Ames Housing Price Prediction

More Power!

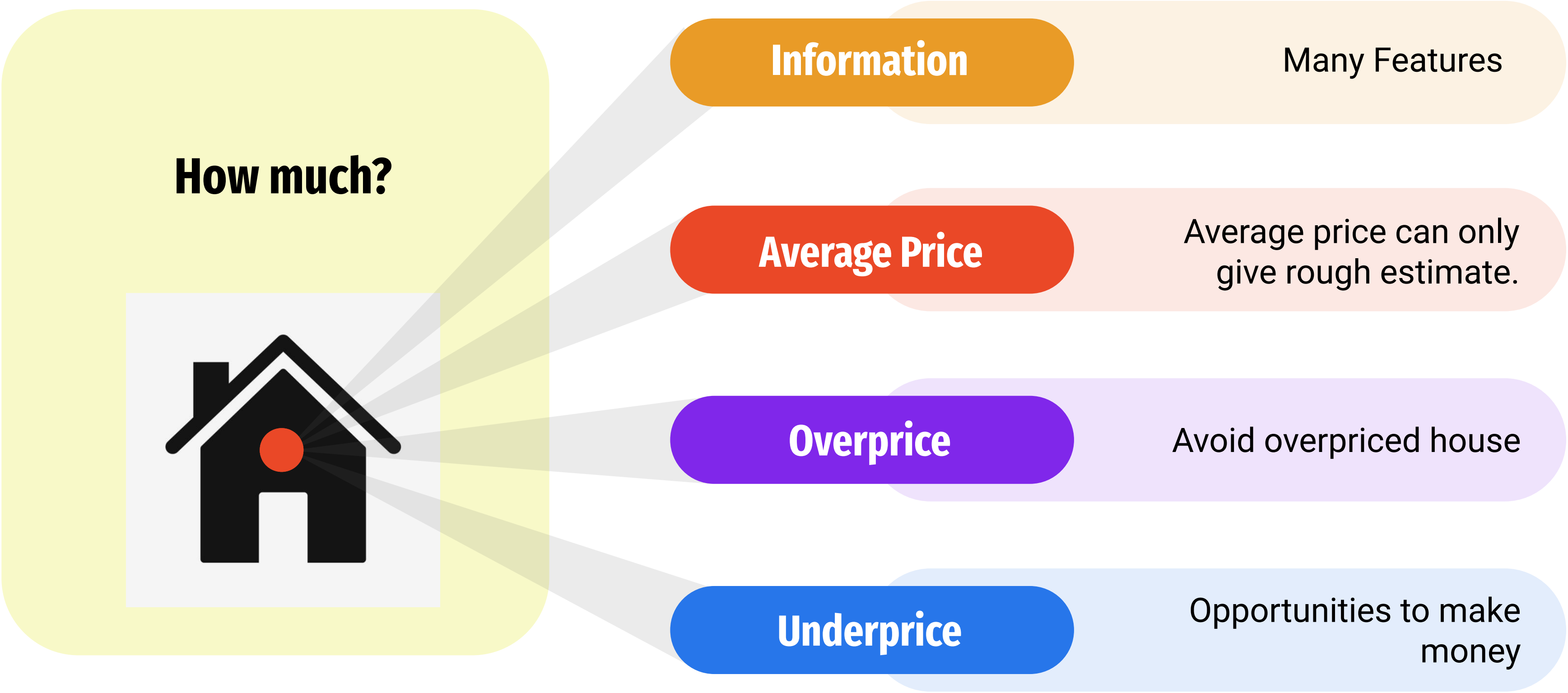
By Im Depends

Introduction

Chalermchon Wongsopa

Kantaphon Vareekasem

Problem Statement



How much?

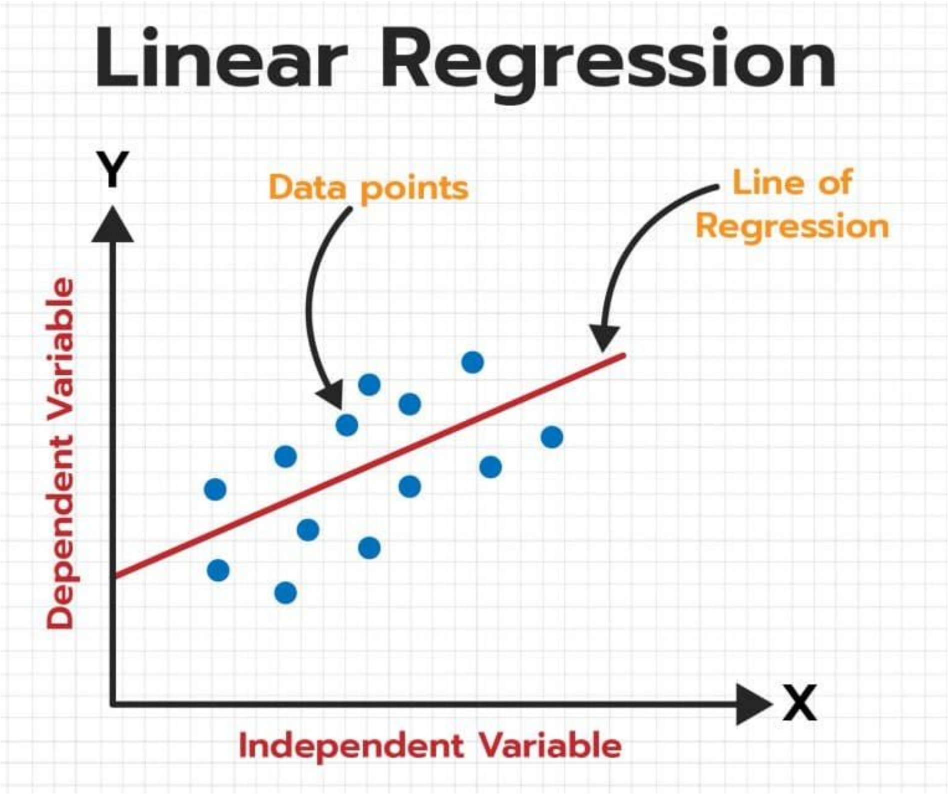
Information — Many Features

Average Price — Average price can only give rough estimate.

