

# **Singapore HDB Flat Resale Prices Prediction (2015-2020)**

**Final Project DS28 - Rivaldo**

# Data Understanding

Dataset berisi informasi tentang harga jual kembali flat HDB (Housing & Development Board) di Singapura mulai tahun 1990 hingga 2020. Pada analisis ini, akan dibatasi mulai tahun 2015 hingga 2020 saja.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	2015-01	ANG MO KIO	3 ROOM	174	ANG MO KIO AVE 4	07 TO 09	60.0	Improved	1986	70	255000.0
1	2015-01	ANG MO KIO	3 ROOM	541	ANG MO KIO AVE 10	01 TO 03	68.0	New Generation	1981	65	275000.0
2	2015-01	ANG MO KIO	3 ROOM	163	ANG MO KIO AVE 4	01 TO 03	69.0	New Generation	1980	64	285000.0
3	2015-01	ANG MO KIO	3 ROOM	446	ANG MO KIO AVE 10	01 TO 03	68.0	New Generation	1979	63	290000.0
4	2015-01	ANG MO KIO	3 ROOM	557	ANG MO KIO AVE 10	07 TO 09	68.0	New Generation	1980	64	290000.0
...	...	...	...	...	...	...	...	...	...	...	...
80369	2020-09	YISHUN	5 ROOM	716	YISHUN ST 71	07 TO 09	131.0	Improved	1987	66 years 03 months	440000.0
80370	2020-09	YISHUN	5 ROOM	760	YISHUN ST 72	07 TO 09	122.0	Improved	1987	65 years 06 months	458000.0
80371	2020-09	YISHUN	5 ROOM	835	YISHUN ST 81	04 TO 06	122.0	Improved	1987	66 years 04 months	490000.0
80372	2020-09	YISHUN	EXECUTIVE	791	YISHUN AVE 2	04 TO 06	146.0	Maisonette	1987	66 years 03 months	558000.0
80373	2020-09	YISHUN	EXECUTIVE	387	YISHUN RING RD	04 TO 06	146.0	Maisonette	1988	66 years 09 months	555000.0

# Data Understanding


- **month**: bulan dan tahun transaksi
- **town**: kota dimana flat berada
- **flat\_type**: tipe flat berdasarkan jumlah ruangan
- **block**: nomor blok flat
- **street\_name**: nama jalan dimana flat berada
- **storey\_range**: lokasi flat berdasarkan tingkat bangunan
- **floor\_area\_sqm**: floor area of flat in square meter
- **flat\_model**: model flat
- **lease\_commence\_date**: tahun dimana flat pertama kali disewakan
- **remaining\_lease**: tahun dan bulan sisa waktu sewa
- **resale\_price**: harga jual flat dalam Singapore Dollars (SGD)

```
<class 'pandas.core.frame.DataFrame'>
Index: 117293 entries, 0 to 80373
Data columns (total 11 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   month               117293 non-null  object
1   town                117293 non-null  object
2   flat_type           117293 non-null  object
3   block               117293 non-null  object
4   street_name         117293 non-null  object
5   storey_range        117293 non-null  object
6   floor_area_sqm      117293 non-null  float64
7   flat_model          117293 non-null  object
8   lease_commence_date 117293 non-null  int64
9   remaining_lease     117293 non-null  object
10  resale_price        117293 non-null  float64
dtypes: float64(2), int64(1), object(8)
memory usage: 10.7+ MB
```

## Objectives / Goals

- Membantu calon pembeli, penjual, dan agen properti dalam memprediksi harga jual unit HDB yang wajar berdasarkan fitur-fitur seperti lokasi, luas lantai, dan sisa masa sewa.
- Membantu calon pembeli maupun pemerintah untuk mengidentifikasi fitur-fitur yang mempengaruhi harga jual flat, serta mengidentifikasi seberapa besar pengaruh fitur-fitur tersebut mempengaruhi harga jual.

# Data Cleaning



	Feature	Missing_Values	Percentage
0	month	0	0.0
1	town	0	0.0
2	flat_type	0	0.0
3	block	0	0.0
4	street_name	0	0.0
5	storey_range	0	0.0
6	floor_area_sqm	0	0.0
7	flat_model	0	0.0
8	lease_commence_date	0	0.0
9	remaining_lease	0	0.0
10	resale_price	0	0.0

