Project 2
HDB Resale Price Prediction

DSIF9 Team 4

Ho Kit Fai
Zhu Ye (Juliana)
Ng Zeng Di
Wong Keng Hui

Agenda

1. Background Situation
2. Approach on Data Cleaning & EDA
3. Feature Engineering & Modelling
4. Model Evaluation
5. Conclusion and Recommendation

How many applicants were successful in their BTO flat application at their first attempt?

A file photo of a Built-To-Order flat being built.

December 1, 2022
https://www.todayonline.com/singapore/20-24-first-time-bto-applicants-mature-estates-succeed-first-try-less-2-need-more-5-tries-desmond-lee-2060306

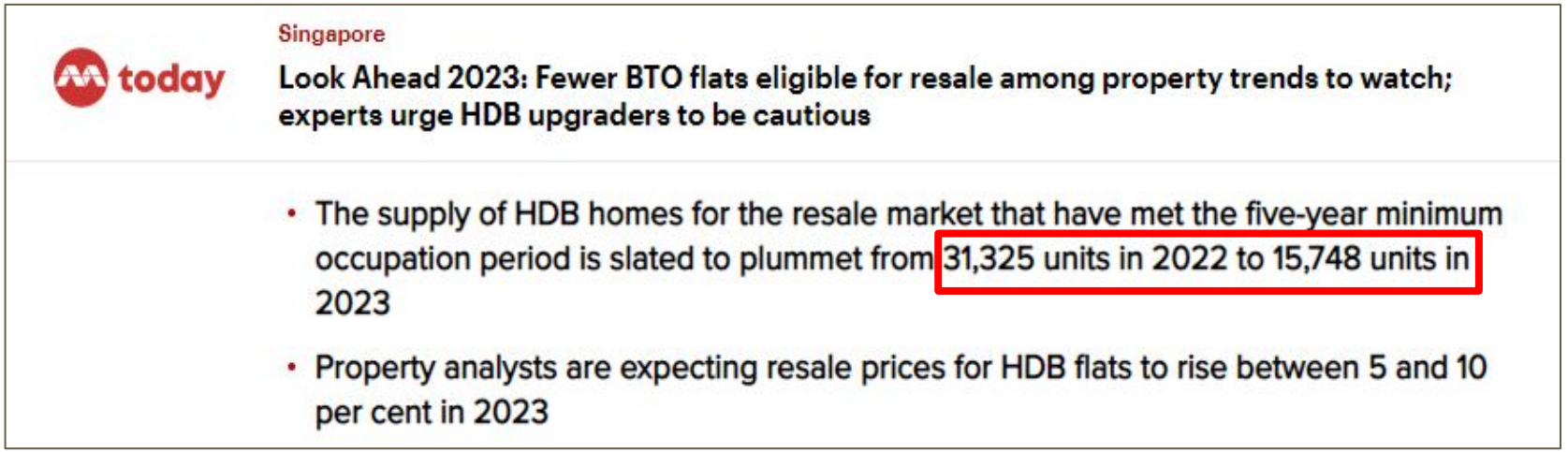
How many applicants were successful in their BTO flat application at their first attempt?

From 2019 to 2021, 20 to 24 per cent of first-time families were successful on their first attempt for a Build-to-Order (BTO) flat in mature estates, says Minister for National Development Desmond Lee.

This is a written response to a parliamentary question by Workers' Party Member of Parliament (MP) He Ting Ru on Nov. 30.

December 1, 2022
https://www.todayonline.com/singapore/20-24-first-time-bto-applicants-mature-estates-succeed-first-try-less-2-need-more-5-tries-desmond-lee-2060306

High resale price for HDB



today

Singapore

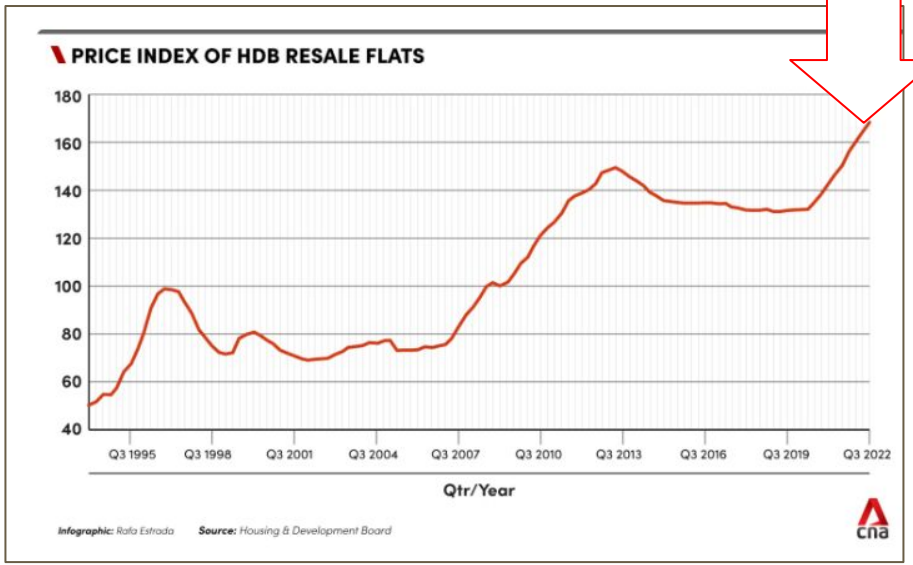
Look Ahead 2023: Fewer BTO flats eligible for resale among property trends to watch; experts urge HDB upgraders to be cautious

- The supply of HDB homes for the resale market that have met the five-year minimum occupation period is slated to plummet from 31,325 units in 2022 to 15,748 units in 2023
- Property analysts are expecting resale prices for HDB flats to rise between 5 and 10 per cent in 2023

January 4, 2023
https://www.todayonline.com/singapore/look-ahead-2023-bto-flats-resale-home-prices-2083416

High resale price for HDB

Q3 2022 more than 160 price index



The following table shows the median resale prices by town and flat type for resale cases registered in the 4th quarter of 2022:

