SC1015 MINI PROJECT

HDB Resale Flat Price Analysis in Singapore

LIM WEI JIAN, DAVID U2430324H
SEE PEI JIN, ERNEST U2430223J
LAB ECDS GROUP 2

CONTENT

INTRODUCTION

- HDB resale flats are a major part of Singapore's housing market

- In 2024, there were 27835 recorded transactions of resale flats in Singapore. That's about 76 transactions a day.

- Resale flat prices are determined by the different attributes of a flat

PROBLEM DEFINITION

Can we

1. Develop a data-driven model to predict HDB resale prices based on relevant features

And

2. Classify if the resale price people pay is fair or unfair

PROBLEM DEFINITION

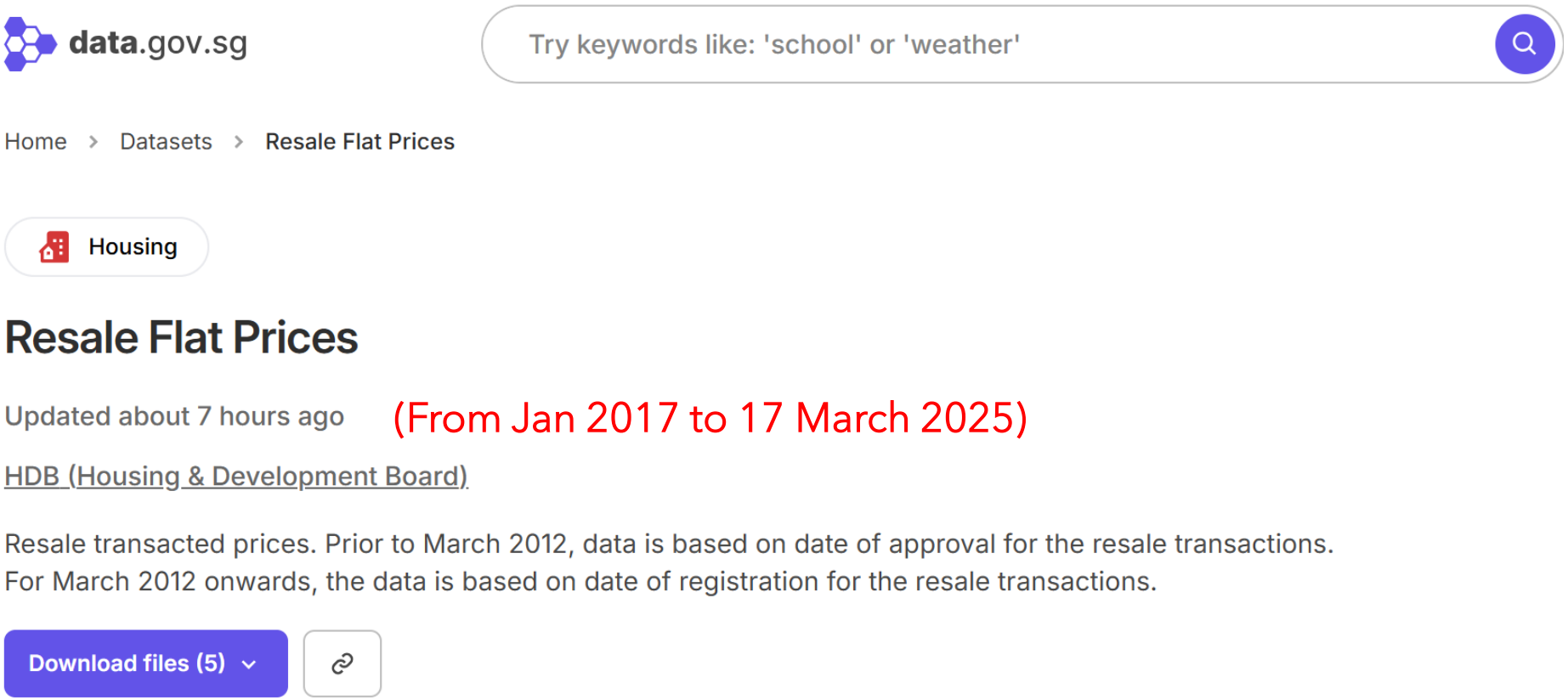
Why do we want to predict the resale price?

- Important for homebuyers, sellers, real estate agents, and policymakers to gauge the value of resale flats

Why is it important to determine if the price is "fair"?

- Allows for informed negotiations to give both buyer and seller confidence in pricing fairness
- Ensuring affordability and accessibility for first time buyers

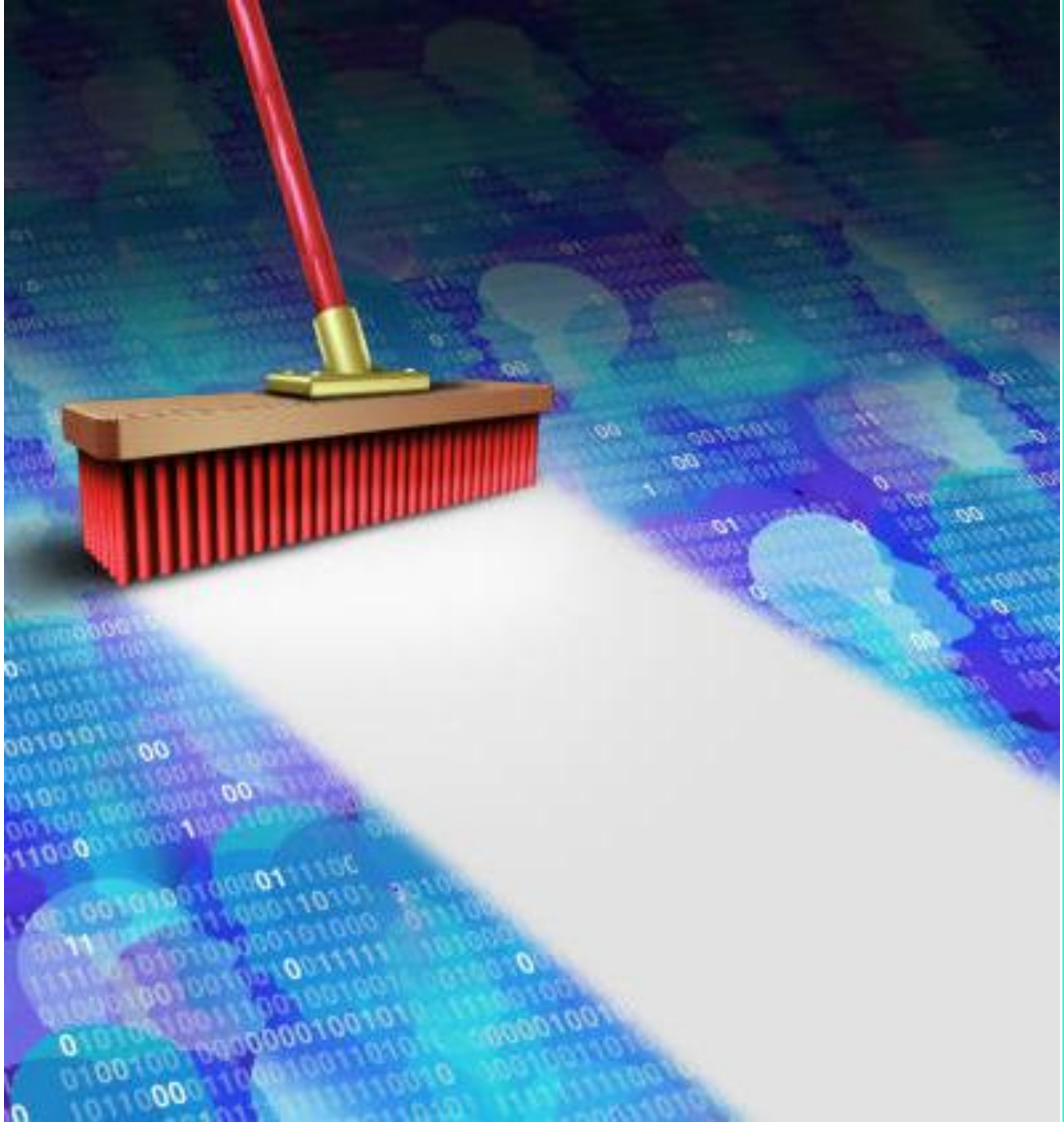
DATA SET USED



DATA SET USED

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2017-01	ANG MO KI(	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	2017-01	ANG MO KI(	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Gener	1978	60 years 07 months	250000
4	2017-01	ANG MO KI(	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	262000
5	2017-01	ANG MO KI(	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1980	62 years 01 month	265000
6	2017-01	ANG MO KI(	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	265000
7	2017-01	ANG MO KI(	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Gener	1981	63 years	275000
8	2017-01	ANG MO KI(	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1979	61 years 06 months	280000
9	2017-01	ANG MO KI(	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Gener	1976	58 years 04 months	285000

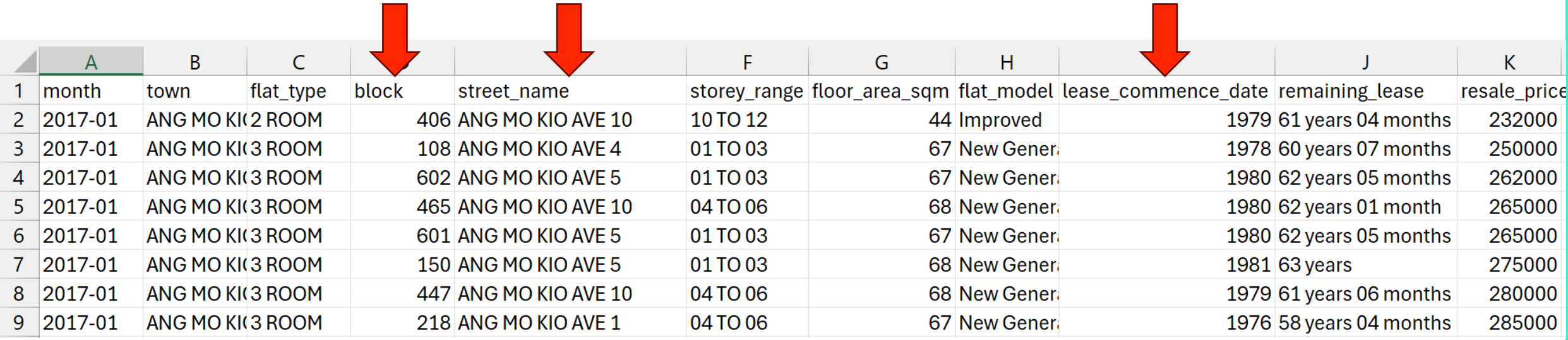
DATA EXTRACTION AND CLEANING



EXTRACTION AND CLEANING OF DATA

- Dropping of irrelevant columns
- Converting Remaining Lease to Float
- Creating a New Variable: Price per square Meter
- Perform One-Hot Encoding on categorical data

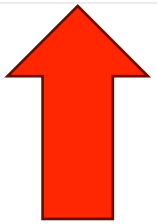
DROPPING OF IRRELEVANT COLUMNS



	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2017-01	ANG MO KI(	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	2017-01	ANG MO KI(	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Gener	1978	60 years 07 months	250000
4	2017-01	ANG MO KI(	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	262000
5	2017-01	ANG MO KI(	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1980	62 years 01 month	265000
6	2017-01	ANG MO KI(	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	265000
7	2017-01	ANG MO KI(	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Gener	1981	63 years	275000
8	2017-01	ANG MO KI(	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1979	61 years 06 months	280000
9	2017-01	ANG MO KI(	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Gener	1976	58 years 04 months	285000

CONVERTING REMAINING LEASE TO FLOAT

	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2017-01	ANG MO KI(	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	2017-01	ANG MO KI(	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Gener	1978	60 years 07 months	250000
4	2017-01	ANG MO KI(	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	262000
5	2017-01	ANG MO KI(	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1980	62 years 01 month	265000
6	2017-01	ANG MO KI(	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Gener	1980	62 years 05 months	265000
7	2017-01	ANG MO KI(	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Gener	1981	63 years	275000
8	2017-01	ANG MO KI(	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Gener	1979	61 years 06 months	280000
9	2017-01	ANG MO KI(	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Gener	1976	58 years 04 months	285000

