Reimagining Real Estate Intelligence:

Building accurate HDB resale valuations through predictive modelling

Presented by SHYE GROUP

Serene Chan | Chen Yunfei | Efydayana Aditijawarman | Teo Hwee Sze

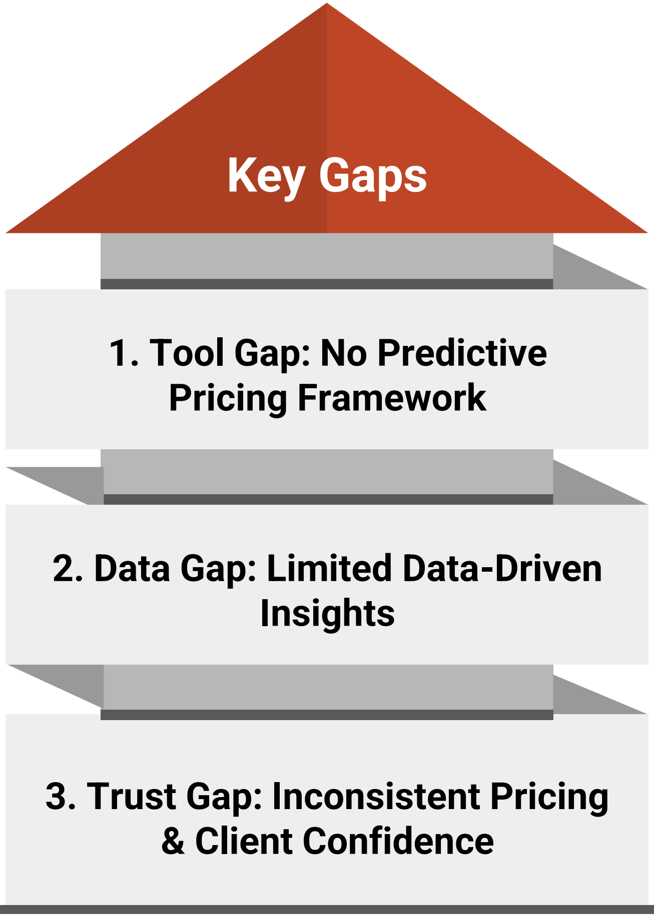
Agenda

1. **Introduction:** Problem Statement, Target Audiences
2. **Our Approach:** Goals, Workflow, Assumptions & Limitations
3. **Exploratory Analysis:** Key factors that Influence Resale Price
4. **Building our Predictive Tool:** Chosen Modeling & Approach, Modeling Results, Recommendations & Product Demo
5. **Reflections**
6. **Q&A**

Problem Statement

The HDB resale market in Singapore is fast-moving and competitive.

WOW! agents currently rely on fragmented data & gut instinct to advise clients, often resulting in pricing inaccuracies and inconsistent client trust.



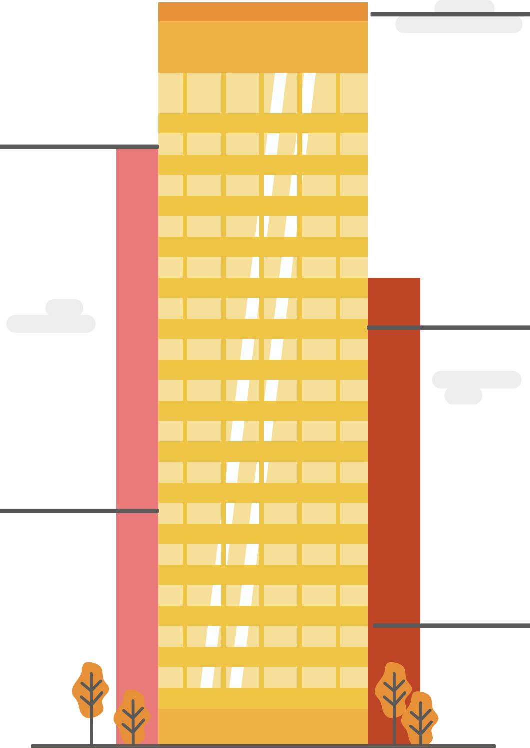
Target Audience

Buyers & Sellers (Clients)

Fair valuations; negotiation confidence
Data-backed price benchmarks

Policy Watchers / Economists

Market stability & trends
Price index analytics



WOW! Real Estate Agents

Accurate, quick pricing advice; client confidence
Data-driven dashboard & price predictor for instant pricing.

Management / Strategy Team

Macro trends, policy impact, market heatmap
Aggregated insights, visual reports

Data Analysts / Data Science Team

Clean, scalable dataset; reproducible analytics
Documented data model & codebase

Our Approach

Project Goals

1. Identify top price drivers to guide agents' pricing strategies

2. Develop two predictive modeling approaches to predict HDB resale value

3. Visualise market insights through interactive tools that help agents simulate pricing scenarios

Project Workflow

Data Market Research to Identify Features
Understand what drives HDB resale prices in Singapore.

