



# HDB Resale Price Analysis

via Multivariable Linear Regression

Presented and prepared by:  
JunQi, Jessica, Ziyang

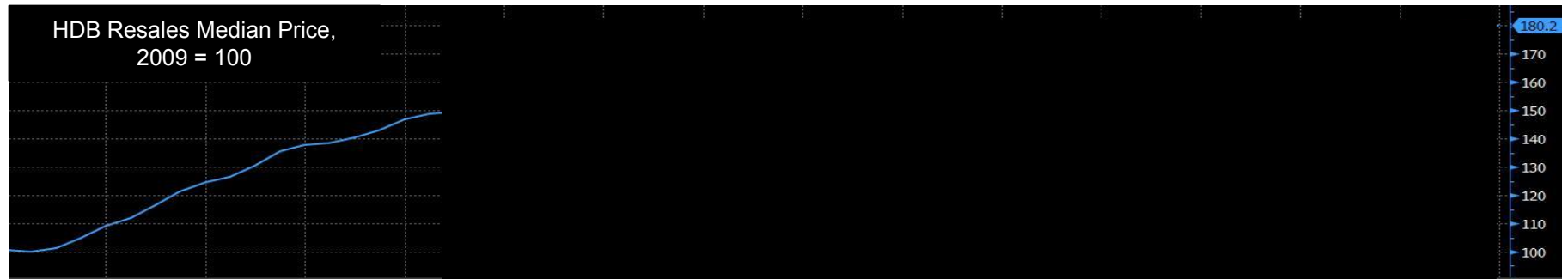
# Agenda

- ❖ Macroeconomics Background
- ❖ Exploratory Data Analysis
- ❖ Common Questions answered
- ❖ Modeling
  - Features
  - Data Cleaning
  - Model A vs Model B
- ❖ Limitation
- ❖ Conclusion

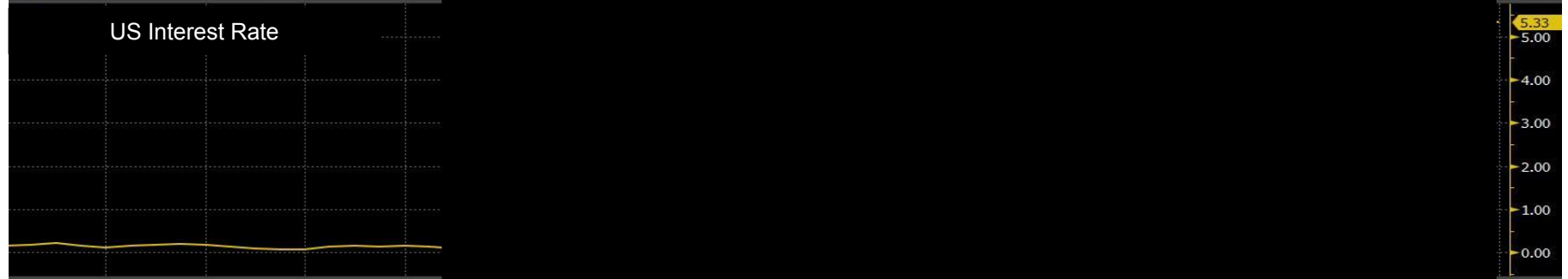


Image source: <https://satoshispeaks.com/the-2008-financial-crisis-peak-of-the-modern-financial-system/>

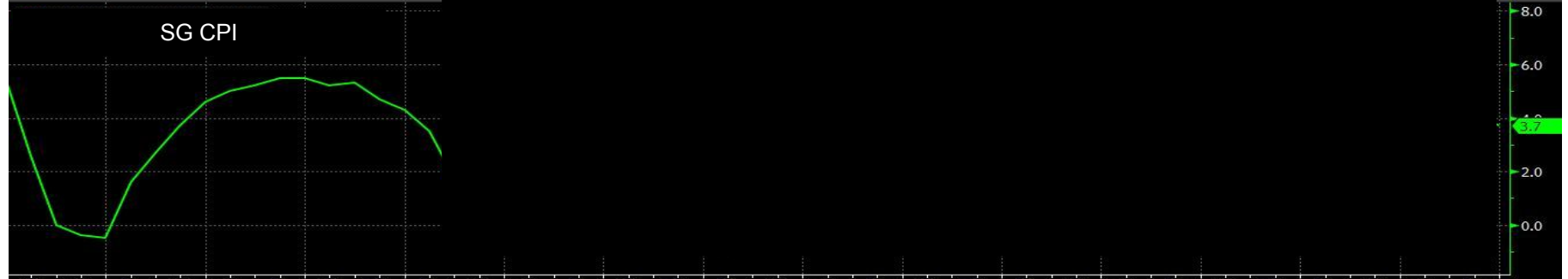
HDB Resales Median Price,  
2009 = 100



US Interest Rate



SG CPI



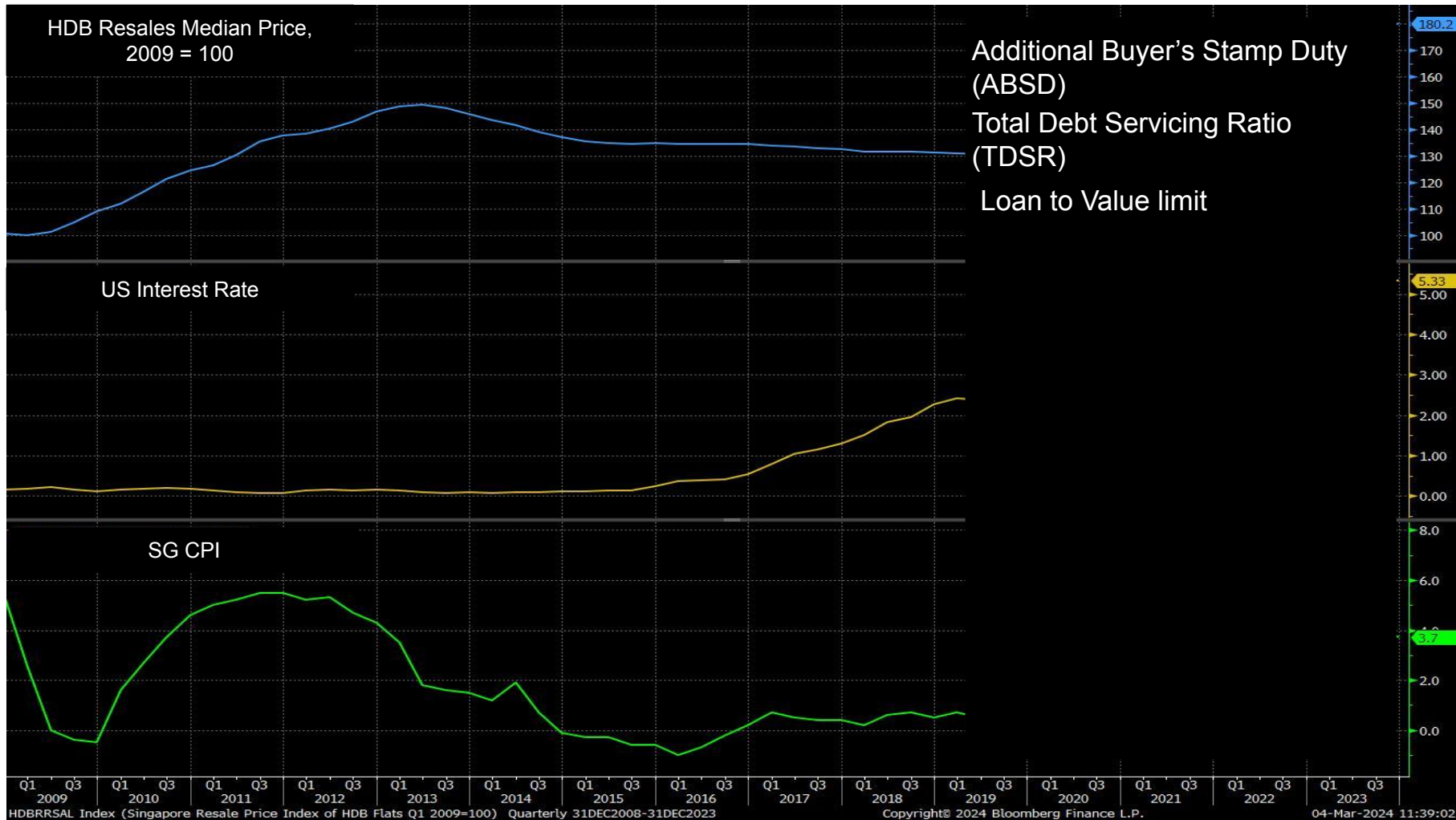


HDB Resales Median Price,  
2009 = 100

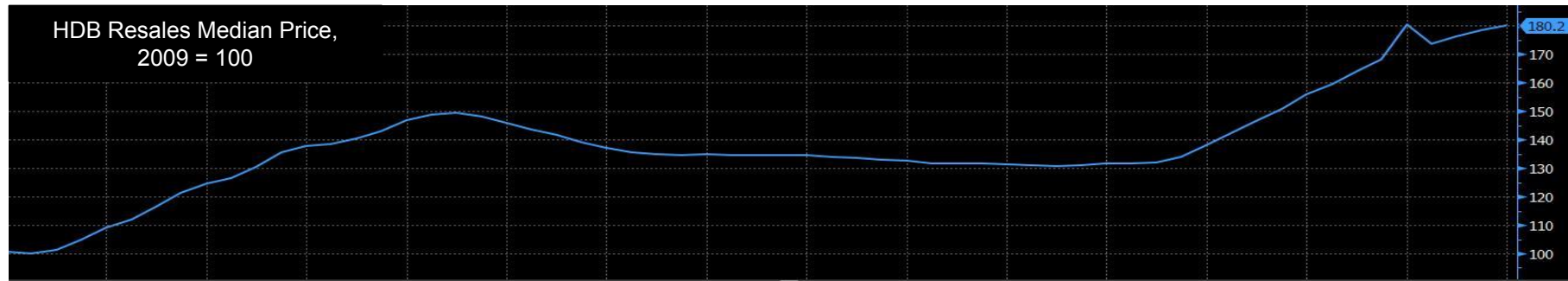
Additional Buyer's Stamp Duty  
(ABSD)  
Total Debt Servicing Ratio  
(TDSR)  
Loan to Value limit

