Project 2 - Singapore Housing Data and Kaggle Challenge

Presented by: Tan Jun Jie, Timothy Chan and Kho Guan Guo
Presented on: 03 March 2023

Agenda

- Problem statement
- Factors for resale price
- Analysis of features affecting resale price
- Modelling: Methodology and Findings
- Model Conclusions
- Model Limitations
- Recommendations

Introduction: Backdrop

1. Singapore’s HDB resale market is becoming increasingly competitive.

2. Buyer, seller and real estate agents are all competing for information.

3. Accurate information is important for decision making.

Problem Statement

Real Estate company needs a model for their agents to

1. Help themselves and clients better understand the **key features** that **affect pricing**.

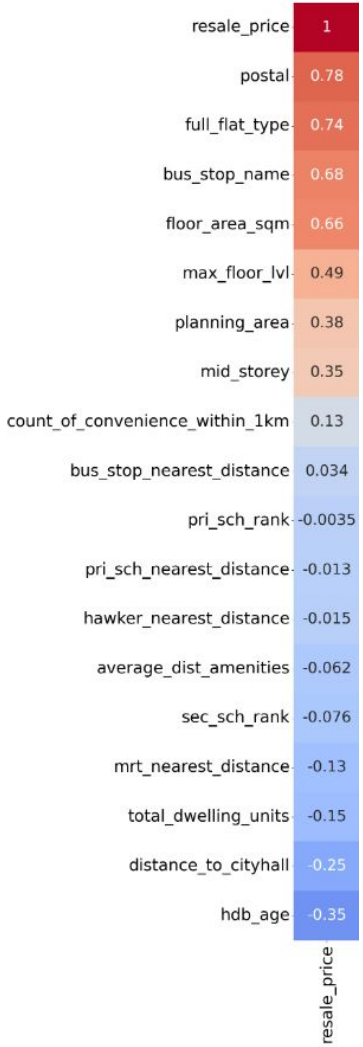
2. Help **selling clients** know what to expect in terms of pricing in order to **list**.

3. Help **buying clients budget** in terms of what features they want from their purchase.

Factors Affecting Resale Price

1. Size
2. Location
3. Age
4. Amenities

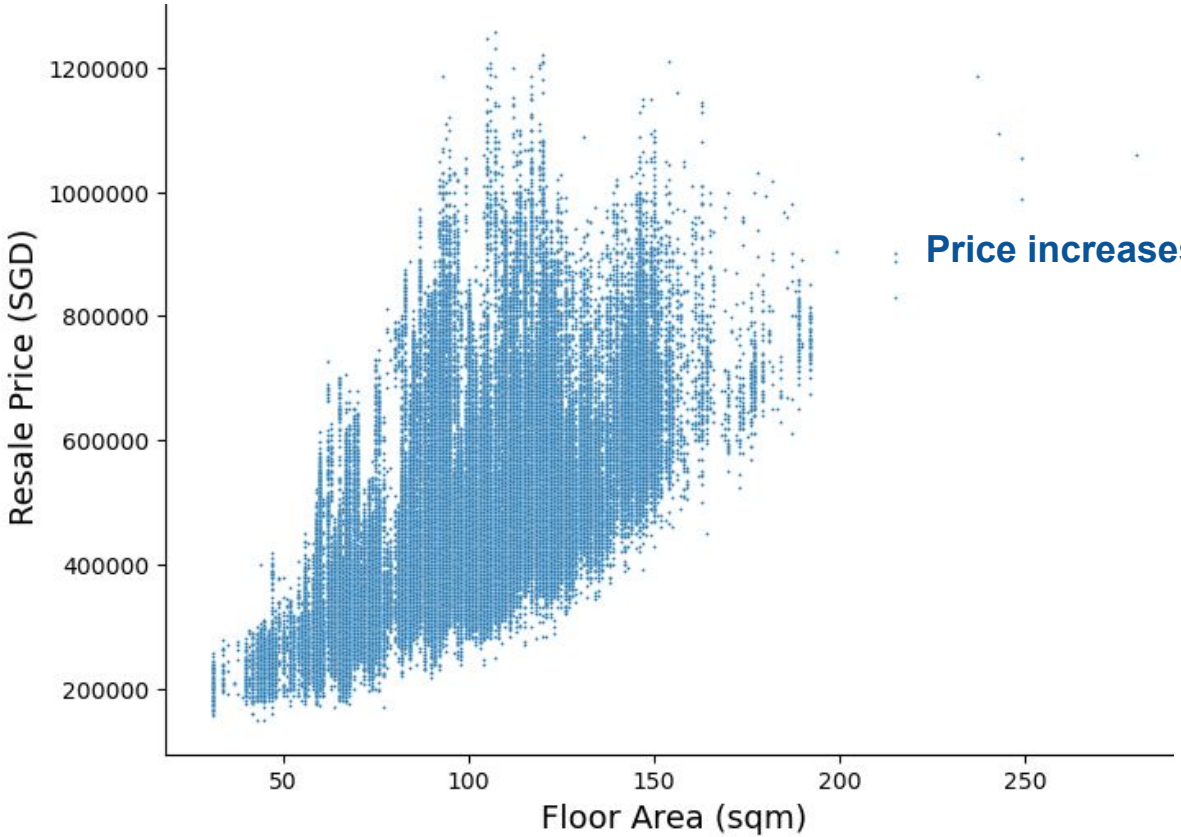
Analysis of features affecting resale price