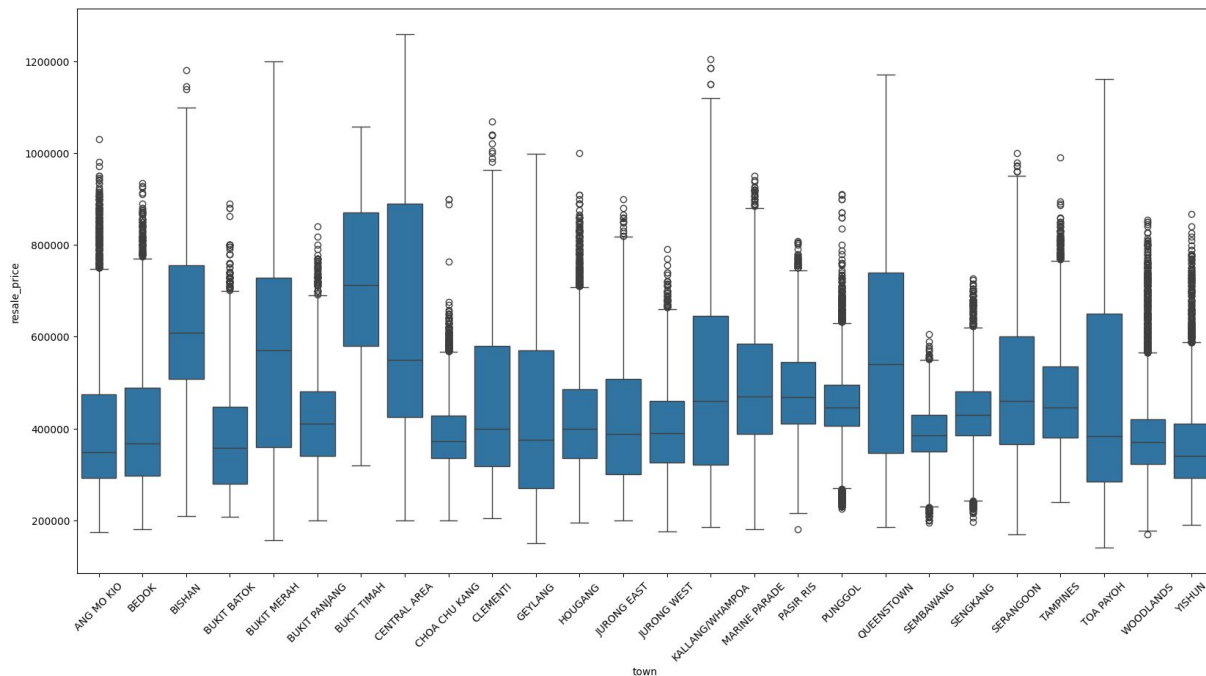
## Missing Values

Tidak ada

## Duplicated Data

Ada, di-handle dengan drop rows yang duplikat karena proporsi data hilang sangat sedikit

# Town vs Resale Price

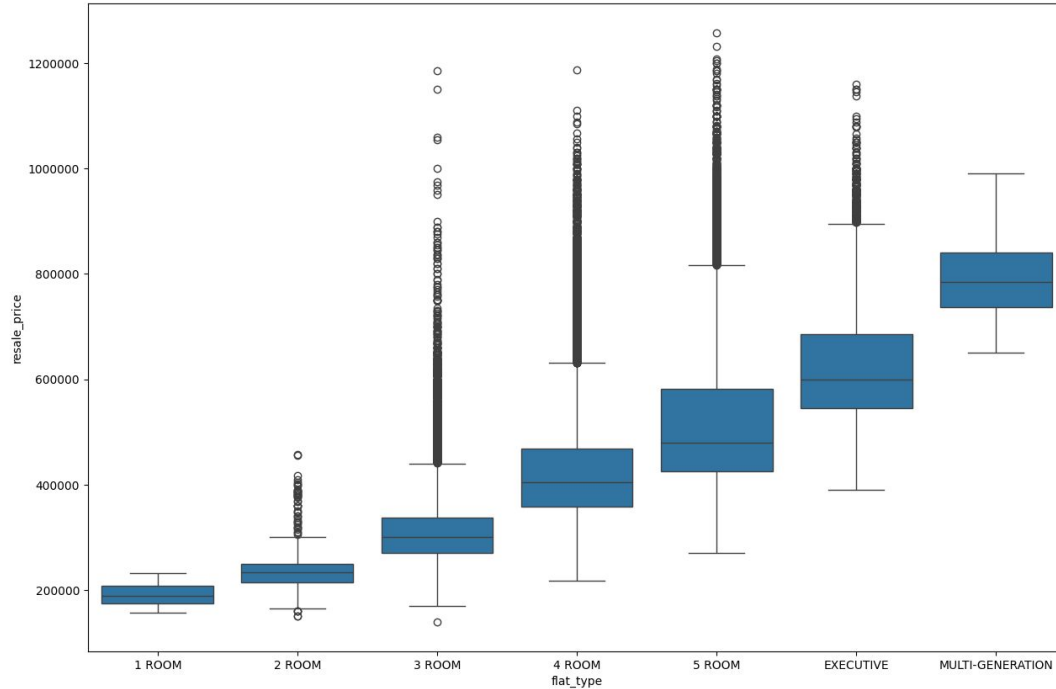


## Observasi:

Kota dengan Median  $\geq$  500k dollars:

- Bishan
- Bukit Merah
- Bukit Timah
- Central Area
- Queenstown

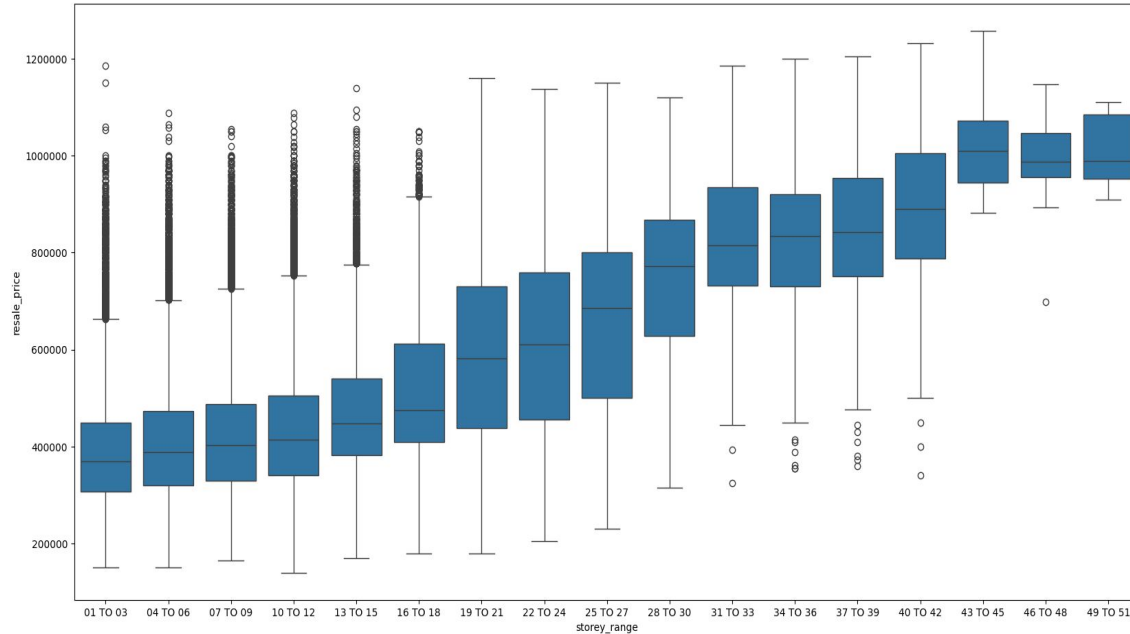
# Flat Type vs Resale Price



## Observasi:

- Semakin banyak ruangan dalam flat, maka harga jualnya semakin tinggi.
- Ada beberapa outliers, artinya Flat Type bukan satu-satunya fitur yang mempengaruhi Resale Price.

# Storey Range vs Resale Price

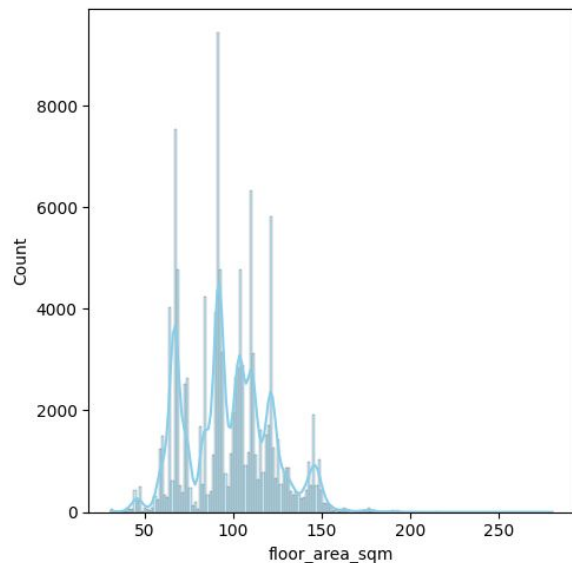


## Observasi:

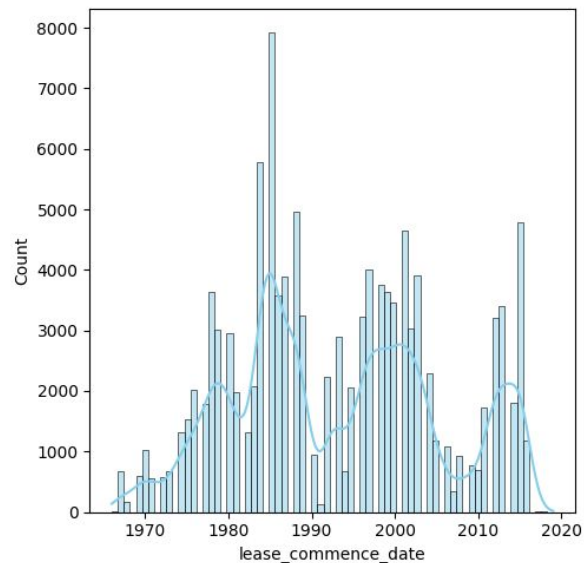
- Semakin tinggi lokasi flat, maka harga jualnya semakin tinggi.
- Ada beberapa outliers pada storey 01 hingga 18, artinya Storey Range juga bukan satu-satunya fitur yang mempengaruhi Resale Price.



