We Care About
Your Future

# Predict Housing Prices Studi Case : California Housing Prices


Final Project Oleh :

- Syifa Auliyah Hasanah
- Adamita Ruswir
- Laurenzius Julio
- Martinus Sapto Nugroho

We Care About
Your Future

# LATAR BELAKANG MASALAH

Keinginan untuk memiliki rumah merupakan impian dari banyak orang. Tetapi pada kenyataannya, banyak rumah yang dijual mahal namun tidak sesuai dengan spesifikasinya. Maka dari itu, kami mencoba untuk membuat sistem yang kiranya dapat memberikan referensi untuk memprediksi harga rumah.

**Problem :**

Adanya ketidak merataan daerah-daerah tertentu untuk tempat tinggal. (Ada yang sangat laku ada yang tidak laku)

Tujuan dari analisis statistik ini adalah untuk membantu memahami hubungan antara fitur/lokasi rumah dan bagaimana variabel tersebut digunakan untuk memprediksi harga rumah.

**Objective :**

- Memprediksi harga rumah.
- Menggunakan model yang berbeda untuk meminimalkan perbedaan antara prediksi dan aktualnya.

We Care About
Your Future

# PRE-PROCESSING

## Keterangan

- longitude: seberapa jauh ke barat sebuah rumah; nilai yang lebih tinggi lebih jauh ke barat
- latitude: seberapa jauh ke utara sebuah rumah; nilai yang lebih tinggi lebih jauh ke utara
- housingMedianAge: Usia rata-rata sebuah rumah dalam satu blok; angka yang lebih rendah adalah bangunan baru
- totalRooms: Jumlah total kamar dalam satu blok
- totalBedrooms: Jumlah total kamar tidur dalam satu blok
- population: Jumlah total orang yang tinggal dalam satu blok
- households: Jumlah total rumah tangga, sekelompok orang yang tinggal dalam satu unit rumah, untuk satu blok
- medianIncome: Pendapatan rata-rata untuk rumah tangga dalam satu blok rumah (diukur dalam puluhan ribu Dolar AS)
- medianHouseValue: Nilai median rumah untuk rumah tangga dalam satu blok (diukur dalam Dolar AS)
- oceanProximity: Lokasi rumah dengan laut/laut

We Care About
Your Future

# PRE-PROCESSING

## 1. DATA CLEANING

### RAW DATA

```
[26] df
```

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | median_house_value | ocean_proximity |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -122.23 | 37.88 | 41.0 | 880.0 | 129.0 | 322.0 | 126.0 | 8.3252 | 452600.0 | NEAR BAY |
| **1** | -122.22 | 37.86 | 21.0 | 7099.0 | 1106.0 | 2401.0 | 1138.0 | 8.3014 | 358500.0 | NEAR BAY |
| **2** | -122.24 | 37.85 | 52.0 | 1467.0 | 190.0 | 496.0 | 177.0 | 7.2574 | 352100.0 | NEAR BAY |
| **3** | -122.25 | 37.85 | 52.0 | 1274.0 | 235.0 | 558.0 | 219.0 | 5.6431 | 341300.0 | NEAR BAY |
| **4** | -122.25 | 37.85 | 52.0 | 1627.0 | 280.0 | 565.0 | 259.0 | 3.8462 | 342200.0 | NEAR BAY |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **20635** | -121.09 | 39.48 | 25.0 | 1665.0 | 374.0 | 845.0 | 330.0 | 1.5603 | 78100.0 | INLAND |
| **20636** | -121.21 | 39.49 | 18.0 | 697.0 | 150.0 | 356.0 | 114.0 | 2.5568 | 77100.0 | INLAND |
| **20637** | -121.22 | 39.43 | 17.0 | 2254.0 | 485.0 | 1007.0 | 433.0 | 1.7000 | 92300.0 | INLAND |
| **20638** | -121.32 | 39.43 | 18.0 | 1860.0 | 409.0 | 741.0 | 349.0 | 1.8672 | 84700.0 | INLAND |
| **20639** | -121.24 | 39.37 | 16.0 | 2785.0 | 616.0 | 1387.0 | 530.0 | 2.3886 | 89400.0 | INLAND |