Strongest +ve corr: location, type of flats, size

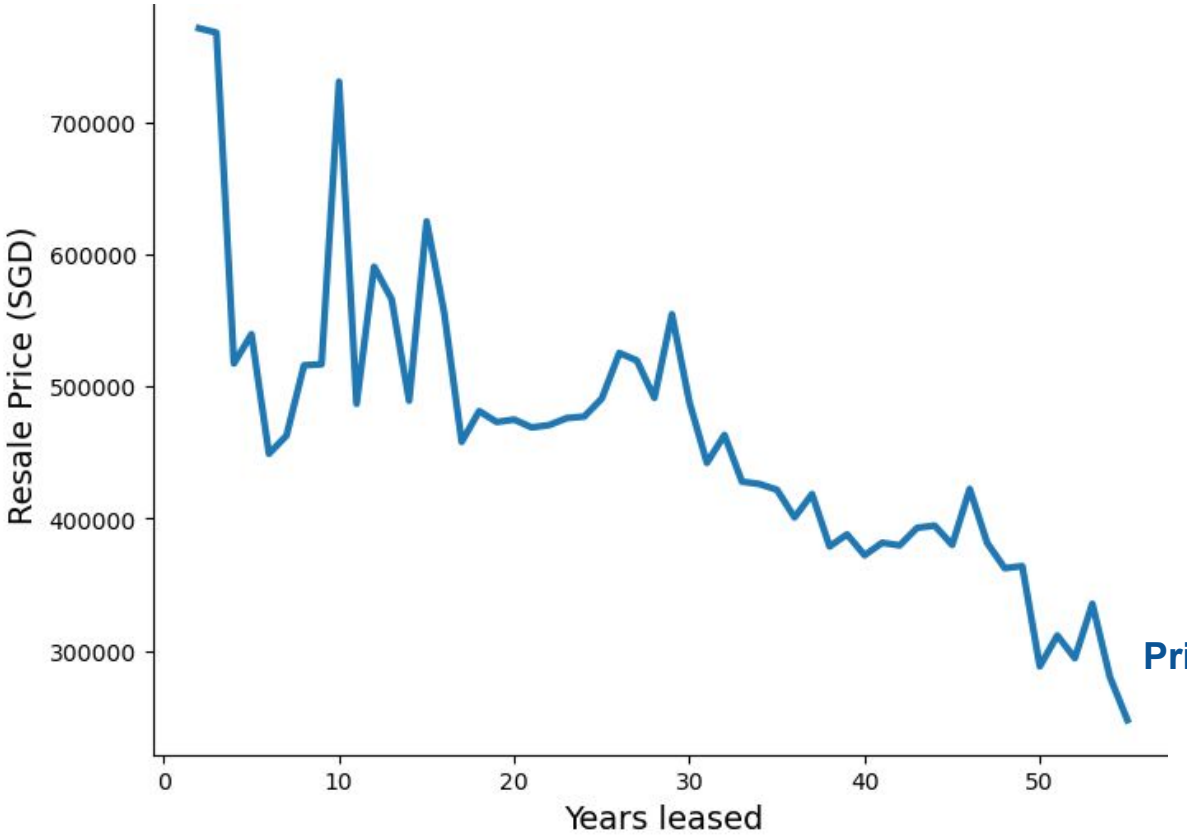
Strongest -ve corr: age and distance to CBD

