TIM BEBAS

TELKOM UNIVERSITY

GLORIA NATASYA IRENE SIDEBANG
FAHMI AGUNG MAULANA
MUHAMMAD RAMDHAN FITRA HIDAYAT

DAFTAR ISI

BUSINESS UNDERSTANDING

Faktor

Penentuan faktor utama yang memengaruhi harga rumah

Harga Rumah

Mengetahui harga rumah sesuai dengan faktor penentu

TUJUAN :

Memprediksi harga rumah dengan **menganalisis** faktor utama yang memengaruhi menggunakan **model machine learning** yang **paling tepat**

TEKNIK :

Menggunakan model **Random Forest**

INDIKATOR KEBERHASILAN :

Akurasi model >80%

DATA UNDERSTANDING

ATRIBUT

Numerik Feature :

- price_in_rp (interval)
- lat (interval)
- long (interval)
- bedrooms (rasio)
- bathrooms (rasio)
- land_size_m2 (rasio)
- building_size_m2 (rasio)
- carports (rasio)
- maid_bedrooms (rasio)
- maid_bathrooms (rasio)
- building_age (rasio)
- year_built (interval)
- garages (rasio)

Total : 13 Features (3 interval + 10 rasio)

Kategorik Feature :

- url (nominal)
- title (nominal)
- address (nominal)
- district (nominal)
- city (nominal)
- facilities (nominal)
- property_type (nominal)
- ads_id (nominal)
- certificate (nominal)
- electricity (nominal)
- floors (nominal)
- property_condition (nominal)
- building_orientation (nominal)
- furnishing (nominal)

Total : 14 Features

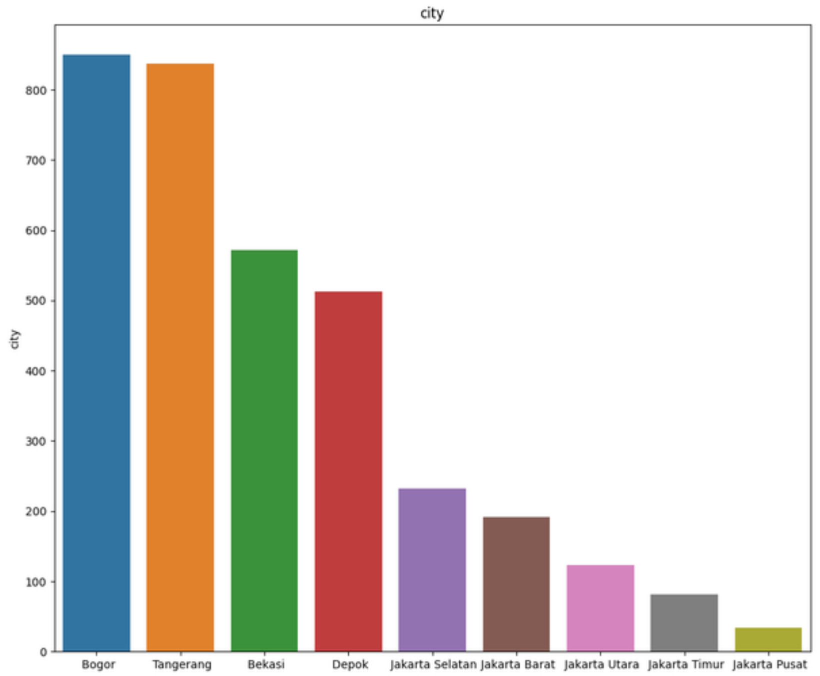
DATA UNDERSTANDING

Unique Value Counts In Each Column

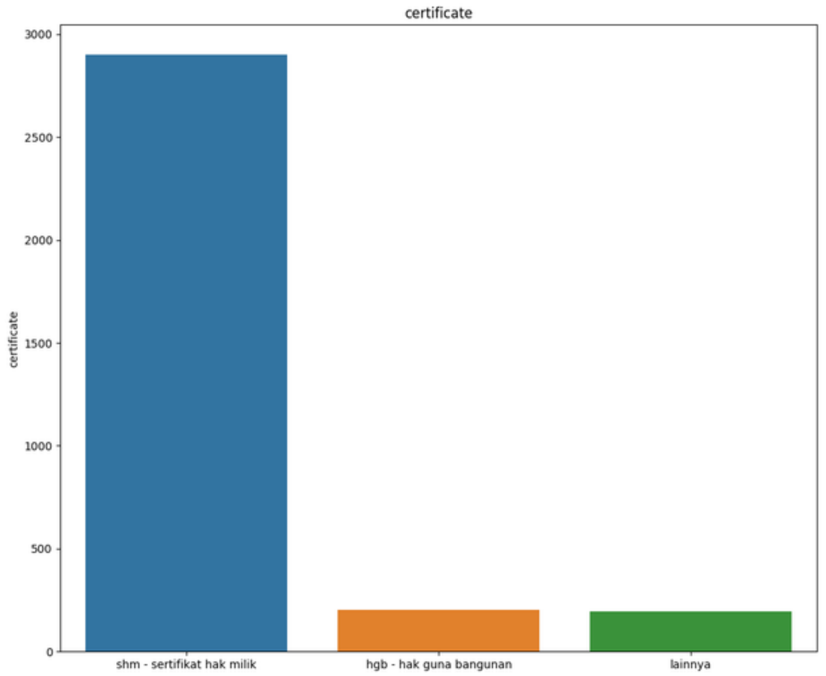
	Unique Value Count
url	3435
price_in_rp	660
title	3341
address	397
district	380
city	9
lat	389
long	390
facilities	2004
property_type	1
ads_id	3434
bedrooms	22
bathrooms	22
land_size_m2	481
building_size_m2	358
carports	13
certificate	3
electricity	29
maid_bedrooms	8
maid_bathrooms	6
floors	5
building_age	42
year_built	46
property_condition	5
building_orientation	8
garages	11
furnishing	3

