PROJECT 2:
YOUR HOUSE, YOUR FUTURE: MAKING INFORMED REAL ESTATE DECISIONS

TEAM: DATA-NINE-NINE
Ming Fatt, Jasmine, Jin Jin, Wen Xi, Willson

DID YOU KNOW?

Expected Prices Increase of HDB

2%- 8% for year 2023

Whereas Price Hike of HDB

10.4% for year 2021

12.7% for year 2022

Source: https://www.channelnewsasia.com/singapore/cooling-measures-singapore-hdb-resale-prices-towns-property-map-3499961#:~:text=Analysts%20expect%20a%20one%2Ddigit,12.7%20per%20cent%20in%202021.

Problem Statement

The general public may not be well-equipped with the information needed to aid their **Real Estate** decision making process. Some of the common questions they might have are

(1) What are the available options given my current budget?

(2) Which flat type and where can I afford?

(3) What price should I set when I sell my flat?

(4) How to market my flat to increase its selling price?

General public

FRET NOT!

REAL ESTATE START-UP COMPANY, DATA NINE-NINE IS HERE TO HELP!

With our state-of-the-art data driven HDB resale price prediction model, your real estate issues shall be a thing of the past.

FLOW OF MODEL BUILDING PROCESS

Understanding the model building process

OUR PROCESS IS EASY

Define the problem

- Identify market need
- Serve the need
- Through **Automated valuation process**

Gather data

- Gather the necessary raw data
- Data cleaning works
- Feature Engineering

Explore data

- Study correlation between features
- Verify reliability of correlation
- **Select features** for our model

OUR PROCESS IS EASY

Produce a model with the data

- Construct Model
- **Linear, Ridge, and Lasso Regression models**
- Optimise models

Evaluate the model

- Evaluate models with Cross Validation, RMSE and R Squared
- Best model among the **9** are deployed

Providing recommendation to the problem with the model

- Made **customised Predictions**
- Data-Driven Recommendations will be provided

EXPLORATORY DATA ANALYSIS PROCESS

Data Janitorial Work

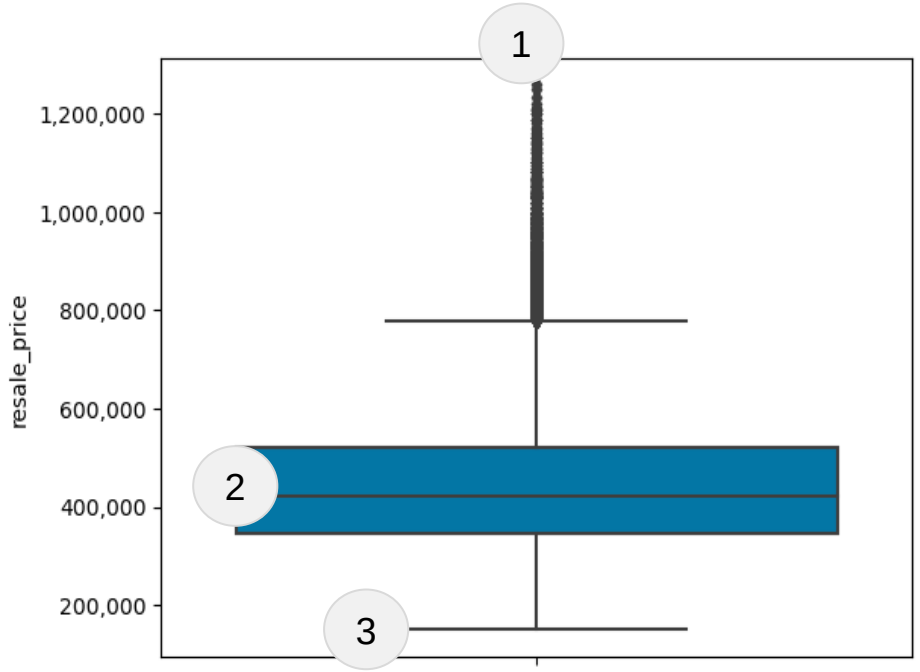
Datasets used contained

150,634
HDB flat resale transaction

77
Features

Mar ‘12 – Apr ‘21
Duration of dataset

Overview of sale pricing



1

Most expensive sale

$1,258,000

5 room flat in Central Area (in 2020)

2

Avg sale price

$449,162

3

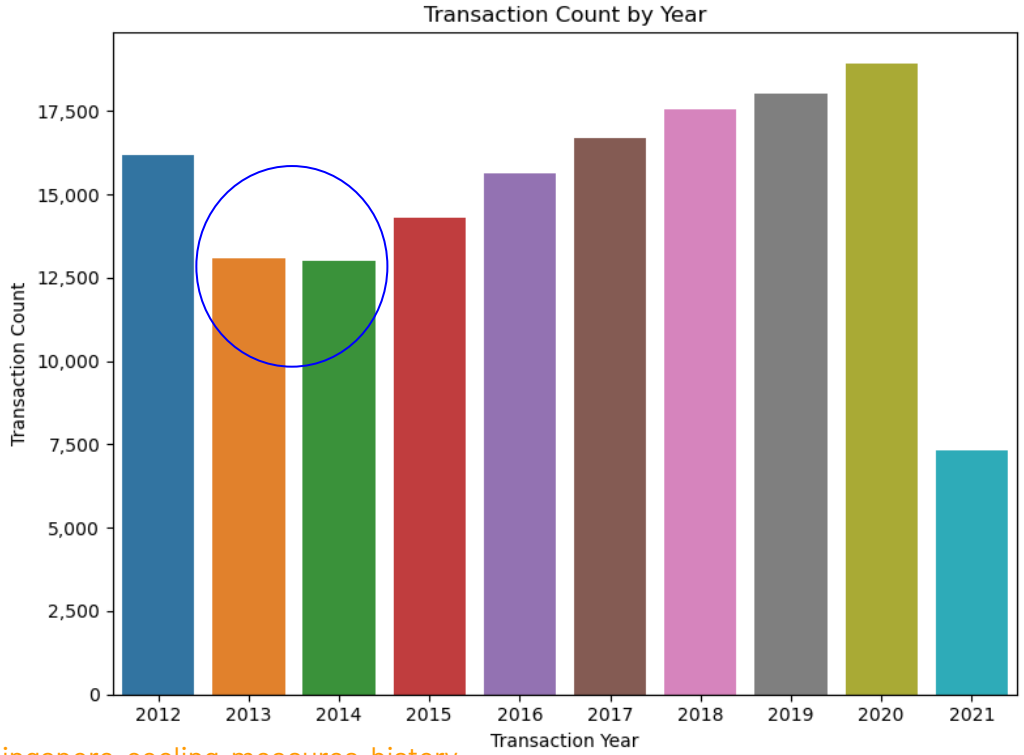
Cheapest sale

$150,000

2 room flat in Toa Payoh (2020) & Geylang (in 2019)

Exploratory Data Analysis

- The dip of transaction count of HDB resale flats in the year of 2013 & 2014 is noted, and may be related to cooling measures introduced in 2013*.
- 2021 data is only up till Apr; if pro-rated for 2021 (entire year) it is on track to be higher than in 2020



