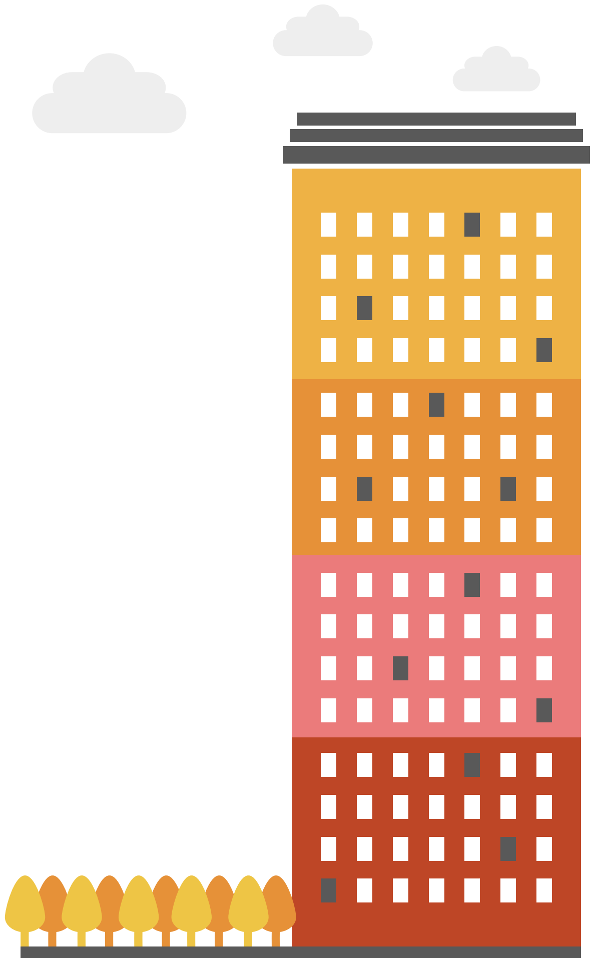
# HDB Resale Price Predictions

Jervin Seow
DSI-29

# Table of Contents





# Background

With more than 1 million flats spread across 24 towns and 3 estates, the Singapore brand of public housing is uniquely different. These flats spell home for over 80% of Singapore's resident population. Singaporeans have two main options when it comes to purchasing a HDB flat; they can either choose to buy a new flat from HDB (BTO) or a resale flat from the open market. Naturally, both types of flats come with their own pros and cons.

In recent years, there are reports stating that demand for resale flats has spiked, resulting in a reactionary increase in resale flat prices. With this increased interest in resale flats, I decided to base my capstone project on them.

# HDB Valuation

- Official valuation can only be obtained partway through (Step 4) a typical sale transaction of a resale flat
- If selling price > official valuation, the difference = Cash on Valuation (COV)
- COV can only be paid in Cash
- Only way to get an estimate of the COV through personal due diligence

# Problem Statement

The aim of this project is to create a model that will give an accurate prediction of the actual valuation that can be utilized by the people involved in a resale flat transaction, namely the buyers, sellers and property agents). This would inform the relevant parties whether a particular flat is undervalued or overvalued, and in turn give a good estimate of its potential COV.

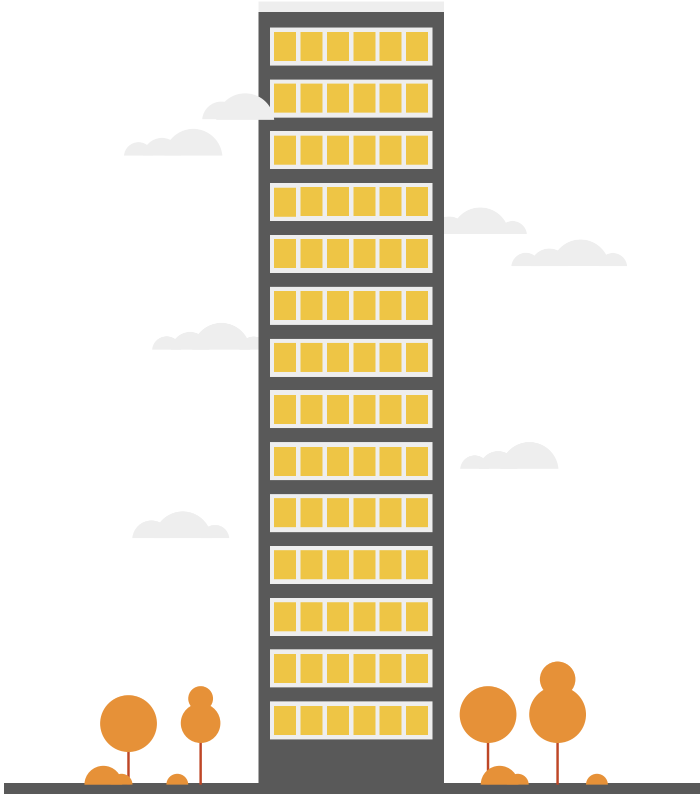
# Data Used

**Main Dataset:**

Combination of datasets downloaded from data.gov.sg . These datasets contain resale flat transaction history from January 2000 to July 2022.

**Supplementary Data used for Feature Engineering:**

1. MRT Locations
2. List of Shopping Malls in Singapore
3. Listing of Licensed Supermarkets
4. Locations of Hawker Centres and Markets (KML file)
5. Locations of Parks (KML file)
6. School Directory and Information

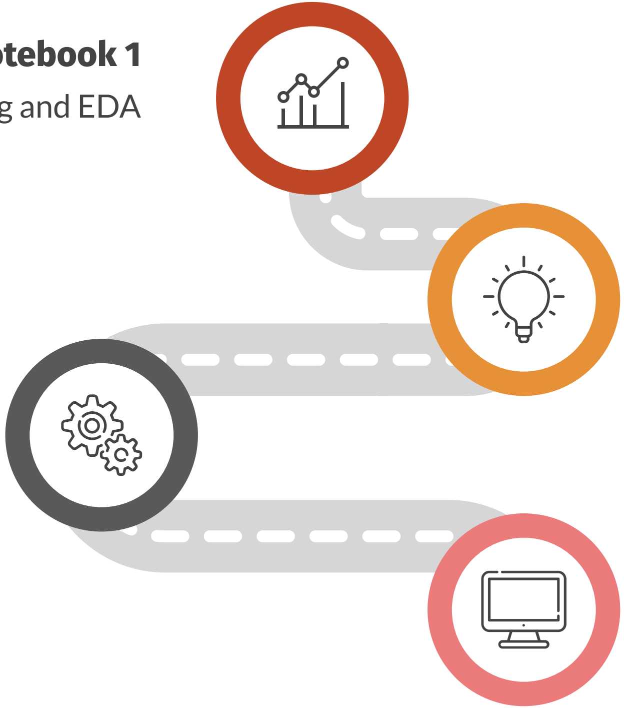


# Features of Main Dataset

- Month and Year of Sale
- Town
- Flat Type
- Flat Model
- Block and Street Name
- Storey Range
- Floor Area (sqm)
- Lease Commencement Date
- Remaining Lease
- Resale Price