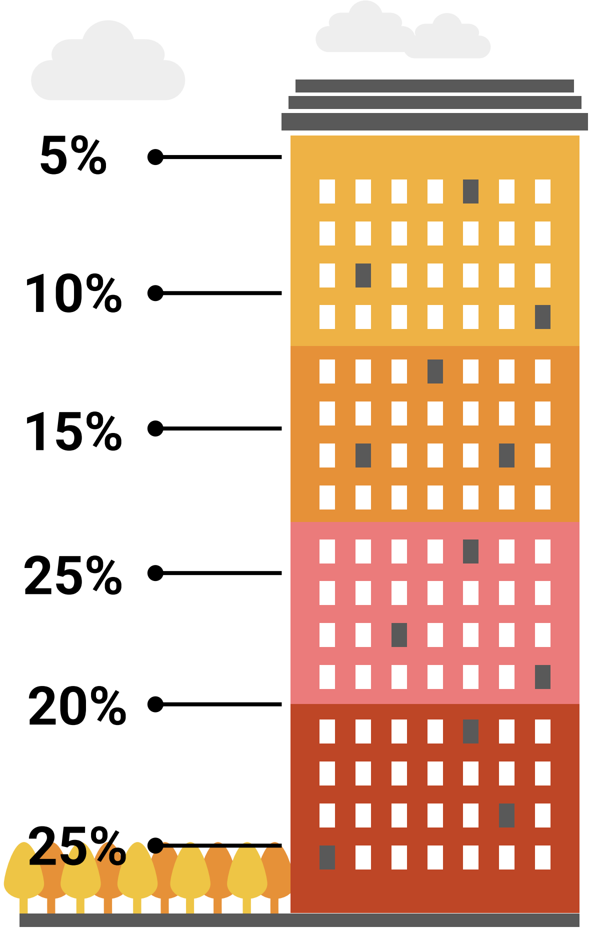
Dataset Collection
Focus on relevant variables for analysis.

Data Cleaning & Wrangling
Prepare the dataset for reliable modeling.

Exploratory Analysis & Correlation Study
Identify features that influences resale price.

Model Development to analyse the results
Train and test a predictive model to estimate resale prices.

Dashboard & Web App Deployment
Deliver a practical tool for real estate professionals.



Data Assumptions, Limitations and Challenges

ASSUMPTIONS

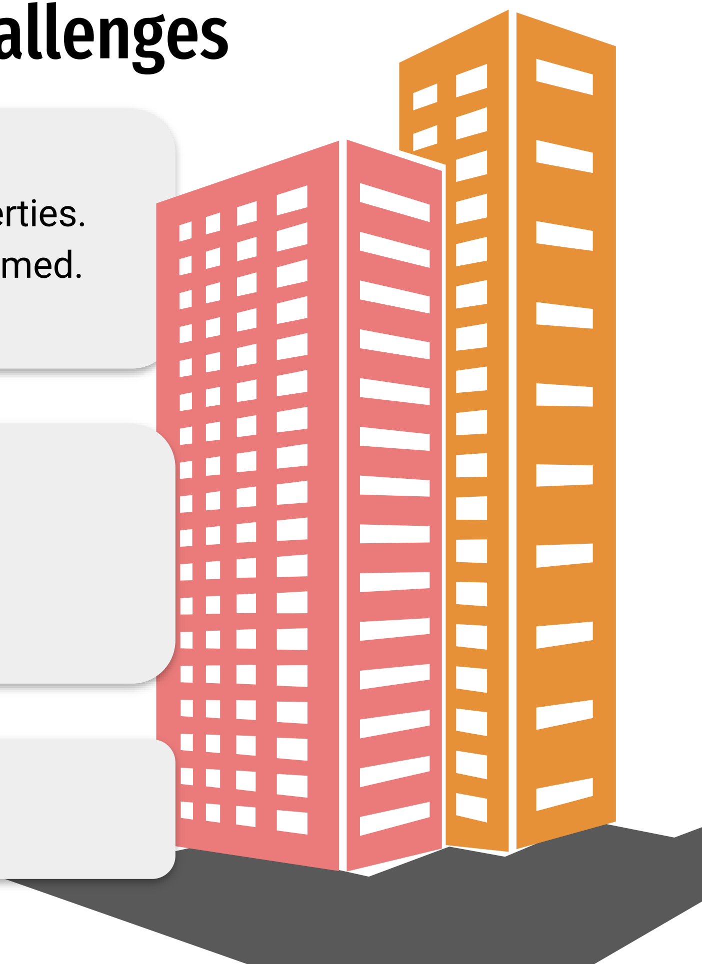
1. HDB resales treated as residential, not investment properties.
2. No liquidity constraints — active buyers and sellers assumed.
3. Demand remains constant within the dataset.

LIMITATIONS

1. Missing demand factors
   (e.g. LTV limits, cooling measure dates)
2. Analysis limited to available dataset features.

CHALLENGES

1. Difficulty identifying impactful features for analysis.



Exploratory Analysis

78 Variables in total - 34 Variables for Analysis - 4 subcategory

Macro Factors

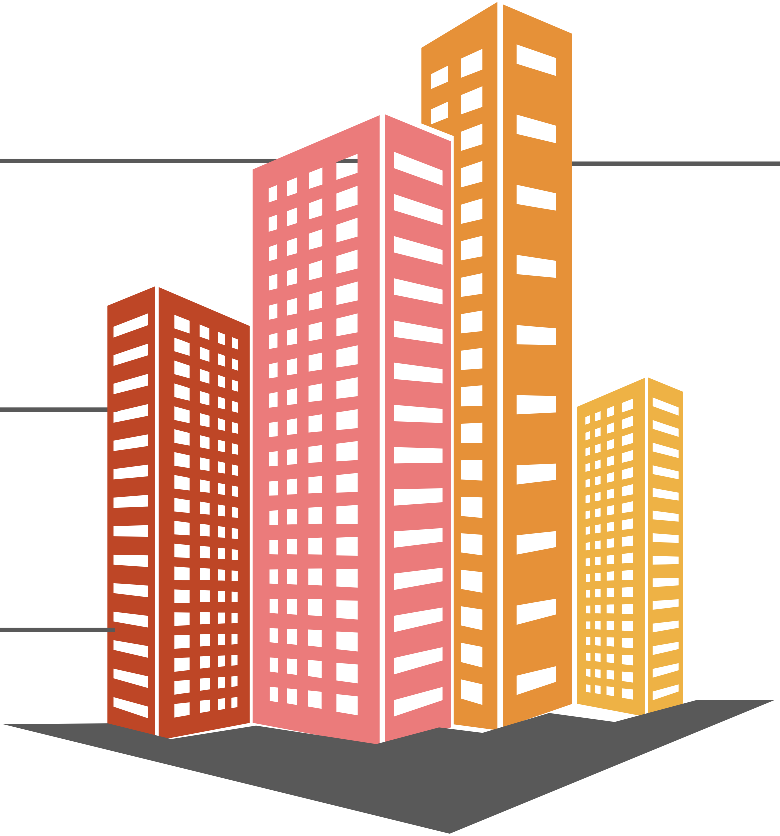
Tranc_Year, Tranc_Month, year_completed and total_dwelling_units

