Dream House Realty Group

WE MAKE YOUR DREAMS COME TRUE!

2023

dhrgroup.com

ABOUT US

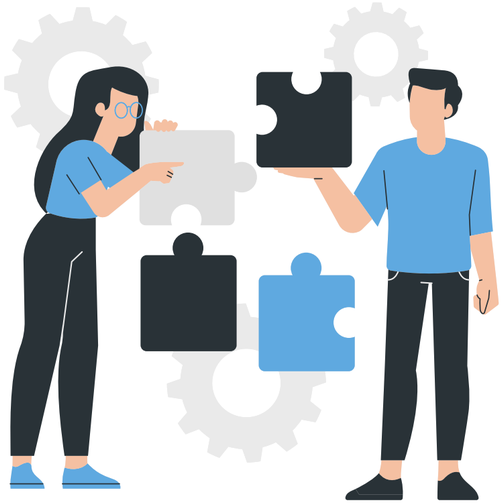
Dream House Realty Group is a real estate agency company focused on assisting clients in selling and finding dream properties in California. With an experienced team and extensive knowledge of the California housing market, Dream House Realty Group is committed to providing quality services and effective solutions in the housing business.

dhrgroup.com

BUSINESS PROCESS

PROBLEM STATEMENT

How to give prediction of property prices appropriately based on available features?



DATA PREPARATION

HANDLING MISSING VALUE

Because the missing value contained in the given dataset is only **0.948%** (**less than 1% of the total data**), the method of deleting rows / columns containing missing value can be a faster process option, and is not problematic for the entire data.


```
longitude             0.000000
latitude              0.000000
housing_median_age    0.000000
total_rooms           0.000000
total_bedrooms        0.948228
population            0.000000
households            0.000000
median_income         0.000000
ocean_proximity       0.000000
median_house_value    0.000000
dtype: float64
```

DATA PREPARATION

FEATURE ENGINEERING

With the aim that the data can be better understood, new features are made based on existing data. The new features are as follows:

- **'person_in_house'** = Represents the number of people in 1 house.
- **'bedroom_in_house'** = Represents the number of bedrooms in 1 house.
- **'room_in_house'** = Represents the number of rooms in 1 house.

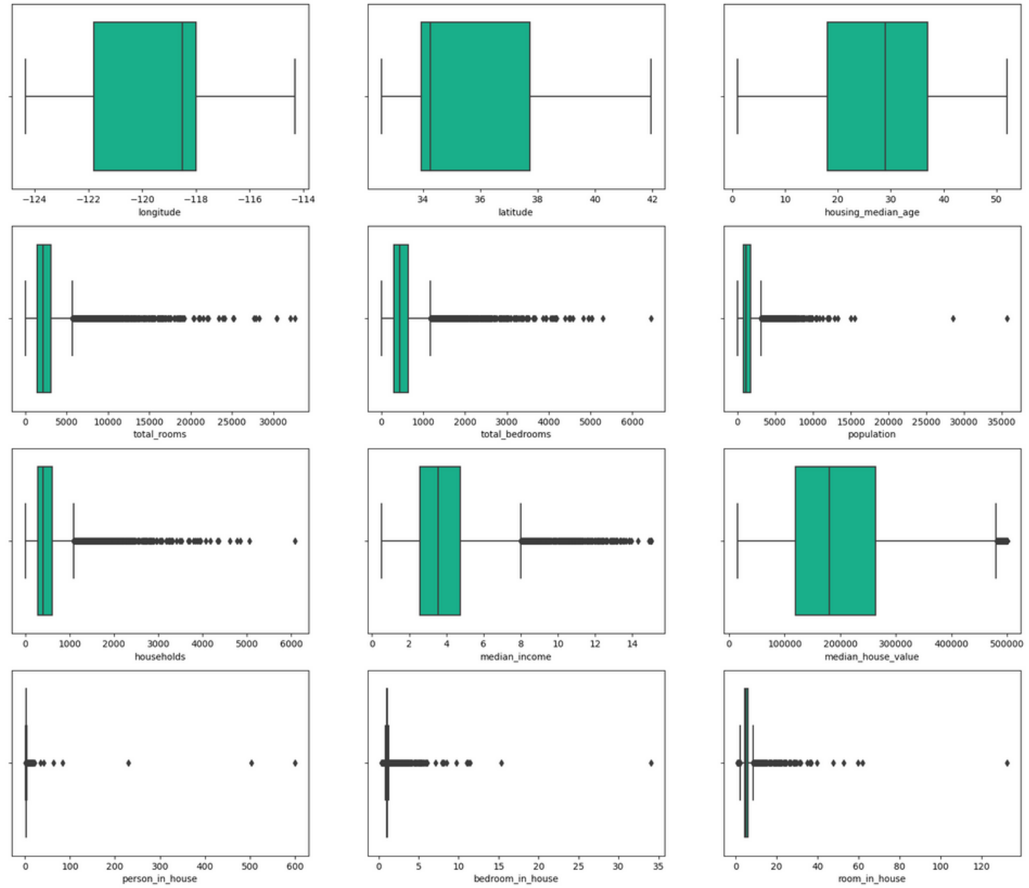
The new feature is created by dividing the corresponding column by other relevant columns.