US Interest Rate

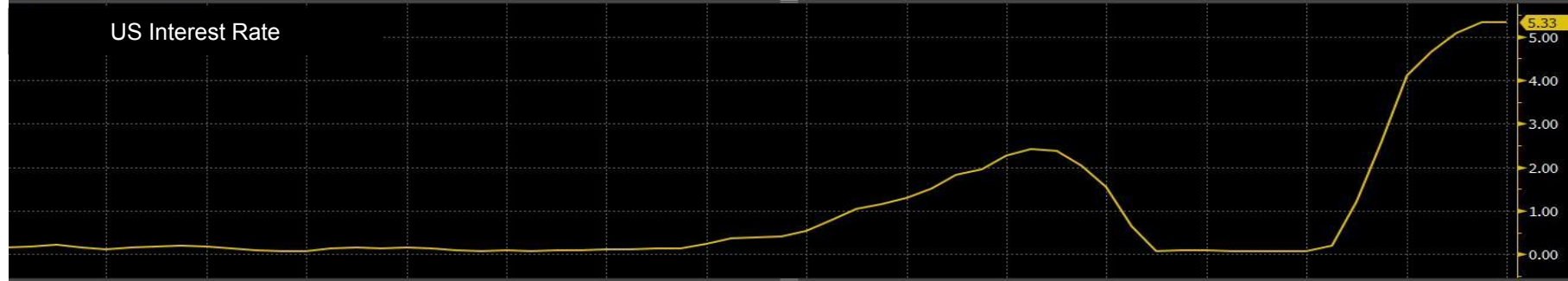
SG CPI



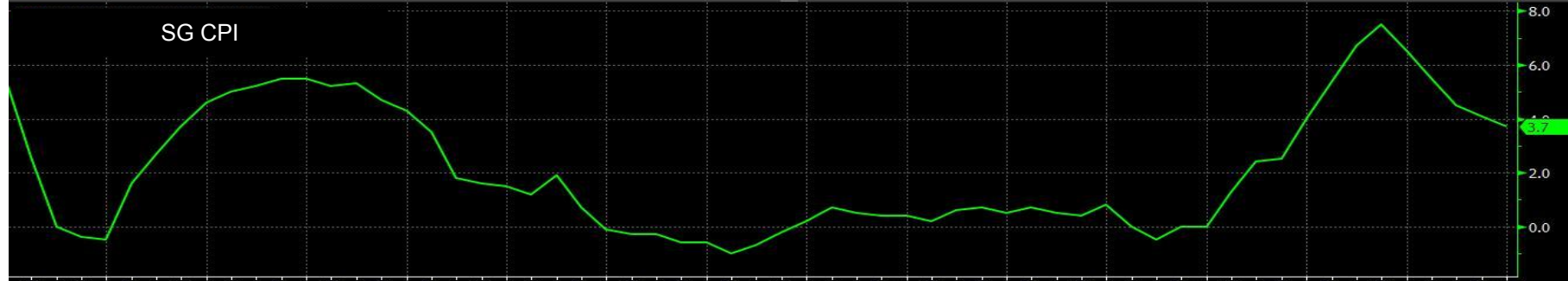
HDB Resales Median Price,  
2009 = 100



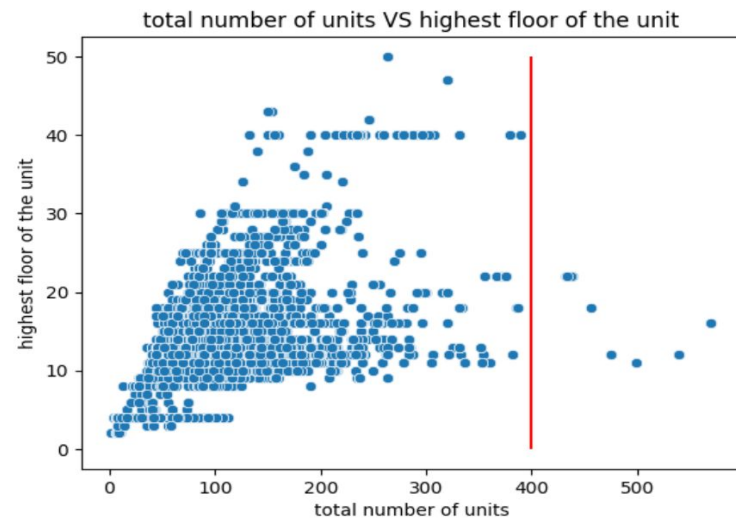
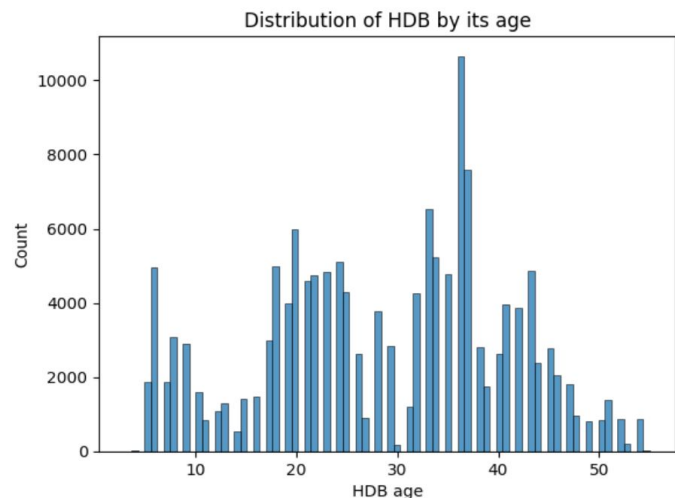
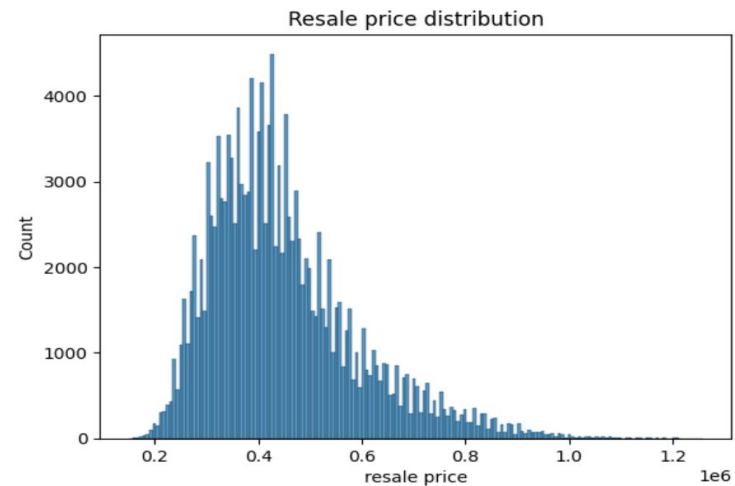
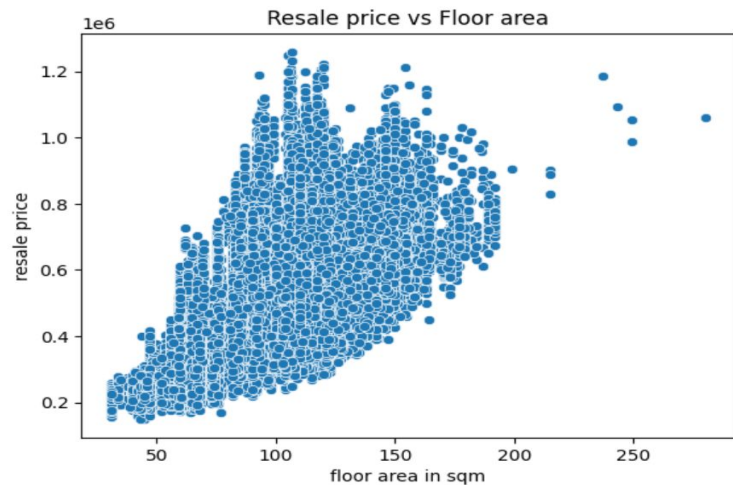
US Interest Rate