Location Attributes

town and planning_area

Flat Attributes

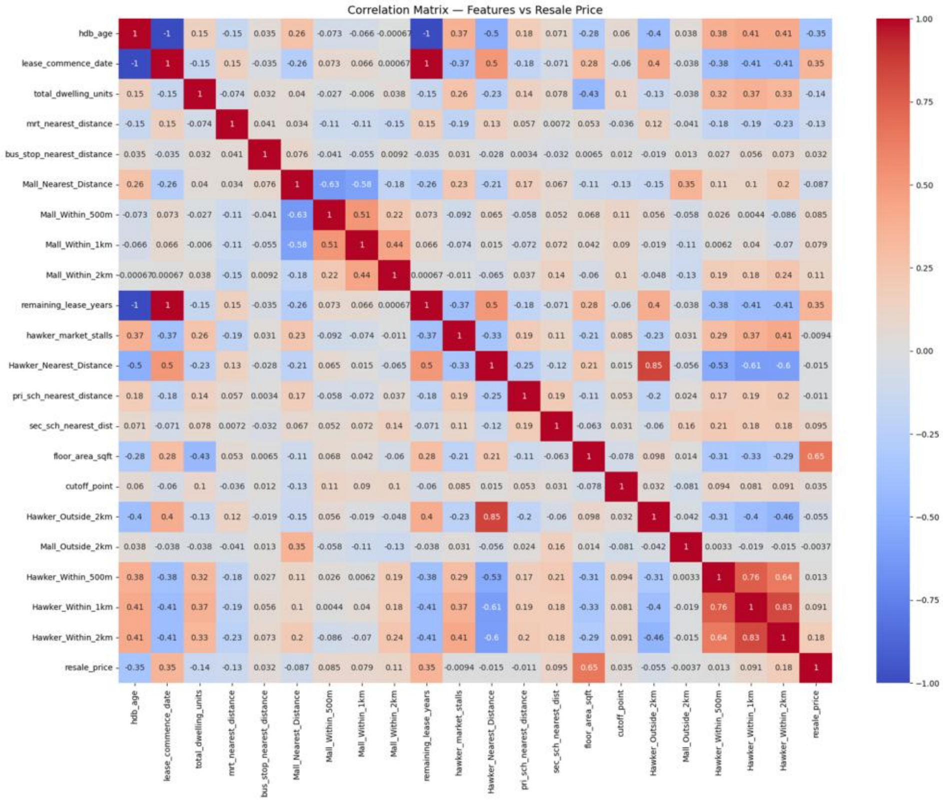
flat_type, flat_model, floor_area_sqft, storey_range, and lease_commence_date



Infrastructure & Accessibility

mrt_nearest_distance, mrt_interchange, bus_stop_nearest_distance, bus_interchange, commercial, market_hawker, mall_nearest_distance, mall_within_500m, mall_within_1km, mall_within_2km, hawker_market_stalls, hawker_food_stalls, hawker_nearest_distance, hawker_within_500m, hawker_within_1km, hawker_within_2km, pri_sch_nearest_distance, pri_sch_name, pri_sch_affiliation, sec_sch_nearest_dist, sec_sch_name, and cutoff_point

Numerical Variables With Highest Correlation Coefficient

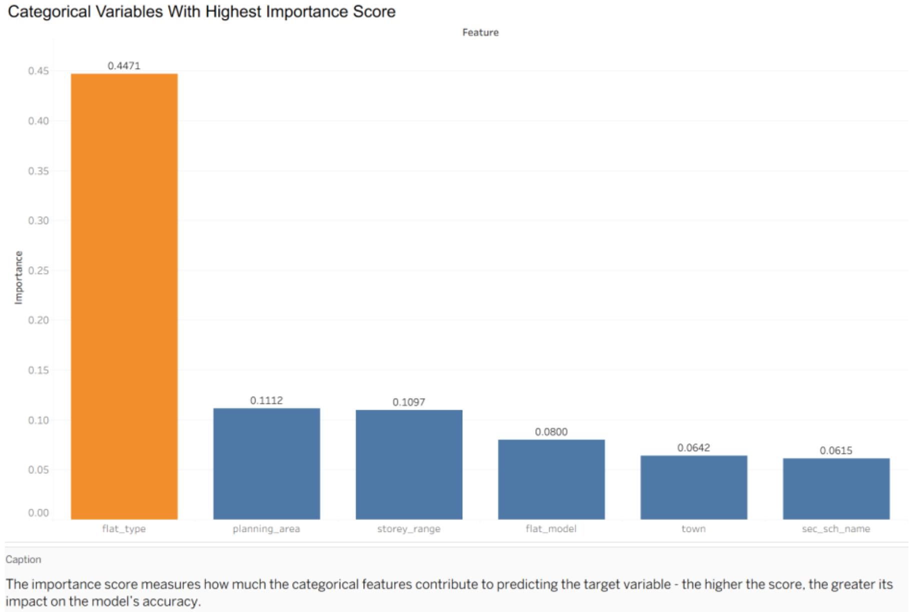


Feature	Correlation Coefficient
`floor_area_sqft`	0.65
`hdb_age`	-0.35

Variable **correlation** measures the strength of the relationship between two variables. Higher correlation (close to +1 or -1) means that a target variable is more associated with changes to the variable.