Overprice — Avoid overpriced house

Underprice — Opportunities to make money

Solution



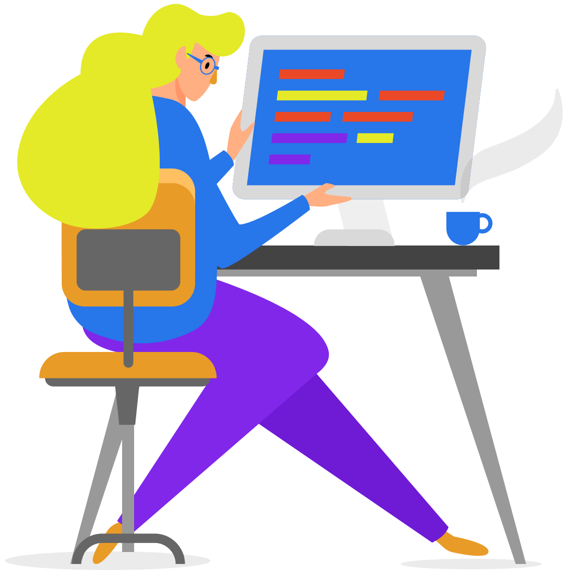
Best Model

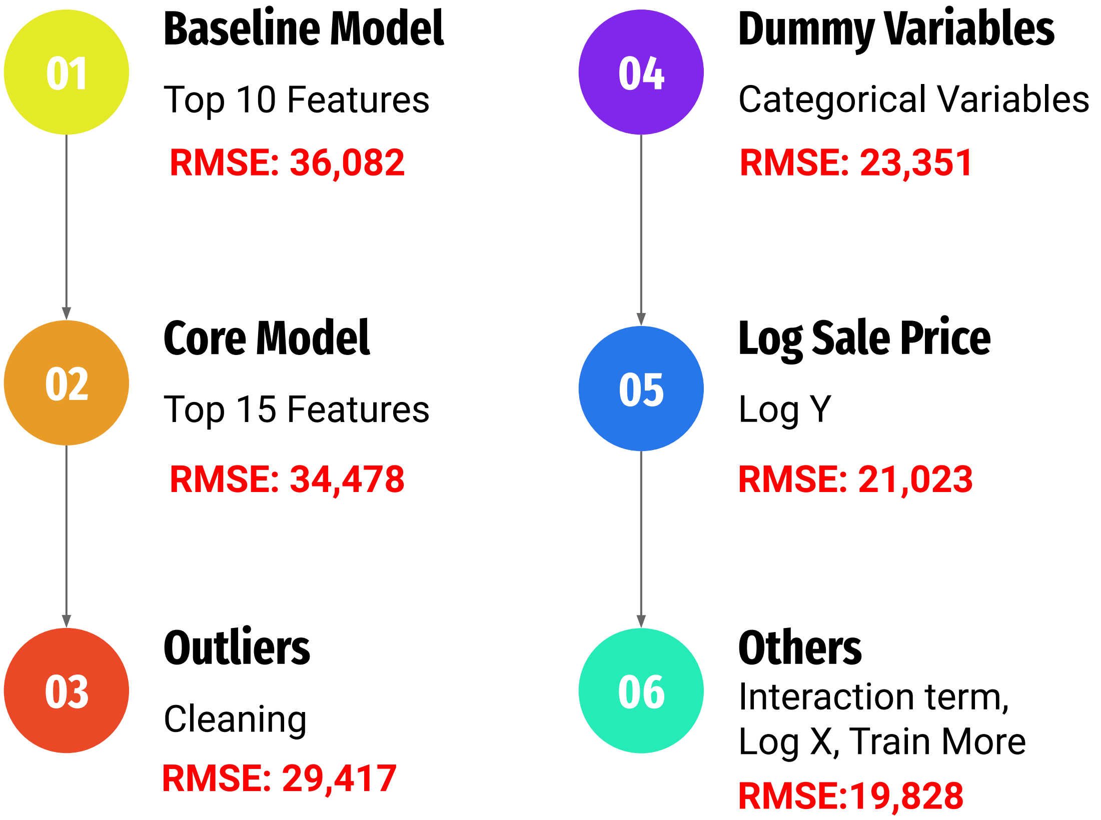
15 Numerical + 20 Dummies

RMSE: **19,828** USD

R Squared: **93%**

Models Improvement Process





01 **Baseline Model**

Top 10 Features

RMSE: 36,082

02 **Core Model**

Top 15 Features

RMSE: 34,478

03 **Outliers**

Cleaning

RMSE: 29,417

04 **Dummy Variables**

Categorical Variables

RMSE: 23,351

05 **Log Sale Price**

Log Y

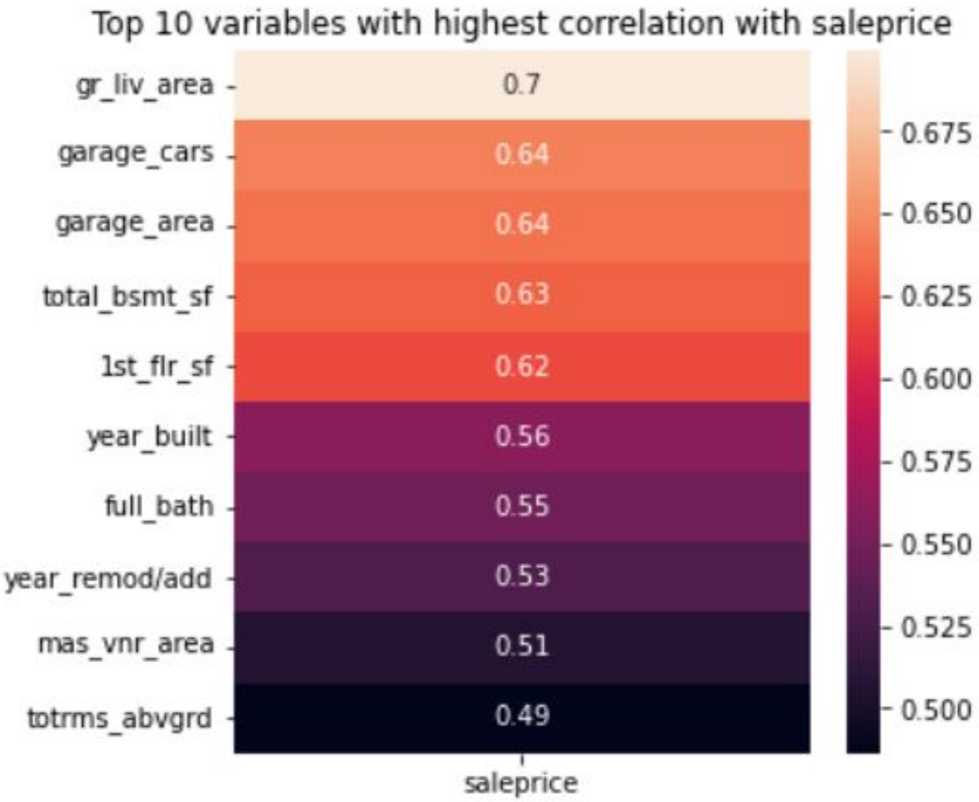
RMSE: 21,023

06 **Others**

Interaction term,
Log X, Train More

RMSE:19,828

Baseline Model

