Predicting HDB Resale Prices

Russell, Simon, Clifton, Kathy

Problem Introduction

Problem Introduction

Singapore HDBs are resold at various prices

The resale price is affected by:

- floor area
- lease year
- flat type
- and many more!

Two Goals

PREDICTION
Predict resale prices

INFERENCE
Analyze how each feature contributes to resale price

Dataset Description

Dataset Description

Government resale flat data from **data.gov.sg** managed by Housing Development Board (HDB)

4410 resale transactions taken from Jan - Feb 2023

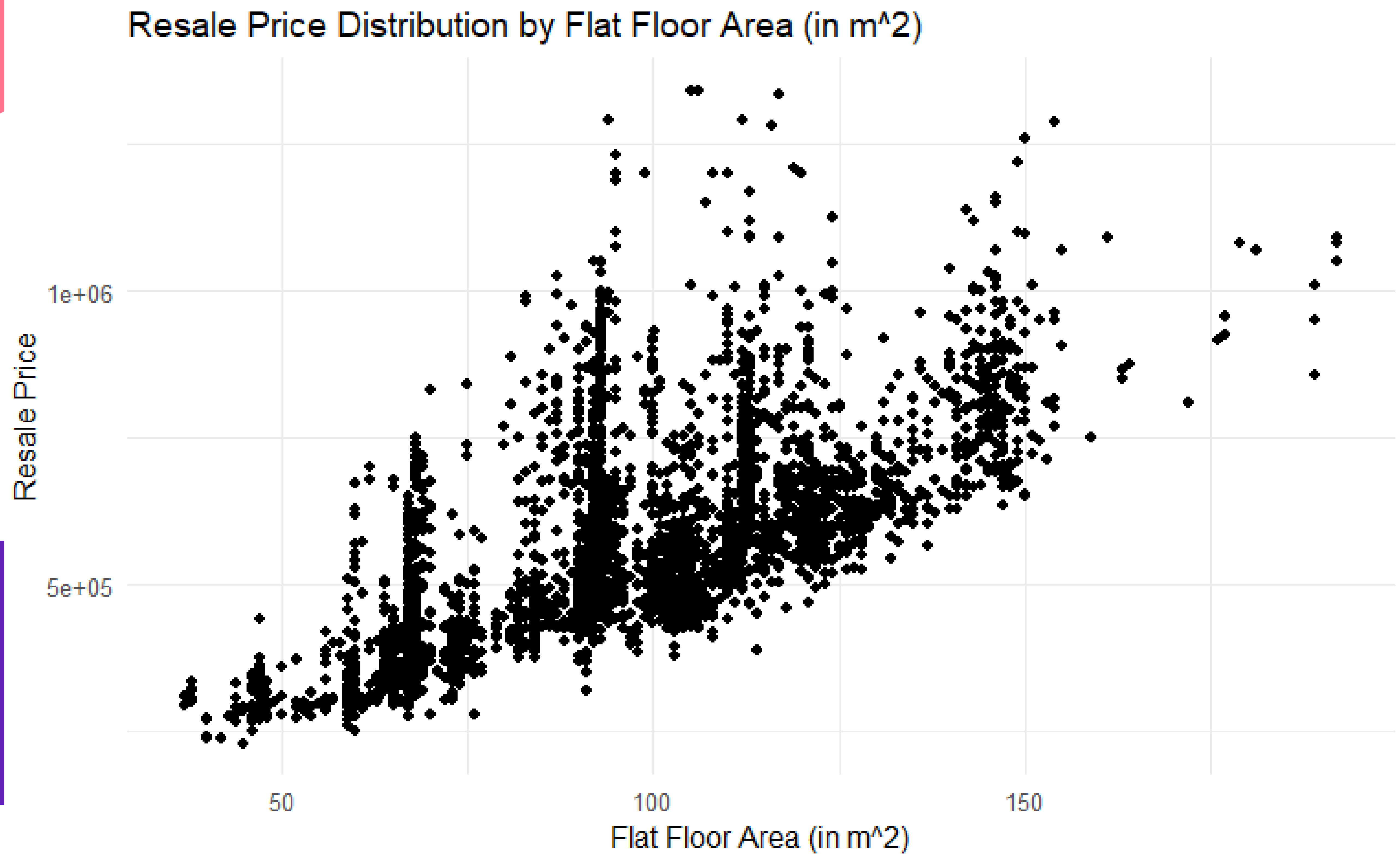
Dataset Description

11 variables:

1. month
2. town
3. flat_type
4. block
5. street_name
6. storey_range
7. floor_area_sqm
8. flat_model
9. lease_commence_date
10. remaining_lease
11. resale_price (response)

