Task 1
Predicting HDB Resale Prices

Table of Contents

1 – Aim & Data of Analysis

2 – Methodology

3 - Exploratory Data Analysis

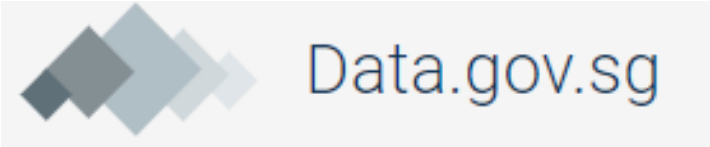
4 – Model Development Summary

5 – Future Work

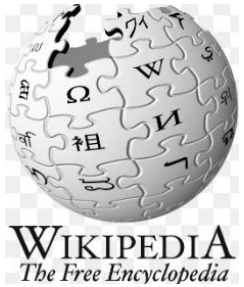
Aim & Data

Aim: This deck outlines the findings from the analysis of building a predictive algorithm to determine the factors affecting prices of residential properties in Singapore.

Data:



117,527 transactions of HDB resale flat prices from Jan 2015 to Dec 2020



To extract the list of the Public Health Institutions (i.e., Hospitals & Polyclinics)

To extract the coordinates of the HDB Resale Flats

To extract the list of Shopping Malls in Singapore

Methodology

Step 1: Data Preprocessing

Transforming columns in the HDB Resale Dataset to provide more meaningful insights to the model.

(a) Handling remaining lease

Remaining lease was in years and months which is hard for the prediction model to interpret. We will convert to years.

remaining_lease	remaining_lease_temp
61 years 04 months	61
60 years 07 months	60
62 years 05 months	62
62 years 01 month	62
62 years 05 months	62

(b) Removing transaction with remaining lease above 95 years

In Singapore's context, there is a minimum occupancy period (MOP) of 5 years before one can sell their HDB unless there are exceptional circumstances.

month	town	flat_type	block	street_name	remaining_lease
2015-01	JURONG WEST	3 ROOM	339A	KANG CHING RD	96
2015-01	KALLANG/WHAMPOA	4 ROOM	38D	BENDEMEER RD	96
2015-01	KALLANG/WHAMPOA	4 ROOM	5A	UPP BOON KENG RD	96

(c) Extracting median storey

Storey range in the dataset was a range which we extracted the median storey from.

storey_range	storey_median
10 TO 12	11.0
01 TO 03	2.0

Step 2: Feature Extraction

Latitudes & Longitudes of the HDB blocks, and distances to CBD, nearest healthcare facilities, and shopping malls were obtained OneMap API.

(a) Geolocation of HDB blocks

In the dataset, only the block number and street name were provided. In order to calculate the distance to the nearest expressway, latitudes & longitudes have to be obtained via OneMap API.

block	street_name
406	ANG MO KIO AVE 10
108	ANG MO KIO AVE 4

→ OneMap API →

Latitude	Longitude
1.368949	103.856787
1.377019	103.839552

(b) Geolocation of CBD, healthcare instituitions, shopping malls

They were obtained via WIkiPedia and Kaggle.

(c) Proximity of HDB block to facilities

For each of the HDB block, the haversine distance was calculated against the distance to CBD, nearest shopping mall, and healthcare facilities based on their geolocation coordinates.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

Where: r = radius of the earth ($\approx$ 6366 km)
φ_1, φ_2 = latitude of points 1 and 2, in radians
λ_1, λ_2 = longitude of points 1 and 2, in radians

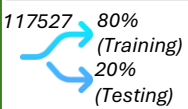
Note that φ and λ can be converted from degrees to radians by multiplying by $\pi/180$

Step 3: Model Training

A total of 3 models were built with the aim to predict the resale price and identify the variables that are of high importance.

(a) Data Preparation

Categorical variables have to be one-encoded as regression models cannot handle categorical representation. Numerical variables have to be normalize so that their scales don't disproportionately influence the model.



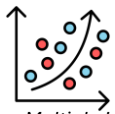
flat_type	town_BEDOK	town_BISHAN
2.0	0.0	0.0
2.0	0.0	0.0

One-hot encoding

(b) Model Training

3 models were built with resale price as the target variable:

1. Multiple Linear Regression
2. Random Forest
3. LightGBM



Multiple Linear Regression

Random Forest

LightGBM

Step 4: Model Evaluation & Recommendations

Models were evaluated based on RMSE and R^2, and important variables were identified.

