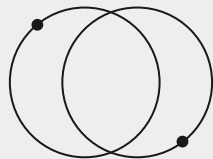Darwis | DSI-38

HDB Resale Price Prediction

Analysis of HDB resale prices based on various contributing factors
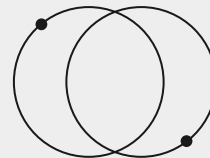
✳ ✳ ✳ ✳

Table of contents

01

Project
Background & Goal

02

Dataset

03

Exploratory Data
Analysis

04

Model &
Evaluation

05
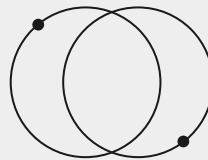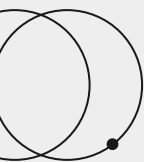
Conclusion &
Recommendation

❊ ❊ ❊ ❊

01

Project Background & Goal

Project Background

HDB resale price has been steadily increasing over the decades, especially post COVID-19 pandemic. Understanding key factors that influence the price is crucial to maximize return of investment (ROI).

Main Goals

UNDERSTAND HDB FEATURES AFFECTING RESALE PRICE

SUCCESSFULLY PREDICT HDB RESALE PRICE BASED ON PROPOSED MODEL

HDB resale price index

Index

176*

180
160
140
120
100
80
60
40

Q1 1994   Q1 1998   Q1 2002   Q1 2006   Q1 2010   Q1 2014   Q1 2018   **Q2 2023**

*Flash estimate

Source: HOUSING BOARD
STRAITS TIMES GRAPHICS
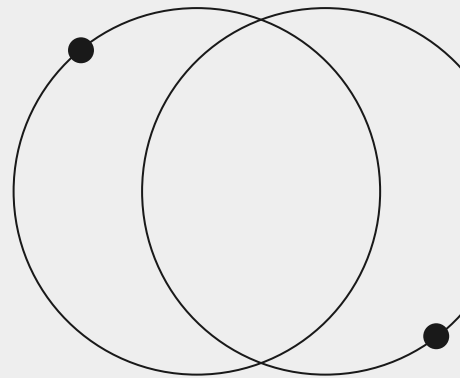
02

Dataset

Image Source: _Unsplash_
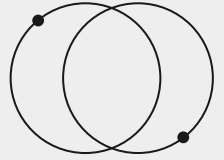
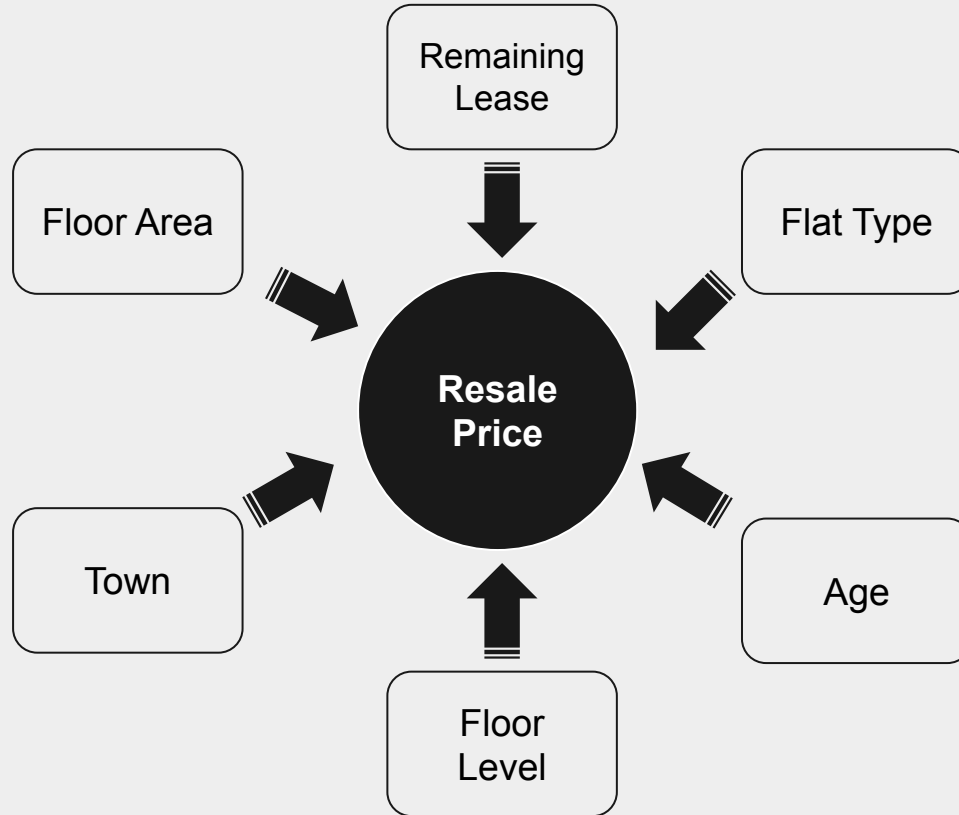# 150,000

