Project 2: Creating a HDB Regression Model

- Samuel
- Ron
- Leila
- Benjamin

Agenda

1. Problem Statement
2. Exploratory Data Analysis
3. Feature Selection / Engineering
4. Model Evaluation
5. Interpretation of Results
6. Recommendations
7. Conclusion

Problem Statement

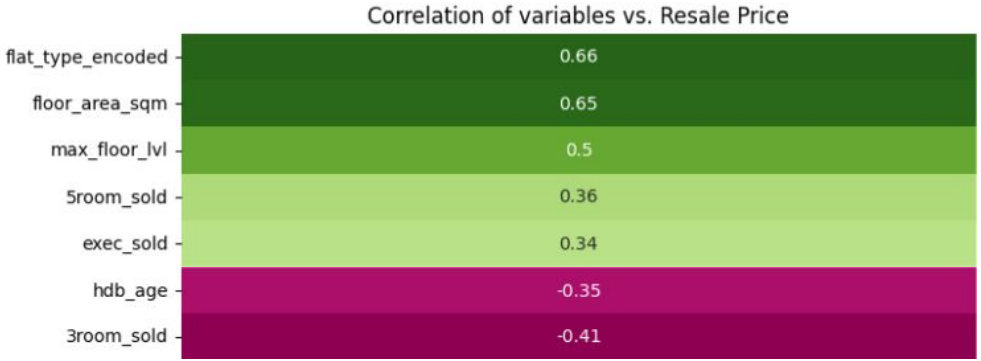
As real estate analysts, we are tasked with creating a regression model based on a Singapore HDB dataset. Through this analysis, our goal is to:

- Educate buyers/sellers if a house is being priced fairly
- Understand the main factors affecting HDB prices
- Predict HDB prices given a set of variables

Exploratory Data Analysis

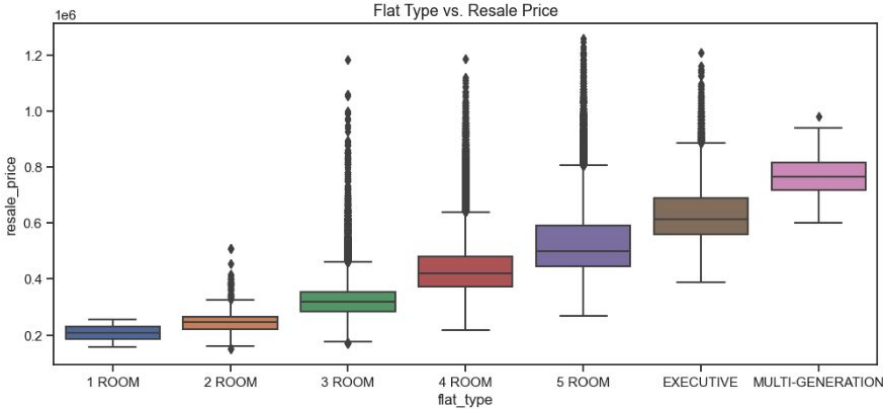
- We identified the positive and negative variables with the highest correlation (>0.3) to Resale Price

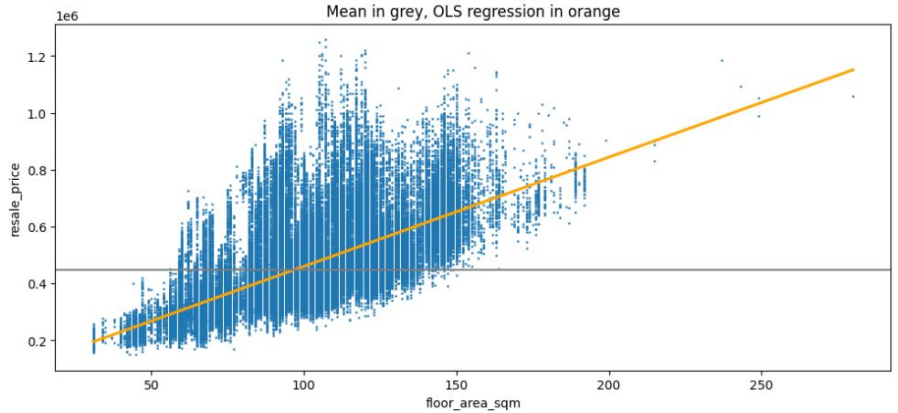
- Due to the huge presence of outliers, “exec_sold” will be removed