(a) Model Evaluation

The 3 models were evaluated based on RMSE and R^2, with the lowest RMSE and highest R^2 as the best performing model.

Model	RMSE	R^2
Model 1 - Multiple Linear Regression	52445.64	0.8736
Model 2 - Random Forest	**25648.67**	**0.9698**
Model 3 - LightGBM	31334.87	0.9549

(b) Features Importance

For model 1, the coefficients of the features were analysed. For model 2 and 3, the feature importance was analysed to identify top few data variables important in predicting resale price.

	coef	std err	t	P>|t|
const	-1.282e+05	3.74e+04	-3.426	0.001
flat_type	1.522e+04	852.769	17.842	0.000
town_BEDOK	563.2175	1109.710	0.508	0.612
town_BISHAN	8.425e+04	1620.406	51.991	0.000
town_BUKIT BATOK	-3.469e+04	1478.148	-23.471	0.000

Output of Regression Model

	Feature	Importance
0	town	655
5	dist_to_cbd	500
4	remaining_lease	393
2	floor area sqm	382

Feature Importance Ranking

Exploratory Data Analysis

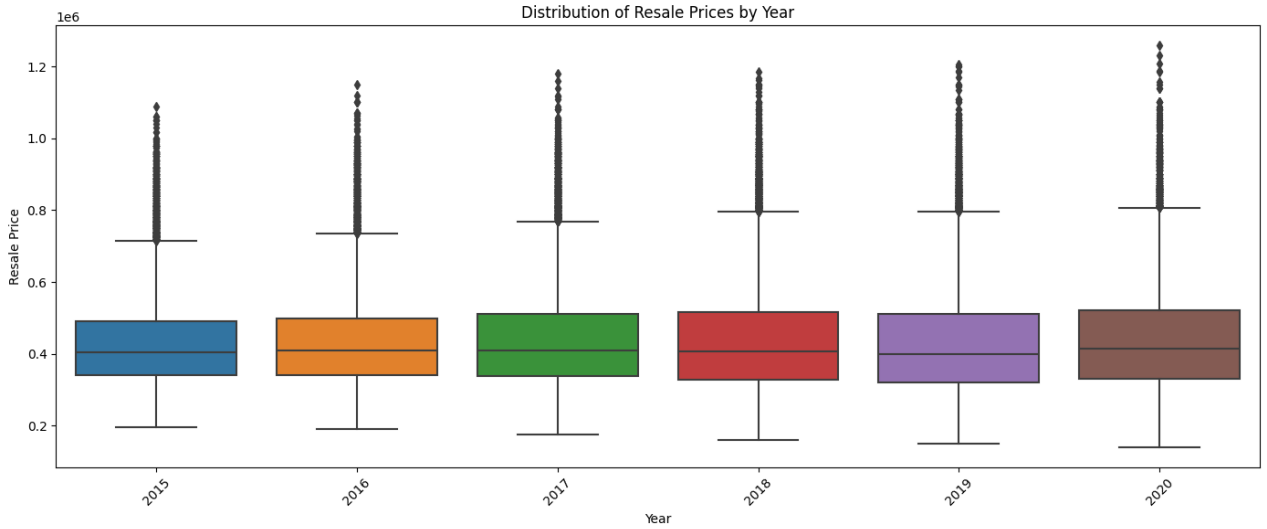


Figure 1 – Box Plot of Resale Price Against Transaction Year

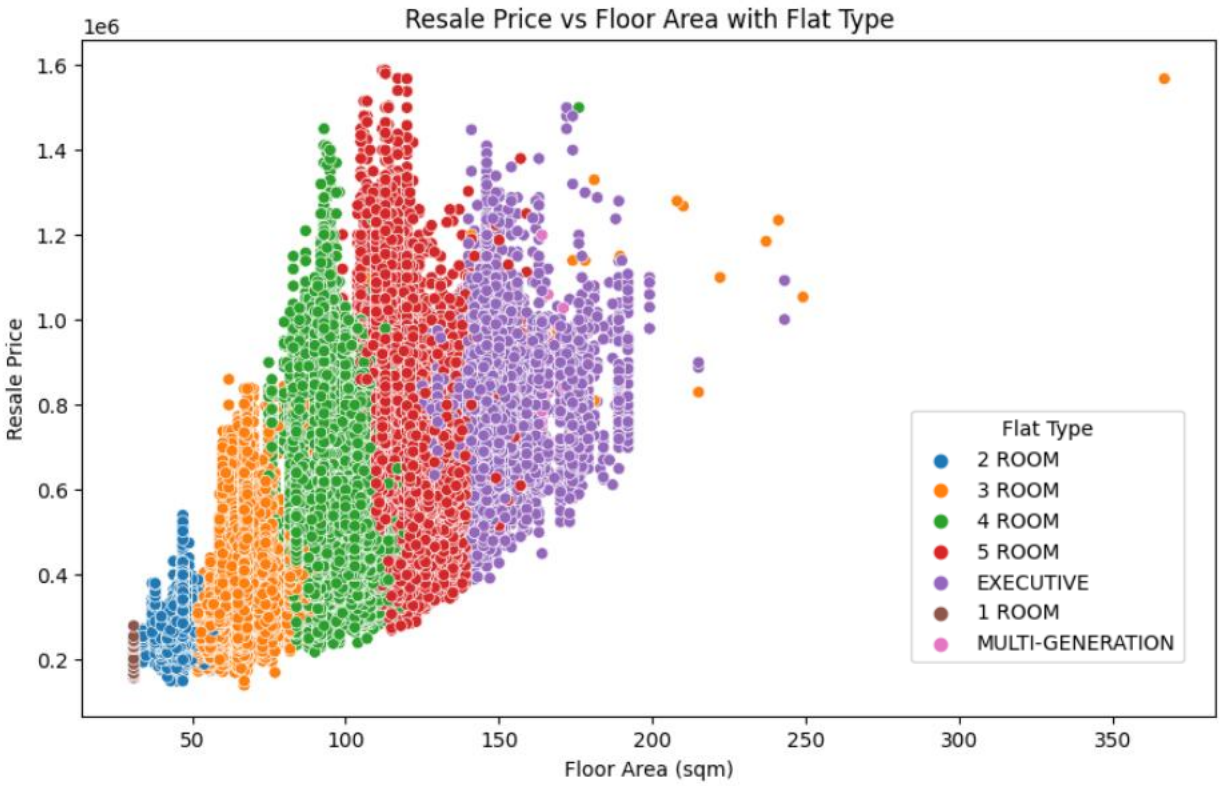


Figure 2 – Scatter Plot of Resale Price Against Floor Area & Flat Type

- Figure 1 suggests the resale prices seem to be **generally consistent throughout the years**, with **more resale flats reaching above $1million dollars in the later years**.
- Figure 2 indicates that **other factors such as floor area and flat type also influence resale prices**, with larger flats generally commanding higher prices.

Model Development Summary

Model	RMSE	R^2	Impt Variable 1	Impt Variable 2	Impt Variable 3
Model 1 - Multiple Linear Regression	52445.64	0.8736	flat model	town	nearest_shopping_distance
Model 2 - Random Forest	**25648.67**	**0.9699**	floor_area_sqm	dist_to_cbd	remaining_lease
Model 3 - LightGBM	31334.87	0.9549	town	dist_to_cbd	remaining_lease

Figure 1 – Model Summary of the 3 models

- A total of 3 models were developed with the aim of predicting Resale Price:
  1. Multiple Linear Regression – Linear Model
  2. Random Forest – Non-Linear Model
  3. LightGBM – Gradient Boosting Model

- **Model 2 - Random Forest** achieved the best performance with the highest *R^2 (0.9699) and the lowest ^RMSE.