Nilai unik untuk setiap feature

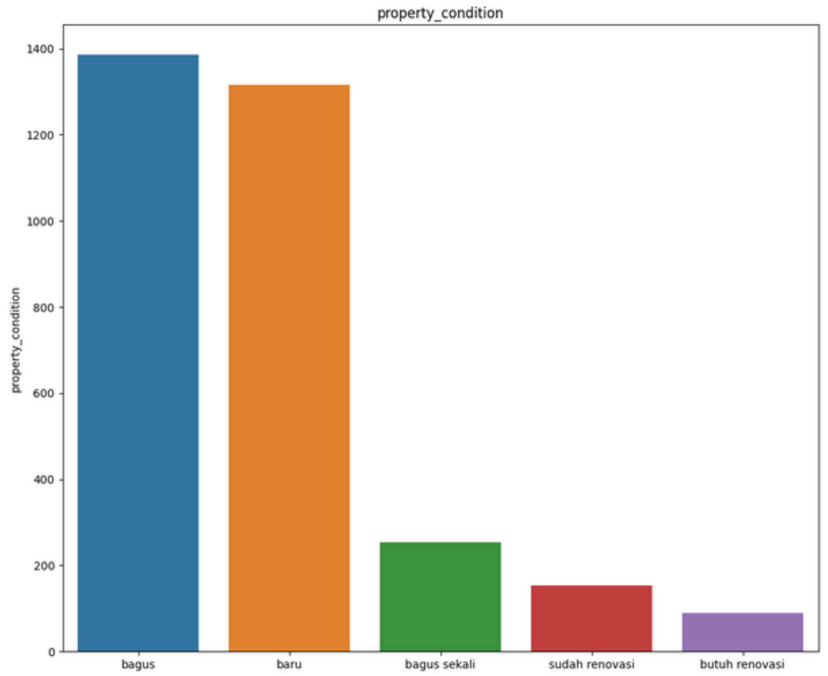
DATA UNDERSTANDING



Histogram plot feature **city**

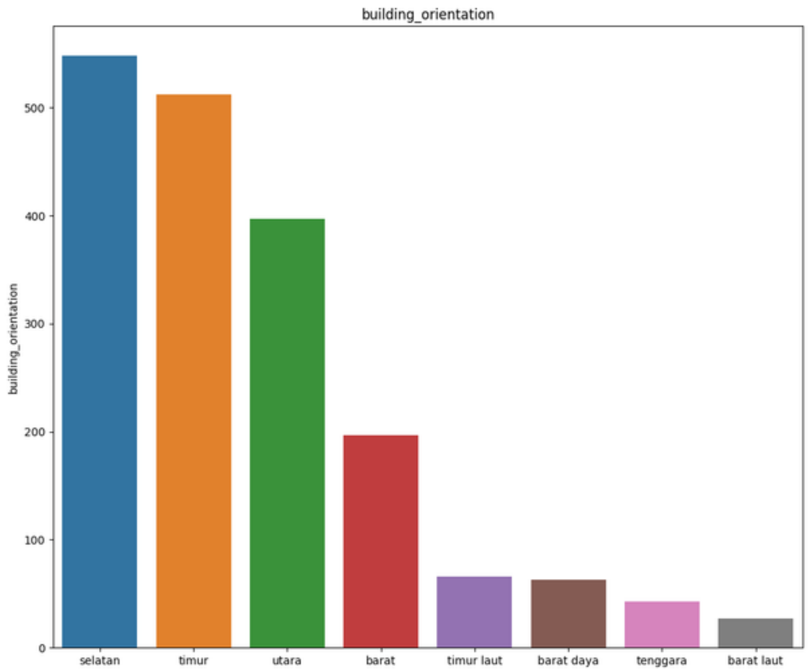


Histogram plot feature **certificate**

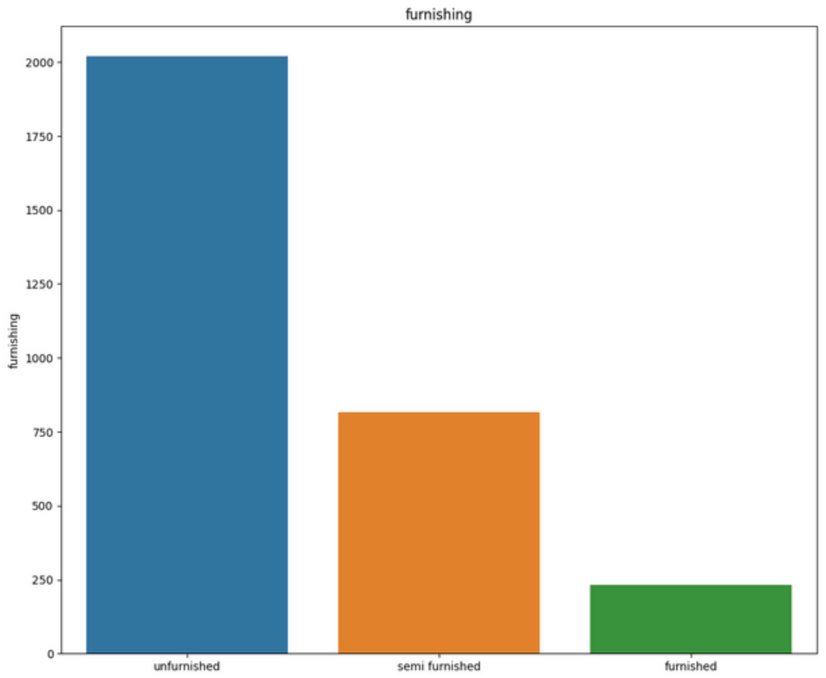


Histogram plot feature **proprrty_condition**

Beberapa histogram plot untuk ketagorik feature



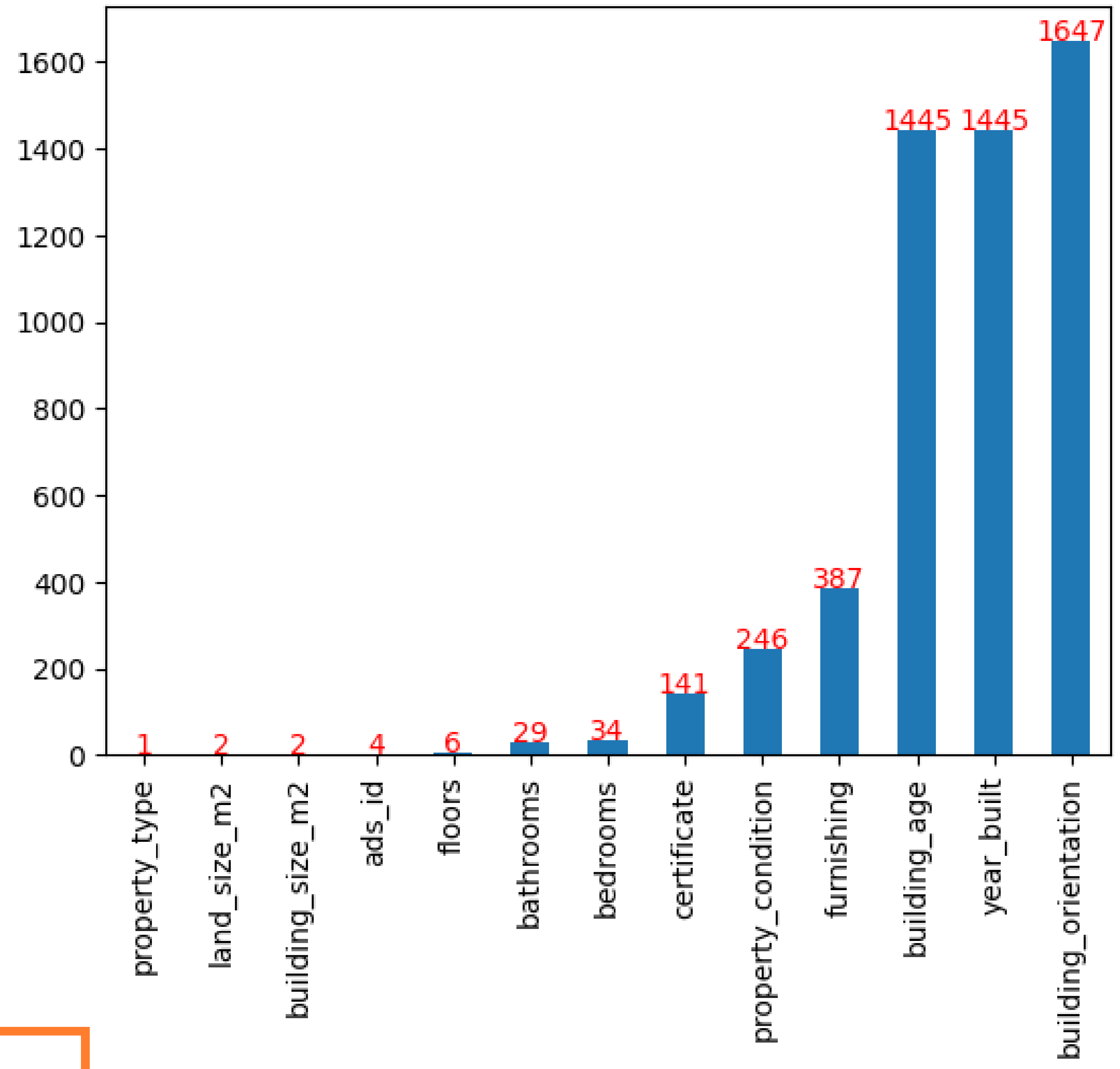
Histogram plot feature **building orientation**



Histogram plot feature **furnishing**

DATA PREPARATION

PENGECEKAN & HANDLING MISSING VALUE



Plot Jumlah Niali Hilang untuk Setiap Feature

Metode Handling:

- **Drop** feature with **missing value > 1/3** of total record
- **Numeric** missing value handle with **KNNImputer**
- **Categorical** missing value handle with **mode**

DATA PREPARATION

DATA DUPLICATE

Terdapat 115 record duplikat berdasarkan 'title', 'address', 'district','price_in_rp','electricity', 'bedrooms', 'bathrooms','floors','year_built'

	url	price_in_rp	title	address	district	city	lat	long	facilities	property_type	...
99	https://www.rumah123.com/properti/bekasi/hos11...	2.150000e+09	Di Jual Rumah Siap Huni di Cluster Asera Harap...	Harapan Indah, Bekasi	Harapan Indah	Bekasi	-6.181752	106.973684	AC	rumah	...
100	https://www.rumah123.com/properti/bekasi/hos11...	2.150000e+09	Di Jual Rumah Siap Huni di Cluster Asera Harap...	Harapan Indah, Bekasi	Harapan Indah	Bekasi	-6.181752	106.973684	AC	rumah	...

Contoh record yang memilik kesamaan nilai pada beberapa fitur

DATA PREPARATION

DATA DUPLICATE

Terdapat 3 duplikat data berdasarkan ads_id

```
# cek data duplikat berdasarkan ads id
sum_duplicated = data.duplicated(subset = ['ads_id']).sum()
print(f"Terdapat {sum_duplicated} record duplikat berdasarkan ads_di")
```

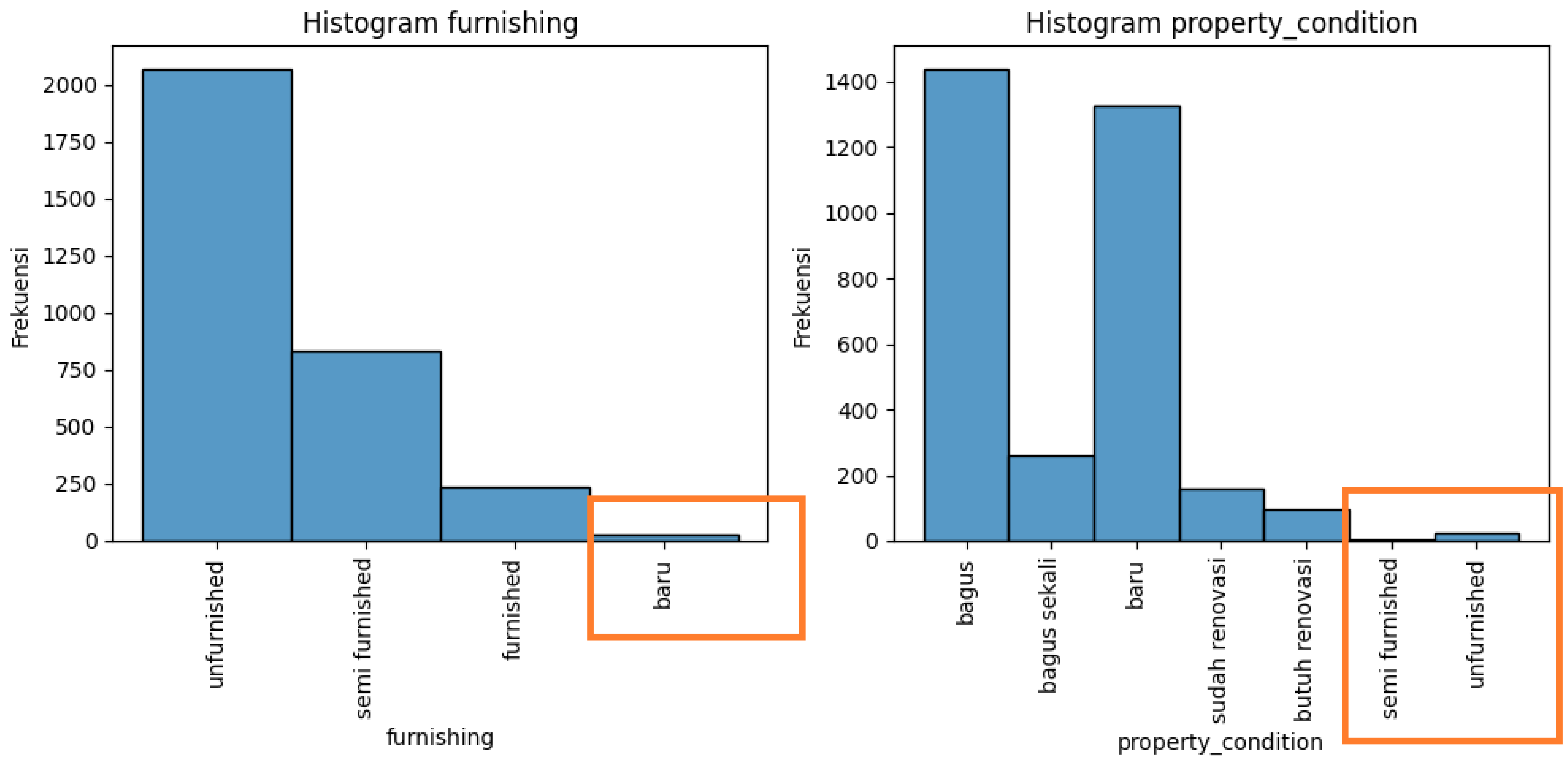
```
Terdapat 3 record duplikat berdasarkan ads_di
```

```
data.drop_duplicates(subset = ['ads_id'], inplace=True)
```