Exploratory Data Analysis - Positively Correlated

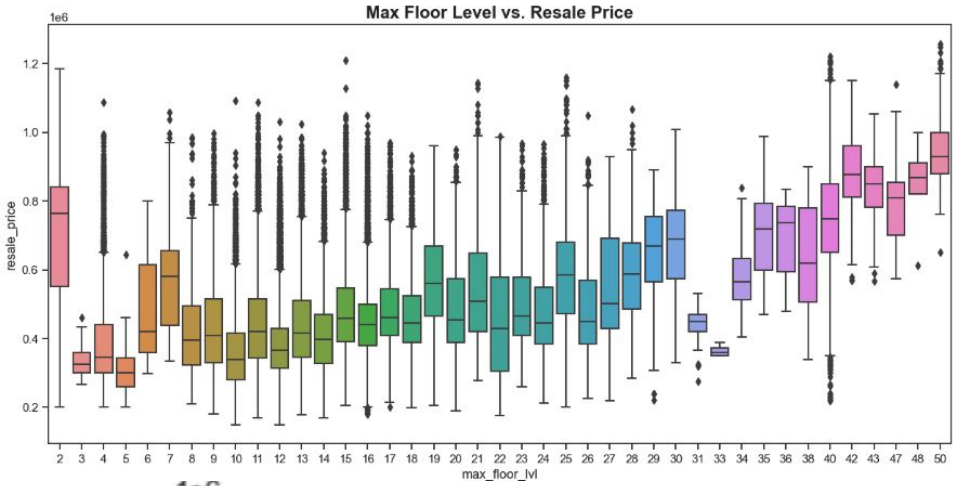
- Flat Type
  - Encoded as an ordinal variable with a score from 0 to 6
- Floor Area (sqm)