20640 rows × 10 columns

```
[6] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

We Care About
Your Future

# PRE-PROCESSING

- Cek and handling missing values (Mengisi variabel yang memiliki nilai yang hilang dengan median agar proses pemodelan bisa dilakukan)
- Cek and handling outlier (Data yang berada diluar batas wajar (outlier) bisa mempengaruhi proses prediksi yang akan dilakukan, oleh karena itu penghapusan nilai-nilai yang berada diluar batas quartil 1 dan 3 akan dihapus)

```
[10] #Cek Missing Values
     print('Missing Values pada Data :')
     print(df.isnull().sum())
```

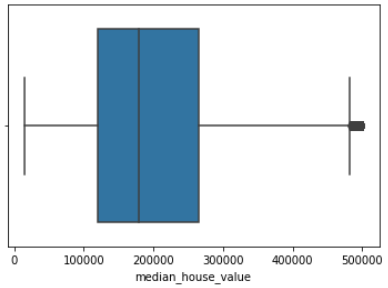
```
Missing Values pada Data :
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

```
[11] #Handling Missing Value
     print('=====================================\n')
     print('Handling Missing Values variabel total_bedrooms pada Data :')
     df['total_bedrooms'] = df['total_bedrooms'].fillna(df['total_bedrooms'].median())
     df.isna().sum()
```

```
=====================================

Handling Missing Values variabel total_bedrooms pada Data :
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

```
[12] #Cek Outlier
     sns.boxplot(df['median_house_value'])
     plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43:
  FutureWarning
```




```
[13] #Define Q1 and Q3

     Q1 = df['median_house_value'].quantile(0.25)
     Q3 = df['median_house_value'].quantile(0.75)
     IQR = Q3 - Q1
     Lower_Whisker = Q1-(1.5*IQR)
     Upper_Whisker = Q3+(1.5*IQR)
     print(Upper_Whisker)
     print(Lower_Whisker)
```

```
482412.5
-98087.5
```

We Care About
Your Future

# PRE-PROCESSING

2. Feature Engineering

One Hot Encoding pada variabel ocean_proximity (bertujuan agar variabel kategori memiliki nilai numerik yang nantinya bisa digunakan pada proses pemodelan)

One Hot Encoding

```
[ ] le_ocean_pro = LabelEncoder()
    df["ocean_proximity"] =le_ocean_pro.fit_transform(df["ocean_proximity"])
    df.head()
```

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | median_house_value | ocean_proximity |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -122.23 | 37.88 | 41.0 | 880.0 | 129.0 | 322.0 | 126.0 | 8.3252 | 452600.0 | 3 |
| **1** | -122.22 | 37.86 | 21.0 | 7099.0 | 1106.0 | 2401.0 | 1138.0 | 8.3014 | 358500.0 | 3 |
| **2** | -122.24 | 37.85 | 52.0 | 1467.0 | 190.0 | 496.0 | 177.0 | 7.2574 | 352100.0 | 3 |
| **3** | -122.25 | 37.85 | 52.0 | 1274.0 | 235.0 | 558.0 | 219.0 | 5.6431 | 341300.0 | 3 |
| **4** | -122.25 | 37.85 | 52.0 | 1627.0 | 280.0 | 565.0 | 259.0 | 3.8462 | 342200.0 | 3 |

