TERM PROJECT
HOUSE PRICE
PREDICTION

INSTRUCTORS:
DR. PHAUK SOKKHEY AND MR. CHAN SOPHAL

Presenter: KHUN Dararith

CONTENT

01. INTRODUCTION

Base on the source of the dataset which state that about growth of real estate markets in Sydney and Melbourne. A study start to seeking for opportunities to take advantage of this growth. It makes the real estate markets present an interesting opportunity for data analysts.

02. OBJECTIVE

Since the price of property are keep increasing

This project is proposed to:

Predict of property price

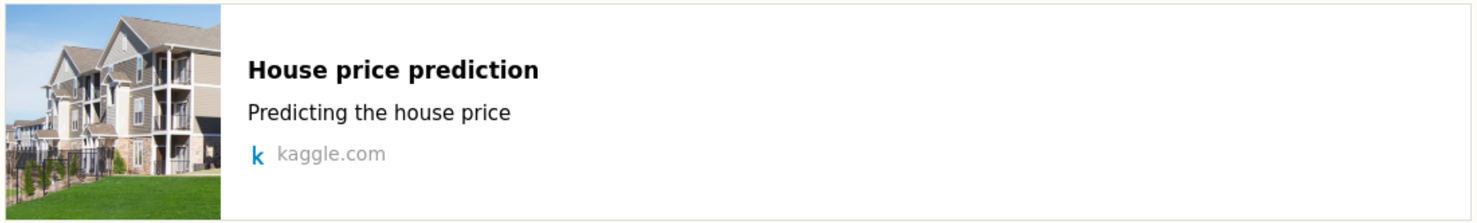
Make a good indicator of market condition

Give a high efficiently prediction

03. DATASET

This dataset is base on real estate market in Sydney and Melbourne. This data was collected since 2014 from unknown source that consist of 4600 rows and 18 columns.