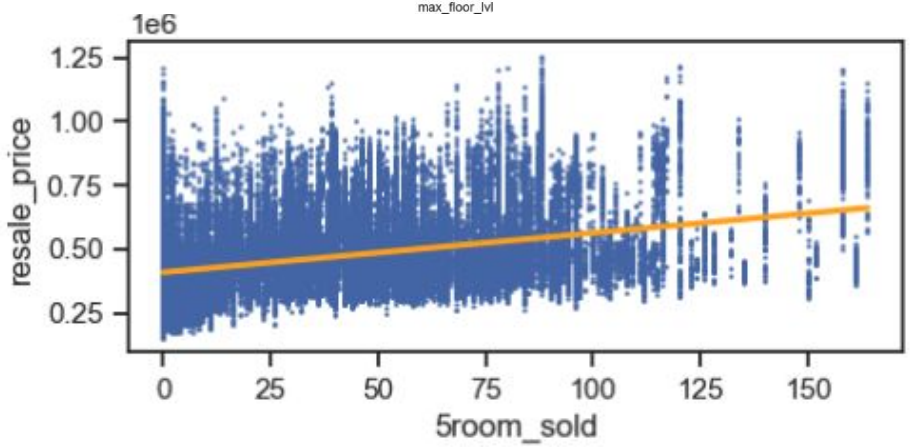




Exploratory Data Analysis - Positively Correlated

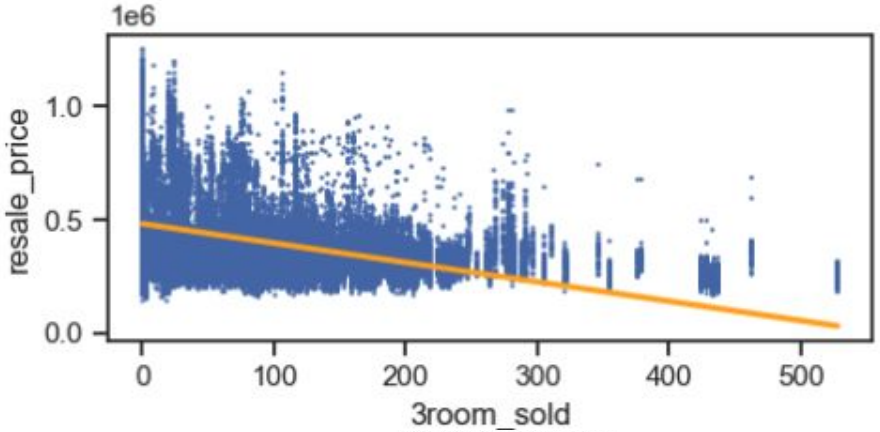
- Max Floor Level
- No. of 5-room units

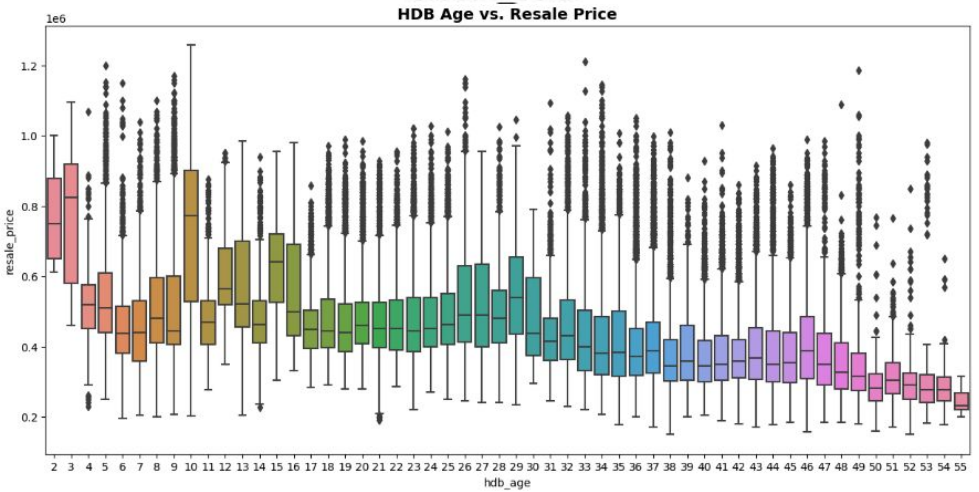




Exploratory Data Analysis - Negatively Correlated

- No. of 3-room units
- HDB Age

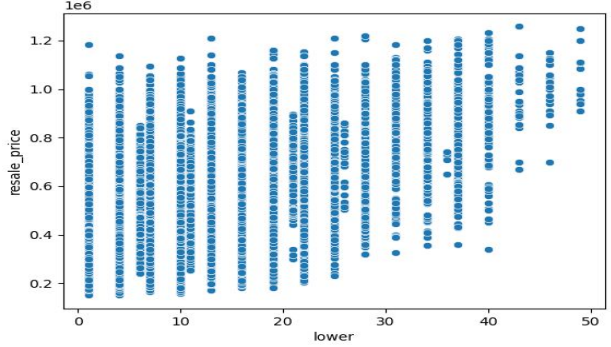


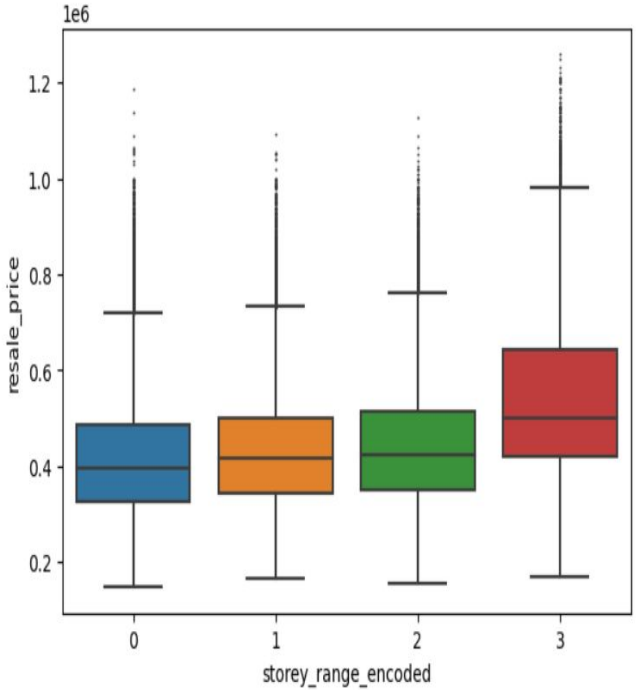


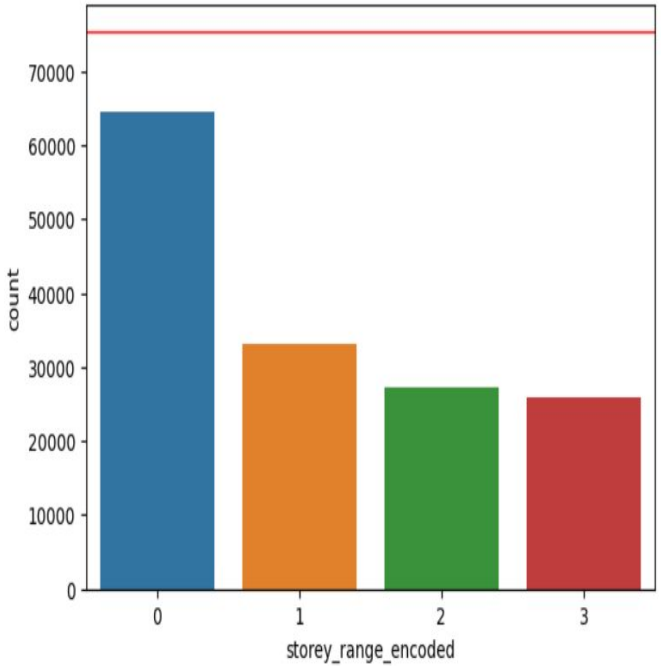
Storey Range

Quartile Groups:

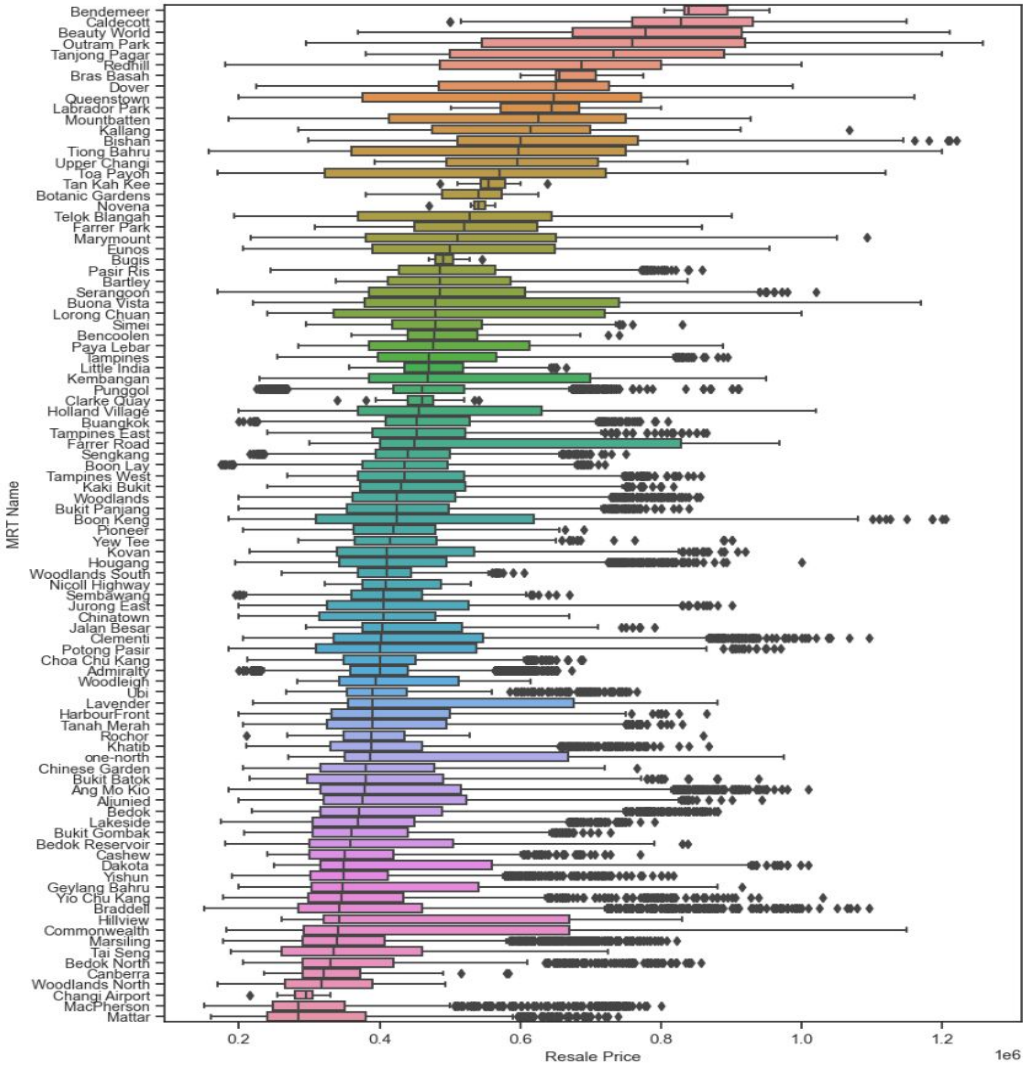
storey 1-4, 5-7, 8-10, 11-49



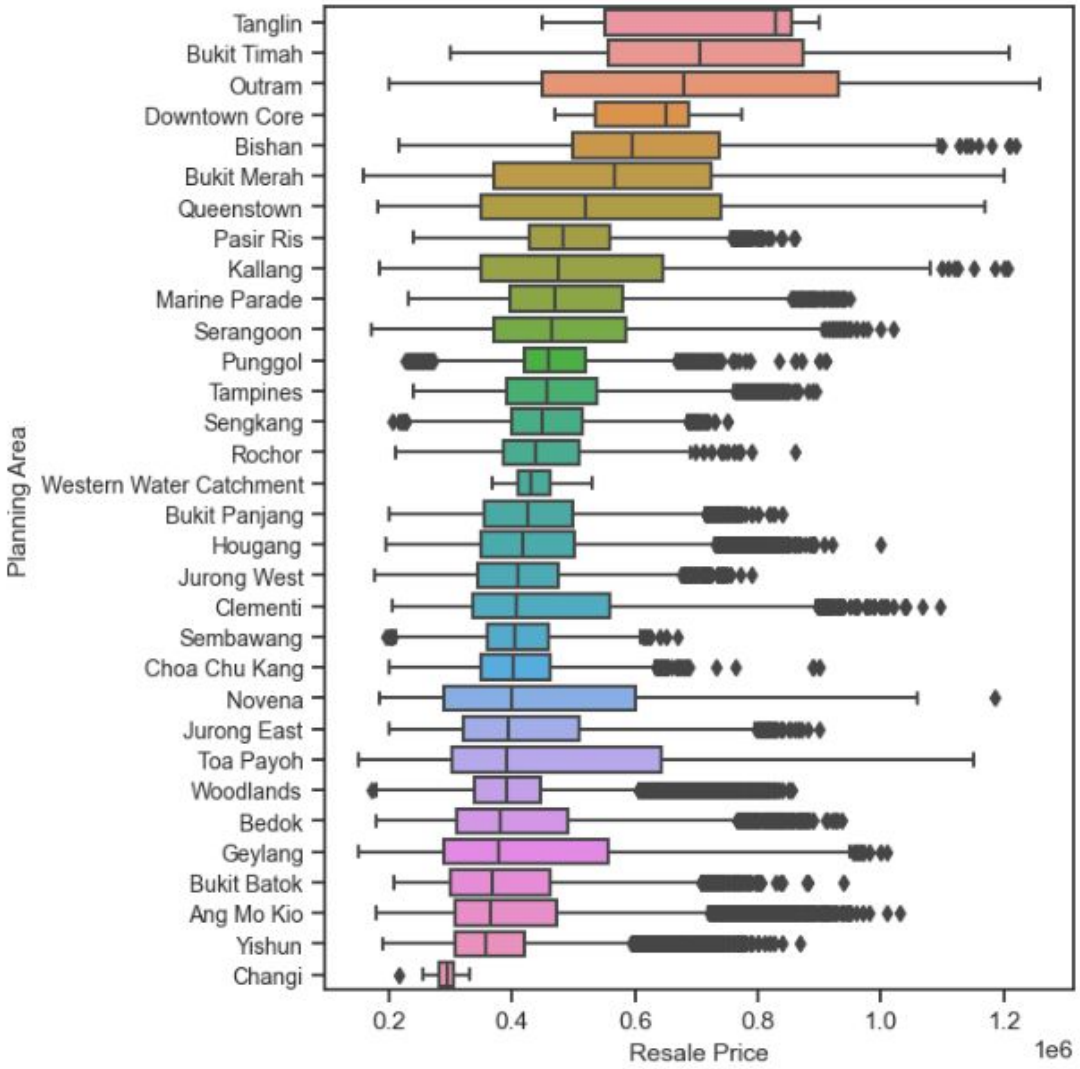




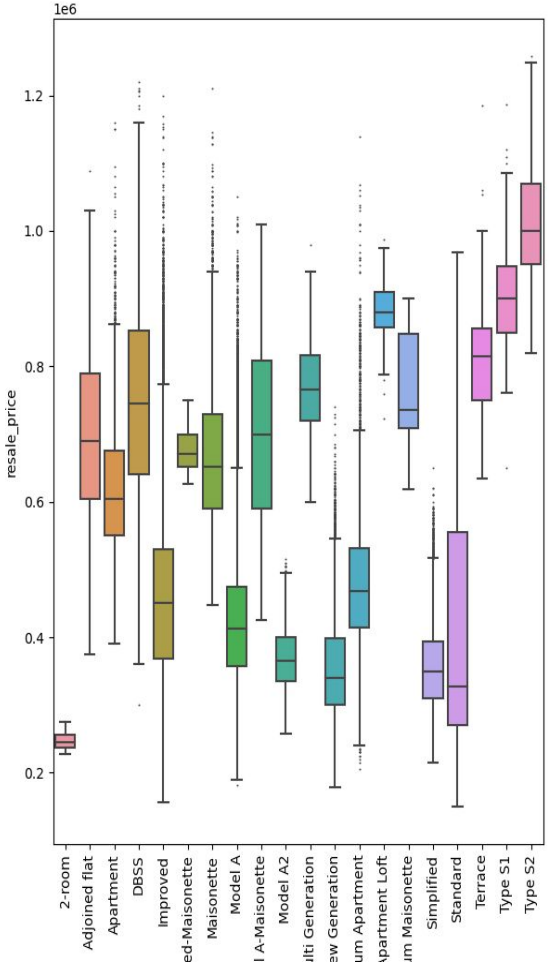
MRT NAME

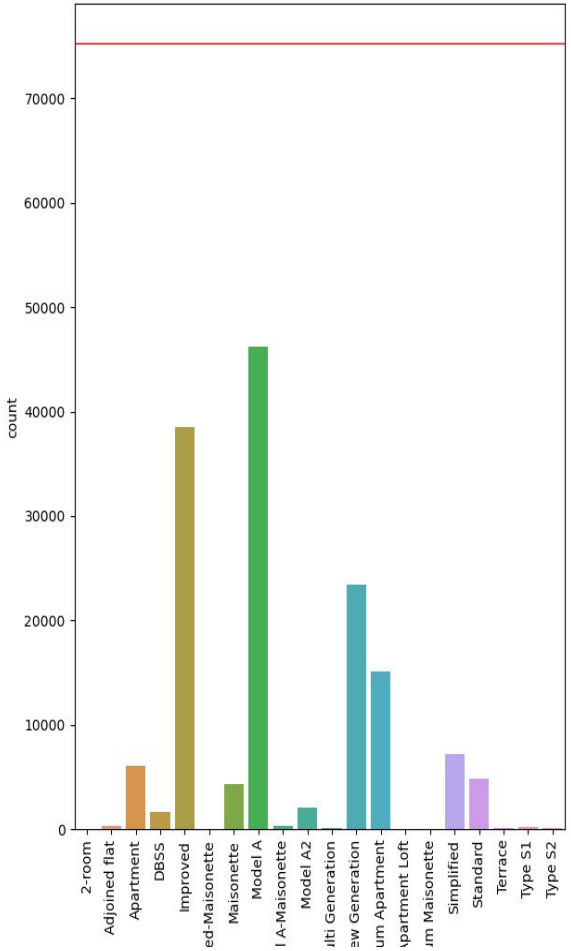

MRT Name
Bendemeer
Caldecott
Beauty World
Outram Park
Tanjong Pagar
Redhill
Bras Basah
Dover
Queenstown
Labrador Park
Mountbatten
Kallang
Bishan
Tiong Bahru
Upper Changi
Toa Payoh
Tan Kah Kee
Botanic Gardens
Novena
Telok Blangah
Farrer Park
Marymount
Eunos
Bugis
Pasir Ris
Bartley
Serangoon
Buona Vista
Lorong Chuan
Simei
Bencoolen
Paya Lebar
Tampines
Little India
Kembangan
Punggol
Clarke Quay
Holland Village
Buangkok
Tampines East
Farrer Road
Sengkang
Boon Lay
Tampines West
Kaki Bukit
Woodlands
Bukit Panjang
Boon Keng
Pioneer
Yew Tee
Kovan
Hougang
Woodlands South
Nicoll Highway
Sembawang
Jurong East
Chinatown
Jalan Besar
Clementi
Potong Pasir
Choa Chu Kang
Admiralty
Woodleigh
Ubi
Lavender
HarbourFront
Tanah Merah
Rochor
Khatib
one-north
Chinese Garden
Bukit Batok
Ang Mo Kio
Aljunied
Bedok
Lakeside
Bukit Gombak
Bedok Reservoir
Cashew
Dakota
Yishun
Geylang Bahru
Yio Chu Kang
Braddell
Hillview
Commonwealth
Marsiling
Tai Seng
Bedok North
Canberra
Woodlands North
Changi Airport
MacPherson
Mattar
0.2
0.4
0.6
0.8
1.0
1.2
Resale Price
1e6