DATA PREPARATION

HANDLING OUTLIERS

The steps to be carried out consist of several processes, namely as follows:

- Calculate IQR (Interquartile Range).
- Set an Upper Limit and a Lower Limit.
- Identify Outliers that are outside the Upper and Lower Limits of the Data.
- Elimination of data that has been identified as Outlier.



DATA PREPARATION

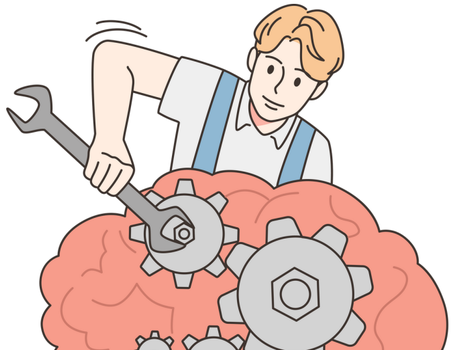
HANDLING OUTLIERS

After handling outliers, there are now 10996 data left from 14448 total initial data.

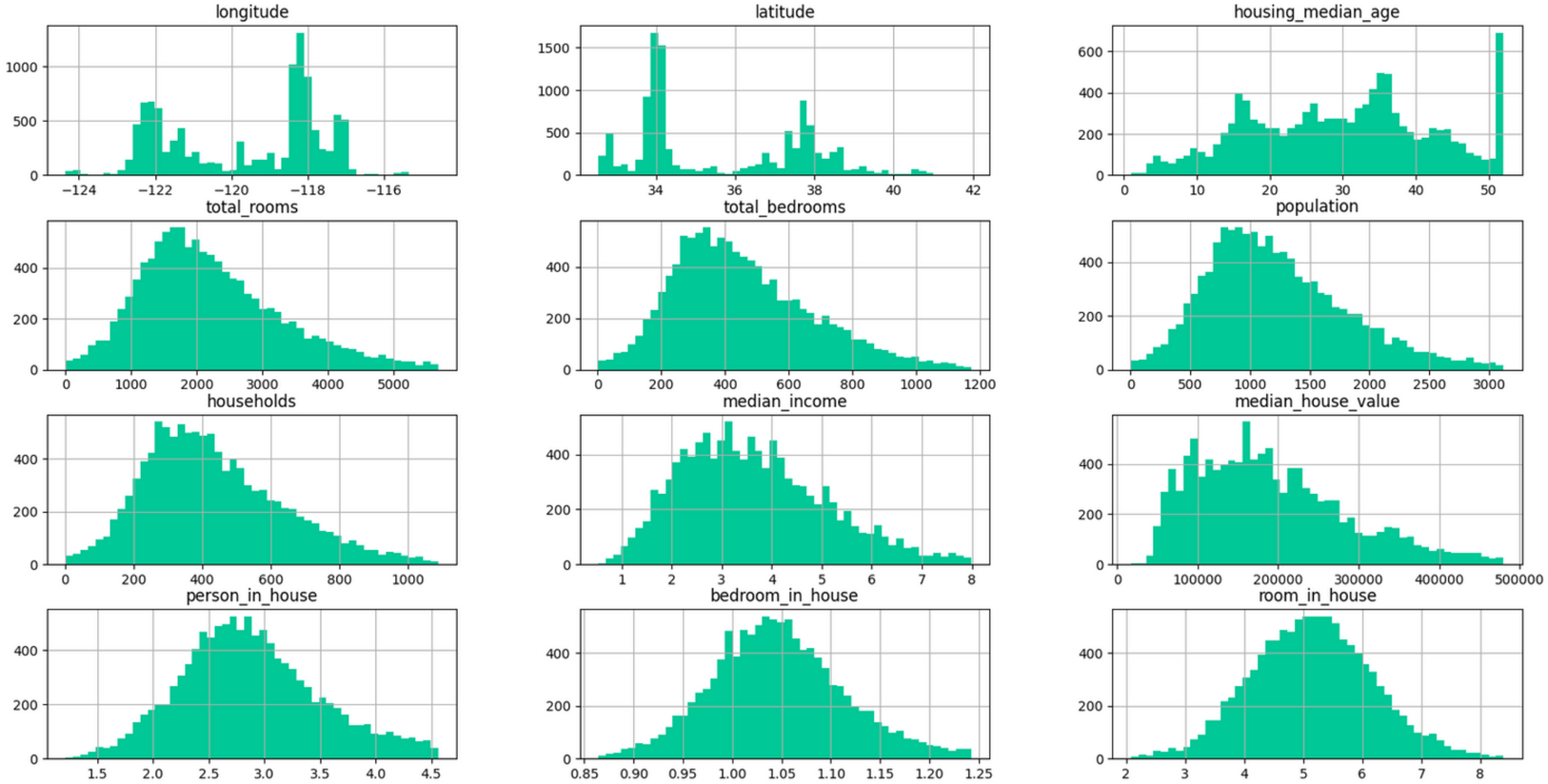
- Deleted data is about **23.90%** of the total initial data.
- Final data used **76.10%** of the total initial data.