# Methodology

**Notebook 1**
Data Cleaning and EDA



**Notebook 2**
Feature Engineering

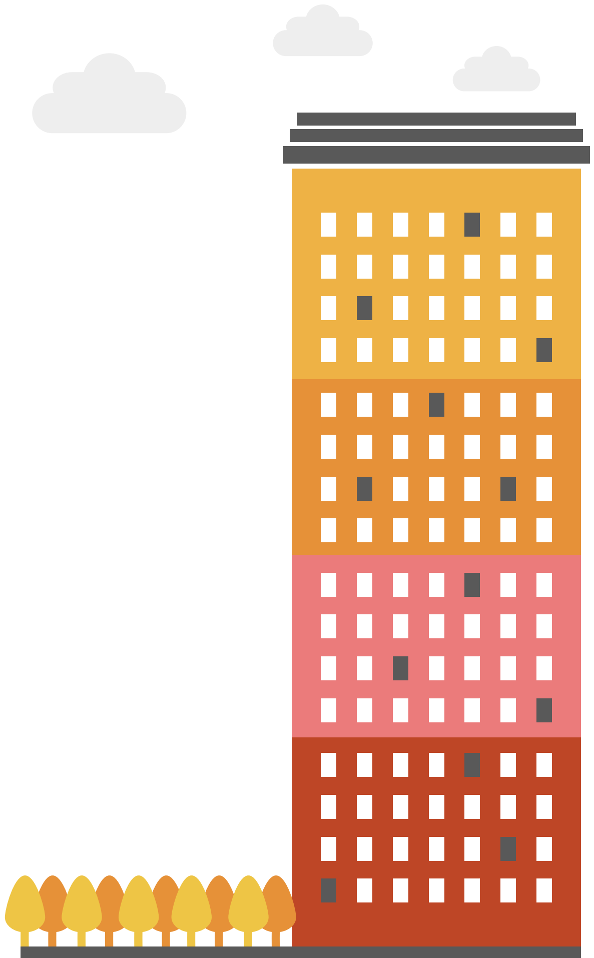
**Notebook 3**
Preprocessing

**Notebook 4**
Modeling

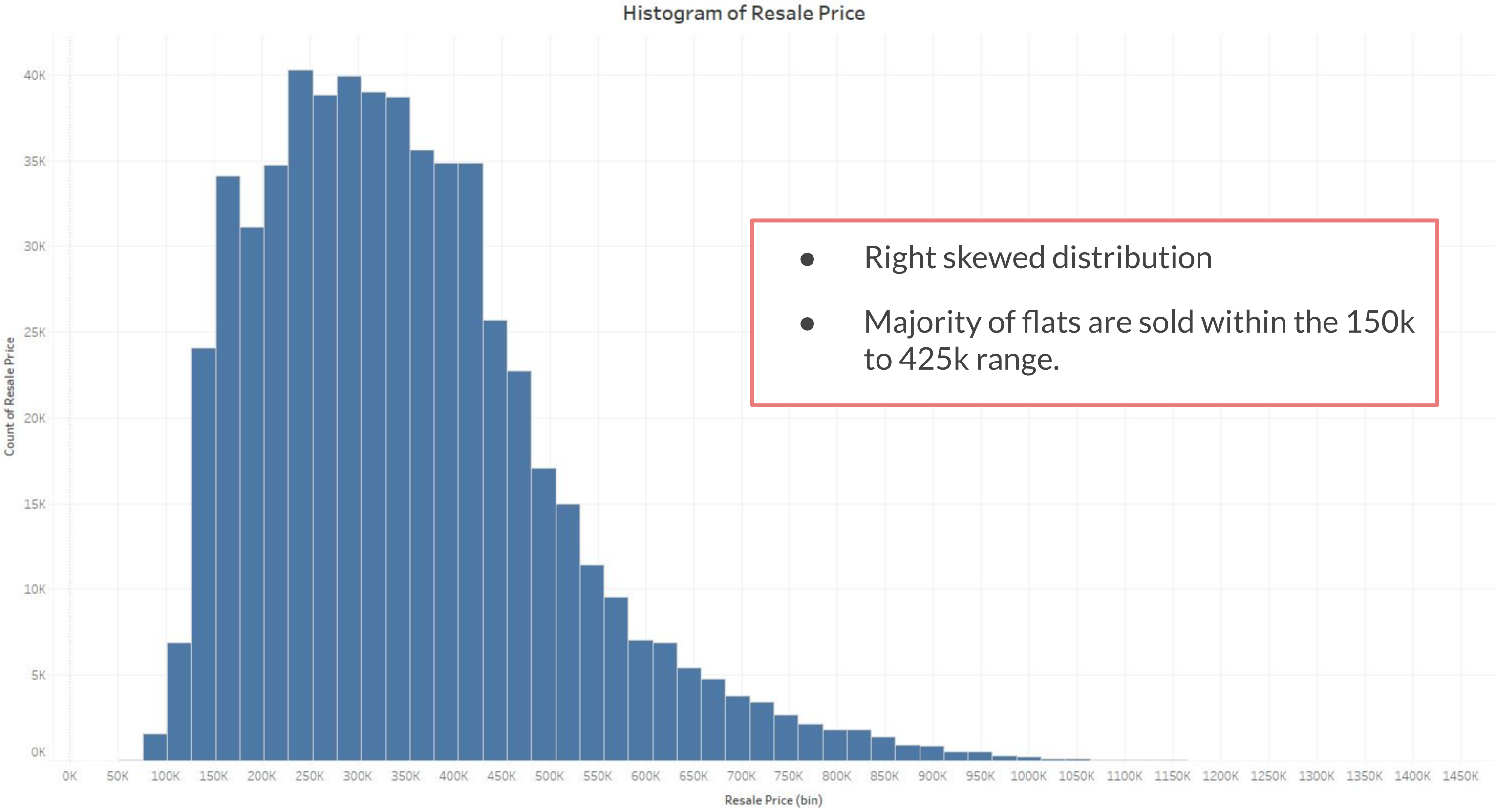
# Data Cleaning and EDA

# Data Cleaning

- Data was mostly clean with no missing values
- `remaining_lease` feature was only tracked from 2015 onwards, can be estimated by (99 - year of sale - lease commence date)
- `remaining_lease` values were also in string format. Needed to convert them into float, with the unit being amount of years.
- `flat_type`: Multi-Generation, 2 room and 1 room resale flats make up only 1.2% of the dataset  (7109 of 588338 entries) and were dropped.