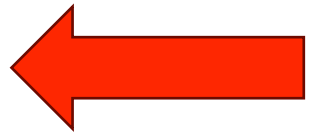


	remaining_lease_years
0	61.333333
1	60.583333
2	62.416667
3	62.083333
4	62.416667

CREATING A NEW VARIABLE: PRICE PER SQUARE METER

$$\text{price\_per\_sqm} = \frac{\text{resale\_price}}{\text{floor\_area\_sqm}}$$

	resale_price	floor_area_sqm	psm
0	232000.0	44.0	5272.727273
1	250000.0	67.0	3731.343284
2	262000.0	67.0	3910.447761
3	265000.0	68.0	3897.058824
4	265000.0	67.0	3955.223881

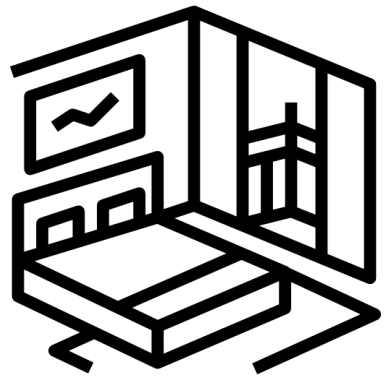


ONE HOT ENCODING FOR CATEGORICAL DATA

- One Hot Encoding is a method for converting categorical variables into a binary format.
- The primary purpose of One Hot Encoding is to ensure that categorical data can be effectively used in machine learning models.