SG CPI

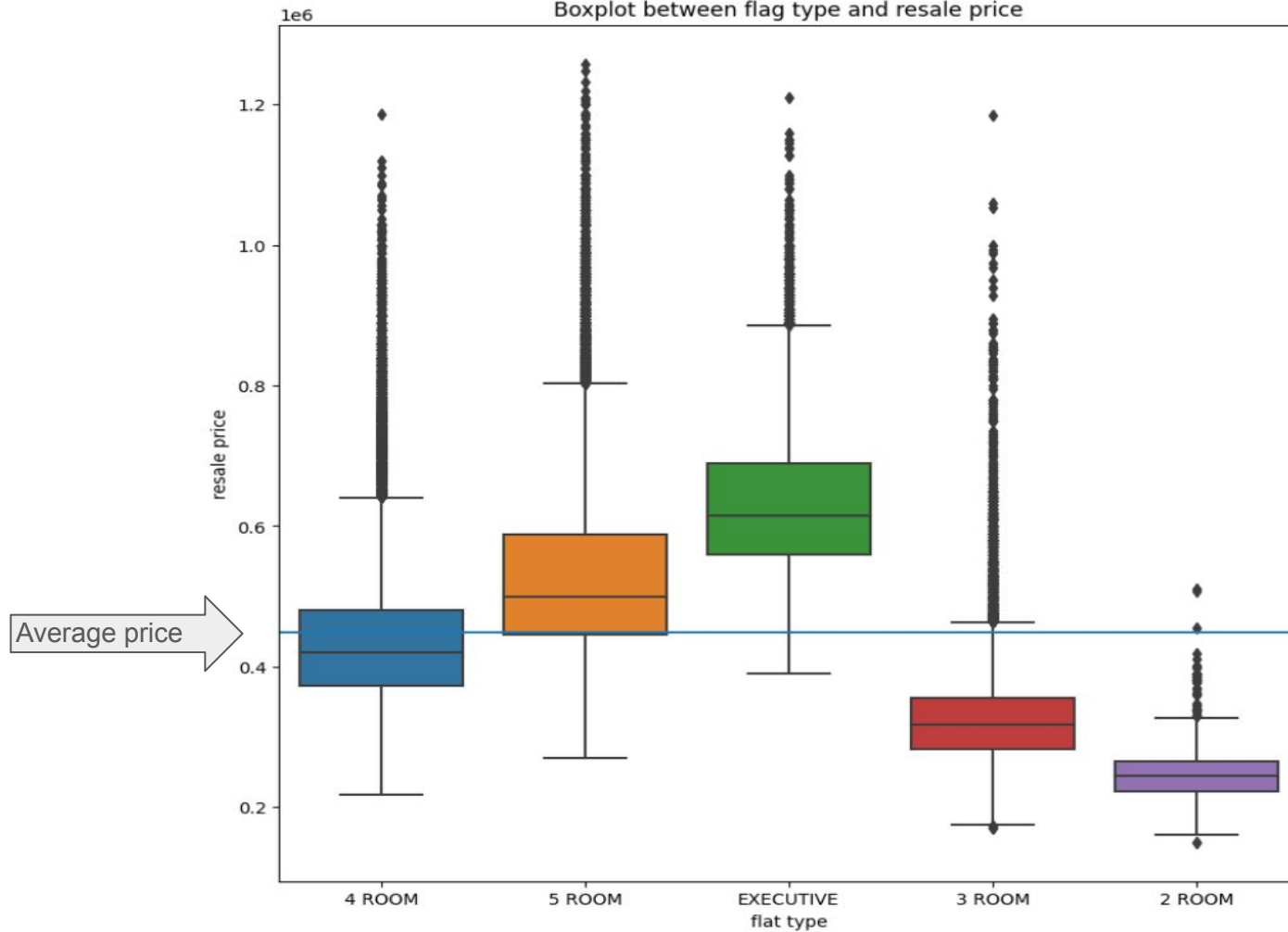


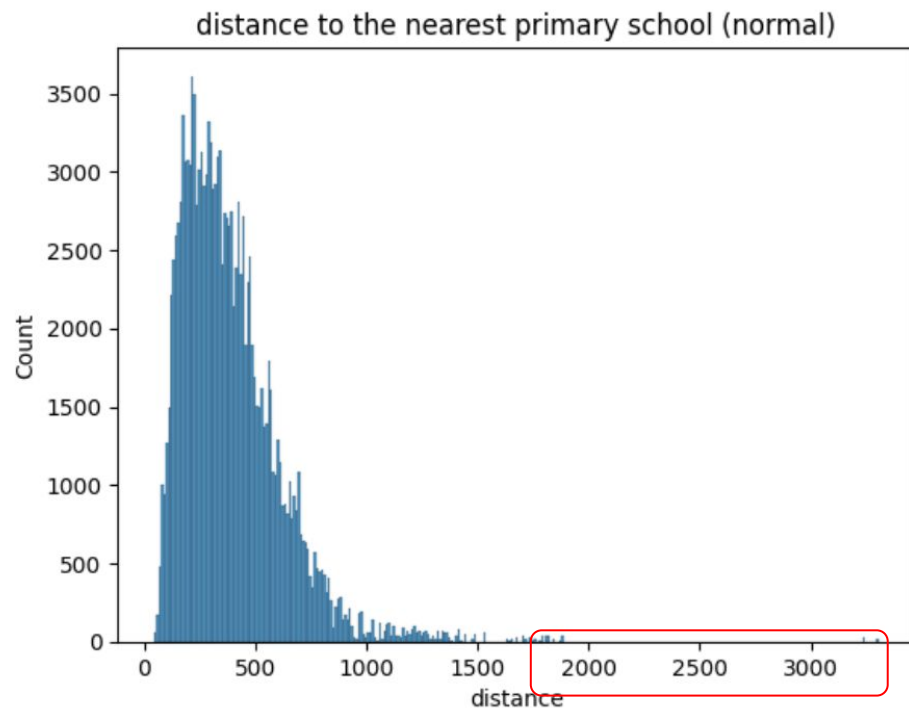
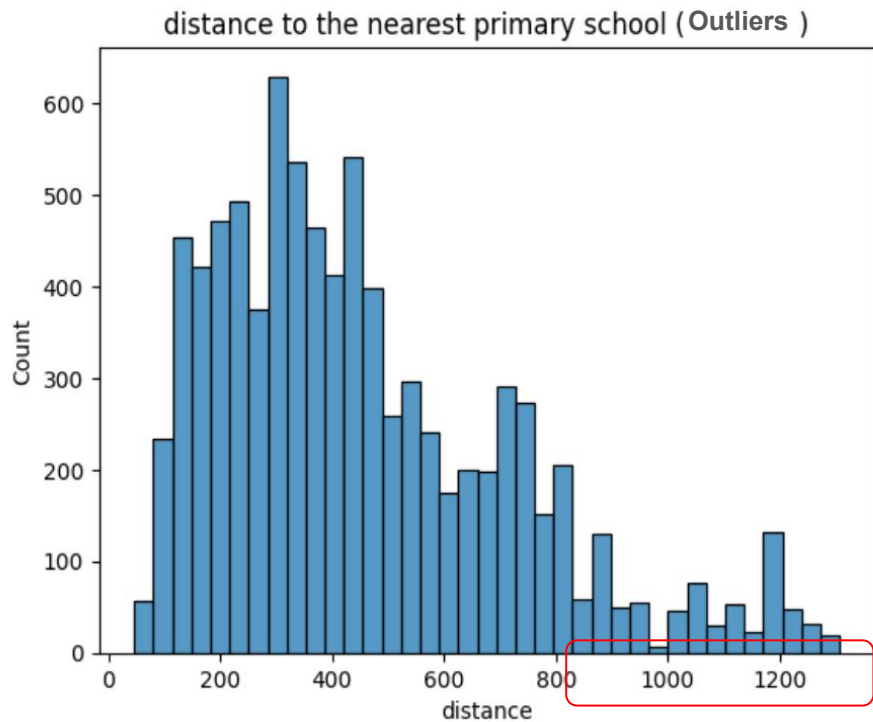
# Exploratory Data Analysis





Boxplot between flag type and resale price





Could the distance between nearest primary school makes the outliers “outliers”?

# Common Asked Questions

Is Central The Best Area to Invest ?



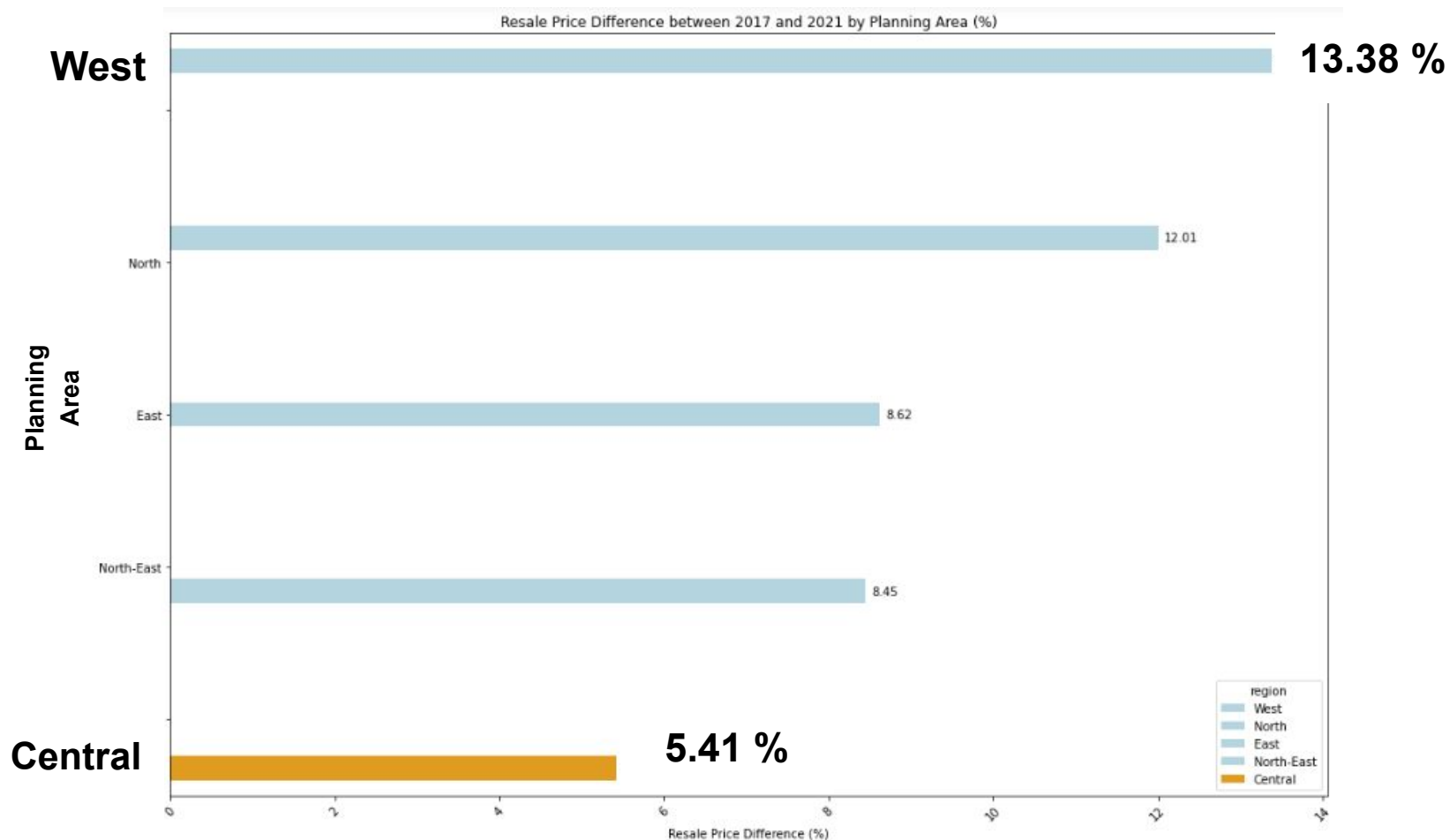
How about mature estate?



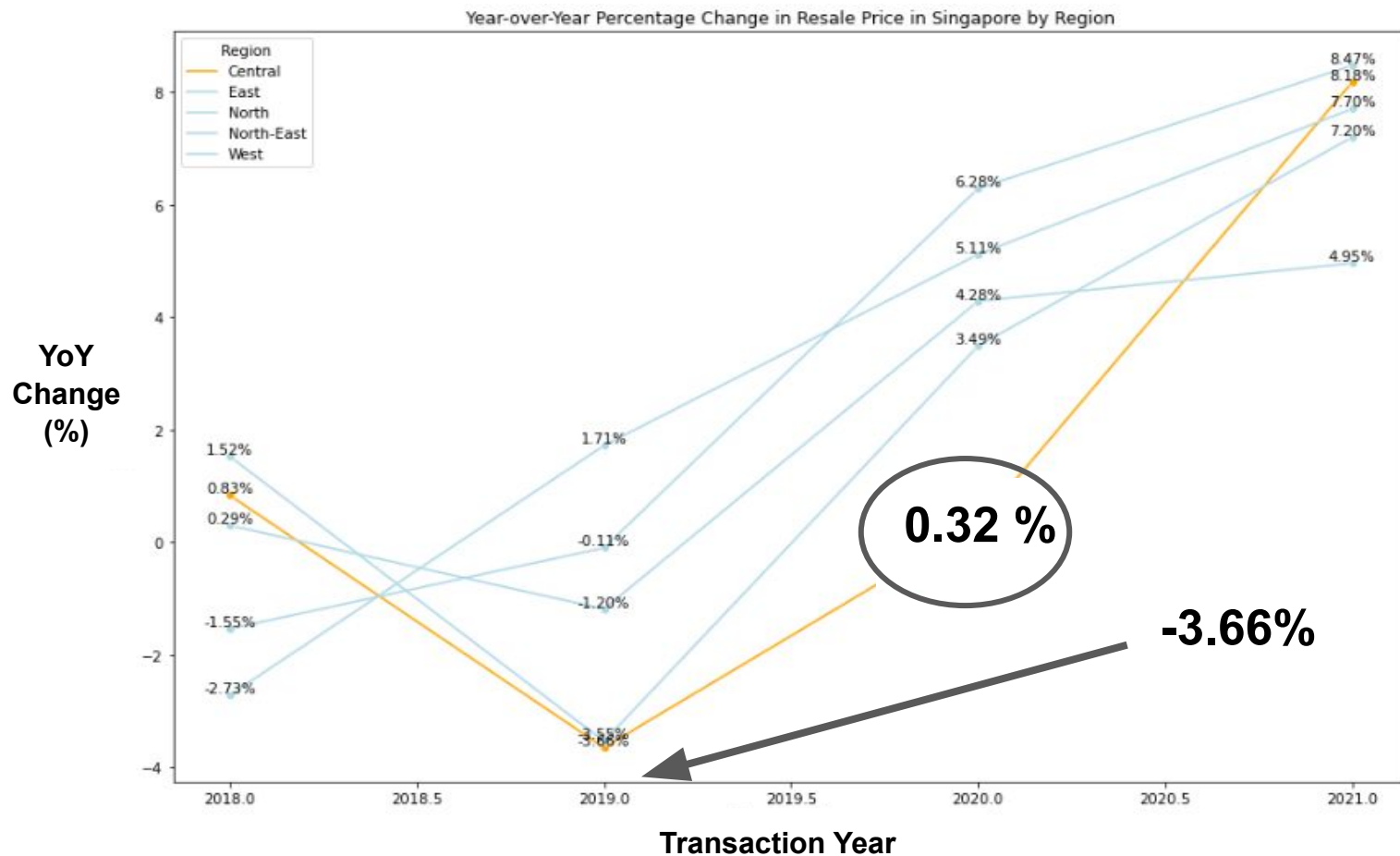
Where is the Best Planning Area to invest?