Jumlah data duplikat berdasarkan ads_id

DATA PREPARATION

RECORD TERBALIK



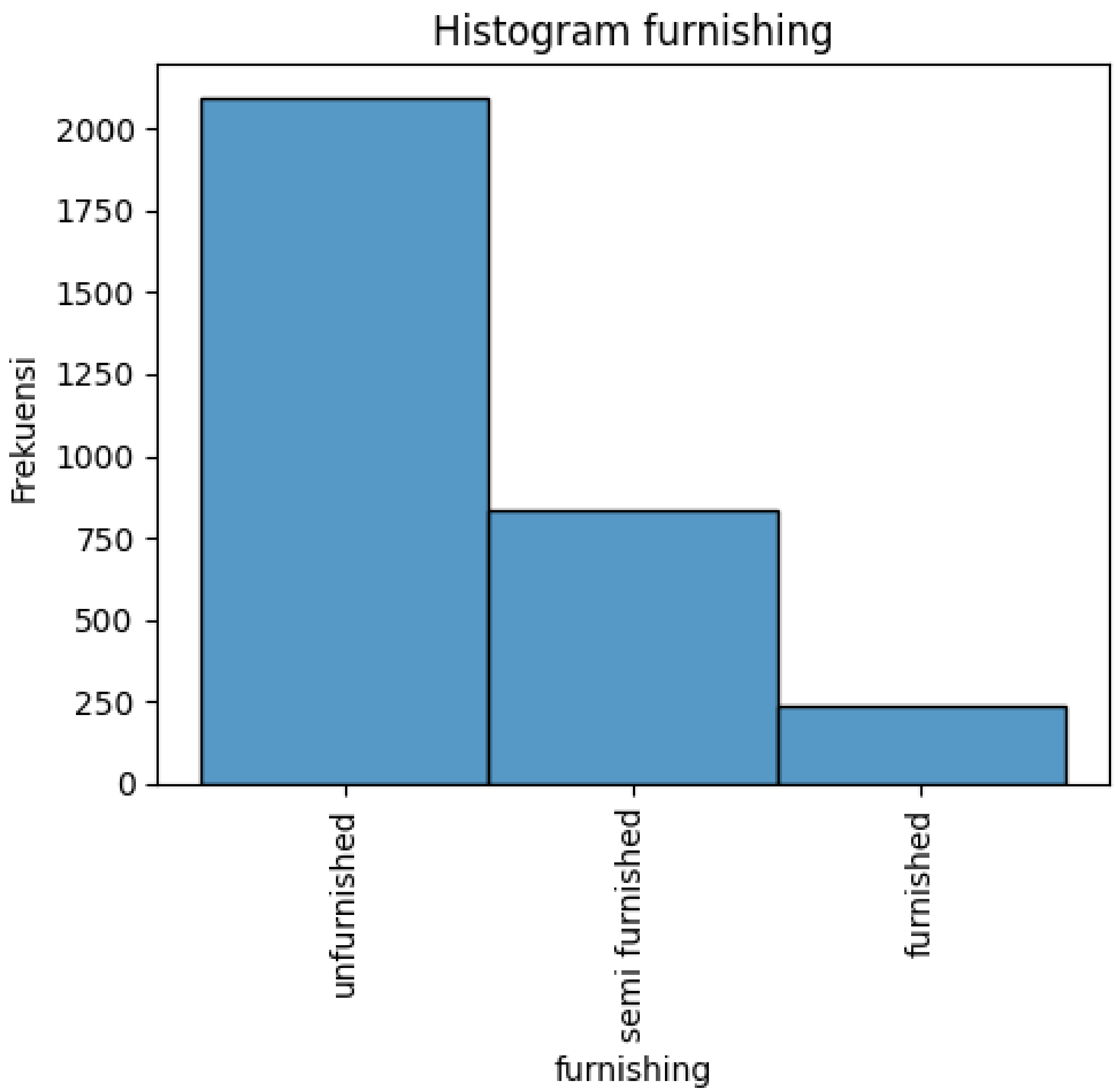
Histogram plot feature furnishing dan property_condition

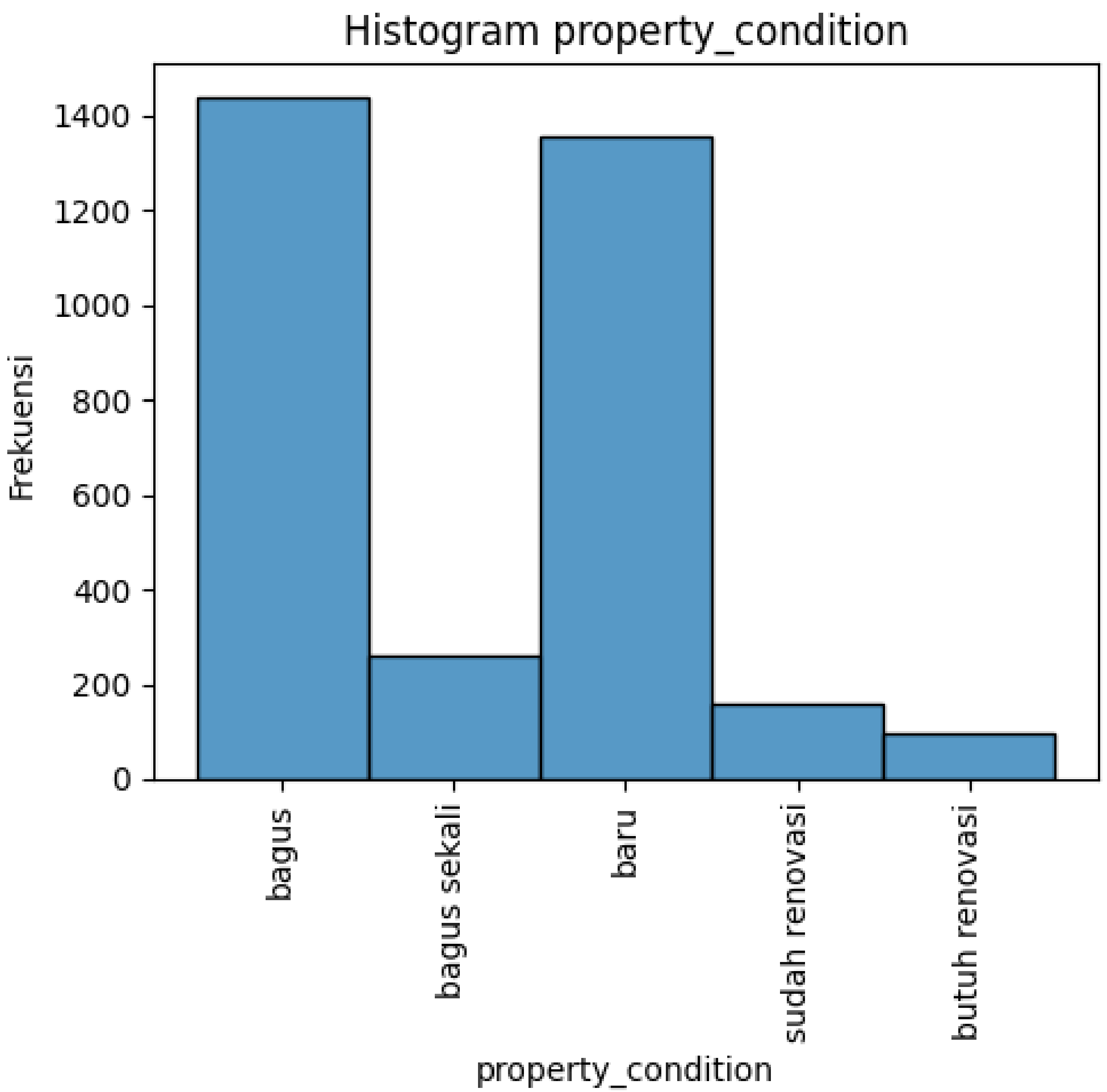
Tedapat data yang pengisiannya terbalik, sehingga dilakukan pertukaran pada data yang terbalik