TOWNS	1-ROOM	2-ROOM	3-ROOM	4-ROOM	5-ROOM	EXECUTIVE
ANG MO KIO	-	*	$378,000	$555,000	$868,000	*
BEDOK	-	*	$360,000	$480,000	$650,000	*
BISHAN	-	-	*	$669,400	$877,500	*
BUKIT BATOK	-	*	$360,000	$514,000	$700,000	*
BUKIT MERAH	*	*	$400,000	$775,400	$899,000	-
BUKIT PANJANG	-	*	$375,000	$480,000	$590,000	$780,000
BUKIT TIMAH	-	-	*	*	*	*
CENTRAL	-	*	$450,000	*	*	-
CHOA CHU KANG	-	*	$389,500	$500,000	$588,000	$679,000
CLEMENTI	-	*	$378,000	$574,900	*	*
GEYLANG	-	*	$335,000	$576,500	$781,500	*
HOUGANG	-	*	$380,000	$518,000	$627,500	$845,900
JURONG EAST	-	*	$360,000	$465,000	$643,000	*
JURONG WEST	-	*	$358,000	$481,500	$566,000	$685,000
KALLANG/WHAMPOA	-	-	$380,000	$788,000	$800,000	*
MARINE PARADE	-	-	$430,000	*	*	-
PASIR RIS	-	*	*	$520,000	$641,000	$796,000
PUNGGOL	-	*	$451,000	$575,900	$700,000	*
QUEENSTOWN	-	*	$390,000	$870,000	*	-
SEMBAWANG	-	$327,500	$432,500	$535,000	$579,000	*
SENGKANG	-	*	$443,400	$548,000	$580,000	$710,000
SERANGOON	-	*	$386,500	$539,000	*	*
TAMPINES	-	*	$410,000	$536,500	$652,500	$850,000
TOA PAYOH	-	*	$340,000	$780,000	$865,000	*
WOODLANDS	-	*	$395,000	$495,000	$580,000	$762,500
YISHUN	-	*	$372,000	$474,000	$620,000	$788,000

https://www.hdb.gov.sg/residential/selling-a-flat/overview/resale-statistics

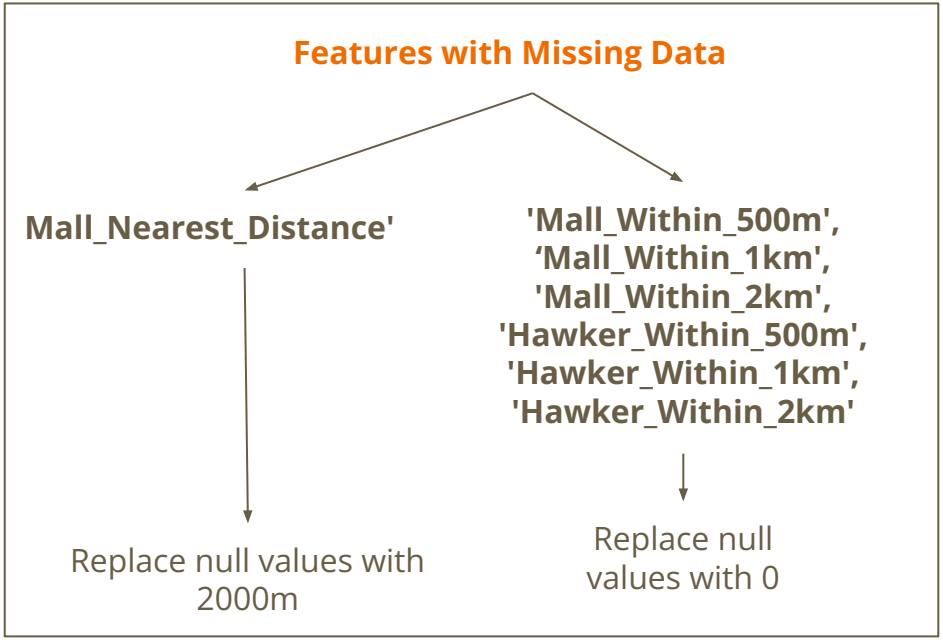
Situation

- Not easy to apply for BTO flat due to low successful application rate
- Increasing resale price for HDB
- Lesser HDB resale supply added in 2023

Resolution

- Find out what features have an influence on the market resale value of HDB flats
- Aim to predict the resale price of an HDB flat for Buyer and Seller make better decision on getting the right price for transaction.

Approach for Data Cleaning



Feature Selection

Dropped features* based on:

1. Presence of an existing substitutable feature or high correlation with another feature
2. Sparsity or lack of variance (all 1s or Ys)
3. Being used in creating new features
4. Limited correlation with resale price

*Dropped features: Tranc_YearMonth, block, street_name, storey_range, floor_area_sqm, lease_commence_date, lower, upper, mid, full_flat_type, address, residential, postal, Latitude, Longitude, planning_area, mrt_name, mrt_latitude, mrt_longitude, bus_stop_name, bus_stop_latitude, bus_stop_longitude, pri_sch_name, pri_sch_latitude, pri_sch_longitude, sec_sch_name, sec_sch_latitude, sec_sch_longitude

Approach for Exploratory Data Analysis (EDA)