Correlation coefficients above 0.5 (or below -0.5) **suggest stronger relationship worth exploring in prediction models.**

16 Categorical Variables - 6 With Highest Importance Score



Variable **importance score** measures how much variable contributes to predicting the target variable - the higher the score, the greater its impact on the model's accuracy. **importance score of ≥ 0.05 (5 %) suggest that the feature is important to the model accuracy.**

Key Variables that Influence Resale Pricing

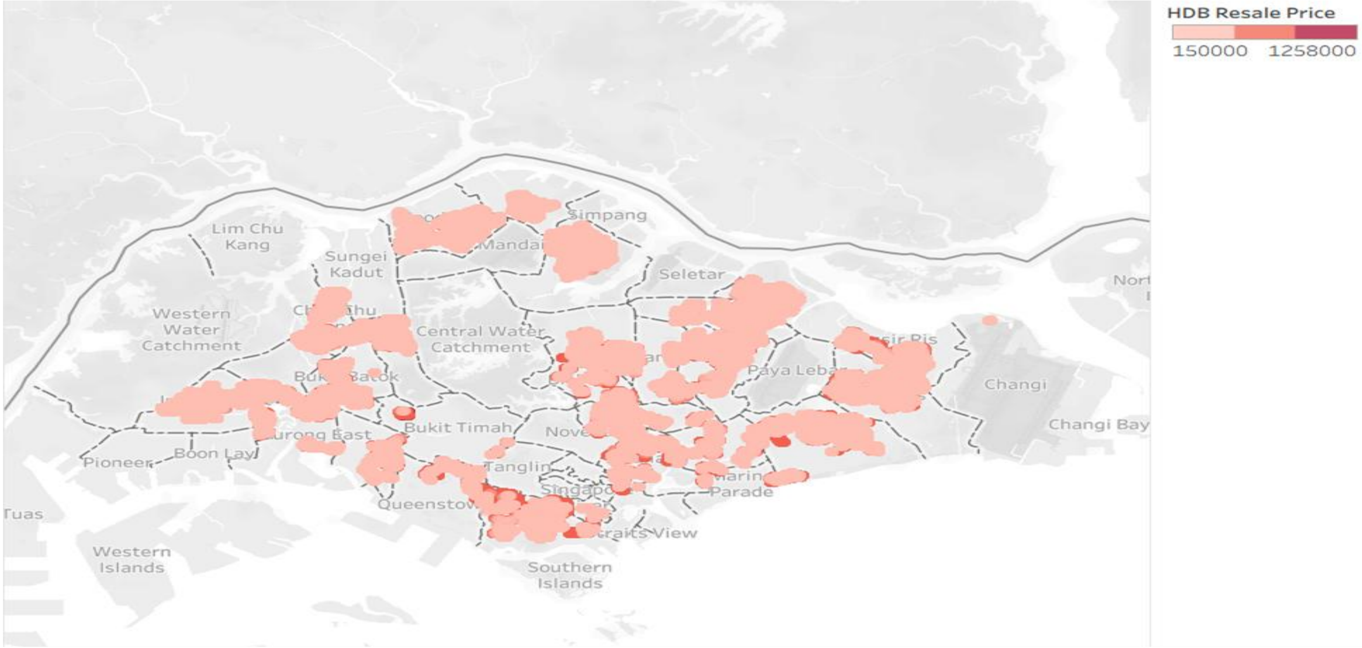
Location Variables

`planning_area`

Flat Variables

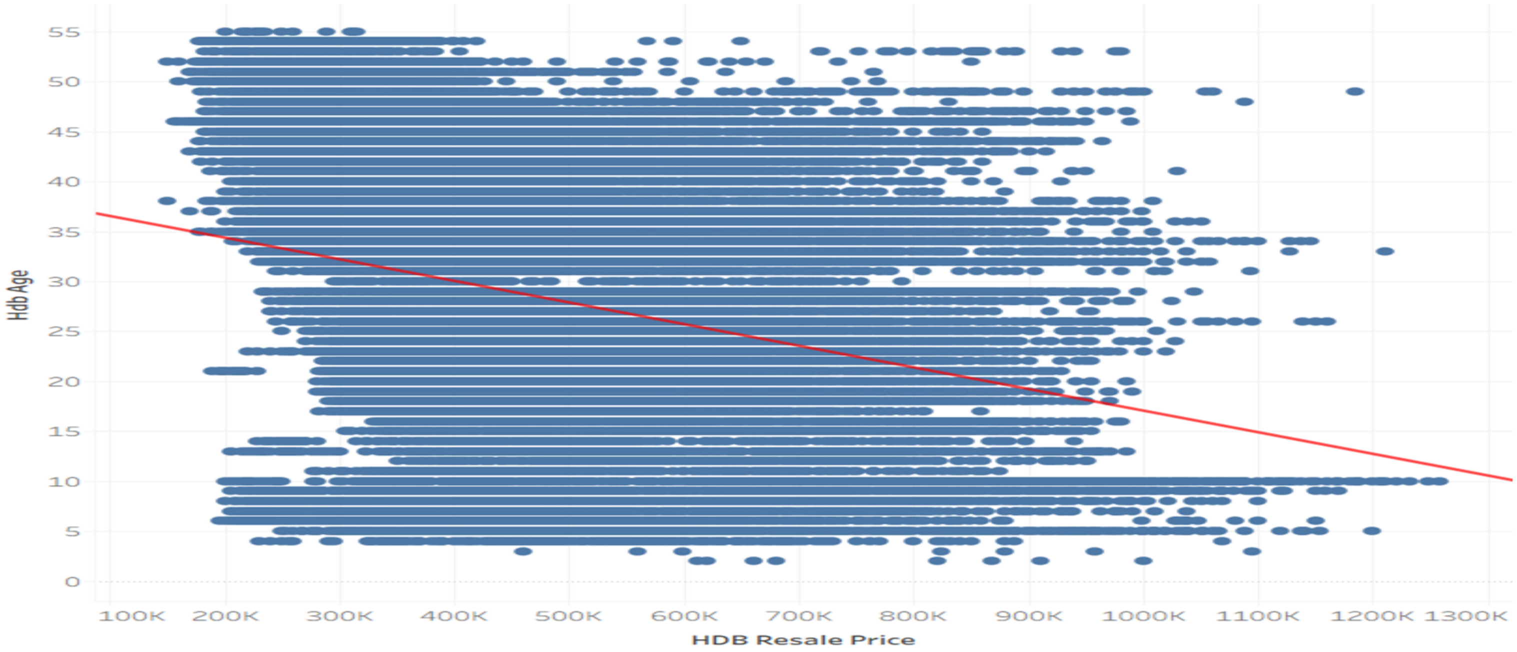
`hdb_age, flat_type, floor_area_sqft, storey_range`