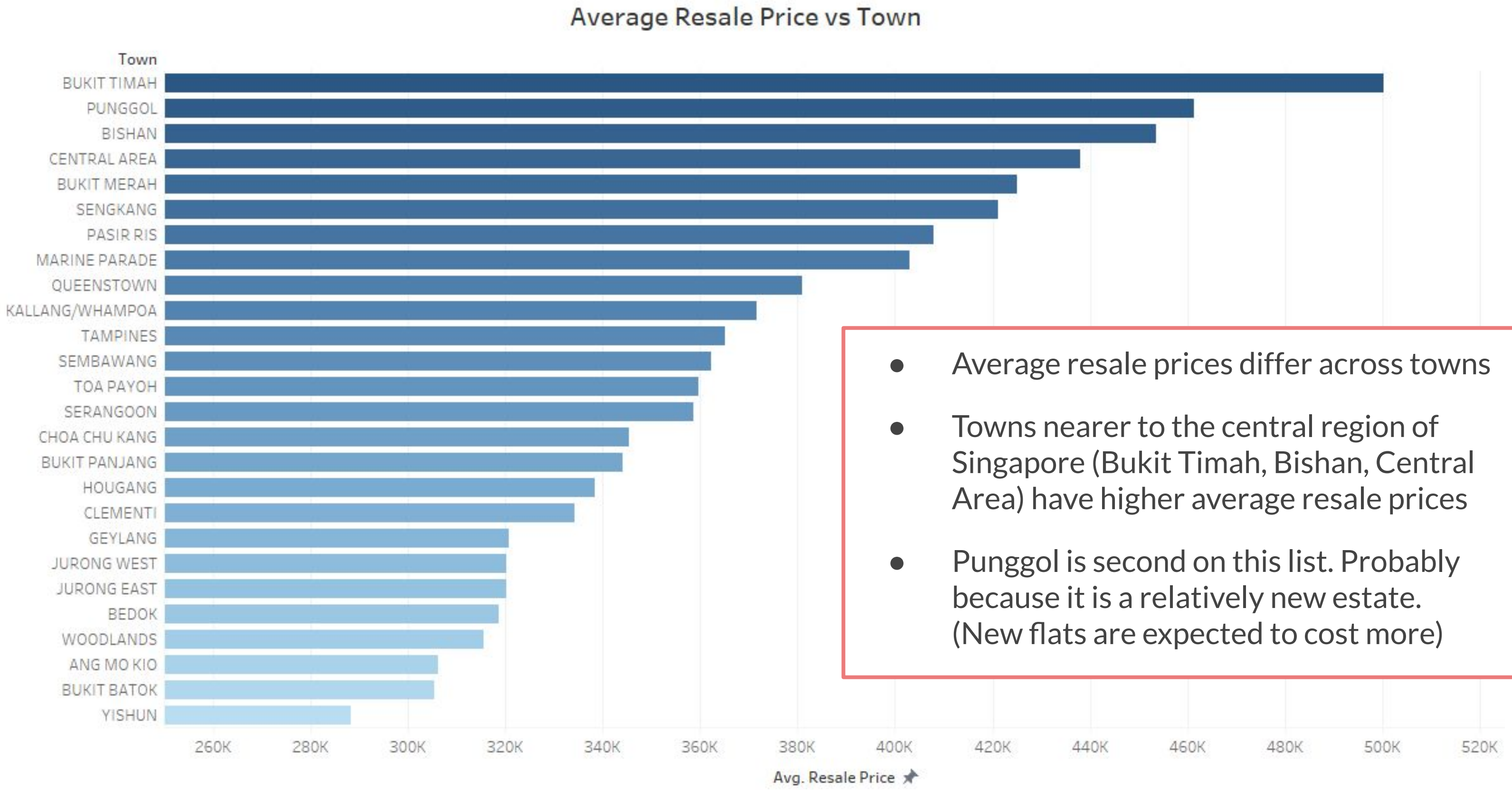


# EDA



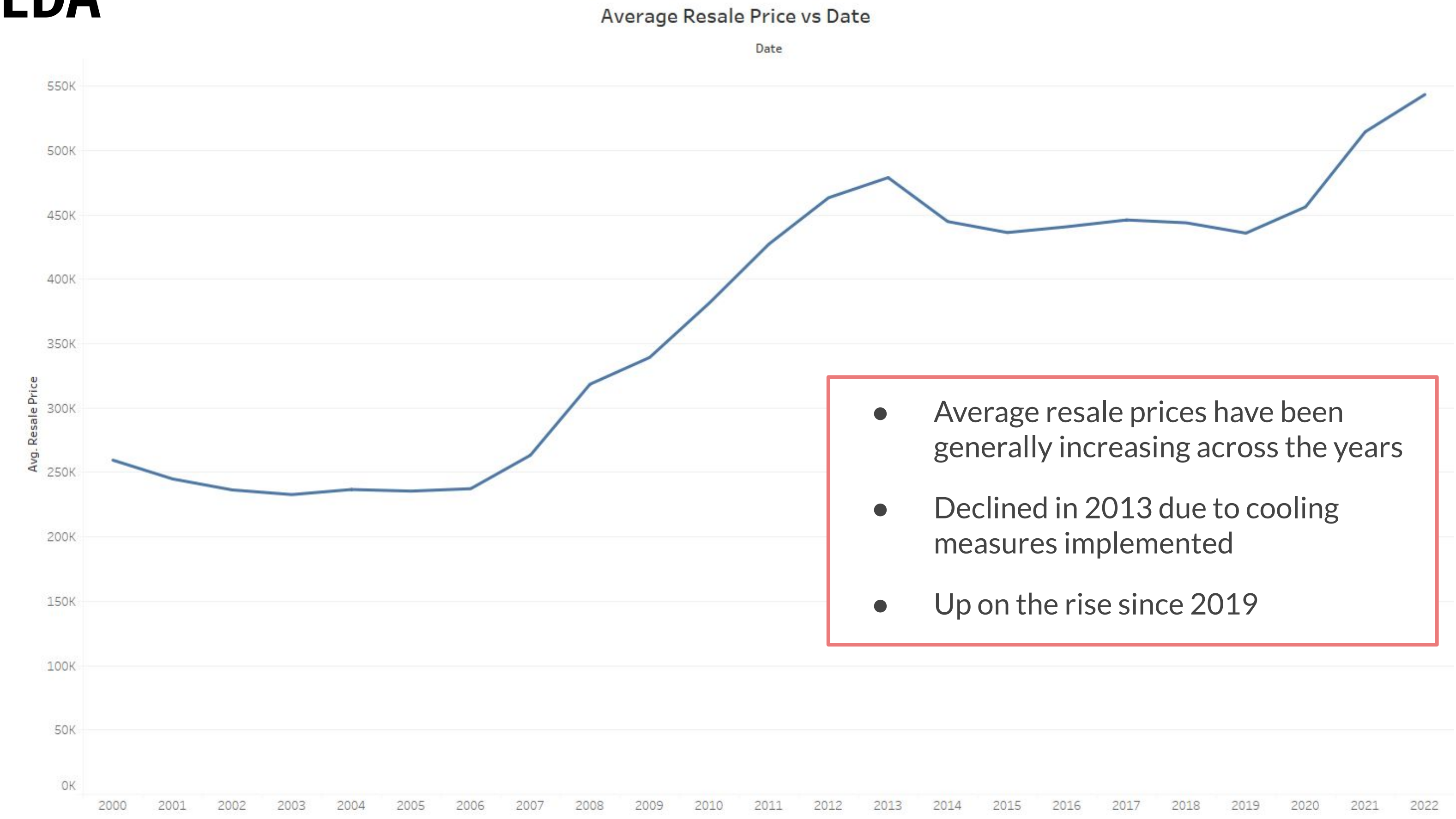

- Right skewed distribution
- Majority of flats are sold within the 150k to 425k range.