# Is Central The Best Area To Invest?



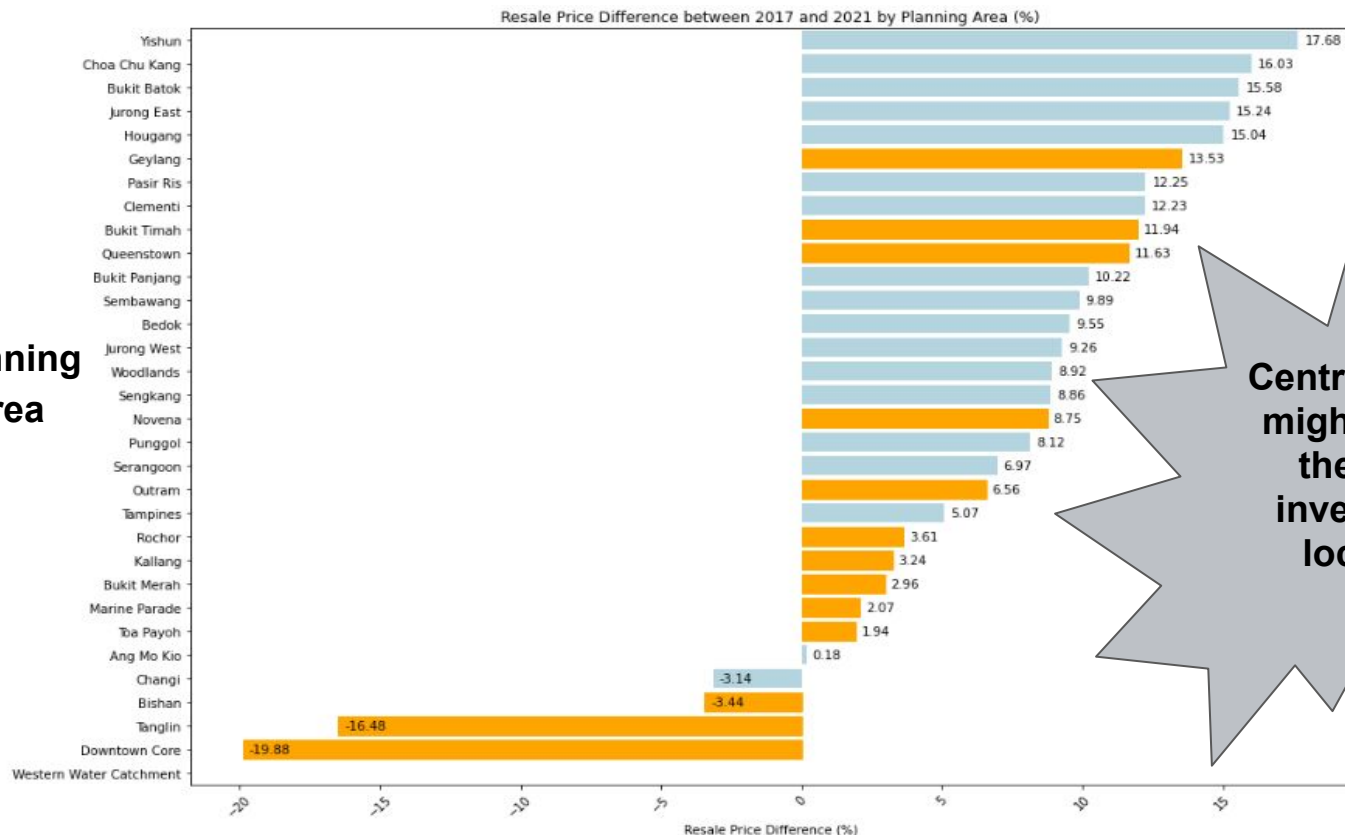
# Is Central The Best Area To Invest?