Impact of Planning Area on Resale Price:



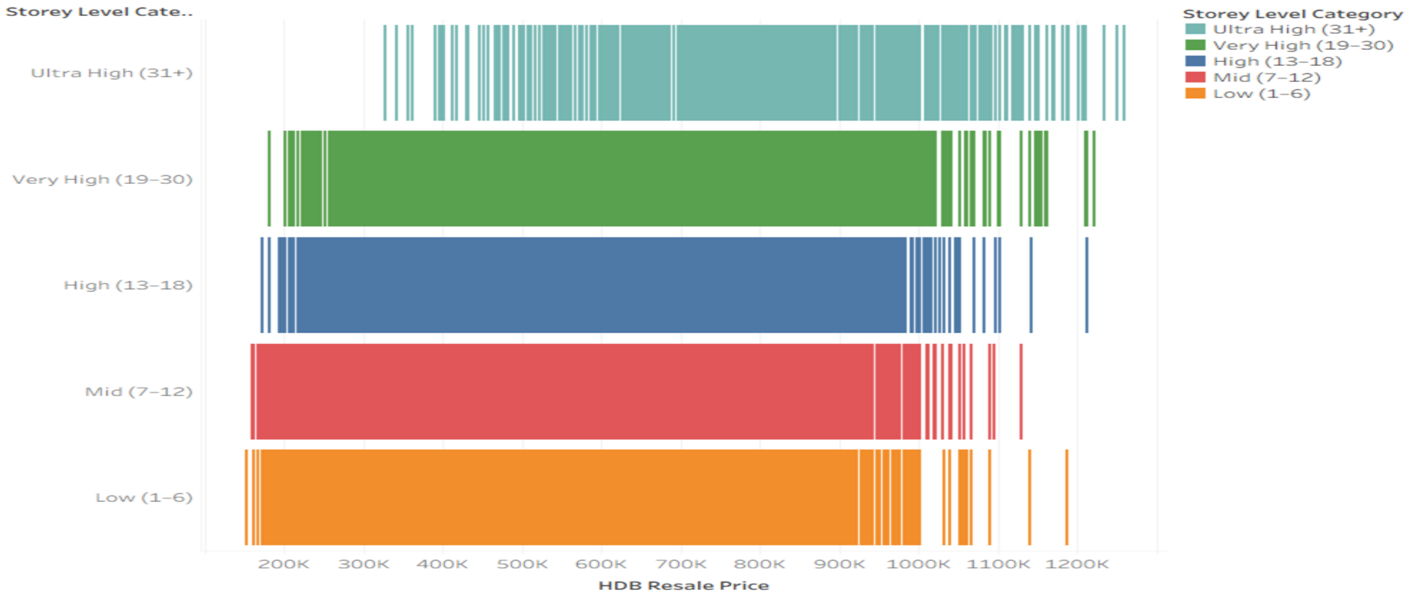
The darker the shade, the higher the price. Higher resale price flats are mainly concentrated in central-east areas

Impact of HDB age on Resale Price:



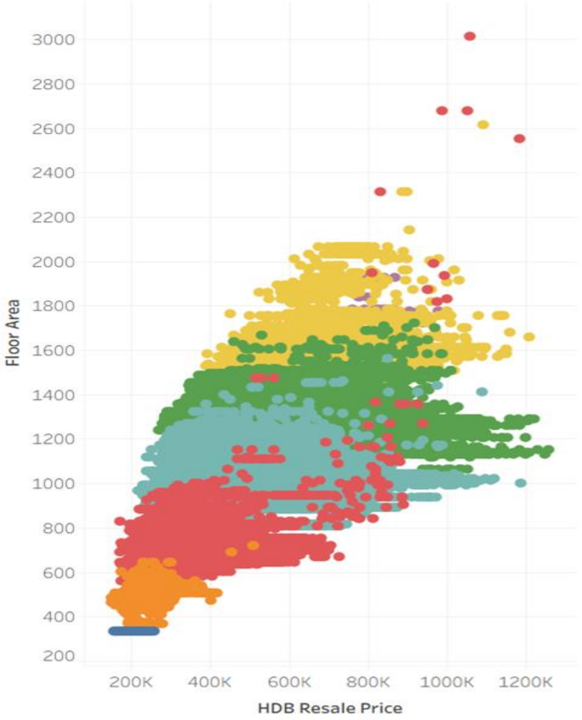
Each additional year of flat age generally reduces resale price, holding other factors constant.