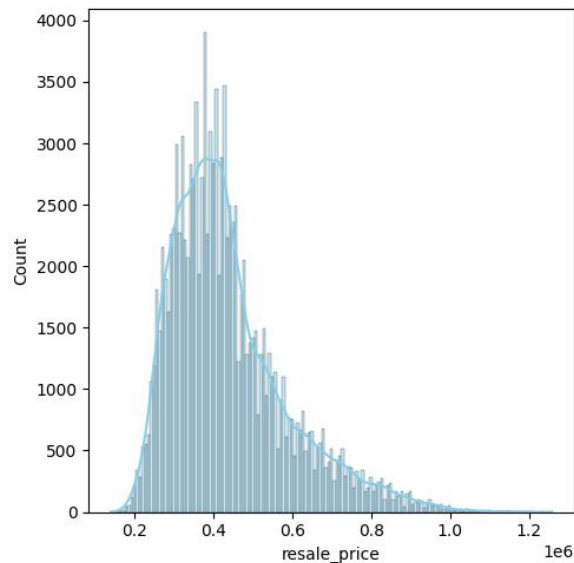
# KDE Plot to Check Distribution Form



Positively-skewed

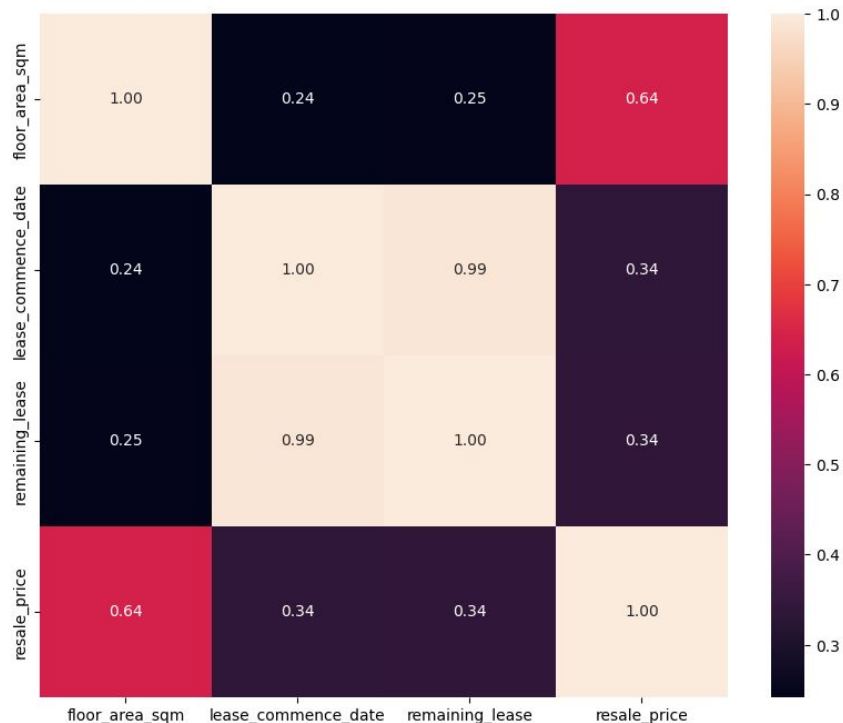


Multimodal



Positively-skewed

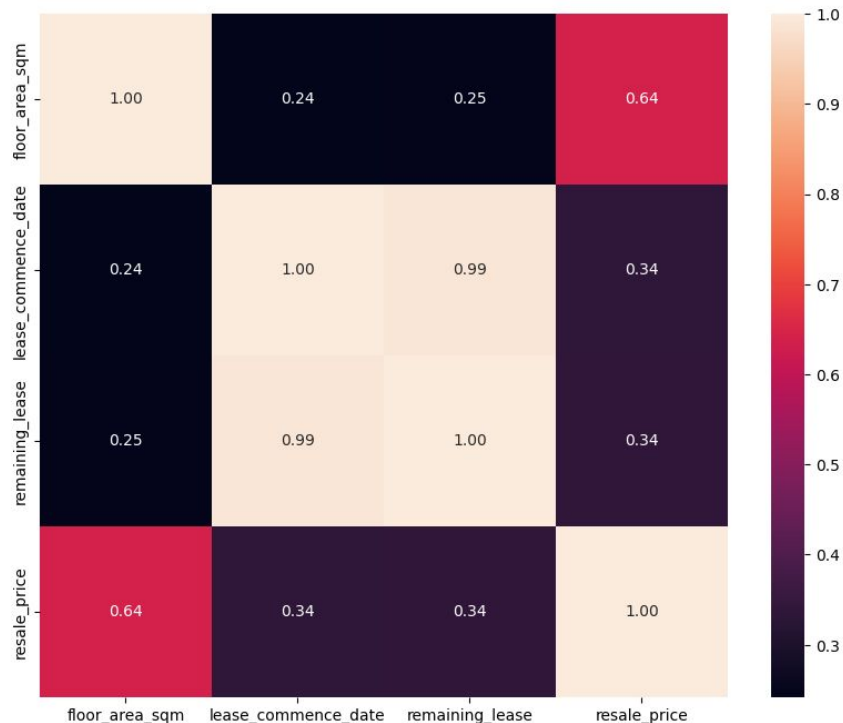
# Multicollinearity Study (1/2)



## Korelasi antar fitur:

- Korelasi antara **luas flat** (`floor_area_sqm`) terhadap **tahun mulai sewa** (`lease_commence_date`) hanya sebesar 0.24 (sangat lemah)
- Korelasi antara **luas flat** (`floor_area_sqm`) terhadap **siswa masa sewa** (`remaining_lease`) hanya sebesar 0.25 (sangat lemah)
- Korelasi antara **tahun mulai sewa** (`lease_commence_date`) terhadap **siswa masa sewa** (`remaining_lease`) adalah sebesar 0.99 (sangat tinggi)