Number of Transaction

70+

Number of Features

10 years

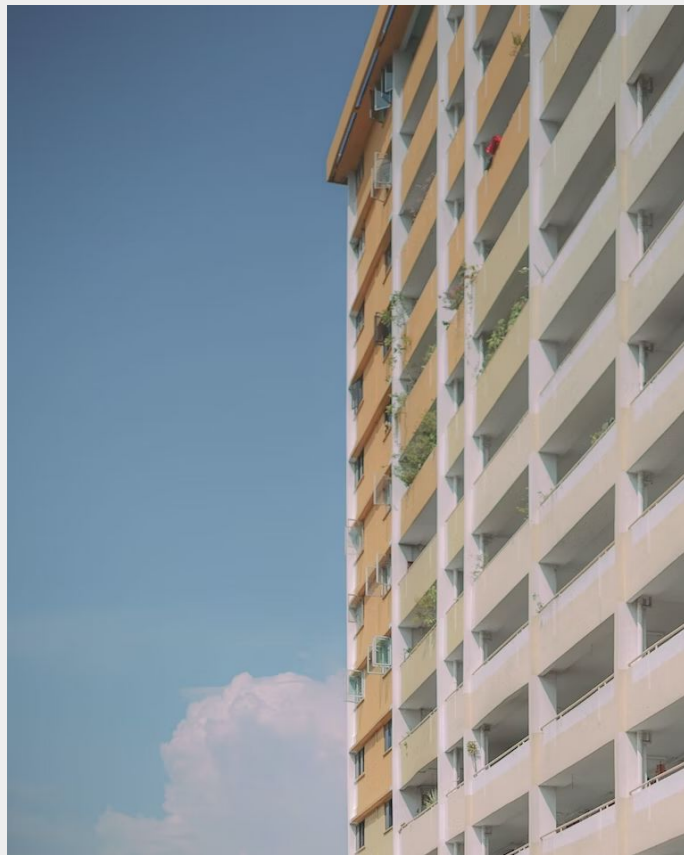Period of Data Collection (2012-2022)

Key Features

03

Exploratory Data Analysis

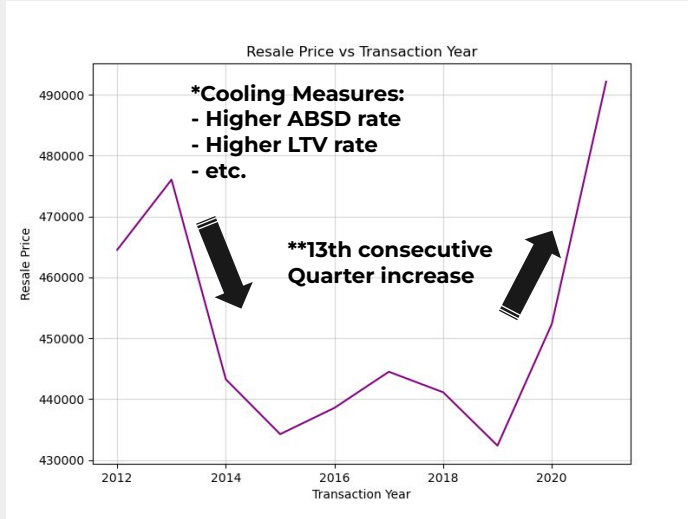*Image Source: Unsplash*

#1: Annual Resale Price

Resale Price vs Transaction Year

*Cooling Measures:
- Higher ABSD rate
- Higher LTV rate
- etc.