Planning Area



Flat Model


1e6
1.2
1.0
0.8
0.6
0.4
0.2
resale_price
2-room
Adjoined flat
Apartment
DBSS
Improved
Maisonette
Model A
Model A2
Simplified
Standard
Terrace
Type S1
Type S2


70000
60000
50000
40000
30000
20000
10000
0
count
2-room
Adjoined flat
Apartment
DBSS
Improved
Maisonette
Model A
Model A2
Simplified
Standard
Terrace
Type S1
Type S2

Feature Selection / Engineering

- Features selected in EDA process:
  - Categorical Features
    - One Hot Encoding - Flat Model
    - Target Encoding - Planning Area, MRT Name
    - Label Encoding - Flat Type
    - Binning - Storey Range
  - Continuous Features
    - Floor Area in sqm
    - HDB Age
    - Max Floor Level
    - Number of 3R Flat in the Block of Resale Flat
    - Number of 5R Flat in the Block of Resale Flat
- Standard Scalar

Modeling

Baseline Model	First Iteration	Second Iteration
>0.5 corr 1. Flat Type 2. Floor Area	**Selected in EDA Process** 1. Flat Model (20 variables) 2. Planning Area 3. MRT Name 4. Flat Type 5. Storey Range 6. Floor Area 7. HDB Age 8. Max Floor Level 9. Number of 3R in the Block 10. Number of 5R in the Block	**K-best Feature Selection, K=15** **(Additional 10 variables)** 1. Total Dwelling Unit 2. Number of 2R in the Block 3. Number of Exec in the Block 4. Number of Hawker within 2km 5. MRT Nearest Distance 6. Latitude 7. MRT Latitude 8. Bus Stop Latitude 9. Pri Sch Latitude 10. Sec Sch Latitude

Model Evaluation

3 models:

- Baseline
    - LinearRegression (OLS)
    - 2 features
- First attempt
    - RidgeCV (L2)
    - 29 features selected from EDA
- Second attempt
    - RidgeCV (L2)
    - 39 features after K-best feature selection (K=15)

Baseline

We chose the LinearRegression model as the baseline.

- 2 features, both of which have relatively high correlation with resale price.
- Benchmark for trained models

Scores:

- Cross Val score (R2): 0.44
- Training RMSE: 106, 713
- Validation RMSE: 106, 783

First feature iteration

29 features identified in the EDA process are added into our model:

- Ridge and LinearRegression did not have any real difference. Ridge was chosen at random

Scores:

- Cross Val score (R2): 0.77
- Training RMSE: 68, 183
- Validation RMSE: 68, 033

Second feature iteration

K-best feature selection found 15 features:

- Applied only to continuous features
- 5 of the features can already be found on the previous model, for a total of 39 features
- Ridge was chosen again

Scores:

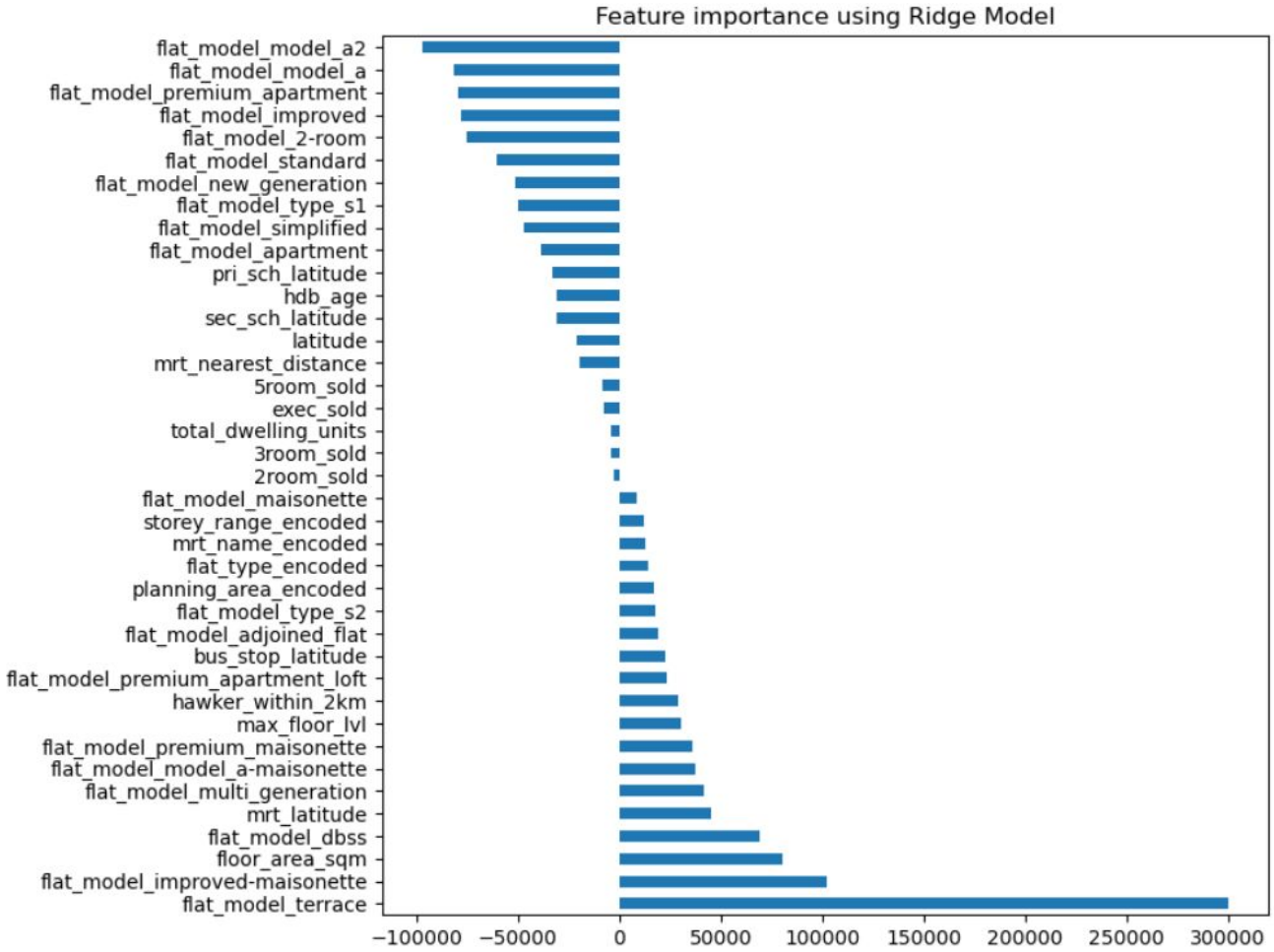
- Cross Val score (R2): 0.83
- Training RMSE: 58, 176
- Validation RMSE: 58, 281