DATA PREPARATION

RECORD TERBALIK

Hasil dari pertukaran data

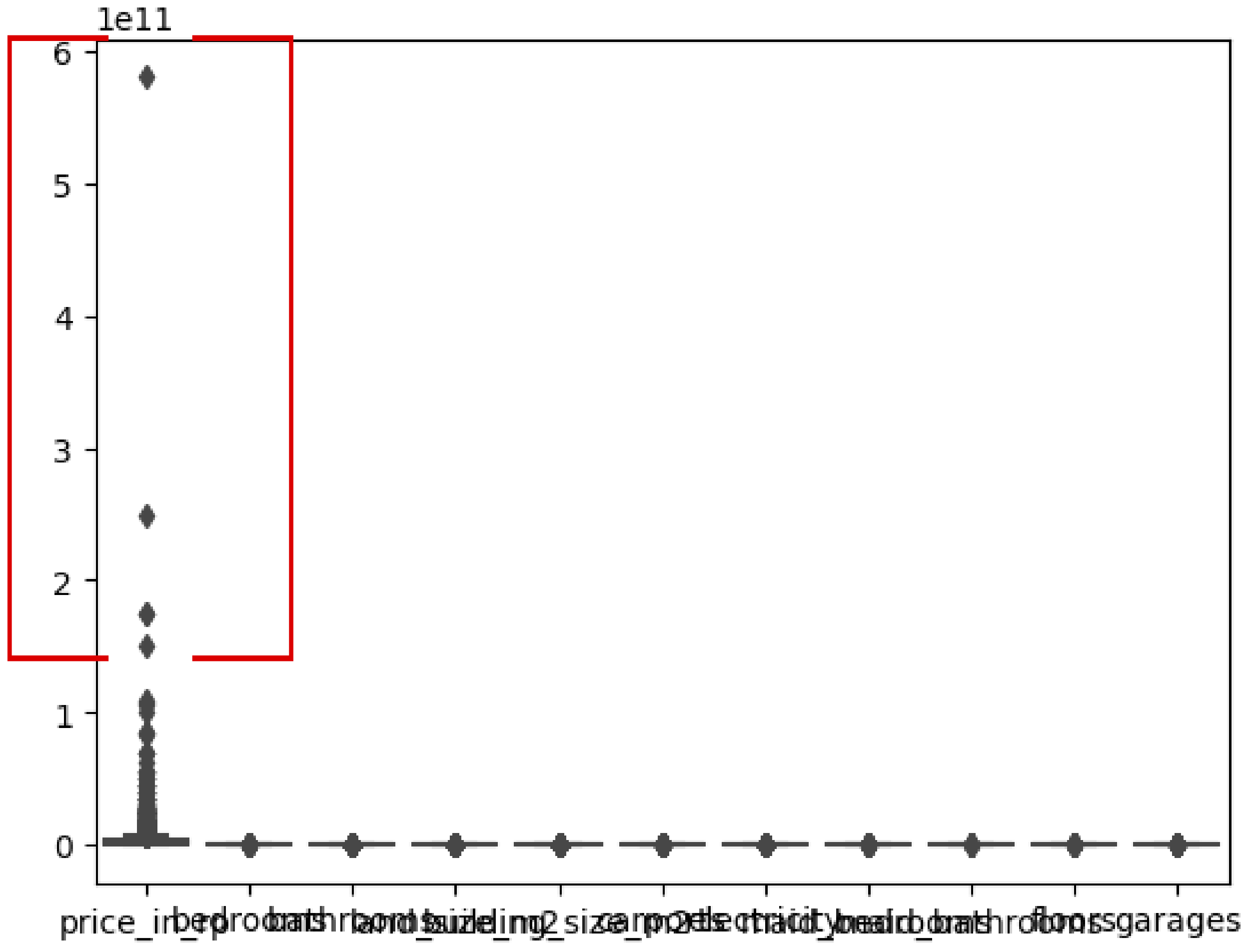




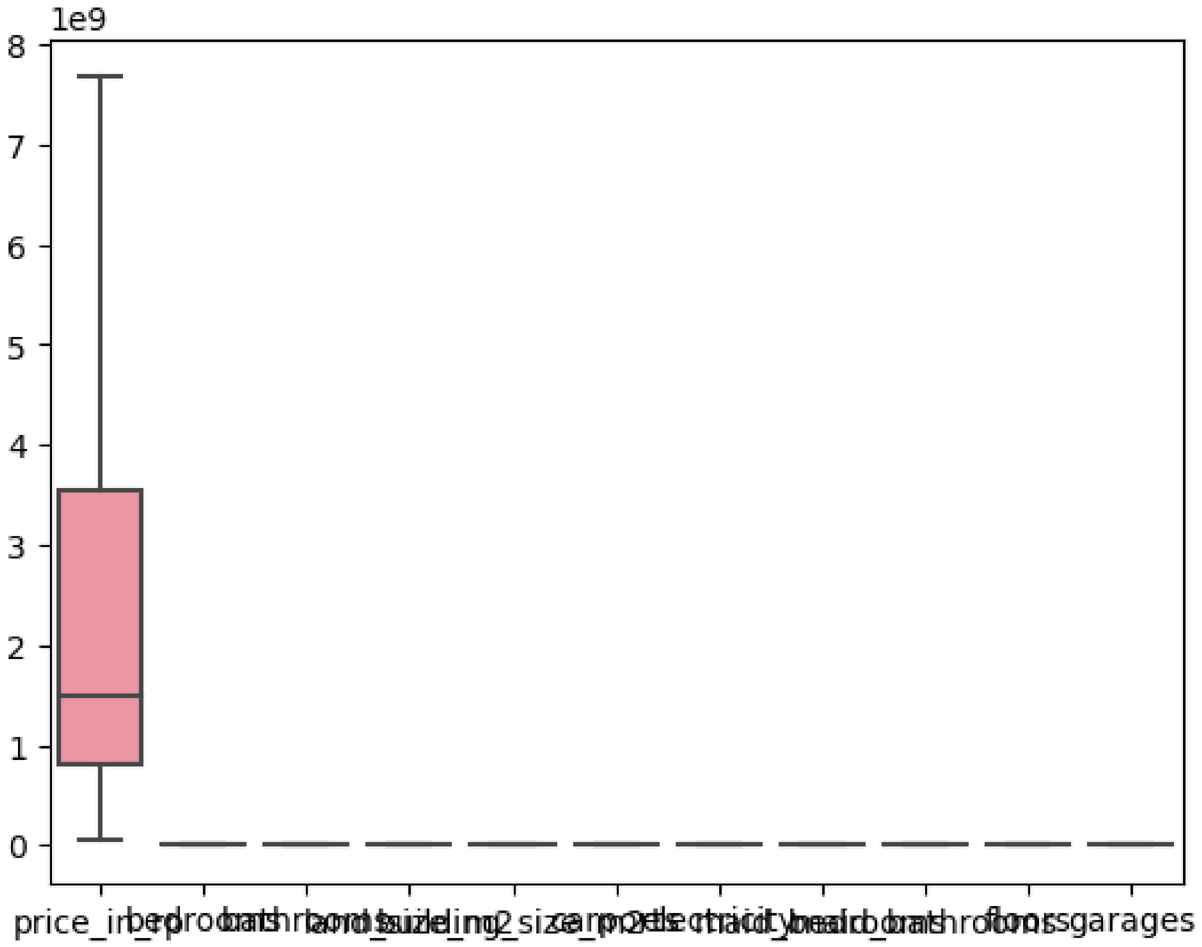
DATA PREPARATION

OUTLIER

Terdeteksi adanya **outlier pada price_in_rp**



Boxplot Numerik Feature



Boxplot Numerik Feature setelah dilakukan handling outlier

DATA PREPARATION

OUTLIER

Mengatasi outlier menggunakan **teknik winsorize**

```
def winsorize_column_iqr(df, column, multiplier):
    q1 = df[column].quantile(0.25)
    q3 = df[column].quantile(0.75)
    iqr = q3 - q1
    lower_limit = q1 - multiplier * iqr
    upper_limit = q3 + multiplier * iqr
    df[column] = np.where(df[column] < lower_limit, lower_limit, df[column])
    df[column] = np.where(df[column] > upper_limit, upper_limit, df[column])
    return df
```

DATA PREPARATION

HIGH CARDINALITY

Unique Value Counts In Each Column

	Unique Value Count
url	3435
price_in_rp	660
title	3341
address	397
district	380
city	9
lat	389
long	390
facilities	2004
property_type	1
ads_id	3434
bedrooms	22
bathrooms	22

Drop kolom yang memilik nilai unik yang banyak **(high cardinality)** dan yang hanya memiliki **1 nilai unik**

```
# drop kolom 'url', 'title' , 'property_type', 'address','lat','long','ads_id', facilities'
list_to_drop = ['url', 'property_type', 'address','ads_id','lat','long','facilities']
data.drop(list_to_drop, axis=1, inplace = True )
```

DATA PREPARATION

KONVERSI DATA

mengubah tipe data atribut electricity menjadi numerik

electricity
2200 mah
2200 mah
2200 mah
lainnya mah

```python
# mengubah tipe data dari feature electricity
data['electricity'] = data['electricity'].str.slice(stop=-4)
data['electricity'] = data['electricity'].replace('lainnya', np.nan)
data['electricity'] = pd.to_numeric(data['electricity'])

unique_values = data['electricity'].unique()
print(f"Fitur 'electricity': {unique_values}")
```

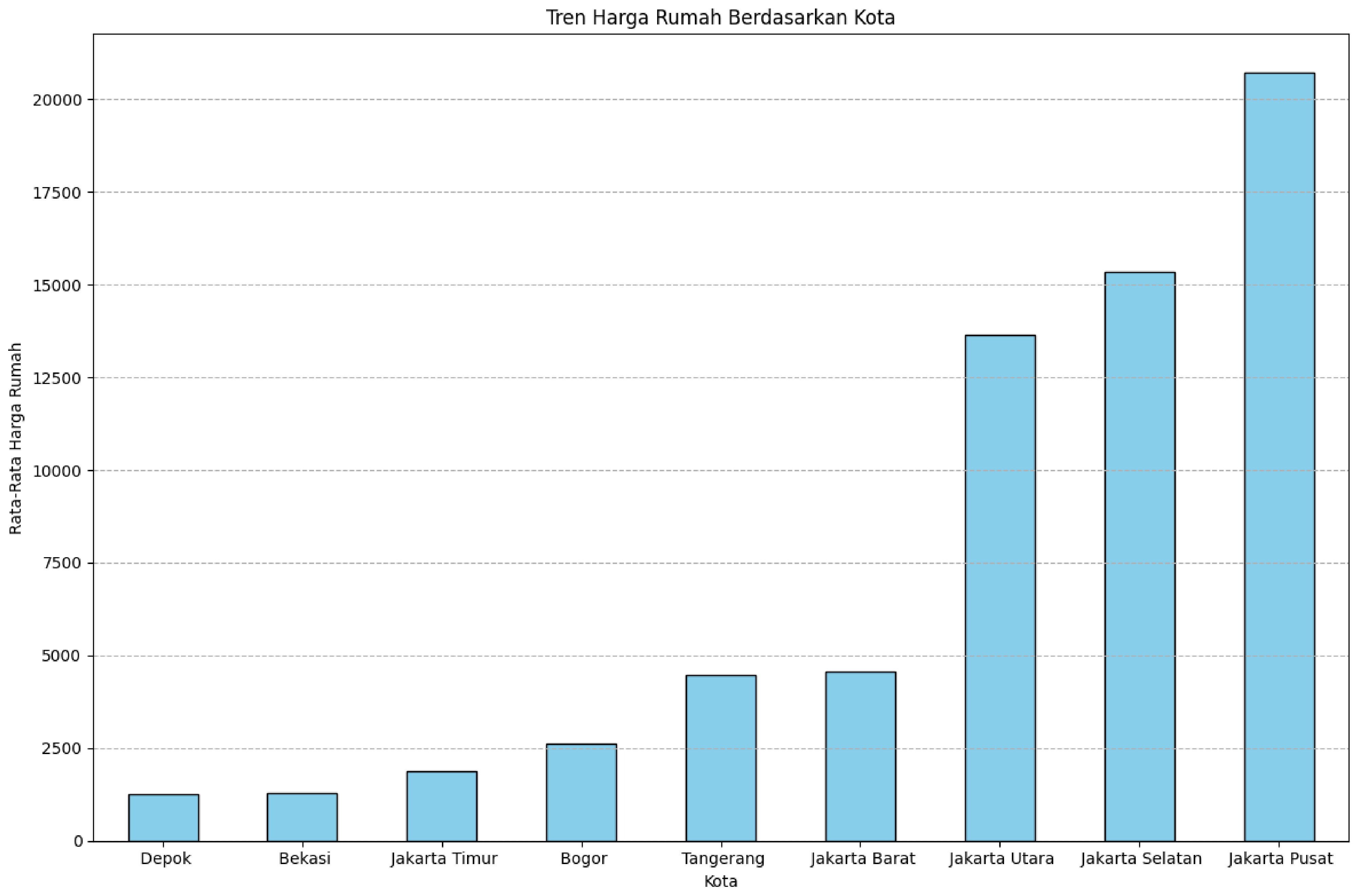
```
Fitur 'electricity': [ 4400.  2200.  3500.  1300.    nan  5500.  6600.  7700.  3300.  7600.
 10600.   900. 47500. 11000.  8000.   450. 10000. 53000. 16500. 13200.
 13900. 17600. 23000. 41500. 12700. 13300. 33000. 24000. 22000.  9500.]
```

EXPLORATORY DATA ANALYSIS

DATA INSIGHT

TREN HARGA RUMAH

berdasarkan lokasi, tren harga rumah **cenderung meningkat semakin dekat dengan pusat kota**

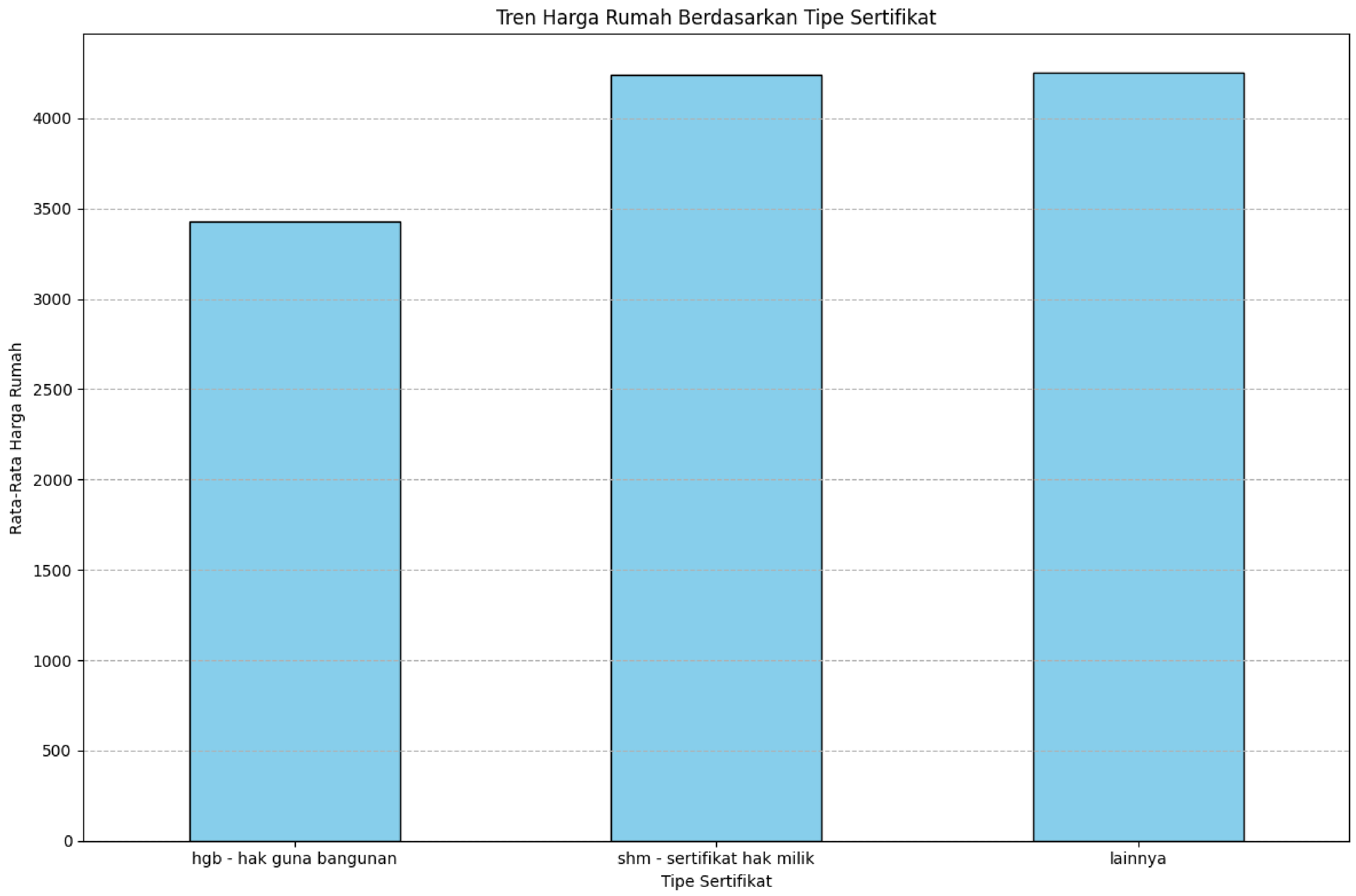


EXPLORATORY DATA ANALYSIS

DATA INSIGHT

TREN HARGA RUMAH

Harga rumah dengan sertifikat **SHM(Sertifikat Hak Milik) meniliki harga jual yang lebih mahal** dibanding dengan rumah bersertifikat HGB(Hak Guna Bangunan)