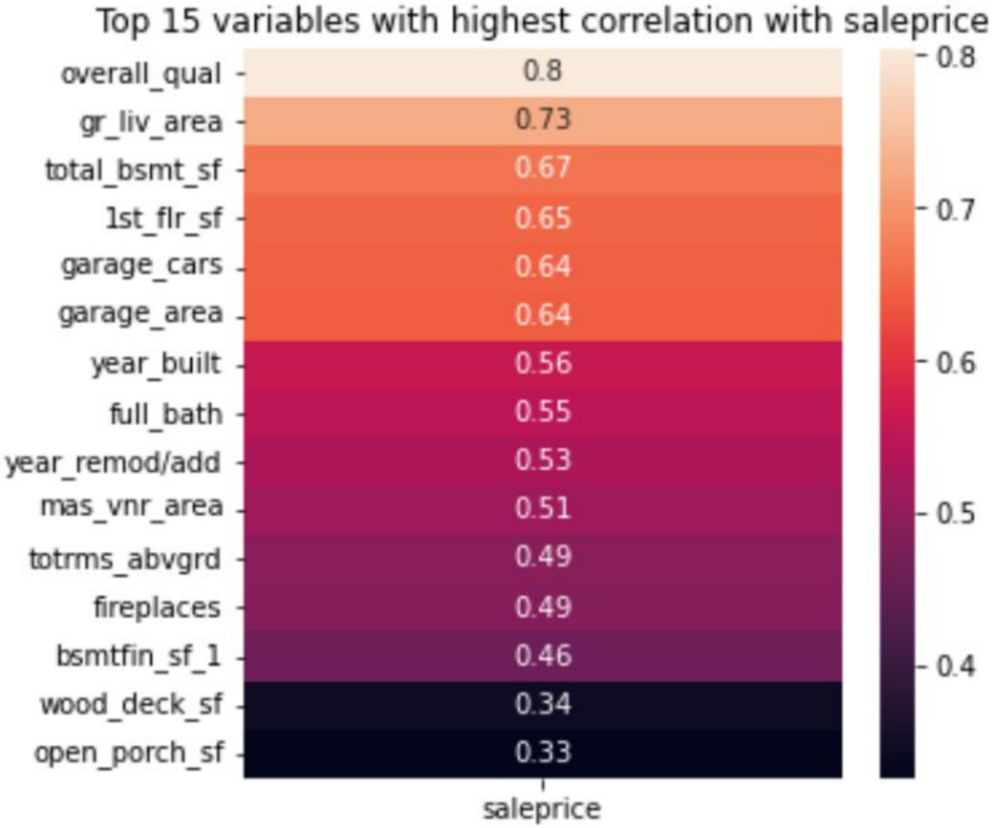


Linear Regression

- Top 10 variables
- Fill Missing Values with 0
- Target Variable: Sale Price

RMSE: 36,082

Core Model



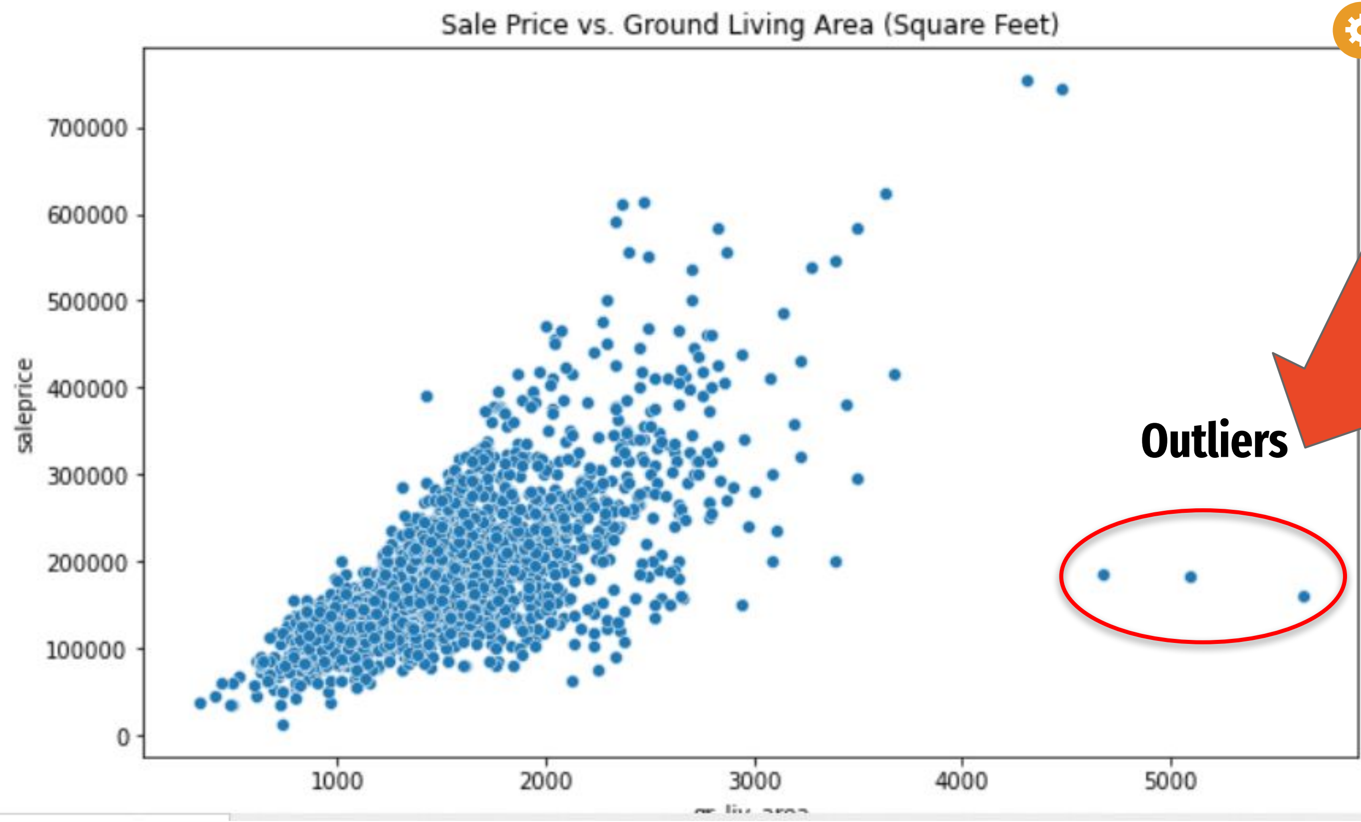
Linear Regression

- Top 15 variables
- Adjust number of top correlated variables using RMSE

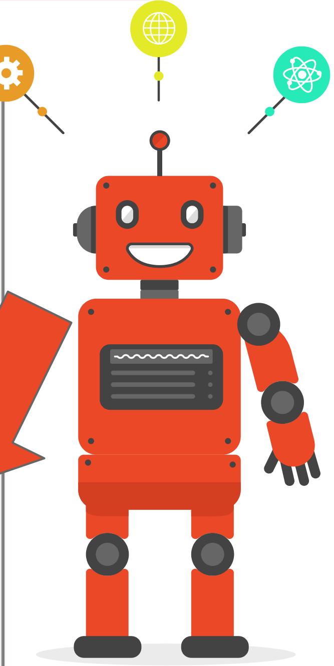
RMSE: 34,478

Outliers

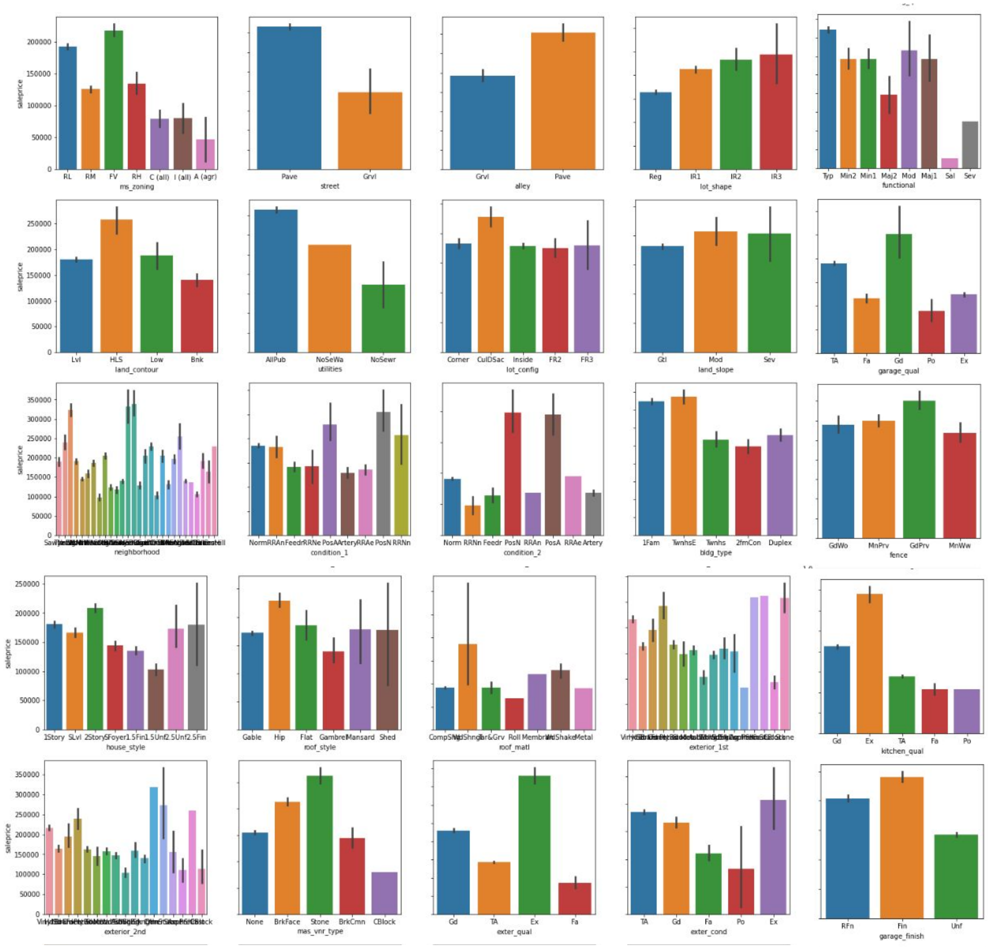
RMSE: 29,417



Outliers



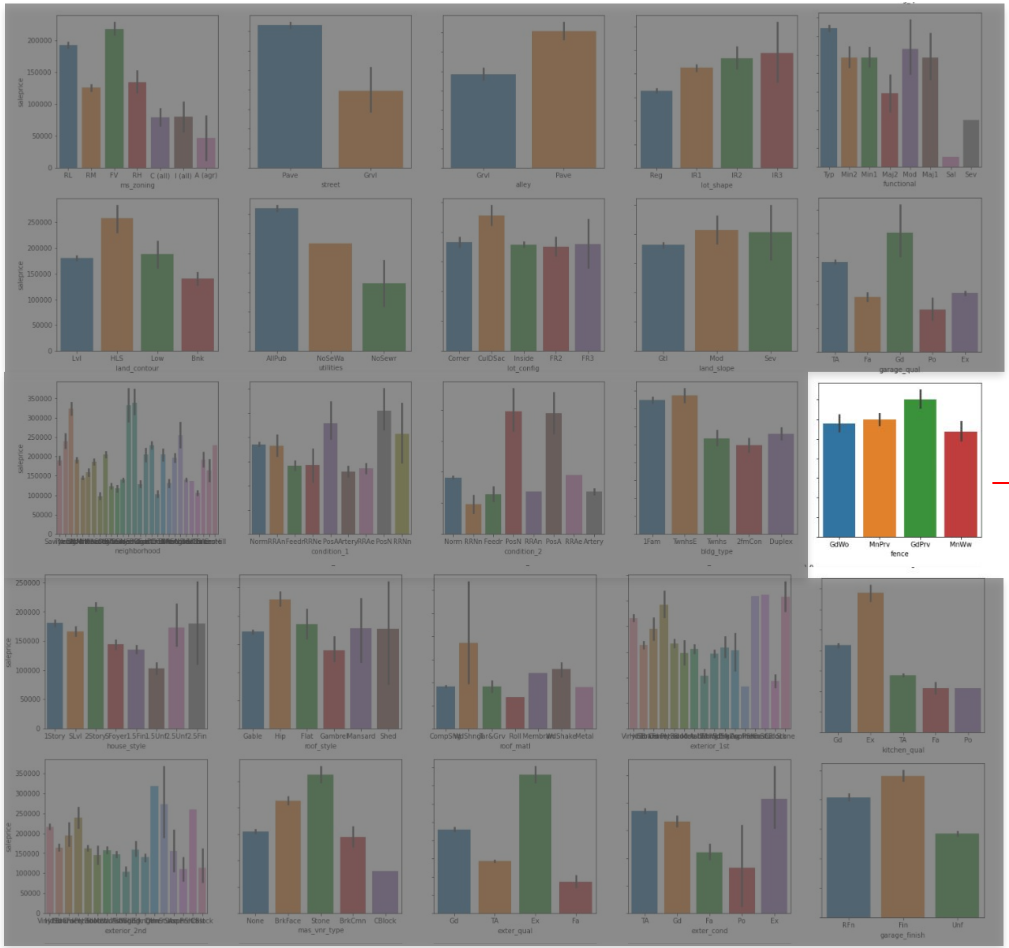
Categorical Variables

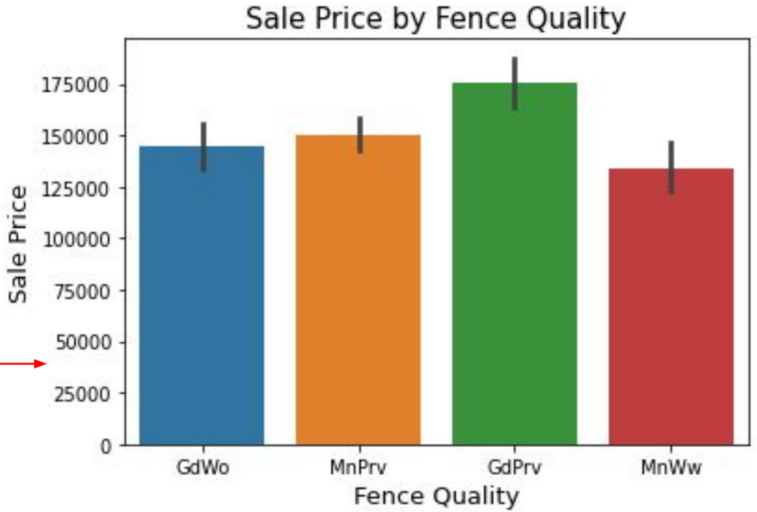


- 43 Variables are available for the model
- We distinguished ‘Signal’ from ‘Noise’ using 2 Indicators