Resale Price vs Floor Area



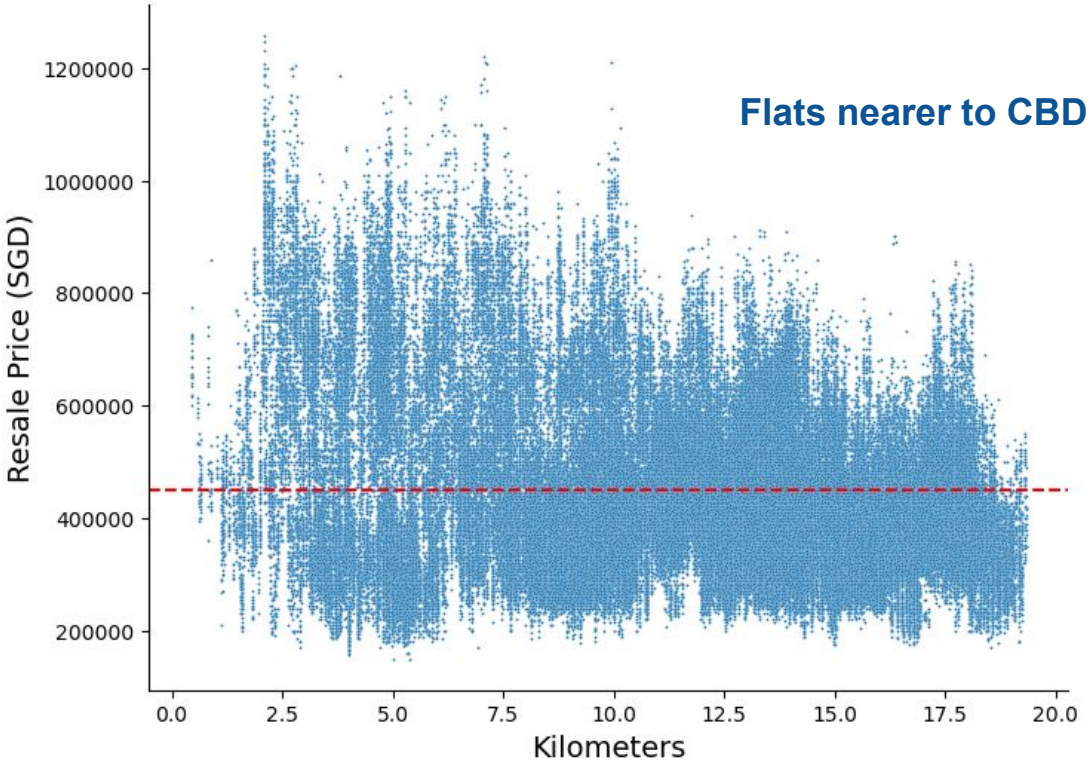
Price increases with size of flat

Average Resale Price vs HDB age



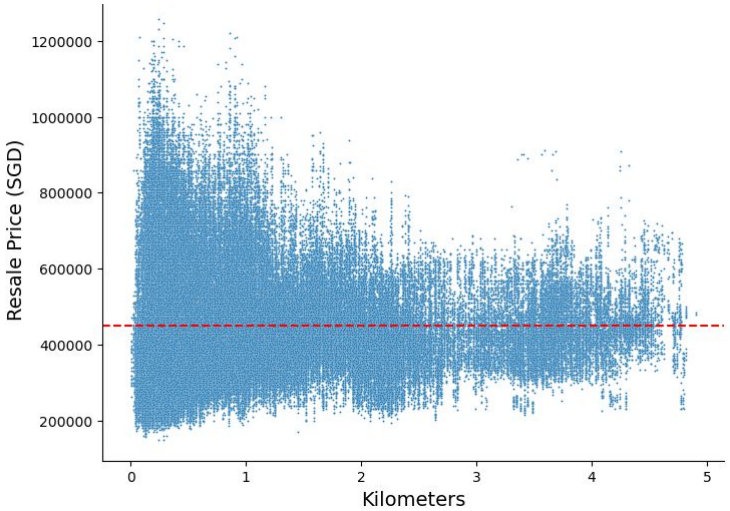
Price decreases as HDB ages