```
[ ] df['ocean_proximity'].value_counts()
```

```
0    8552
1    6519
4    2419
3    2074
2       5
Name: ocean_proximity, dtype: int64
```

We Care About
Your Future

# PRE-PROCESSING

Dataset setelah melalui proses preprocessing dan feature Engineering

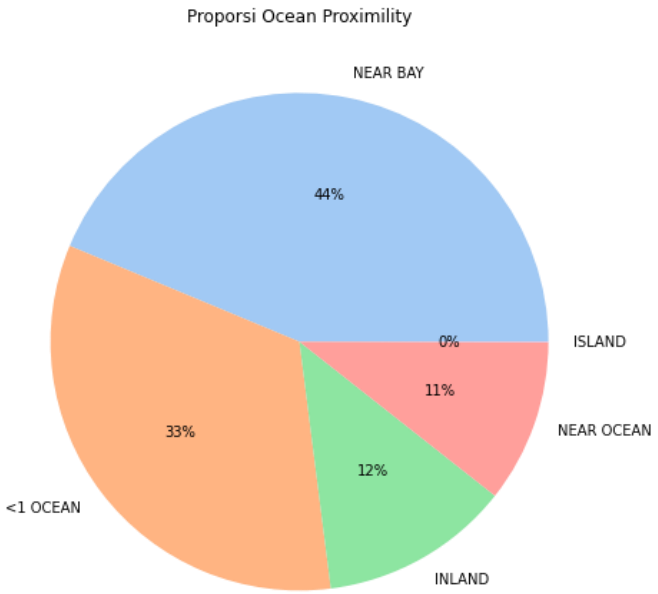
```
[56] df
```

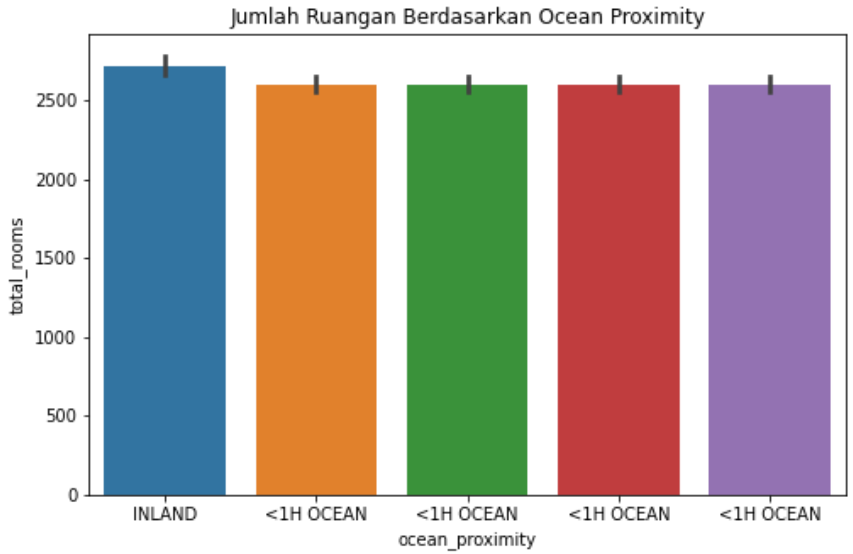
| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income | median_house_value | ocean_proximity |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -122.23 | 37.88 | 41.0 | 880.0 | 129.0 | 322.0 | 126.0 | 8.3252 | 452600.0 | 3 |
| **1** | -122.22 | 37.86 | 21.0 | 7099.0 | 1106.0 | 2401.0 | 1138.0 | 8.3014 | 358500.0 | 3 |
