Bangkok Housing Price Prediction

What is the price of my property?

Problem Statement

Property Type

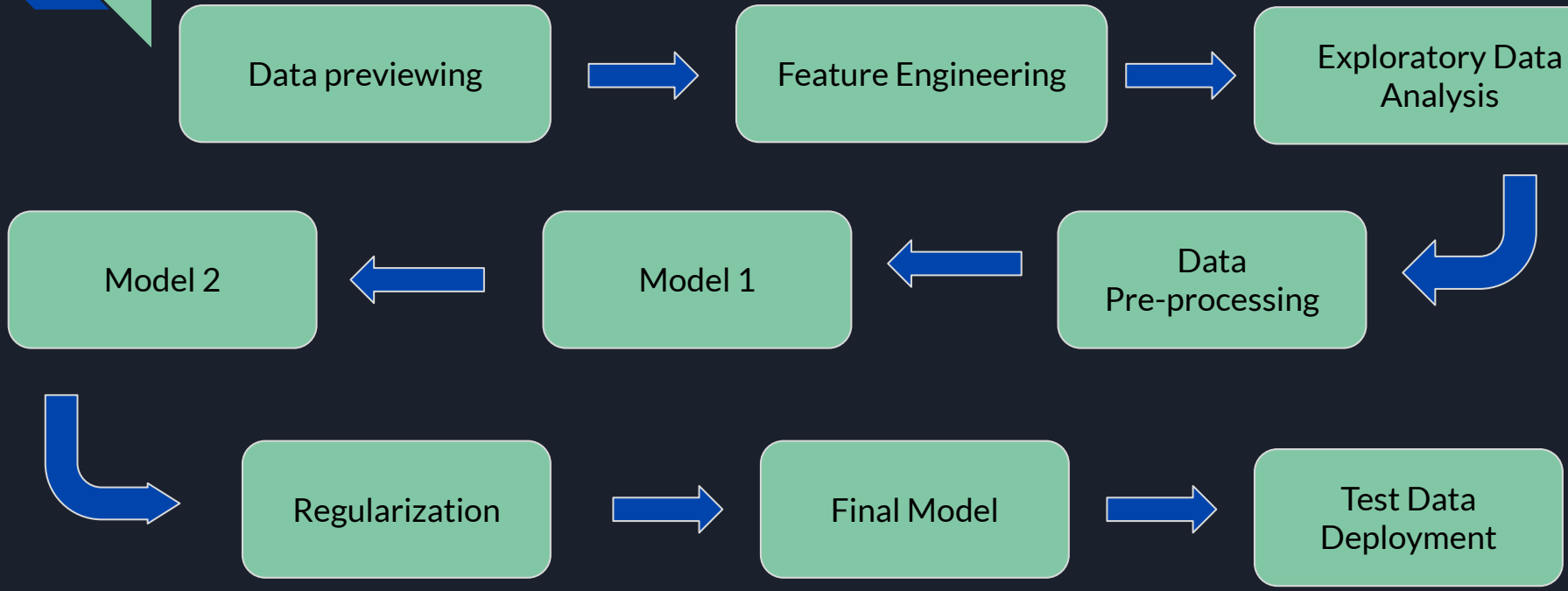
Facilities

Housing plan

Transportation Proximity

Shop / Market Conveniencies

Model Building Journey



Data Previewing

- Multiple missing values
- 8 object data columns

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        14271 non-null  int64
 1   province                  14271 non-null  object
 2   district                  14271 non-null  object
 3   subdistrict               14260 non-null  object
 4   address                   14271 non-null  object
 5   property_type             14271 non-null  object
 6   total_units               10509 non-null  float64
 7   bedrooms                  14228 non-null  float64
 8   baths                     14236 non-null  float64
 9   floor_area                14271 non-null  int64
 10  floor_level               8093 non-null   float64
 11  land_area                 4917 non-null   float64
 12  latitude                  14271 non-null  float64
 13  longitude                 14271 non-null  float64
 14  nearby_stations           14271 non-null  int64
 15  nearby_station_distance   7228 non-null   object
 16  nearby_bus_stops          6009 non-null   float64
 17  nearby_supermarkets       13885 non-null  float64
 18  nearby_shops              14271 non-null  int64
 19  year_built                14271 non-null  int64
 20  month_built               8397 non-null   object
 21  facilities                14271 non-null  object
 22  price                     14271 non-null  int64
dtypes: float64(9), int64(6), object(8)
```