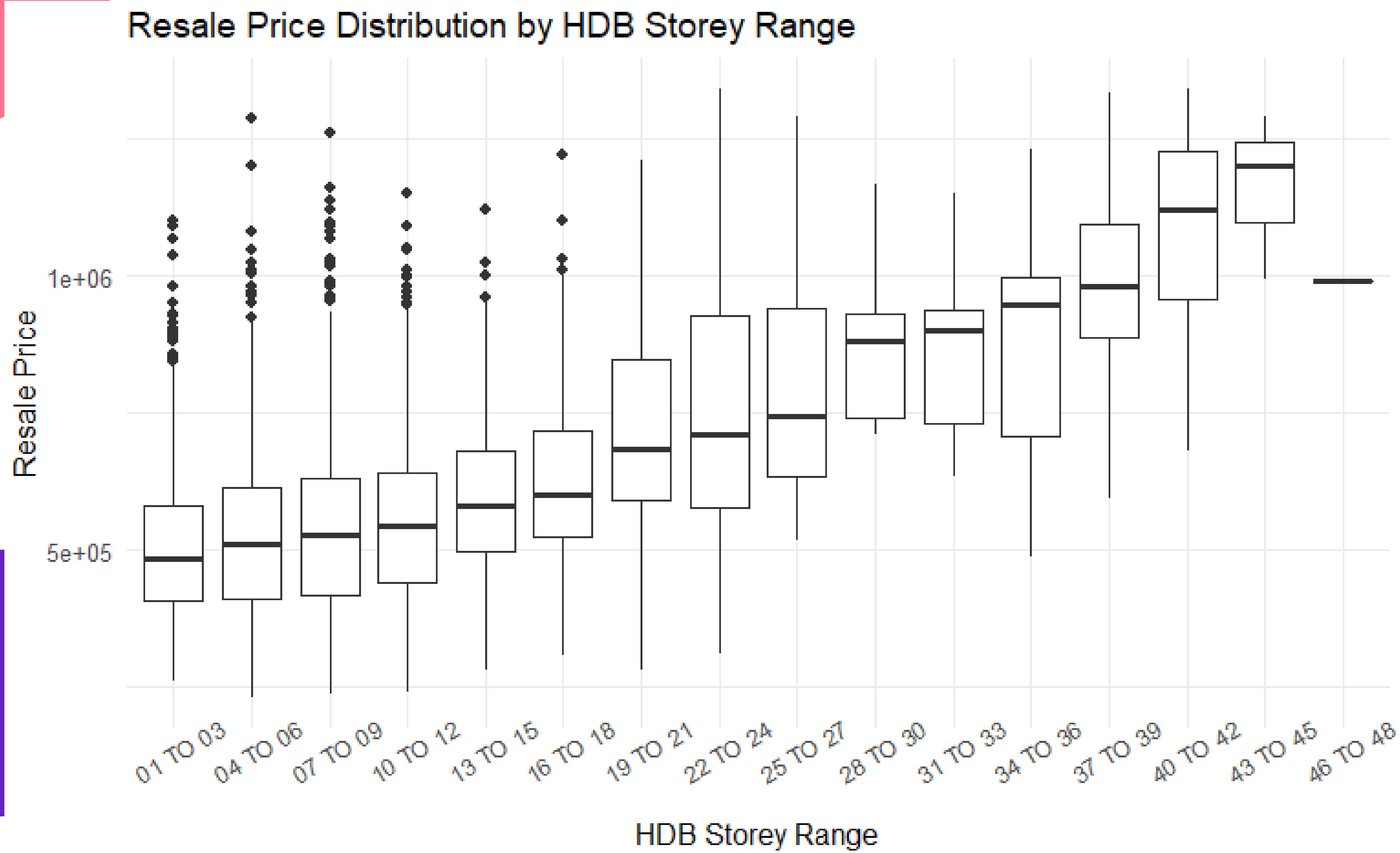
Exploratory Data Analysis

Exploratory Data Analysis



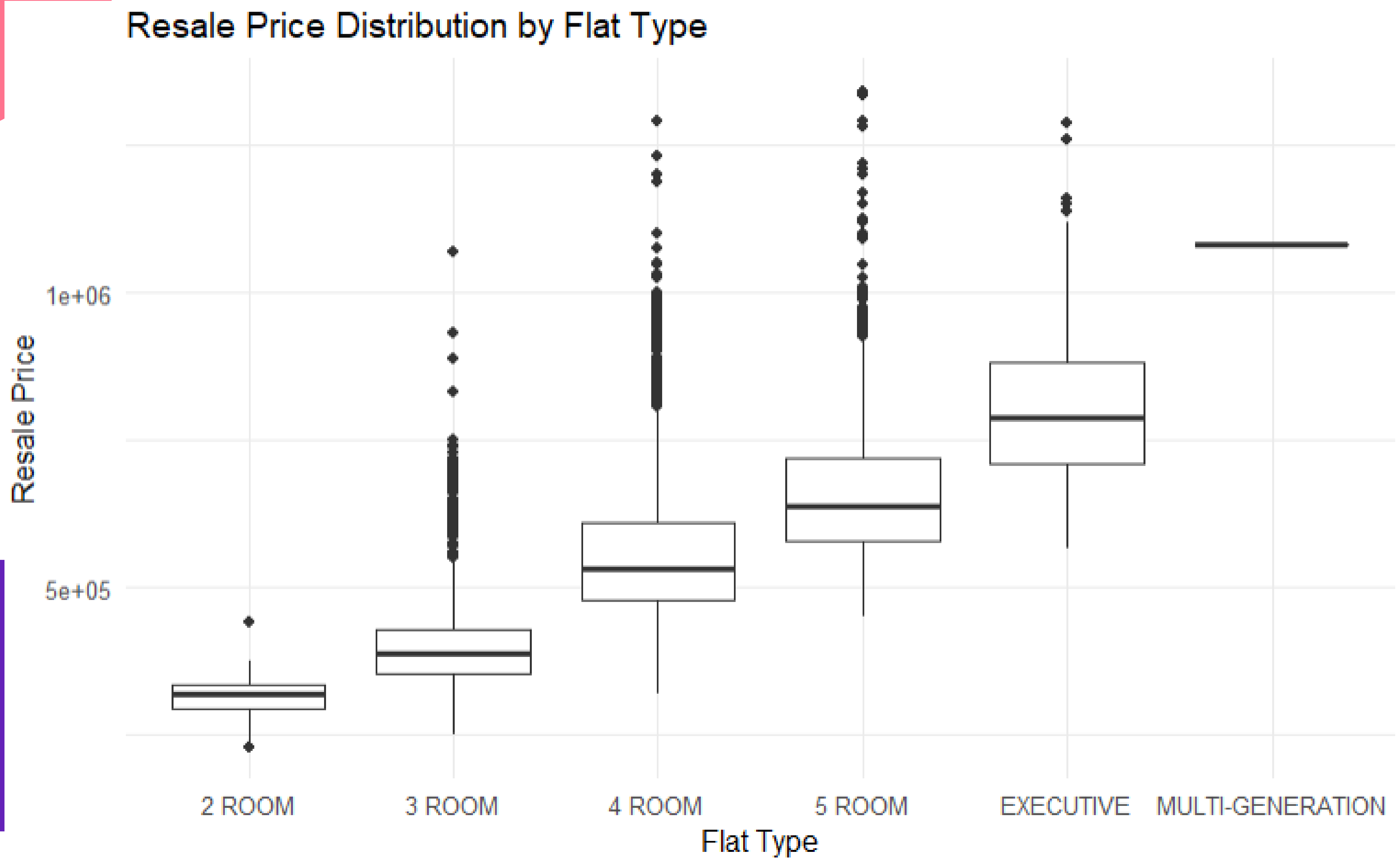
Noticeable upward trend!