  - Average sale price
  - Number of observations

RMSE: 23,351

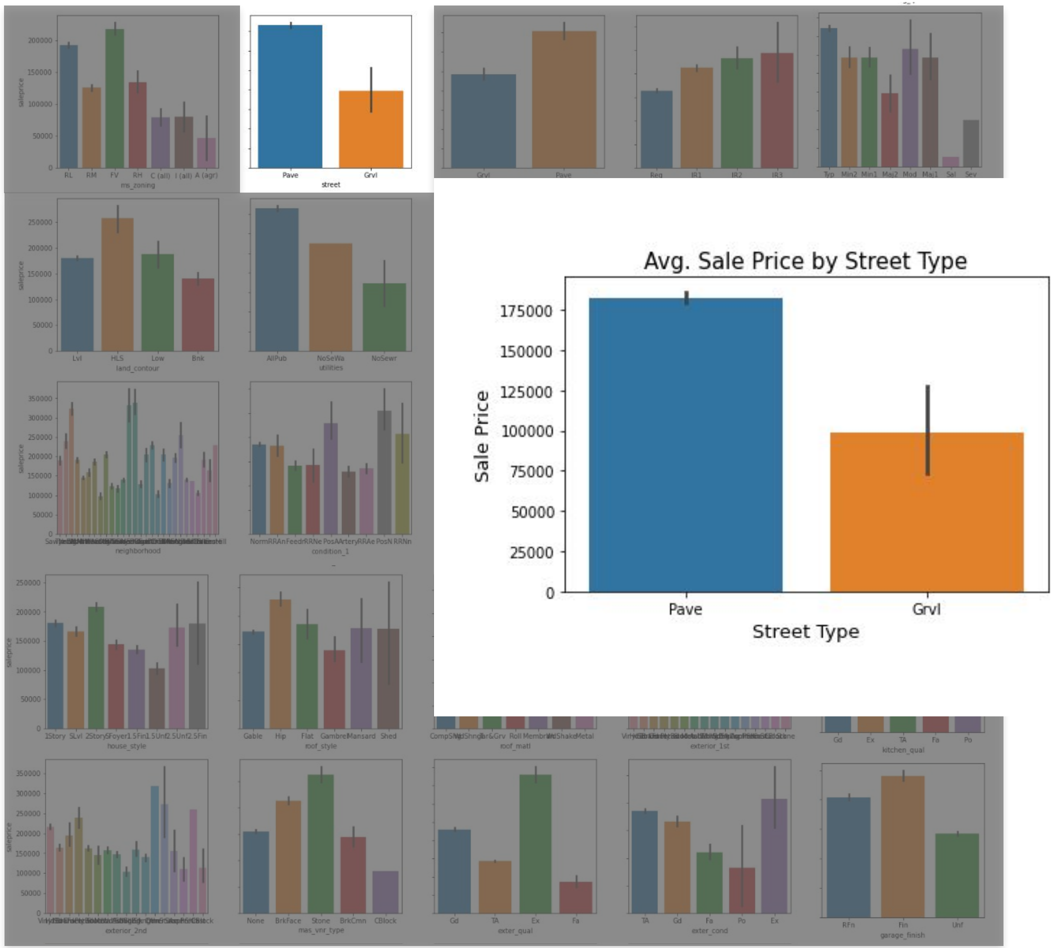
Excluded Variables - Type I

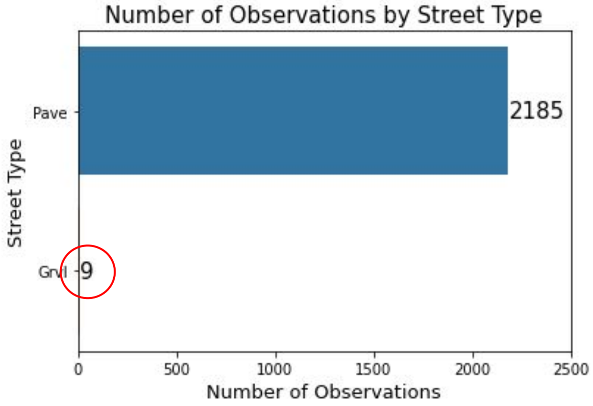




Indistinguishable “Sale Price”
for each “Fence Quality”

Excluded Variables - Type II





Insufficient data for **gravel street type** to train the model

Selected Categorical Variables

Group I

01

Applied “get_dummies” function

1. Locations within Ames city
2. Type of dwelling
3. Exterior quality
4. Condition of sale
5. Fireplace quality
6. Flatness of the property
7. Home functionality
8. General shape of property
9. Paved driveway
10. Central air conditioning