```
id                          0
province                    0
district                    0
subdistrict                11
address                     0
property_type               0
total_units              3762
bedrooms                   43
baths                      35
floor_area                  0
floor_level              6178
land_area                9354
latitude                    0
longitude                   0
nearby_stations             0
nearby_station_distance  7043
nearby_bus_stops         8262
nearby_supermarkets       386
nearby_shops                0
year_built                  0
month_built              5874
facilities                  0
price                       0
dtype: int64
```

Feature Engineering

- Target Encoding: District, Subdistrict, Province, Property_type columns
- Numeric Encoding Month_built column
- Count_facilities column: Count each facility
- Closest_station_name and Closest_station_distance

```
{'Nonthaburi': 0, 'Samut Prakan': 1, 'Bangkok': 2}
```

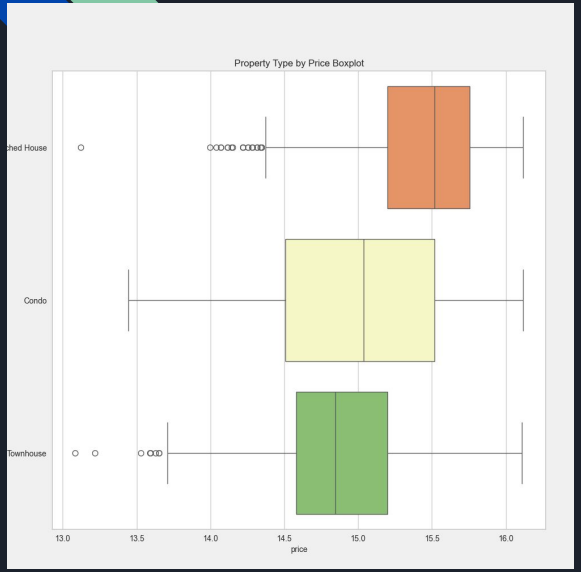
```
province
0    3.286570e+06
1    3.316038e+06
2    4.464201e+06
```

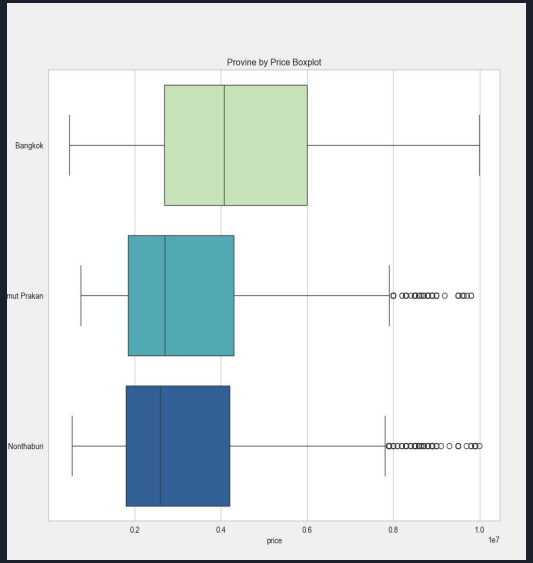
```
{'Townhouse': 0, 'Condo': 1, 'Detached House': 2}
```

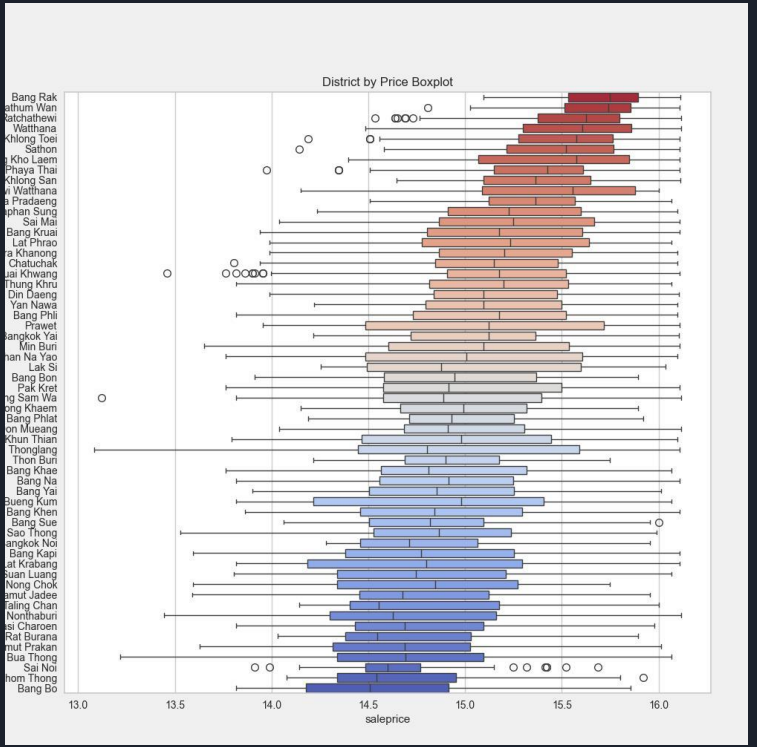
```
property_type
0    3.376431e+06
1    3.898478e+06
2    5.594899e+06
```

nearby_station_distance	closest_station_name	closest_station_distance
[[E7 Ekkamai BTS, 270], [E6 Thong Lo BTS, 800]]	E7 Ekkamai BTS	270
[[E7 Ekkamai BTS, 270], [E6 Thong Lo BTS, 800]]	E7 Ekkamai BTS	270
[[E9 On Nut BTS, 110]]	E9 On Nut BTS	110

Exploratory Data Analysis





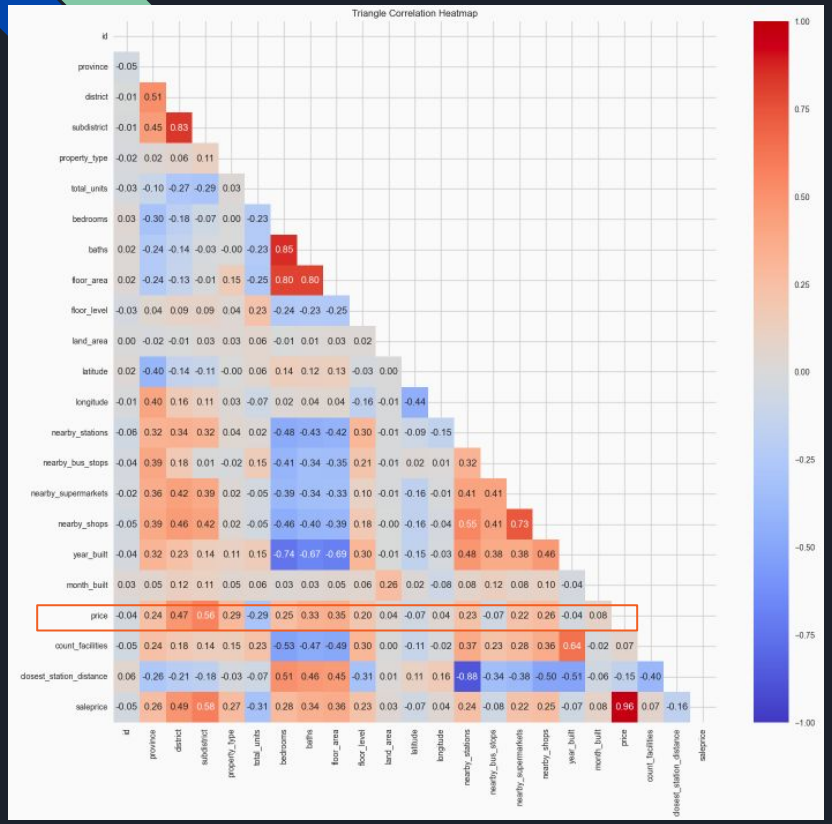


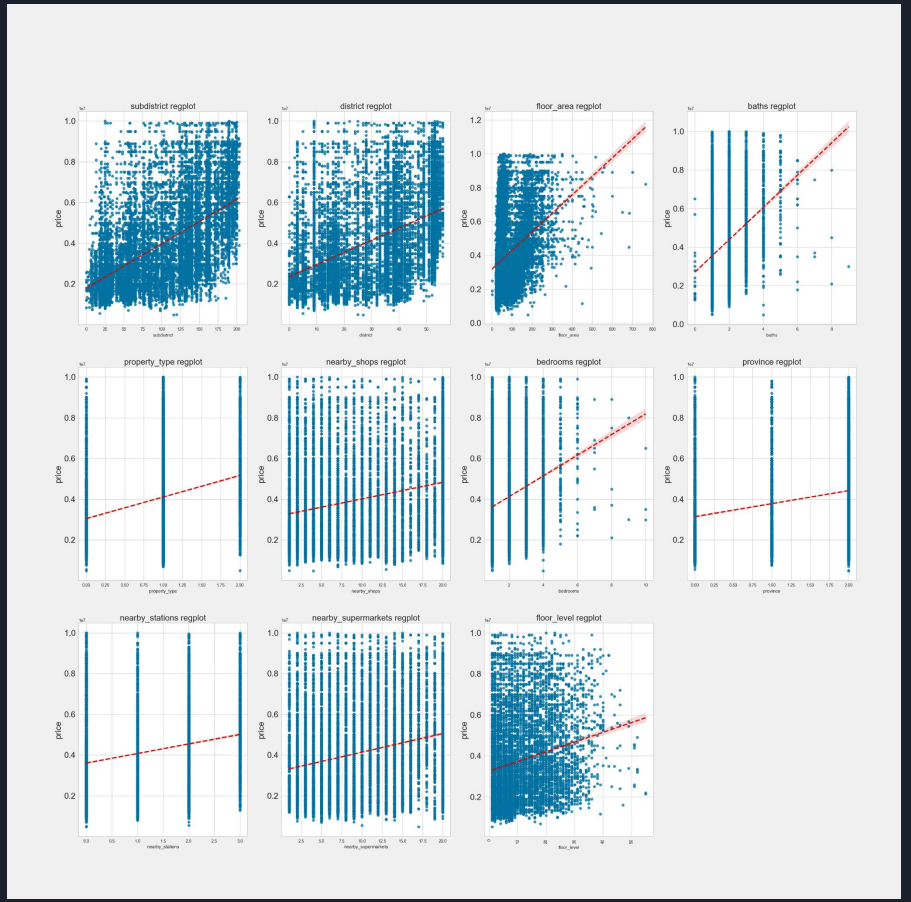
nearby_stations	saleprice
3	15.458860
2	15.129510
0	14.982282
1	14.941575

bedrooms	saleprice
6.0	15.630790
5.0	15.552675
7.0	15.539077
9.0	15.404475
8.0	15.366126
4.0	15.358923
10.0	15.222124
2.0	15.208586
3.0	15.195864
1.0	14.913156

count_facilities	saleprice
31	15.953069
21	15.472844
57	15.424749
13	15.419966
18	15.413153
28	15.402691
25	15.367762
15	15.212087
10	15.142996
6	15.128390

Correlation Analysis





Feature Selection

price	1.000000
saleprice	0.962380
subdistrict	0.564869
district	0.473325
floor_area	0.351357
baths	0.334650
property_type	0.285423
nearby_shops	0.257855
bedrooms	0.254158
province	0.239900
nearby_stations	0.232143
nearby_supermarkets	0.224702
floor_level	0.198645
month_built	0.079473
count_facilities	0.069456
longitude	0.037417
land_area	0.036735
year_built	-0.042557
id	-0.044732
latitude	-0.065560
nearby_bus_stops	-0.066557
closest_station_distance	-0.152466
total_units	-0.285896