| **2** | -122.24 | 37.85 | 52.0 | 1467.0 | 190.0 | 496.0 | 177.0 | 7.2574 | 352100.0 | 3 |
| **3** | -122.25 | 37.85 | 52.0 | 1274.0 | 235.0 | 558.0 | 219.0 | 5.6431 | 341300.0 | 3 |
| **4** | -122.25 | 37.85 | 52.0 | 1627.0 | 280.0 | 565.0 | 259.0 | 3.8462 | 342200.0 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **20635** | -121.09 | 39.48 | 25.0 | 1665.0 | 374.0 | 845.0 | 330.0 | 1.5603 | 78100.0 | 1 |
| **20636** | -121.21 | 39.49 | 18.0 | 697.0 | 150.0 | 356.0 | 114.0 | 2.5568 | 77100.0 | 1 |
| **20637** | -121.22 | 39.43 | 17.0 | 2254.0 | 485.0 | 1007.0 | 433.0 | 1.7000 | 92300.0 | 1 |
| **20638** | -121.32 | 39.43 | 18.0 | 1860.0 | 409.0 | 741.0 | 349.0 | 1.8672 | 84700.0 | 1 |
| **20639** | -121.24 | 39.37 | 16.0 | 2785.0 | 616.0 | 1387.0 | 530.0 | 2.3886 | 89400.0 | 1 |

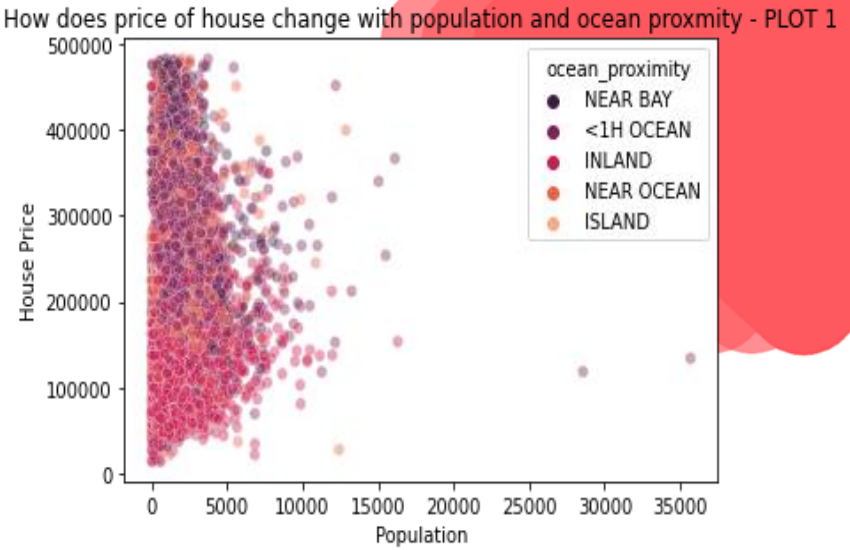
19569 rows × 10 columns

We Care About
Your Future

# INSIGHT




- 44% orang akan memilih rumah yang dekat dengan Bay
- Tidak ada orang yang ingin memiliki rumah hanya di pulau saja (proporsi = 0%)

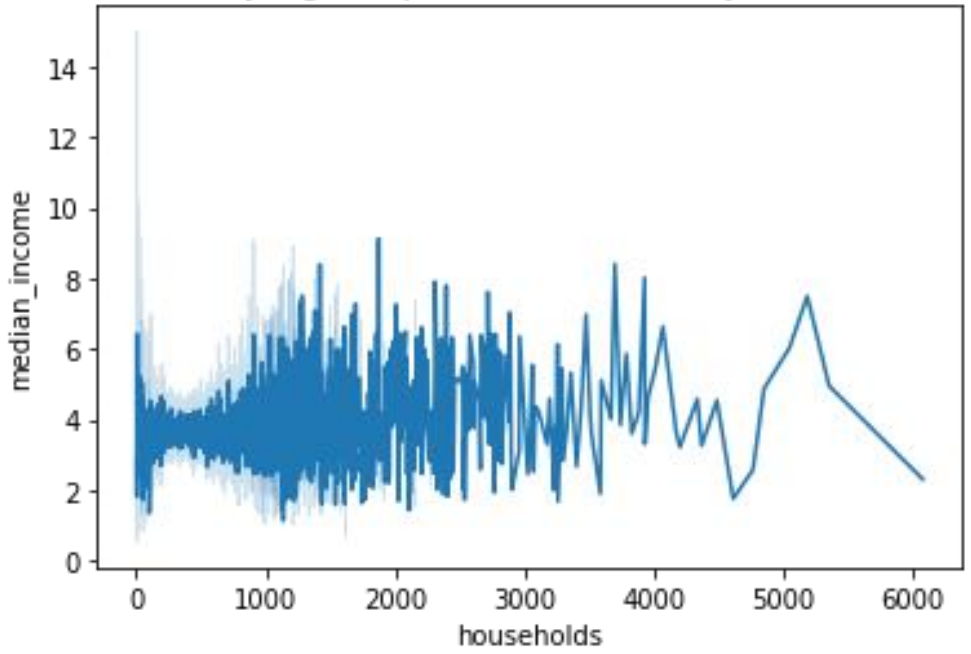
Jumlah total ruangan terbanyak berada pada rumah-rumah yang berada di pedalaman, sedangkan untuk kategori lainnya cenderung memiliki total ruangan yang sama

Rumah yang berada di pedalaman cenderung memiliki harga yang lebih murah dibanding rumah yang berada di dekat ocean atau bay