**13th consecutive Quarter increase

Resale Price vs Transaction Month (by Year)

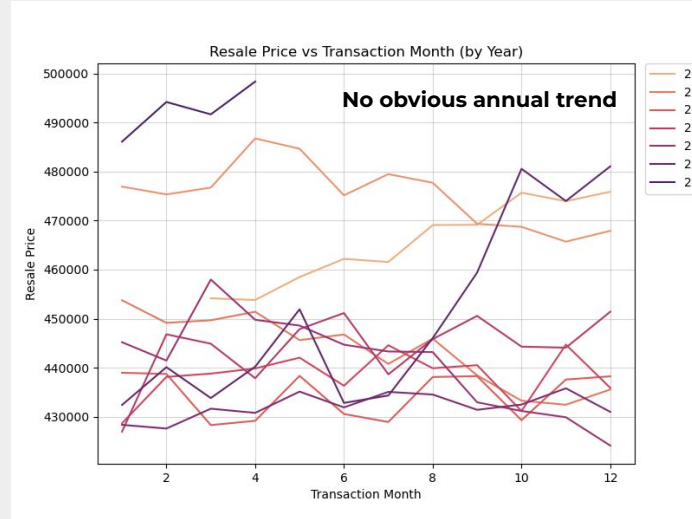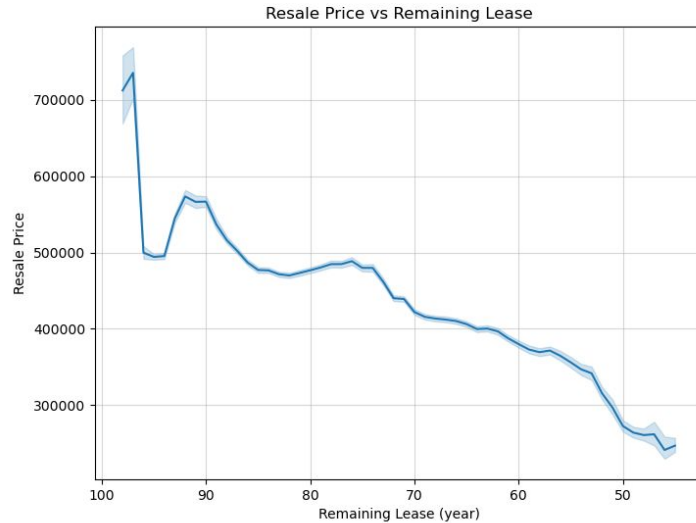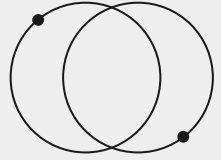No obvious annual trend
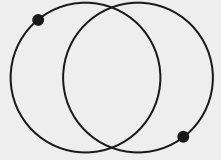
*Source: Stackhomes
**Source: The Straits Times