1

Descriptive statistics

2

Change data types to better suit linear regression

3

Examine dependent variable

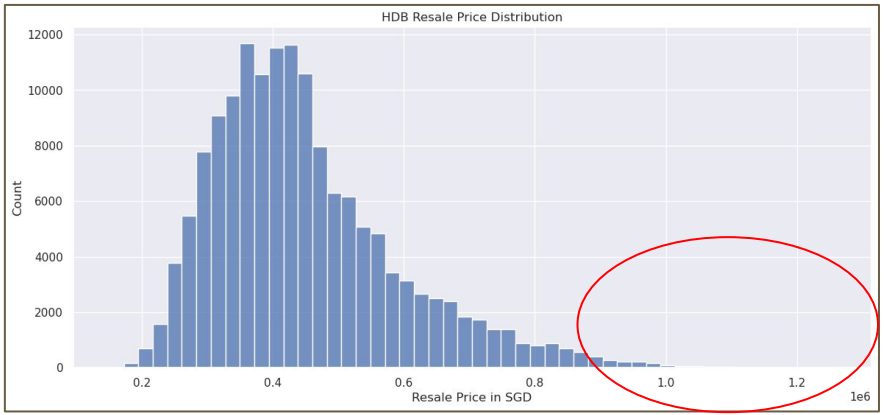
4

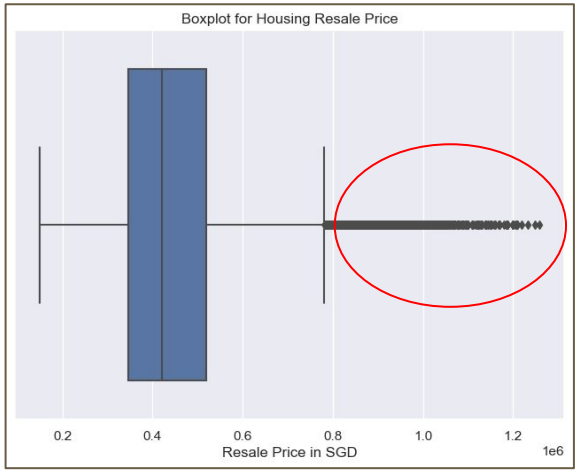
Relationships between variables and with resale price

5

Test hypothesis on whether different unit types are predicted by different factors

EDA: Examining ‘Resale Price’

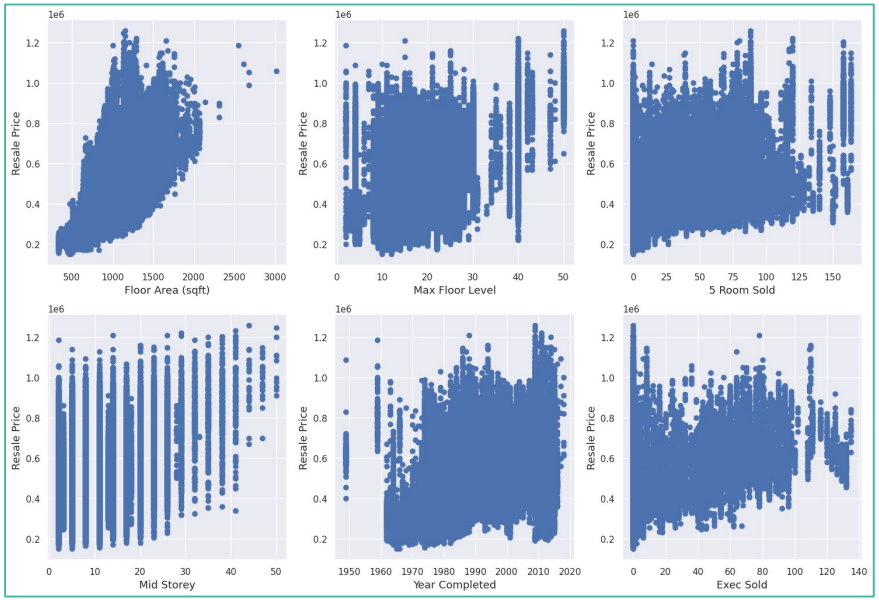


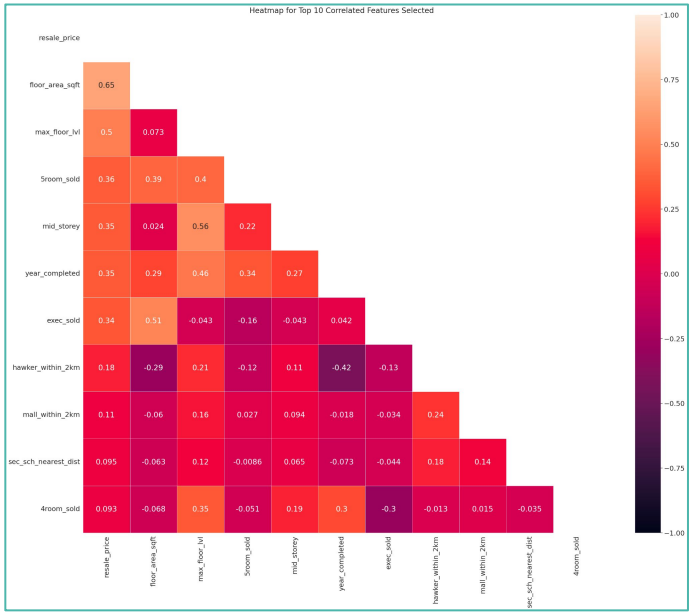


Observations:

1. Right-skewed indicating a small number of homes were sold at exceptionally high prices
2. The box-plot shows outliers priced above SGD 800,000
3. Choose not to exclude outliers since they reflect actual market conditions

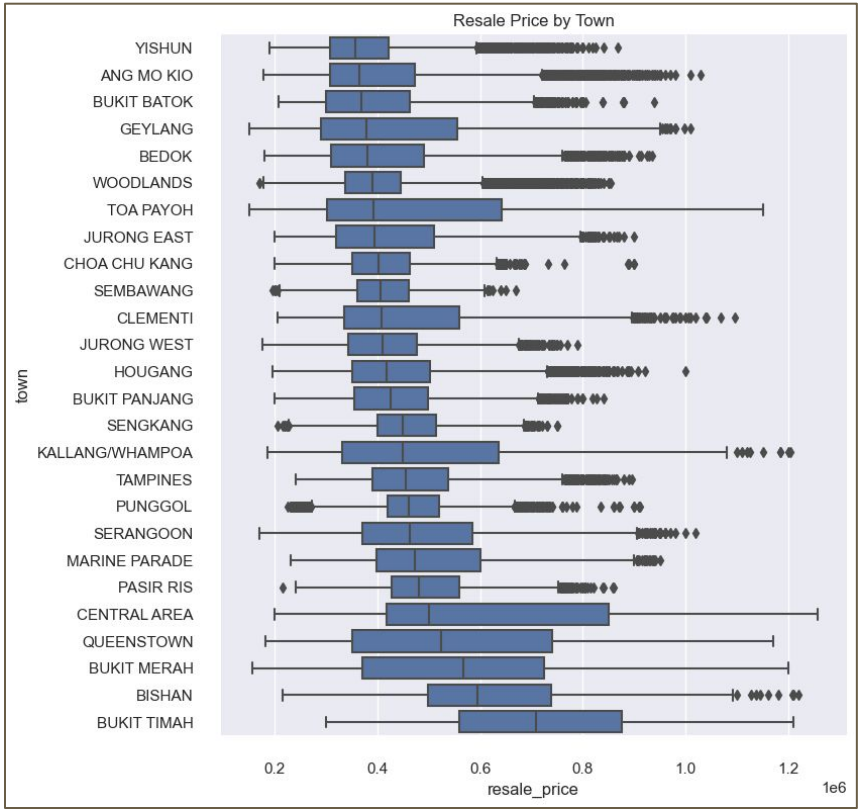
EDA: Examining Relationships among Numerical Features