Dataset Source: **Kaggle/*House Price Predection***

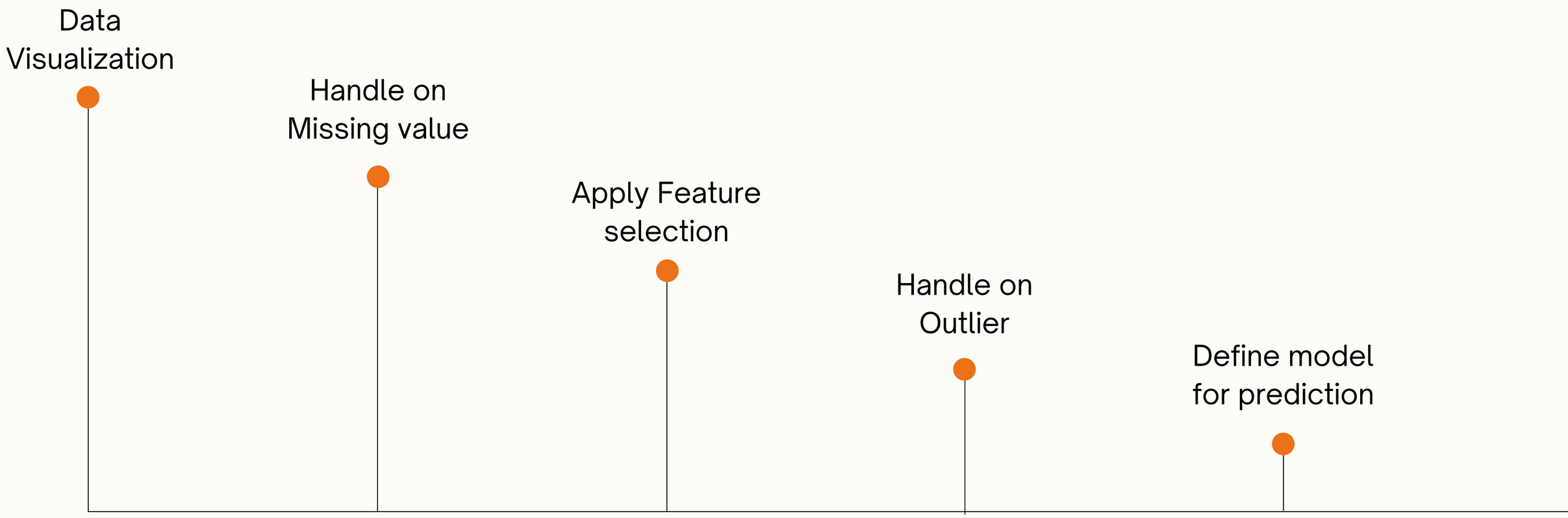


Here is how briefly looks of the dataset:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	str
0	2014-05-02 00:00:00	313000.0	3.0	1.50	1340	7912	1.5	0	0	3	1340	0	1955	2005	188 Densm Ave
1	2014-05-02 00:00:00	2384000.0	5.0	2.50	3650	9050	2.0	0	4	5	3370	280	1921	0	709 Blaine
2	2014-05-02 00:00:00	342000.0	3.0	2.00	1930	11947	1.0	0	0	4	1930	0	1966	0	262(262 143rd A
3	2014-05-02 00:00:00	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000	1963	0	857 17(Pl
4	2014-05-02 00:00:00	550000.0	4.0	2.50	1940	10500	1.0	0	0	4	1140	800	1976	1992	91 170th A
5	2014-05-02 00:00:00	490000.0	2.0	1.00	880	6380	1.0	0	0	3	880	0	1938	1994	522 88th
6	2014-05-02 00:00:00	335000.0	2.0	2.00	1350	2560	1.0	0	0	3	1350	0	1976	0	26 174th A