# EDA



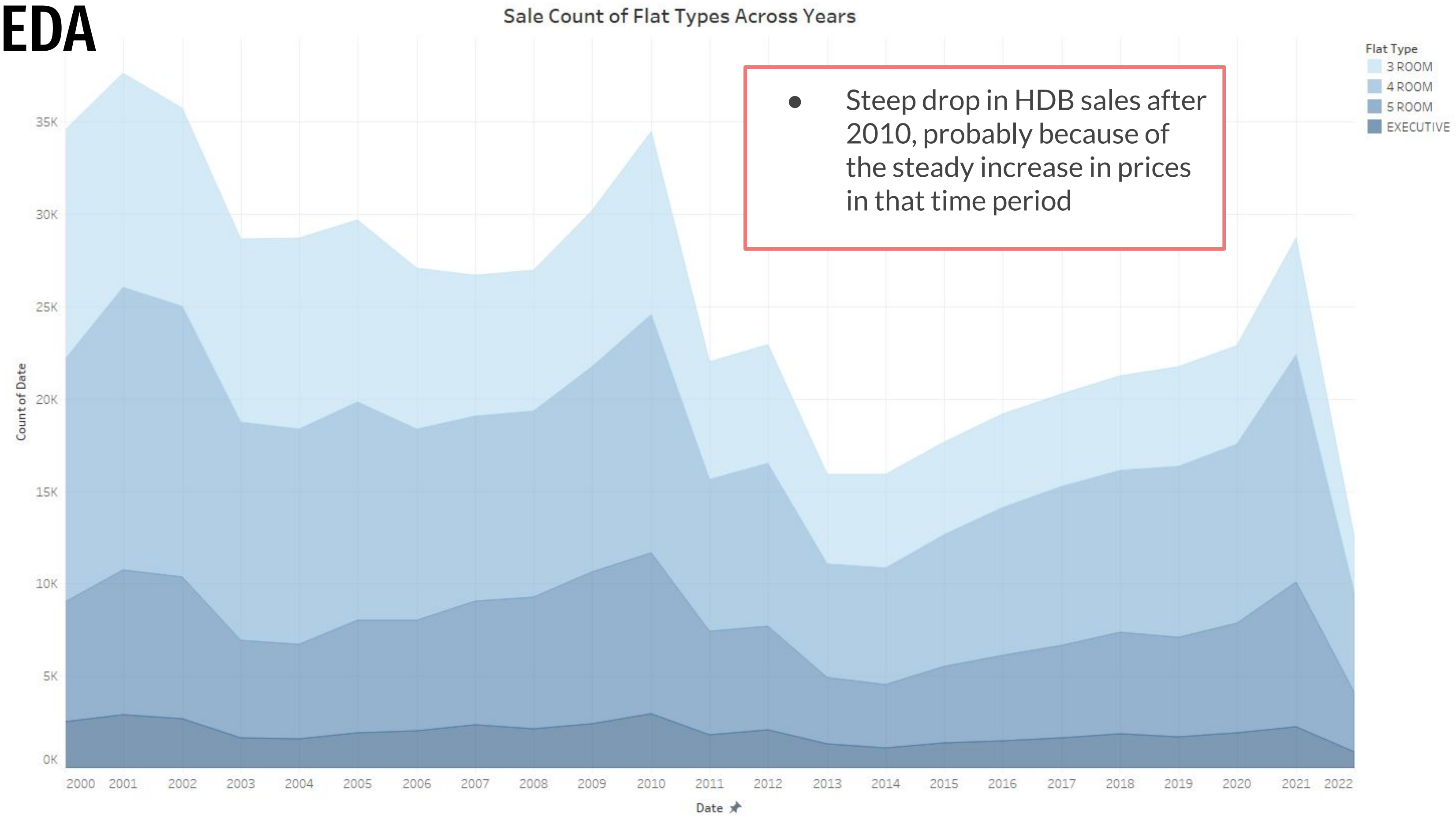

- Average resale prices differ across towns
- Towns nearer to the central region of Singapore (Bukit Timah, Bishan, Central Area) have higher average resale prices
- Punggol is second on this list. Probably because it is a relatively new estate. (New flats are expected to cost more)

# EDA




- Average resale prices have been generally increasing across the years
- Declined in 2013 due to cooling measures implemented
- Up on the rise since 2019