We Care About
Your Future

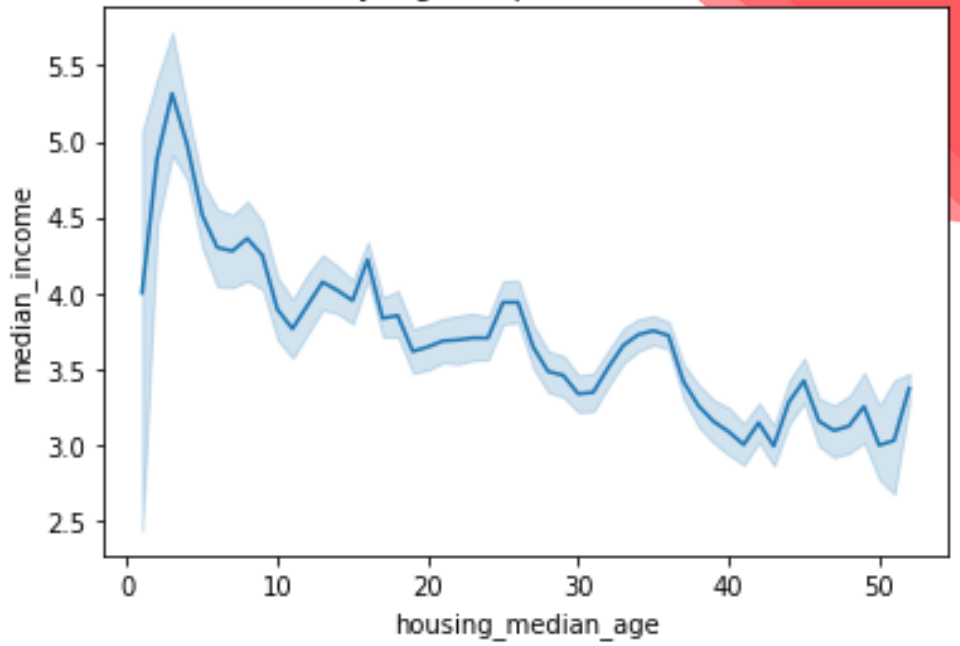
# INSIGHT

Pertumbuhan Income yang Didapatkan Berdasarkan Jumlah Total Rumah Tangga




Jumlah households tidak terlalu mempengaruhi median incomenya karena terlihat dari grafik diatas tidak terdapat suatu tren tertentu dan data cenderung naik turun.

Pertumbuhan Income yang Didapatkan Berdasarkan Umur Rumah



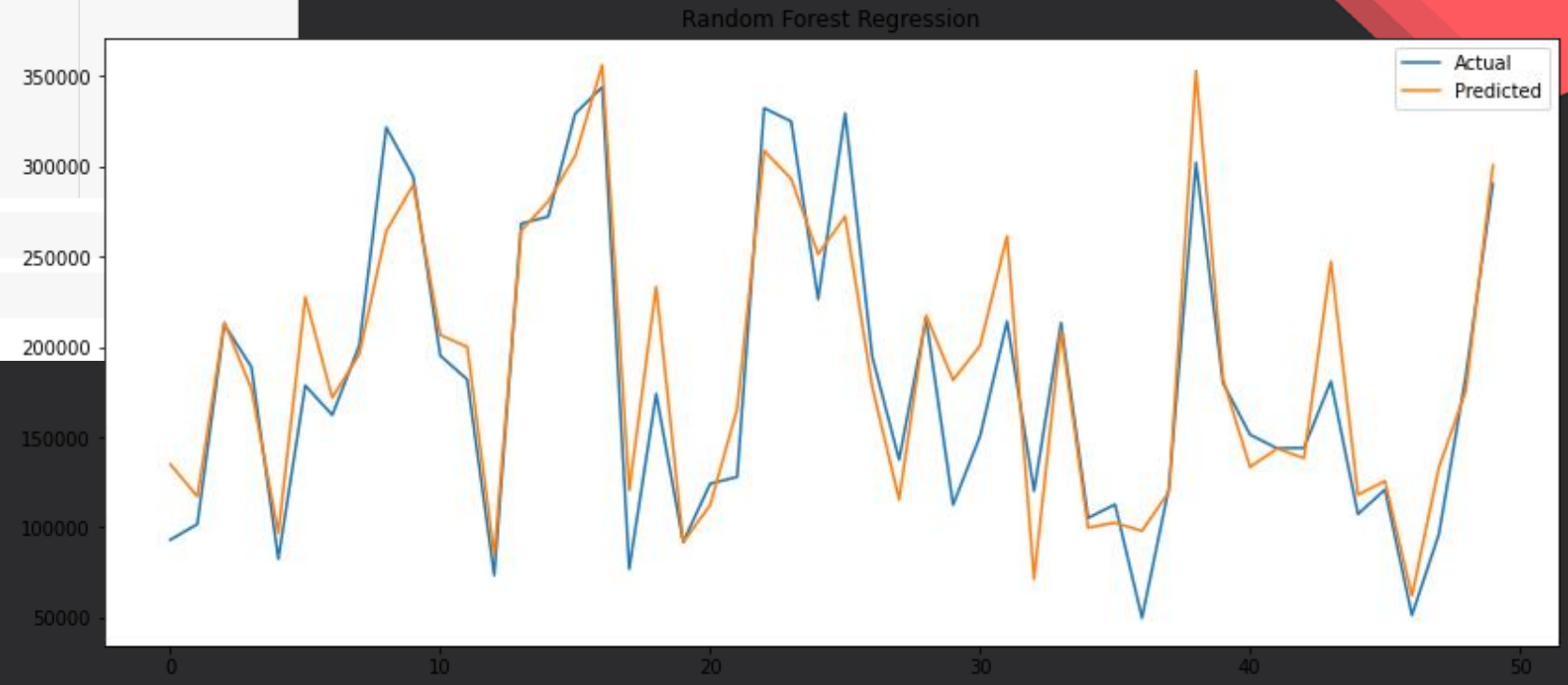

Umur hunian akan berpengaruh terhadap incomenya karena semakin tua suatu hunian, maka akan semakin kecil pula income yang didapatkan.

We Care About
Your Future

# MODELLING EXPERIMENTS

## RANDOM FOREST