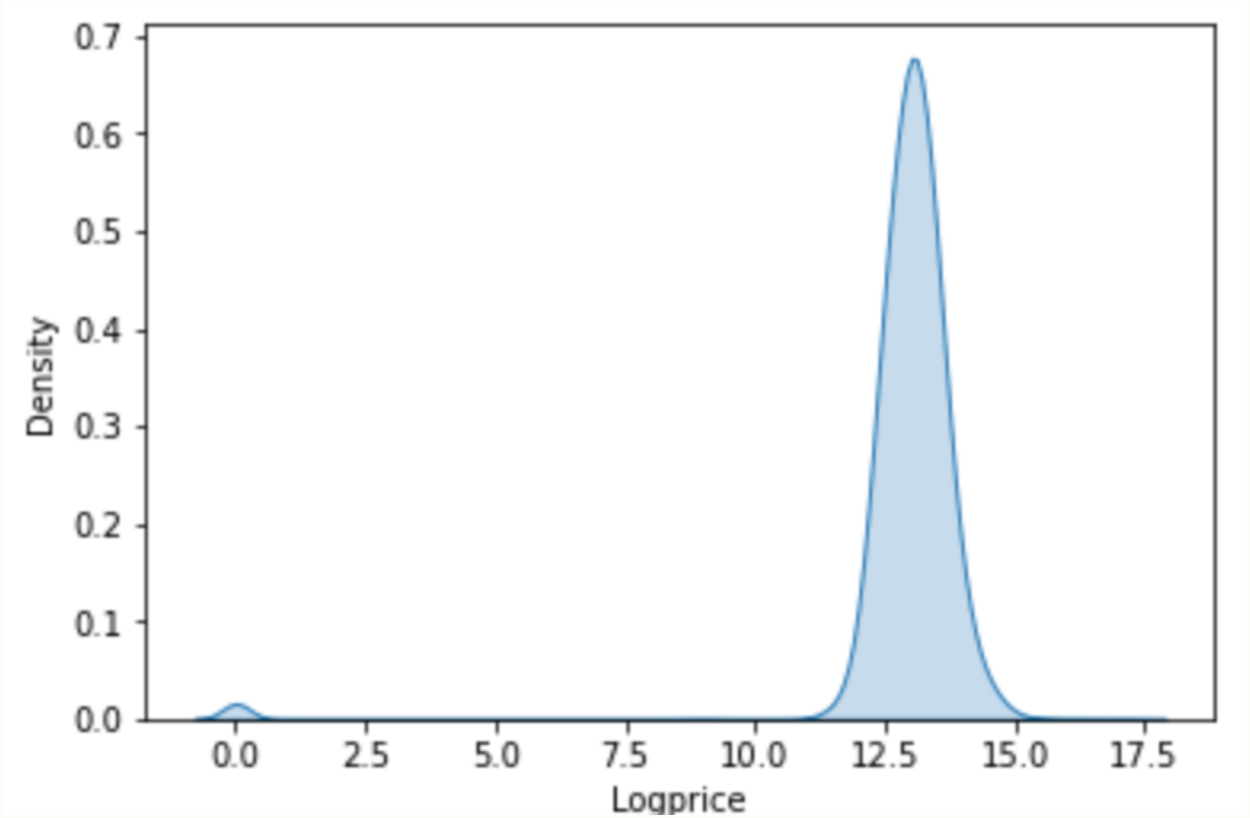
04. METHODOLOGY

Amount Techniques were used in this project



DATA VISUALIZATION

After checking on the distribution of our target column [price].