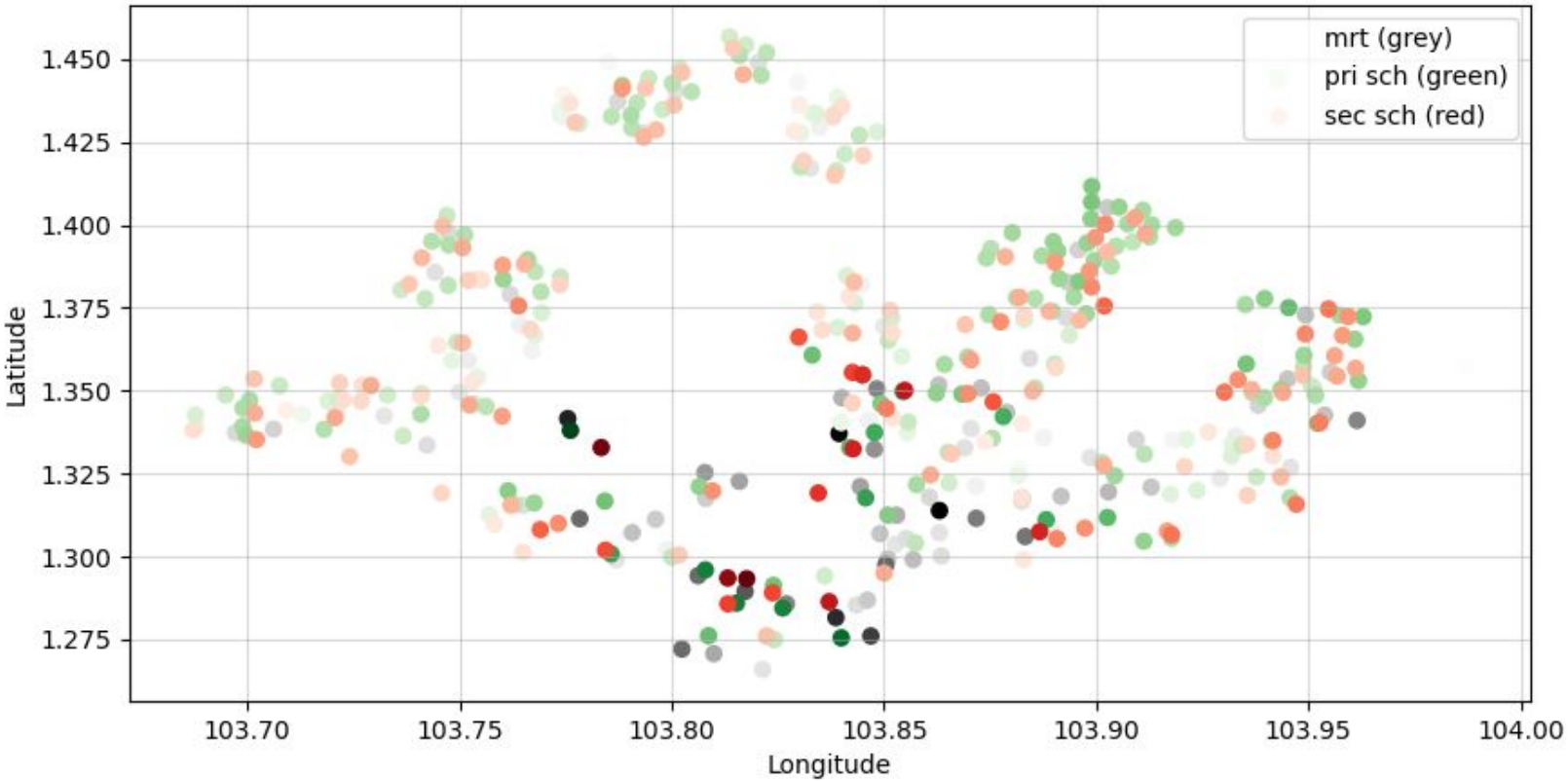
Interpretation of Results

In the final model:

- Cross val score of 0.83 means that these 39 features explain 83% of the variation of the predicted price
- Training RMSE of 58, 176 as compared to Validation RMSE of 58, 281 indicates that bias is high, and hence the model is still under-fitted.
- Add more features, create new features to avoid under-fitting

Interpretation of Results




mrt (grey)
pri sch (green)
sec sch (red)
1.450
1.425
1.400
1.375
1.350
1.325
1.300
1.275
Latitude
103.70
103.75
103.80
103.85
103.90
103.95
104.00
Longitude

Conclusion

Order of Importance

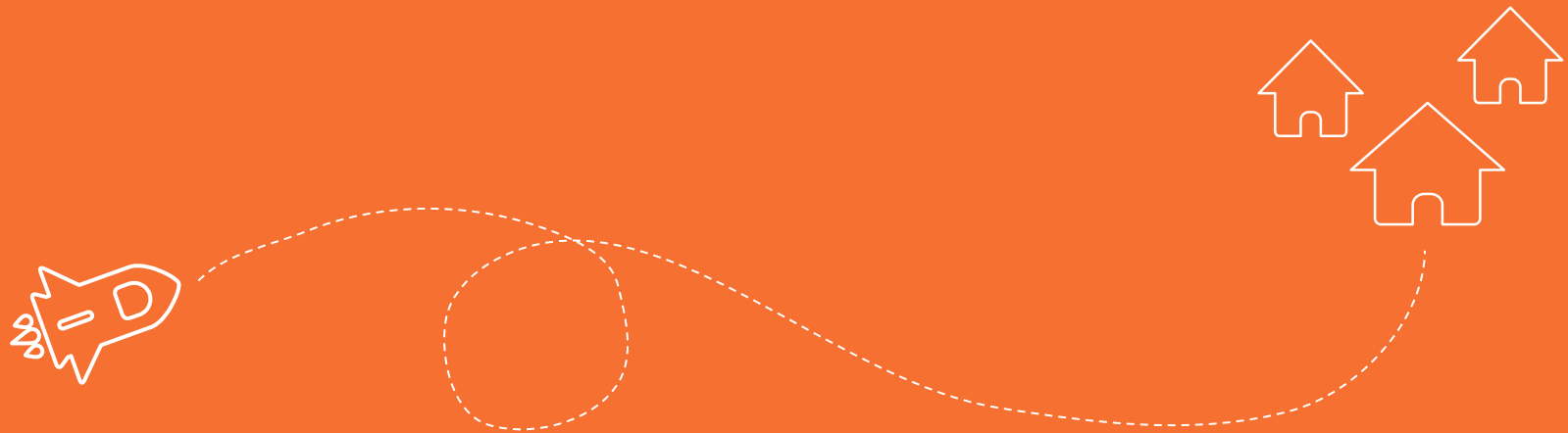
⇩

flat model (particularly terrace)
floor area in sqm
mrt latitude
pri school latitude
hdb age
sec school latitude
max floor lvl
hawker within 2km
bus stop latitude
latitude based on postal code
mrt nearest distance
planning area
flat type

Model	CV R2 Score (Train)	CV RMSE (Train)
Baseline Model	0.4453	106873.66
---	---	---
OLS using EDA feature selection	0.7734	68205.74
Ridge Regression using EDA feature selection	0.7734	68205.58
Lasso Regression using EDA feature selection	0.7713	68519.83
---	---	---
OLS using K-best (k=15, f-reg)	0.8350	58200.76
Ridge Regression using K-best (k=15, f-reg)	0.8350	58200.69
Lasso Regression using K-best (k=15, f-reg)	0.8329	58562.00
---	---	---
Kaggle Submission Score (public)	-	57657.40
Kaggle Submission Score (private)	-	59412.28

Recommendations

- Improve Model Complexity
  - Polynomial features
  - Additional dataset: GDP, interest rate related to Macroeconomic factors
- Target Encoding Method (MRT Name, Planning Area)
  - Some variables have disproportionate groups. Consider apply weightage in each group for target encoding.
- Method to Manage Outliers
  - Use Robust Scaling instead of Standard Scaler
- Applied K-best on Categorical Variables to Rely More on Stats
  - Using chi-square method



Thank you