# Multicollinearity Study (2/2)



## Korelasi fitur terhadap target:

- Korelasi antara **luas flat** (`floor_area_sqm`) terhadap **harga jual** (`resale_price`) adalah sebesar 0.64 (cukup kuat)
- Korelasi antara **tahun mulai sewa** (`lease_commence_date`) dan **sisa masa sewa** (`remaining_lease`) terhadap **harga jual** adalah sebesar 0.34 (cukup lemah)

# Feature Engineering (1/2)

Menggunakan **OneMap** untuk mendapatkan data Latitude dan Longitude. Data ini digunakan untuk menghitung jarak antara flat HDB ke sebuah lokasi (Misal MRT Station).

- Main Web: <https://www.onemap.gov.sg/>
- API Docs: <https://www.onemap.gov.sg/apidocs/>





# Multicollinearity Study After Feature Engineering



## Korelasi antar fitur (dengan fitur baru):

- Korelasi antara **jarak ke stasiun MRT** (**distance\_to\_mrt**) terhadap **jarak ke Mall** (**distance\_to\_mall**) sebesar 0.99 (sangat tinggi).

Sehingga, diputuskan untuk menggunakan salah satu fitur saja dalam pemodelan agar tidak redundan.

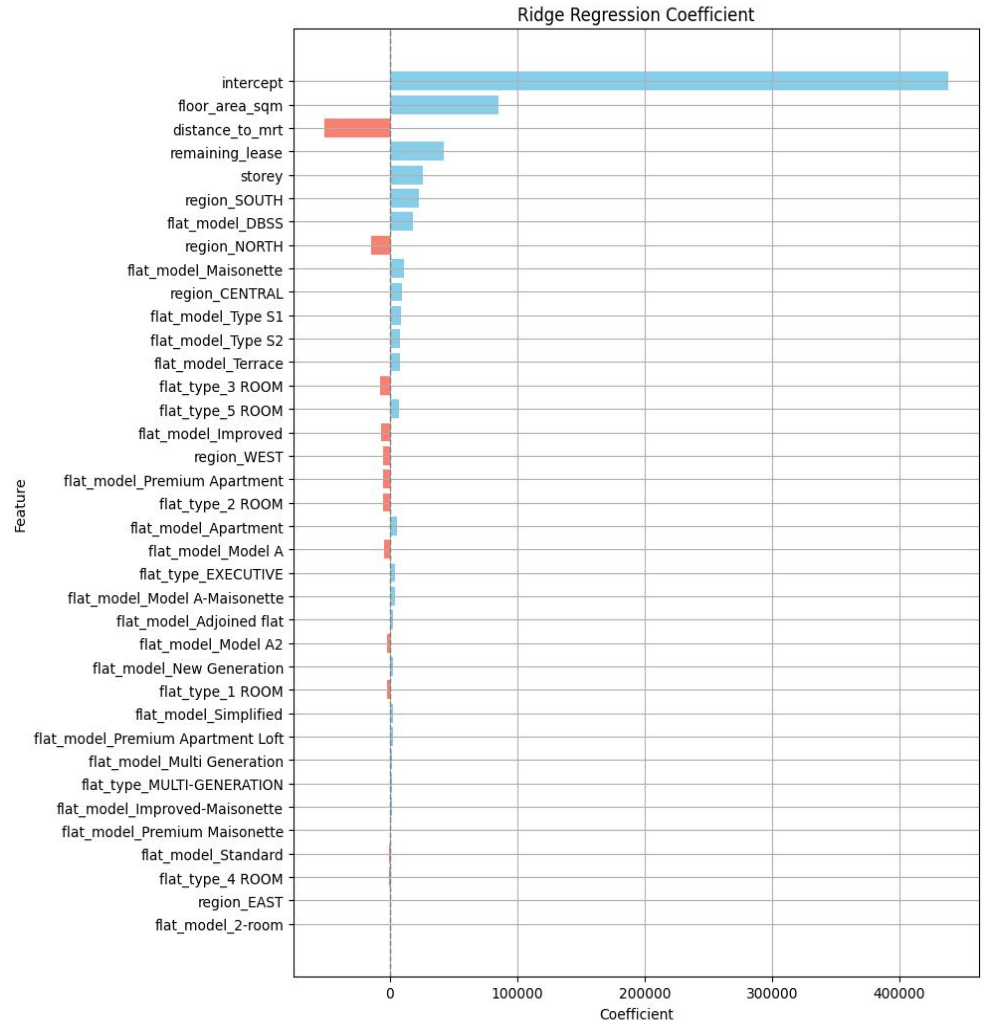
# Modeling

- Linear Regression (Baseline Model)
- Ridge Regression
- Random Forest Regression
- XGBoost Regression

# Model Interpretation (1/3)

## [Interpretable Model]

- **floor\_area\_sqm:**  
Luas flat memiliki pengaruh yang kuat terhadap harga jual. Semakin luas, akan semakin mahal (begitupun sebaliknya).
- **distance\_to\_mrt:**  
Flat yang dekat dengan stasiun MRT memiliki harga jual yang lebih mahal.
- **remaining\_lease:**  
Semakin panjang sisa waktu sewa sebuah flat, maka harga jualnya semakin mahal.
- **storey:**  
Semakin tinggi lokasi flat pada bangunan tinggi, maka harga jualnya semakin mahal.