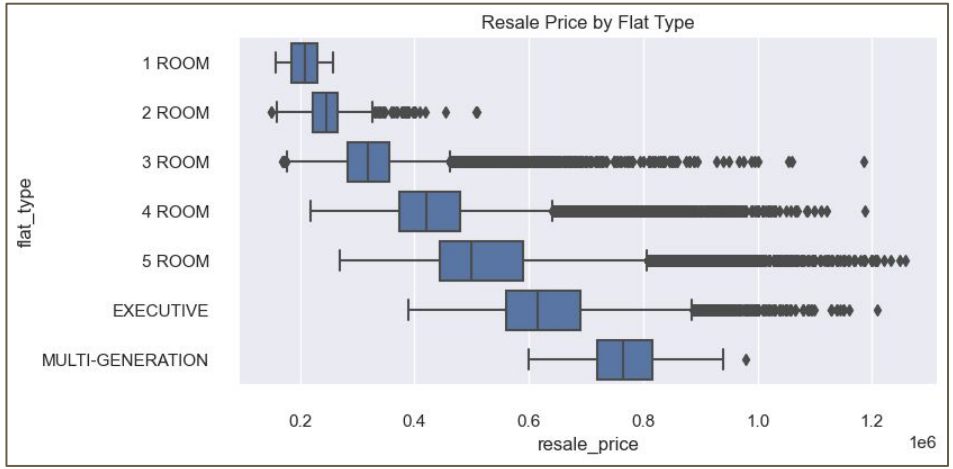




- Scatterplots and correlation heatmaps were used to study each feature's relationship with resale price and other predictor variables (i.e. multicollinearity).
- Not all numerical features had linear relationships with resale price. We note which ones did not, and think about transforming them so that they will make sense if included in a linear regression model.
- High multicollinearity creates interpretation problems and increases risk of overfitting for linear regression models. We consider combining variables and dropping the ones that have substitutes.

EDA: Examining Categorical Variables

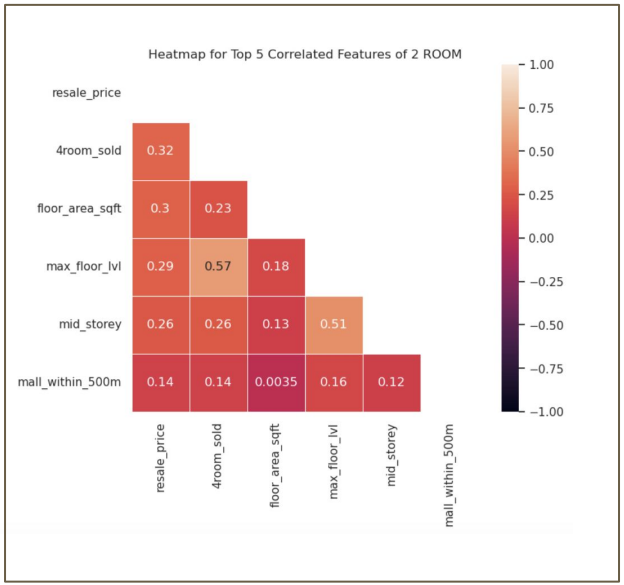


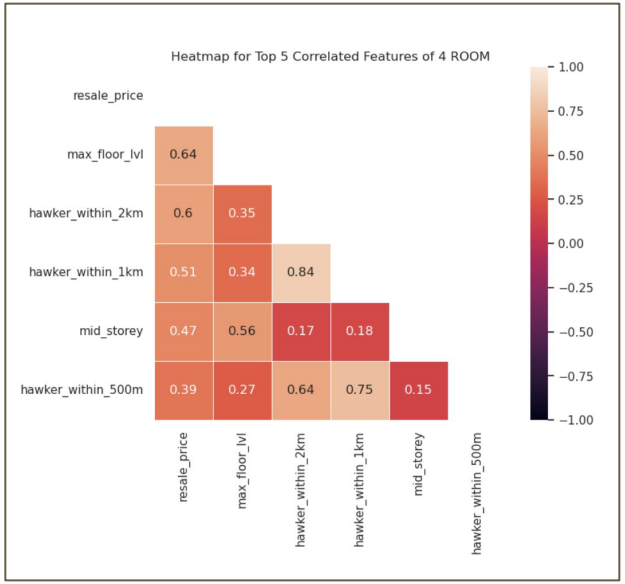


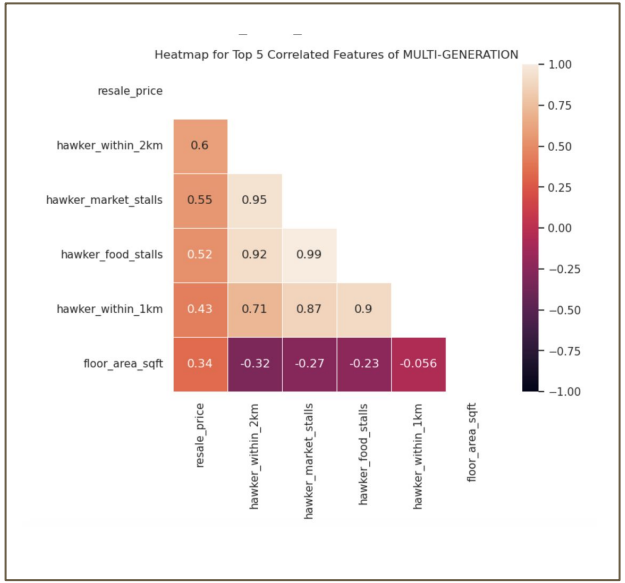
Boxplots and bar charts address some of our initial hypotheses for categorical variables such as town and flat type:

1) Whether they vary with resale price
2) The median resale prices for each category
3) Distribution and presence of outliers

EDA: Correlation with Resale Price by Flat Type



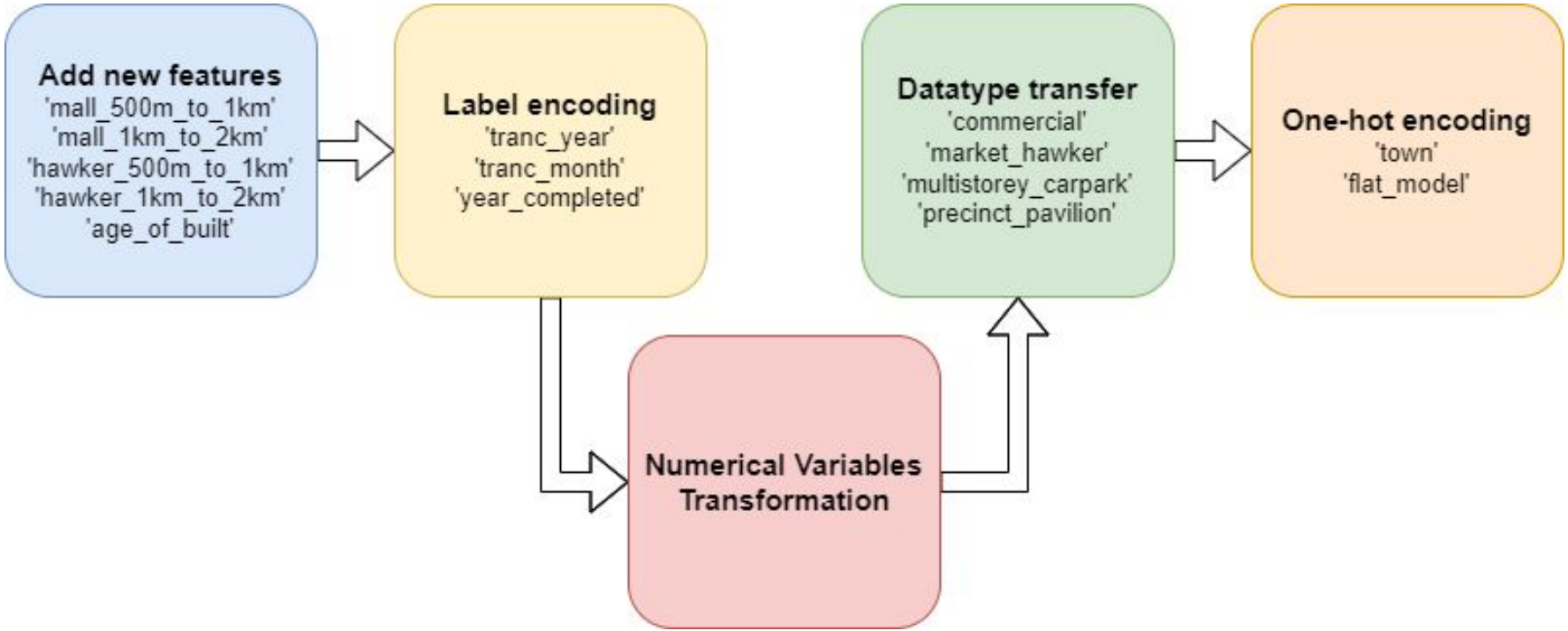




The top 5 correlated features by room type (showing a sample of rooms) are quite different.

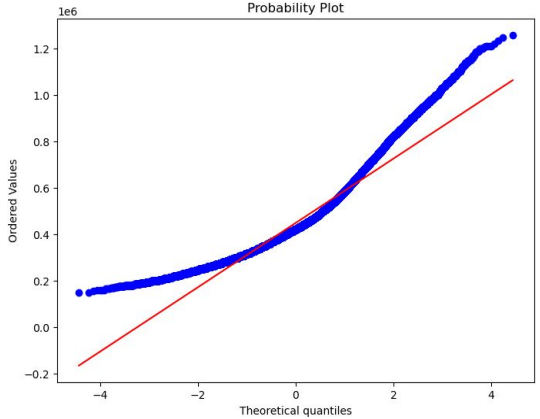
Should we do different models for each room type to predict resale prices?

Feature Engineering Steps

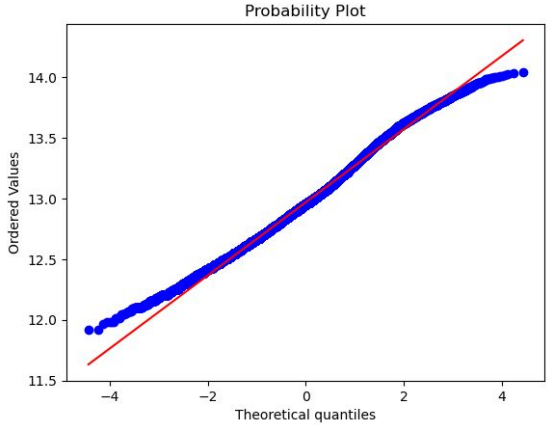


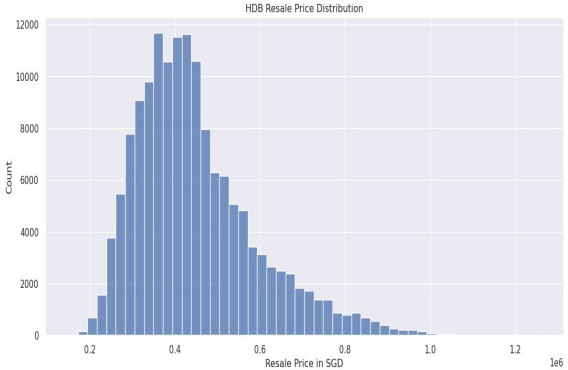
Feature Engineering: Numerical Variables Transformation