Group II

02

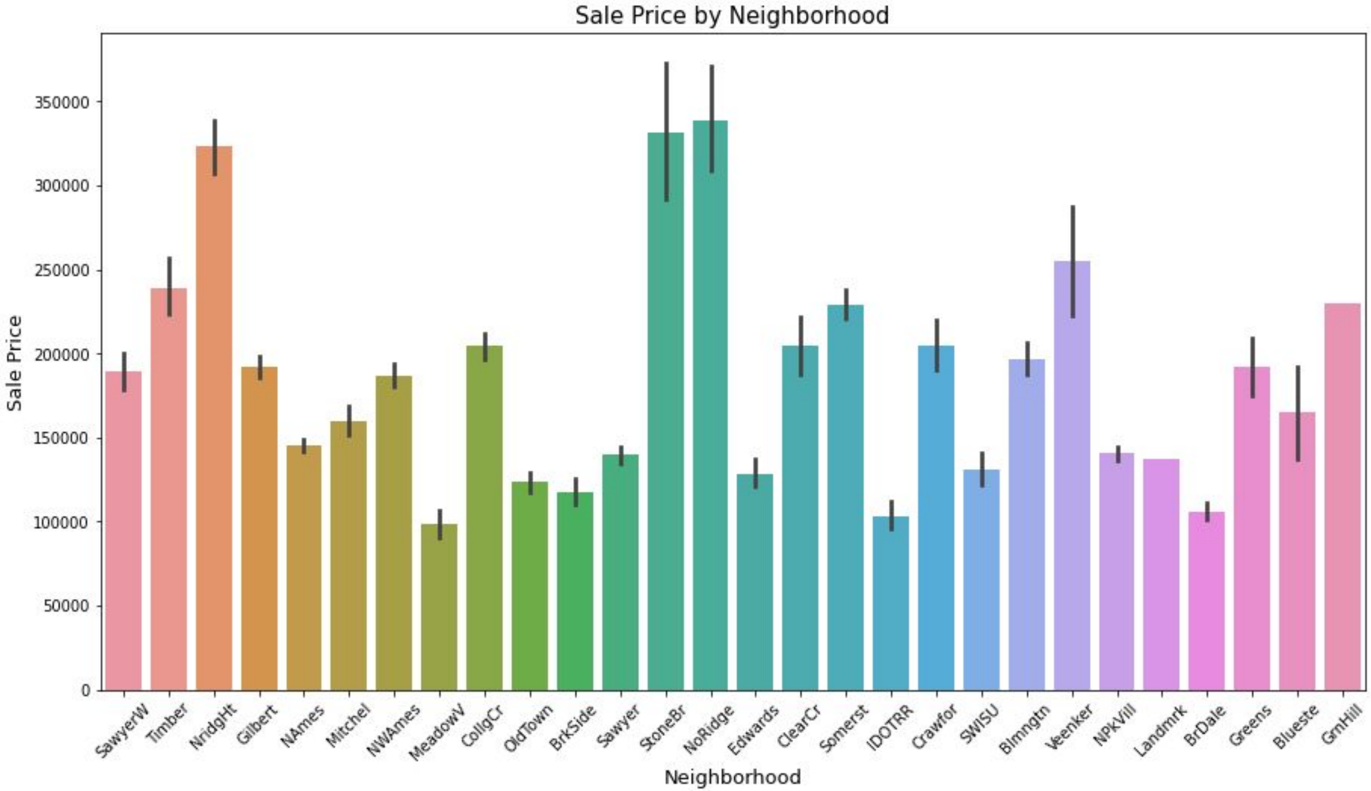
Grouped to binary form

1. Kitchen quality
2. Height of the basement
3. Type of sale
4. Exterior covering on house
5. Heating quality
6. Zoning classification
7. Proximity to various conditions
8. Garage condition
9. Garage quality
10. Miscellaneous feature

Group I - Applied “get_dummies” function

👍 1st Most Impact - Neighborhood

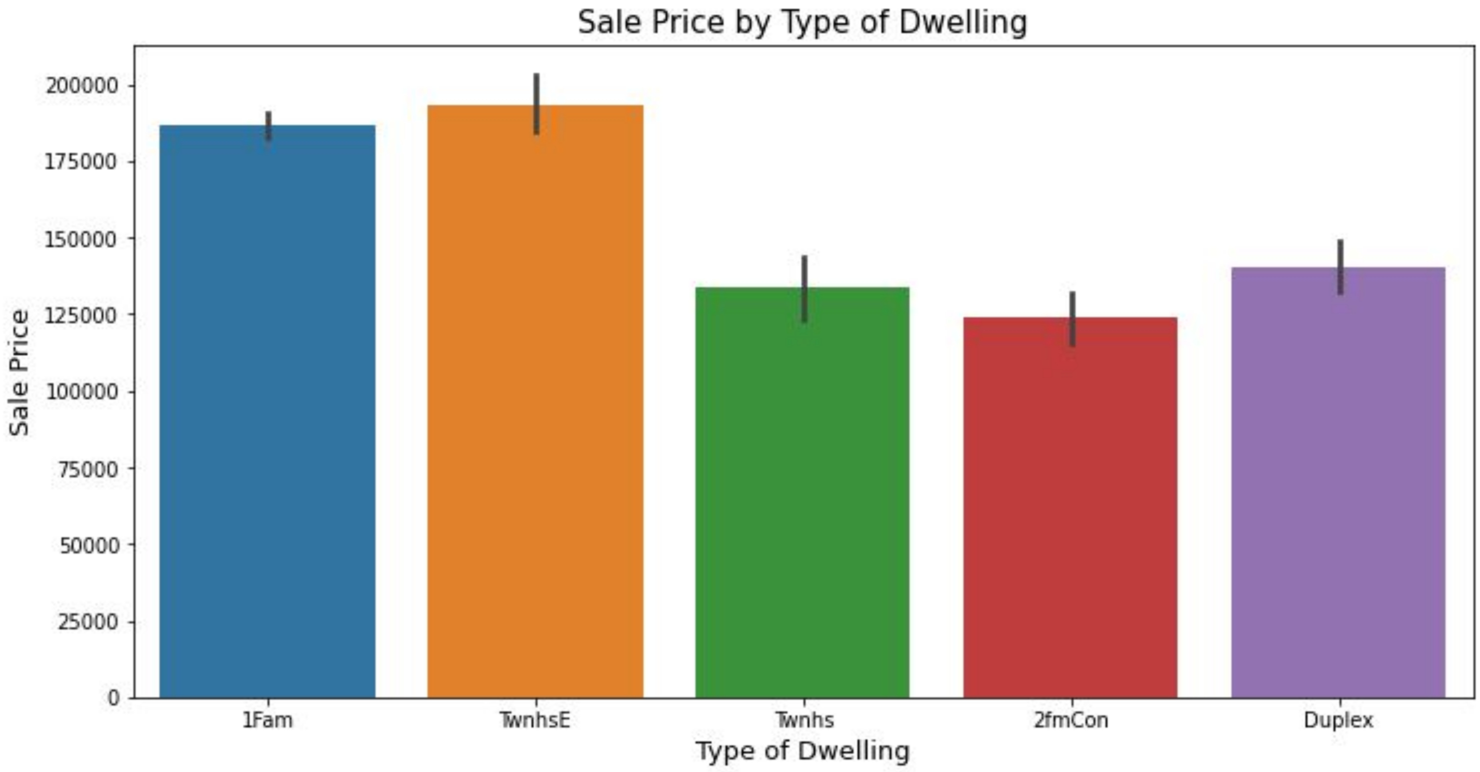
Sale Price by Neighborhood



Around **1,400 USD of RMSE was decreased** after including the variable to the model