* https://stackedhomes.com/editorial/singapore-cooling-measures-history

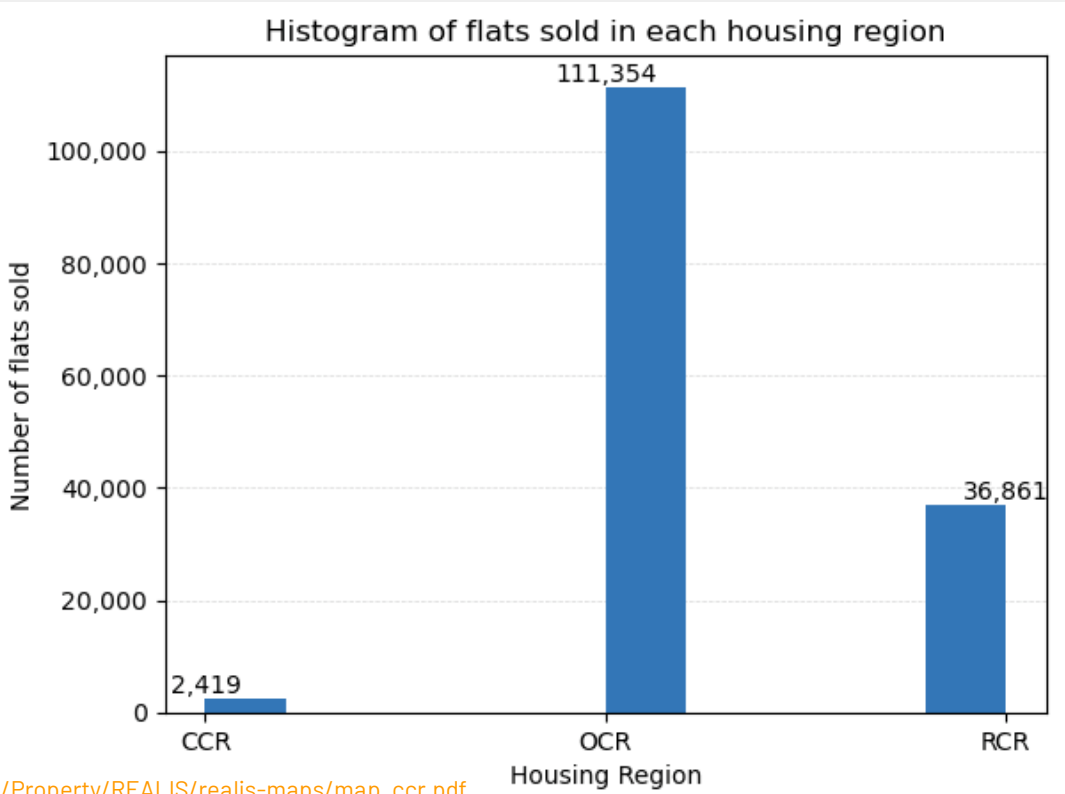
Exploratory Data Analysis

Legend*:

CCR = Core Central Region

RCR = Rest of Central Region

OCR = Outside Central Region



Most houses sold were in the OCR

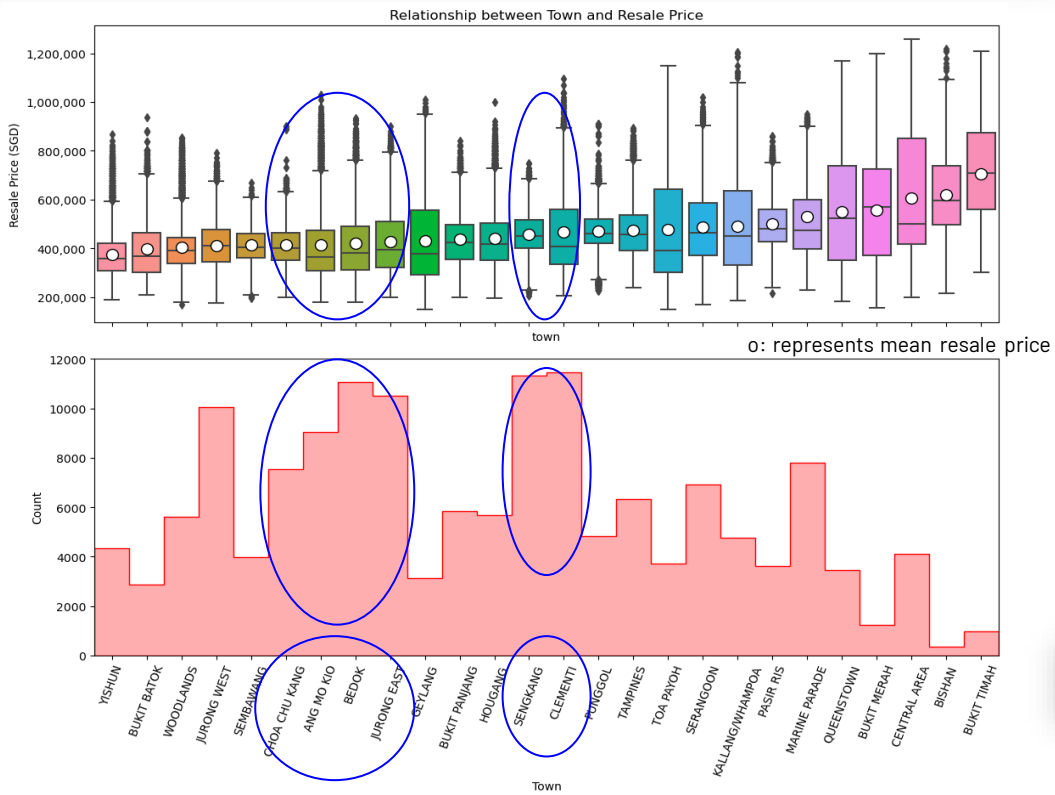
(Outside Central Region)

https://www.ura.gov.sg/-/media/Corporate/Property/REALIS/realis-maps/map_ccr.pdf

Exploratory Data Analysis

Most houses sold from 2012 to 2021 are in the OCR region, not so much within central region

Towns with the more expensive resale prices had the lowest number of transactions