Resale Price vs Distance to City Hall

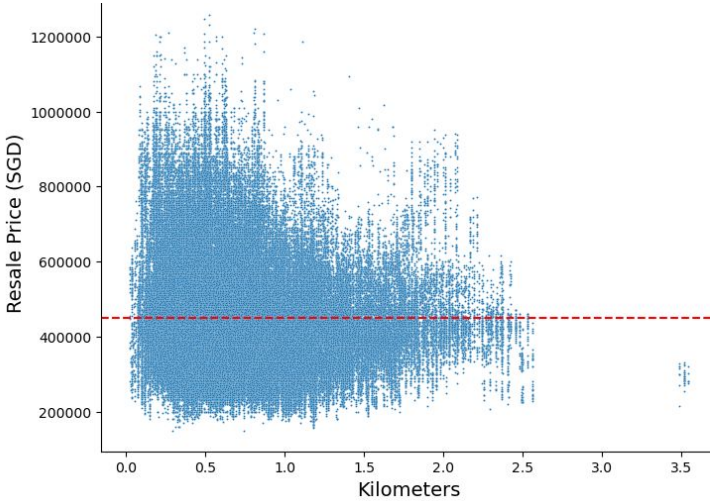


Flats nearer to CBD can command higher prices

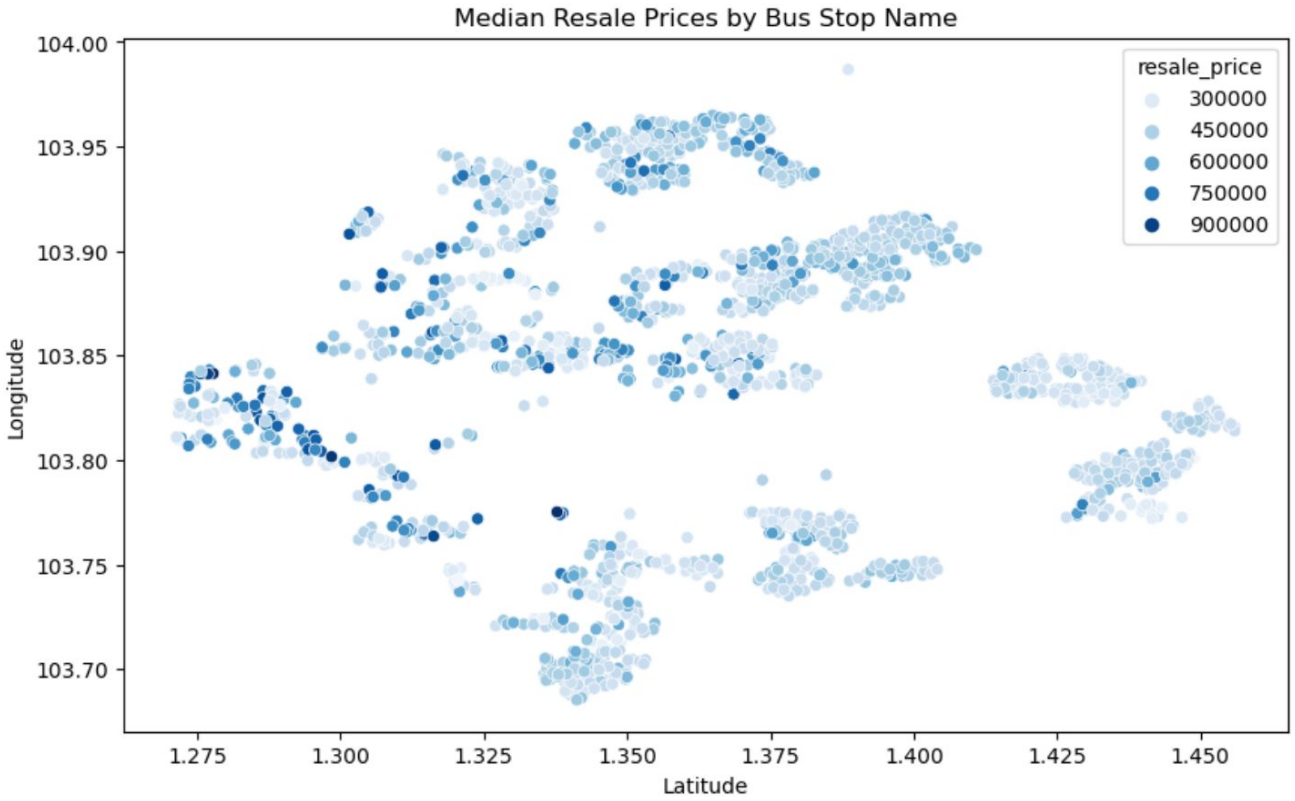
Resale Price vs Distance to Nearest Hawker