# EDA

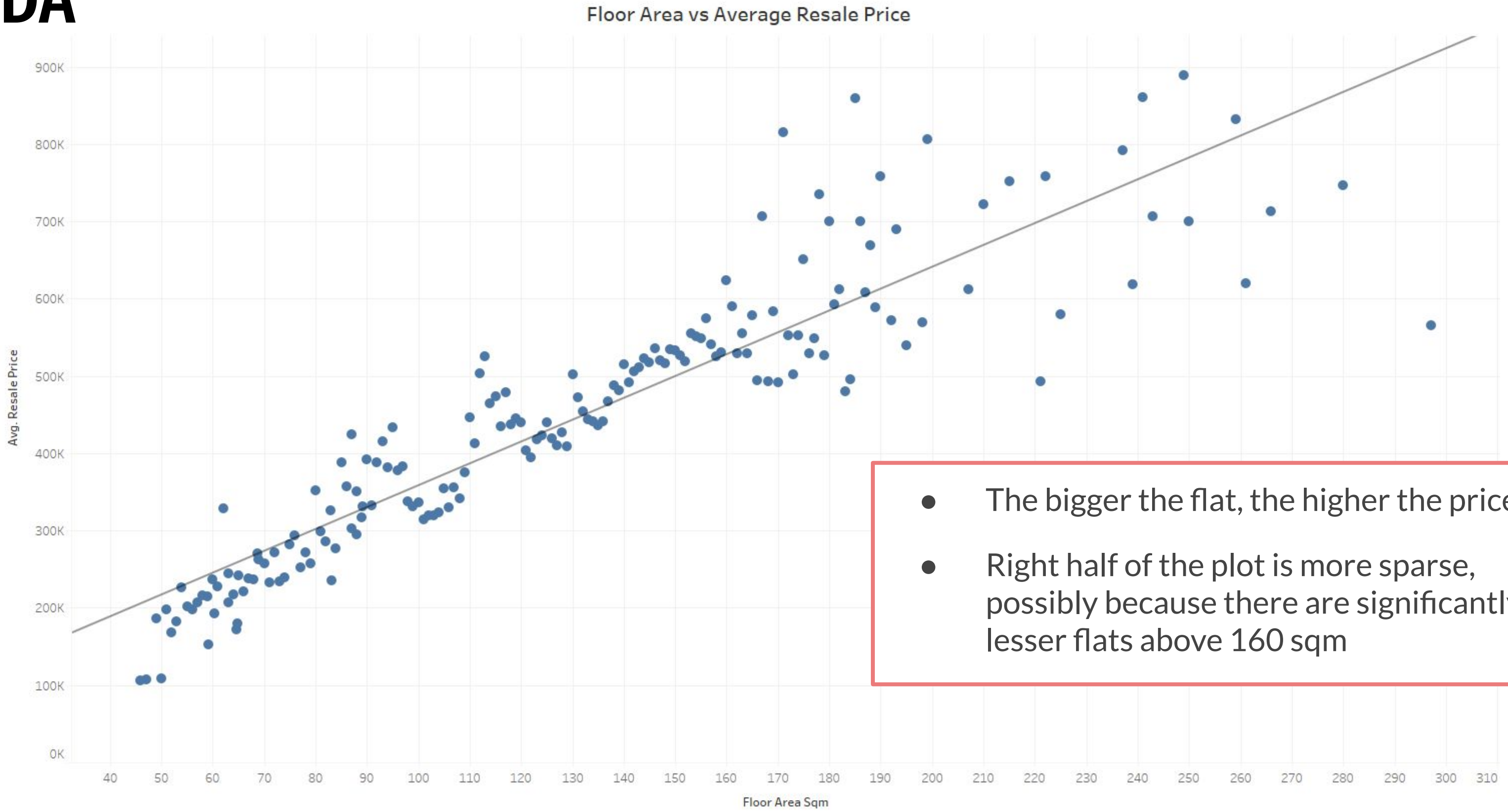



- Steep drop in HDB sales after 2010, probably because of the steady increase in prices in that time period

# EDA

- Average prices also differ across different flat models
- Expected observation as Apartment, Maisonette and Special are generally considered more premium compared to the other 3 models