# Is Central The Best Area To Invest?



Central  
Non-Central

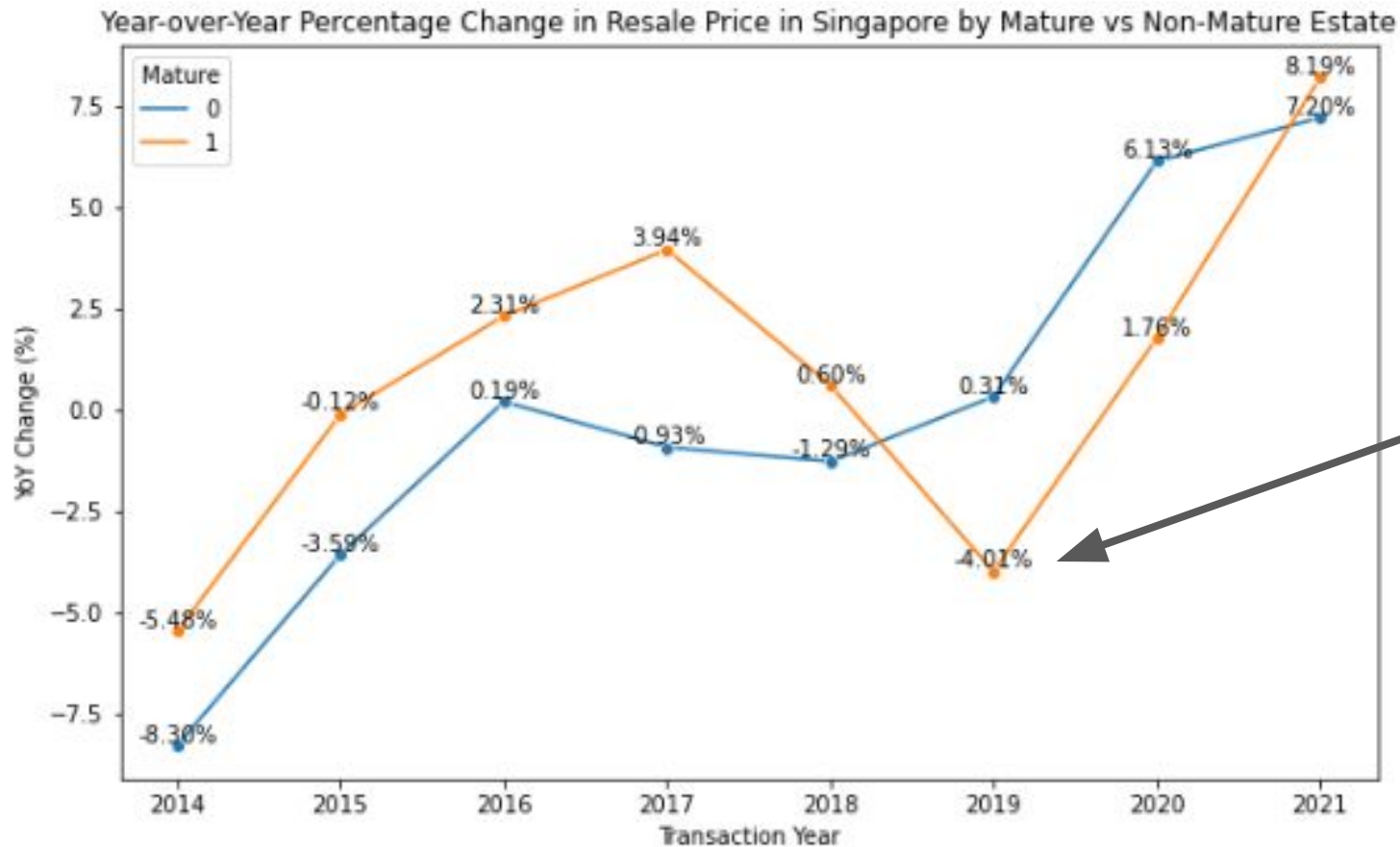


Central region  
might not be  
the ideal  
investment  
location



# Is Mature Estate The Best Area To Invest?

 Mature  
 Non-Mature

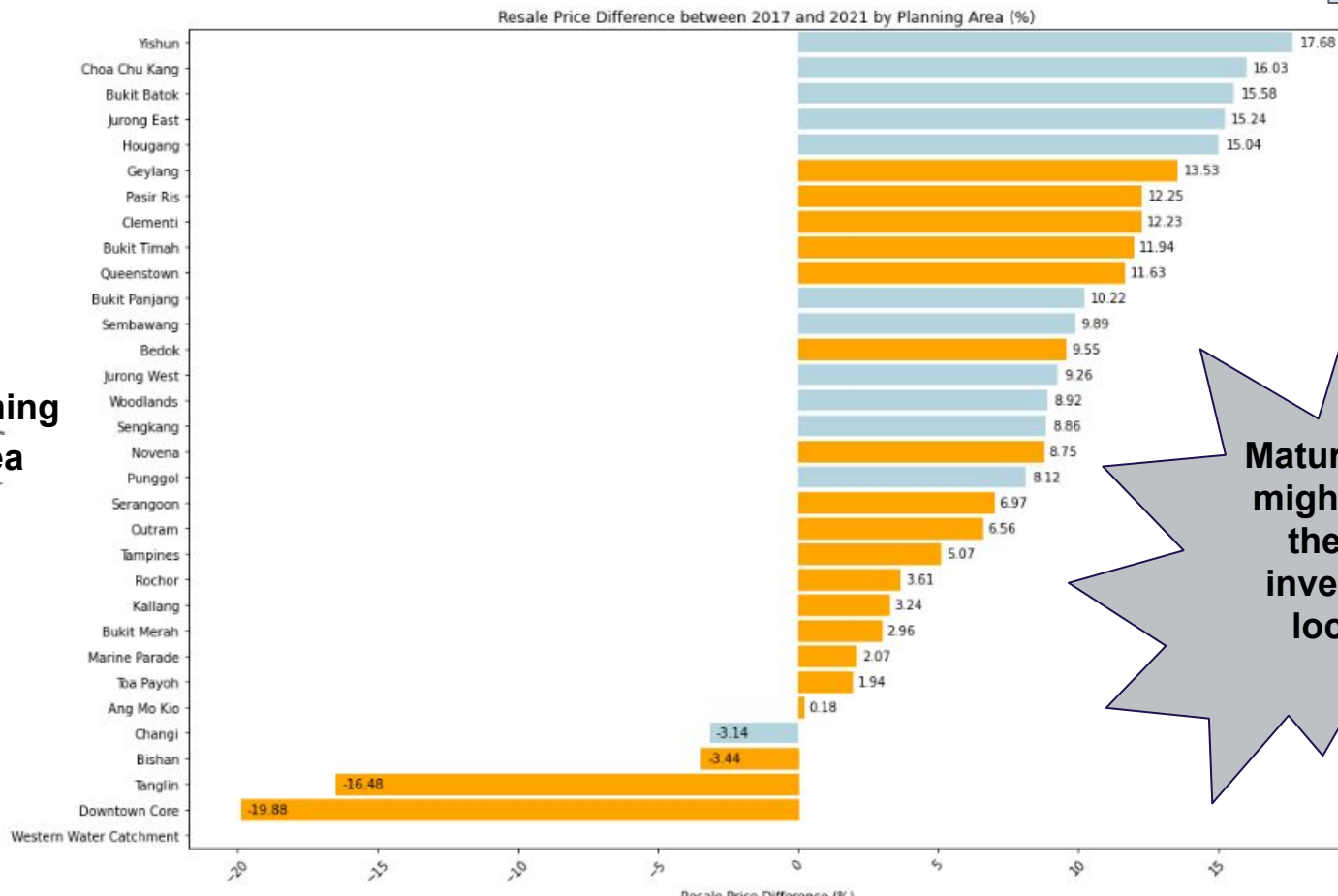


**-4.01%**

# Is Mature Estate The Best Area To Invest?

 Mature  
 Non-Mature

Planning  
Area

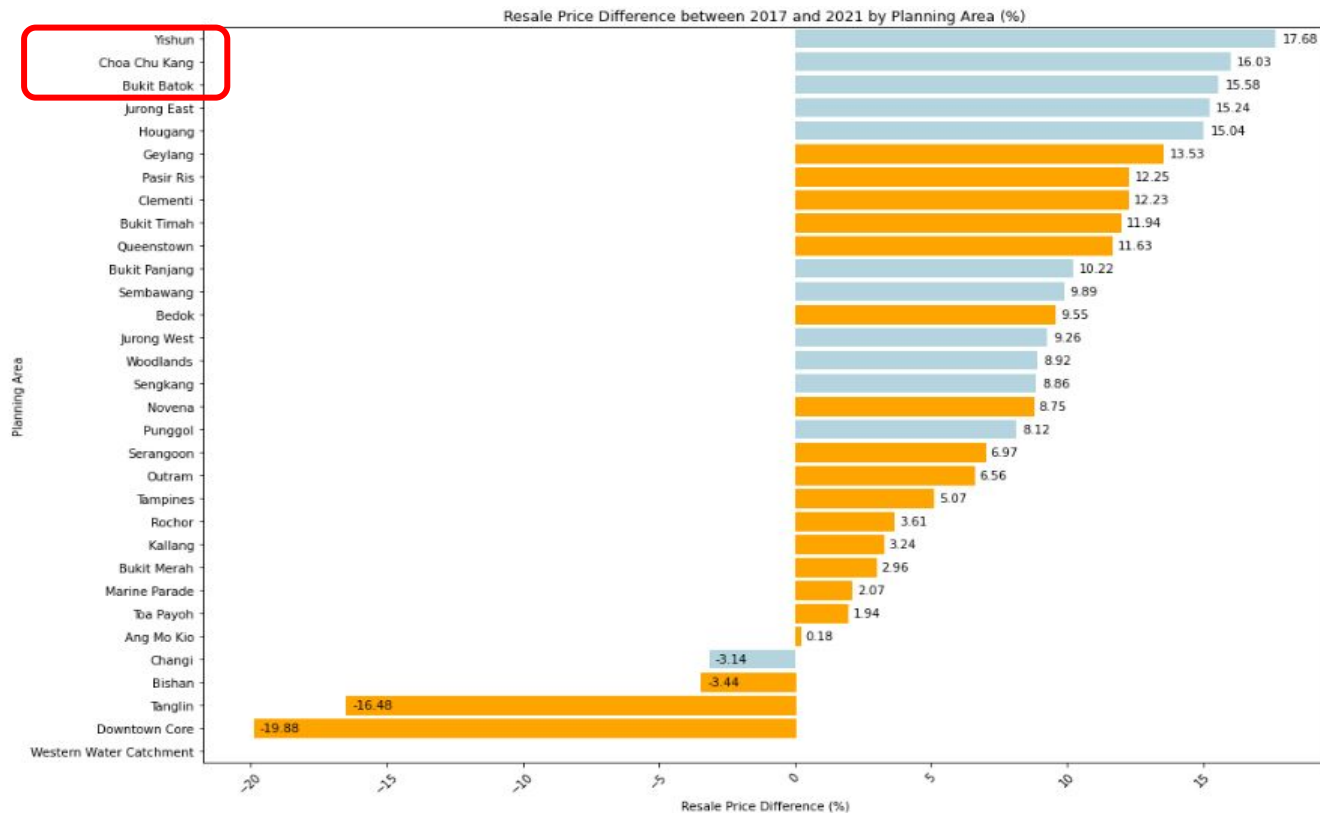


**Mature estate  
might not be  
the ideal  
investment  
location**



# Top 3 Planning Area

None of the top 3 planning areas belong to either Central region or Mature estate



**Yishun**  
**17.68%**



**Choa Chu  
Kang**  
**16.03%**



**Bukit Batok**  
**15.58%**

# Top 3 Planning Area : Commonality

None of the top 3 planning areas belong to either Central region or Mature estate



**Yishun**  
**17.68%**



**Choa Chu  
Kang**  
**16.03%**



**Bukit Batok**  
**15.58%**



# Data Cleaning and Feature Engineering

# Features Available

- 75 features in the dataset - 4 main categories

## Transaction Details

- Transaction Date
- Transaction Price

## Description of flat unit

- Storey
- Size of unit
- Type of unit

## Description of neighbourhood

- Amenities (Mall, Hawker Centre)
  - Distance to nearest amenity
  - Count within 500m, 1km and 2km
- Transport (MRT, bus stops)
  - Distance to nearest transport
  - Name of stop
  - Latitude and Longitude
- Schools (Primary, Secondary)
  - Distance to nearest school
  - School Name
  - School Affiliation
  - Latitude and Longitude

## Description of flat block

- Location
  - Exact location
  - Neighbourhood
- Age
- Max floor level
- Total Dwelling Units
- Type of flat
- Presence of other amenities in the same block
- Count of units sold in the same block, by unit type
- Count of rental units in the same block, by unit type

# Data Cleaning

- 75 features in the dataset - 4 main categories

## Transaction Details

- Transaction Date
- Transaction Price

- Drop Tranc\_YearMonth
  - Only used for transactions that occurred in the specific **month** and **year**
- Treat **Tranc\_Year** and **Tranc\_Month** as **categorical** variables
  - Represent the **macro conditions** in those years and months (eg. **interest rate**, **inflation rate**, **cooling measures**)
  - Remove the assumption that year and month has a linear relationship with resale price

# Data Cleaning

- 75 features in the dataset - 4 main categories

## **Description of flat unit**

- Unit Storey
- Unit Size
- Unit Type

- Drop overlapping features
  - Floor area sqft vs floor area sqm
  - Flat type and flat model vs full flat type
  - Storey range vs lower, mid and upper
- Avoid running into high multicollinearity

# Data Cleaning

- 75 features in the dataset - 4 main categories

## Description of flat block

- Location
  - Exact location
  - Neighbourhood
- Age
- Max floor level
- Total Dwelling Units
- Type of flat
- Presence of other amenities in the same block
- Count of units sold in the same block, by unit type
- Count of rental units in the same block, by unit type

- Drop Exact Location
  - Too specific as a feature - will lead to overfitting
- Summarise numerical variables
  - Putting meaning to numbers eg. percent of rental units in the block, percent of 4 room and above units



# Data Cleaning

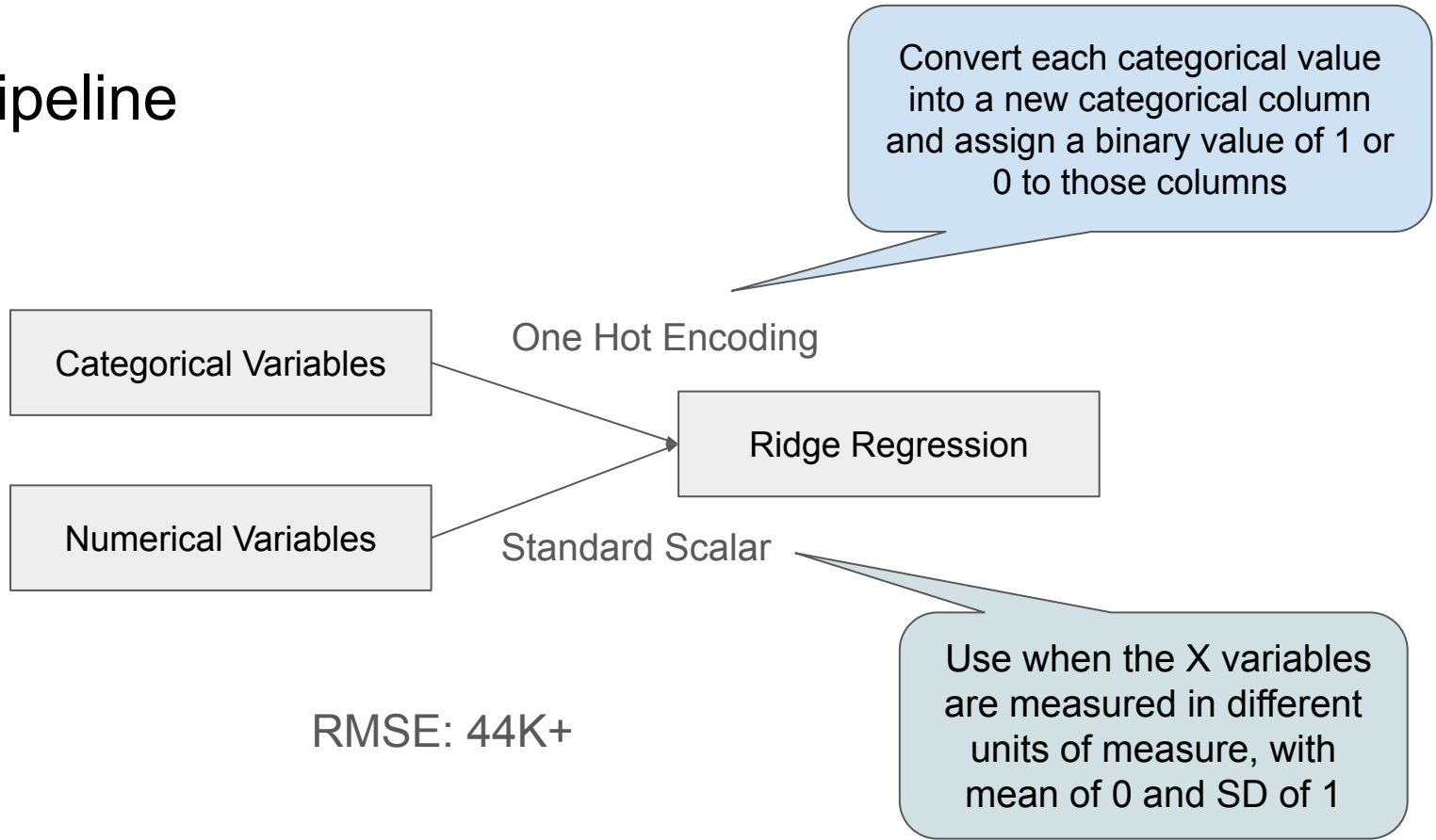
- 75 features in the dataset - 4 main categories