Exploratory Data Analysis



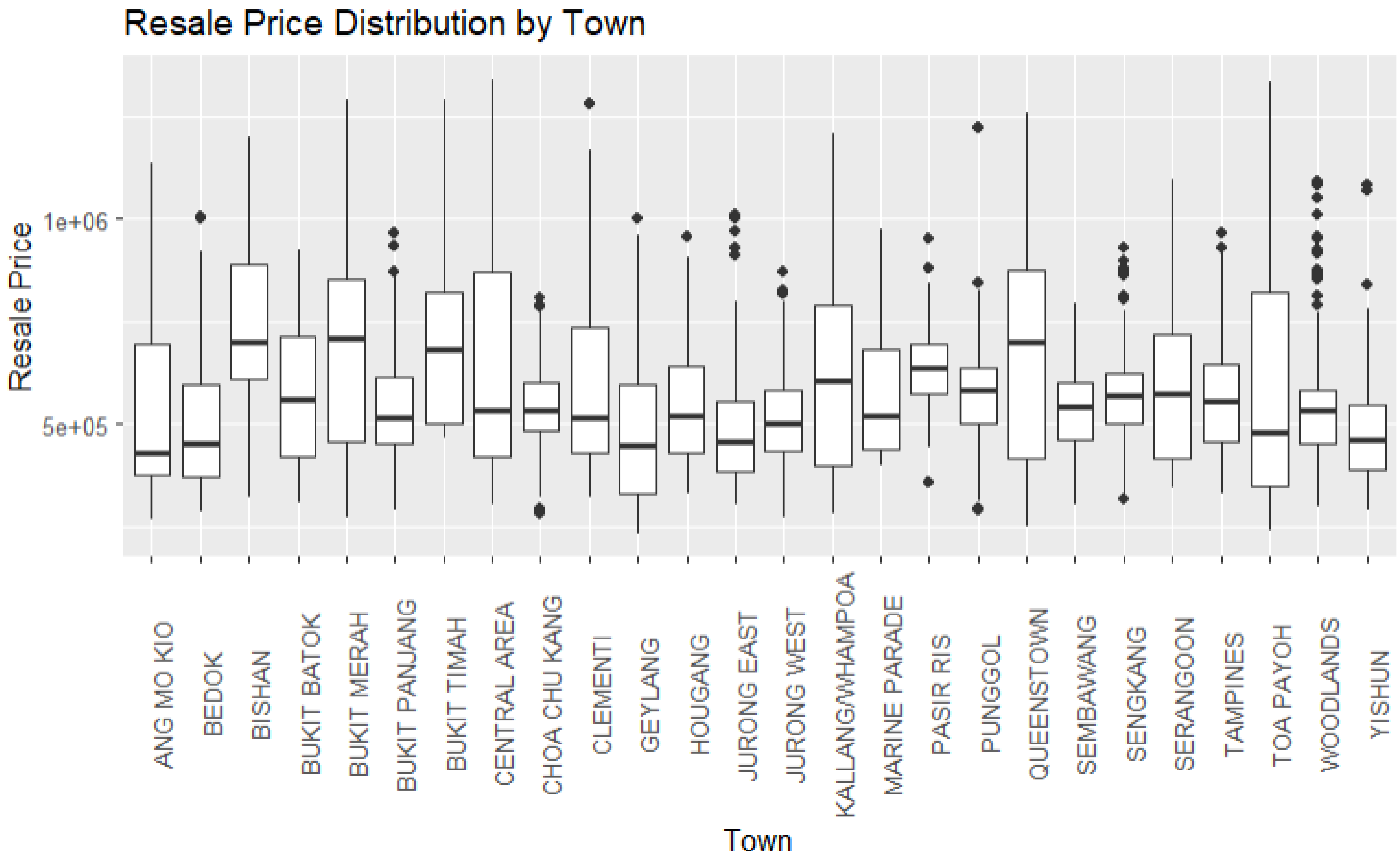
Higher storey -> higher price?

Exploratory Data Analysis



Better flat type -> higher price?

Exploratory Data Analysis



Various distributions -> feature engineering?