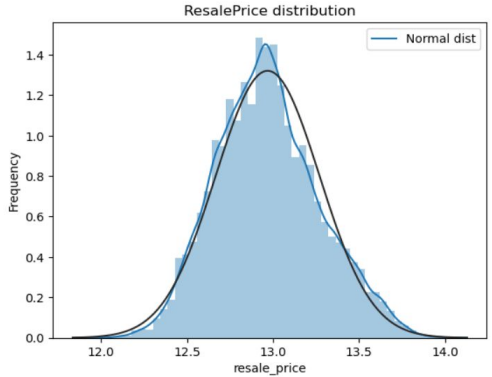
Original y_train



Transformed y_train



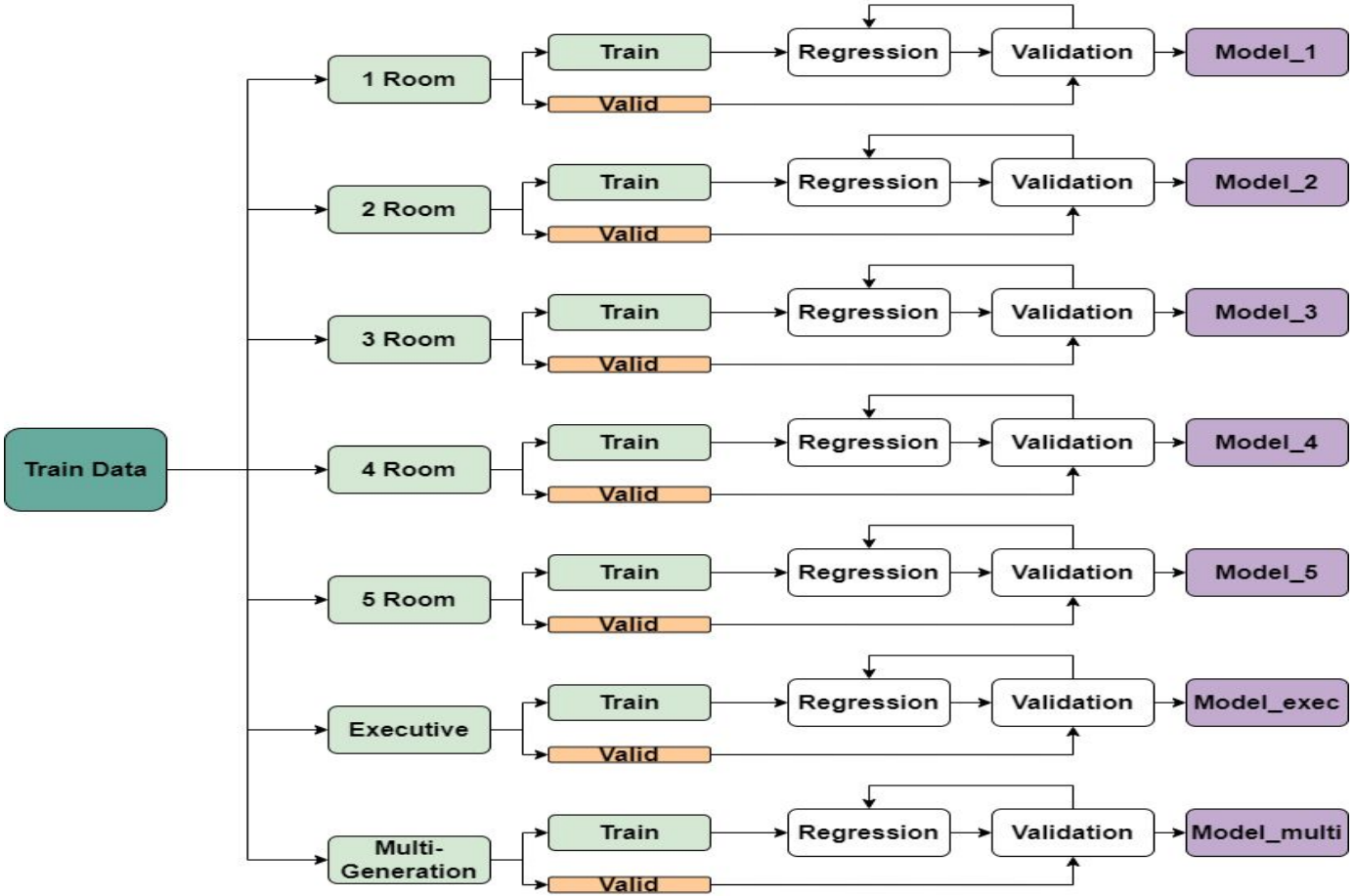




Skew in numerical features

X	Skew
other_room_rental	79.521010
multigen_sold	50.660446
1room_sold	43.330977
1room_rental	41.446940
2room_rental	27.034726
3room_rental	22.261928
studio_apartment_sold	15.331996
2room_sold	8.680007
mall_within_2km	4.628154
mall_1km_to_2km	3.872227