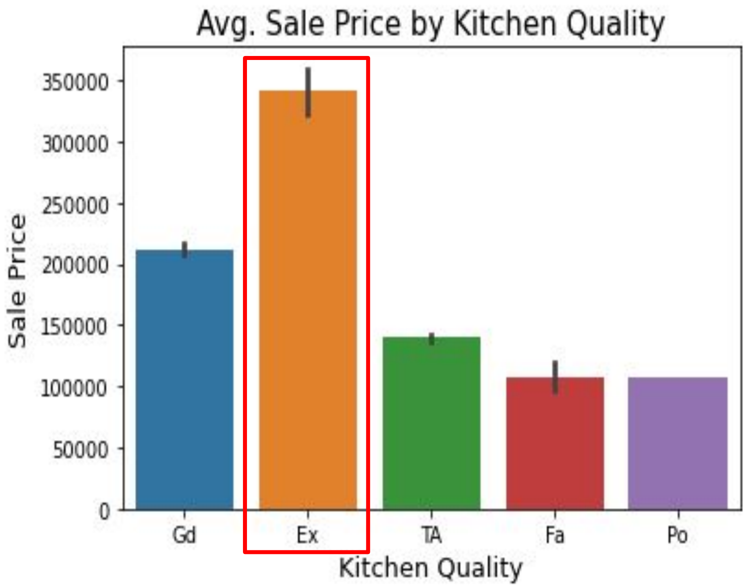
Group I - Applied "get_dummies" function