## **Description of neighbourhood**

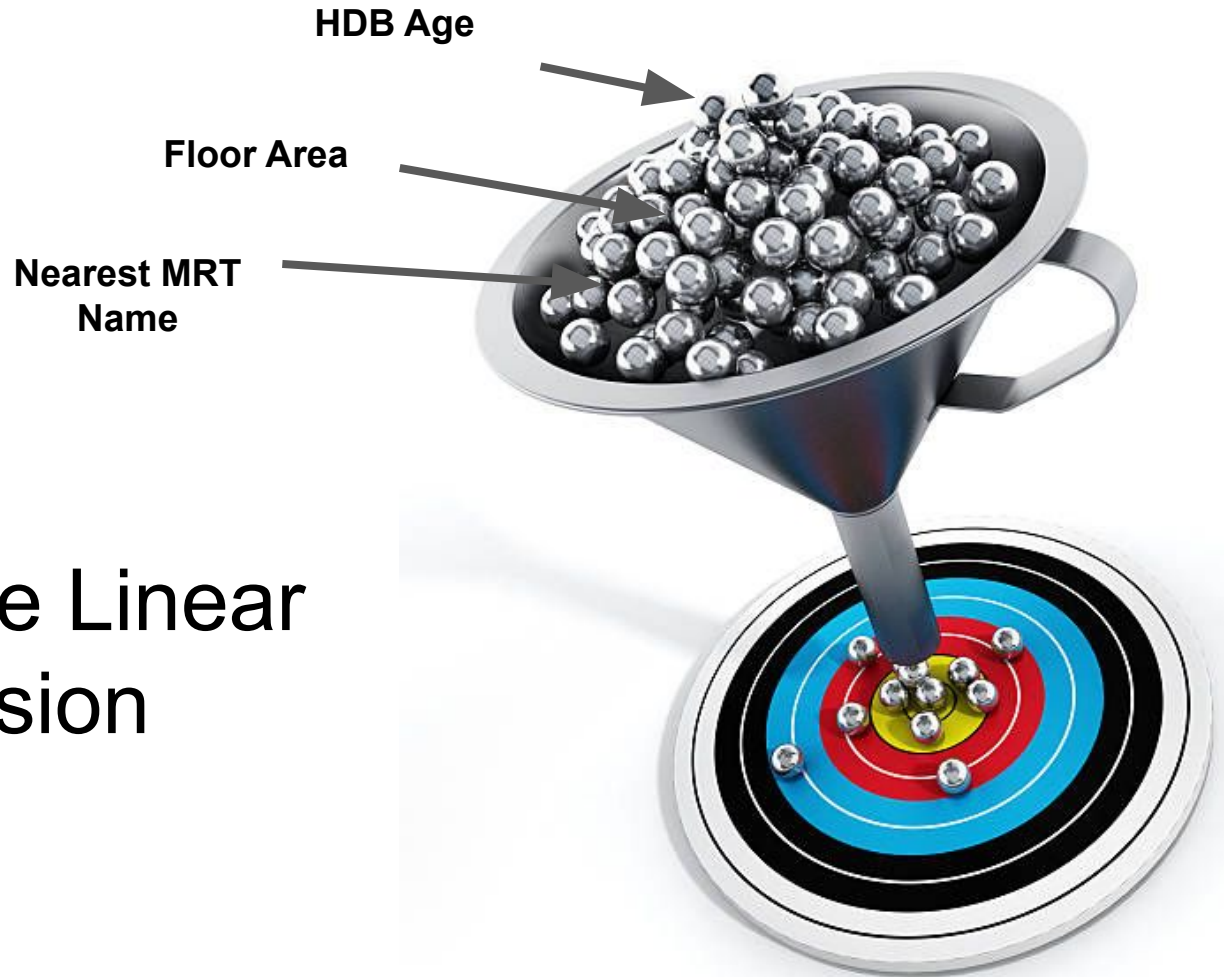
- Amenities (Mall, Hawker Centre)
  - Distance to nearest amenity
  - Count within 500m, 1km and 2km
- Transport (MRT, bus stops)
  - Distance to nearest transport
  - Name of stop
  - Latitude and Longitude
- Schools (Primary, Secondary)
  - Distance to nearest school
  - School Name
  - School Affiliation
  - Latitude and Longitude

- Combine numerical variables (interaction term)
  - Join 2 weak variables into a strong variable  
eg. distance to nearest mall and distance to nearest MRT
- Categorise names
  - Putting meaning to names eg. top 10 primary schools