```
[60] #Function Random Forest
     def forest(df):
       #Split Data
       train_X, test_X, train_y, test_y = train_test_split(X, y, test_size=0.25, random_state=43)

       #Model Random Forest

       #create the model
       rf=RandomForestRegressor()

       #fit the model
       rf_fit = rf.fit(train_X,train_y)

       #prediction value
       rf_predict = rf.predict(test_X)

       #score random forest
       rf_score = rf.score(train_X, train_y), rf.score(test_X, test_y)
       return rf_predict, rf_score, test_y

[61] rf_predict, rf_score, test_y = forest(df)

[62] print("Random Forest Score: {}".format(rf_score))

     Random Forest Score: (0.9699280753770038, 0.7900394506389028)
```



We Care About
Your Future

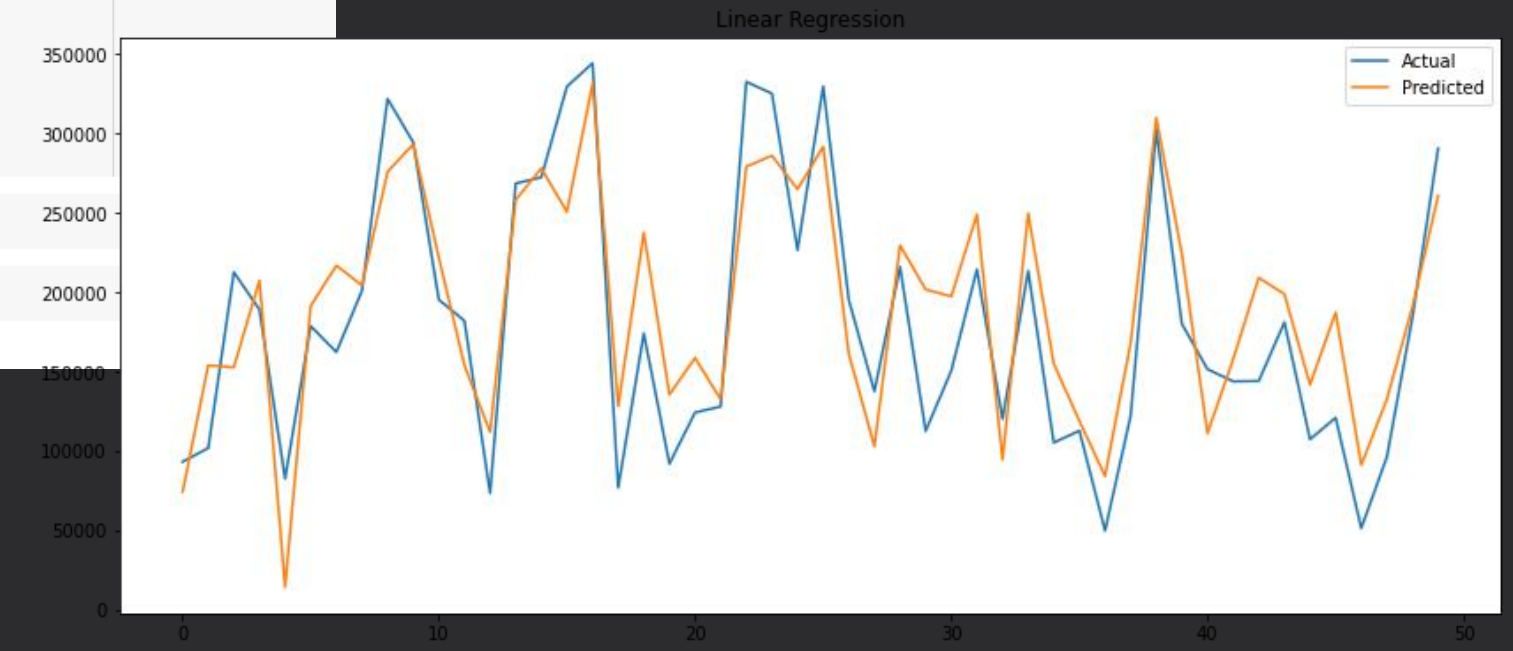
# MODELLING EXPERIMENTS

## LINEAR REGRESSION

```
[63] #Function Linear Regression
     def linear(df):
       #Split Data
       train_X, test_X, train_y, test_y = train_test_split(X, y, test_size=0.25, random_state=43)

       #Model Random Forest

       #create the model
       lr = LinearRegression()

       #fit the model
       lr_fit = lr.fit(train_X,train_y)

       #prediction value
       lr_predict = lr.predict(test_X)