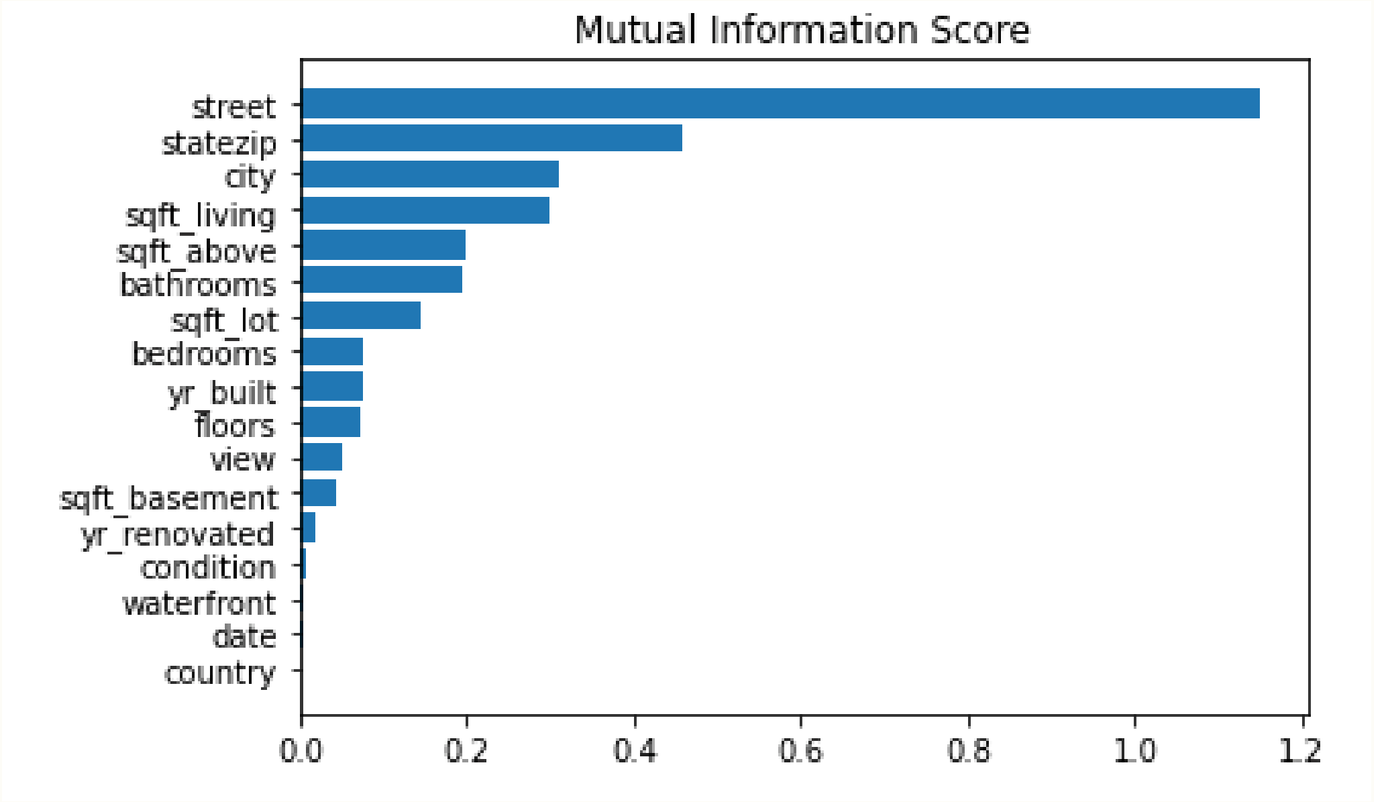
HANDLE ON MISSING VALUE

Mean Imputation should be used to handle on missing value. As a result there is no null data to handle.

```
# Sum of null data
data.isnull().sum()
```

```
date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
street           0
city             0
statezip         0
country          0
dtype: int64
```

APPLY FEATURE SELECTION



10 out of 18 as new features for this project.

Features were selected based on Mutual info score.

HANDLE ON OUTLIER

Z-Score is used to perform outlier detection. When z-score were calculate, the data point are too far from zero.

- Define threshold 3 or -3
- Checking if z-score value greater than 3 or smaller than -3, it is the outlier.

273 rows data were found as outlier.

```
new_dataset.shape
```

```
(4435, 10)
```

DEFINE MODEL FOR PREDICTION

Before define model for prediction, The data were split into

- 80% as training set
- 20% as test set

To make a better perform on prediction, Some column such as [street, statezip, city] were encoded inform of number.

	street	statezip	city	sqft_living	sqft_above	bathrooms	yr_built	sqft_lot	bedrooms
0	3.859344e+05	375656.132823	421503.366508	1340	1340	1.50	1955	7912	3.0
1	1.766601e+06	849969.156900	570200.891592	3650	3370	2.50	1921	9050	5.0
2	4.052677e+05	321296.502254	439062.351650	1930	1930	2.00	1966	11947	3.0
3	4.572677e+05	605283.972110	794664.954324	2000	1000	2.25	1963	8030	3.0

DEFINE MODEL FOR PREDICTION

There are 5 Models were selected:

- Linear Regression
- Robust Regression
- Lasso Regression
- Support Vector Machine
- Random Forest