Outlier data cleaning can be done **up to 30%** of the initial amount of data in fulfilling the test carried out. So, **Data can still be used** for the Model to be applied.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10996 entries, 1 to 14447
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           10996 non-null  float64
 1   latitude            10996 non-null  float64
 2   housing_median_age  10996 non-null  float64
 3   total_rooms         10996 non-null  float64
 4   total_bedrooms      10996 non-null  float64
 5   population          10996 non-null  float64
 6   households          10996 non-null  float64
 7   median_income       10996 non-null  float64
 8   ocean_proximity     10996 non-null  object
 9   median_house_value  10996 non-null  float64
 10  person_in_house     10996 non-null  float64
 11  bedroom_in_house    10996 non-null  float64
 12  room_in_house       10996 non-null  float64
dtypes: float64(12), object(1)
memory usage: 1.2+ MB
```



DATA PREPARATION DISTRIBUTION CHECK



DATA PREPARATION

BINNING

A way to categorize the age of property buildings is carried out, which allows binning of the **housing_median_age** column, in the hope of gaining deeper insights.

housing_age_category	housing_age
New Property	0 - 10 Years
Small Renovation Property	10 - 20 Years
Medium Renovation Property	20 - 30 Years
Big Renovation Property	30 - 40 Years
Antique / Old Property	>40 Years

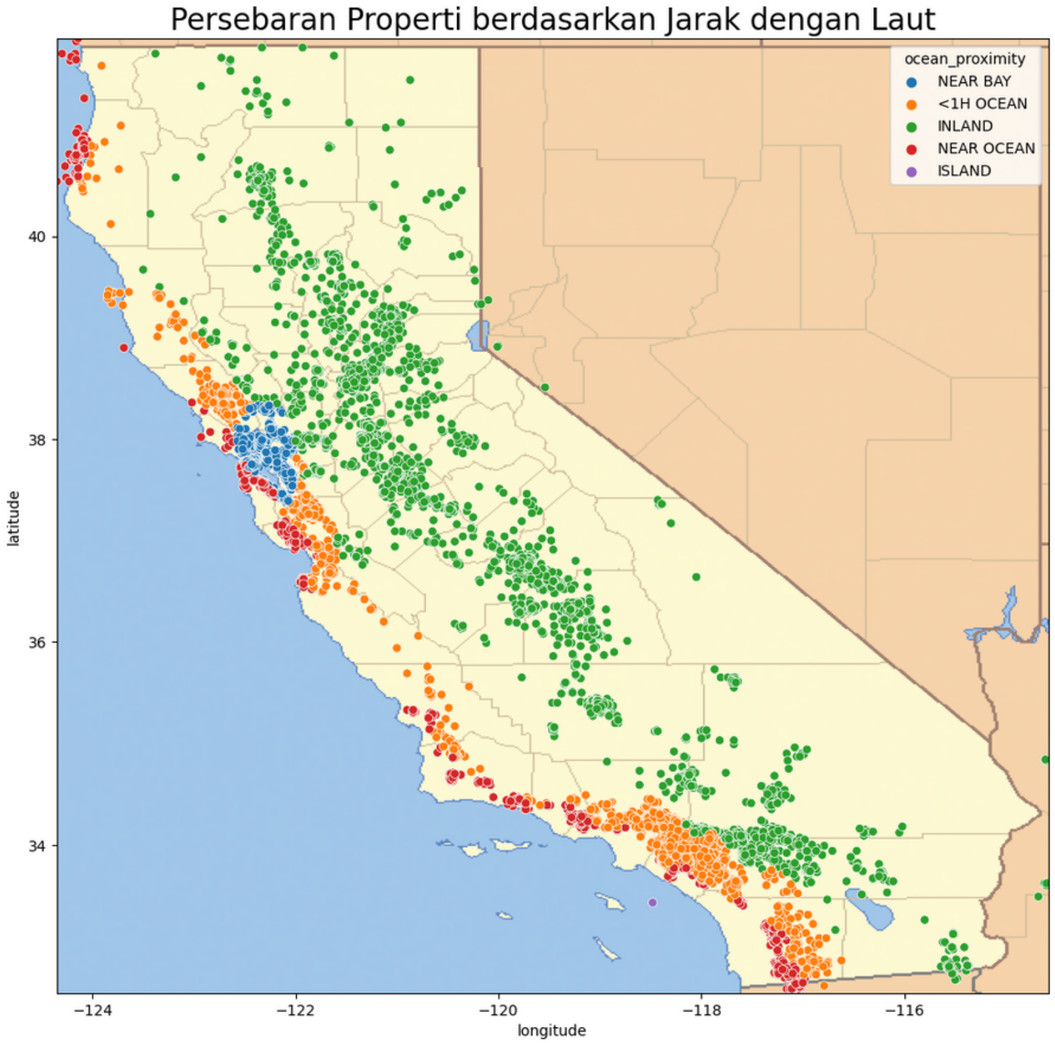
EXPLORATORY DATA ANALYSIS

It can be considered why properties that are in the Inland category/area are **more spread** compared to other areas in California can involve the following factors:

1. LAND AVAILABILITY