# Model Pipeline

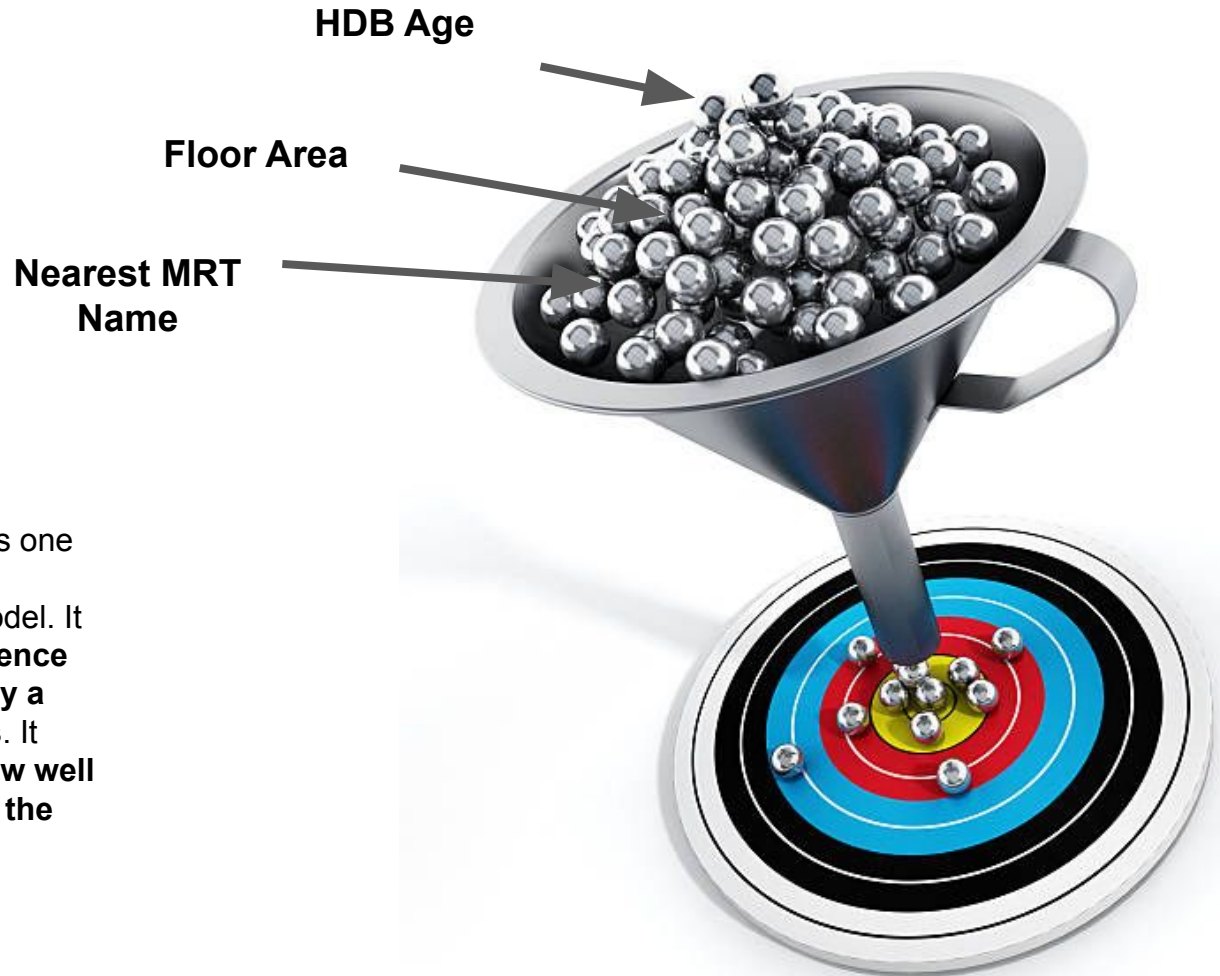


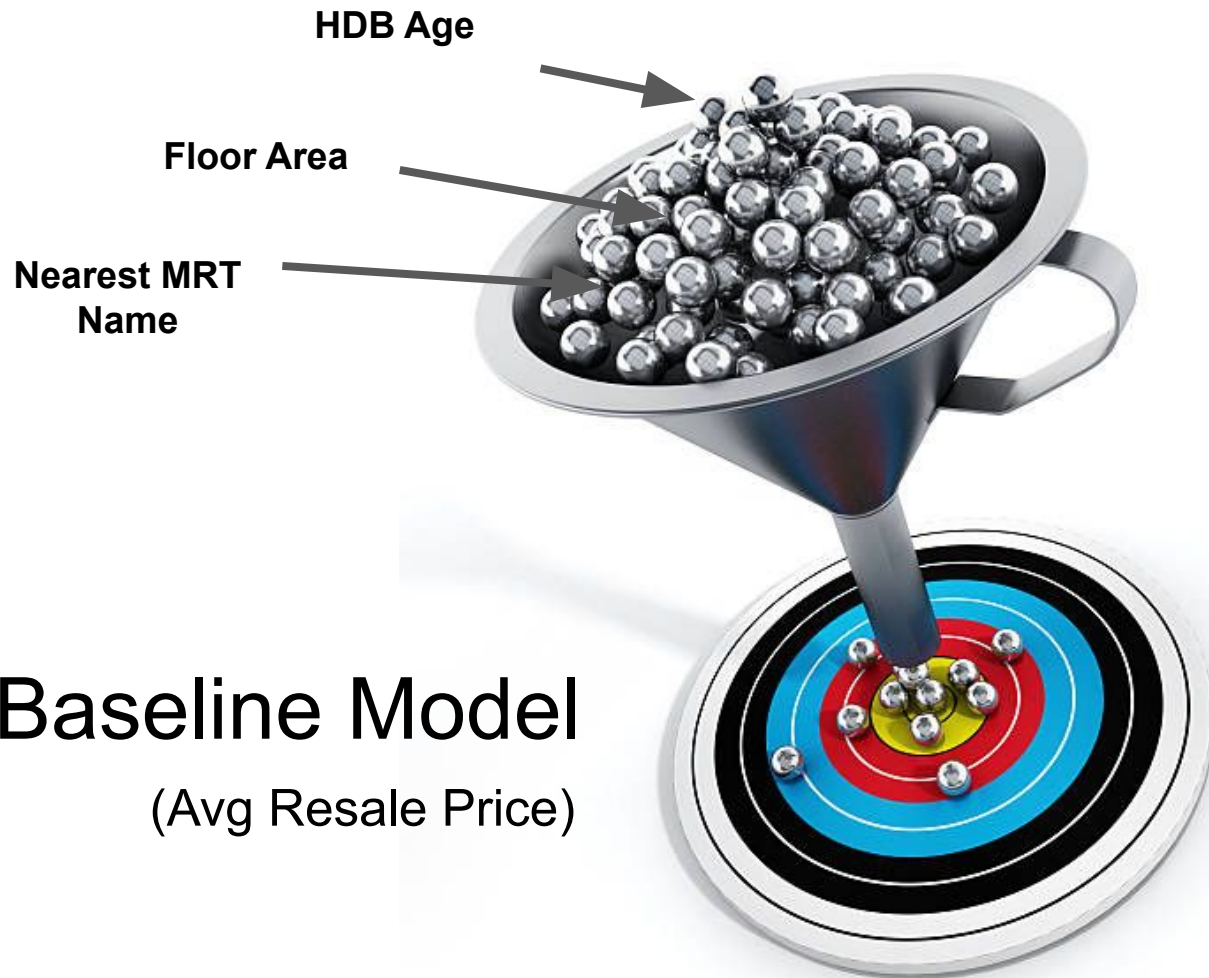
# Multivariable Linear Regression



# RMSE

**Root Mean Squared Error** is one of the two main performance indicators for a regression model. It measures the **average difference between values predicted by a model and the actual values**. It provides an **estimation of how well the model is able to predict the resale price**





**RMSE: 44K**



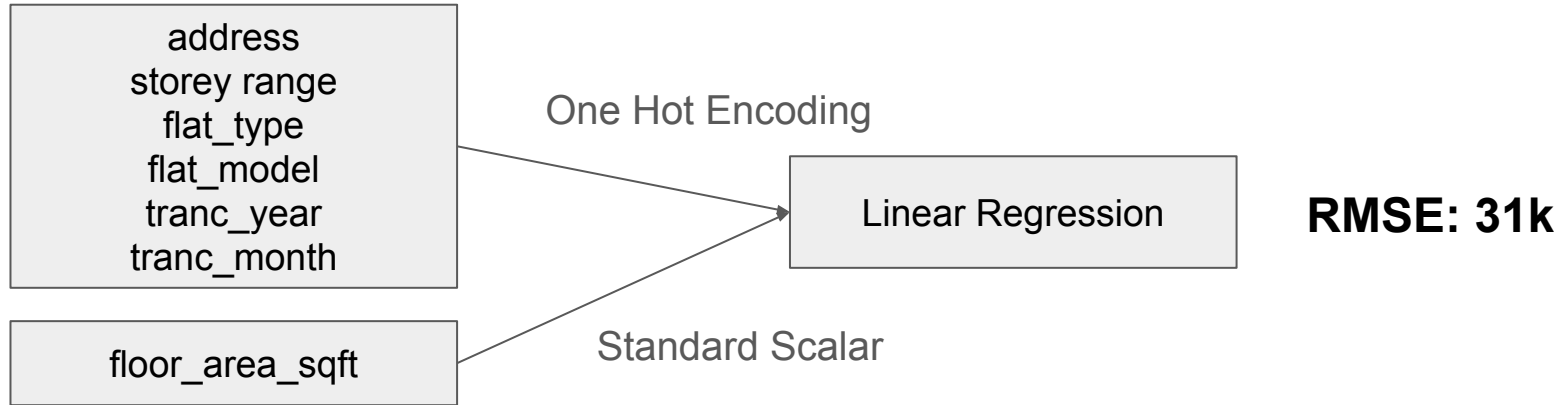
**210% vs Baseline Model**

(Avg Resale Price)

And this is not all, we had a surprising  
result...

## 2nd Best Performing Model

The 2nd best performing model is to include the exact location by using the “address” feature and flat unit features and transaction date features



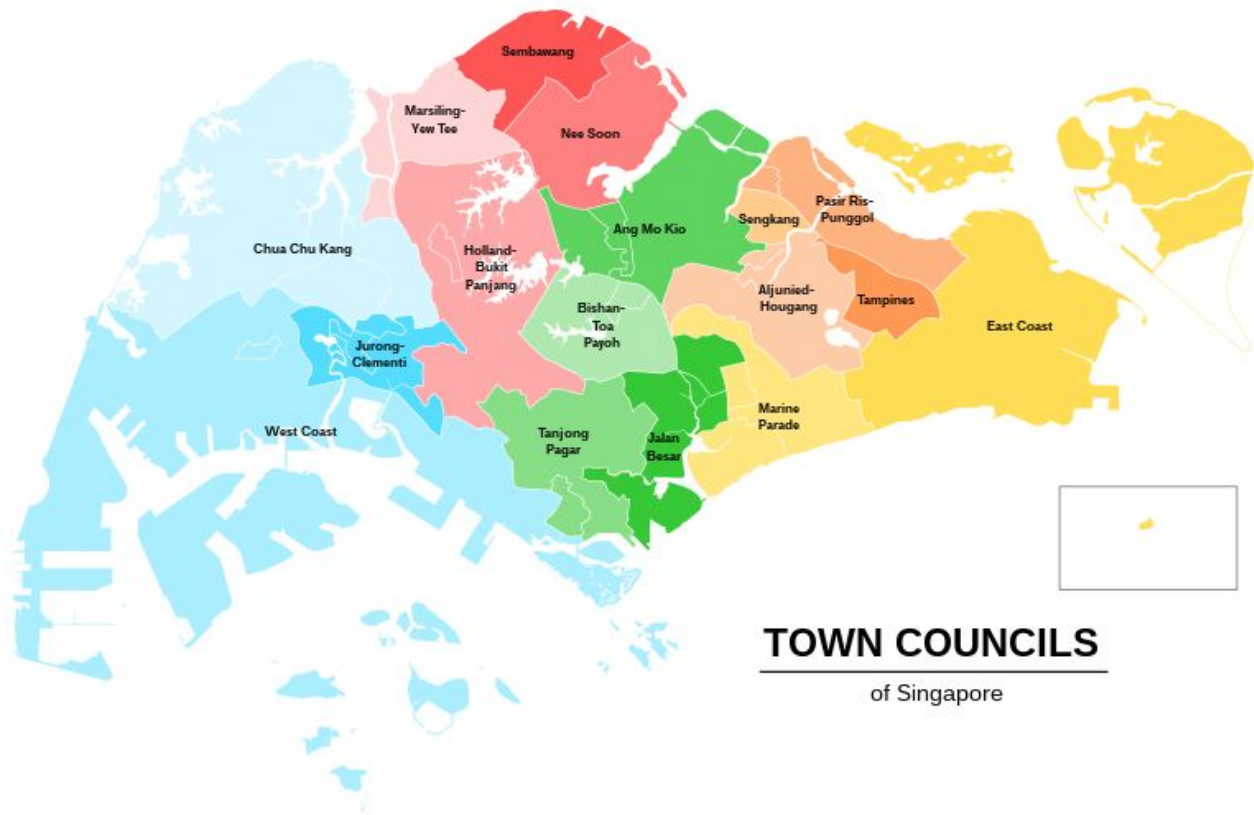
But why? As noted previously, the exact location represents flat block features (eg. age, rental units etc) as well as neighbourhood features (eg. amenities, transport, schools).

However, this model has low explainability.

How do we replicate the 2nd best performing model without using the exact location?