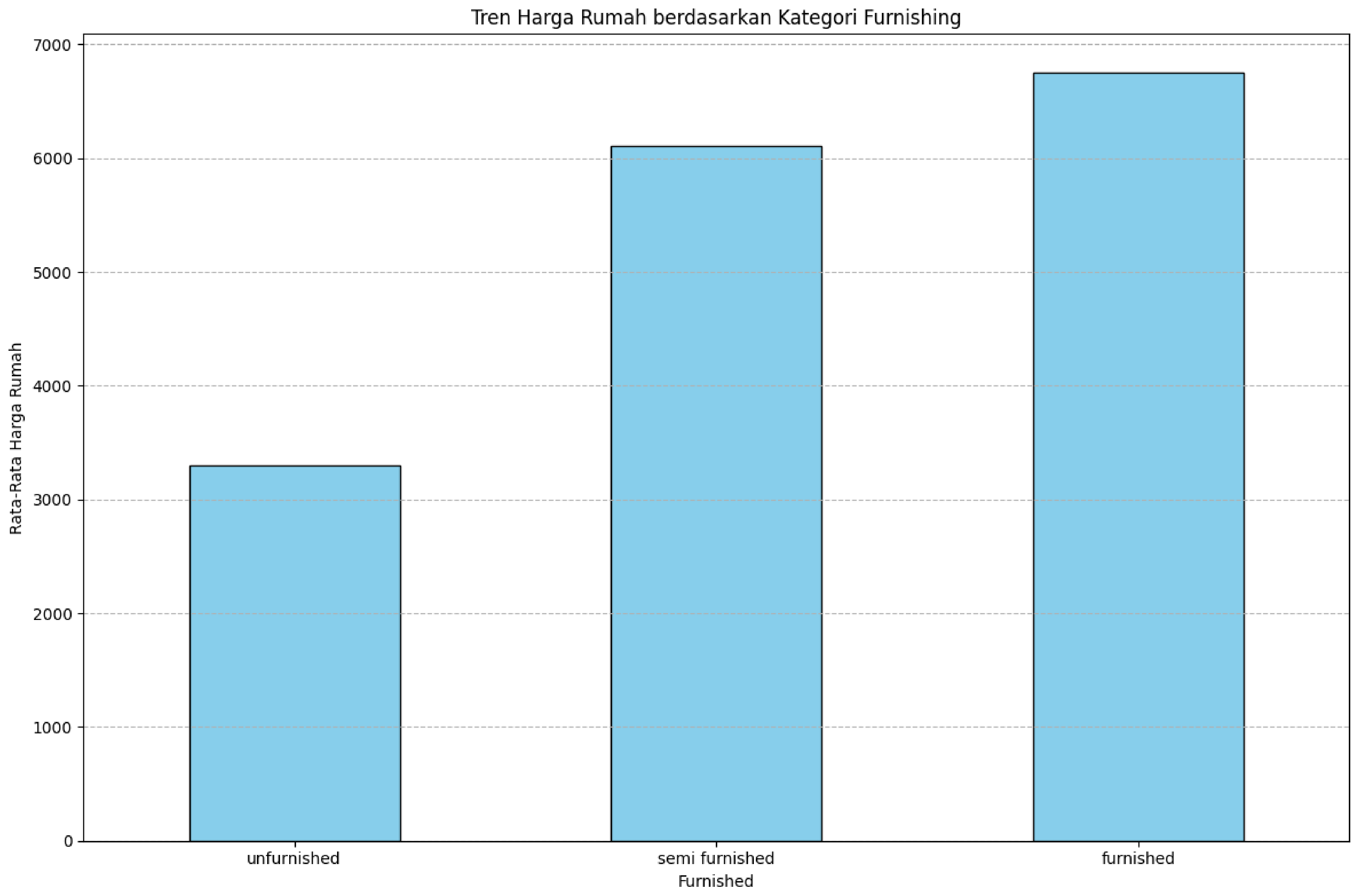
EXPLORATORY DATA ANALYSIS

DATA INSIGHT

TREN HARGA RUMAH

Harga rumah bertipe **furnished** memiliki harga yang **paling mahal**.

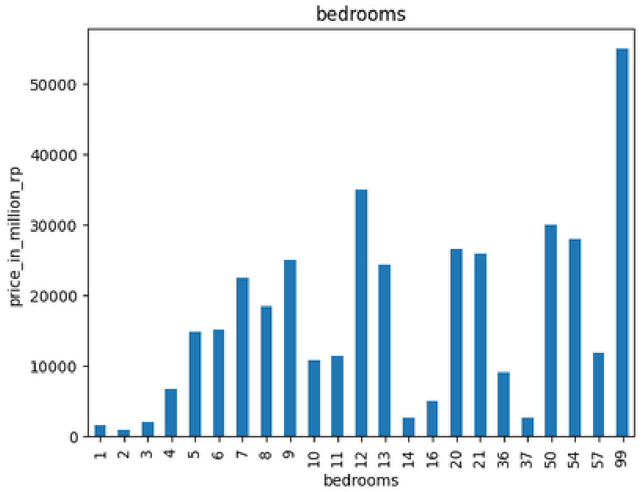
Terdapat **kenaikan harga yang signifikan** antar rumah bertipe unfurnished dan semi furnished

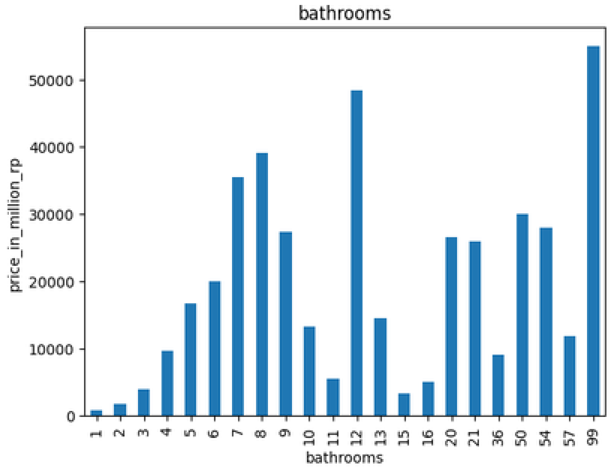


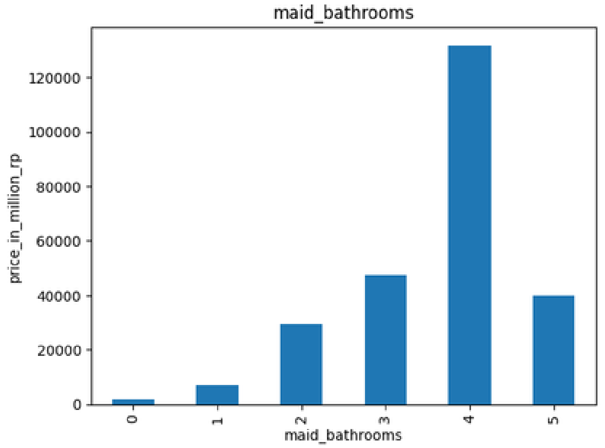
EXPLORATORY DATA ANALYSIS

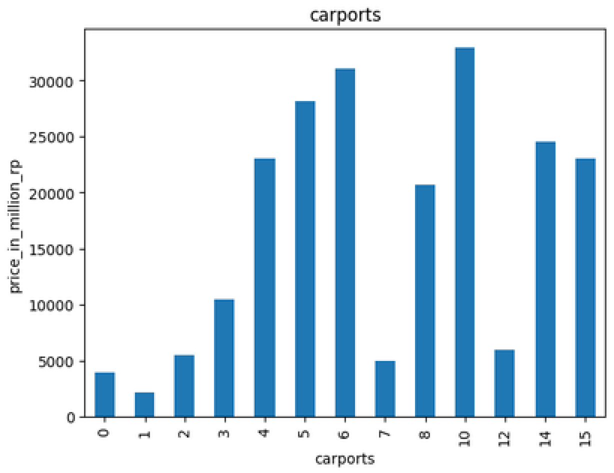
DATA INSIGHT

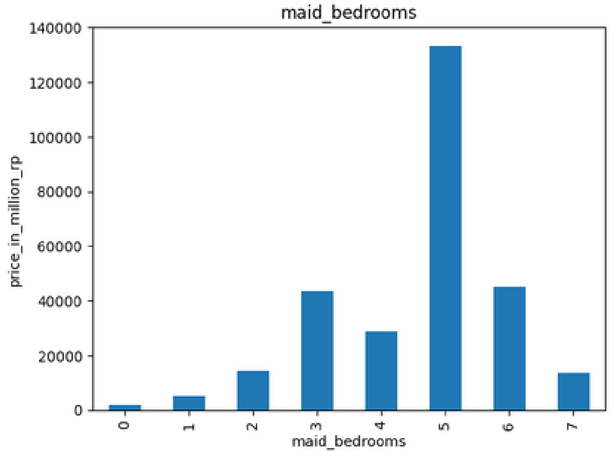
- Feature 'bedrooms', 'bathrooms', 'maid_bedrooms', 'maid_bathrooms', 'carports', 'garages' dengan harga rumah **tidak memiliki hubungan yang linear.**
- **Tren harga cenderung fluktuatif** untuk jumlah 'bedrooms', 'bathrooms', 'maid_bedrooms', 'maid_bathrooms', 'carports', 'garages' diatas 4.
- **Hubungan linear terlihat pada feature 'floors'** dengan harga rumah sehingga dapat disimpulkan salah satu faktor yang memengaruhi harga rumah adalah tingkat bagunan