Feature Engineering

Feature Engineering

Nearest, Distance to Nearest, and Total Nearby

1. MRTs
2. Bus Stops
3. Schools
4. Primary Schools
5. Malls

Data from data.busrouter.sg and data.gov.sg
Latitude Longitude data (including HDBs) from OneMap SG API

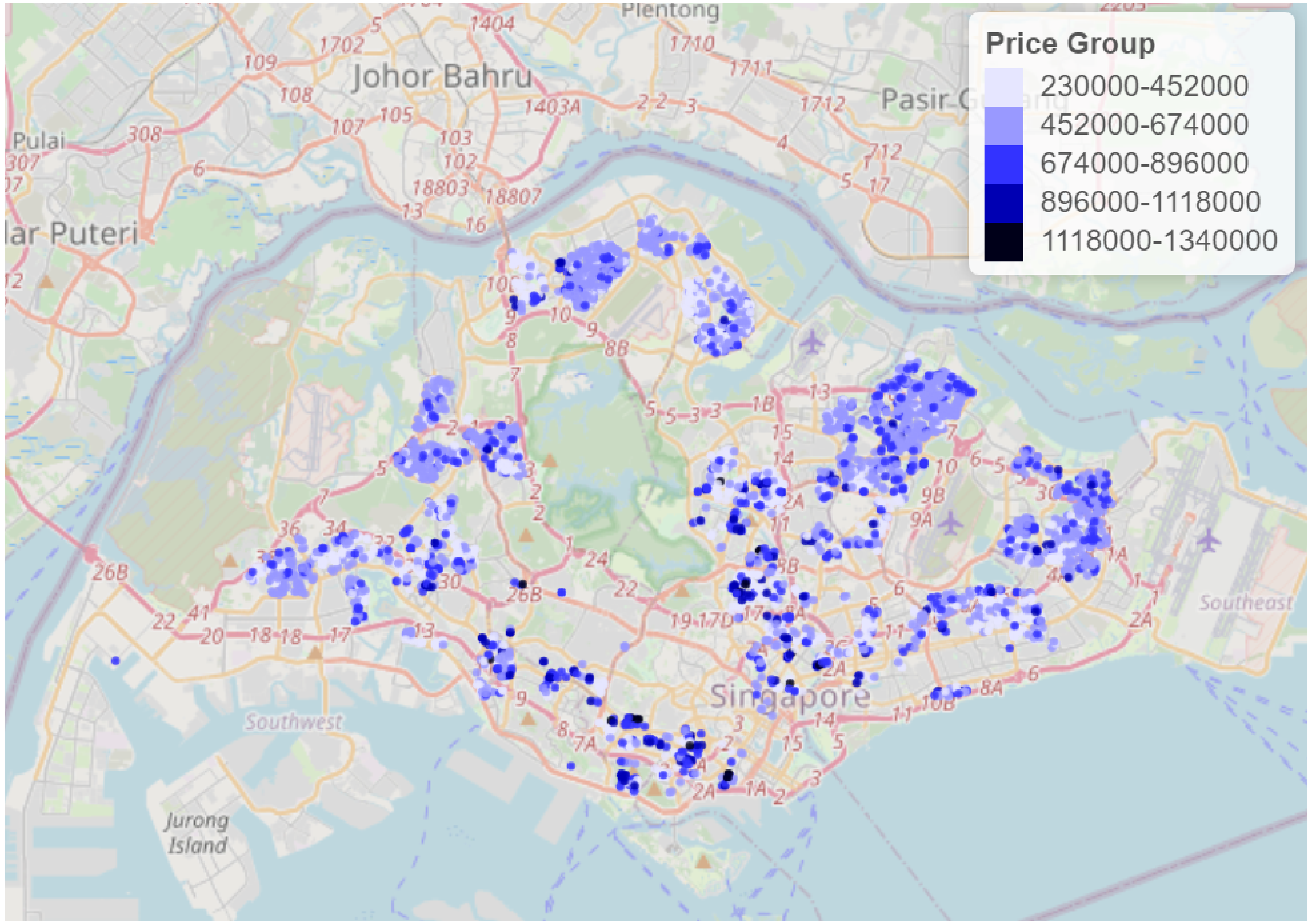
Feature Engineering

Split into 80% train 20% test

After splitting, we added 3 more variables:
1. total resales in town
2. total resales in block
3. total resales in street

Then, we did one-hot encoding for categorical variables
Lastly, we standardized the predictors to mean 0 and variance 1
Result: 4181 variables (including response variable)