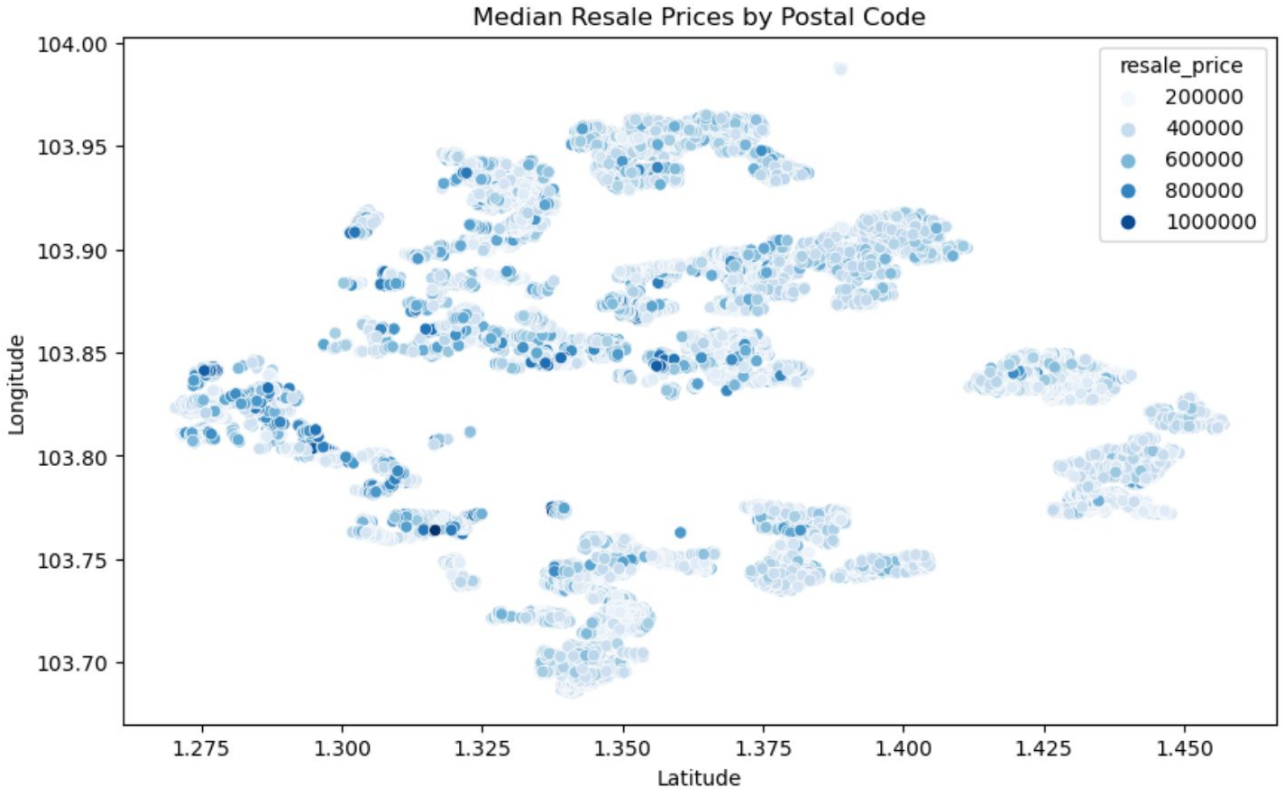


Resale Price vs Distance to Nearest MRT

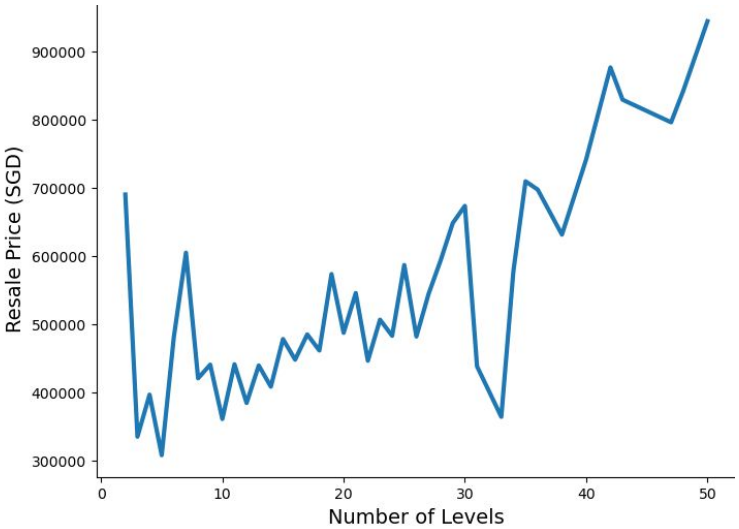



Median Resale Prices by Planning Area
resale_price
300000
400000
500000
600000
700000
800000
104.00
103.95
103.90
103.85
103.80
103.75
103.70
Longitude
1.275
1.300
1.325
1.350
1.375
1.400
1.425
1.450
Latitude

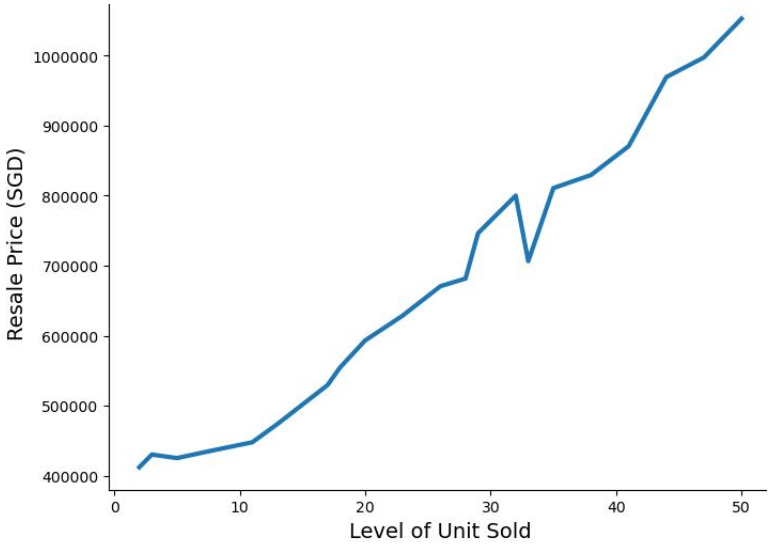

Median Resale Prices by Bus Stop Name
resale_price
300000
450000
600000
750000
900000
104.00
103.95
103.90
103.85
103.80
103.75
103.70
Longitude
1.275
1.300
1.325
1.350
1.375
1.400
1.425
1.450
Latitude


Median Resale Prices by Postal Code
resale_price
200000
400000
600000
800000
1000000
104.00
103.95
103.90
103.85
103.80
103.75
103.70
Longitude
1.275
1.300
1.325
1.350
1.375
1.400
1.425
1.450
Latitude

Avg Resale Price vs Number of Levels



Avg Resale Price vs Level Of Unit Sold



Modelling

Linear Regression Model Benchmarking

Model	Number_of_ Features	training_ RMSE	baseline_ RMSE	RidgeCV_ RMSE	LassoCV_ RMSE	Polynomial_4 _train_RMSE	Polynomial_4 _baseline_ RMSE	Number_of_ Poly_Feature s	Recommend
1	18	$ 54,860	$ 54,630	$ 54,834	$ 54,859	$ 37,760	$ 37,766	7,315	Yes
2	21	$ 54,849	$ 54,609	$ 54,821	$ 54,849	-	-	-	
3	28	$ 54,571	$ 54,314	$ 54,550	$ 54,571	-	-	-	