#2: Resale Price vs 99-year Remaining Lease

Remaining Lease ↘

Resale Price ↘

#3: Resale Price vs HDB Storey/Size

Resale Price vs storey_range

Resale Price vs flat_type
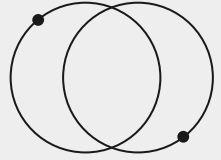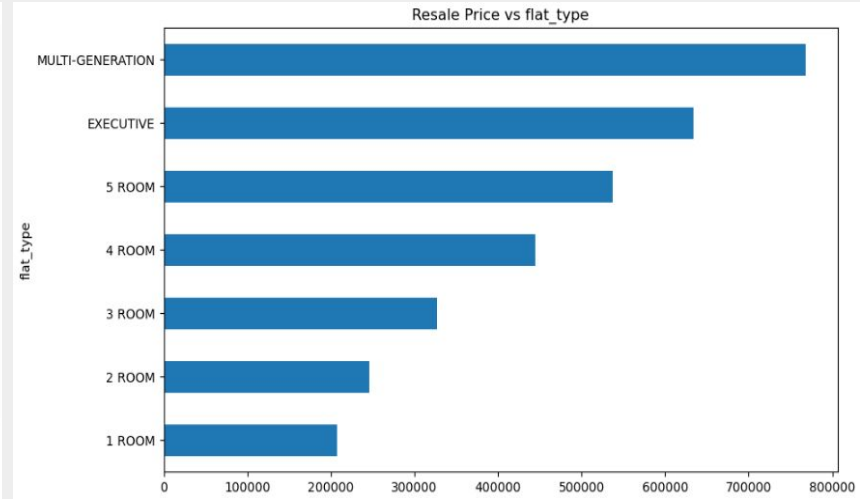
Floor ↗ Resale Price ↗

Unit Size ↗ Resale Price ↗

04
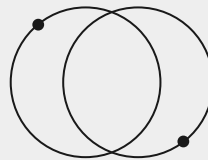
Model & Evaluation

Data Preparation & Modeling

Handling Missing Data

Replace missing data with 0s

Data Separation

into binary/numerical/ categorical

Feature Engineering

optimize feature and drop redundant ones

Train-Test Split

60% train 40% test

One-Hot Encoding

translate categorical features to 0/1s

Standard Scaler

standardize numerical features

Model Data

using linear/ridge/ lasso regression model

Evaluate Model

assess R^2 and RMSE

Model Evaluation

	Linear Regression	Ridge Regression	Lasso Regression
Model Accuracy (R2)	-5.8 x 10^16	**0.915**	0.897
Error in Prediction (RMSE)	3.4 x 10^13	**41,625**	45,664
Model fitting (Train-Test % difference)	100,000,000,000%	0.1%	0.1%

⇒ **Ridge Regression** has the highest model accuracy and demonstrate repeatability.

⇒ **Linear regression** is worst with significantly varying result during cross-validation (model not robust).

Ridge Regression (actual vs prediction)
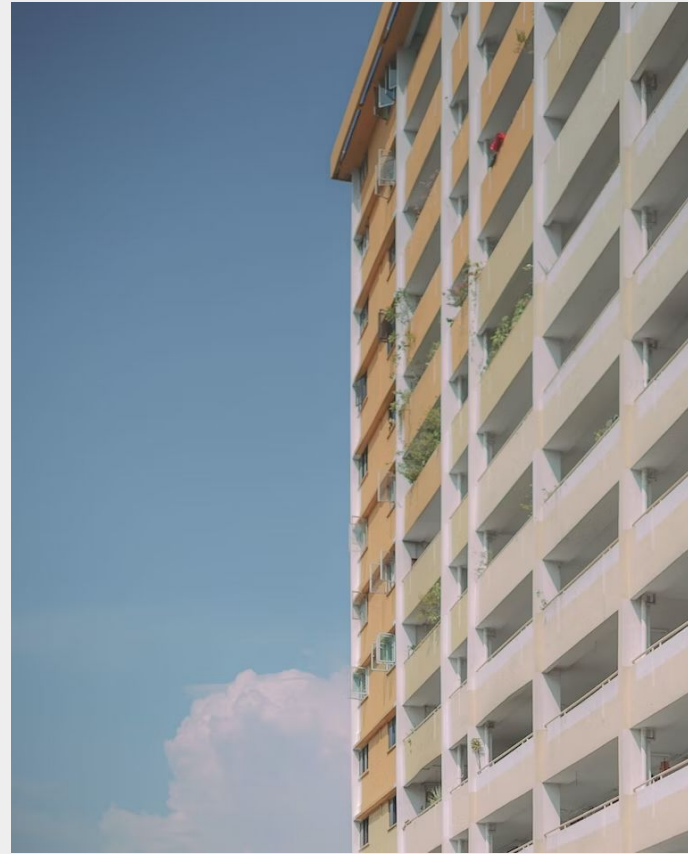
05
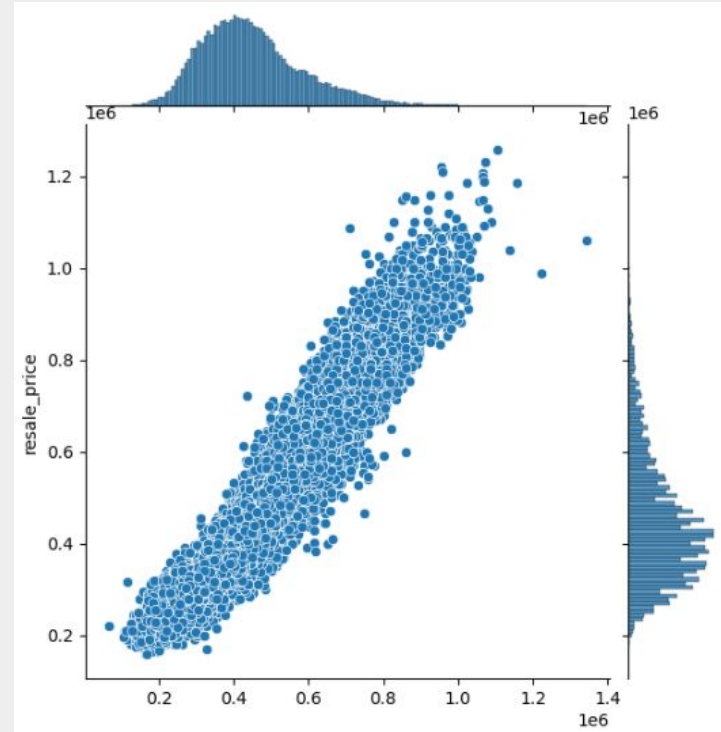
Conclusion & Recommendation