town

storey_range

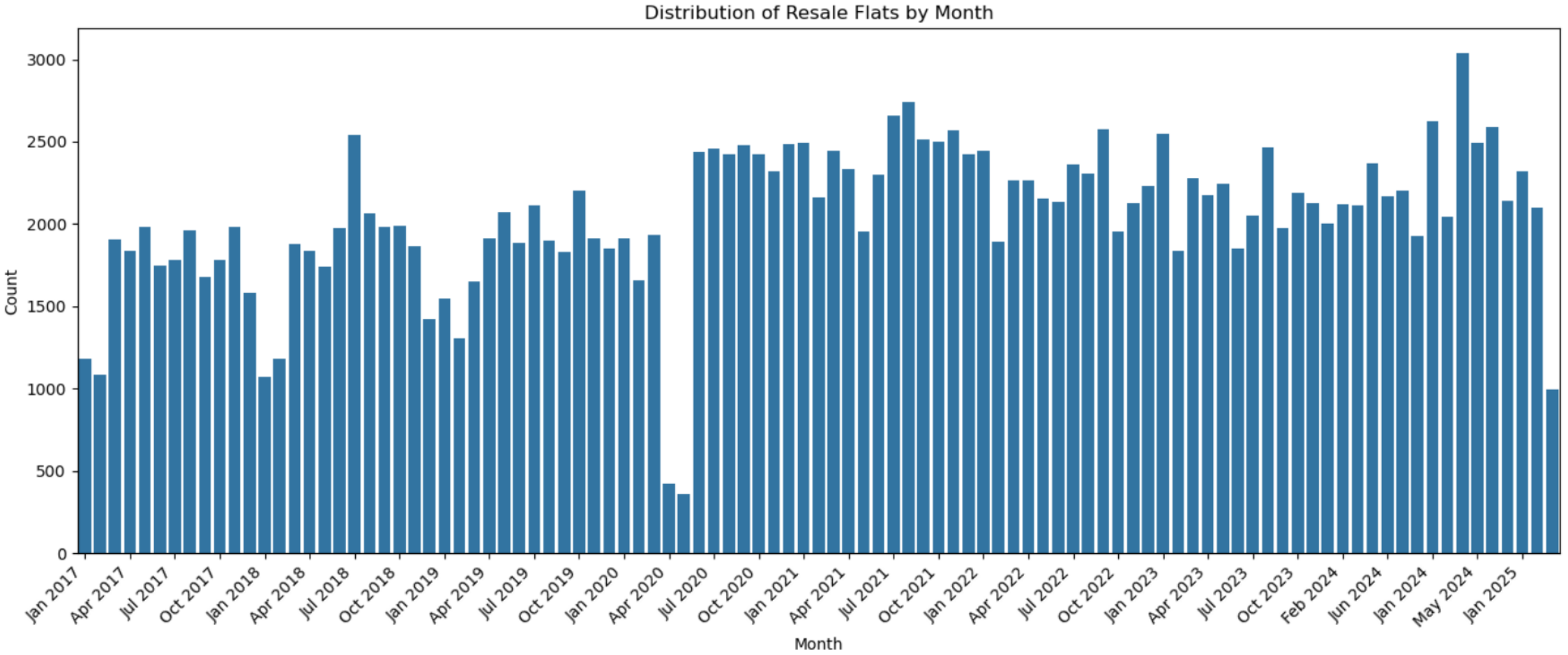


flat_type

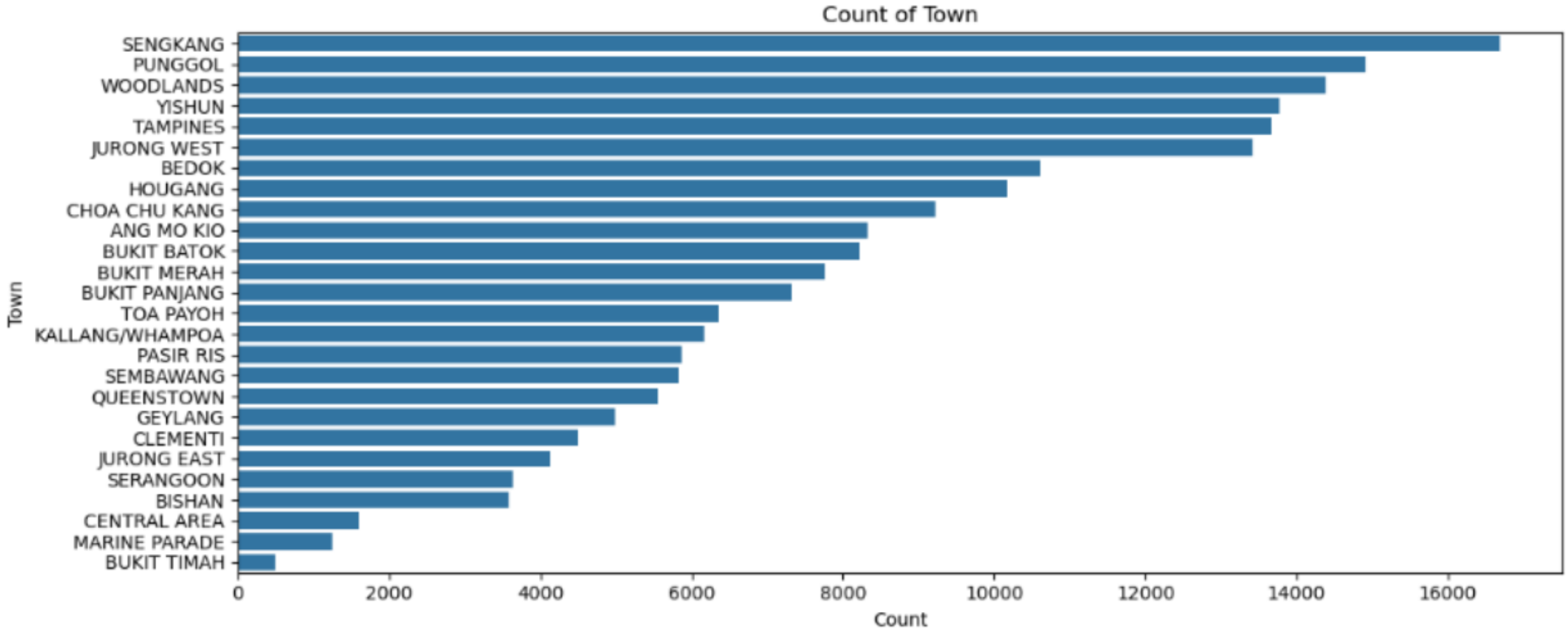
EXPLORATORY DATA ANALYSIS

1. Overview of Resale Flat Distribution (Univariate Analysis)
2. Bivariate Analysis with Respect to Resale Price

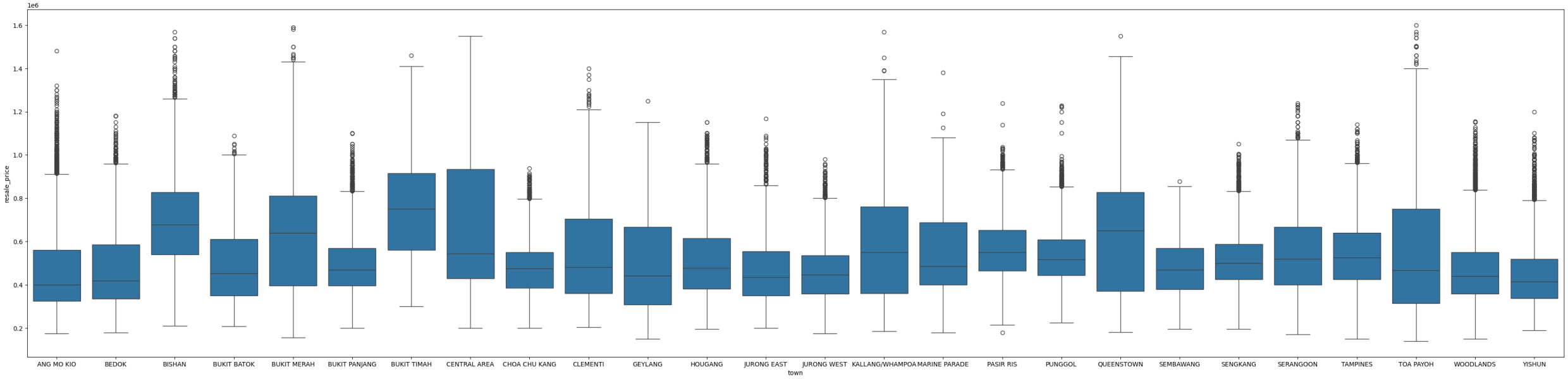
RESALE FLATS SOLD BY MONTH



RESALE FLATS SOLD BY TOWN

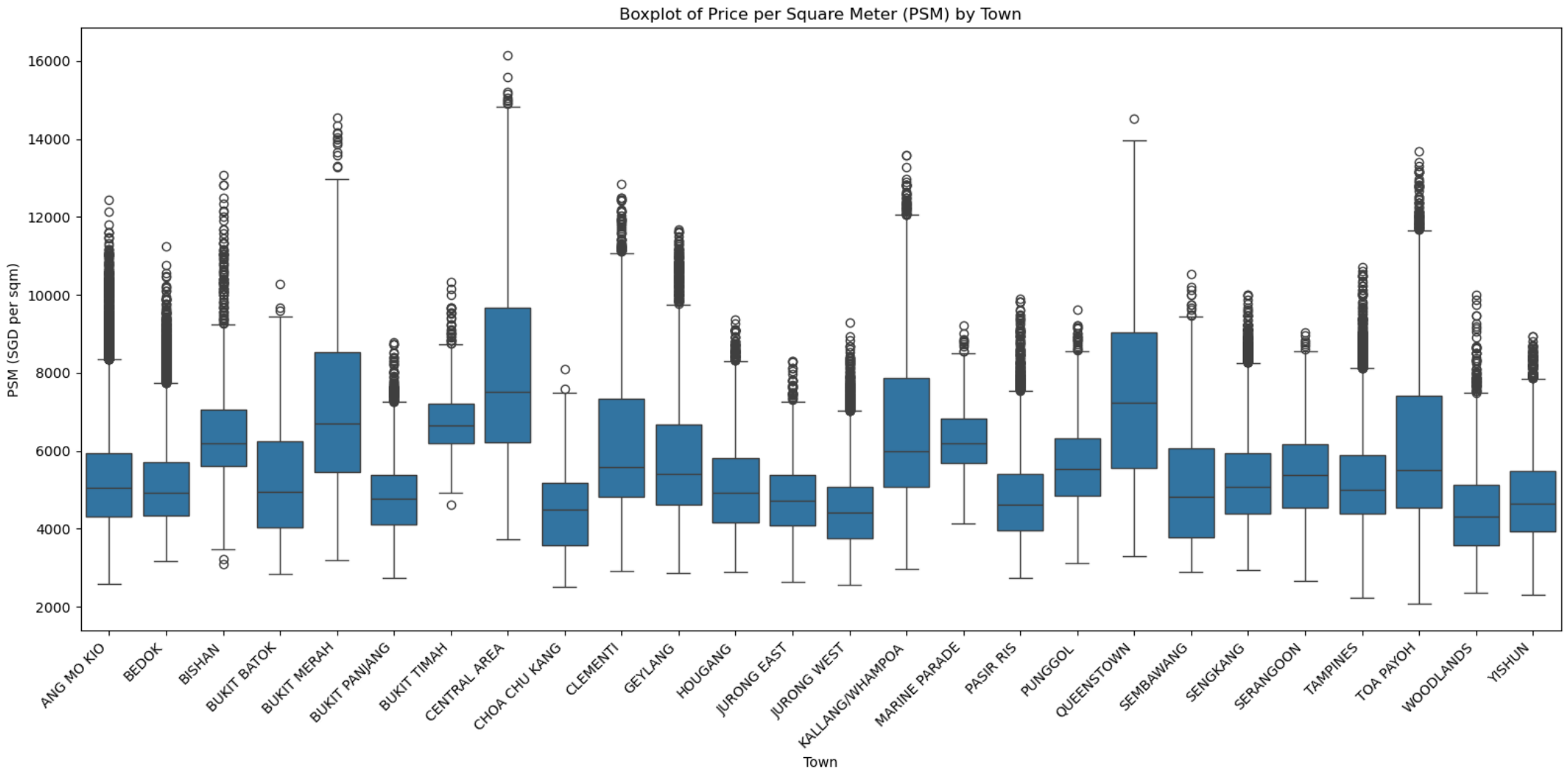


RESALE PRICE IN EACH TOWN

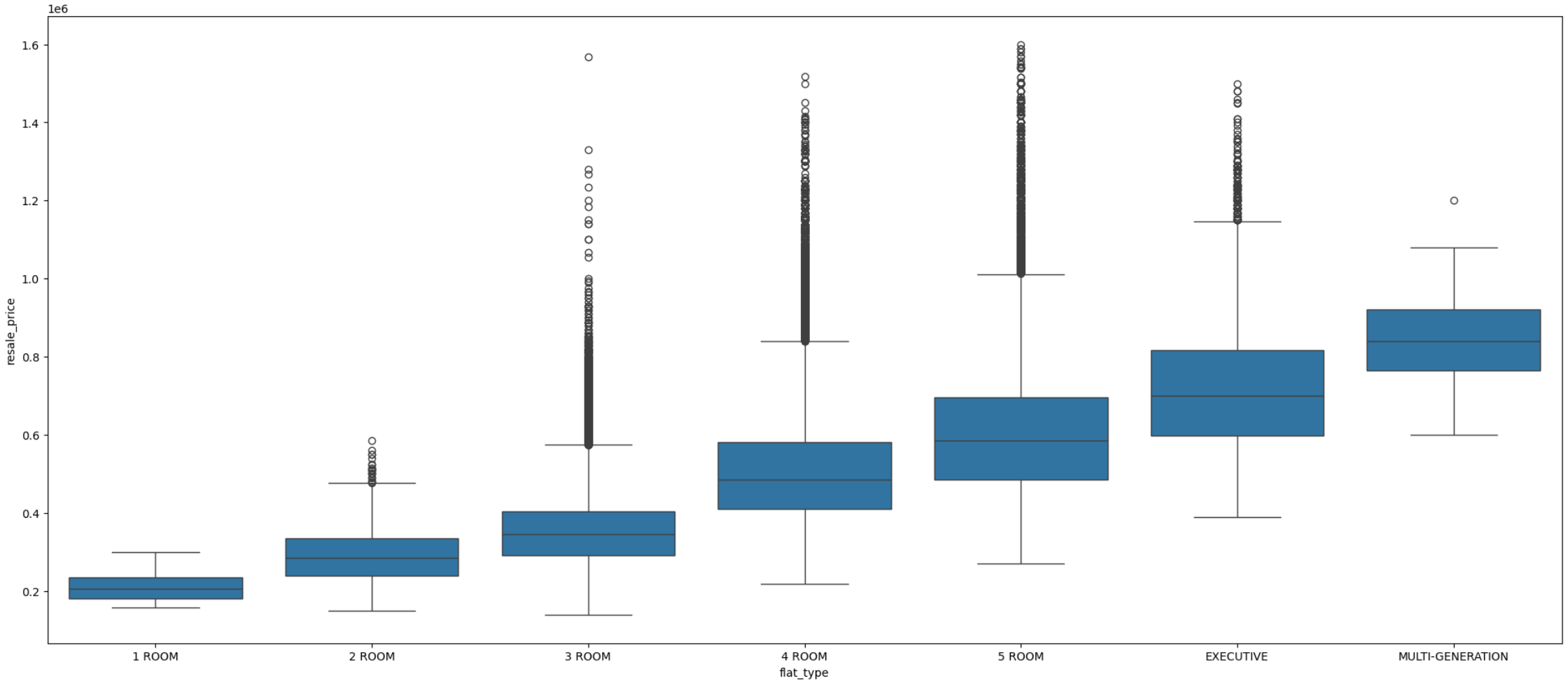


Exploratory Data Analysis (Bivariate Analysis with Respect to Resale Price)

RESALE PRICE IN EACH TOWN

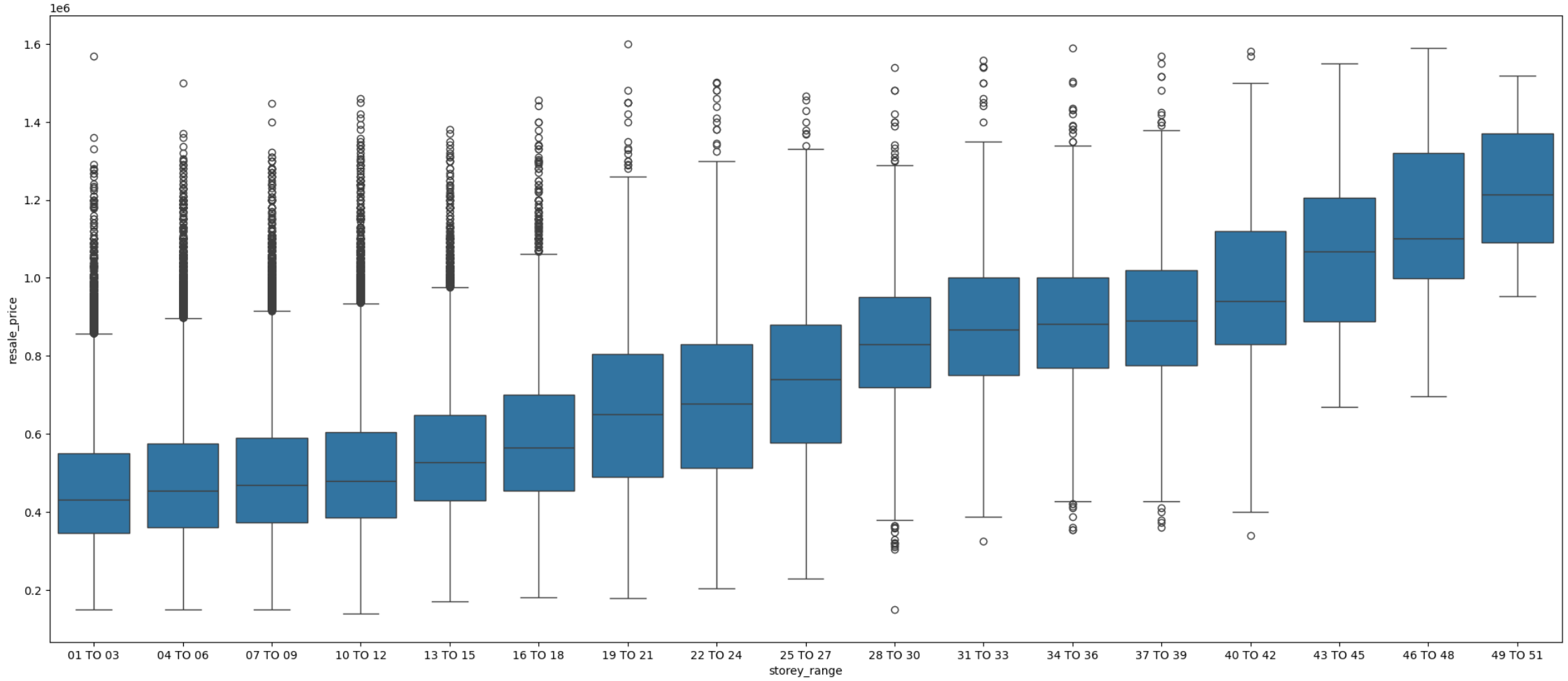


RESALE PRICE FOR EACH FLAT TYPE

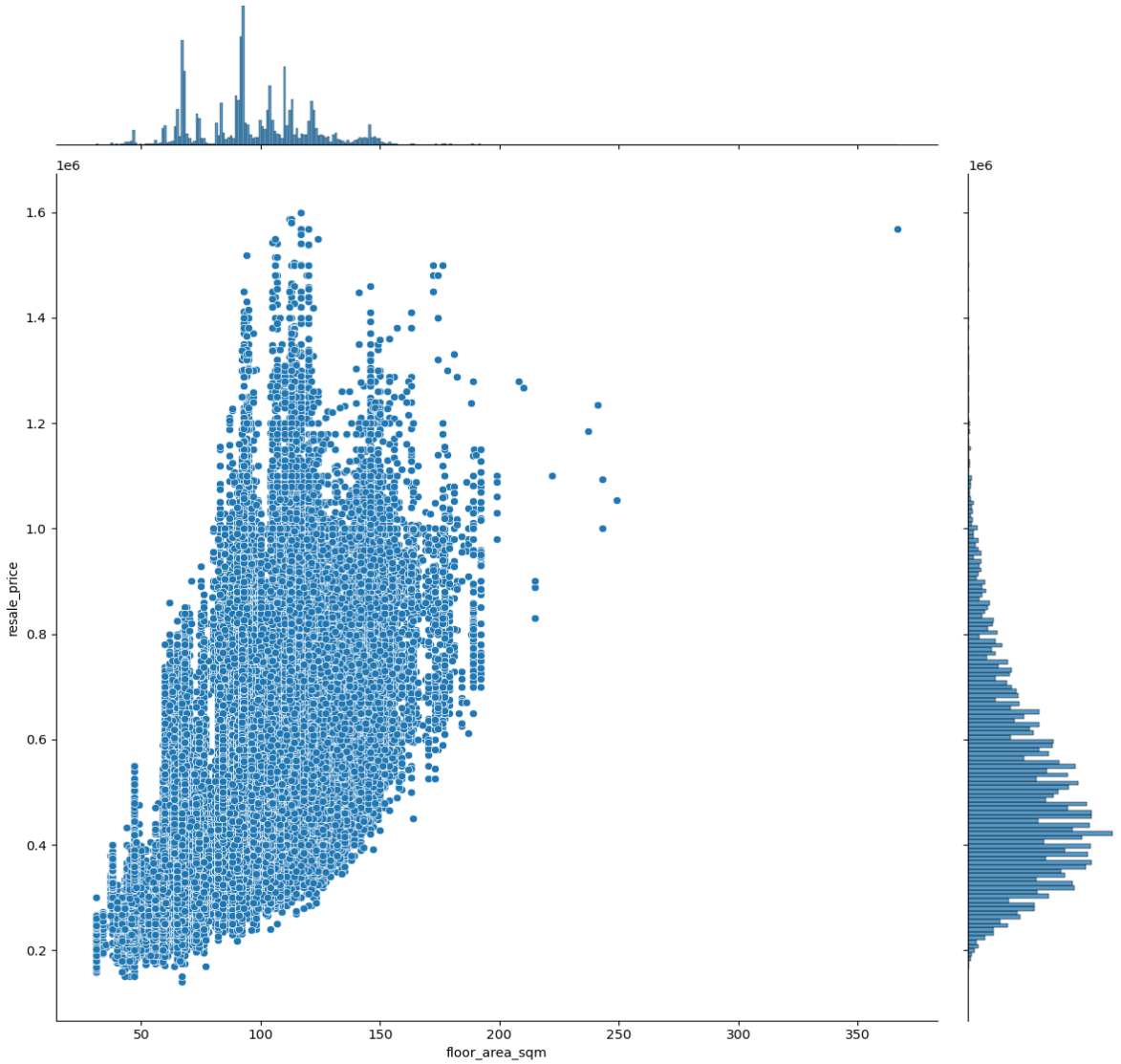


Exploratory Data Analysis (Bivariate Analysis with Respect to Resale Price)