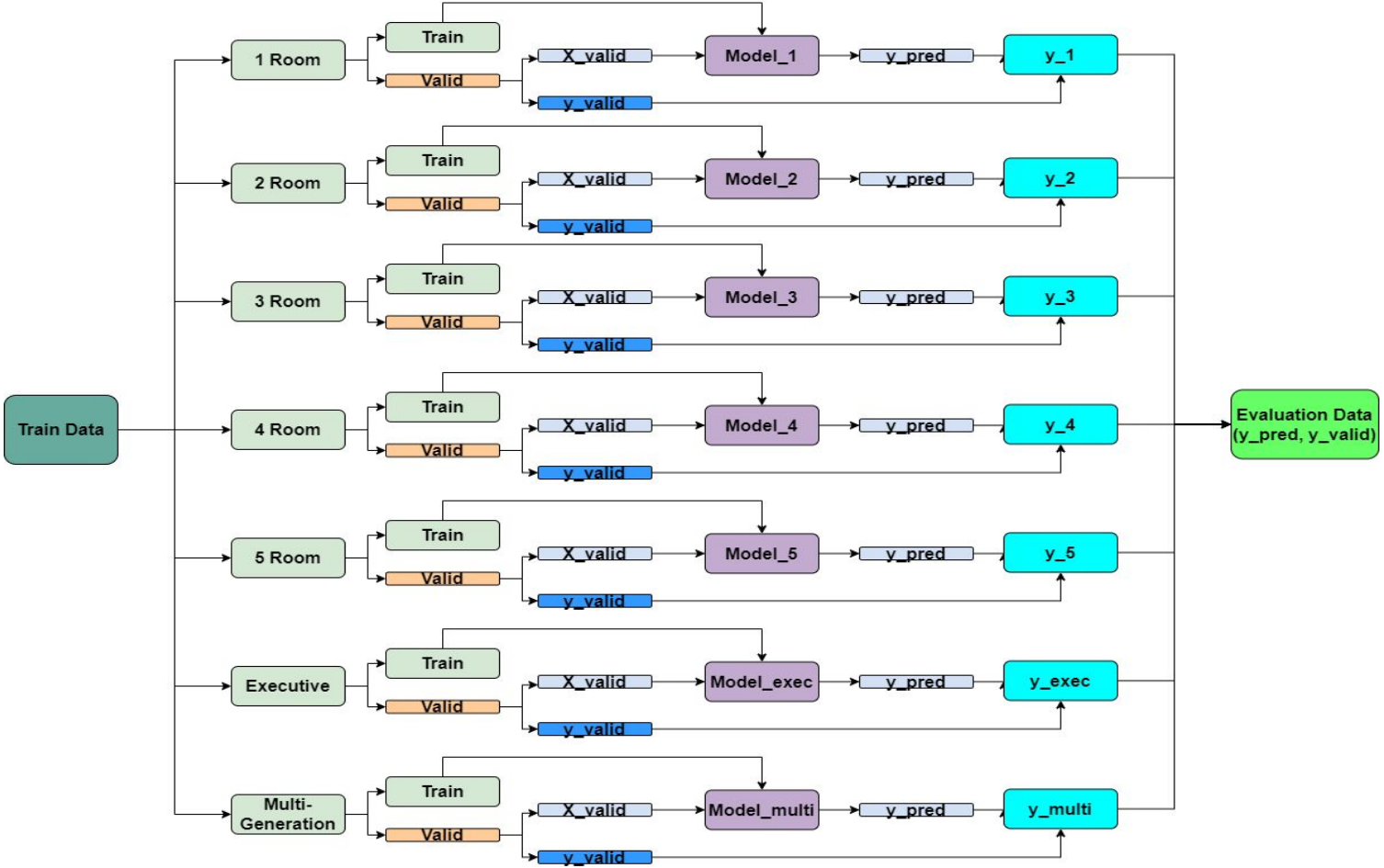
Modelling Methodology: Split Models for Different Flat Types



Summary of Models for Different Flat Types

Model Name	Flat Type	Regression Model	Normalized RMSE on Validation Set
Model_1	1 ROOM	LinearRegression	0.06683
Model_2	2 ROOM	LinearRegression	0.07313
Model_3	3 ROOM	LASSORegression	0.08614
Model_4	4 ROOM	LinearRegression	0.07959
Model_5	5 ROOM	RidgeRegression	0.08074
Model_6 (Model_exec)	EXECUTIVE	LinearRegression	0.07393
Model_7 (Model_multi)	MULTI-GENERATION	LinearRegression	0.05287

Model Evaluation with Validation Set



Model Evaluation Results

Baseline Model - Computation of Mean Resale Price

- **RMSE of Resale price: 143307.1$**
- Mean Resale Price: 449161.5$

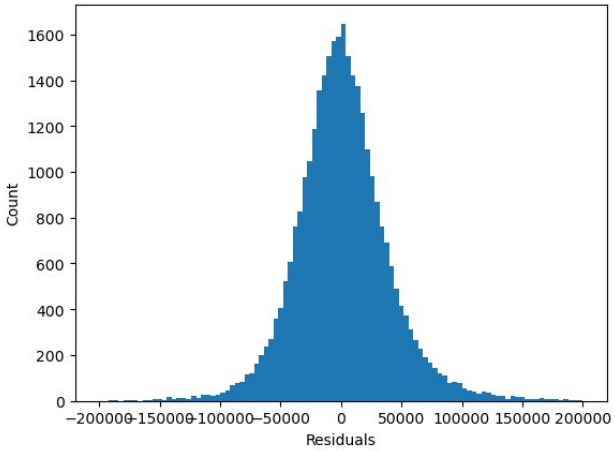
Machine Learning Model - Amalgamation of 7 models

- **RMSE of Resale price: 39415.3$**
- R2 of Resale Price: 0.928

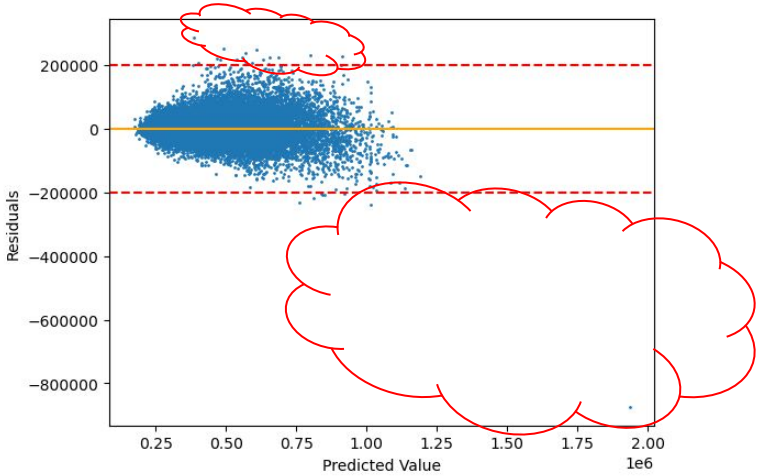
It’s a marked improvement from the Baseline in terms of predictive capability