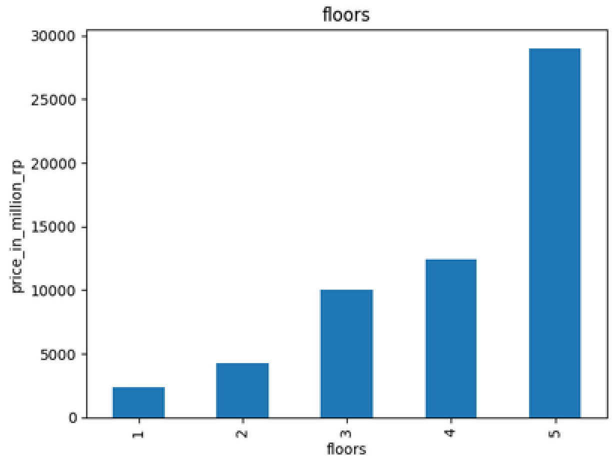




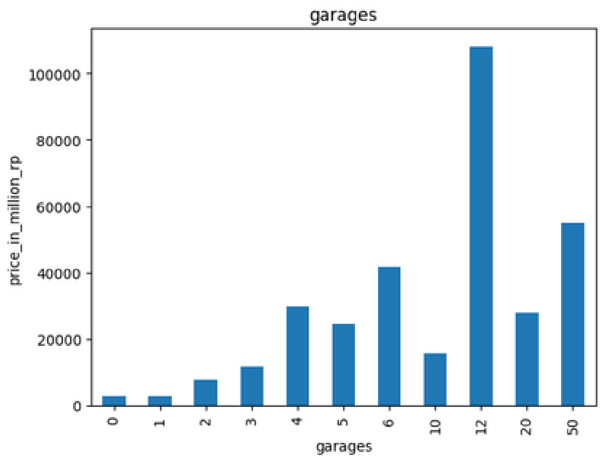








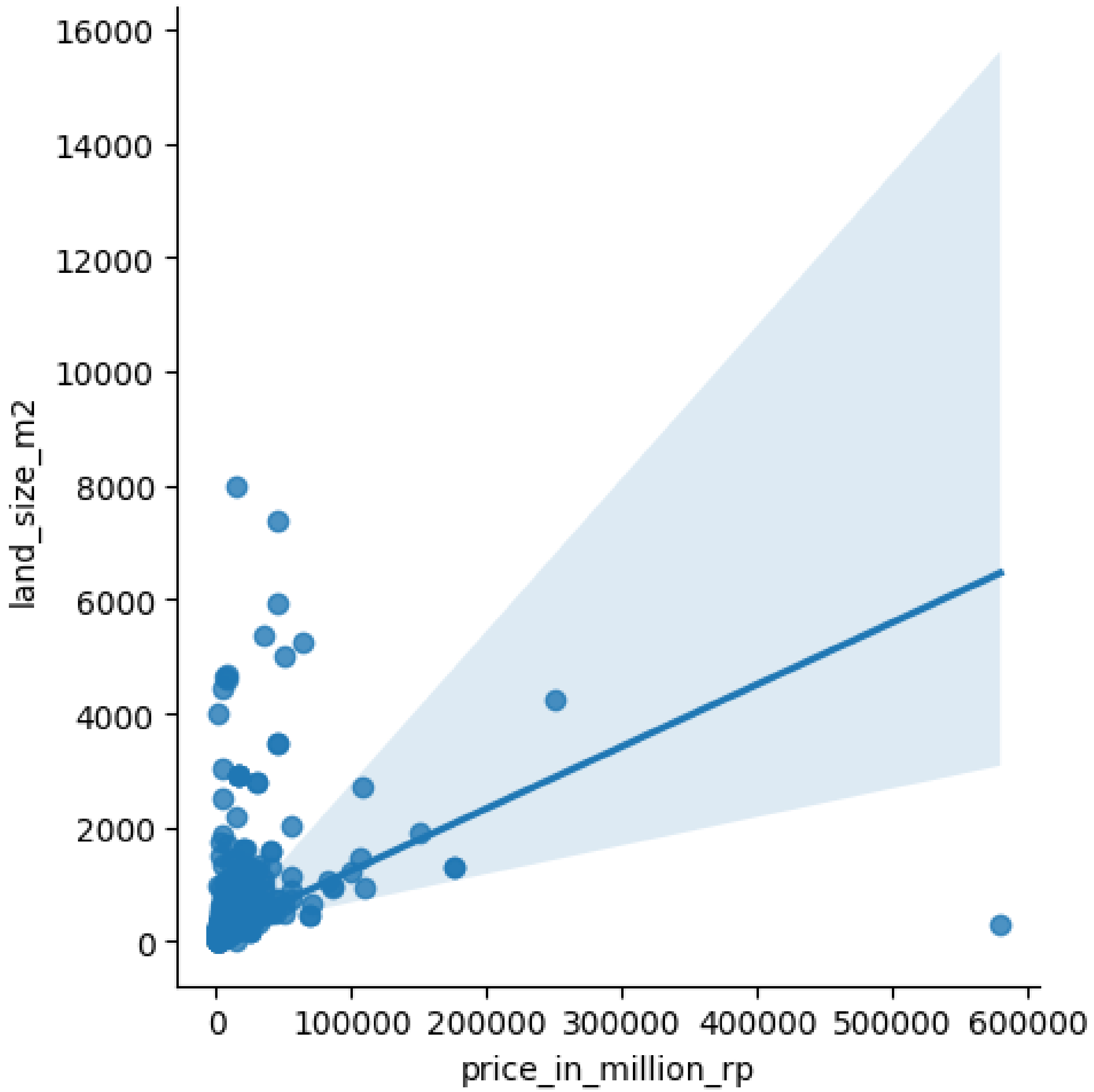
Beberapa plot feature kategorik dengan price_in_rp



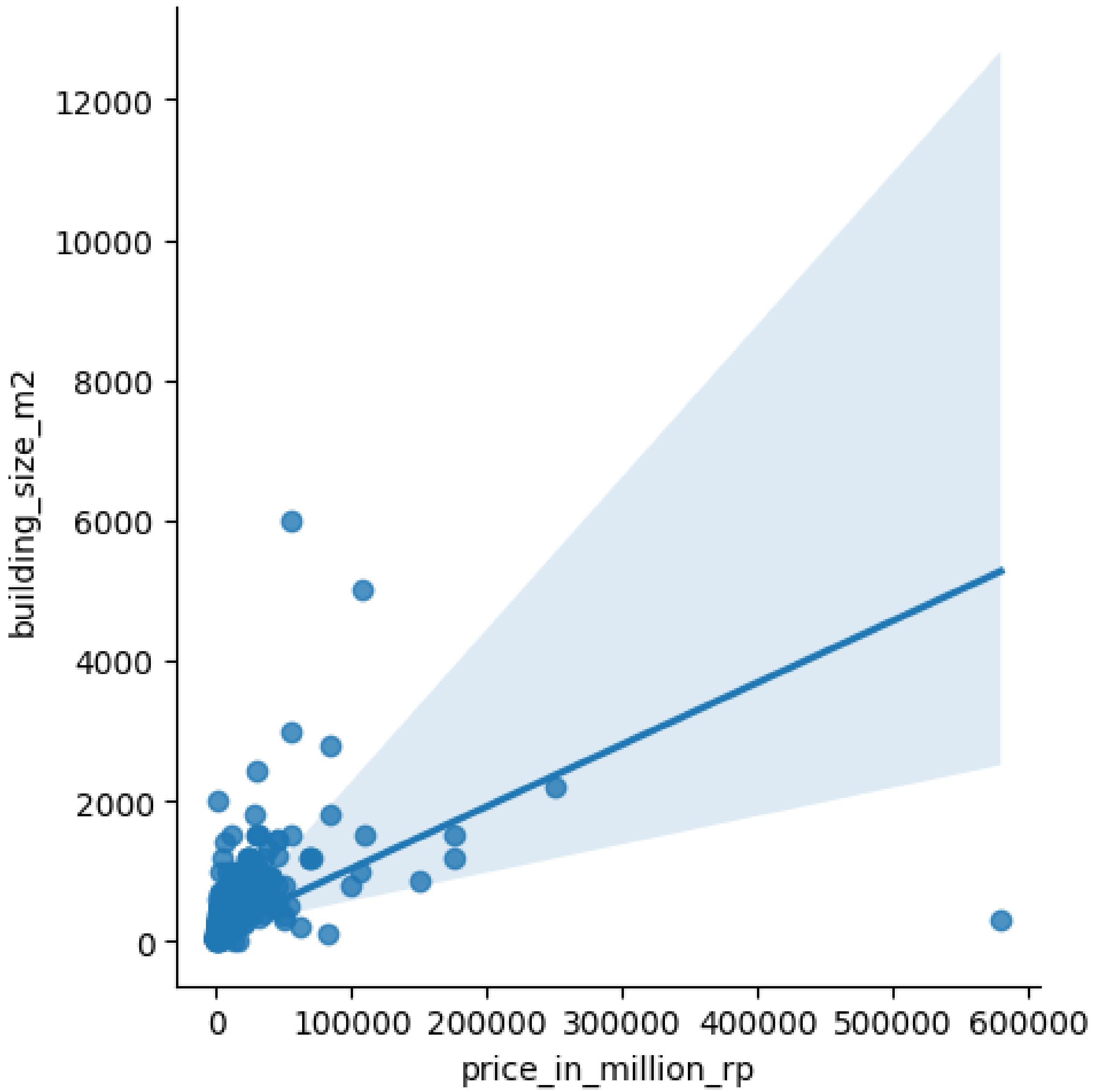
EXPLORATORY DATA ANALYSIS

DATA INSIGHT

Tren harga terhadap luas lahan



Tren harga terhadap luas bangunan



Terdapat rumah yang berharga sekitar 600 M sehingga diasumsikan terdapat anomali pada data