Feature	Coefficient
flat_model_Terrace	330213.484473
town_BUKIT TIMAH	240720.136312
flat_model_Type S2	198134.167596
flat_model_Improved-Maisonette	174203.444412
flat_model_DBSS	136447.402058
flat_model_Type S1	132854.392173
flat_model_Premium Apartment Loft	122834.320721
town_MARINE PARADE	118063.671511
flat_model_Model A-Maisonette	116480.652882
flat_model_Multi Generation	111556.104004

Figure 2 – Top 10 Feature Importance of Model 1

Feature	Importance
floor_area_sqm	0.457480
dist_to_cbd	0.302031
remaining_lease	0.081119
flat_type	0.029473
nearest_shopping_dist	0.024693
storey_median	0.024474
nearest_healthcare_dist	0.022896
flat_model_DBSS	0.015270
flat_model_Model A	0.005514
town_BISHAN	0.002818

Figure 3 – Top 10 Feature Importance of Model 2

Feature	Importance
town	655
dist_to_cbd	500
remaining_lease	393
floor_area_sqm	382
nearest_healthcare_dist	350
nearest_shopping_dist	304
flat_model	195
storey_median	166
flat_type	55

Figure 4 –Feature Importance of Model 3

- Across the 3 models, **floor_area_sqm**, **dist_to_cbd**, **remaining_lease**, and **town** have been identified as top important variables for determining resale price.

- **Larger floor areas** generally correlate with **higher property values**.
- Distance to the Central Business District (CBD) is often a key factor in housing pricing, as **properties closer to the city center tend to be more expensive**.
- Remaining lease is crucial for leasehold properties, as **shorter leases can significantly impact valuation**.
- Location plays a major role in determining the resale price. **Different towns may have distinct pricing trends, amenities, and desirability** especially the Central (i.e., Bukit Timah, Marine Parade) which are a shorter distance away from CBD.

** R^2 represents the proportion of the variance in the target variable that is explained by the features in the model, with a value closer to 1 indicating a better fit.*

^ Mean-Square Error measures the average of the squared differences between the actual and predicted values, indicating how close the predictions are to the real data (lower values are better).

Future Work

Adding on more data variables as HDB resale prices are influenced by multiple factors. The following elements can also play a significant role:

- **Proximity to Public Transport**: Being near MRT stations or bus stops significantly increases convenience and can positively affect resale prices. *[Data can be obtained from Land Transport Authority (LTA)'s website]*

- **Access to Schools**: Living near schools is an important factor for parents to consider, as being within a certain proximity increases the chances of their children securing enrollment. *[Data might be obtainable from OneMap]*

- **Maturity of the Town**: Older, well-established towns like Ang Mo Kio and Toa Payoh tend to have higher resale values due to their infrastructure, history, and established communities. *[Data can be obtained from Singapore Department of Statistics (SingStat)]*

- **Demographic Composition**: The makeup of an estate, including the percentage of working adults, elderly residents, and families with children, can influence housing demand. Different demographics may prioritize certain features like proximity to schools or eldercare facilities, which could affect the value of properties in the area. *[Data can be obtained from Singapore Department of Statistics (SingStat)]*

Task 2
Factors and considerations in building an in-house predictive model for users

Factors & Considerations

- Knowing your stakeholders
  - **Different stakeholders** have **distinct use cases** for predictive models.
  - Using HDB resale prices as an example, **potential new buyers** can benefit from such a tool to **determine whether they are paying a fair market rate or being overcharged**.
  - On the other hand, **HDB planners** can leverage the model to analyse key factors influencing resale prices and **identify adjustments that could help regulate rising costs more effectively**.

- Interpretability of model
  - **Not all** machine learning models are **easily interpretable**.
  - While a Neural Network may produce highly accurate predictions, its complexity can make it **difficult for non-technical stakeholders to understand** or even **take actions on the results**.
  - This lack of understanding can **lead to confusion** or **even misinterpretation of results**, potentially affecting decision-making.
  - Therefore, efforts must be made in **striking a balance between accuracy and interpretability**, to ensure the model is both effective and usable.

- Model Deployment & Maintenance
  - To ensure the model **remains relevant** and **continues to generate accurate results**, it must be **regularly retrained** and fine-tuned as **new data becomes available**.
  - Additionally, **some of the data used for training may be confidential** and should be restricted from unauthorized access to protect privacy and security.

- External Factors
  - Machine Learning Model is not everything as there can be **sudden regulations or unforeseen events** that can affect the predictions
  - For example, HDB resale prices were heavily influenced by the **COVID-19 pandemic** as well as **Government Cooling Measures**