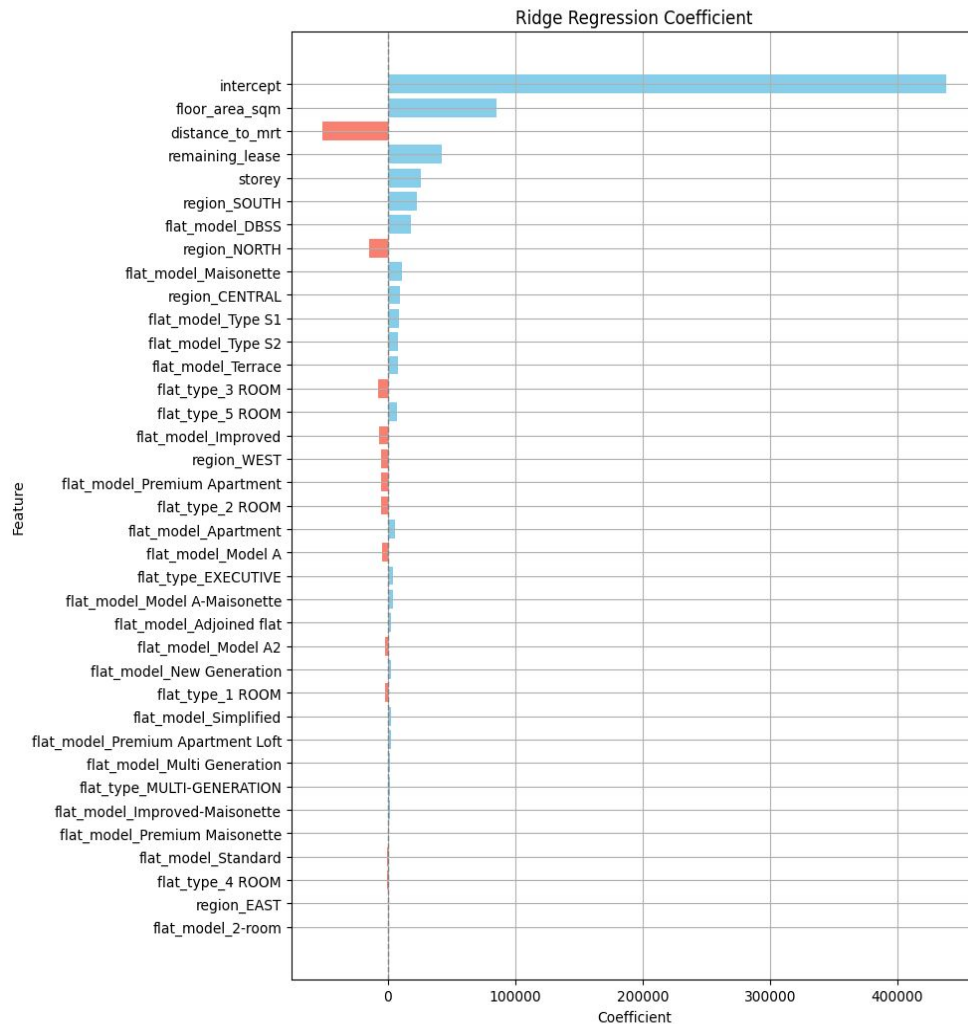




# Model Interpretation (2/3)

## [Interpretable Model]

- **region:**  
Flat di wilayah selatan dan tengah cenderung lebih mahal daripada di wilayah utara dan barat.
- **flat\_model:**  
Flat dengan model DBSS, Maisonette, Type S1, Type S2, and Terrace cenderung lebih mahal.
- **flat\_type:**  
Flat dengan banyak ruangan cenderung lebih mahal.