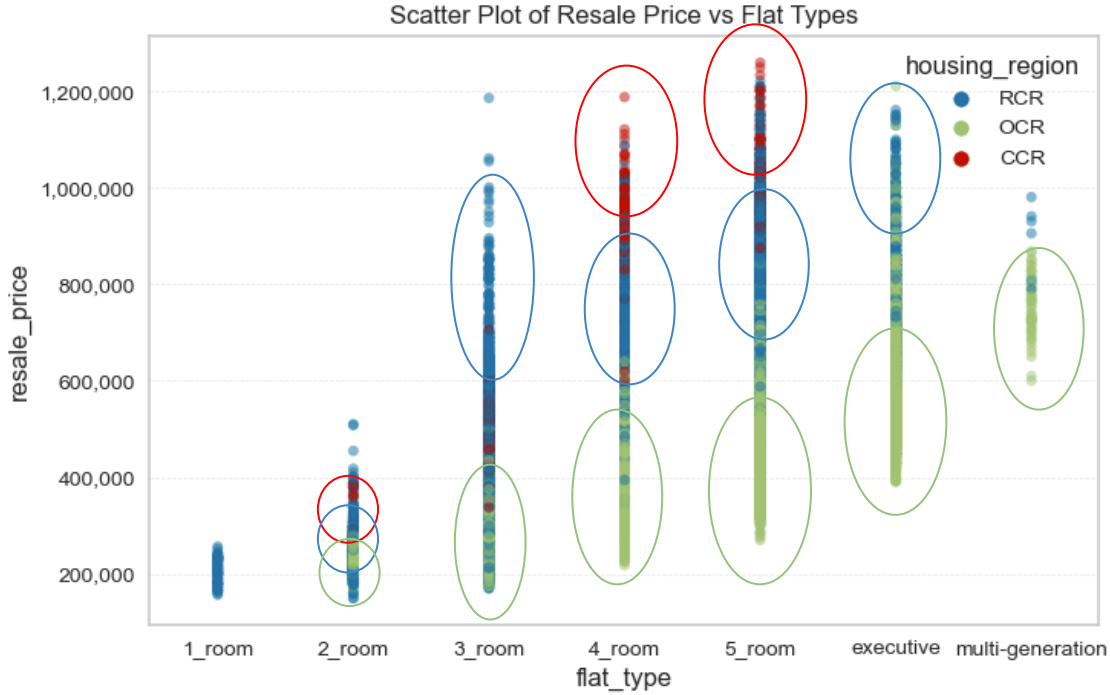
Exploratory Data Analysis

Legend:

CCR = Core Central Region

OCR = Outside Central Region

RCR = Rest of Central Region



CCR: Consistently more expensive

RCR: Middleground

OCR: Generally the least expensive

Exploratory Data Analysis

Low Floor:

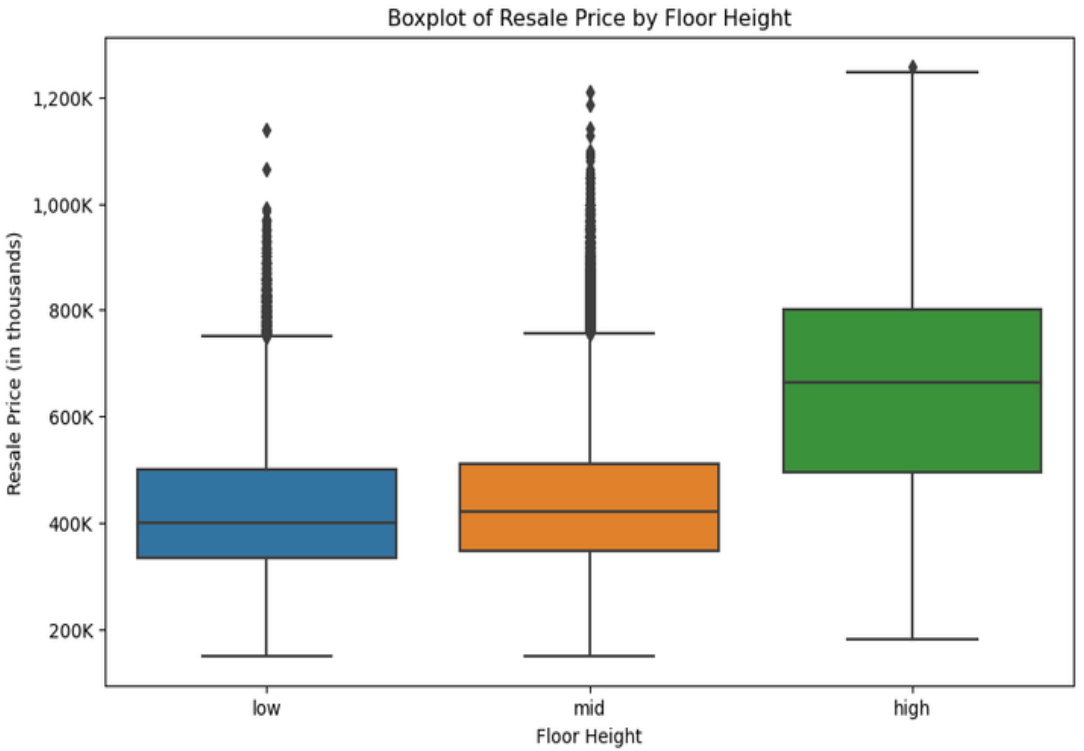
- Below 6 storey
- Less than ⅔ of low rise building
- Less than ⅓ of mid rise building

Mid Floor :

- Between storey 7 - 18
- Above ⅓ of mid rise building
- Above ⅔ of low rise building

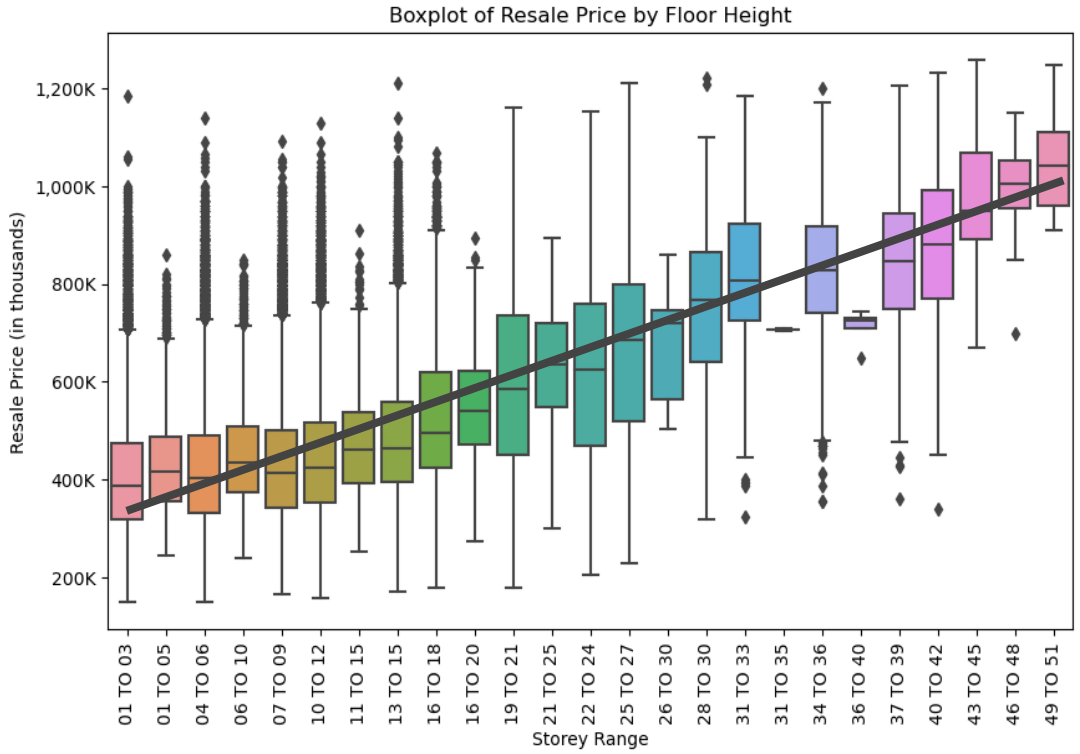
High Floor:

- Above storey 18



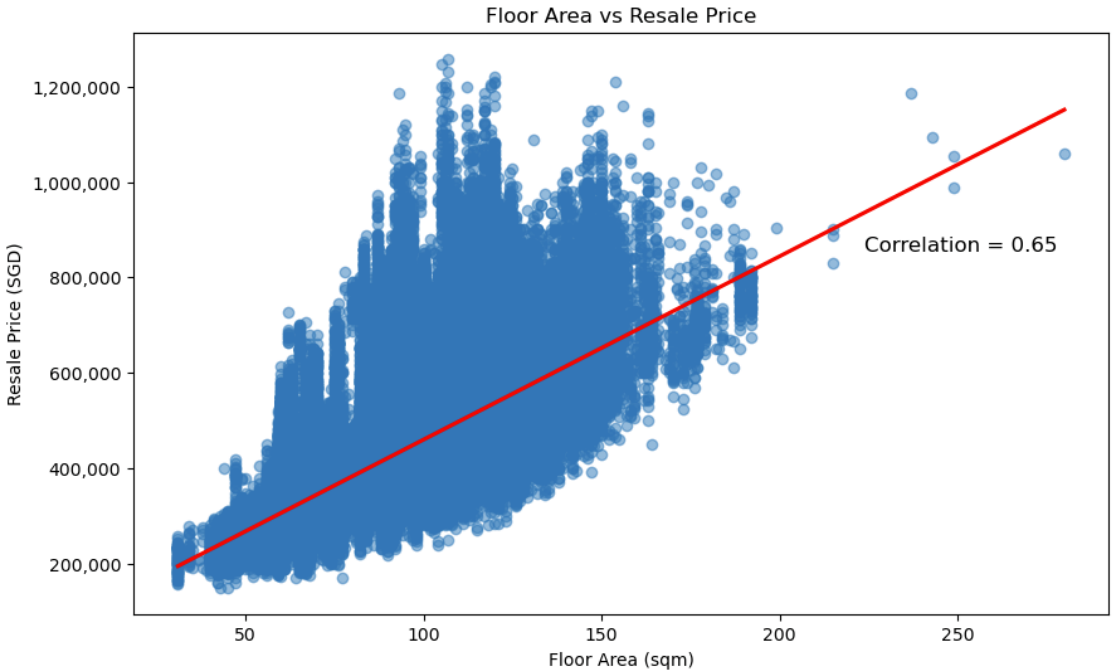
Exploratory Data Analysis

There's a general uptrend on the resale price as the storey range goes up



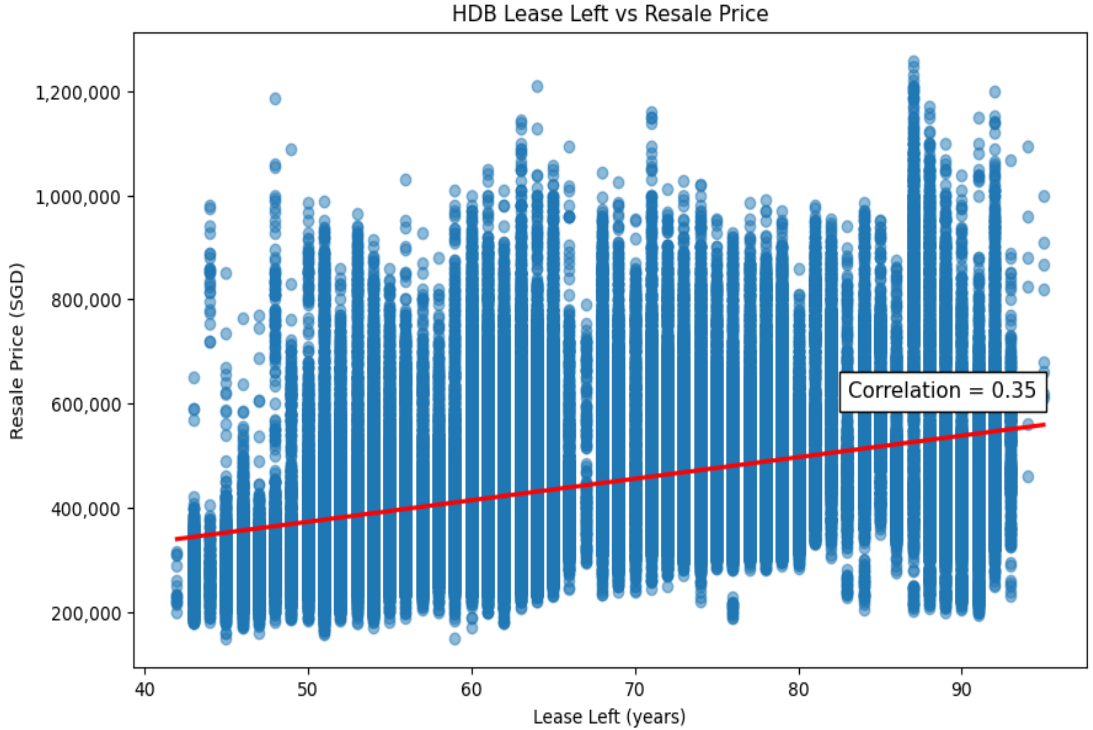
Exploratory Data Analysis

Floor Area of the HDB unit seems to have a strong positive correlation to the resale price of the HDB unit.



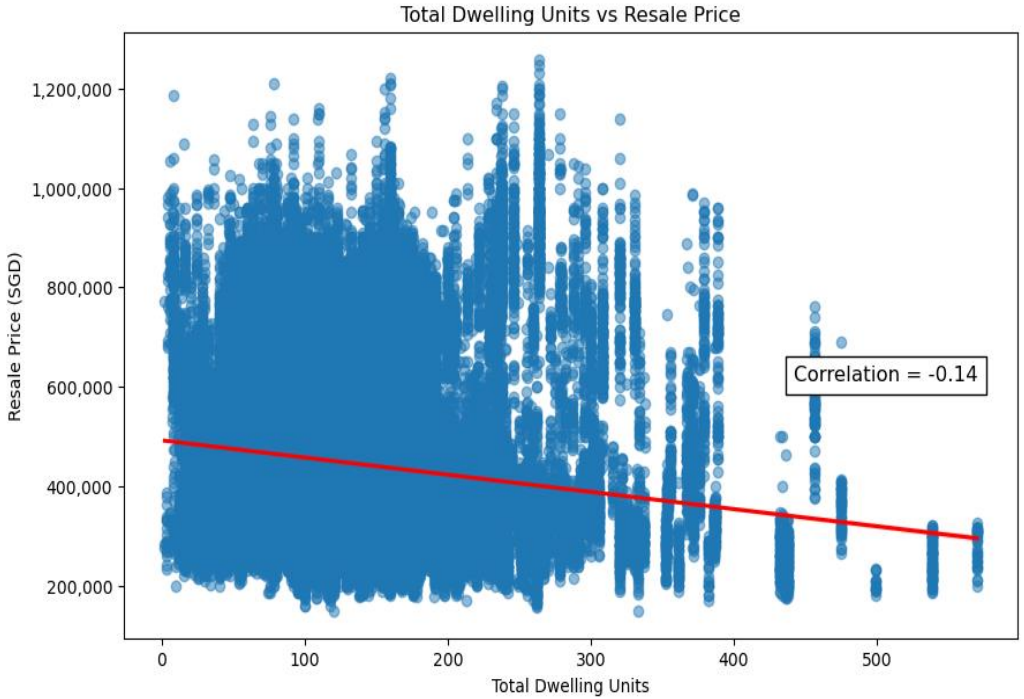
Exploratory Data Analysis

The more years left in the lease of the HDB unit is correlated to how high of a resale price the HDB unit is able to fetch.