```
# Features selection for predictors with price-correlated score above absolute 0.1
features = ['subdistrict', 'district', 'floor_area', 'baths',
            'property_type', 'nearby_shops', 'bedrooms',
            'province', 'nearby_stations', 'nearby_supermarkets',
            'floor_level', 'total_units', 'closest_station_distance']
```

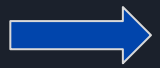
- Selected features are set for X dataframe
- Price for Y dataframe

Pre-processing

KNN-Imputer

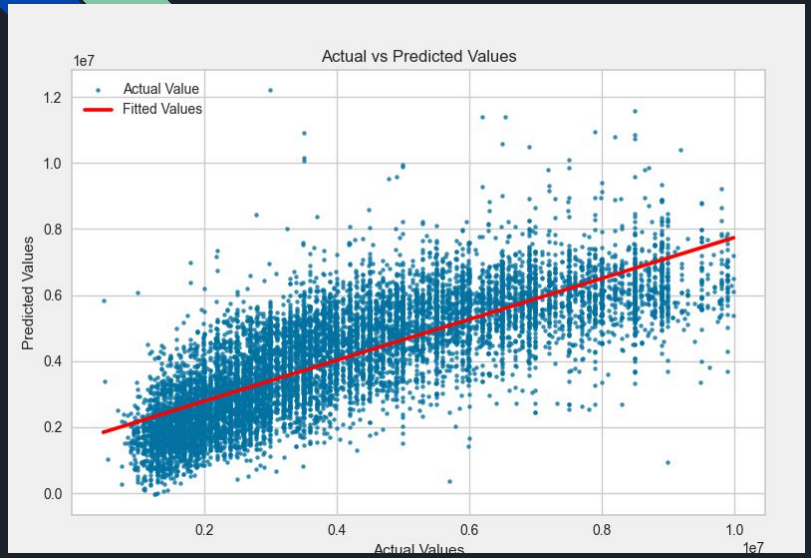
```
[{'K': 1, 'RMSE': 1377283.3377121552},
 {'K': 2, 'RMSE': 1373359.9953259628},
 {'K': 3, 'RMSE': 1373497.8253319028},
 {'K': 4, 'RMSE': 1371375.2827217935},
 {'K': 5, 'RMSE': 1372416.397510357},
 {'K': 6, 'RMSE': 1373266.8728558398},
 {'K': 7, 'RMSE': 1372352.5412769017},
 {'K': 8, 'RMSE': 1372658.577194074},