RESALE PRICE FOR EACH STOREY RANGE

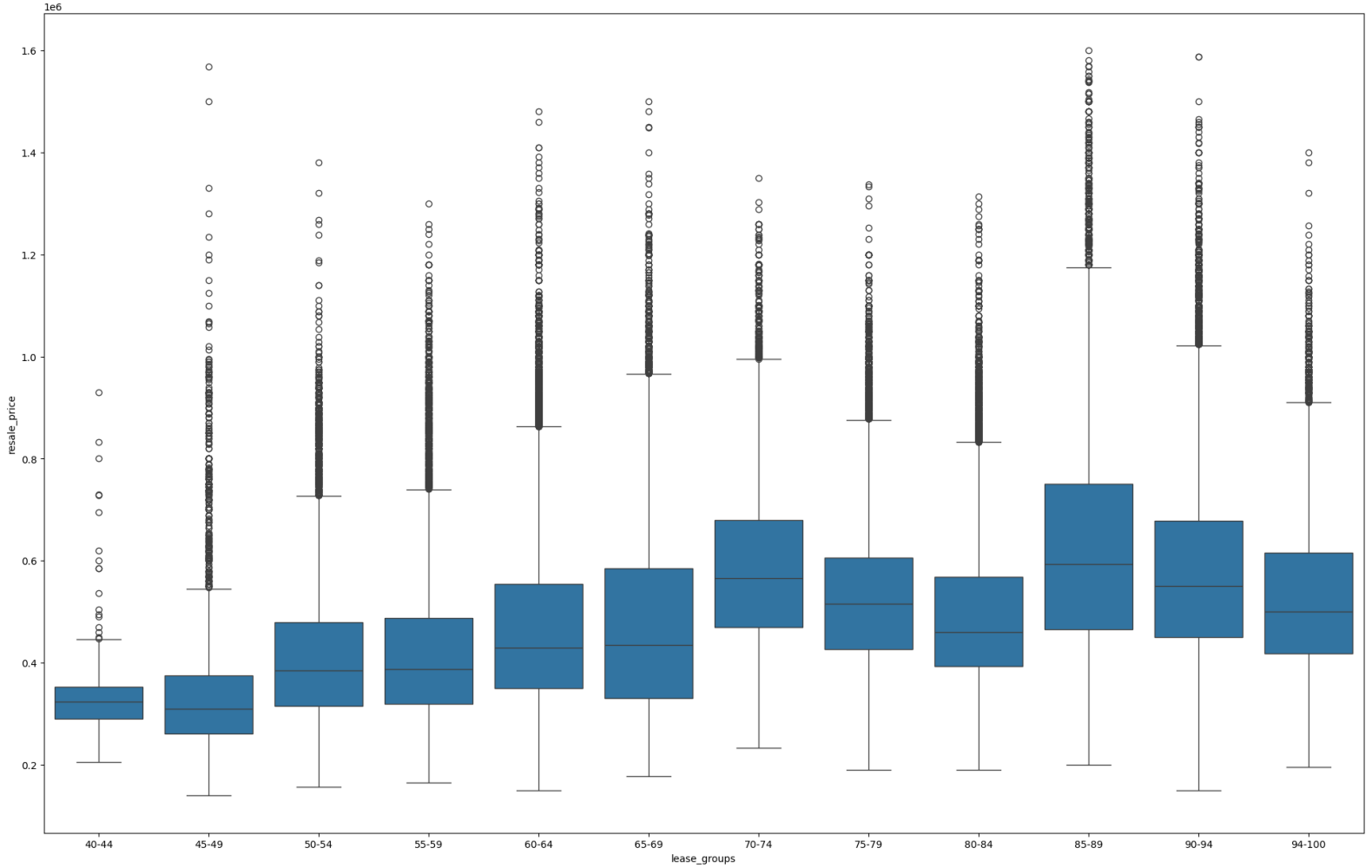


RESALE PRICE FOR FLOOR AREA



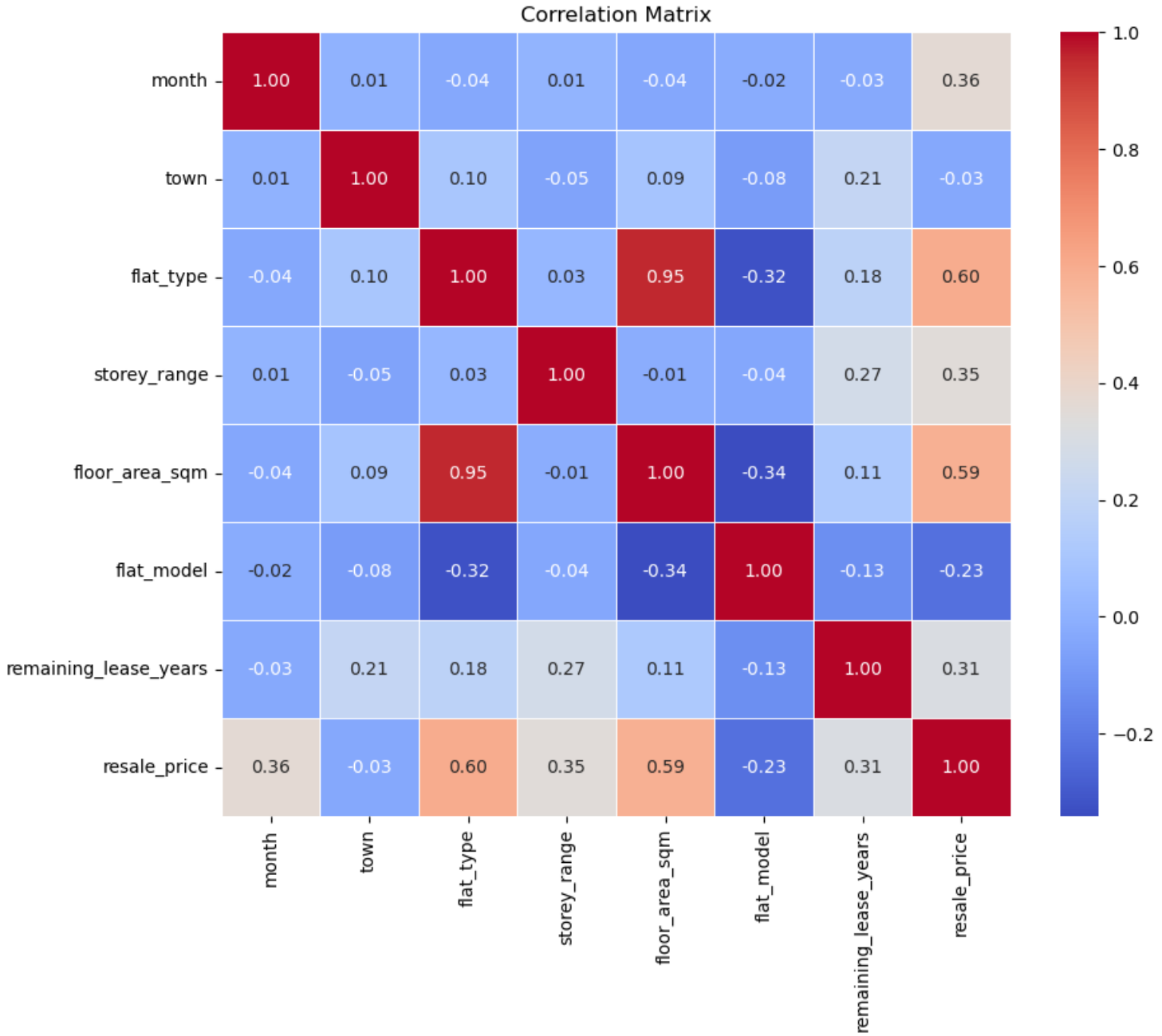
Exploratory Data Analysis (Bivariate Analysis with Respect to Resale Price)

RESALE PRICE FOR REMAINING LEASE YEARS



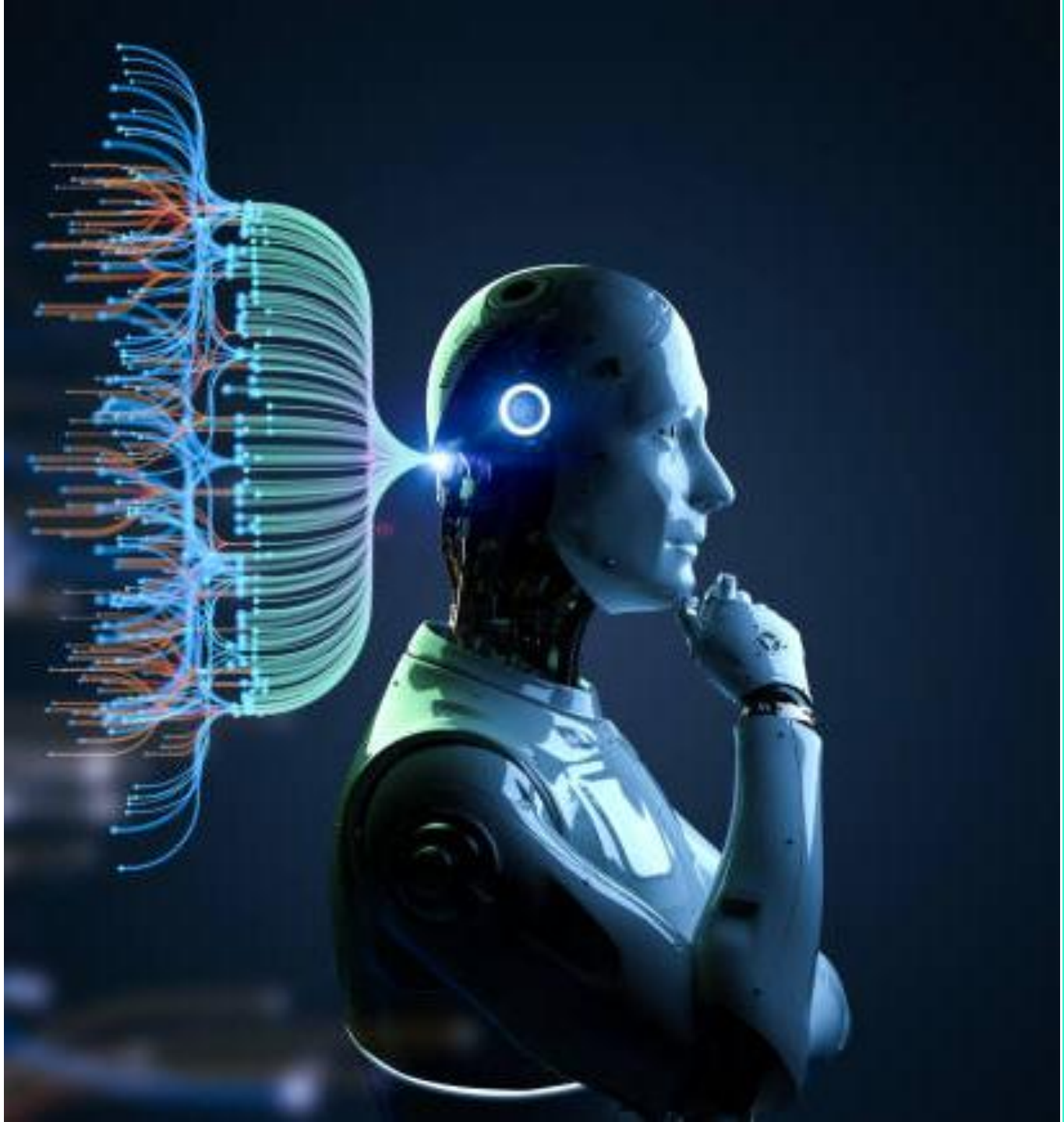
Exploratory Data Analysis (Bivariate Analysis with Respect to Resale Price)