Exploratory Data Analysis

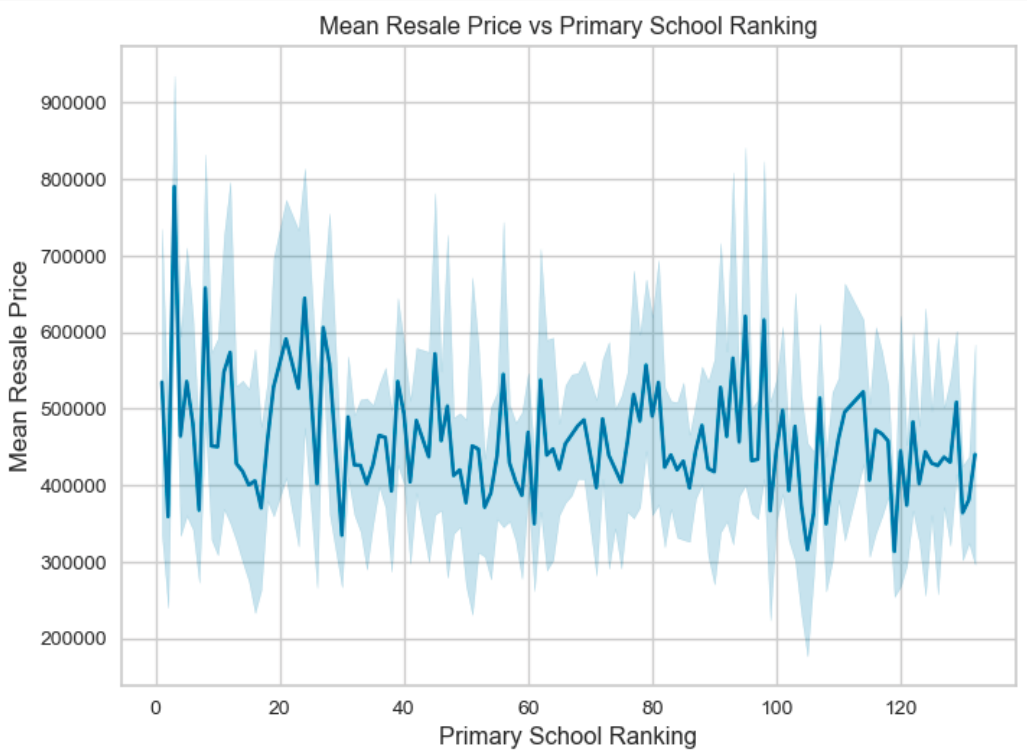
Total dwelling units in a HDB block is not observed to have a strong correlation with the resale price.



Exploratory Data Analysis

Nearby primary school ranking does not seem to have an obvious influence over the resale price of a HDB unit.

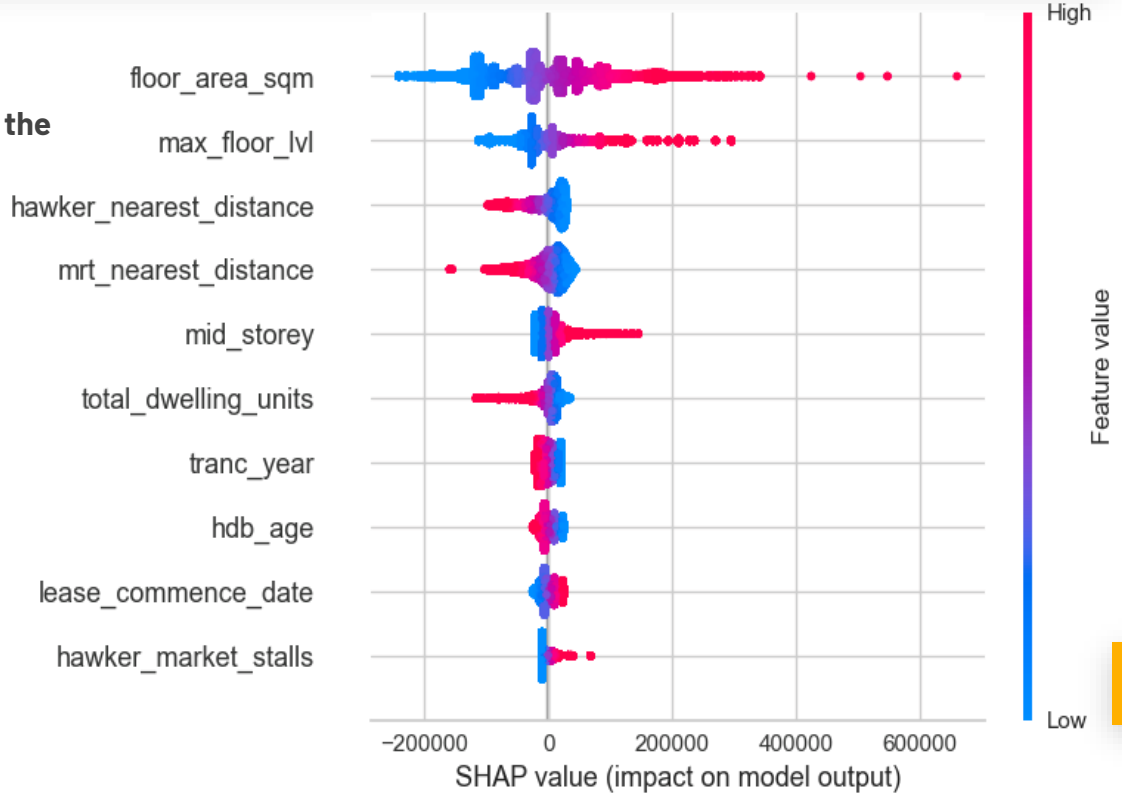
This observation is consistent with the findings of other sources.



Source: https://dollarsandsense.sg/hdb-property-prices-near-popular-primary-schools-really-cost/

Exploratory Data Analysis

Features that were found to have the greatest impact on the model output .



LINEAR REGRESSION MODEL

Building a “linear guideline” for making predictions of future prices of HDB units.