EXPLORATORY DATA ANALYSIS

STATISTIKA DESKRIPTIF

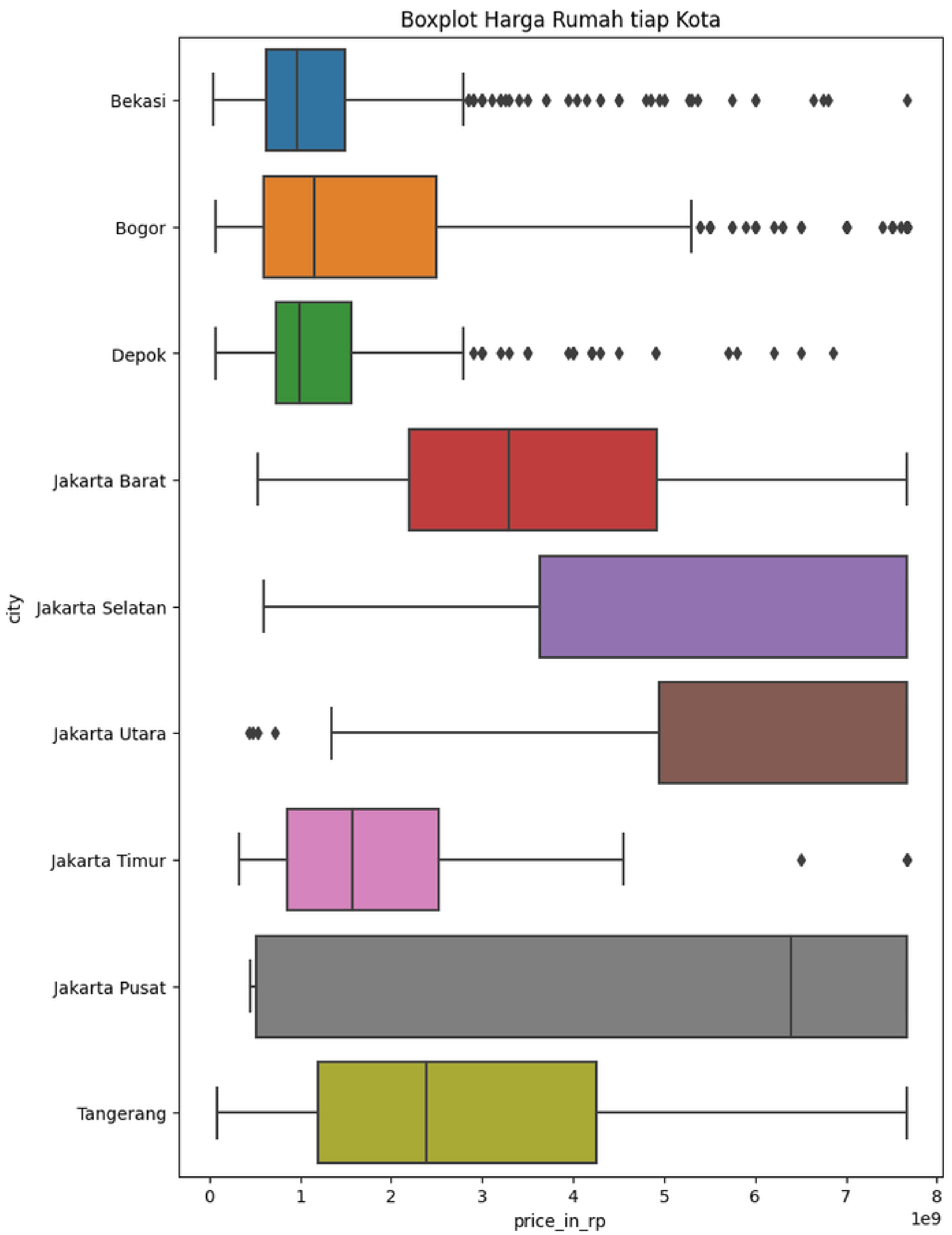
Summary Statistics

	count	mean	std	min	25%	50%	75%	max
price_in_million_rp	3553.000000	4191.684773	13750.673821	42.000000	800.000000	1500.000000	3590.000000	580000.000000
bedrooms	3553.000000	3.323952	2.670015	1.000000	2.000000	3.000000	4.000000	99.000000
bathrooms	3553.000000	2.625668	2.691021	1.000000	2.000000	2.000000	3.000000	99.000000
land_size_m2	3553.000000	204.795947	402.017784	12.000000	75.000000	108.000000	192.000000	8000.000000
building_size_m2	3553.000000	186.588798	248.376707	1.000000	66.000000	112.000000	208.000000	6000.000000
carports	3553.000000	1.197861	1.114996	0.000000	1.000000	1.000000	2.000000	15.000000
electricity	3553.000000	3327.825218	3423.699622	450.000000	2200.000000	2200.000000	3500.000000	53000.000000
maid_bedrooms	3553.000000	0.496482	0.685723	0.000000	0.000000	0.000000	1.000000	7.000000
maid_bathrooms	3553.000000	0.370391	0.536024	0.000000	0.000000	0.000000	1.000000	5.000000
floors	3553.000000	1.763299	0.637584	1.000000	1.000000	2.000000	2.000000	5.000000
garages	3553.000000	0.708978	1.311879	0.000000	0.000000	0.000000	1.000000	50.000000

EXPLORATORY DATA ANALYSIS

STATISTIKA DESKRIPTIF

Harga Rumah di daerah Jakarta Barat cenderung terdistribusi secara merata dibanding kota lain

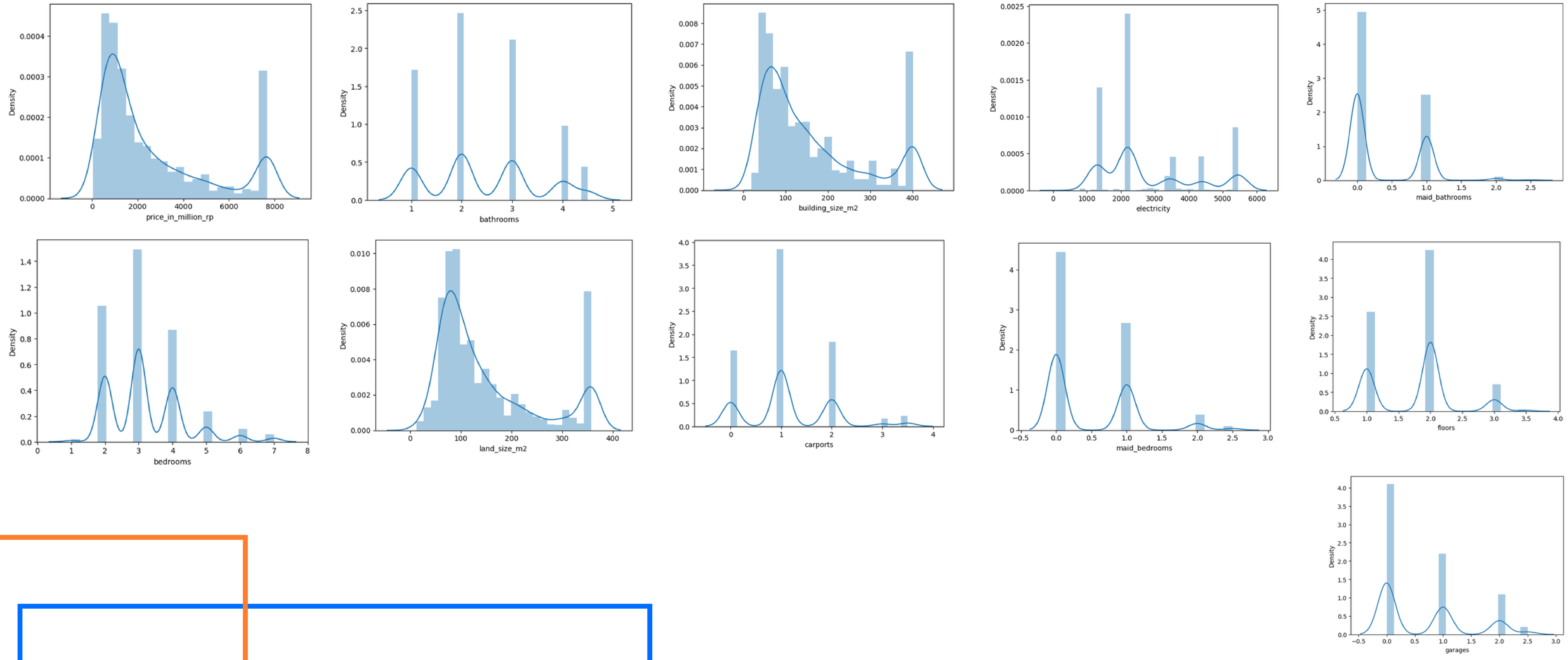


Boxplot evaluasi harga rumah per city

EXPLORATORY DATA ANALYSIS

VISUALISASI VARIABEL DAN PENYEBARAN DATA

Numeric Feature



EXPLORATORY DATA ANALYSIS

VISUALISASI VARIABEL DAN PENYEBARAN DATA

Numeric Feature

Skewness untuk kolom numerik:

- price_in_million_rp : 1.183807
- bedrooms : 1.003497
- bathrooms : 0.247150
- land_size_m2 : 1.133285
- building_size_m2 : 1.022905
- carports : 0.604335
- electricity : 0.797325
- maid_bedrooms : 1.038299
- maid_bathrooms : 0.999499
- floors : 0.277732
- garages : 0.821764

Handling Method:

```
# Handling skewness
data['bedrooms'] = np.sqrt(data['bedrooms'])
data['maid_bedrooms'] = np.sqrt(data['maid_bedrooms'])
data['price_in_million_rp'] = np.sqrt(data['price_in_million_rp'])
data['land_size_m2'] = np.sqrt(data['land_size_m2'])
data['building_size_m2'] = np.sqrt(data['building_size_m2'])
```

EXPLORATORY DATA ANALYSIS

OVERSAMPLING

Dari 11 Feature-Featrue numerik tersebut 5 (price_in_million_rp, land_size_m2 , building_size_m2, maid_bedrooms, bedrooms) diantaranya terdistribusi tidak merata / tidak normal. Oleh karena itu, diperlukan penanganan lebih lanjut

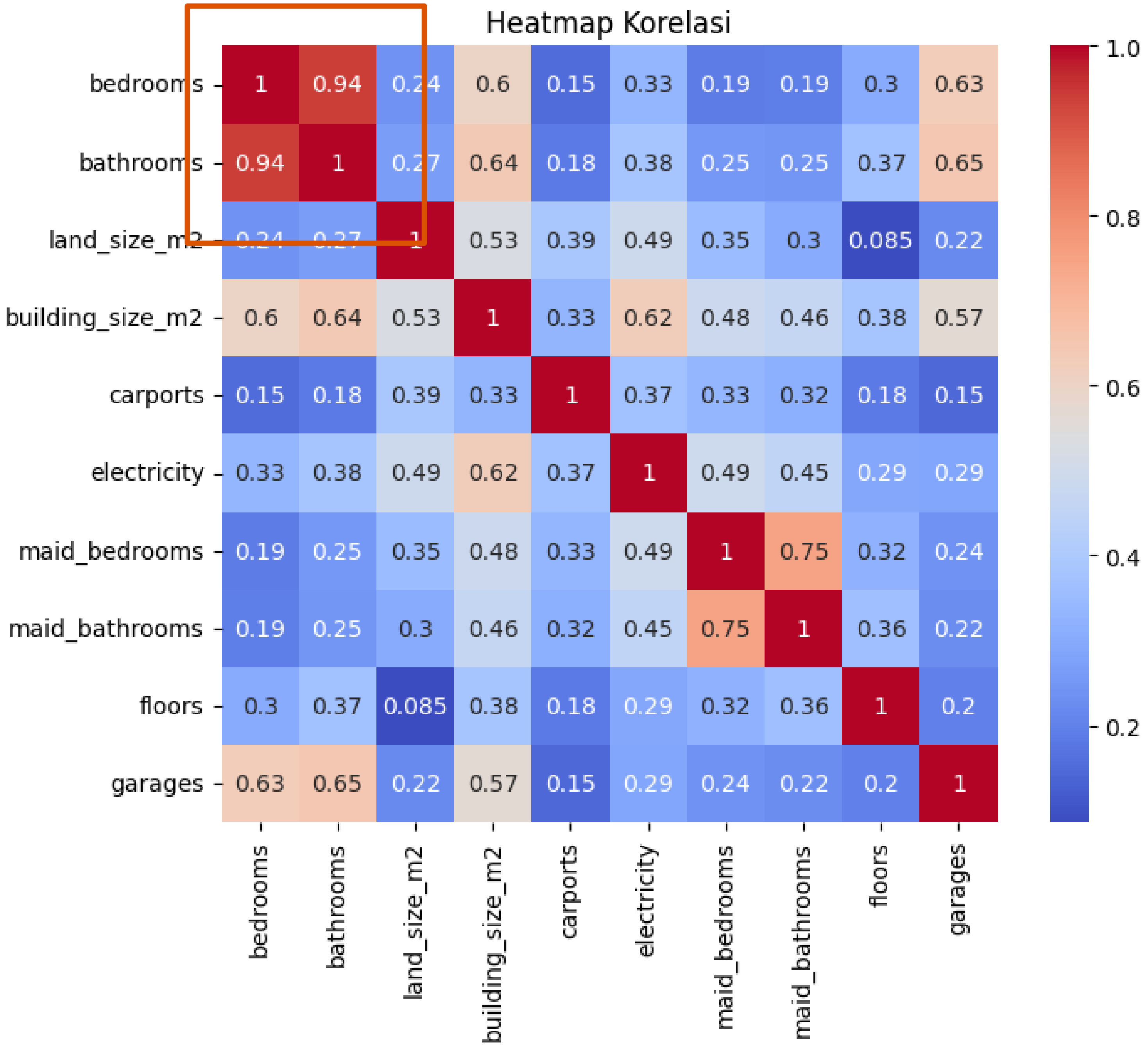
```
# Menginisialisasi dan menerapkan SMOTE hanya pada fitur numerik
ros = RandomOverSampler()
X_numerik_resampled, y_resampled = ros.fit_resample(X_numerik, y)

# Menggabungkan hasil oversampling
data = pd.DataFrame(X_numerik_resampled, columns=X_numerik.columns)
data['price_in_rp'] = y_resampled
```