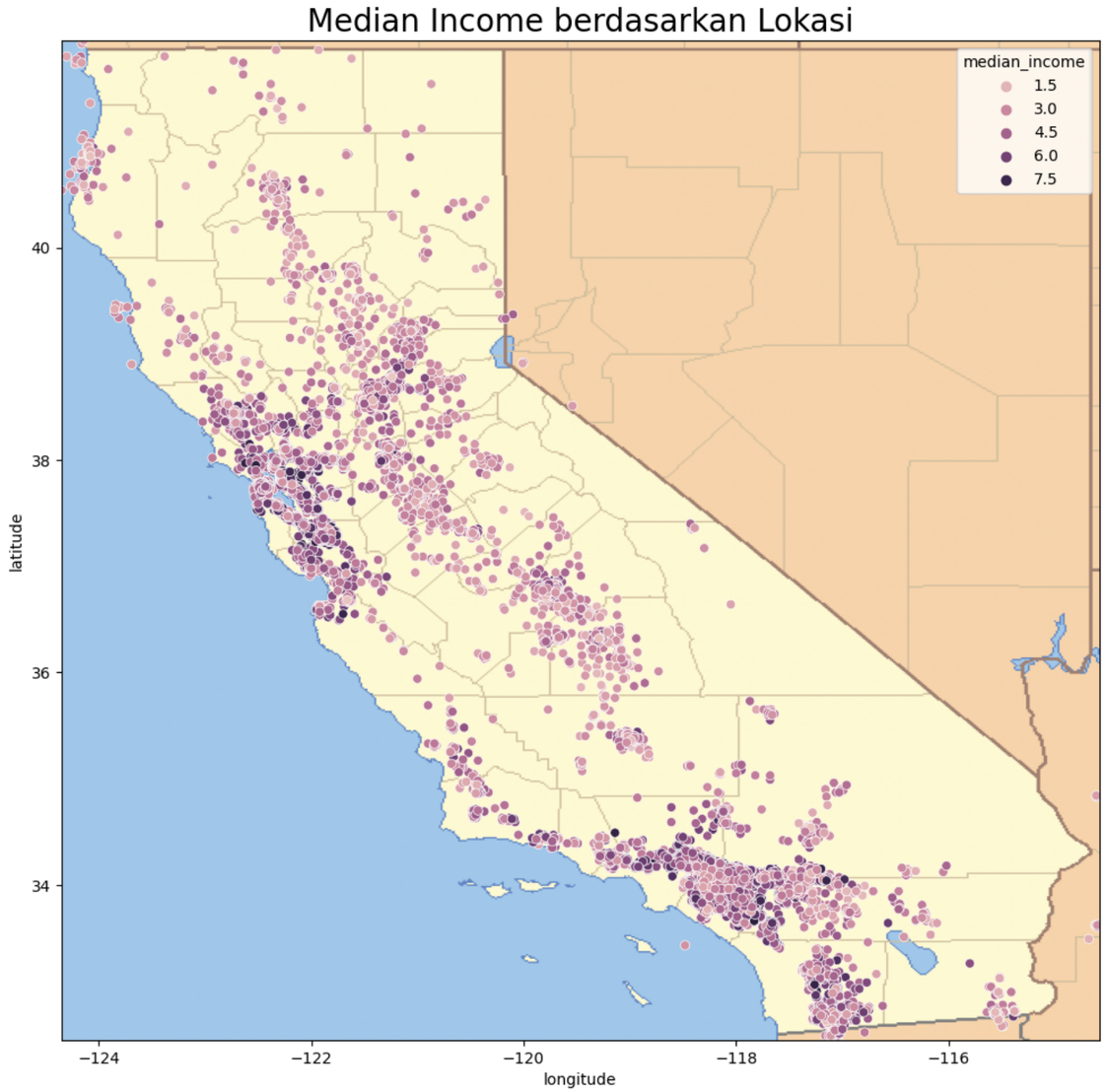
2. PROPERTY PRICE

3. ECONOMIC ACTIVITIES



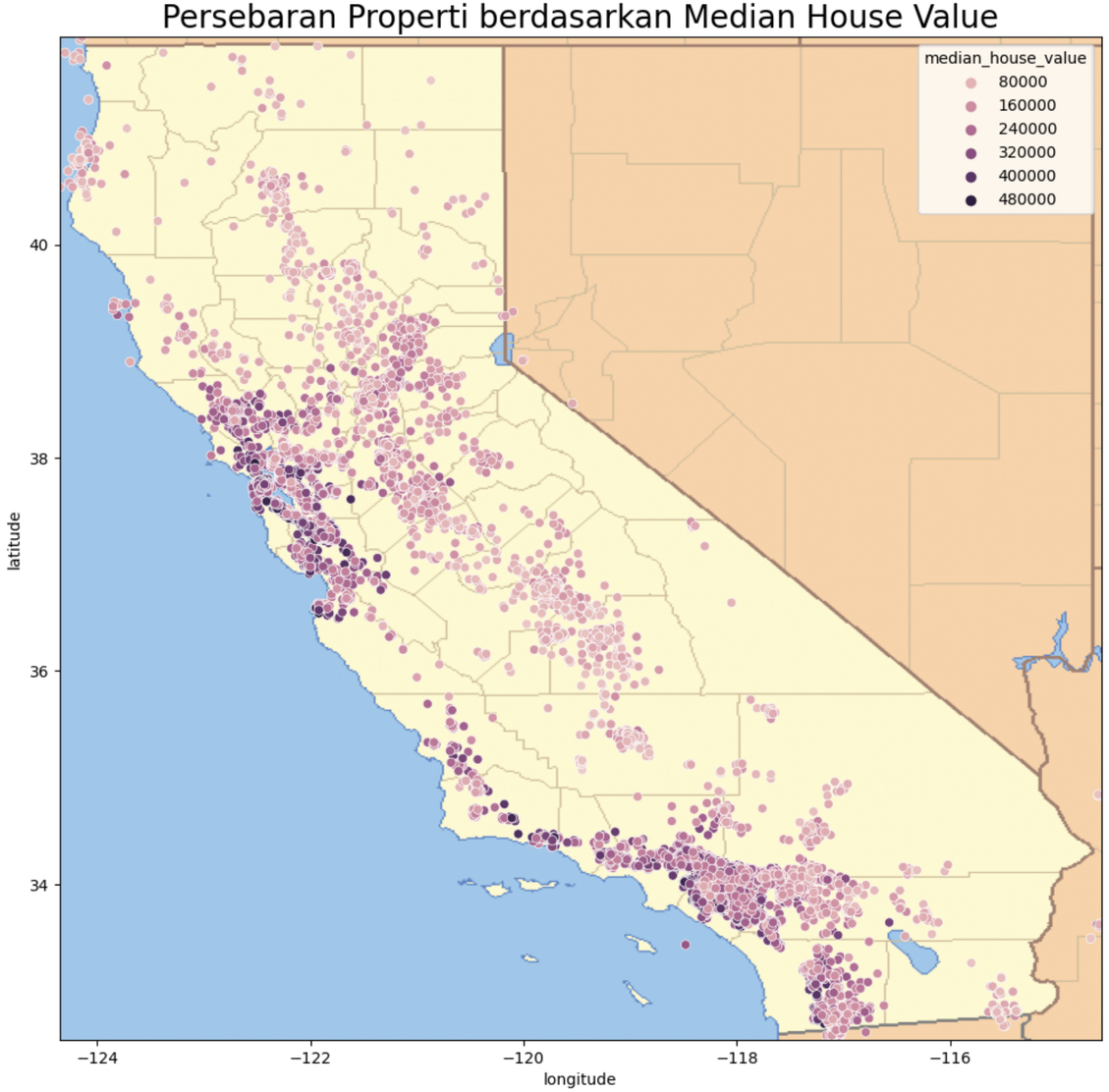
EXPLORATORY DATA ANALYSIS

It can be noted that the distribution of Median Income is quite similar to the distribution of Median House Value where **properties that have a high Median House Value** then the **Median Income around the location of the property is also high.**



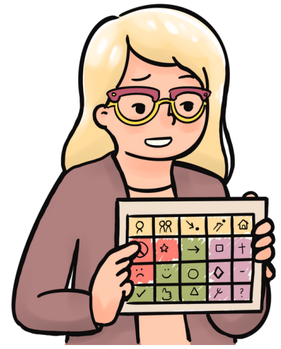
EXPLORATORY DATA ANALYSIS

It can be noted that **properties with higher prices are mostly located closer to the coast / close to the beach**. It can be seen in the visualization that **properties that have a location that is closer to the coast / close to the beach then the value is also increasing**.



dhrgroup.com

DATA MODELLING TRANSFORMER



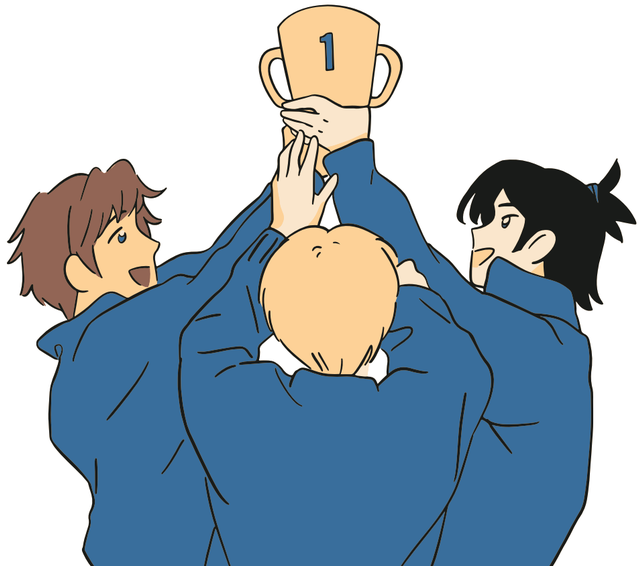
```
# transformer