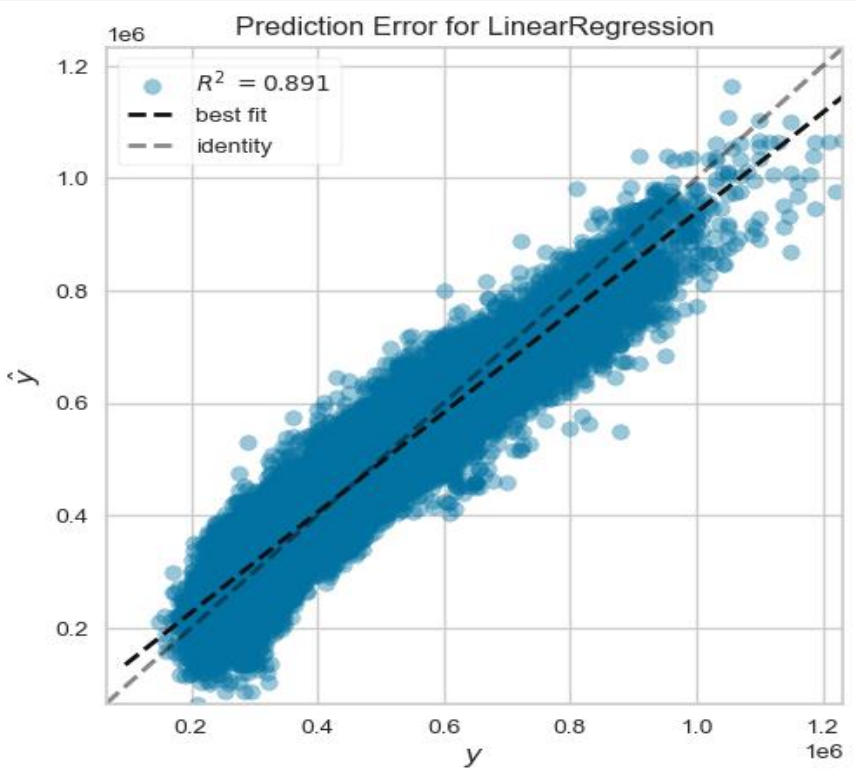
Regression Model Results

Model 1 performed better on unseen data set

Model	R^2 (train)	R^2 (test)	RMSE	RMSE (Unseen data, kaggle)
Baseline	0.8638	0.8610	52,965.62	-
1	0.8911	0.8904	47,162.25	47,149.78
2	0.9147	0.9137	41,837.78	54,155.30
3	0.8927	0.8920	46,806.20	55,781.83

Linear Regression Model

Prediction error plot shows the actual targets from the dataset against the predicted values generated by our model



Linear Regression: Feature Selection

Model	Feature Selection Description
Baseline	• Baseline model runs with all numeric features • Used as a baseline to evaluate model performance
1	• Feature selection based on domain knowledge • Elements that are known to affect housing prices
2	• The features selection are based on features correlation • Feature engineering of region against flat types • Popularity ranking of primary schools • Availability of amenities
3	• Feature selection based on model 1 features and • Feature importance from previous models.

PRODUCT DEMONSTRATION

Live demonstration of your property valuation

RECOMMENDATION & KEY INSIGHTS

So what's the gist of it?

Recommendations

Buyer

- Know your available options given your budget
  - Buyers should be able to make an informed decision that fits their budget.
- Prioritize and personalize your wants
  - Buyers are recommended to straighten out their priority and decide on the factors that fits their needs most.
- Quality home with comfortable price

Seller

- Appraise your property value based on market valuation
  - Sellers are recommended to at least have some understanding of the market resale price of their respective units.
- Pivot your selling strategies
  - Sellers are recommended to be ready to switch up with their selling strategies at any given time due to market volatility.
- Match your property's unique selling points to the right buyers

1 Minute

To have an estimated price for your dream house

89.1%

Prediction accuracy

11 X-Factors

Focus on the factors that matters

KEY LIMITATIONS

If only we had more time and resource.

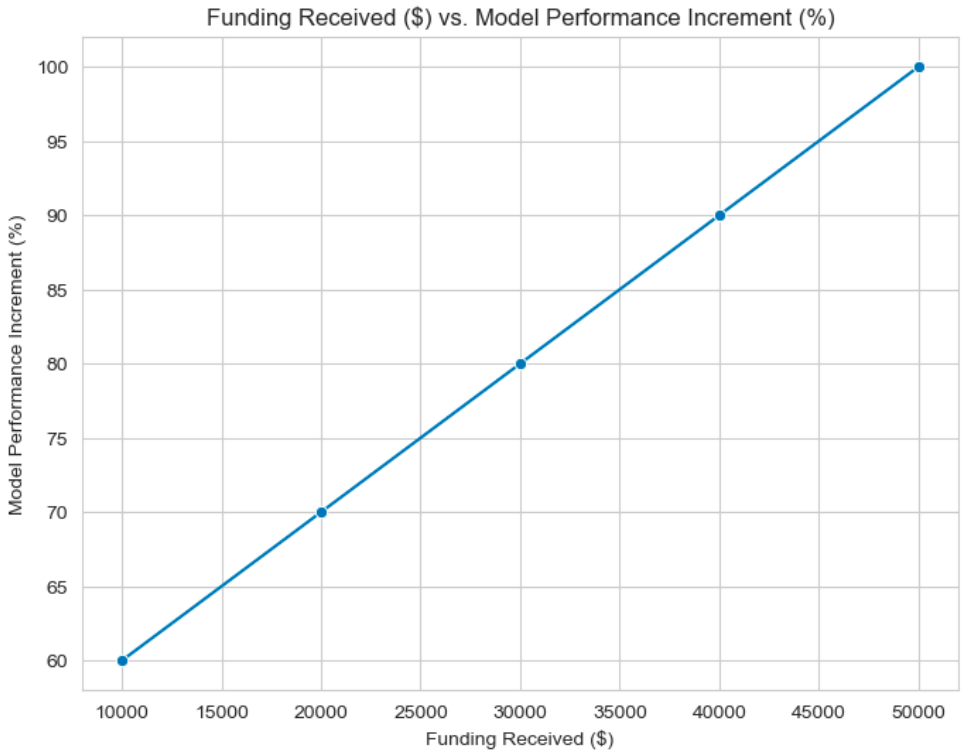
Key Limitations

- Limitation of data
  - Data collected is only up to 2021
- Modelling is limited to only Linear Regression
  - Opportunity to utilize more sophisticated model in the future for better prediction.
- Lack of info on the existing condition of the sold units
- Collinearity does not imply causation

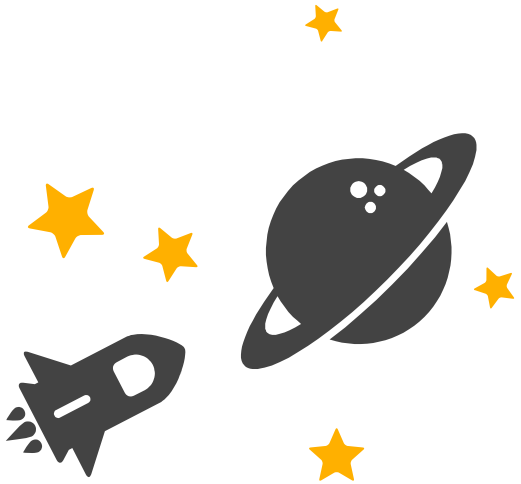
Key Limitations (Joke)

Funding

Severe lack of funding is making it challenging to carry out further refinement works to increase the model performance.