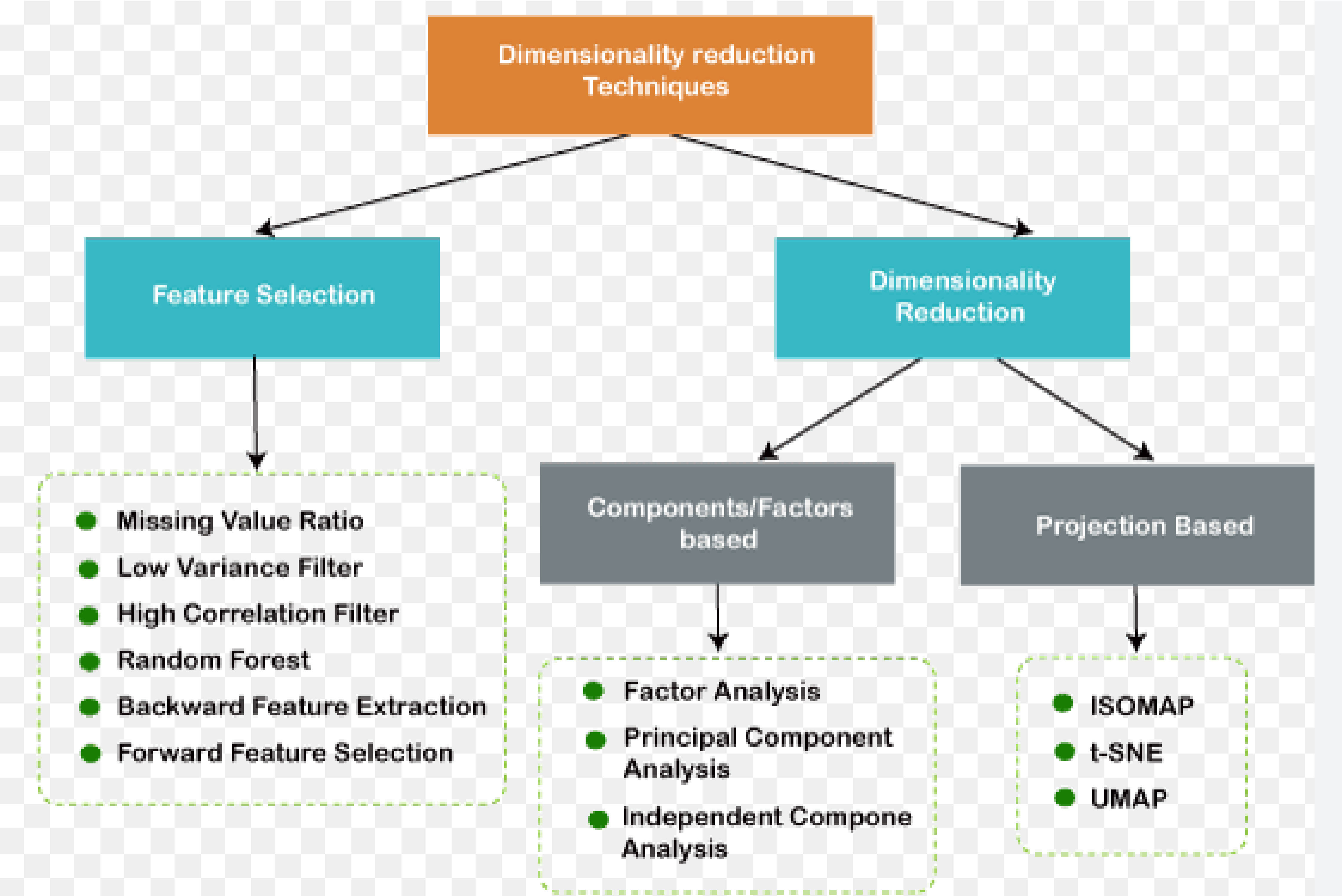
HDB Resale Prices by Location



More expensive unit in the central?

Dimensionality Reduction

Dimensionality Reduction



We don't want to do factor-based or projection-based dimensionality reduction as it makes our models less interpretable for inference

Feature Selection

- Filter Method
- Wrapper Method
- Embedded Method

Filter Method

Variance Threshold (Remove variance 0)

233 predictors removed

Ensemble of 6 Feature Selections

Wrapper Method

Forward Selection

Generate 100 selected variables

Recursive Feature Elimination (100 selected variables)

- Ridge Regression
- Gradient Boosting Regressor

Note: RFE is similar to Backward Selection

Embedded Method

Best Subset Selection (100 selected variables)

- F Regression
- Mutual Info Regression

- F Regression uses F statistics to see a linear relationship
- Mutual Info Regression captures the complex, non-linear relationship of each predictor vs response

Lasso (select 100 nonzero variables)

Majority Rule Voting-Based

Select variables that are selected by >= 3 methods
Total: 74 final predictors

Top 5 variables (selected by all 6 methods):
1. floor_area_sqm
2. total_resales_in_town
3. nearest_mrt_dist
4. remaining_lease
5. town_BUKIT MERAH

Models

Models

- Price
- Price/sqm

Note:
1. All models (except Linear Regression and Neural Network) are finetuned using GridSearchCV
2. Linear Regression uses non-scaled data while other models use scaled data
3. For Price/sqm models we are not using floor_area_sqm as predictor

Models

- Linear Regression
- ElasticNet (Combination of L1 and L2 Penalties)
- Neural Network (3 Hidden Layers w/ ReLu)
- Random Forest Regression
- Gradient Boosted Regression
- XGBoost

Evaluation

Evaluation

1. All metrics reported are using the best parameters after GridSearchCV (except Linear Regressionand Neural Network)
2. Metrics for Price/sqm model are calculated after converting back to price

Evaluation

Metrics	Model	LinReg	ElasticNet	NN	RF	GBR	XGBoost
RMSE	Price	54316	52786	43526	46939	50644	38256
	Price/sqm	49426	49330	44579	42254	35003	34987
MAPE	Price	7.81%	7.53%	5.3%	5.53%	6.19%	4.53%
	Price/sqm	6.62%	6.65%	5.71%	5.09%	4.34%	4.40%
Adj R2	Price	87.51%	89.77%	92.42%	91.91%	87.38%	94.14%
	Price/sqm	89.40%	91.07%	92.06%	93.45%	94.75%	95.00%

Evaluation

Metrics	Model	LinReg	ElasticNet	NN	RF	GBR	XGBoost
RMSE	Price	54316	52786	43526	46939	50644	38256
	Price/sqm	49426	49330	44579	42254	35003	34987
MAPE	Price	7.81%	7.53%	5.3%	5.53%	6.19%	4.53%
	Price/sqm	6.62%	6.65%	5.71%	5.09%	4.34%	4.40%
Adj R2	Price	87.51%	89.77%	92.42%	91.91%	87.38%	94.14%
	Price/sqm	89.40%	91.07%	92.06%	93.45%	94.75%	95.00%