CORRELATION MATRIX



MACHINE LEARNING

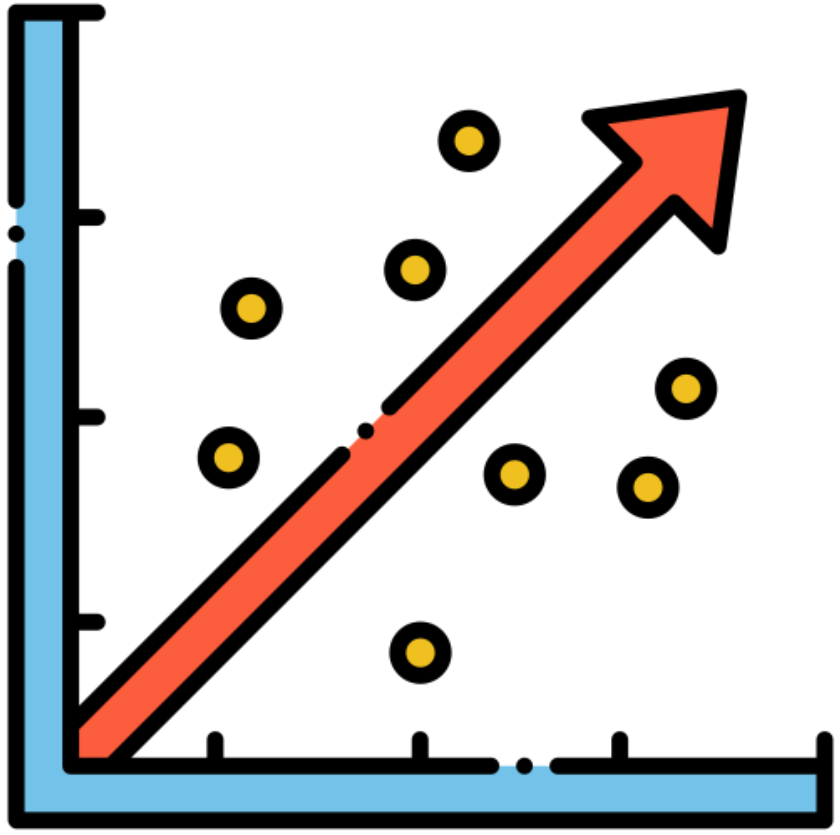
1. Linear Regression
   - Univariate
   - Multivariate
2. Random Forest
   - Regression
   - Classification



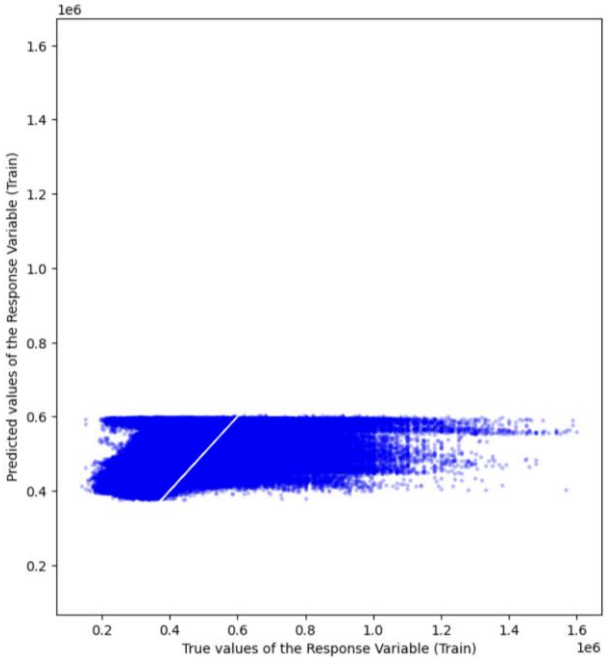
REGRESSION MODELS

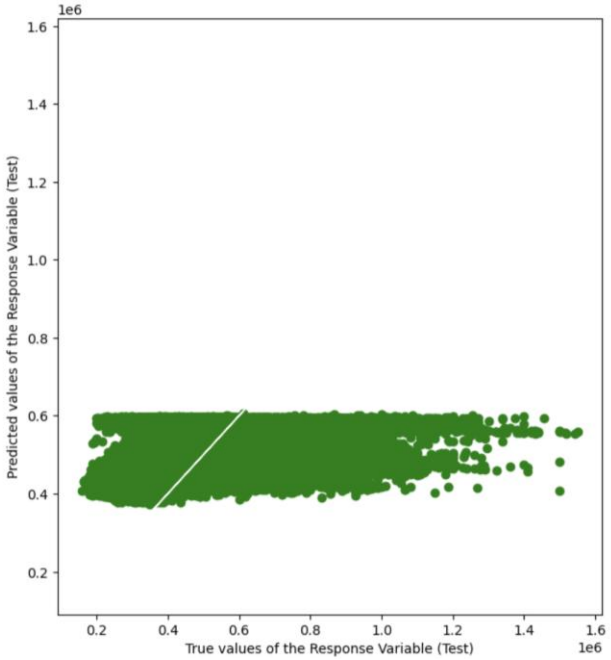
1. Linear Regression
- Univariate
- Multivariate

2. Random Forest Regression



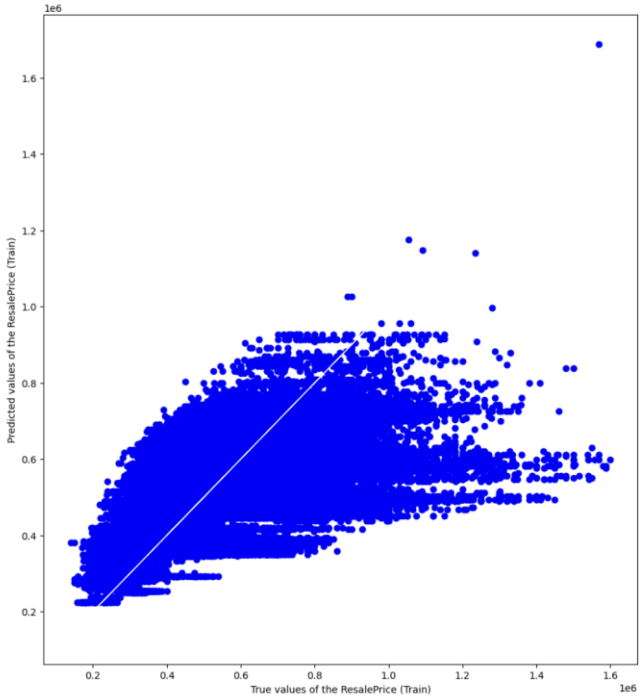
Predictor Feature: Remaining Lease Years

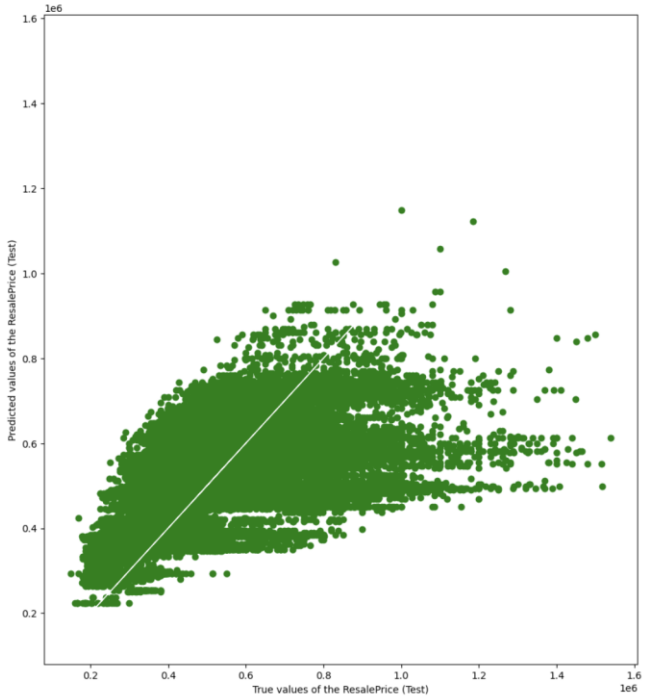





```
Intercept of Regression          : b =  213735.1894158881
Coefficients of Regression       : a =  [4001.62147308]

Goodness of Fit of Model         Train Dataset
Explained Variance (R^2)         : 0.09867421973560608
Root Mean Squared Error (MSE)    : 170626.08049260982

Goodness of Fit of Model         Test Dataset
Explained Variance (R^2)         : 0.09651990981648151
Root Mean Squared Error (RMSE)   : 170217.31058027665
```

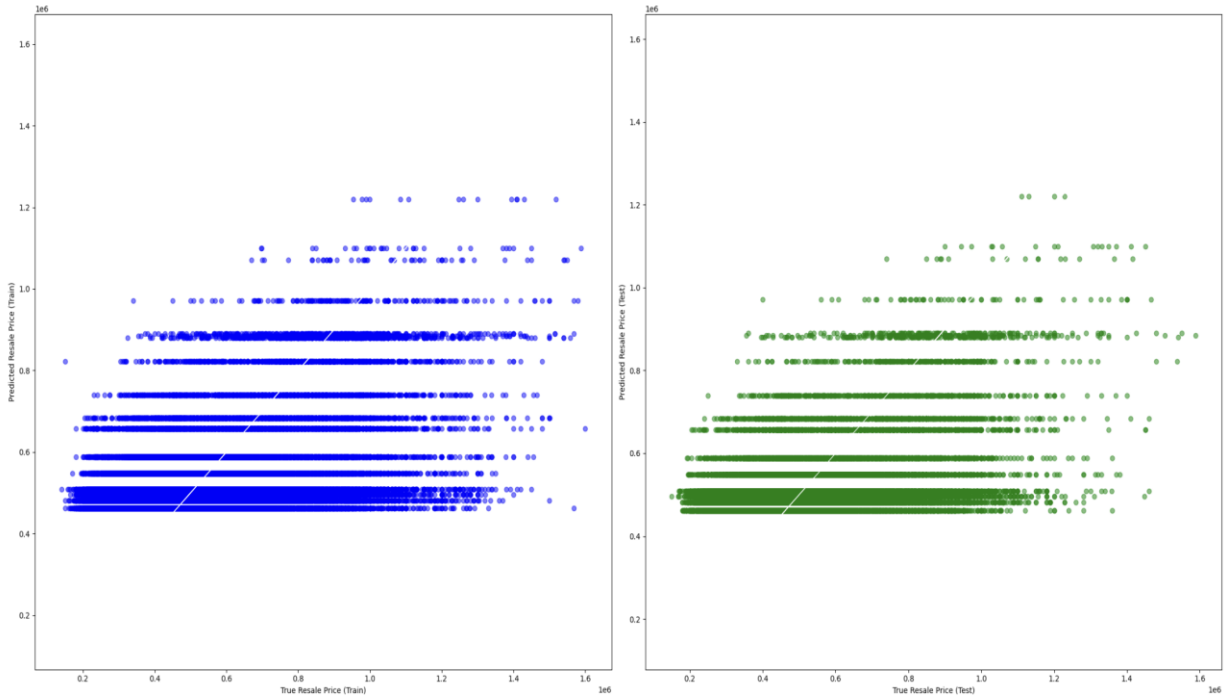
Predictor Feature: Floor Area






```
Intercept of Regression          : b =  [88179.5949147]
Coefficients of Regression       : a =  [[4368.60811938]]

Goodness of Fit of model         Train Dataset
Explained Variance (R^2)         : 0.3410137037289168
Root Mean Squared Error (RMSE)   : 145895.7731101697

Goodness of Fit of model         Test Dataset
Explained Variance (R^2)         : 0.34616025904006287
Root Mean Squared Error (RMSE)   : 144803.84305481025
```