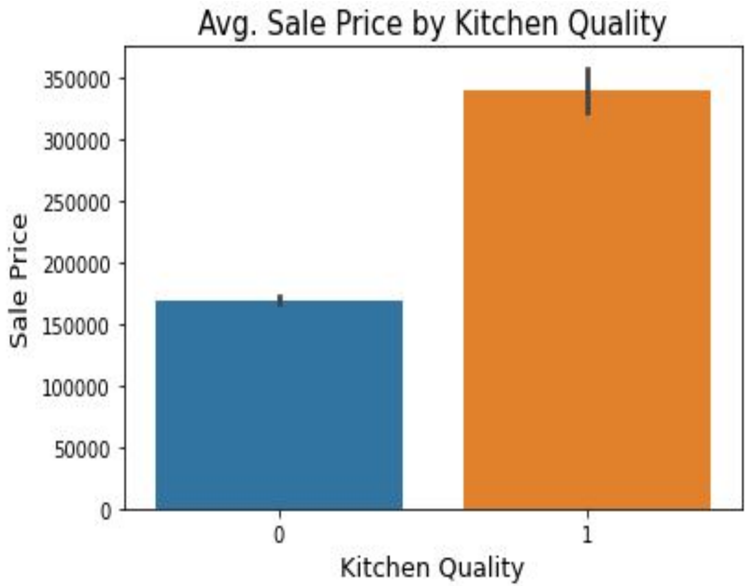
👍 2nd Most Impact - Type of Dwelling



Group II - Grouped to binary form

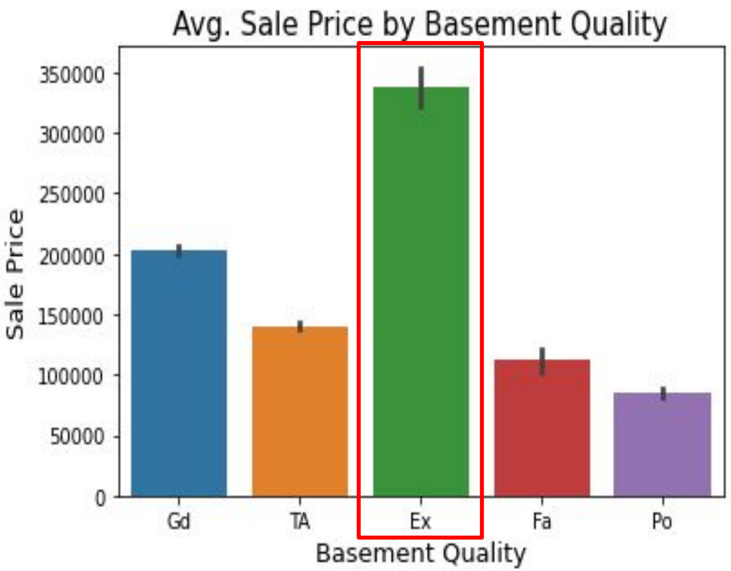
1st Most Impact - Kitchen Quality

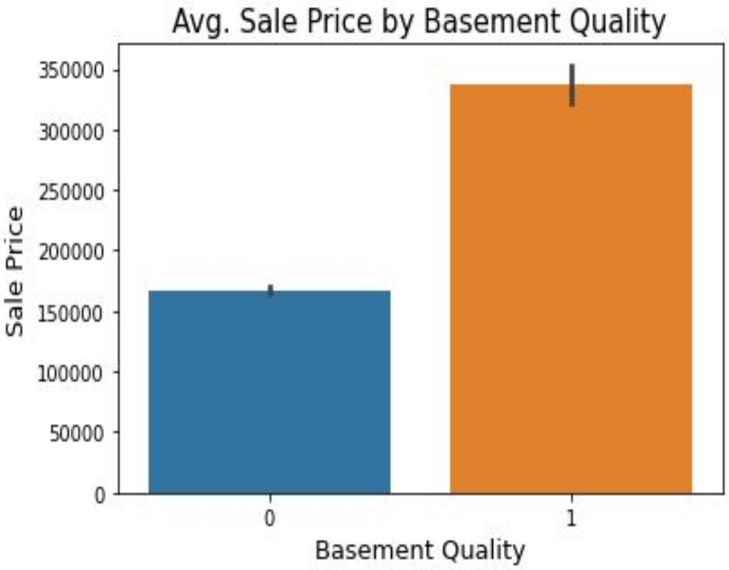




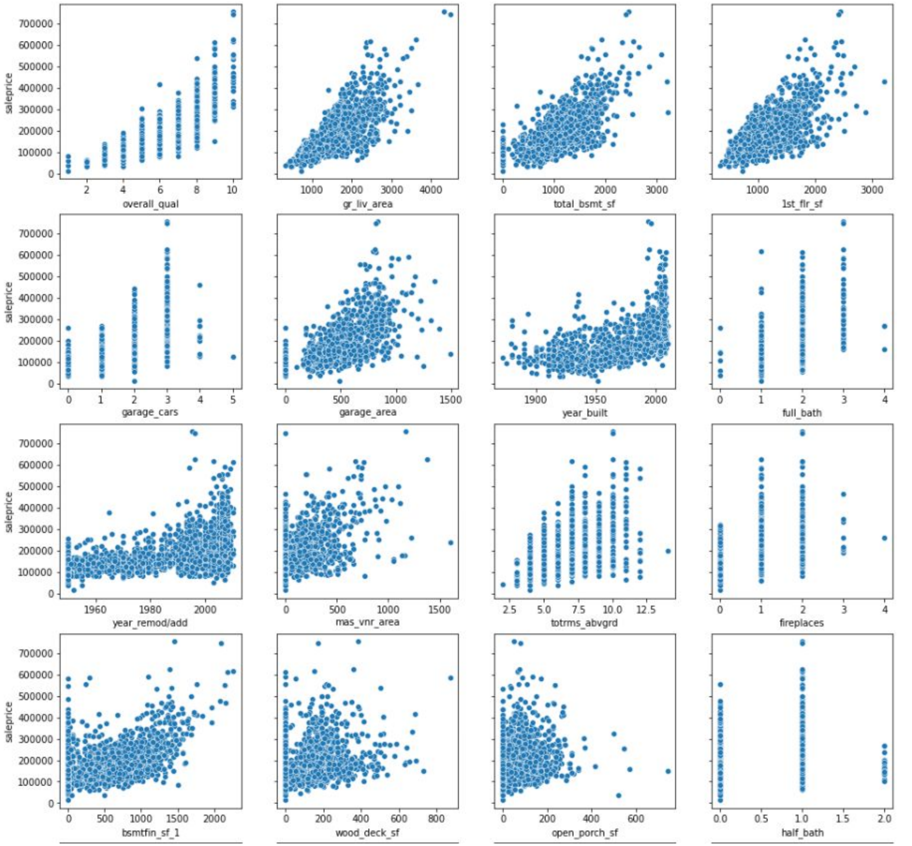
Group II - Grouped to binary form

2nd Most Impact - Basement Quality





Scatter Plot



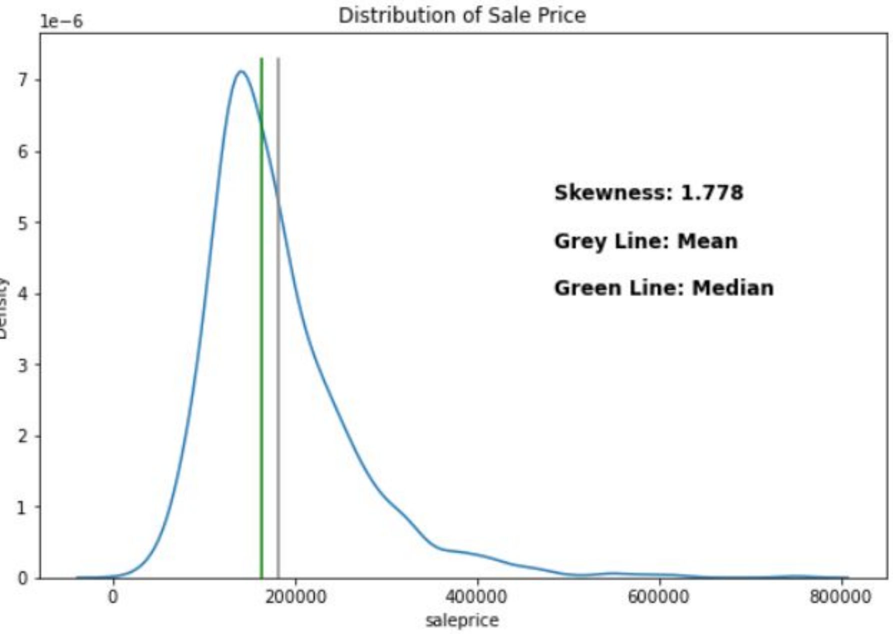
Scatter Plot

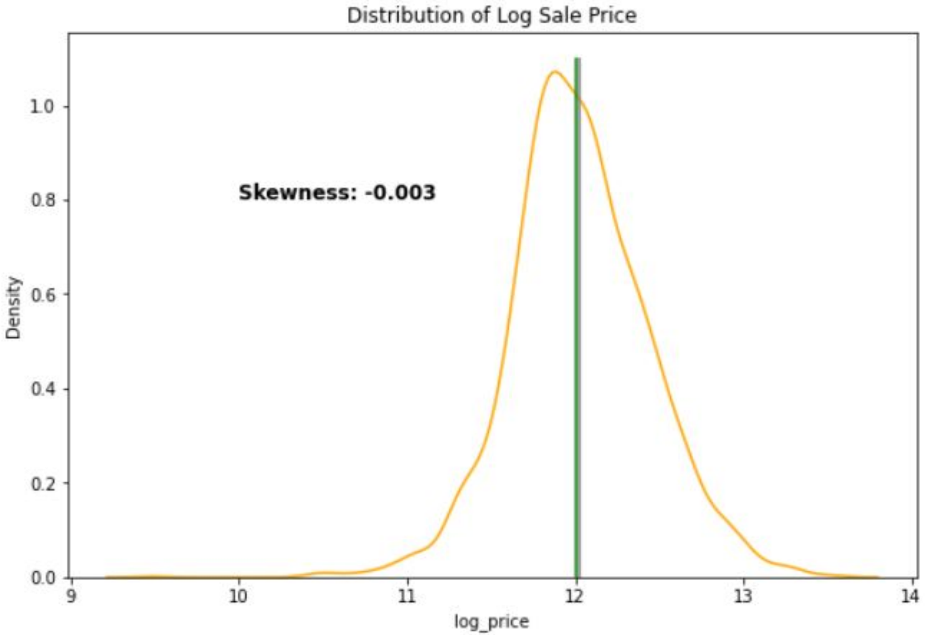
- Check LINE ASSUMPTION
- Linearity
- If not linear, we can transform
- Drop variables: Open-Porch, Total Room

05

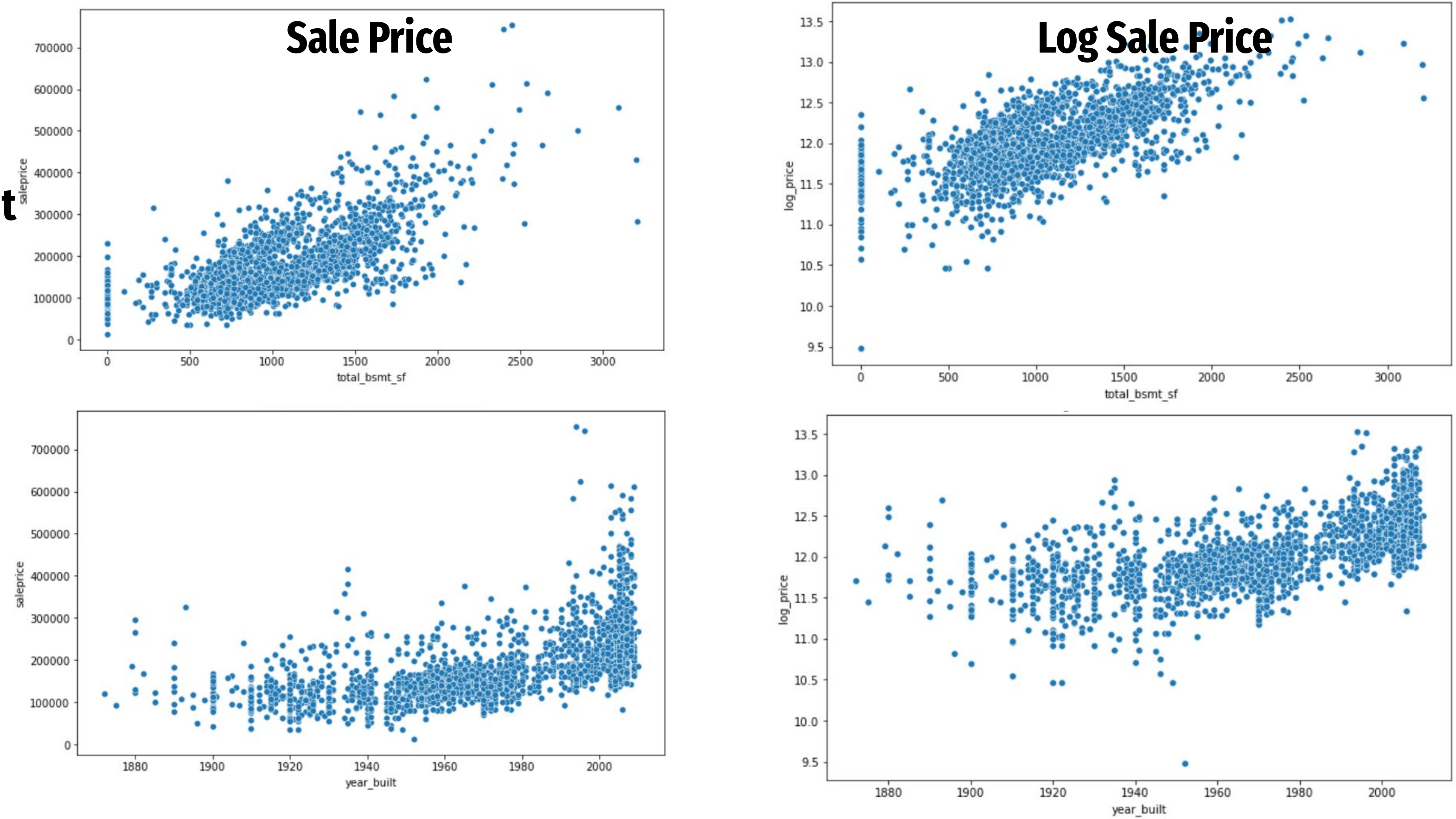
Log Sale Price (Target)