       #score random forest
       lr_score = lr.score(train_X, train_y), lr.score(test_X, test_y)
       return lr_predict, lr_score, test_y
```

```
[64] lr_predict, lr_score, test_y = linear(df)
```

```
[65] print("Linear Regression Score: {}".format(lr_score))

     Linear Regression Score: (0.6011240629891184, 0.6002390063667927)
```



We Care About
Your Future

# MODELLING EXPERIMENTS

```
[74] models_evaluation(compare_rf, "Random Forest Regression")
0s

     ---------------------------------
     Random Forest Regression
     R-Squared: 0.7900394506389028
     MAE: 29162.935583486615
     MSE: 1891367906.6644585
     MAE%: 0.174896270810106
     ---------------------------------
```

```
[75] models_evaluation(compare_linear, "Linear Regression")
0s

     ---------------------------------
     Linear Regression
     R-Squared: 0.6002390063667927
     MAE: 44618.30944440582
     MSE: 3601129431.195121
     MAE%: 0.2882988573087535
     ---------------------------------
```

Result : Modelling yang diambil menggunakan metode **Random Forest Regression** karena memiliki R^2 tertinggi dan nilai MAE terendah

We Care About
Your Future

# EXECUTIVE SUMMARY AND RECOMMENDATION

SUMMARY :

1. Sebagian besar orang memilih rumah yang dekat dengan Bay. Tidak ada orang yang ingin memiliki rumah hanya di pulau saja (proporsi = 0%)
2. Jumlah total ruangan terbanyak berada pada rumah-rumah yang berada di pedalaman, sedangkan untuk kategori lainnya cenderung memiliki total ruangan yang sama
3. Rumah yang berada di pedalaman cenderung memiliki harga yang lebih murah dibanding rumah yang berada di dekat ocean atau bay
4. Jumlah households tidak terlalu mempengaruhi median incomenya karena terlihat dari grafik tidak terdapat suatu tren tertentu dan data cenderung naik turun.
5. Umur hunian akan berpengaruh terhadap incomenya karena semakin tua suatu hunian, maka akan semakin kecil pula income yang didapatkan.

RECOMMENDATION :

1. Meningkatkan marketing khususnya bagi hunian yang berada di pedalaman
2. Menambah jumlah ruangan di kelompok hunian yang berada di semua kategori selain pedalaman
3. Melakukan perbaikan pada hunian yang memiliki umur yang lebih tua agar income yang didapatkan bisa bertambah

We Care About
Your Future

# PEMBAGIAN TUGAS

1. Problem apa yang ingin diselesaikan

Jawab : Membantu memahami hubungan antara fitur/lokasi rumah dan bagaimana variabel tersebut digunakan untuk memprediksi harga rumah.

1. Dataset seperti apa yang kamu miliki

Jawab : Dataset terdiri dari 9 variabel dengan jumlah baris sebanyak 20640.

1. Insight apa saja yang kamu temukan dari data tersebut

Bisa jelaskan 2 insight paling bagus menurutmu serta action apa yang dapat dilakukan setelah mengetahui insight tersebut

Jawab :

- Umur hunian akan berpengaruh terhadap incomenya karena semakin tua suatu hunian, maka akan semakin kecil pula income yang didapatkan sehingga kita bisa melakukan perbaikan pada hunian yang memiliki umur yang lebih tua agar income yang didapatkan bisa bertambah
- Sebagian besar orang memilih rumah yang dekat dengan Bay. Tidak ada orang yang ingin memiliki rumah hanya di pulau saja (proporsi = 0%) sehingga perlu meningkatkan marketing khususnya bagi hunian yang berada di pedalaman

1. Apa saja yang telah dilakukan dalam membuat model?

Jawab : Model Random Forest Regression dengan pembagian data latih dan data uji sebesar 75% dan 25%

We Care About
Your Future

# Thanks For Your Attention.

Follow our social media on :

@data_bangalore  Data Bangalore  Data Bangalore Id