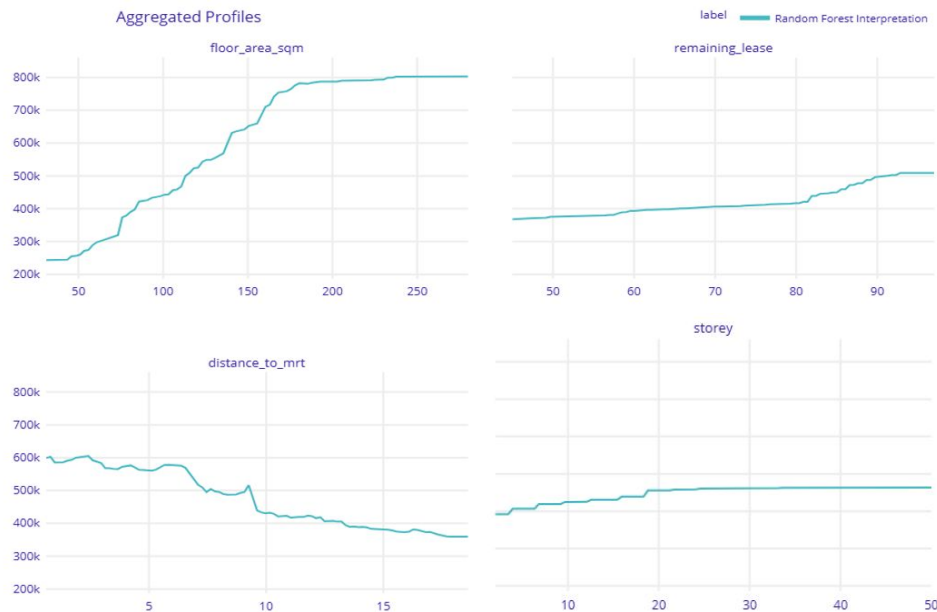


# Model Interpretation (3/3)

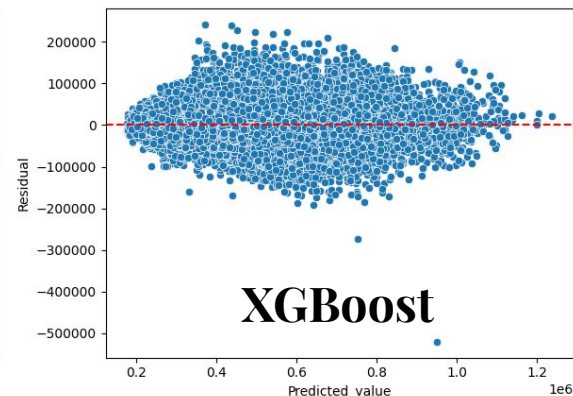
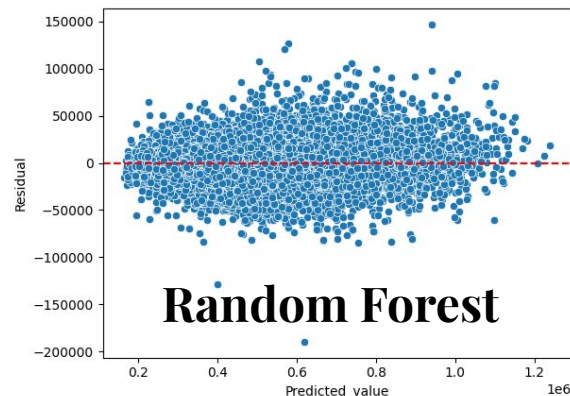
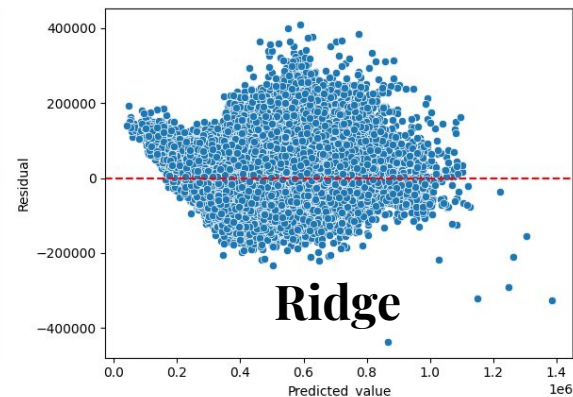
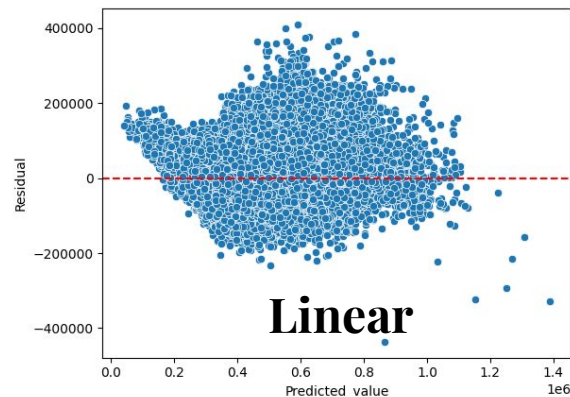
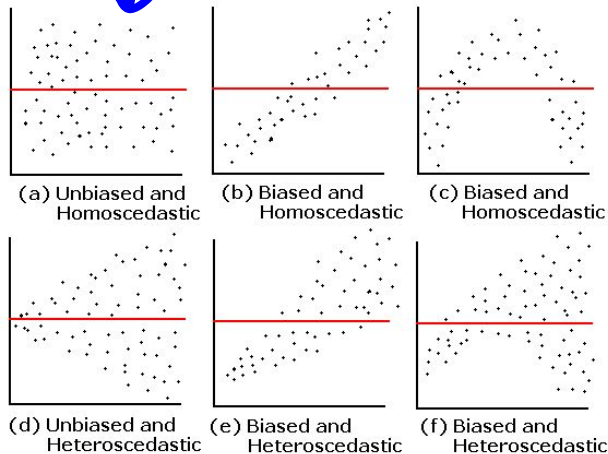
## [Non-Interpretable Model]