Model Evaluation Results

Normality Check (Histogram of Residuals)



Equal Variance Check (Randomness in Residuals)

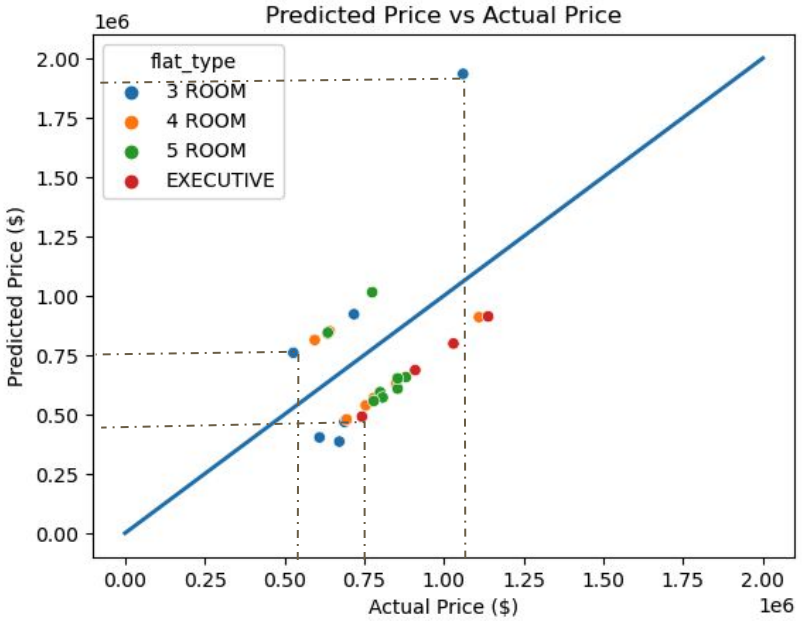


Histogram of Residuals - looks normally distributed, we're good

Randomness in Residuals - It is randomly distributed and not fanning out, however, there are a few outliers beyond +/-200000, let's zoom in

Model Evaluation Results

```
4 ROOM        8
5 ROOM        8
3 ROOM        6
EXECUTIVE     4
Name: flat_type, dtype: int64
```



We have 26 of such cases. This implies that our the models above requires further tuning

Conclusion and Recommendation

Positively correlated features to the resale price for each flat type

	1 Room	2 Room	3 Room	4 Room	5 Room	Exec	Multi
1st	tranc month	4room sold	max floor lvl	max floor lvl	hawker within 2km	hawker within 2km	hawker within_2km
2nd	mid storey	floor area sqft	floor area sqft	hawker within 2km	max floor lvl	hawker within 1km	hawker market stalls
3rd	tranc year	max floor lvl	mid storey	hawker within 1km	hawker within 1km	floor area sqft	hawker food stalls
4th	-	mid storey	hawker within 2km	mid storey	hawker within 500m	hdb age	hawker within 1km
5th	-	mall within 500m	mall within 2km	hawker within 500m	mid storey	affiliation	floor area sqft

Conclusion and Recommendation

Positively correlated features to the resale price for each flat type

	1 Room	2 Room	3 Room	4 Room	5 Room	Exec	Multi
1st	tranc month	4room sold	max floor lvl	max floor lvl	hawker within 2km	hawker within 2km	hawker within_2km
2nd	mid storey	floor area sqft	floor area sqft	hawker within 2km	max floor lvl	hawker within 1km	hawker market stalls
3rd	tranc year	max floor lvl	mid storey	hawker within 1km	hawker within 1km	floor area sqft	hawker food stalls
4th	-	mid storey	hawker within 2km	mid storey	hawker within 500m	hdb age	hawker within 1km
5th	-	mall within 500m	mall within 2km	hawker within 500m	mid storey	affiliation	floor area sqft

Conclusion and Recommendation

Positively correlated features to the resale price for each flat type

	1 Room	2 Room	3 Room	4 Room	5 Room	Exec	Multi
1st	tranc month	4room sold	max floor lvl	max floor lvl	hawker within 2km	hawker within 2km	hawker within_2km
2nd	mid storey	floor area sqft	floor area sqft	hawker within 2km	max floor lvl	hawker within 1km	hawker market stalls
3rd	tranc year	max floor lvl	mid storey	hawker within 1km	hawker within 1km	floor area sqft	hawker food stalls
4th	-	mid storey	hawker within 2km	mid storey	hawker within 500m	hdb age	hawker within 1km
5th	-	mall within 500m	mall within 2km	hawker within 500m	mid storey	affiliation	floor area sqft

Conclusion and Recommendation

Positively correlated features to the resale price for each flat type

	1 Room	2 Room	3 Room	4 Room	5 Room	Exec	Multi
1st	tranc month	4room sold	max floor lvl	max floor lvl	hawker within 2km	hawker within 2km	hawker within_2km
2nd	mid storey	floor area sqft	floor area sqft	hawker within 2km	max floor lvl	hawker within 1km	hawker market stalls
3rd	tranc year	max floor lvl	mid storey	hawker within 1km	hawker within 1km	floor area sqft	hawker food stalls
4th	-	mid storey	hawker within 2km	mid storey	hawker within 500m	hdb age	hawker within 1km
5th	-	mall within 500m	mall within 2km	hawker within 500m	mid storey	affiliation	floor area sqft

Conclusion and Recommendation

Positively correlated features to the resale price for each flat type

	1 Room	2 Room	3 Room	4 Room	5 Room	Exec	Multi
1st	tranc month	4room sold	max floor lvl	max floor lvl	hawker within 2km	hawker within 2km	hawker within_2km
2nd	mid storey	floor area sqft	floor area sqft	hawker within 2km	max floor lvl	hawker within 1km	hawker market stalls
3rd	tranc year	max floor lvl	mid storey	hawker within 1km	hawker within 1km	floor area sqft	hawker food stalls
4th	-	mid storey	hawker within 2km	mid storey	hawker within 500m	hdb age	hawker within 1km
5th	-	mall within 500m	mall within 2km	hawker within 500m	mid storey	affiliation	floor area sqft