Evaluation

Metrics	Model	LinReg	ElasticNet	NN	RF	GBR	XGBoost
RMSE	Price	54316	52786	43526	46939	50644	38256
	Price/sqm	49426	49330	44579	42254	35003	34987
MAPE	Price	7.81%	7.53%	5.3%	5.53%	6.19%	4.53%
	Price/sqm	6.62%	6.65%	5.71%	5.09%	4.34%	4.40%
Adj R2	Price	87.51%	89.77%	92.42%	91.91%	87.38%	94.14%
	Price/sqm	89.40%	91.07%	92.06%	93.45%	94.75%	95.00%

Learnings

Linear Regression Top 5 Features

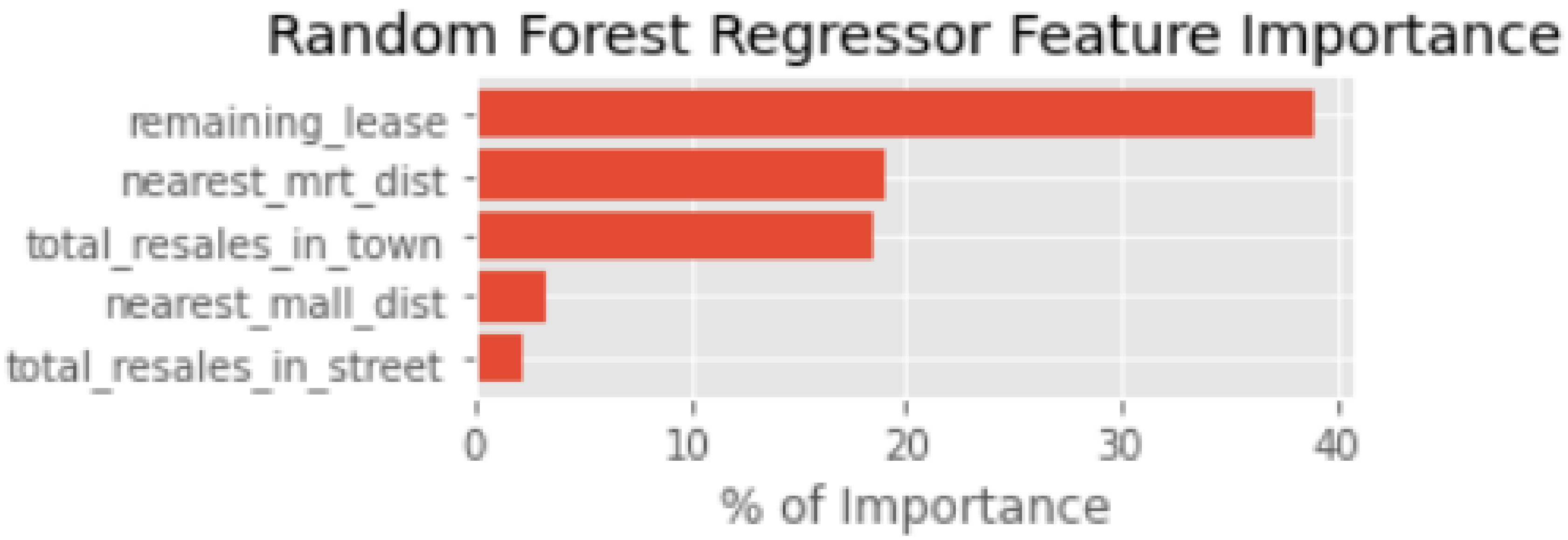
(Price/sqm)

Feature	Coefficient
Intercept	4246.0972
Total resales in town	-5.1221
Remaining lease	64.8747
Nearest mall distance	-159.0245
Total nearby MRTs	87.7080
Nearest MRT distance	-392.0953

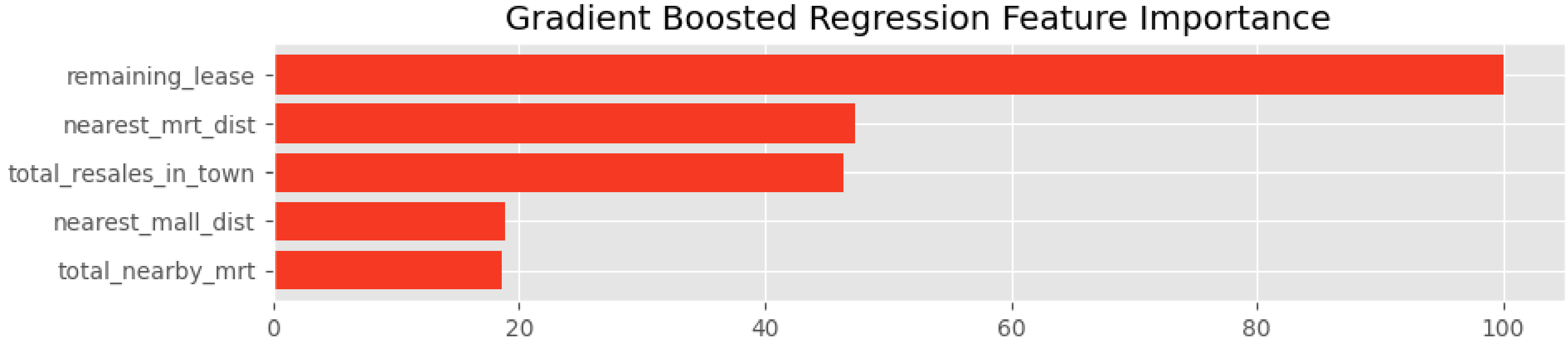
Feature Importance

Top 3 Models: RF, GBR, XGBoost (Price/sqm)