pipeline_robust_scale = Pipeline([
    ('scaler', RobustScaler())
])

pipe_ordinal_scale = Pipeline([
    ('ordinal', OrdinalEncoder(cols=['housing_age_group', 'ocean_proximity'], mapping=ordinal_mapping)),
    ('scaler', RobustScaler())
])

transformer = ColumnTransformer([
    ('pipe_robust_scale', pipeline_robust_scale, ['median_income', 'person_in_house', 'bedroom_in_house', 'room_in_house']),
    ('pipe_ordinal_scale', pipe_ordinal_scale, ['housing_age_group', 'ocean_proximity'])
], remainder='passthrough')
```

MODEL BENCHMARKING

	MAE	RMSE	MAPE
Linear Regression	44847.727445	61089.419600	0.274478
Lasso	44847.645010	61089.414646	0.274477
Ridge	44847.633839	61089.274017	0.274487
Elastic Net	55574.046559	71753.367903	0.383646
KNN Regressor	42815.490909	59136.308628	0.249198
Decision Tree Regressor	54178.045455	75840.302640	0.308953
Random Forest Regressor	40718.320909	56084.346269	0.239095
XGBoost Regressor	41466.492056	57219.456818	0.242485
AdaBoost Regressor	53192.280152	65703.411518	0.367776