Image Source: _Unsplash_
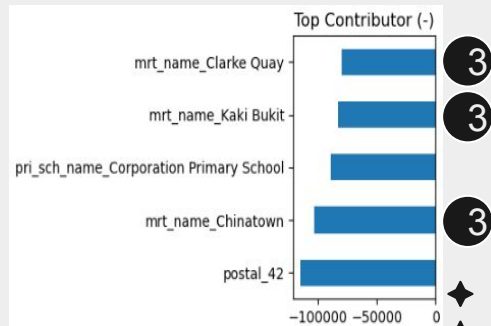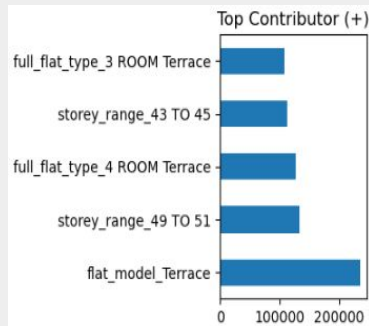
Best Model

Ridge Regression

Most Accurate Model with **~91% accuracy** (R2)

Model able to predict with **~$41k error**

Demonstrate repeatability with **0.1% difference**

Top Contributing Features

1. Flat Model/Type
2. Flat Storey
3. Nearest MRT

Top Contributor (+)

- full_flat_type_3 ROOM Terrace — 1
- storey_range_43 TO 45 — 2
- full_flat_type_4 ROOM Terrace — 1
- storey_range_49 TO 51 — 2
- flat_model_Terrace — 1

0   100000   200000

Top Contributor (-)

- mrt_name_Clarke Quay — 3
- mrt_name_Kaki Bukit — 3
- pri_sch_name_Corporation Primary School
- mrt_name_Chinatown — 3
- postal_42

−100000   −50000   0

Future Improvement

Consider Time Series Testing

- As seen in EDA section, data span across period where price fluctuated.

- Potential time dependency that could skew the conclusion.

- Train data could be based on past and test on the latest time-frame.

Consider More Features

- Economic Indicators (e.g. interest rate, GDP)

- Updated Policies (LTV rate, ABSD rate, MOP*)

- HDB Condition (renovated block)

*MOP (Minimum Occupancy Period) for different locations

Thank You!