05. TOOLS ADN RESULT

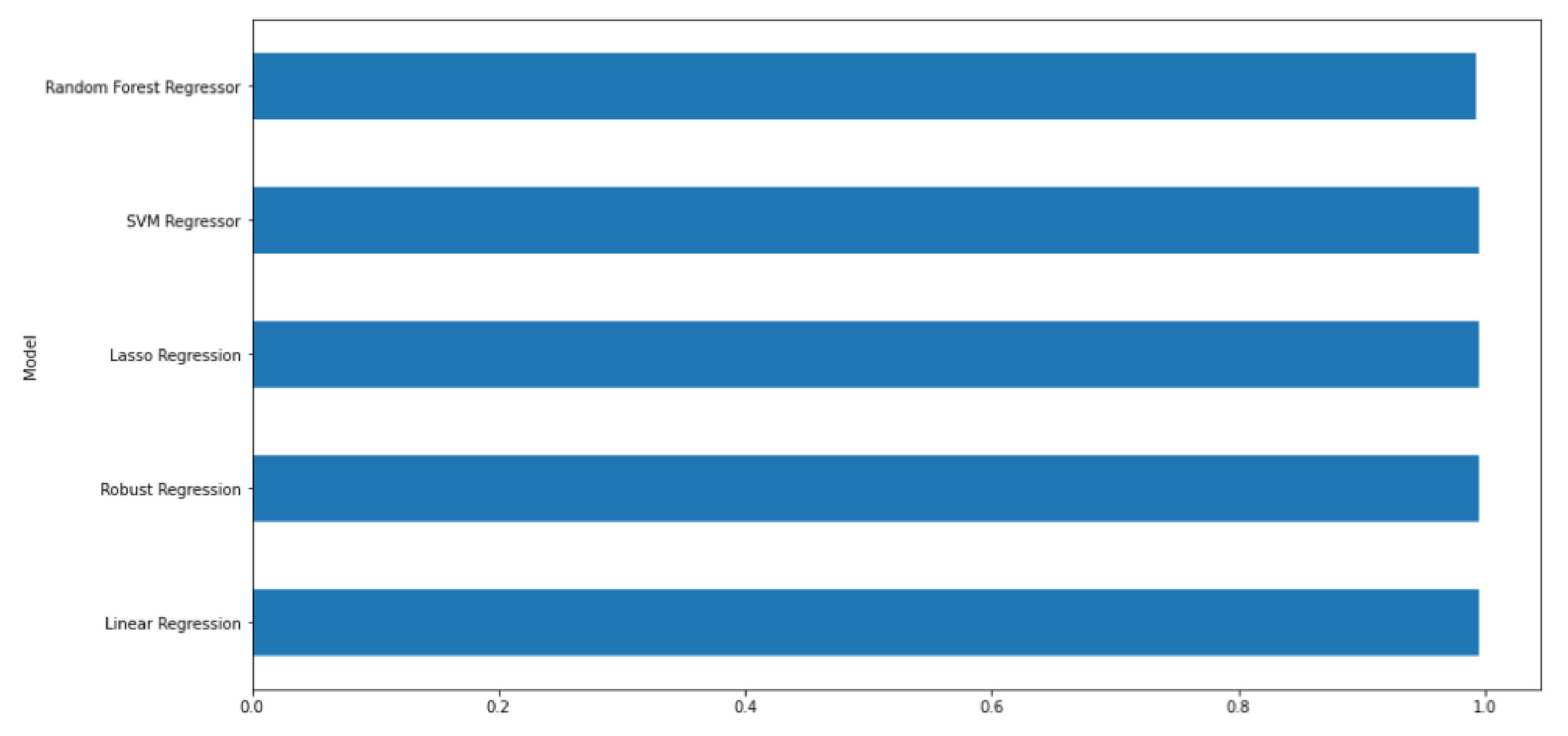
JupyterLab is the latest web-based interactive development environment for notebooks, code, and data.

RESULT

It is a project of prediction, to evaluate the result i use

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R Squared (R2)

	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	2952.223817	4.577064e+08	21394.074029	0.994750
1	Robust Regression	2904.735925	4.581444e+08	21404.309056	0.994745
2	Lasso Regression	2807.911756	4.578024e+08	21396.318222	0.994749
3	SVM Regressor	2427.215018	4.594300e+08	21434.317538	0.994730
4	Random Forest Regressor	3930.331842	6.503838e+08	25502.623595	0.992540


Random Forest Regressor
SVM Regressor
Lasso Regression
Robust Regression
Linear Regression
Model
0.0
0.2
0.4
0.6
0.8
1.0

06. CONCLUSION

In conclusion, after perform the prediction on house price we see that the result of each model are slightly the same (up to 99%) . We notice that Linear Regression is the best among those 5 models.

What have i learnt from this project?

- Improve researching skill
- Seeking for solution
- Getting to know about new model (Lasso, Robust...)
- Learn how to apply what i have learnt in to one one project