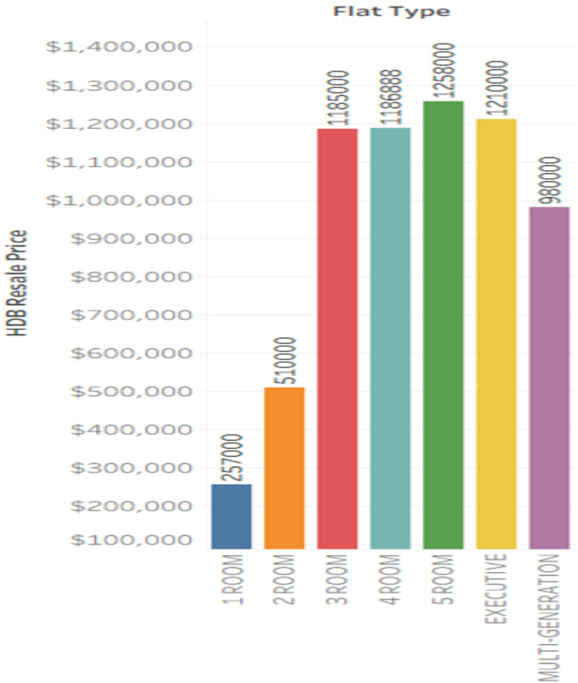
Impact of Storey Range on Resale Price:



Low (1-6): Lowest prices. Narrow spread. | **Mid (7-12):** Most common Range. Wider Spread.
High (13-18): Shift to higher prices. Denser mid-to-high. | **Very High (19-30):** Strong premium. Peaks near $800K - $1M. | **Ultra High (31+):** Highest prices, few but very high value-units.

Impact of Floor Area & Flat Type on Resale Price





Floor area: Positive correlation between floor area & resale price. Each flat type forms a distinct horizontal band.

Flat Type: resale price increases steadily with flat size, from 1ROOM to 5ROOM. Executive & Multi-generational flats command premium prices due to larger space & specialised design. Larger or more premium types will command higher value.

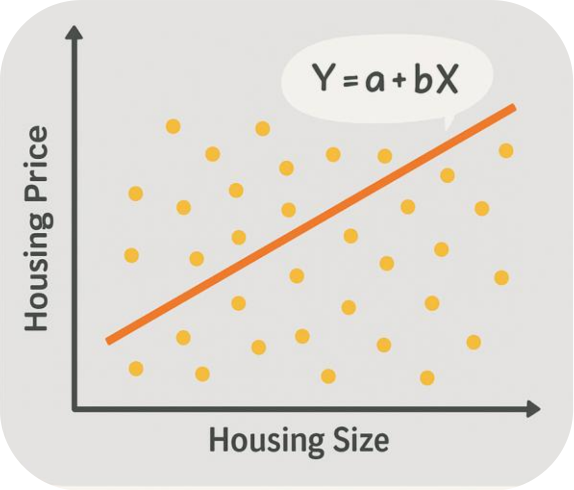
Building our Predictive Pricing Tool

Linear Regression as Chosen Predictive Modeling

A simple yet powerful predictive model that predicts a continuous numeric value — in this case, HDB resale prices — based on different features.

The Goal: Find the best-fitting straight line that predicts resale price based on a set of features. Each feature contributes a specific weight (or coefficient) to the final price prediction.

How it Works: By training on thousands of past resale transactions, the model learns the *average impact* of each feature and applies that knowledge to estimate the likely resale price for unseen flats.



Approach for Building & Evaluating Linear Regression

1. Split dataset into 80% train & 20% test
2. Use Linear Regression on 80% train data
3. Test Linear Regression on 20% test
4. Assess model based on root mean squared error & mean absolute error

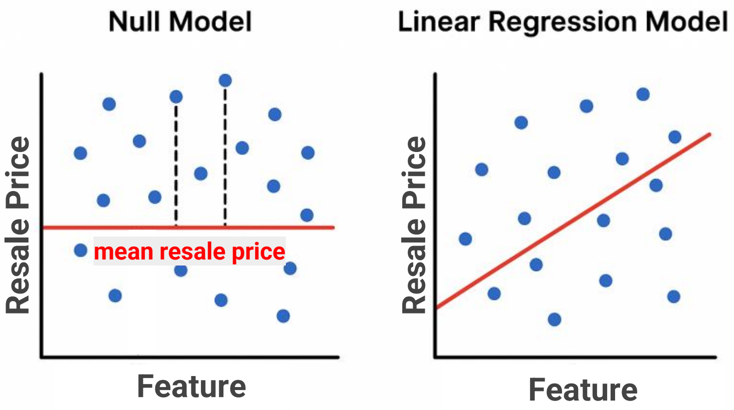
How to determine if model is good?

Compare the model's RMSE and MAE against null benchmark model. The lower, the better!

<u>Null Benchmark Model</u>
➡️ Simplest baseline that predicts the same mean resale price for everyone — no features/predictors.

<u>RMSE (Root Mean Squared Error)</u>
➡️ Shows how far off the predictions are on average, giving more weight to larger errors.

<u>MAE (Mean Absolute Error)</u>
➡️ Shows the average size of prediction errors, treating all errors equally.