EXPLORATORY DATA ANALYSIS

KORELASI

Korelasi antar feature



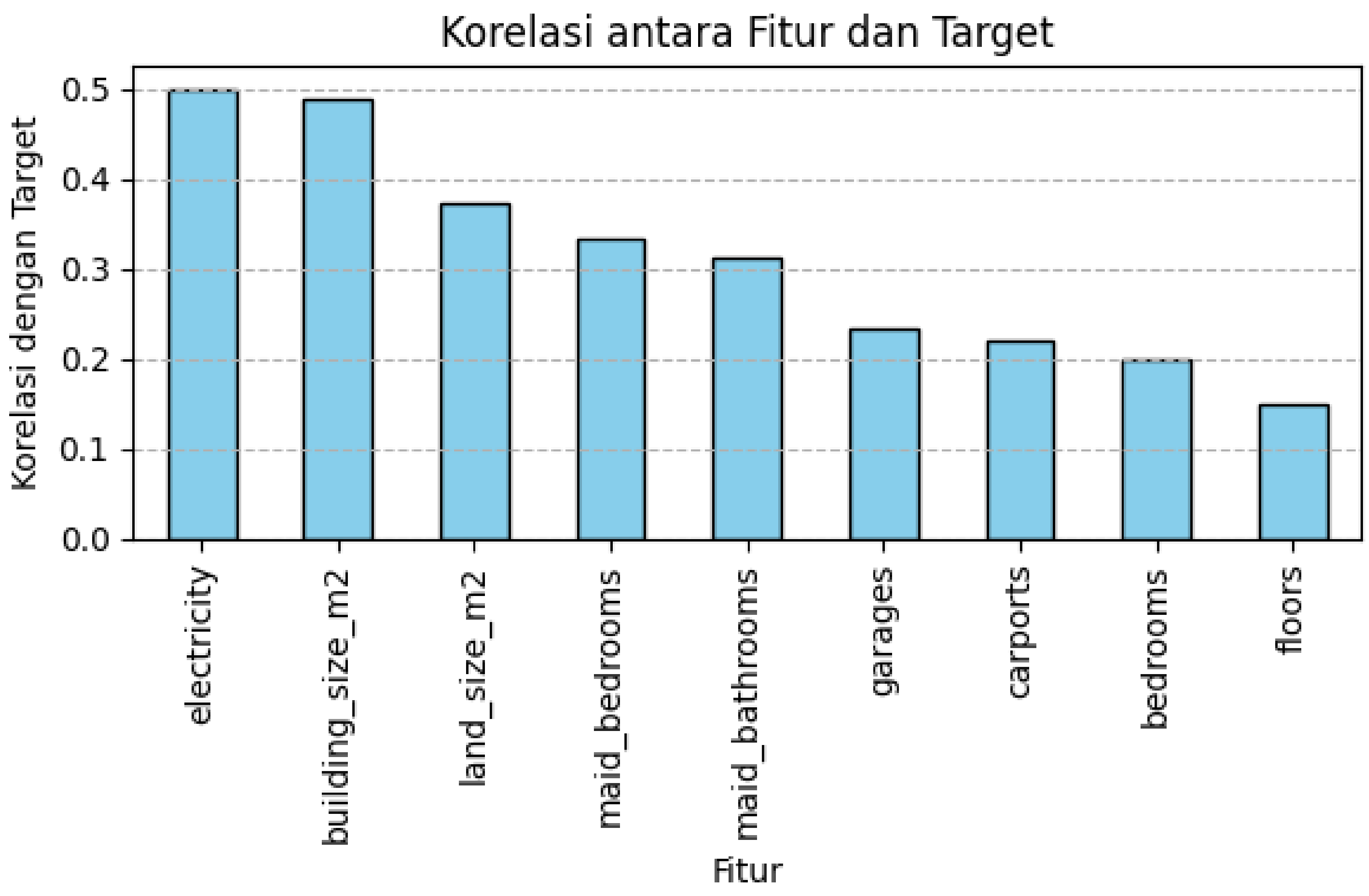
Drop salah satu feature bedrooms atau bathrooms karena memiliki **nilai korelasi > 0.9**

```
[536]
      data.drop('bathrooms',axis=1, inplace =True)
```

EXPLORATORY DATA ANALYSIS

KORELASI

Korelasi feature dengan target



Feature yang akan dipilih :

- **'price_in_rp'**
- **'district',**
- **'city'**
- **'building_size_m2'**
- **'certificate'**
- **'property_condition'**
- **'land_size_m2'**
- **'electricity'**
- **'bedrooms'**

MODELING

1

ORDINAL ENCODER

Melakukan ordinal encoder untuk data categoric

2

TRAIN TEST SPLIT

Membagi data menjadi train dan test

3

MODELING

Membuat model prediksi berupa **Linear regression, Ridge Regression, Lasso Regression, KNN, Decision Tree, dan Random Forest**

4

CROSS VALIDATION

Melakukan 10 cross-validation untuk menghindari underfitting dan overfitting

MODELING

AKURASI

1	2	3	4	5	6
Linear regression	**Ridge regression**	**Lasso regression**	**KNN**	**Decision Tree**	**Random Forest**
R2 score: 0.8254132696388159 MAE: 480485075.8628521 MSE: 5.903981914031519e +17	R2 score : 0.8254160019895221 MAE: 480468438.70356447 MSE: 5.903889514402295e +17	R2 score: 0.8254132696388272 MAE: 480485075.8626162 MSE: 5.903981914031136e +17	R2 score: 0.9905118881030175 MAE: 30012070.707070712 MSE: 3.208585264309764e +16	R2 score: 0.949793519723526 MAE: 235010630.11732408 MSE: 235010630.11732408	R2 score: 0.9941381785696896 MAE: 46087254.769230954 MSE: 1.982286261747275e+ 16

EVALUATION

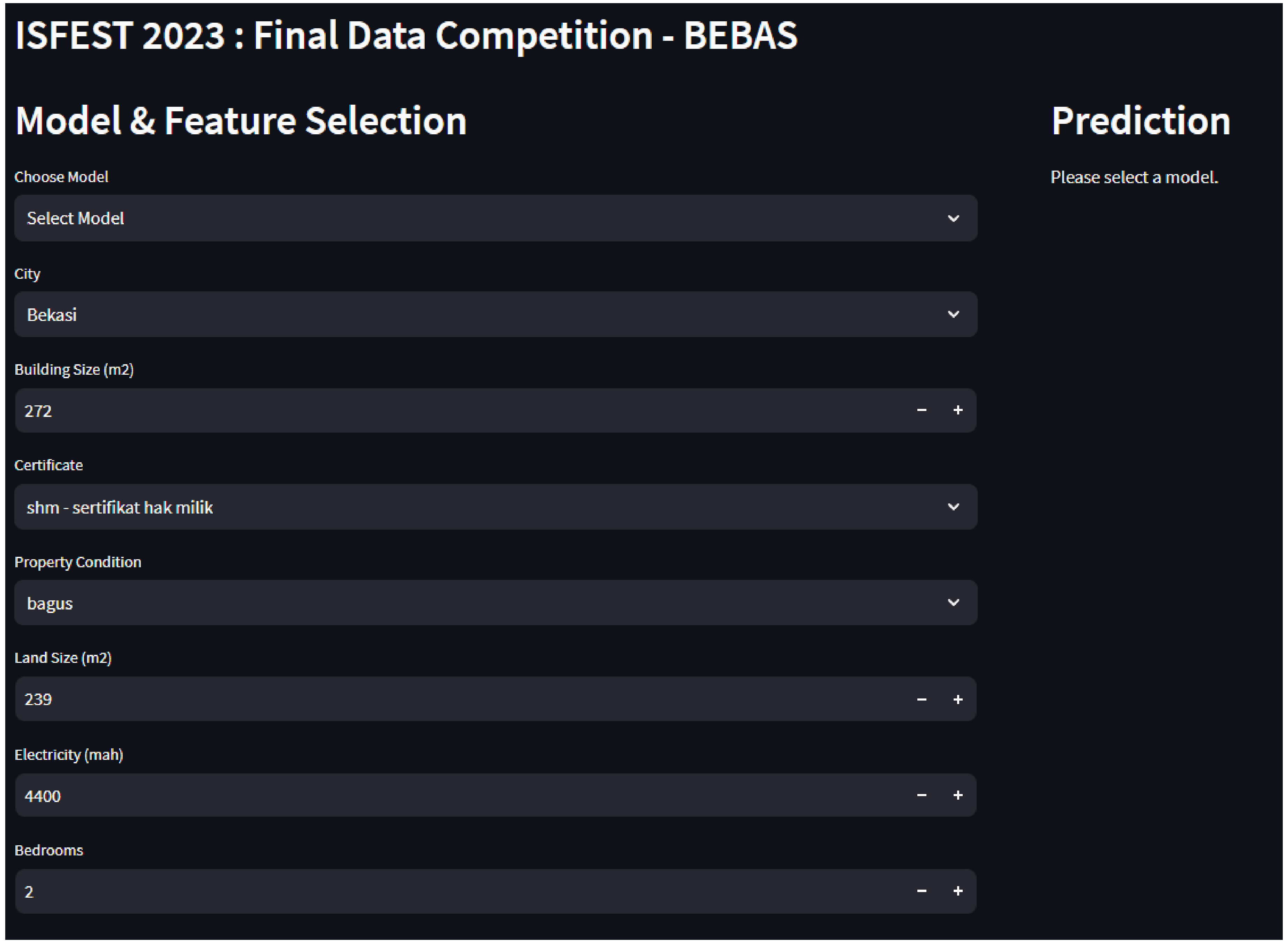
Menggunakan 10 fold cross validation

	Model	RME	RMS	R2
0	Linear Regression	4.797761e+08	5.804952e+17	0.829184
1	Ridge Regression	4.797617e+08	5.804953e+17	0.829184
2	Lasso Regresion	4.797761e+08	5.804952e+17	0.829184
3	KNN	2.935740e+07	3.119692e+16	0.990799
4	Desicion Tree	2.384413e+08	1.712274e+17	0.949659
5	Random Forest	4.390240e+07	1.876441e+16	0.994471

Akurasi terbaik sebesar **0.994471** dengan model **Random Forest**

IMPLEMENTATION

berdasarkan model yang telah dibuat, maka dapat diimplementasikan prediksi harga rumah menggunakan web



Link Implementasi :

[bebas-isfest-final-2023.streamlit.app]

CONCLUSION

- Feature yang signifkan memengaruhi harga rumah :
  - 'district',
  - 'city'
  - 'building_size_m2'
  - 'certificate'
  - 'property_condition'
  - 'land_size_m2'
  - 'electricity'
  - 'bedrooms'

- Model **Random Forest** menghasilkan prediksi yang paling bagus

- building_orientation, building_age, garages, maid_bedrooms, maid_bathrooms tidak memiliki pengaruh yang signifikan terhadap harga rumah

- Implementasi model untuk memprediksi harga rumah di daerah JABODETABEK : bebas-isfest-final-2023.streamlit.app

- Visualisasi data harga rumah di daerah JABODETABEK: House Price Jabodetabek Visualization with Tableau

LIST PERUBAHAN

- EDA
- Feature Selection

THANK YOU