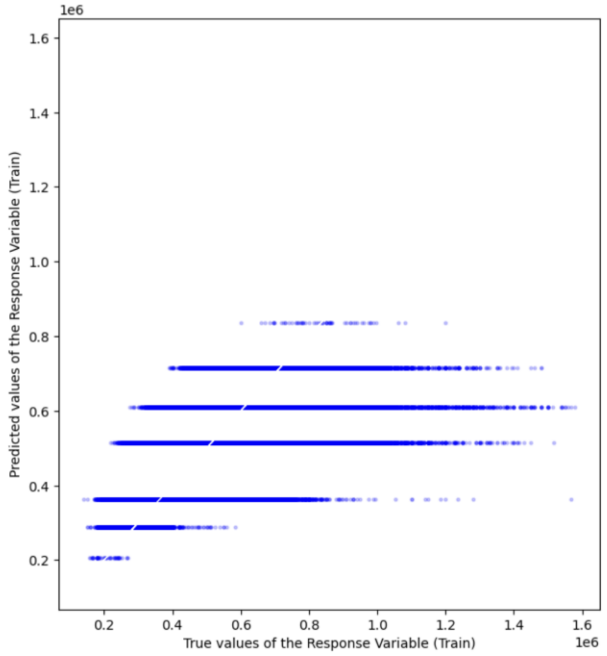
Predictor Feature: Storey Range

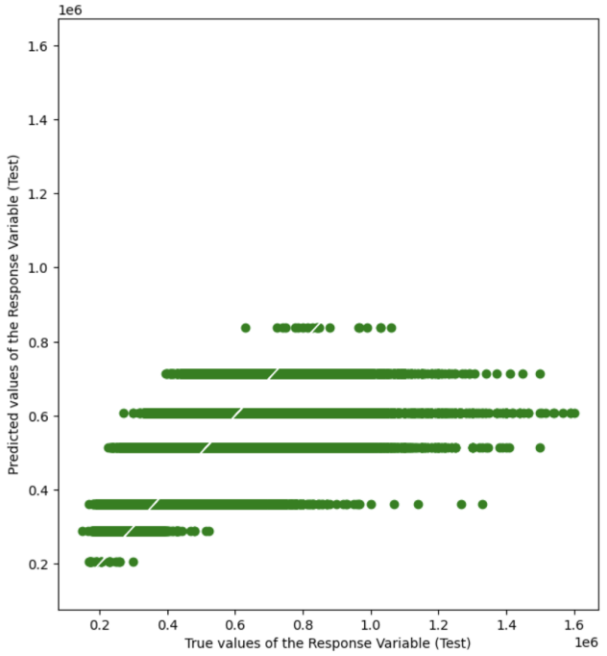



```
Intercept of Regression         : b =  186047341583084.22
Coefficients of Regression      : a =
 storey_range_01 TO 03   -1.860473e+14
storey_range_04 TO 06    -1.860473e+14
storey_range_07 TO 09    -1.860473e+14
storey_range_10 TO 12    -1.860473e+14
storey_range_13 TO 15    -1.860473e+14
storey_range_16 TO 18    -1.860473e+14
storey_range_19 TO 21    -1.860473e+14
storey_range_22 TO 24    -1.860473e+14
storey_range_25 TO 27    -1.860473e+14
storey_range_28 TO 30    -1.860473e+14
storey_range_31 TO 33    -1.860473e+14
storey_range_34 TO 36    -1.860473e+14
storey_range_37 TO 39    -1.860473e+14
storey_range_40 TO 42    -1.860473e+14
storey_range_43 TO 45    -1.860473e+14
storey_range_46 TO 48    -1.860473e+14
storey_range_49 TO 51    -1.860473e+14
dtype: float64

Goodness of Fit of model        Train Dataset
Explained Variance (R²)         : 0.13535325785691277
Root Mean Squared Error (RMSE)  : 167118.24748966802

Goodness of Fit of model        Test Dataset
Explained Variance (R²)         : 0.1373880993740212
Root Mean Squared Error (RMSE)  : 166322.9402548964
```

Predictor Feature: Flat Type

```
Intercept of Regression        : b =  260514827122005.34
Coefficients of Regression     : a =  [-2.60514827e+14 -2.60514827e+14 -2.60514827e+14 -2.60514827e+14
 -2.60514827e+14 -2.60514826e+14 -2.60514826e+14]
```

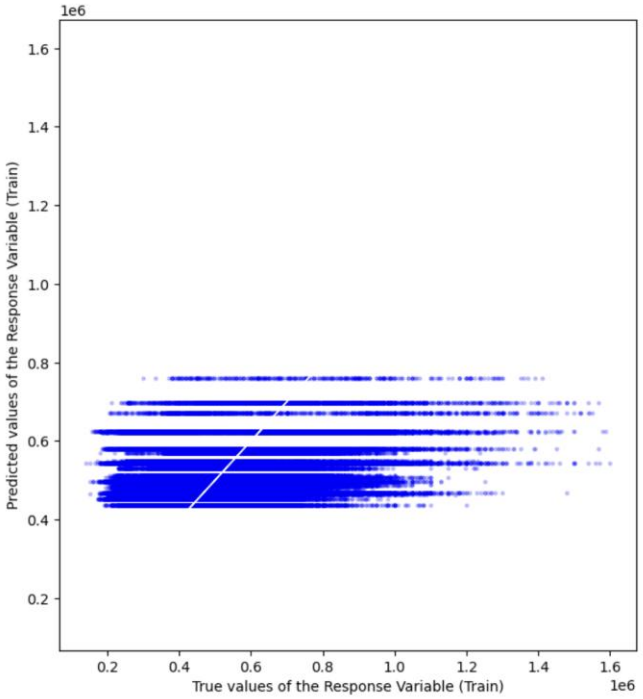


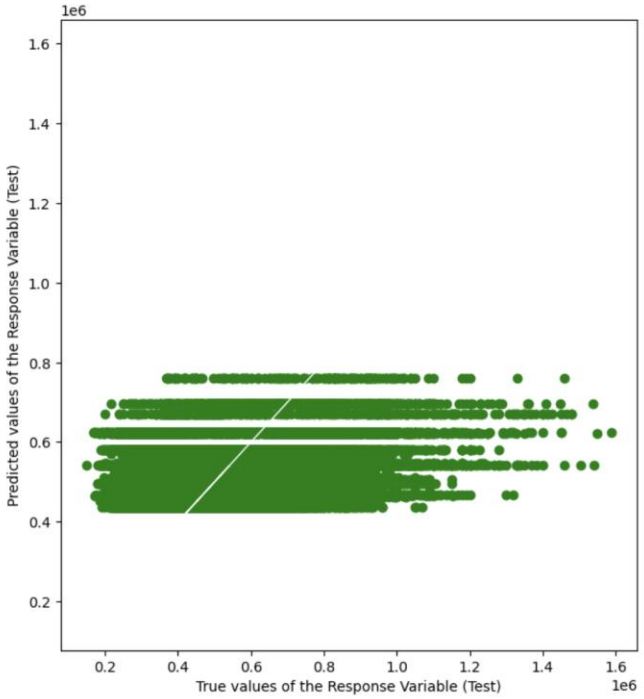



```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.36184510469252396
Root Mean Squared Error (RMSE)  : 143571.27940032273

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.3686097554285708
Root Mean Squared Error (RMSE)  : 142296.2204470233
```

Predictor Feature: Town

```
Intercept of Regression         : b =  -318128937125997.5
Coefficients of Regression      : a =  [3.18128938e+14 3.18128938e+14 3.18128938e+14 3.18128938e+14
 3.18128938e+14 3.18128938e+14 3.18128938e+14 3.18128938e+14
 3.18128938e+14 3.18128938e+14 3.18128938e+14 3.18128938e+14
 3.18128938e+14 3.18128938e+14 3.18128938e+14 3.18128938e+14
 3.18128938e+14 3.18128938e+14 3.18128938e+14 3.18128938e+14
 3.18128938e+14 3.18128938e+14]
```





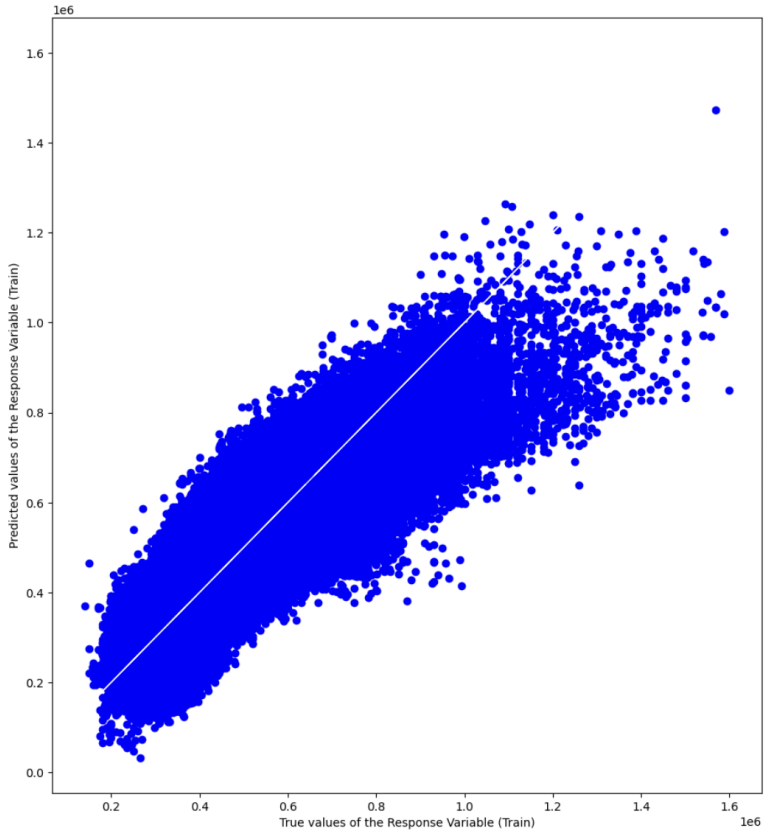

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.10230689153276218
Root Mean Squared Error (RMSE)  : 170281.89065604398