RMSE: 21,023





Sale Price VS Log Sale Price



Sale Price

Log Sale Price

Total Basement Area

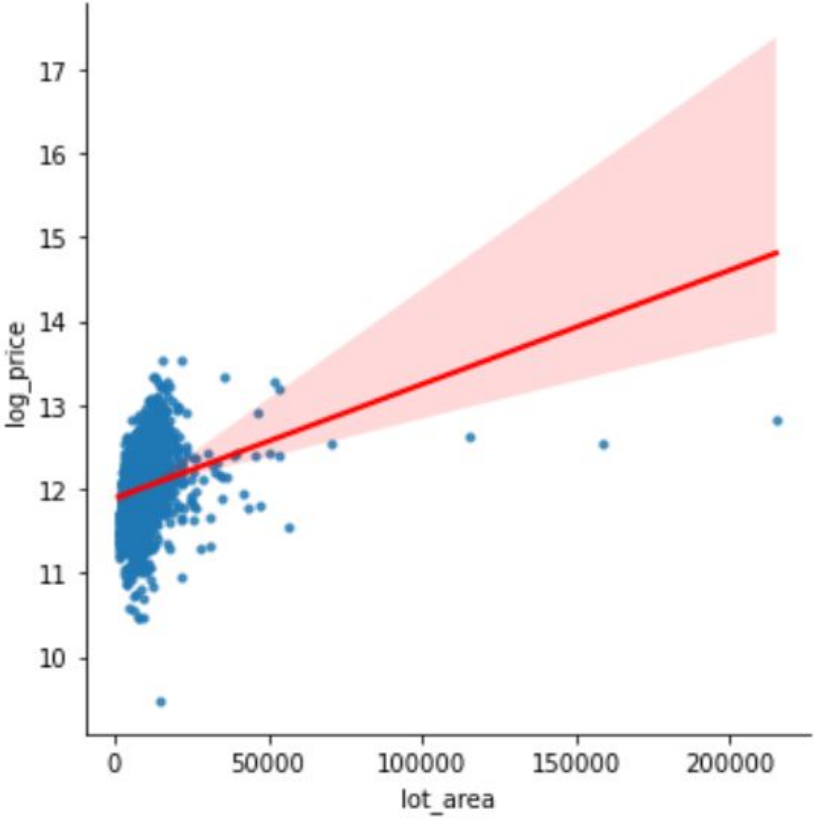
Year Built

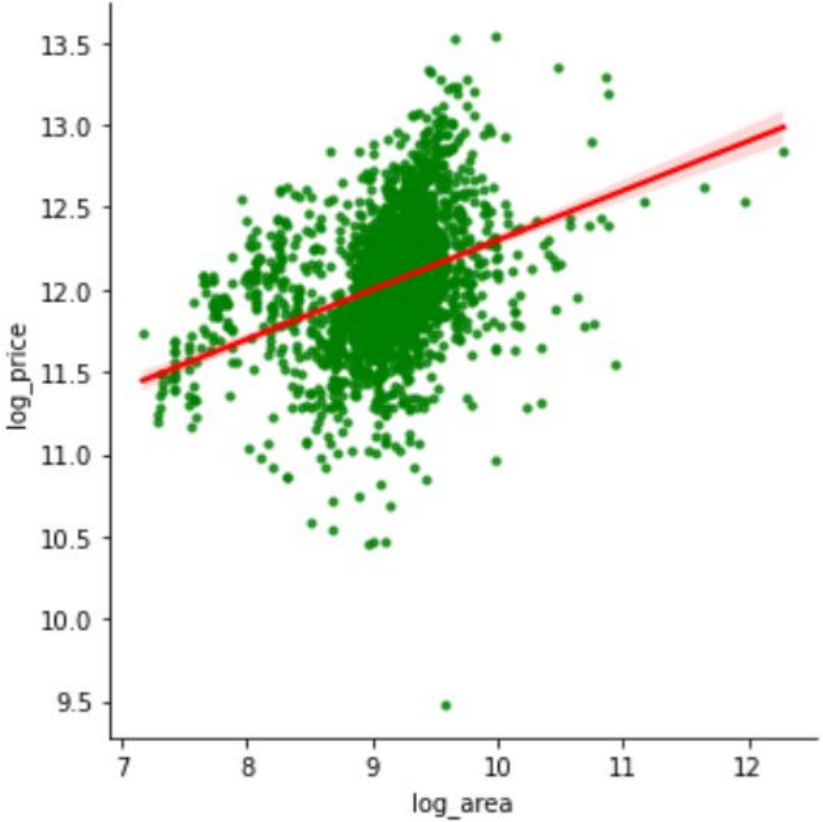
Interaction Terms

Every Good House Need a Good Fireplace!

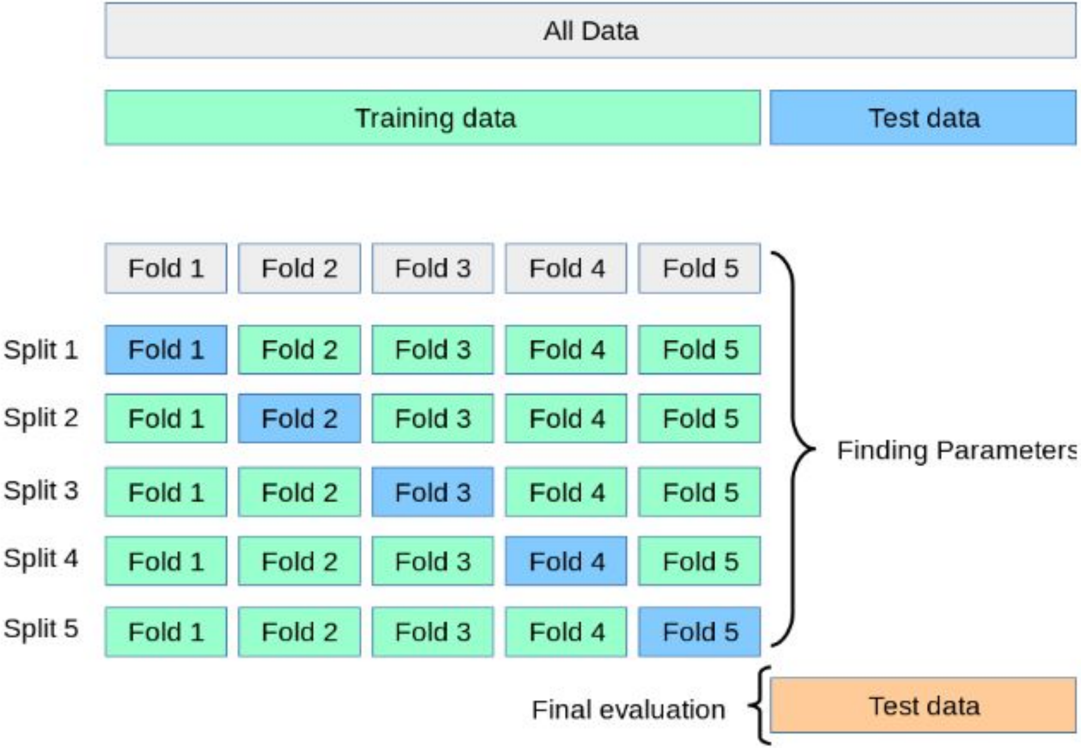
Interaction Terms

Log feature


17
16
15
14
13
12
11
10
log_price
0
50000
100000
150000
200000
lot_area


13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
log_price
7
8
9
10
11
12
log_area

More Power!



All Data

Training data | Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding Parameters

Final evaluation | Test data

Use all data to training

RMSE: 23,351

Summary

Top 15 features

Correlation with sale price

01

Data Cleaning

Remove outliers

02

Pattern identification

Group Dummy Variables

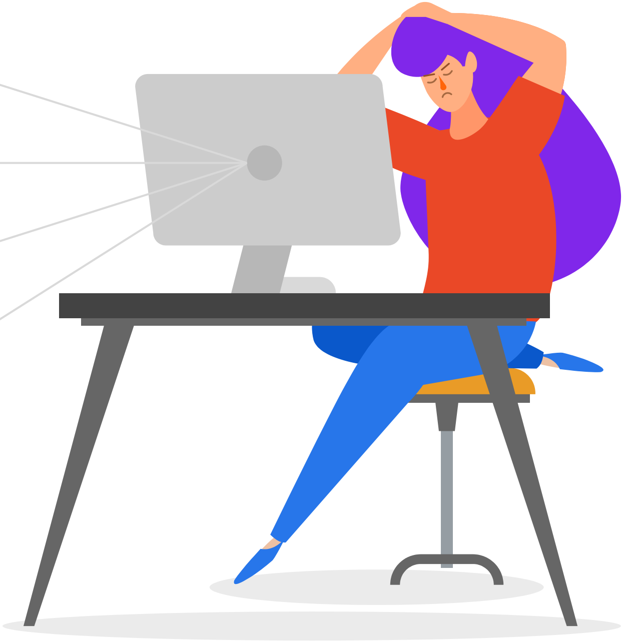
03

Relationship between X&Y

Log Transformation

04

Thought Process



THANK YOU!

Victory
is sweet!