 {'K': 9, 'RMSE': 1372868.397009458},
 {'K': 10, 'RMSE': 1372597.0730115585},
 {'K': 11, 'RMSE': 1372455.7726511175},
 {'K': 12, 'RMSE': 1372345.5270683996},
 {'K': 13, 'RMSE': 1372567.3684354573},
 {'K': 14, 'RMSE': 1372808.0569123065},
 {'K': 15, 'RMSE': 1372833.6433913363},
 {'K': 16, 'RMSE': 1372629.328407314},
 {'K': 17, 'RMSE': 1373212.4456665257},
 {'K': 18, 'RMSE': 1373402.4419220332},
 {'K': 19, 'RMSE': 1373374.8489120326}]
```

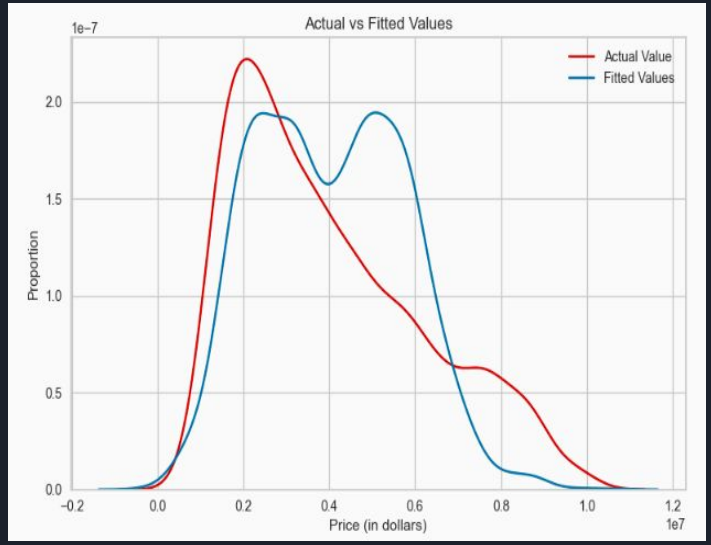
- Perform KNN Imputer from K:1-19, search the K with the lowest RMSE
- Step1: Transforming X_train with KNN Imputer
- Step2: Fit Train KNN with Linear Regression model and evaluate its RMSE



Standard-Scaler

Model1: Linear Regression

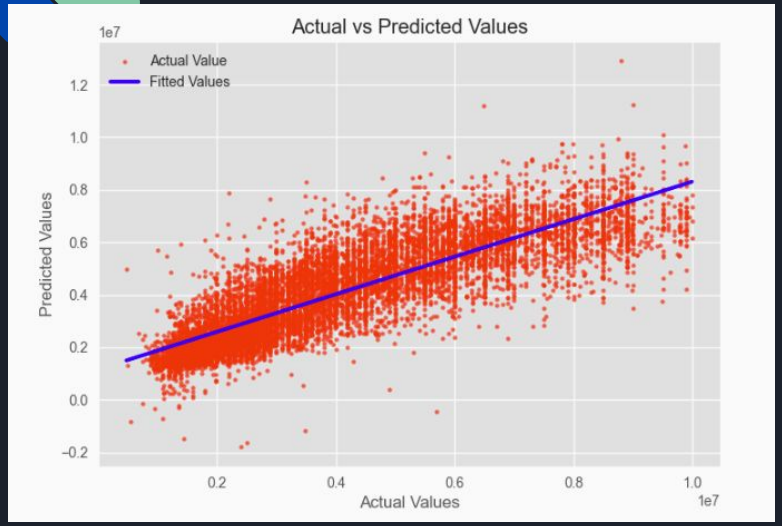


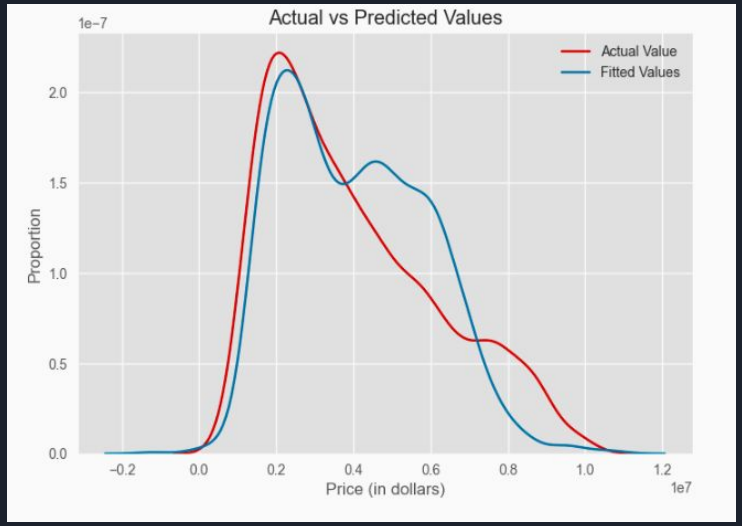