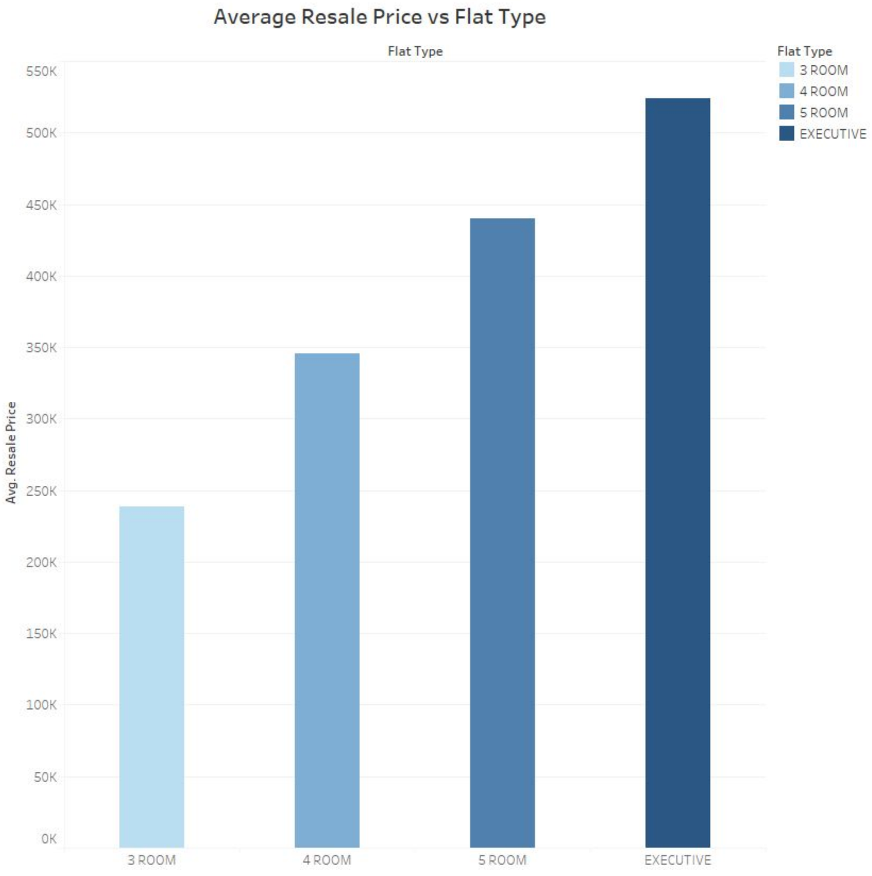
# EDA



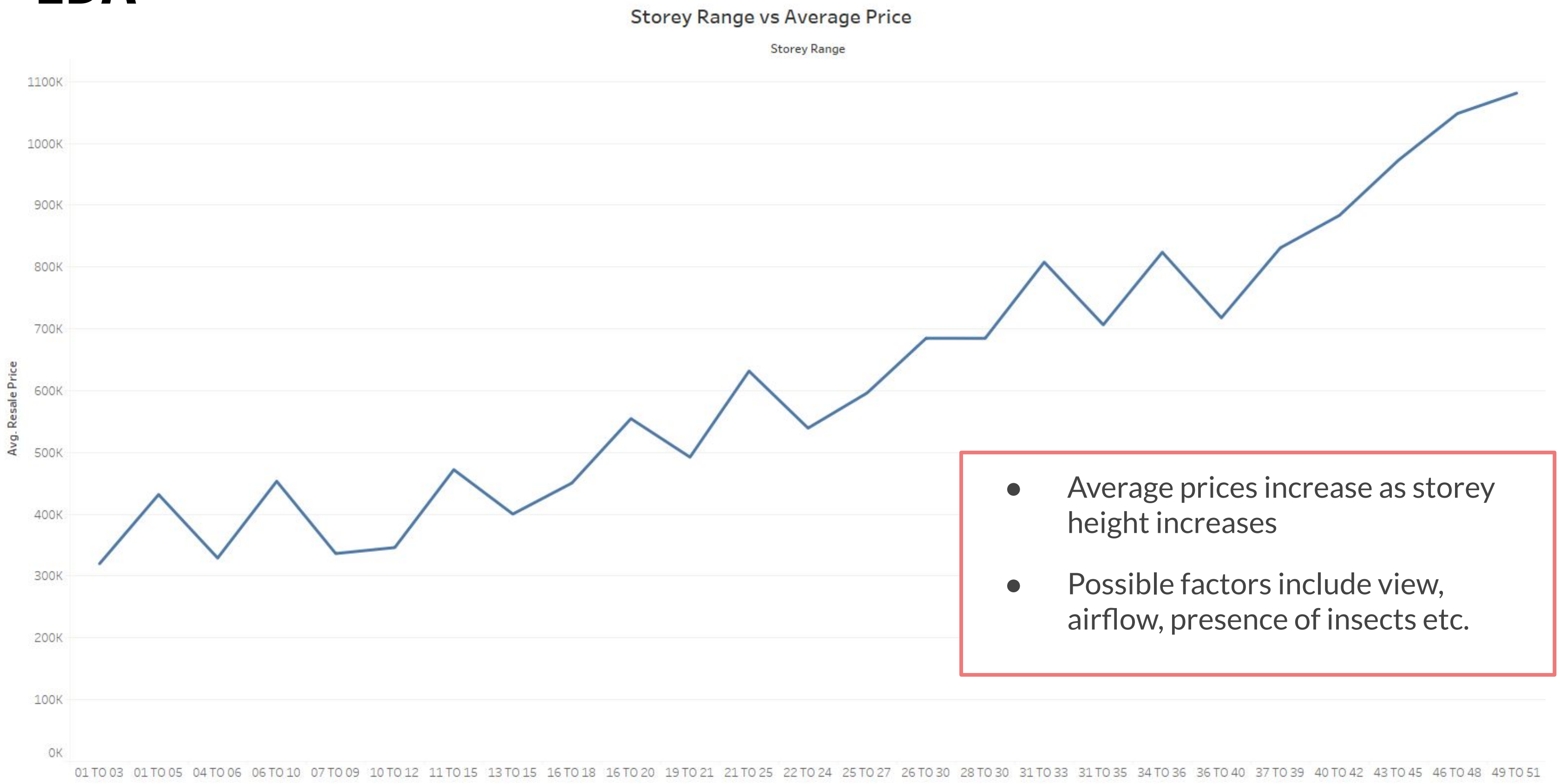

- The bigger the flat, the higher the price.
- Right half of the plot is more sparse, possibly because there are significantly lesser flats above 160 sqm

# EDA

- Flat types are highly correlated to floor area
- Similar observation is to be expected

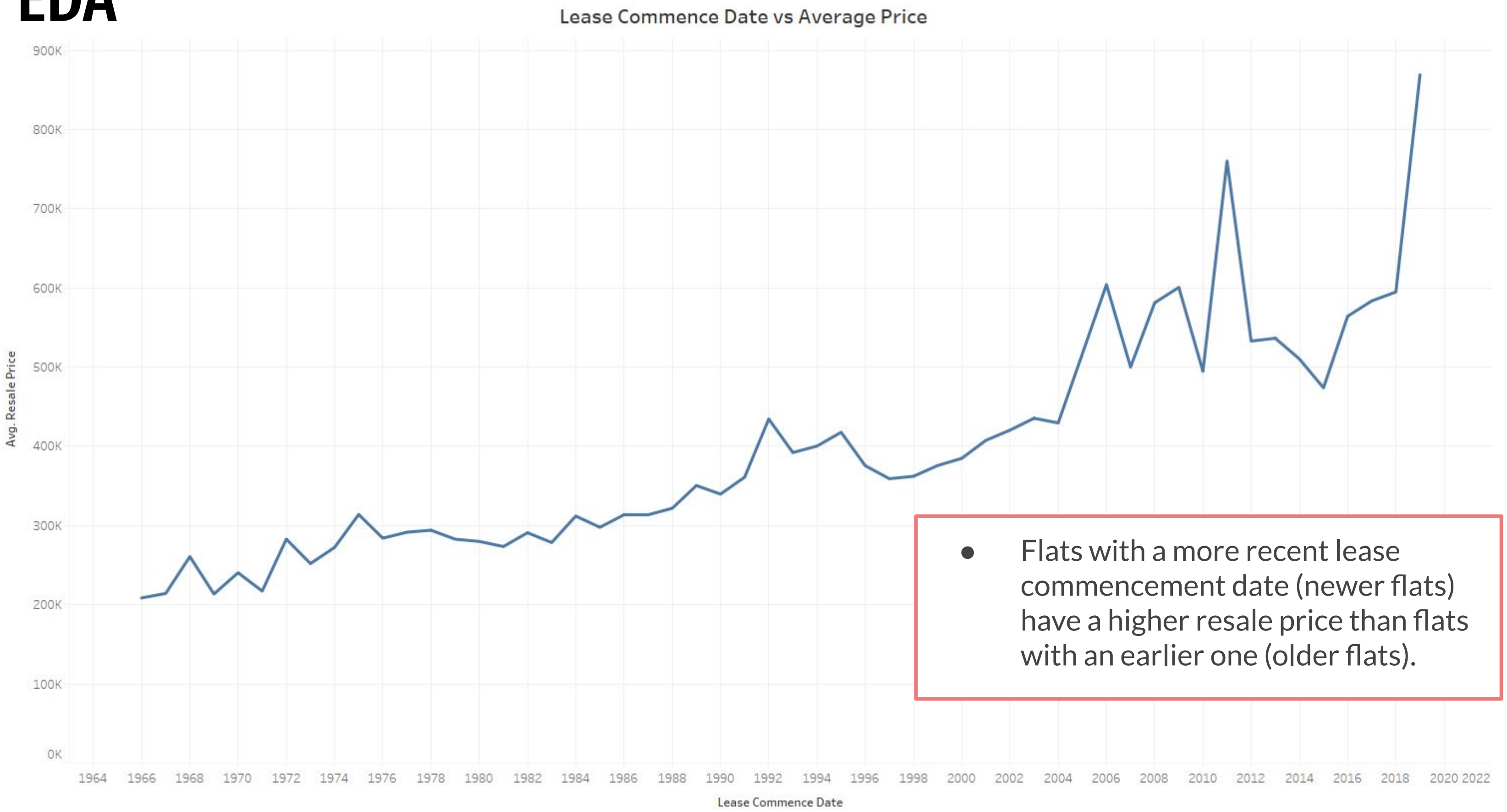


# EDA




- Average prices increase as storey height increases
- Possible factors include view, airflow, presence of insects etc.

# EDA




- Flats with a more recent lease commencement date (newer flats) have a higher resale price than flats with an earlier one (older flats).

# Feature Engineering

# Feature Engineering

| MRT Stations, Shopping Malls, Hawker Centres and Markets, Parks | |
|---|---|
| Distance of nearest amenity | No. of amenities within 1km radius |

| Primary Schools | | |
|---|---|---|
| Distance of nearest Primary School | No. of Primary Schools within 1km radius | No. of Primary Schools between 1km and 2km radius |

# Workflow

**Step 1:**

Scrape coordinates of all unique addresses and amenity locations using OneMap API.

**Step 2:**

For each amenity type, calculate geodesic distance from nearest amenity to each address using Geopy.

**Step 3:**

Calculate the geodesic distance from each address to City Hall MRT.