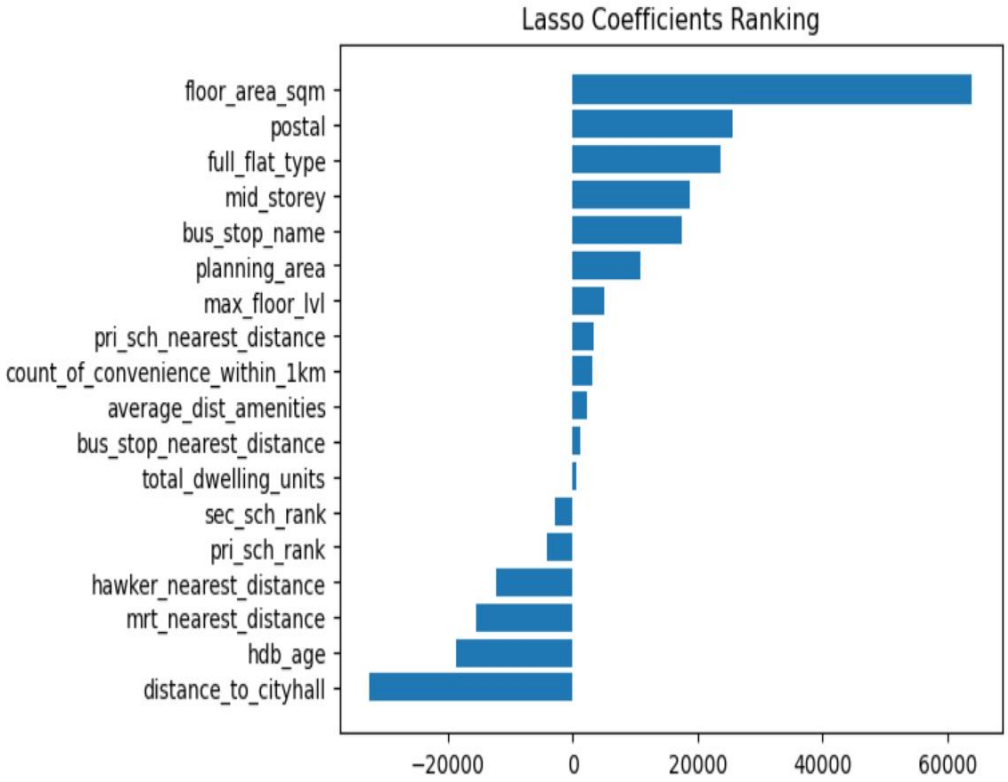
**All model feature inputs have been standard-scaled* **RMSE - root mean squared error*

Simplicity or Accuracy?

	Simple Model	Complex Model
Number of Features	18	7,315
Model Type	Ordinary Linear Regression	OLS with polynomial degree=4
Interpretability	Intuitive	Only a machine can!
Average Accuracy Score^	± 10% error	± 7% error
Types of features	Location, size, amenities, age factors	Location, size, amenities, age, plus interaction features up to a combination of **four* features
Utility	Sufficiently Simple	Accuracy

^Average accuracy score is based on the R-squared score of each model
**The optimal polynomial degree that produces the minimum error based on the set of 18 features is 4.*