CatBoost Regressor	39454.269065	54688.782536	0.230751
SVR	75163.516045	95682.503686	0.507255

TOP 2

1. CATBOOST REGRESSOR
2. RANDOM FOREST REGRESSOR



MODEL PERFORMANCE COMPARISON

BEFORE HYPERPARAMETER TUNING

	MAE	RMSE	MAPE
CatBoost Regressor	39454.269065	54688.782536	0.230751

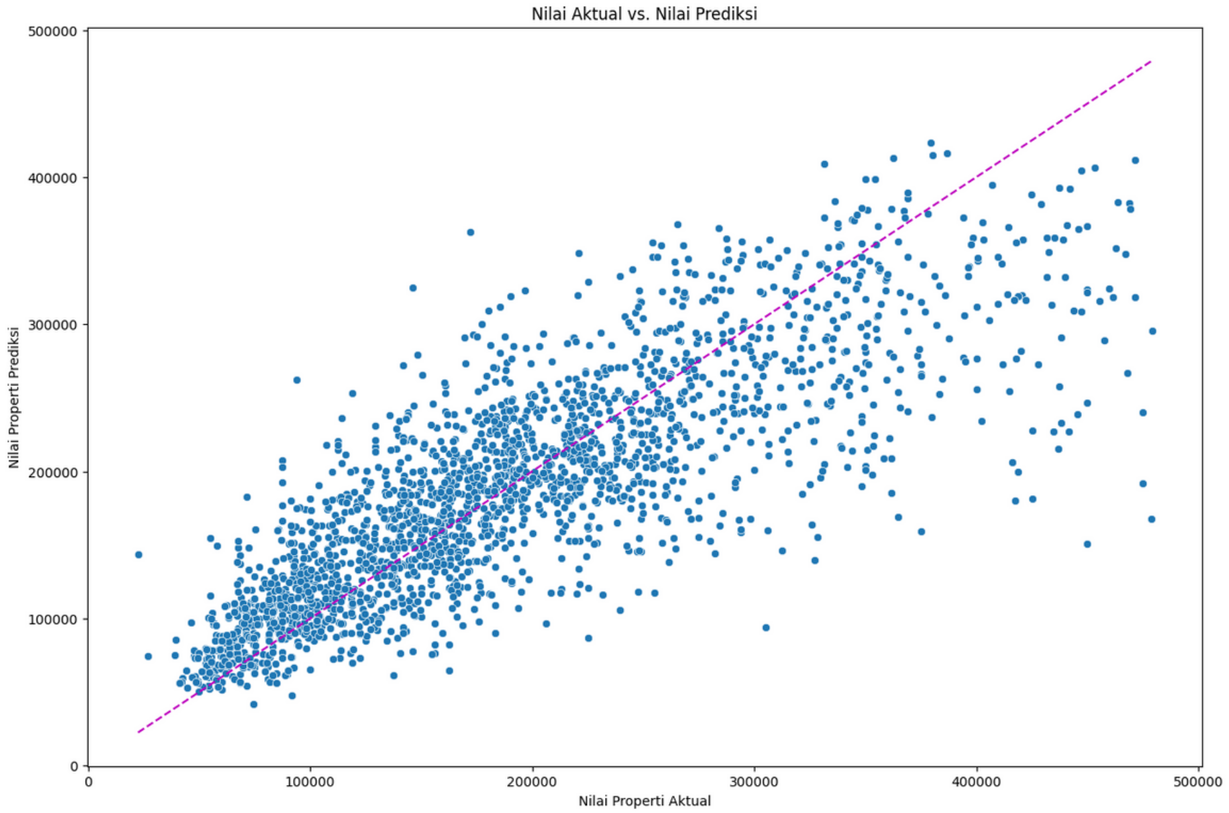
	MAE	RMSE	MAPE
Random Forest Regressor	40718.320909	56084.346269	0.239095

AFTER HYPERPARAMETER TUNING

	MAE	RMSE	MAPE
CatBoost Regressor	39113.078841	54223.162189	0.228225

	MAE	RMSE	MAPE
Random Forest Regressor	40013.198863	55417.224428	0.233952

PREDICTED VS ACTUAL



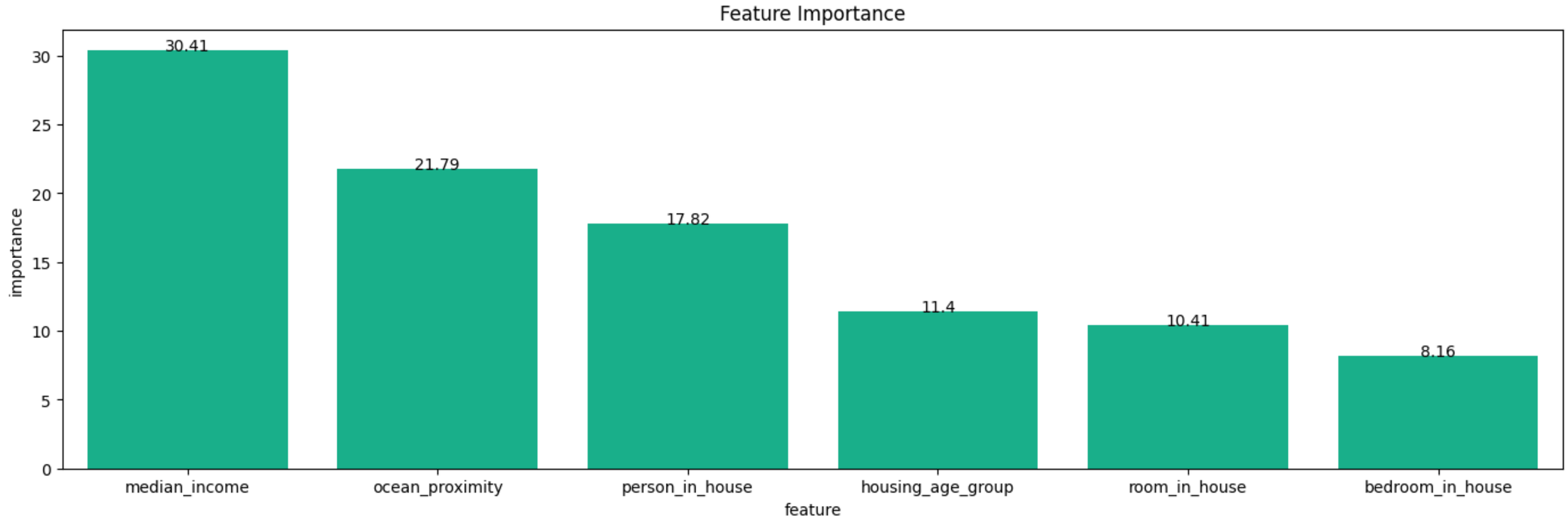
The **data simply forms a linear gradient pattern**.

The **Predicted Value quite close to the Actual Value** to arrive at a property with a **price of around '300,000'**.

Models can **work well enough** to predict property prices up to properties that have a **price of less than '250,000'**.

The **MAPE** value obtained ranges from **'22%'** making this model can be categorized into **'Reasonable Forecasting'**. (Lewis, 1982)