# Preprocessing

# Encoding Categorical Features

**Town**

Nominal Variable - One-Hot Encoding

**Storey Range**

Flats on higher floors cost more than those on lower floors - Ordinal Encoding

**Flat Type**

Ordinal Variable - Ordinal Encoding

**Flat Model**

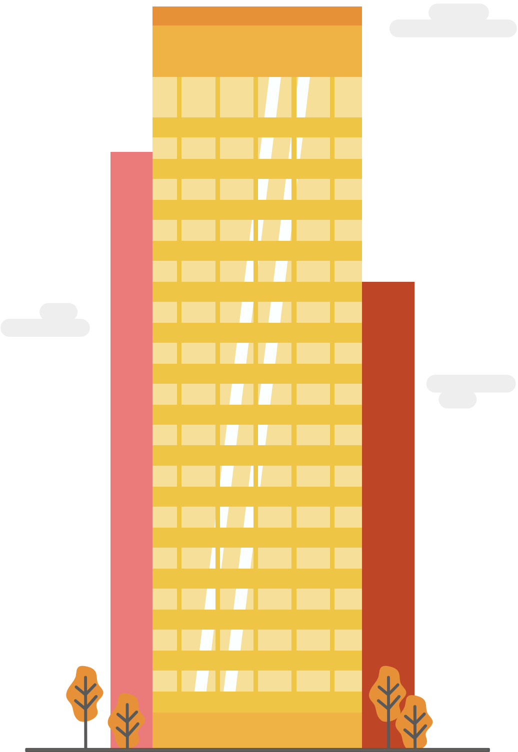
Higher ranks were given to flat models that are considered more 'premium'

- Standard : 1
- New Generation : 1
- Model A : 1
- Apartment : 2
- Maisonette : 2
- Special : 2

# Modeling

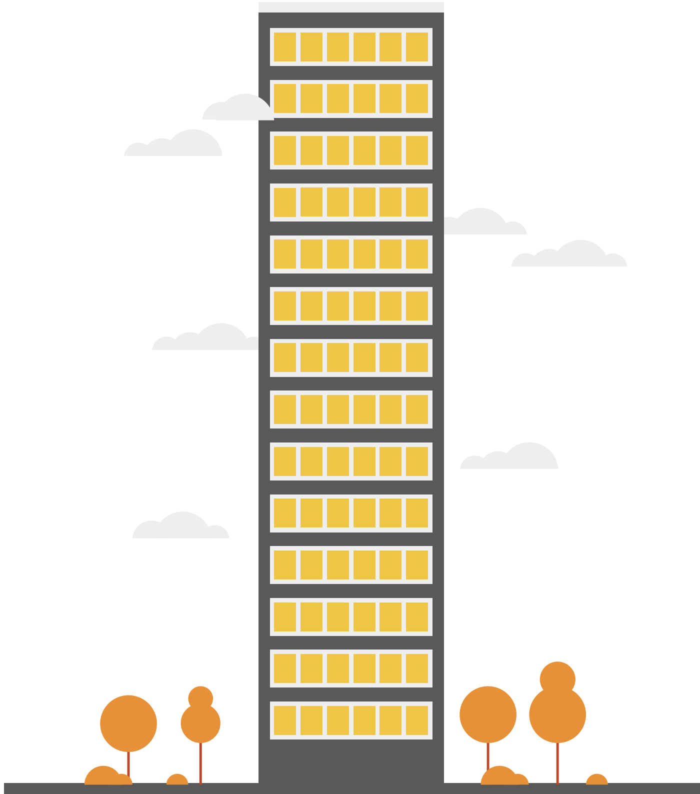
# Metrics Used

1. **Root Mean Squared Error (RSME):**
    a. Tells us how far apart predicted values are from the observed values.
    b. Useful when large errors are particularly undesirable. It is also measured in the same units as our target variable.
    c. Lower = Better
2. **R-Squared (R2) Score:**
    a. Tells us how well a model can predict the value of the target variable in percentage terms.
    b. (Higher = Better)



# Baseline Model: Linear Regression

- Training Scores:
  - RMSE: 62742.26
  - R2 Score:  0.8330
- Testing Scores
  - RMSE: 62480.08
  - R2 Score:  0.8338



# Modeling with PyCaret

- Improved scores for Linear Regression, R2 score increased by 0.02 and RSME decreased by ~3k. Shows that PyCaret environment was useful in improving the accuracy of our predictions.
- Extra Trees is our best scoring model, with RSME ~21k and R2 score ~0.98.
- Took around >100 times the amount of time compared to our 3rd highest scoring model, LightGBM
- LightGBM was chosen for tuning because it is much more efficient.

| Model | RMSE | R2 | TT (Sec) |
|---|---|---|---|
| Extra Trees Regressor | 21290.4761 | 0.9808 | 512.1920 |
| Random Forest Regressor | 21750.5305 | 0.9799 | 463.4090 |
| Light Gradient Boosting Machine | 29145.4967 | 0.9640 | 4.3950 |
| Decision Tree Regressor | 30255.2884 | 0.9612 | 16.4440 |
| Gradient Boosting Regressor | 40440.8772 | 0.9306 | 222.8510 |
| K Neighbors Regressor | 54080.7033 | 0.8760 | 173.4850 |
| Huber Regressor | 58837.8301 | 0.8532 | 87.5250 |
| Linear Regression | 59031.0465 | 0.8522 | 2.3370 |
| Ridge Regression | 59031.1184 | 0.8522 | 0.8560 |
| Bayesian Ridge | 59031.9915 | 0.8522 | 6.3700 |

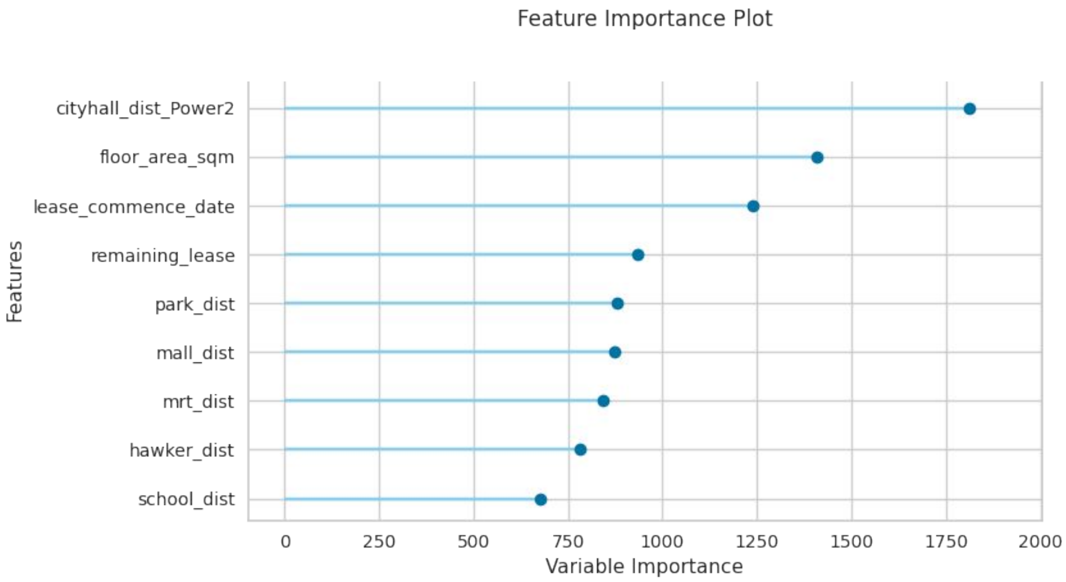
# Model Tuning with PyCaret

- Model was tuned to optimize RSME
- Optuna, an open source hyperparameter optimization framework was used to automate hyperparameter search
- **Tuned Model's RSME:** 21,419 .36 (~7,700 lower after tuning)
- Retrained on hold-out set to prepare for future deployment