### Using Partial Dependence Plot

Grafik memiliki kesimpulan yang sama dengan interpretable model sebelumnya



# Model Evaluation - Residual Plot



# Model Evaluation - Regression Metrics (1/2)

- $R^2$  Score:  
Model terbaik adalah Random Forest Regressor dengan  $R^2$  Score sebesar 96%.  $R^2$  Score yang tinggi menandakan bahwa model ini dapat **menangkap tren** dan **menjelaskan variansi dengan sangat baik**.
- Mean Absolute Error (MAE):  
Rata-rata kesalahan absolut antara nilai aktual dan prediksi adalah sekitar 19.977 SGD. Karena harga jual kembali berada dalam rentang 140.000 SGD hingga 1.258.000 SGD, maka kesalahan sekitar 19.977 SGD dikategorikan relatif kecil. Hal ini menunjukkan bahwa **performa model sudah baik**.

Model	Linear Regression	Ridge Regression	Random Forest	XGBoost
MAE_train	48099.825255	48098.477914	8200.660061	24007.283707
MAE_test	48763.665727	48762.544182	19977.862911	25356.964145
MAPE_train	0.114946	0.114940	0.019303	0.056231
MAPE_test	0.116026	0.116020	0.046312	0.058535
RMSE_train	61875.926937	61875.938920	11699.454127	32536.248288
RMSE_test	62576.822117	62577.278416	28394.543279	34673.125284
$R^2_{train}$	0.825439	0.825439	0.993759	0.951734
$R^2_{test}$	0.822453	0.822450	0.963444	0.945490

## Model Evaluation - Regression Metrics (2/2)

- Mean Absolute Percentage Error (MAPE): Mengukur kesalahan relatif sebagai persentase dari nilai aktual. MAPE dari prediksi model adalah 4,31%, artinya model hanya memiliki selisih 4,31% dari harga jual kembali yang sebenarnya (aktual). Jadi, bisa disimpulkan bahwa model **memiliki akurasi yang tinggi**.
- Root Mean Squared Error (RMSE): Mengukur akar dari rata-rata kesalahan kuadrat. Karena RMSE jauh lebih kecil dari harga jual kembali rata-rata (~438.533), model ini **dapat bekerja dengan baik**.
- Model tidak mengalami overfitting karena score antara train dan test mirip yang artinya **model sudah baik dalam menangani data baru**.

Model	Linear Regression	Ridge Regression	Random Forest	XGBoost
MAE_train	48099.825255	48098.477914	8200.660061	24007.283707
MAE_test	48763.665727	48762.544182	19977.862911	25356.964145
MAPE_train	0.114946	0.114940	0.019303	0.056231
MAPE_test	0.116026	0.116020	0.046312	0.058535
RMSE_train	61875.926937	61875.938920	11699.454127	32536.248288
RMSE_test	62576.822117	62577.278416	28394.543279	34673.125284
R2_train	0.825439	0.825439	0.993759	0.951734
R2_test	0.822453	0.822450	0.963444	0.945490

# Summary & Recommendations (1/2)

## Untuk Pembeli:

- Luas Lantai (**floor\_area\_sqm**): Luas lantai merupakan faktor paling signifikan yang mempengaruhi harga jual kembali. Jika pembeli menginginkan apartemen yang murah, mereka sebaiknya mencari unit dengan luas lantai yang lebih kecil.
- Jarak ke MRT (**distance\_to\_mrt**): Apartemen yang dekat dengan stasiun MRT cenderung memiliki harga jual kembali yang lebih tinggi. Jika pembeli menginginkan apartemen yang murah, mereka bisa mempertimbangkan untuk mencari unit yang lebih jauh dari MRT. Namun, hal ini akan berdampak pada waktu perjalanan.
- Sisa Masa Sewa (**remaining\_lease**): Jika pembeli menginginkan apartemen yang lebih terjangkau, mereka sebaiknya mencari unit dengan sisa masa sewa yang lebih pendek, tetapi tidak disarankan terlalu pendek karena dapat mempengaruhi kenyamanan di masa depan.
- Tingkat Lantai (**storey**): Unit di lantai yang lebih tinggi cenderung memiliki harga jual kembali yang lebih tinggi, karena dapat memberikan tingkat kenyamanan yang lebih tinggi, mengingat di lantai atas lebih sedikit orang berlalu-lalang.

# Summary & Recommendations (2/2)

## Untuk Pembeli:

- Wilayah (**region**): Apartemen di wilayah selatan dan pusat cenderung lebih mahal dibandingkan dengan yang berada di wilayah utara dan barat.
- Model Apartemen (**flat\_model**): Apartemen dengan model seperti DBSS, Maisonette, Type S1, Type S2, dan Terrace umumnya memiliki harga jual kembali yang lebih tinggi.
- Tipe Apartemen (**flat\_type**): Apartemen dengan jumlah kamar lebih banyak umumnya memiliki harga jual kembali yang lebih tinggi.

## Untuk Pemerintah:

Pemerintah sebaiknya mempertimbangkan untuk membangun lebih banyak pilihan transportasi (seperti jalur MRT), pusat perbelanjaan, sekolah, dan fasilitas publik lainnya guna membantu meningkatkan harga jual kembali apartemen.



# Thank You

**Any Questions?**