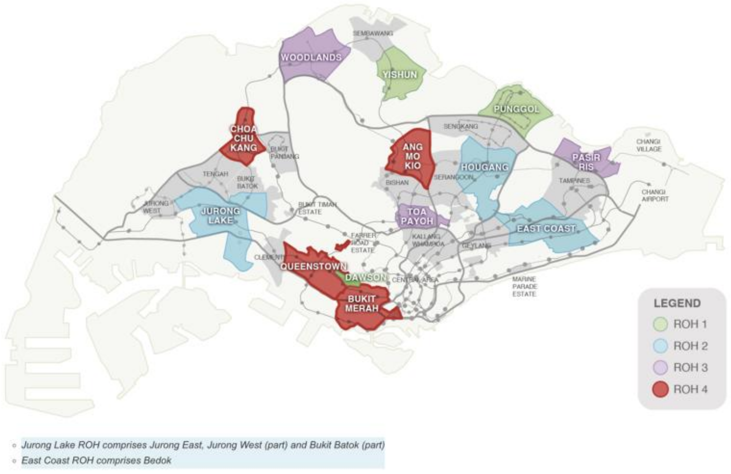
High RMSE & MAE (big errors)

Lower RMSE & MAE (smaller errors)

Two Approaches to Validate

MODEL 1: PREDICTING BASED ON SPECIFIC PLANNING AREA

MODEL 2: PREDICTING BASED ON URA CLASSIFIED REGIONS



Selected Variables for Model 1

```
[('hdb_age', np.float64(-4213.337124287925)),
 ('floor_area_sqft', np.float64(308.1516509218447 5)),
 ('flat_type_1 ROOM', np.float64(-94703.23889673832)),
```

Flat Attributes

`hdb_age,flat_type,floor_area_sqft,`
`storey_range`

Location Attributes

`planning_area`

VARIABLE CHANGE:

`storey_range`: As there are 20+ unique labels with overlapping ranges in the data set, new categories created to standardize range for more stable modeling:

A. Low floor (1–6)
B. Mid floor (7–12)
C. High floor (13–18)
D. Very High floor (19-30)
E. Ultra High floor (31+)

Selected Variables for Model 2

Location Attributes

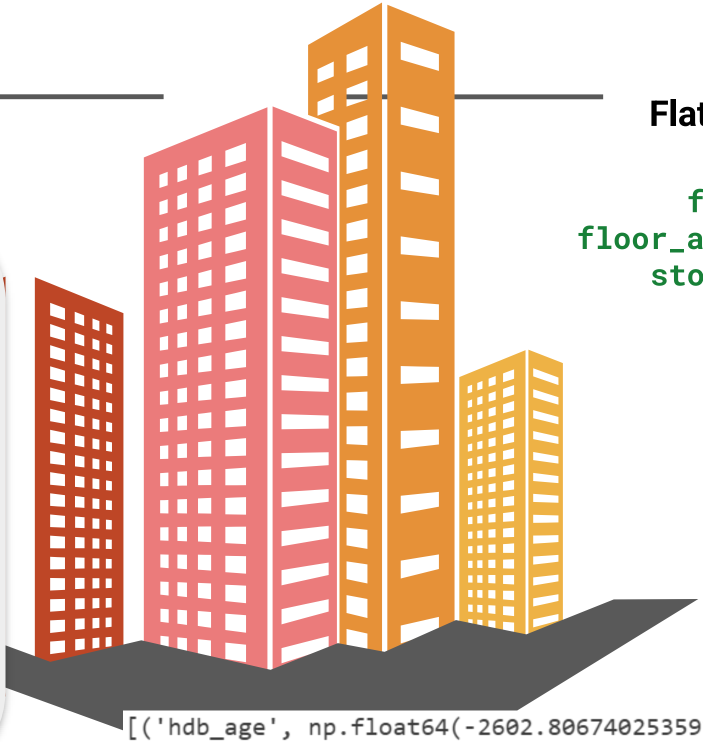
`planning_area`, `region_grouping`

VARIABLE CHANGES:

Added - `region grouping`: to test whether grouping planning areas by URA classification is a more accurate price predicting approach. Planning area reclassified:

Core Central Region (CCR)
Rest of Central Region (RCR)
East Region (ER)
North Region (NR)
North-East Region (NER)
West Region (WR)

Dropped - `planning_area`: As we want to predict resale price based on region, this variable is redundant.



Flat Attributes

`hdb_age`,
`flat_type`,
`floor_area_sqft`,
`storey_range`

```
[('hdb_age', np.float64(-2602.8067402535917)),
 ('floor_area_sqft', np.float64(299.7460764755317)),
 ('flat_type_1 ROOM', np.float64(-87520.11033106191)),
```

Modeling Results

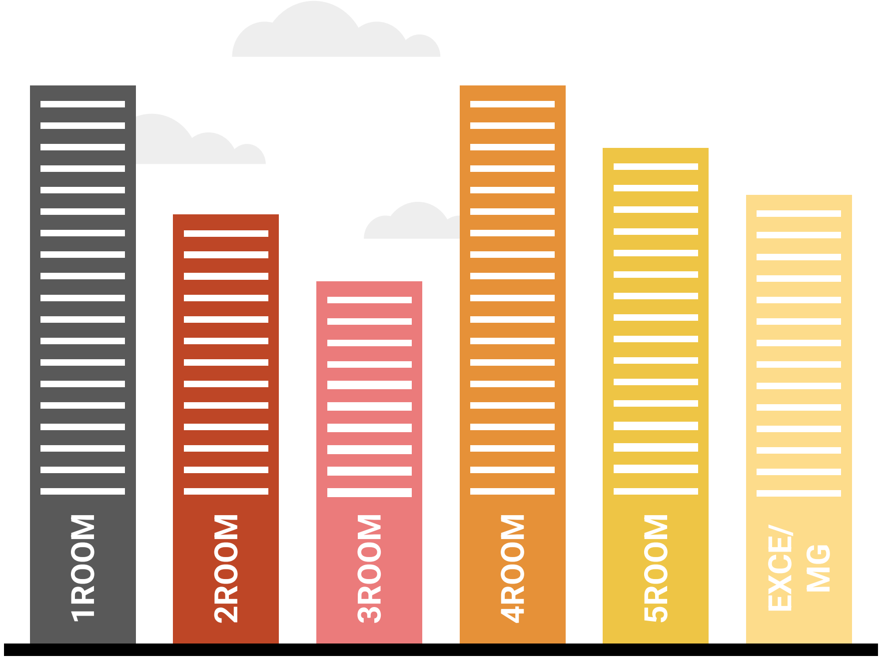
Model	RMSE ($)	MAE ($)	RMSE Improvement (%)	MAE Improvement (%)
Area specific (model 1)	61,772.47	47,956.77	56.98	56.55
Region Grouping (model 2)	74,442.42	56,472.18	48.15	48.83
Benchmark (null)	143,573.95	110,371.43	-	-