FEATURE IMPORTANCE

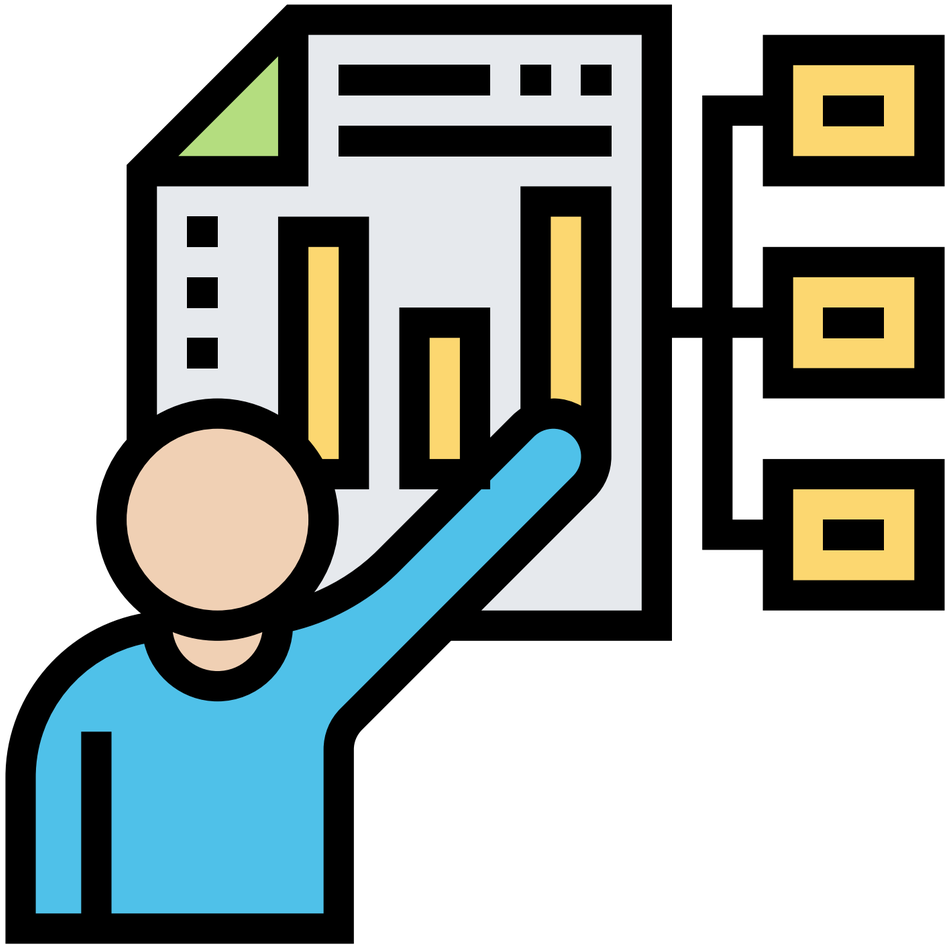


CONCLUSION

- The **'median_income'** and **'ocean_proximity'** features are the **most influential features** on **'median_house_value'**.

- The selected model is **'CatBoost Regressor'**.

- Judging from the value of the **'MAPE'** evaluation metrics that have been generated by the model by **22%**, it can be concluded that if the model is used to estimate property prices in California in the range of values according to those trained on the model, then the **average predicted price can miss approximately '22%'** of the actual price.



RECOMMENDATION

- UPDATE DATA TO THE LATEST CONDITION
- ADD MORE SPECIFIC FEATURES FOR EACH DATA
- DO EXPLORATION WITH HYPERPARAMETER TUNING AND A WIDE VARIETY OF OTHER MACHINE LEARNING MODELS

Dream House Realty Group

THANK YOU!

Contact Us

+62-8222-1155-180

dhrgroup.com

455 N. Rexford Drive
Beverly Hills CA 90210
United States