Housing Price Premia Ranking



“Size” premium

“CityHall” premium

“West” premium

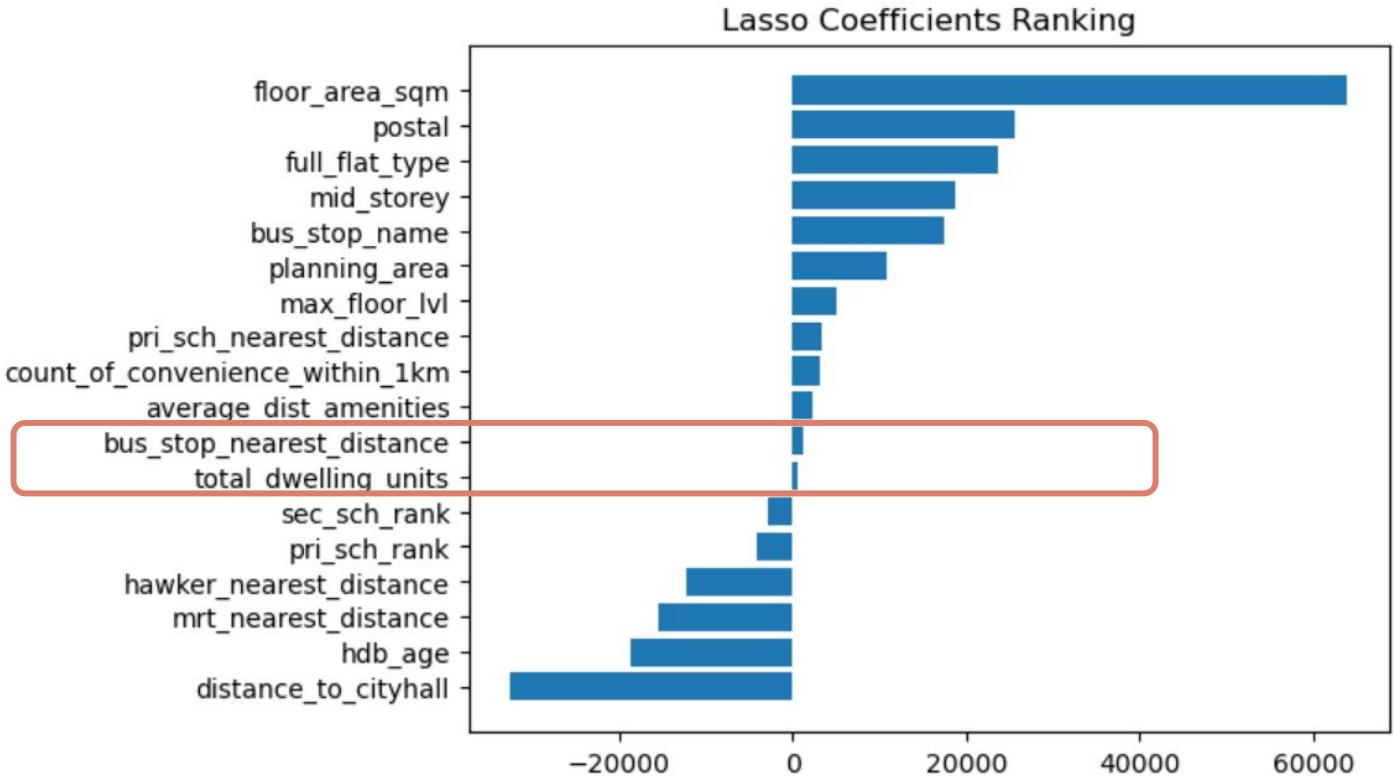
“Height” premium

“Age” premium

“School” premium

“Amenities” premium

Ranking the 18 features, it turns out that total_dwelling_units and bus_stop_nearest_distance are least significant.



Model Robustness

- No overfitting with consistent out-of-sample testing results
- High accuracy value that explains *at least* 80-90% of resale_price
- 18-feature model is simple to use and understand
- Captures the most important features from the four key factors of housing prices
  - Location
  - Size
  - Amenities
  - Age

Model Conclusions

- Location is not King! SIZE IS KING!
- Size > Location > Age > Amenities
- Locations closer to city centre (“City Hall”) are more expensive and taller!
- Locations in the West tend to command a higher premium
- Among the Amenities, proximity to MRT has the biggest premium
- Proximity to good schools is a bonus, not a primary consideration for most

Model Limitations

- Trained on historical data from 2012-2022. Model will need to be retrained on the new market regime.
- The Simple model with just 18 features does not capture sufficient value-add from feature interaction.
- The Complex polynomial model is unintuitive with 7,315 features despite its accuracy.
- Model can only capture 80-90% of price variation meaning clients still have 10-20% buffer with which they will need to budget for.

Recommendations

1. Use simple 18 feature OLS model
   - Clients can expect +/- $55,000 on the predicted price.
   - Simple and easy to use.
   - Accurately captures the predictive value of the 4 mentioned factors.

2. Accurate budgeting for Buy/Sell advice can be given based on the model. (80-90% accuracy)

3. Agents can advise clients based on a structure given in this model. Size > Location > Age > Amenities.

4. Use the 4 main features to help client narrow down search or listing parameters for the best results.