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.1031450431965385
Root Mean Squared Error (RMSE)  : 169592.0686047823
```

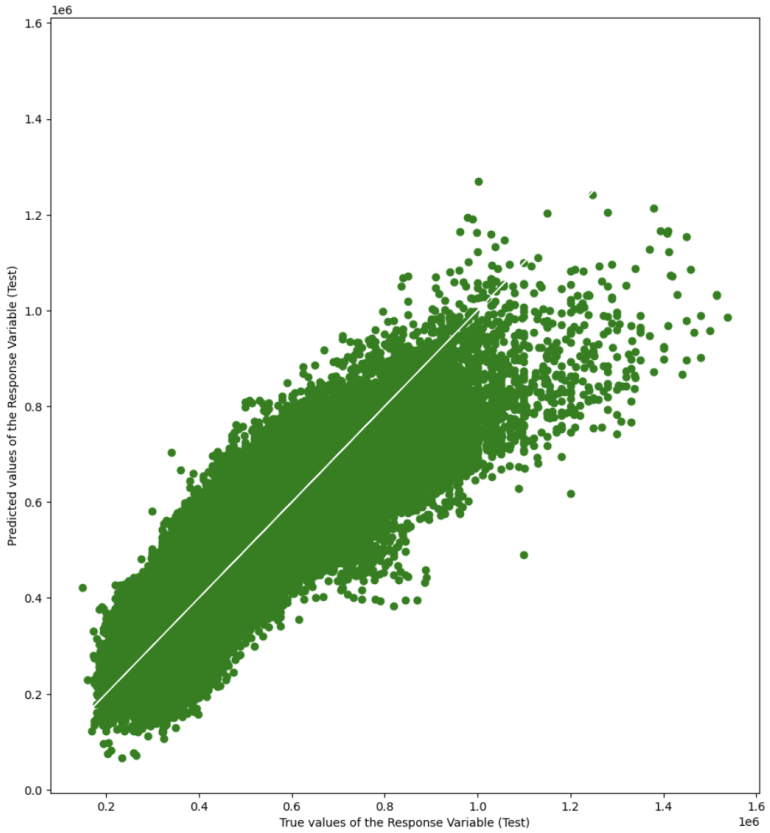
Multi-Variate Linear Regression

Predictor Features:

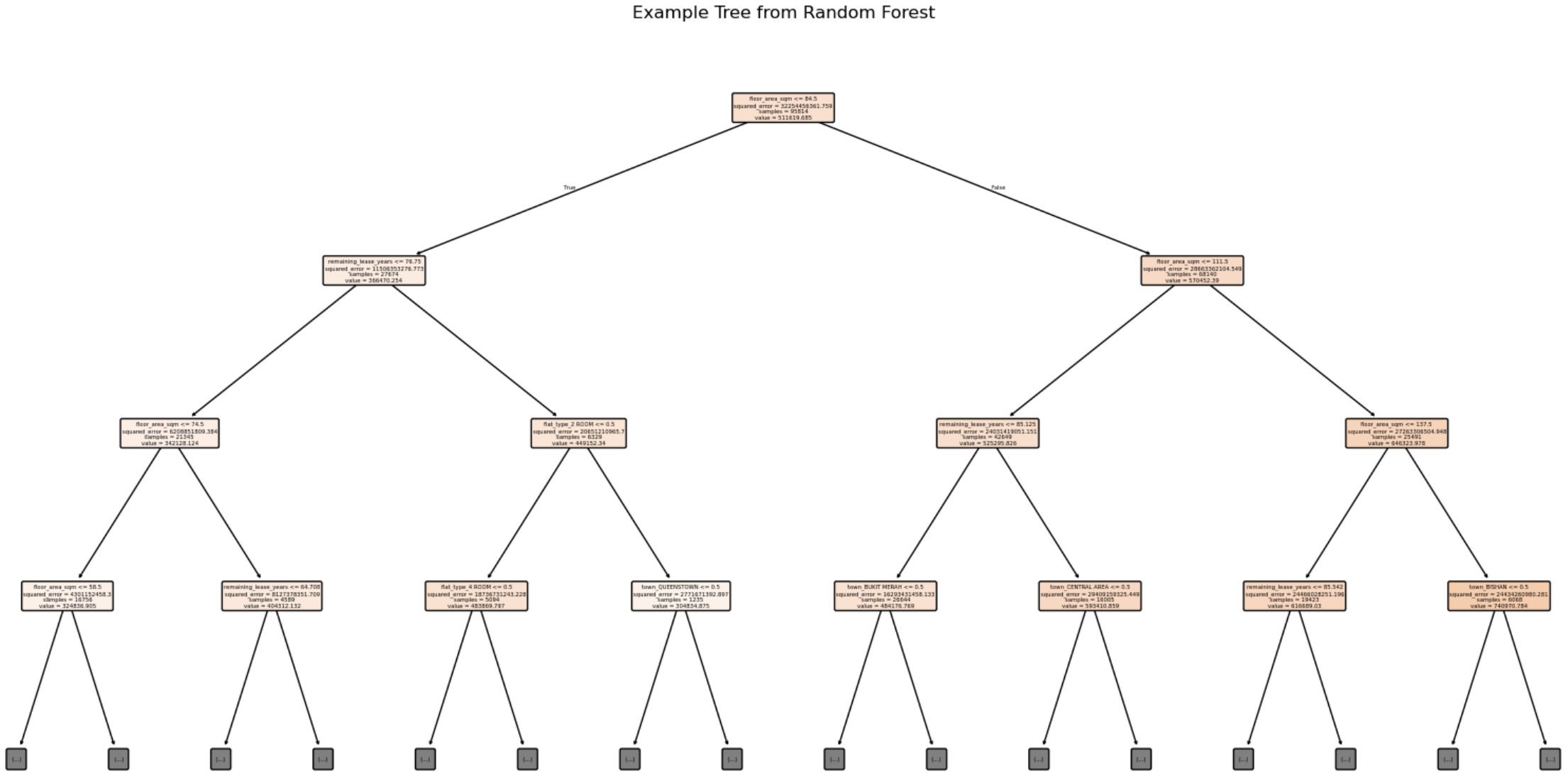
1. Floor Area
2. Remaining Lease Years
3. Storey Range
4. Flat Type
5. Town

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.6819661339147194
Root Mean Squared Error (RMSE)    : 101354.07216340605

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.6842400008377726
Root Mean Squared Error (RMSE)    : 100628.96069128852
```

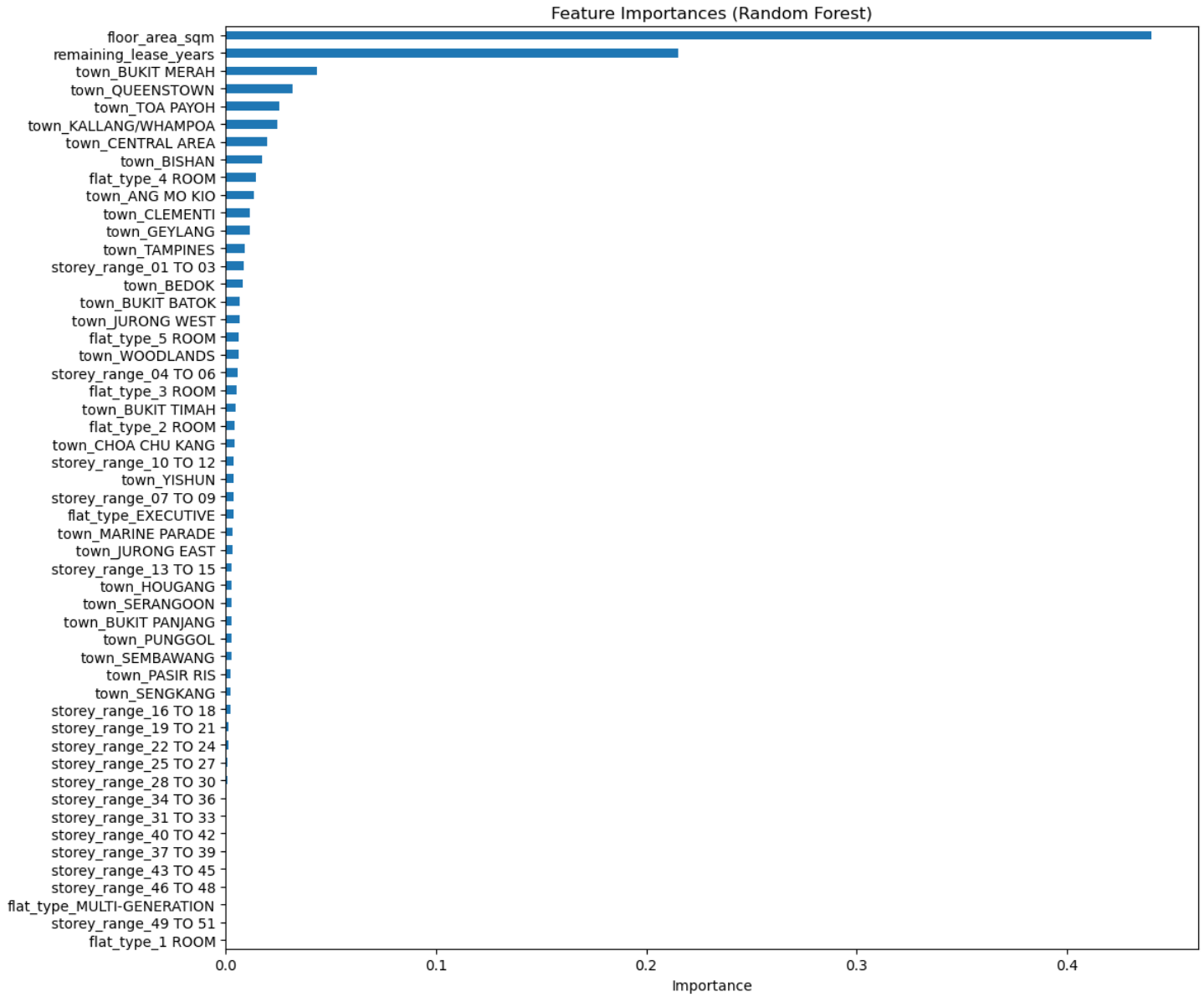




Random Forest Regression



Random Forest Regression



Random Forest Performance Metrics

TRAIN:

R² : 0.9674

MSE : 1052792932.67

TEST:

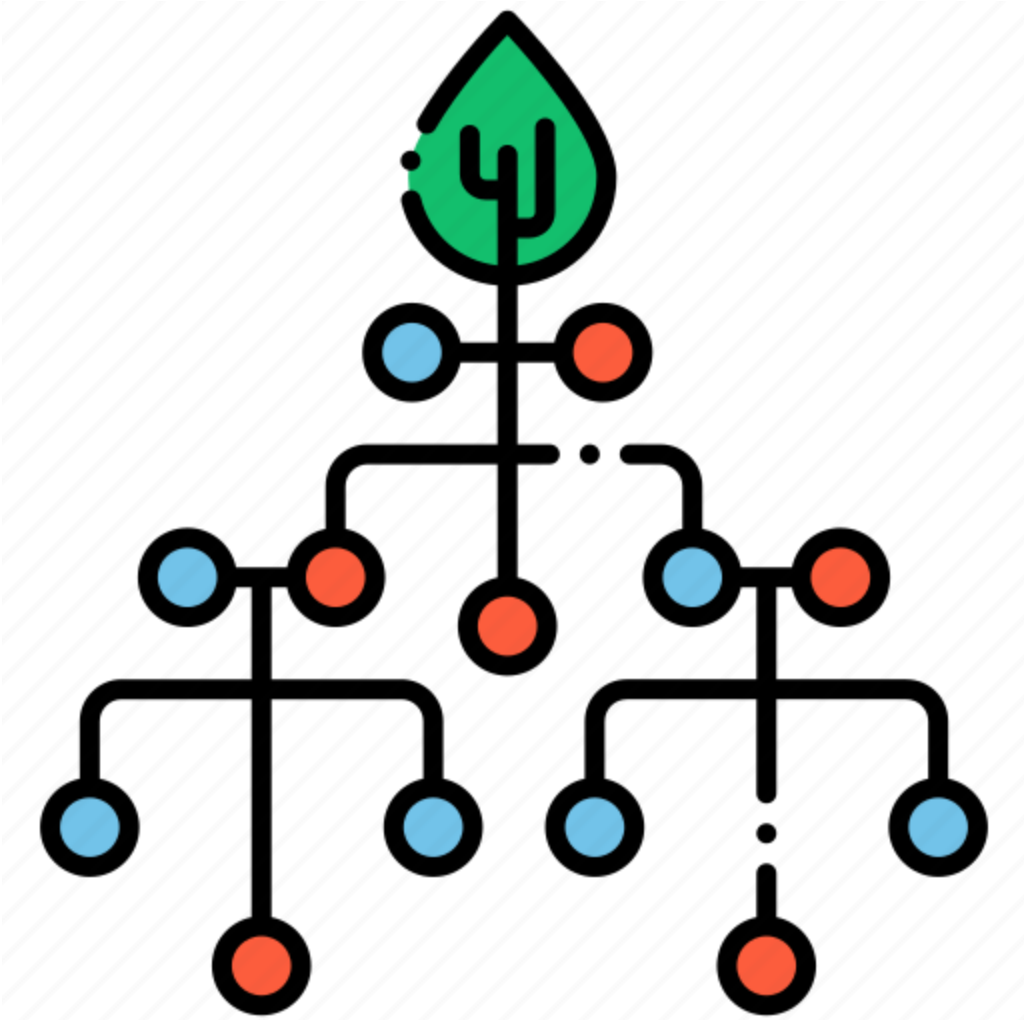
R² : 0.8807

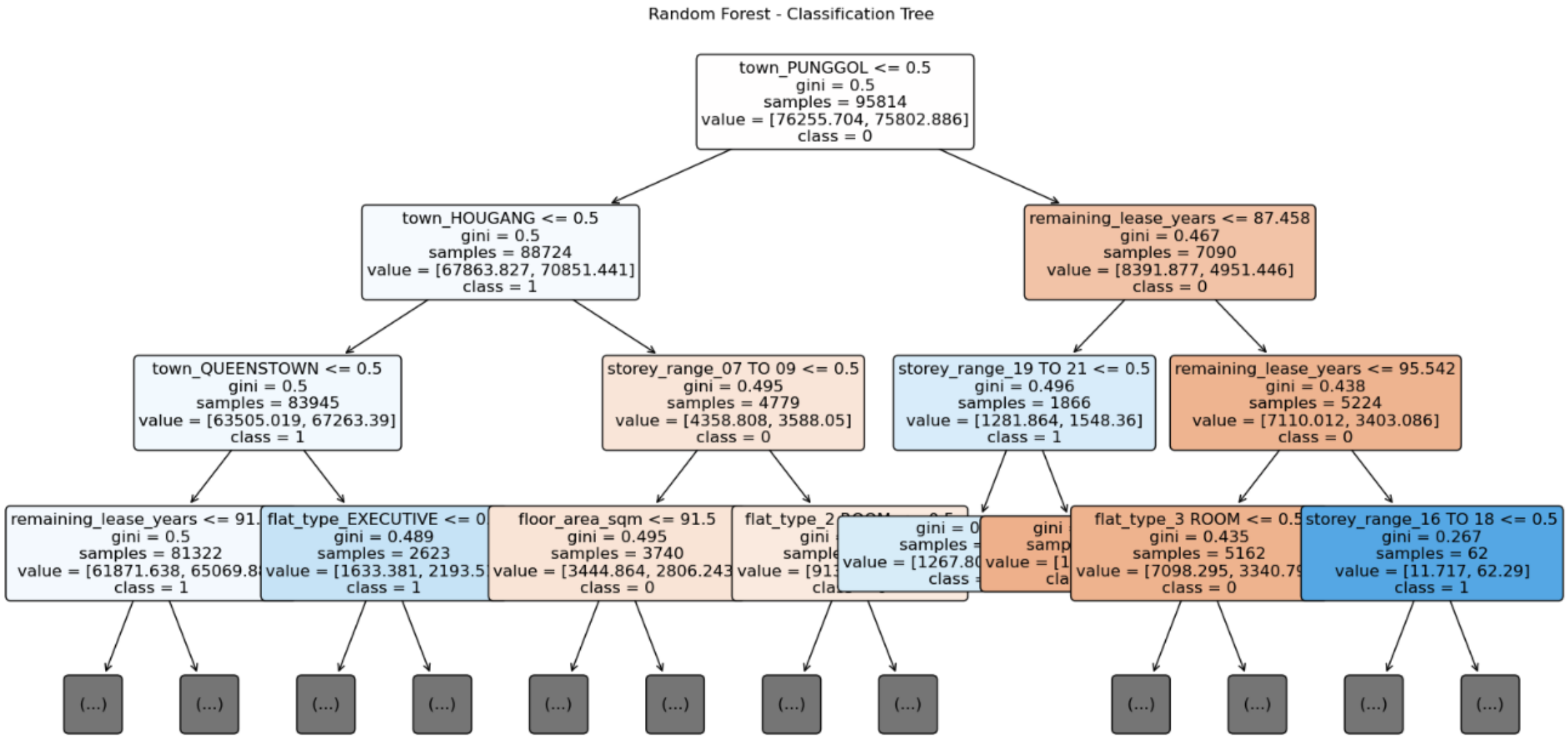
MSE : 3827168889.36

Multi-Variate Linear Regression VS Random Forest Regression

Model	**Multi-Variate Linear Regression**	**Random Forest Regression**
R^2 (Train)	0.682	0.967
RMSE (Train)	101,354	32,446
R^2 (Test)	0.684	0.881
RMSE (Test)	100,628	61,864

RANDOM FOREST CLASSIFICATION




Random Forest - Classification Tree
town_PUNGGOL <= 0.5
gini = 0.5
samples = 95814
value = [76255.704, 75802.886]
class = 0
town_HOUGANG <= 0.5
gini = 0.5
samples = 88724
value = [67863.827, 70851.441]
class = 1
remaining_lease_years <= 87.458
gini = 0.467
samples = 7090
value = [8391.877, 4951.446]
class = 0
town_QUEENSTOWN <= 0.5
gini = 0.5
samples = 83945
value = [63505.019, 67263.39]
class = 1
storey_range_07 TO 09 <= 0.5
gini = 0.495
samples = 4779
value = [4358.808, 3588.05]
class = 0
storey_range_19 TO 21 <= 0.5
gini = 0.496
samples = 1866
value = [1281.864, 1548.36]
class = 1
remaining_lease_years <= 95.542
gini = 0.438
samples = 5224
value = [7110.012, 3403.086]
class = 0
remaining_lease_years <= 91.
gini = 0.5
samples = 81322
value = [61871.638, 65069.8
class = 1
flat_type_EXECUTIVE <= 0
gini = 0.489
samples = 2623
value = [1633.381, 2193.5
class = 1
floor_area_sqm <= 91.5
gini = 0.495
samples = 3740
value = [3444.864, 2806.243
class = 0
flat_type_3 ROOM <= 0.5
gini = 0.435
samples = 5162
value = [7098.295, 3340.7
class = 0
storey_range_16 TO 18 <= 0.5
gini = 0.267
samples = 62
value = [11.717, 62.29]
class = 1
(...)

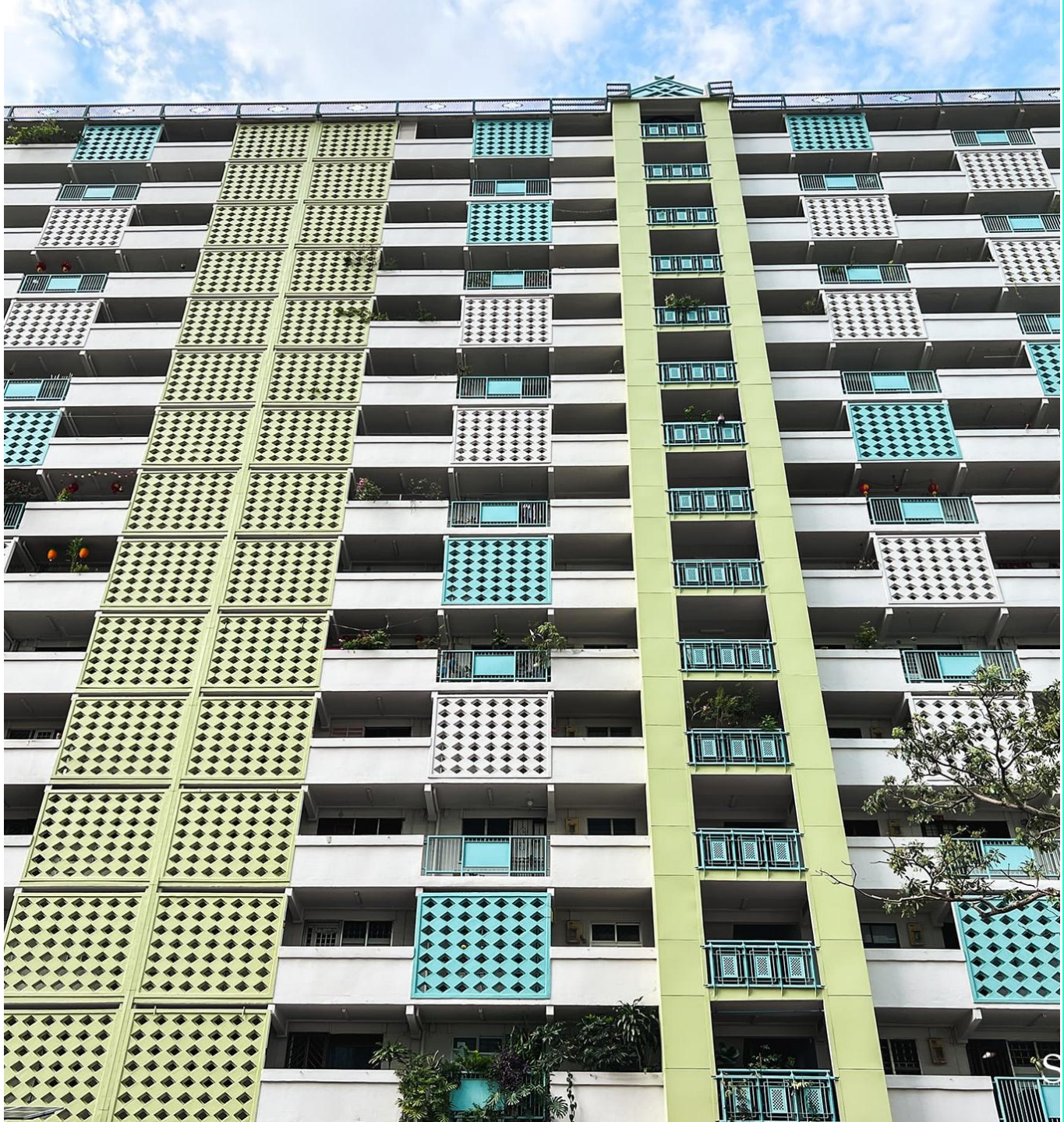
- **Precision**: $\frac{TruePositives}{TruePositives+FalsePositives}$
- **Recall**: $\frac{TruePositives}{TruePositives+FalseNeagtives}$
- **F1 score**: 2 * $\frac{Precision*Recall}{Precision+Recall}$

Classification Report:

	precision	recall	f1-score	support
0	0.39	0.61	0.47	10796
1	0.88	0.73	0.80	39804
accuracy			0.71	50600
macro avg	0.63	0.67	0.64	50600
weighted avg	0.77	0.71	0.73	50600

ROC AUC Score: 0.7442
Average Precision: 0.9060

CONCLUSION



CONCLUSION

Outcome

- We were able to develop a predictive model for HDB Resale prices using a Random Forest Regression model which performed significantly better than Multivariate Linear Regression

```
print("Central Area:", predict_resale_price(107, 84, '22 TO 24', '3 ROOM', 'CENTRAL AREA'))
print("Yishun:", predict_resale_price(64, 60, '06 TO 10', '2 ROOM', 'YISHUN'))
print("Sengkang:", predict_resale_price(141, 74, '06 TO 10', '5 ROOM', 'SENGKANG'))
print("Serangoon:", predict_resale_price(105, 58, '16 TO 18', '3 ROOM', 'SERANGOON'))
```

```
Central Area: 756760.0
Yishun: 404869.99
Sengkang: 769417.71
Serangoon: 724178.04
```

CONCLUSION

Outcome

- We then successfully implemented a Random Forest Classification model that assessed the fairness of our price predictions

```
Actual Price: $450,000.00
Predicted Price: $527,841.20
Fair Price Range: $473,000.00 – $590,000.00
Price Difference: $-77,841.20 (-14.7%)

Verdict: UNFAIR
– The price is OUTSIDE the fair range
```

CONCLUSION

Insights

- Floor area and remaining lease years were the strongest predictors of resale price
- About 78% of homebuyers pay a fair price for their HDB resale flat, suggesting that the real estate market is relatively efficient
- According to our model, the mean absolute difference between predicted and actual resale price was $27,000
- Certain towns such as Bukit Merah and Queenstown were shown to have a noticeable impact on resale prices, likely due to their premium location

CONCLUSION

Recommendations

- Purchasing a resale flat is a lifelong investment that requires careful planning and analysis before making such an expensive decision.
- Overall, we hope to provide homebuyers as well as sellers with a model that helps make rational and informed decisions using past real-life data provided by the housing development board.
- Adding geospatial data points such as proximity to amenities, transportation hubs and schools can help us achieve more accurate predictions

CONCLUSION

Takeaways

- Always room for growth: Initially, we thought that using linear regression was sufficient in predicting resale price. However, we went beyond the syllabus, incorporating random forest regression which was a significantly better model

- The importance of evaluation metrics: Utilizing evaluation metrics is important in helping us understand how well our models are performing and guides improvements

- Not everything is what it seems: According to our model, 22% of transactions in the HDB dataset are “unfair”. When exploring datasets, we should analyse the data closely rather than taking them at face value as some data points may be skewed, even if its data from a reliable source like the government

Thank You!

Presented by

Lim Wei Jian, David (U2430324H)

See Pei Jin, Ernest (U2430223J)