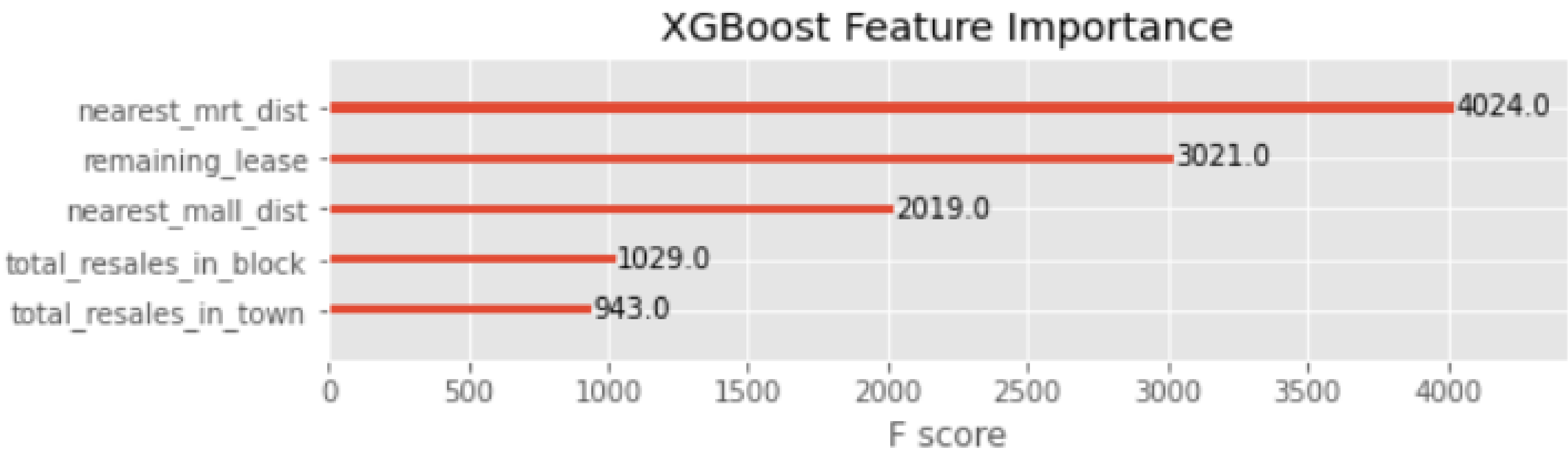
Random Forest Regressor



Gradient Boosting Regressor



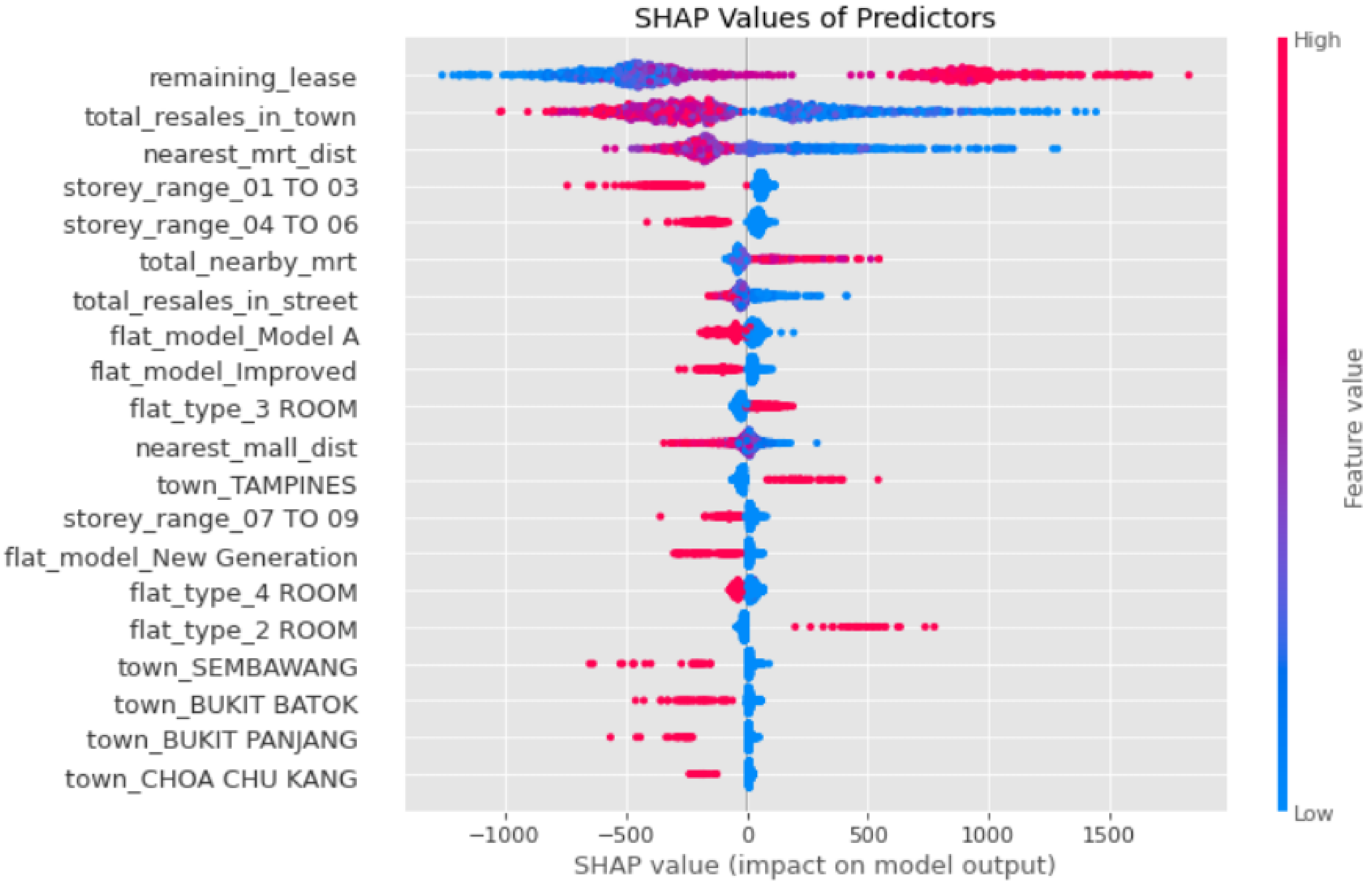
XGBoost



Top 5 Features from Feature Selection:

floor_area_sqm, total_resales_in_town
nearest_mrt_dist, remaining_lease
town_BUKIT MERAH

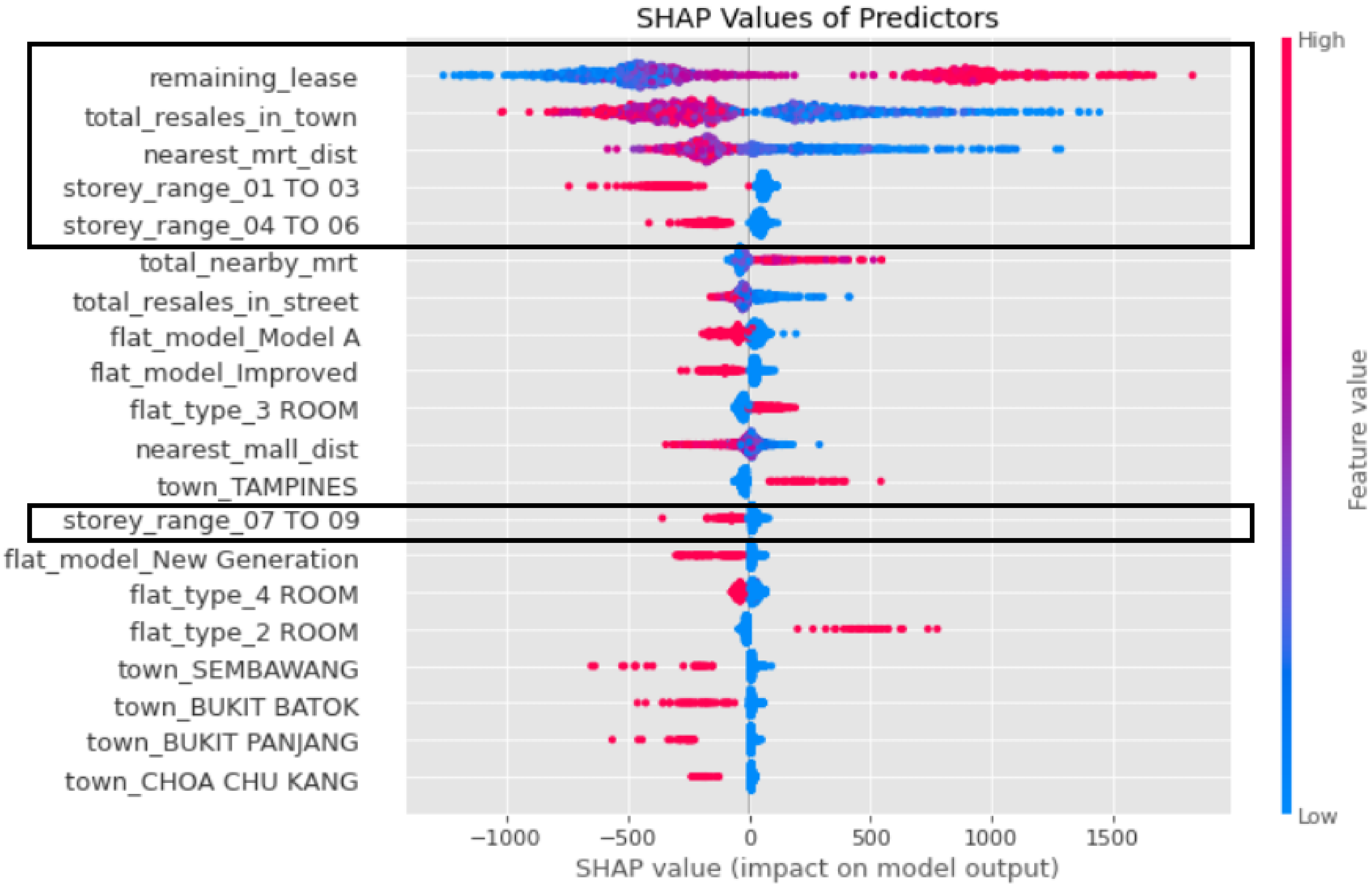
Shapley Values (XGBoost)



Red dots on RIGHT: value of predictor is DIRECTLY proportional to resale pricee price

Red dots on LEFT: value of predictor is INVERSELY proportional to resale price

Shapley Values (XGBoost)



Higher remaining lease -> higher price

Lower total resales in town -> higher price

Nearer MRT -> higher price

HDBs located at storey 1 to 3, 4 to 6, 7 to 9 tend to have lower price

Predictors Effects

Add to overall price
Remaining lease
Total nearby MRTs
Floor number > 20

Subtract from overall price
Nearest MRT distance
Nearest mall distance
Total resales in town

Q & A