```
R-Square Train:  0.6193559381542455
R-Square Validation:  0.6092927658135293
==========================================================
RMSE of Train : 1346780.1298077581
RMSE of Validation  : 1371375.2827217935
```

- The model is underfit
- Performance on train and test data are resemblance

Model2: Polynomial-Featured Linear Regression



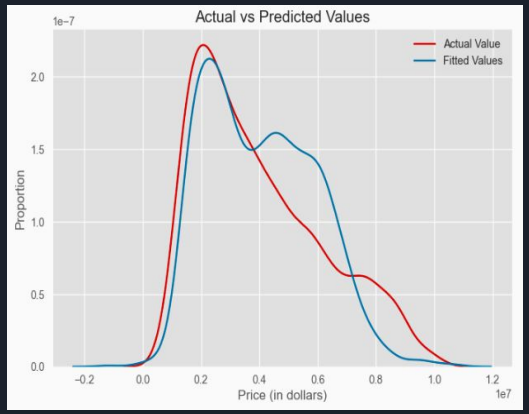



```
R-Square Train:  0.7155729816477747
R-Square Validation:  0.7002739334580338
==================================================
RMSE of Train : 1164186.5969208174
RMSE of Validation  : 1201137.8386152948
```

- The model is underfit
- Performance on train and test data are resemblance
- Performance improved

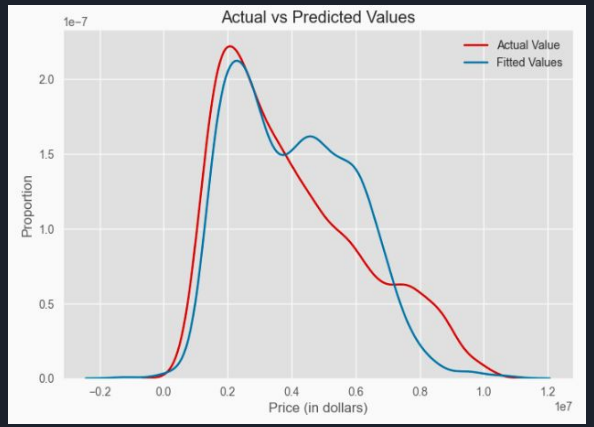
Polynomial Regression Tuning

Ridge Regularization



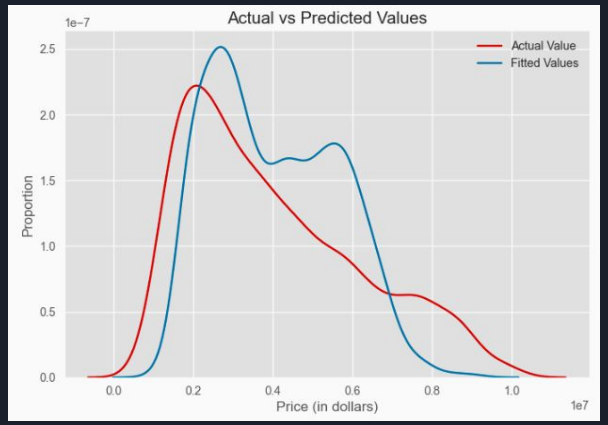

```
R-Square Train:  0.7155349372270392
R-Square Validation:  0.7005436883743024
==========================================
RMSE of Train : 1164264.4540175162
RMSE of Validation  : 1200597.201998452
```

Lasso Regularization




```
R-Square Train:  0.7155729299162299
R-Square Validation:  0.7002846888592379
==========================================
RMSE of Train : 1164186.702791857
RMSE of Validation  : 1201116.2875446773
```

Elastic Net Regularization




```
R-Square Train:  0.6575774224070181
R-Square Validation:  0.6448924379011504
==========================================
RMSE of Train : 1164186.702791857
RMSE of Validation  : 1201116.2875446773
```