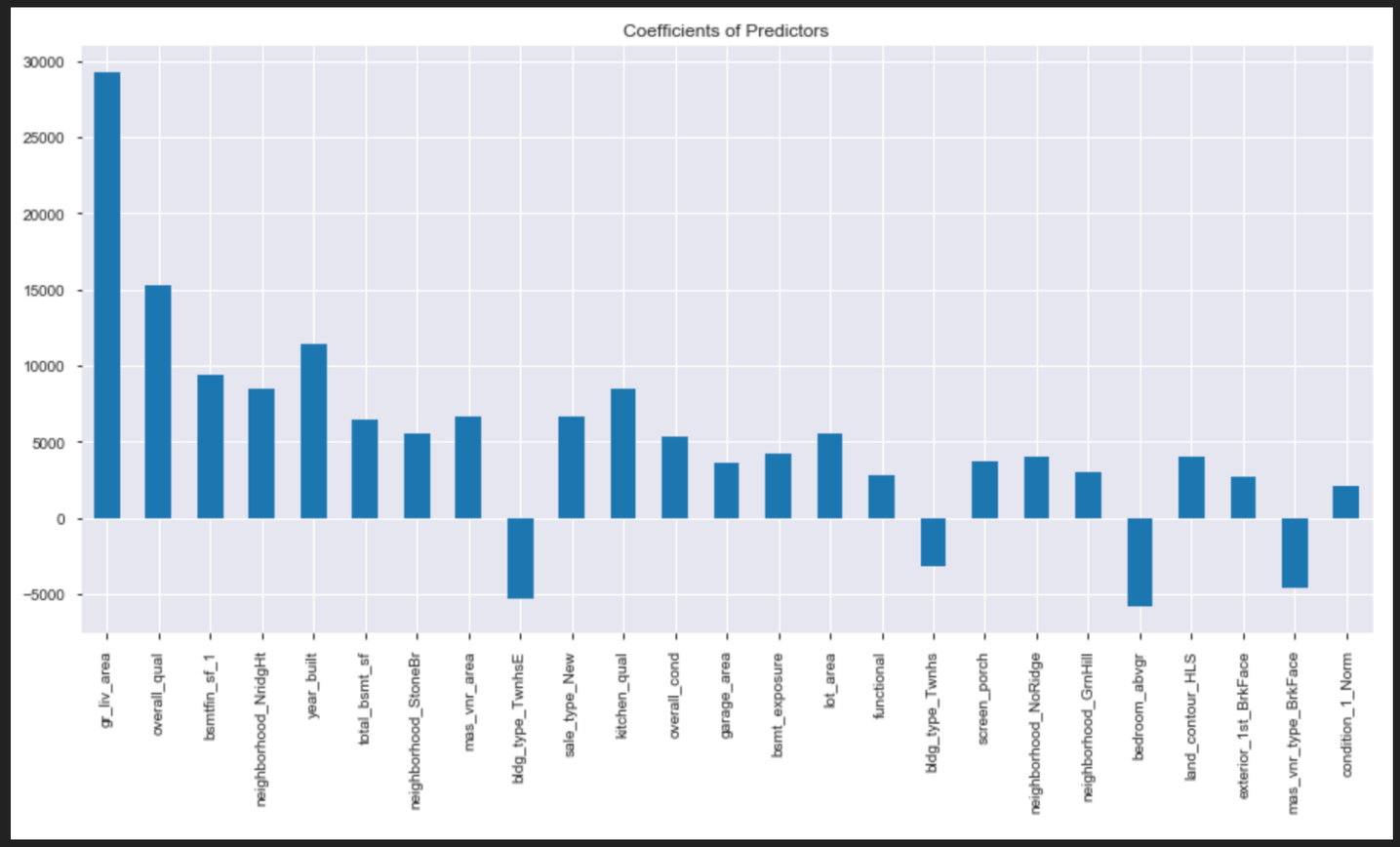
- Ridge Regression scored the best



RECOMMENDATIONS

- Top 3 features:
  General Living Area
  Overall Quality
  Year house was built

- Worst features:
  Number of Bedrooms
  EndUnits Twinhouse



- Controlable features:
  Overall Quality, Kitchen Quality, Overall Condition

- Best Locations:
  Northridge Heights, Stone Brook, Northridge and Green Hills