INSIGHTS

- Both models' predictions are about twice as accurate as the benchmark (as seen in RMSE improvement and MAE improvement), indicating strong performance for noisy and complex datasets.
- **Model 1's RMSE ($61,772.47) is higher than Model 2's ($74,442.42).**
- This also tells us that using region mapping does not give us a higher RMSE and is not as accurate as Model 1.
- **Given the relative performance, we choose Model 1 as the final model to predict resale_price.**

Recommendations for HDB Resale Pricing

RECOMMENDATIONS



Key variables that WOW! agents should focus on to deliver accurate pricing advice:

1. **Flat Type:** Resale prices depends on flat type
2. **Floor Area Sqft:** Resale price increases as floor area sqft increases
3. **Planning Area:** Prime areas like Central fetches higher resale price
4. **HDB Age:** Resale price depreciates as number of years from lease commence date increases.
5. **Storey Range:** Indicates the level like low, mid high to segregate the floor level for pricing

Use our predictive pricing tool to estimate!

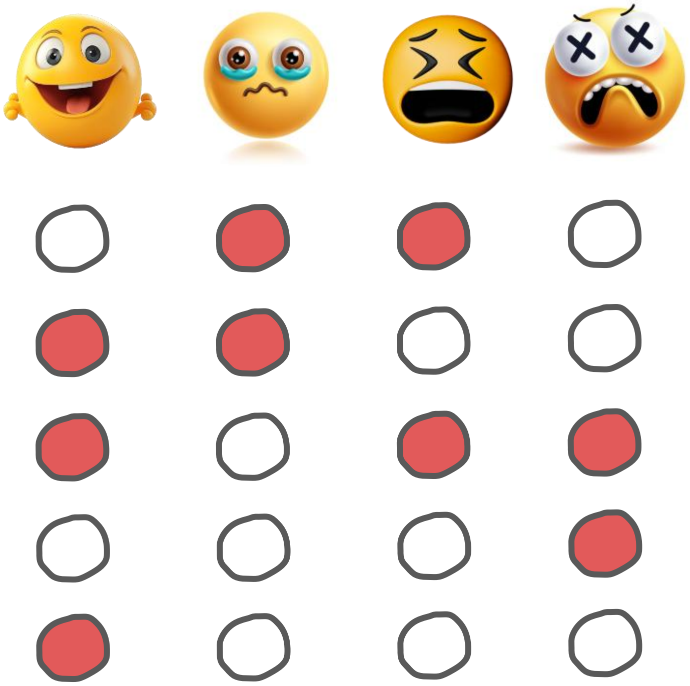
Try Our Pricing Tool!

Please make sure your device is connected to GA’s wifi to access the tool

Try Our Pricing Tool!

Our Reflections

Team Reflections



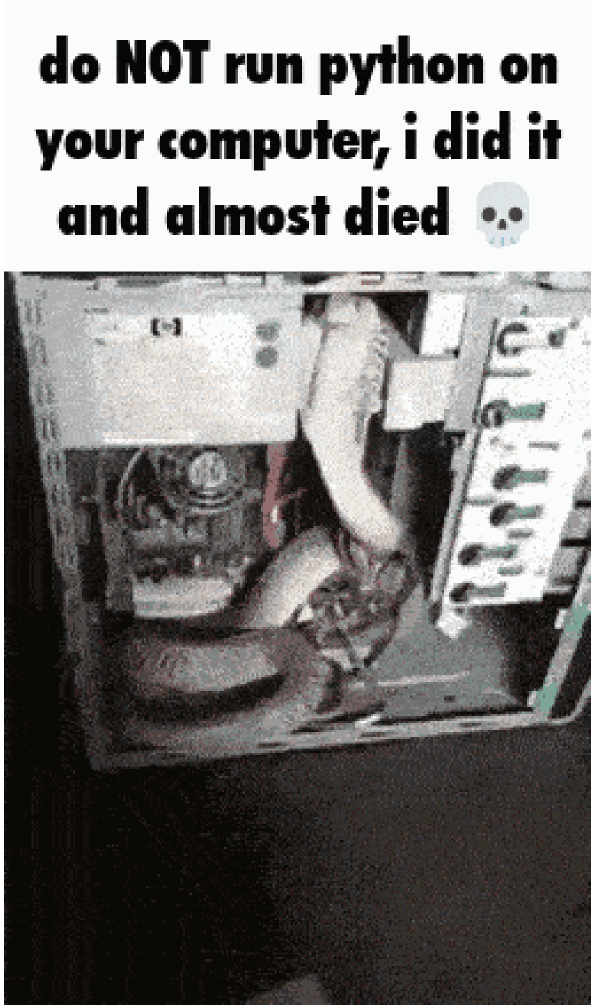
Reviewed our Day 1 sprint and successfully built a proper **work structure and flow**

Adopted a **“divide and conquer”** approach — each task had two members for better collaboration.

Crash course in building a Streamlit web app — intense learning curve under tight time pressure.

Time constraints made debugging and testing stressful.

Improved **team communication and alignment** by setting shared expectations and project tone.


do NOT run python on your computer, i did it and almost died 💀

Q & A

The End.