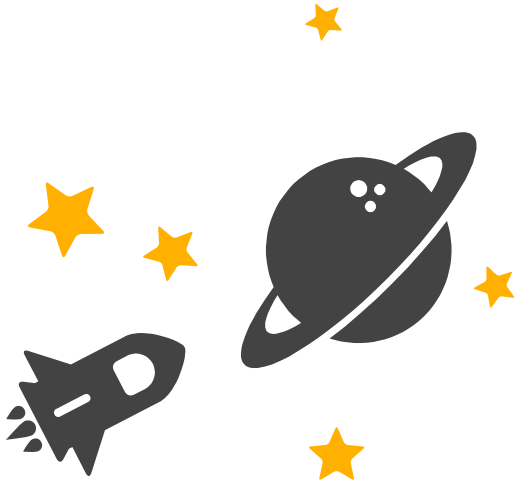


Your House, Your Future



Make your real estate plans with technology of future

THANK YOU!



BACK-UP SLIDES / ANNEX

Extra cheese

Linear Regression Model

Model 1 Features:

Target (y-axis): resale_price

X(axis):

1. `town` (cat)
2. `storey_range` (cat)
3. `full_flat_type` (cat)
4. `pri_sch_name` (cat)
5. floor_area_sqm
6. lease_commence_date
7. mrt_nearest_distance
8. hawker_nearest_distance
9. mall_nearest_distance
10. pri_sch_nearest_distance
11. sec_sch_nearest_dist

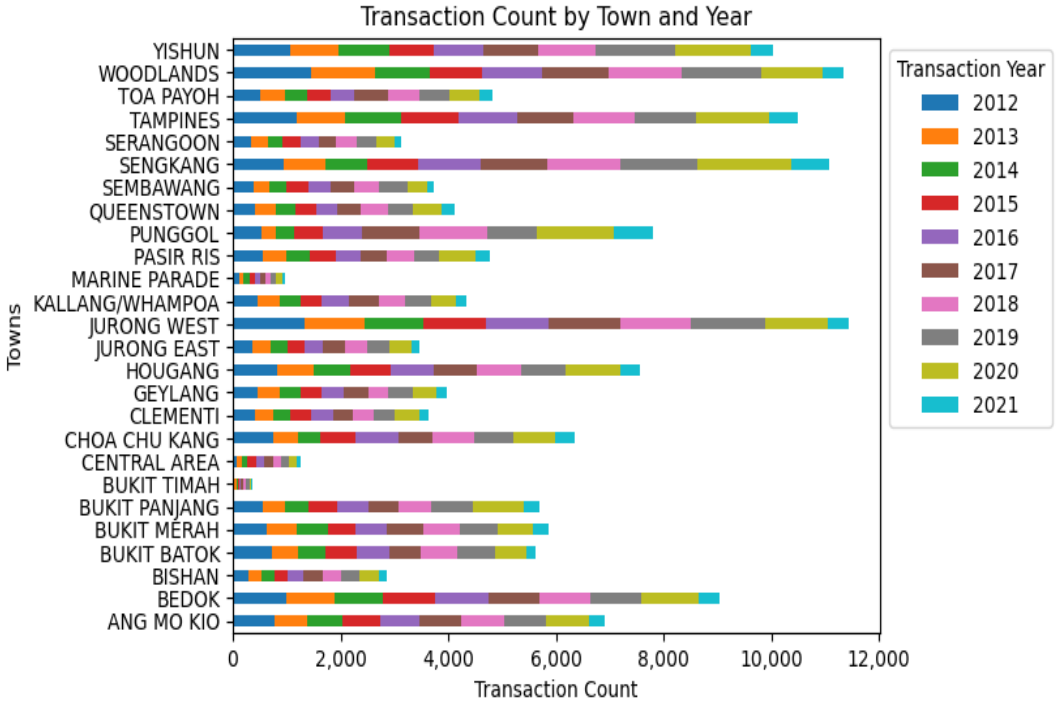
Features

Train.csv features: 78

Dropped: id, price_per_sqft, floor_area_sqft, resale_price (which is target)

Added: pop_ranking, pop_ranking_2cat, housing_region

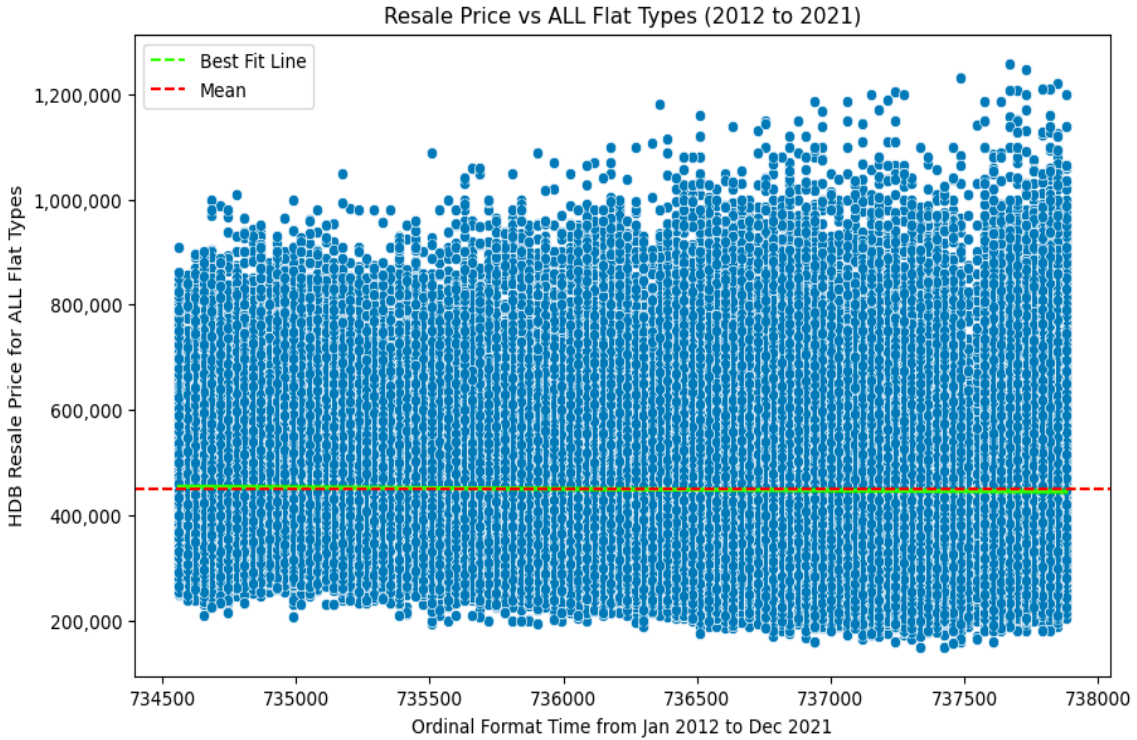
Exploratory Data Analysis



Exploratory Data Analysis

Note:

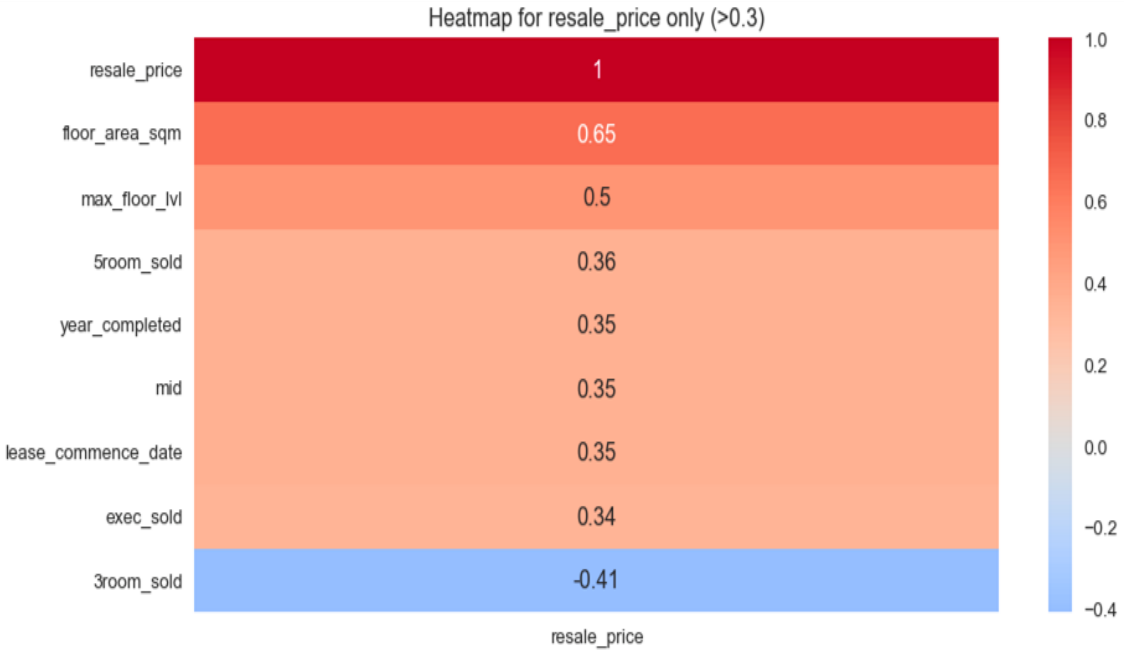
Progression of years is found to not have much effect on the resale prices of HDB units



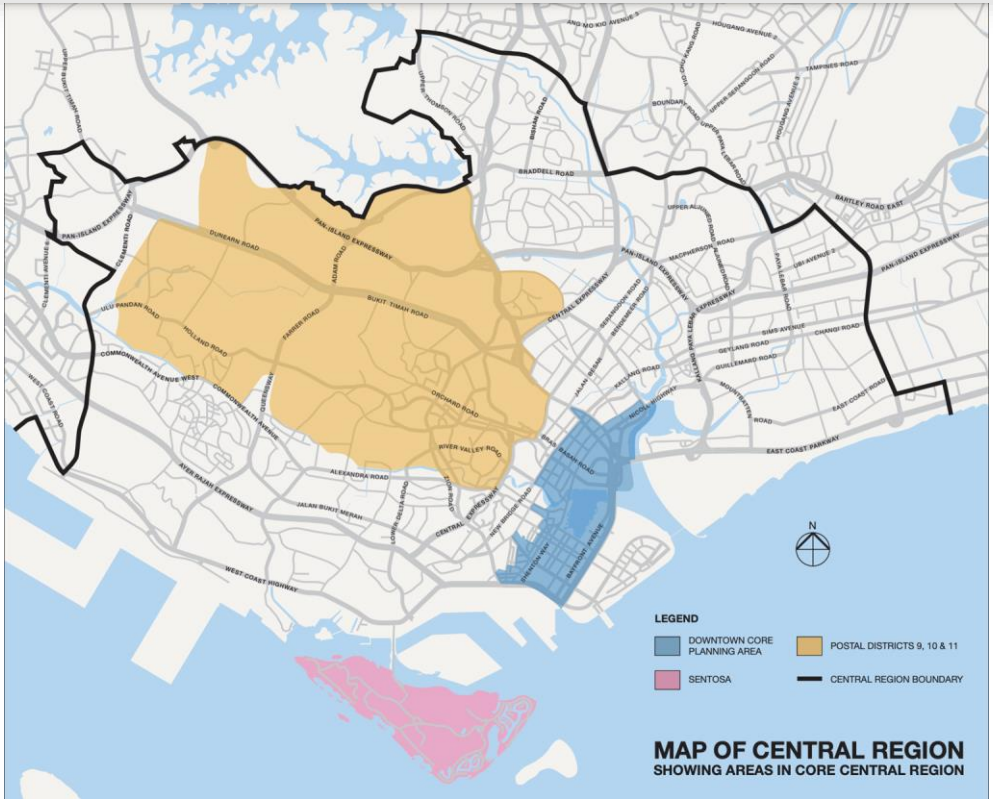
Exploratory Data Analysis

Note:

Heatmap of correlation coefficient of features to resale_price.



Housing Regions: CCR, RCR, OCR



URA Map from:

https://www.ura.gov.sg/-/media/Corporate/Property/REALIS/realis-maps/map_ccr.pdf

Postal Sector

- https://www.mingproperty.sg/singapore-district-code/

Regression Model Results

Linear Regression Model	R^2 (train)	R^2 (test)	RMSE	RMSE (Unseen data, kaggle)
Baseline	**0.8638**	**0.8617**	**52,965.62**	
1	**0.8911**	**0.8904**	**47,162.25**	**47,149.78**
2	**0.9147**	**0.9137**	**41,837.78**	**47,509.99**
3	**0.8917**	**0.8920**	**46,806.20**	**55,781.83**

Regression Model Results

Ridge Regression Model	R^2 (train)	R^2 (test)	RMSE	RMSE (Unseen data, kaggle)
Baseline	**0.8638**	**0.8617**	**52,965.61**	
1	**0.8911**	**0.8903**	**47,163.40**	**47,149.78**
2	**0.9146**	**0.9136**	**41,854.66**	**47,509.99**
3	**0.8927**	**0.8920**	**46,805.96**	**55,781.83**

Regression Model Results

Lasso Regression Model	R^2 (train)	R^2 (test)	RMSE	RMSE (Unseen data, kaggle)
Baseline	**0.8631**	**0.8610**	**53,105.47**	
1	**0.5761**	**0.5751**	**92,861.14**	**47,149.78**
2	**0.7071**	**0.7046**	**77,422.44**	**47,509.99**
3	**0.8920**	**0.8914**	**46,927.46**	**55,781.83**

Linear Regression Model

Note:

Model 2