| Fold | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
|---|---|---|---|---|---|---|
| **0** | 15081.2319 | 450315266.5335 | 21220.6330 | 0.9810 | 0.0612 | 0.0453 |
| **1** | 15148.4986 | 461003348.1430 | 21470.9885 | 0.9803 | 0.0611 | 0.0451 |
| **2** | 15128.2186 | 454942678.1625 | 21329.3853 | 0.9807 | 0.0610 | 0.0450 |
| **3** | 15202.5464 | 463638959.5493 | 21532.2772 | 0.9805 | 0.0610 | 0.0452 |
| **4** | 15079.4048 | 455179501.1742 | 21334.9362 | 0.9810 | 0.0606 | 0.0450 |
| **5** | 15157.3300 | 463650283.3495 | 21532.5401 | 0.9801 | 0.0616 | 0.0455 |
| **6** | 15182.8628 | 461169208.4013 | 21474.8506 | 0.9805 | 0.0614 | 0.0453 |
| **7** | 15085.1678 | 460691746.1750 | 21463.7309 | 0.9803 | 0.0613 | 0.0452 |
| **8** | 15125.0562 | 459132118.1083 | 21427.3684 | 0.9806 | 0.0610 | 0.0453 |
| **9** | 15132.6712 | 458255857.0620 | 21406.9114 | 0.9805 | 0.0617 | 0.0454 |
| **Mean** | 15132.2988 | 458797896.6659 | $\boxed{21419.3622}$ | 0.9805 | 0.0612 | 0.0452 |
| **Std** | 40.0957 | 4017105.5302 | 93.9190 | 0.0003 | 0.0003 | 0.0001 |

# Feature Importance

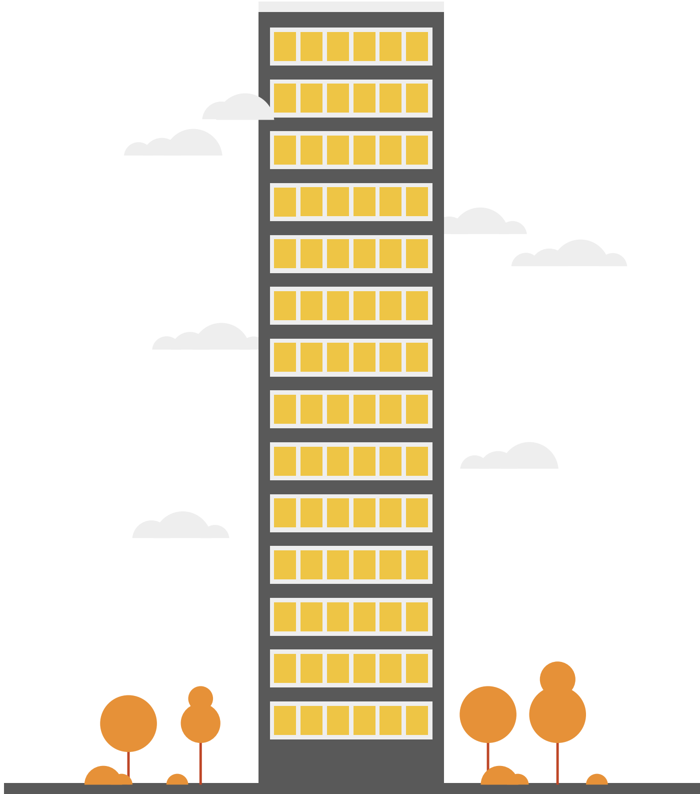
- 7 out of the top 10 features were engineered
- Probably because tree-based models are biased towards continuous variables



# Conclusion and Limitations

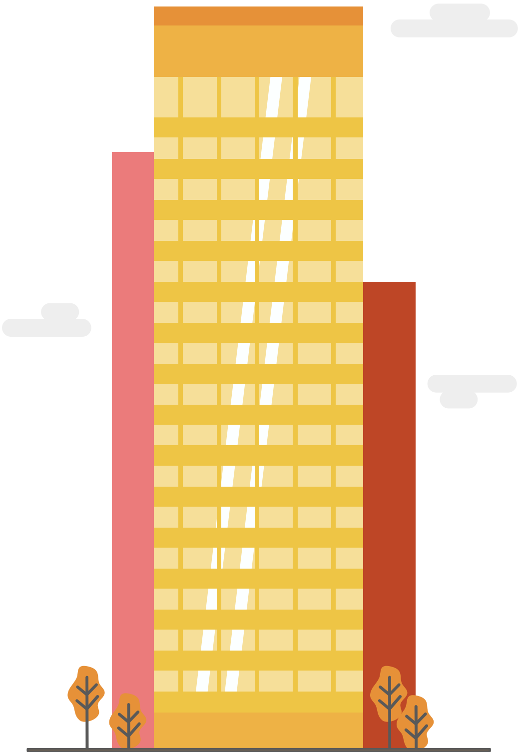
# Conclusion

- The final model has a significantly lower RSME when compared to our baseline. The final RSME is $20,727, which is less than 4% of the current mean resale price of ~$550k.
- Engineered features have a statistically significant relationship with resale flat prices, and were useful in improving the accuracy of our model.
- The model should be able to detect undervalued and overvalued flats, and should be able to give a good estimate of COVs.
- Future plans:
    - Aim to deploy my final model in a web-based application like Streamlit.
    - Test (or train) this model further on future data (HDB updates their datasets regularly)



# Limitations

1. Unable to include other factors that will influence resale flat prices:
    a. **Condition of flats:** Flats with extensive renovations and furnishings or flats which are well maintained tend to fetch a higher price.
    b. **Directions flats are facing:** Higher demand for flats that are North/South facing compared to East/West
2. Trade off between accuracy and efficiency of model:
    a. Models like Extra Trees and Random Forest are more accurate but take longer time to be processed
    b. Deployability was a concern so LightGBM was chosen
3. Geodesic distance from location was used instead of travel time:
    a. Travel time is expected to be a more accurate feature but requires the paid Google Maps API



# References

- HDB Portal
- PropertyGuru
- MOE Portal
- Yahoo! Finance
- Teoalida's Website

# Credits

- Slide Template and Illustrations: **Slidesgo** and **Freepik**

# Thank you!