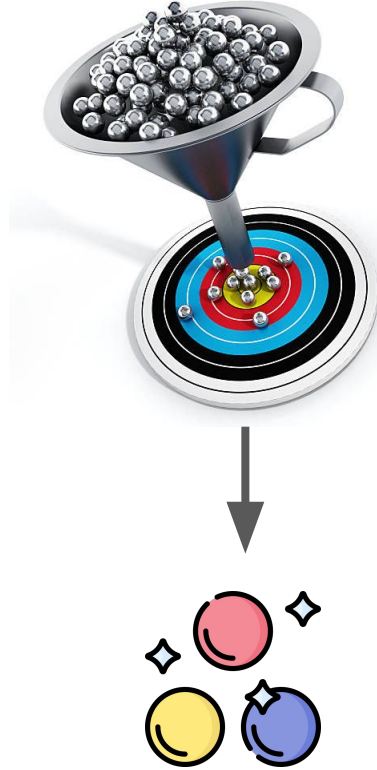




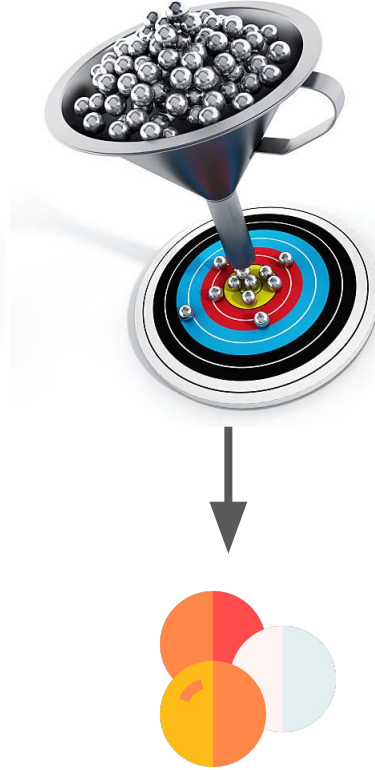


# Splitting By Town

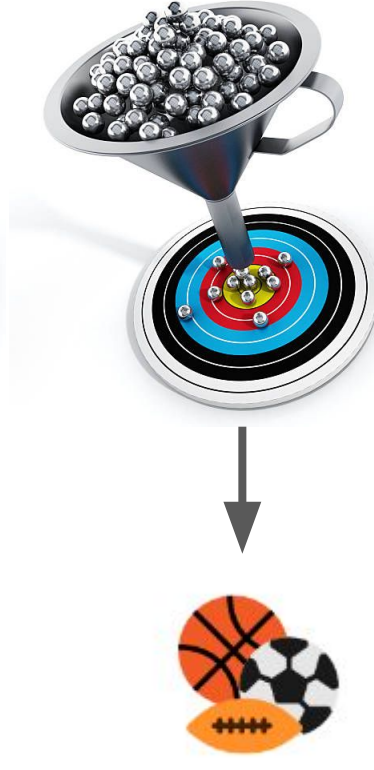
**Kallang/Whampoa**



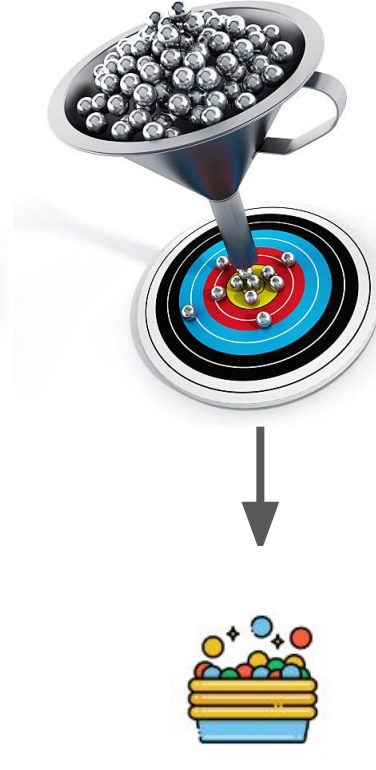
**Yishun**



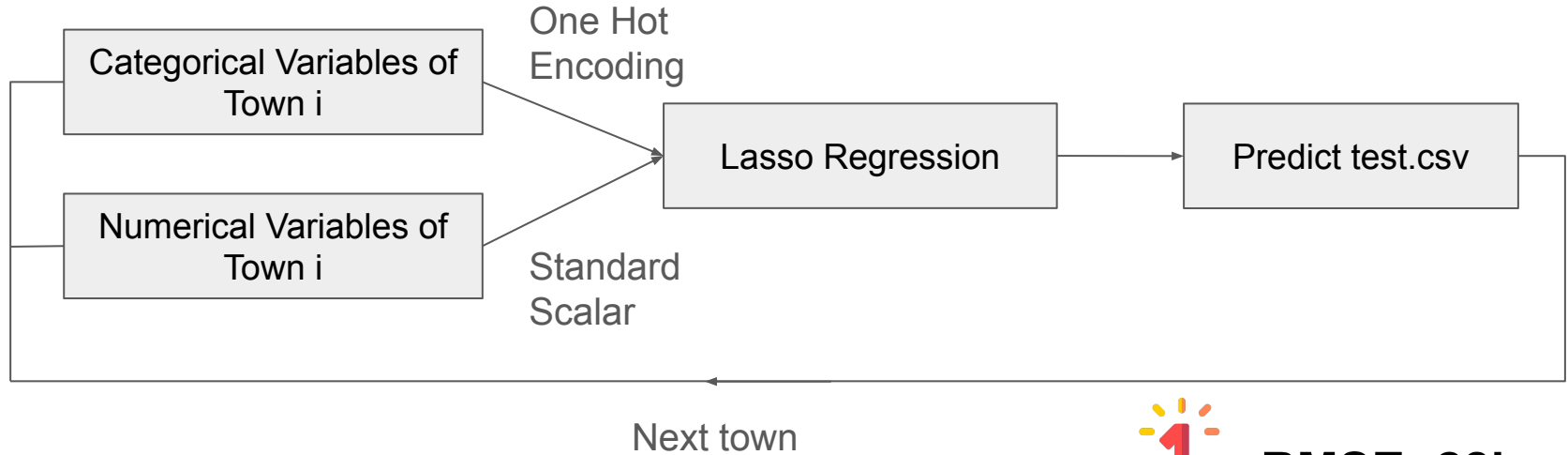
**Ang Mo Kio**



**Model: n**



# 1st Best Performing Model Pipeline



**RMSE: 28k**

# Model Comparison : Summary



**> 1 Models: Each town has  
its own model**  
RMSE: 28K

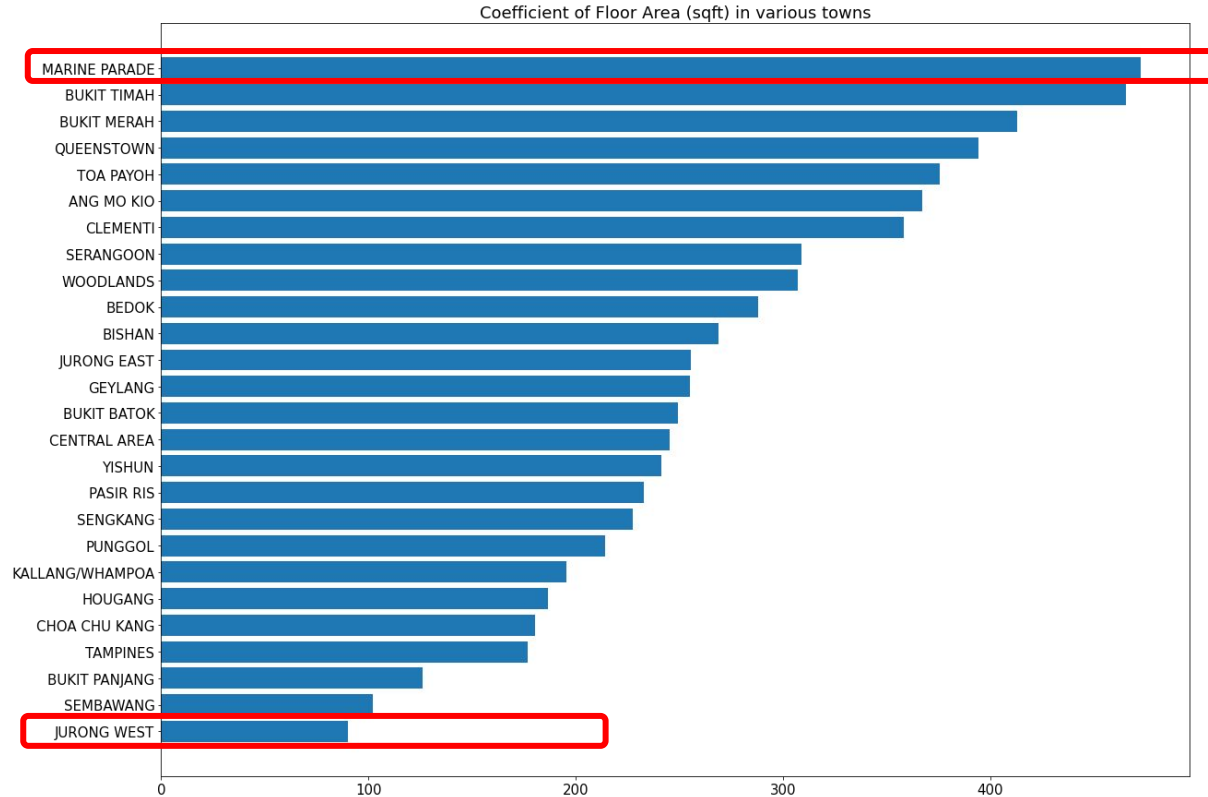


**1 Model - Include Address**  
RMSE: 31K



**1 Model - Use Other location  
variables**  
RMSE: 44K

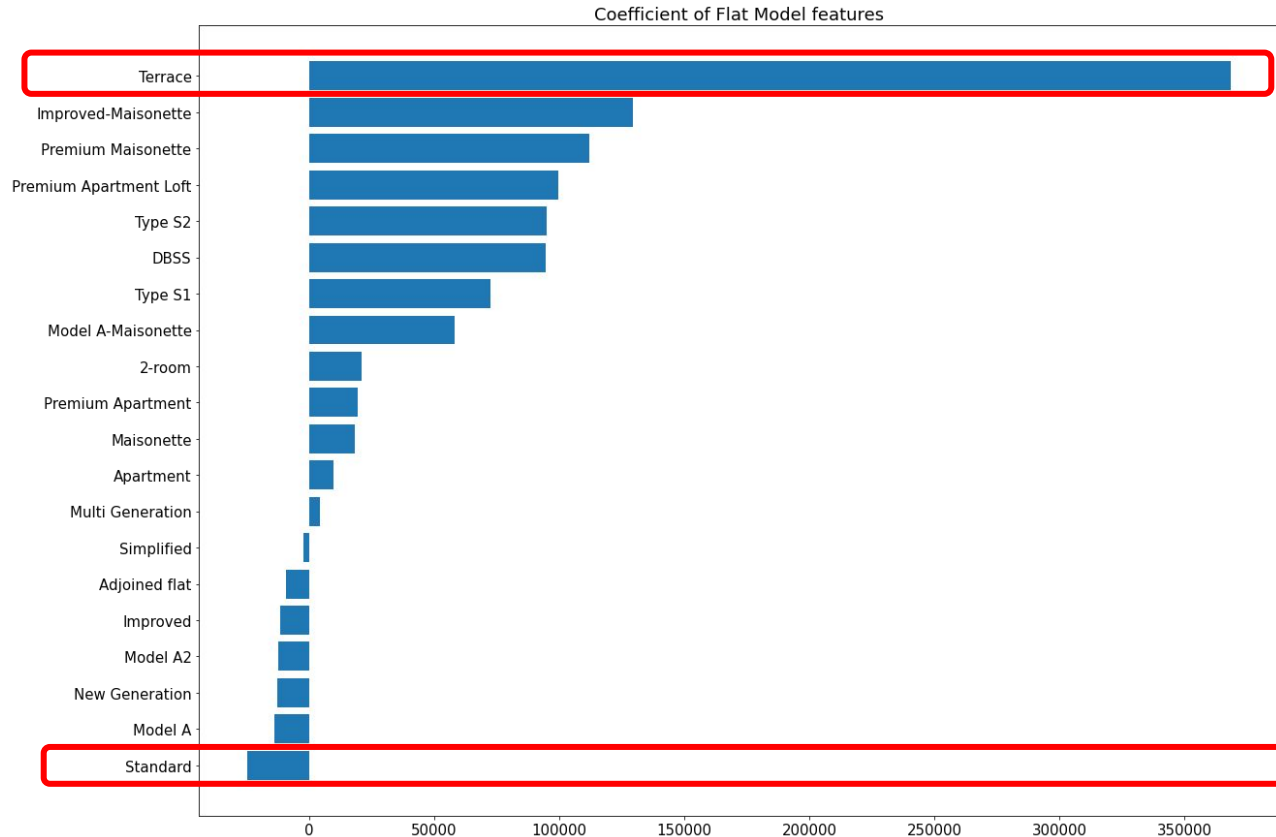
# Most Important Feature: Floor area (sqft)



Floor area (sqft) is most important in Marine Parade where every 1sqft **increases** the resale price by **\$472**.

Interestingly, this is even though the flats in Marine Parade are not the biggest (915sqft) - they are in fact lower than average (1046 sqft).

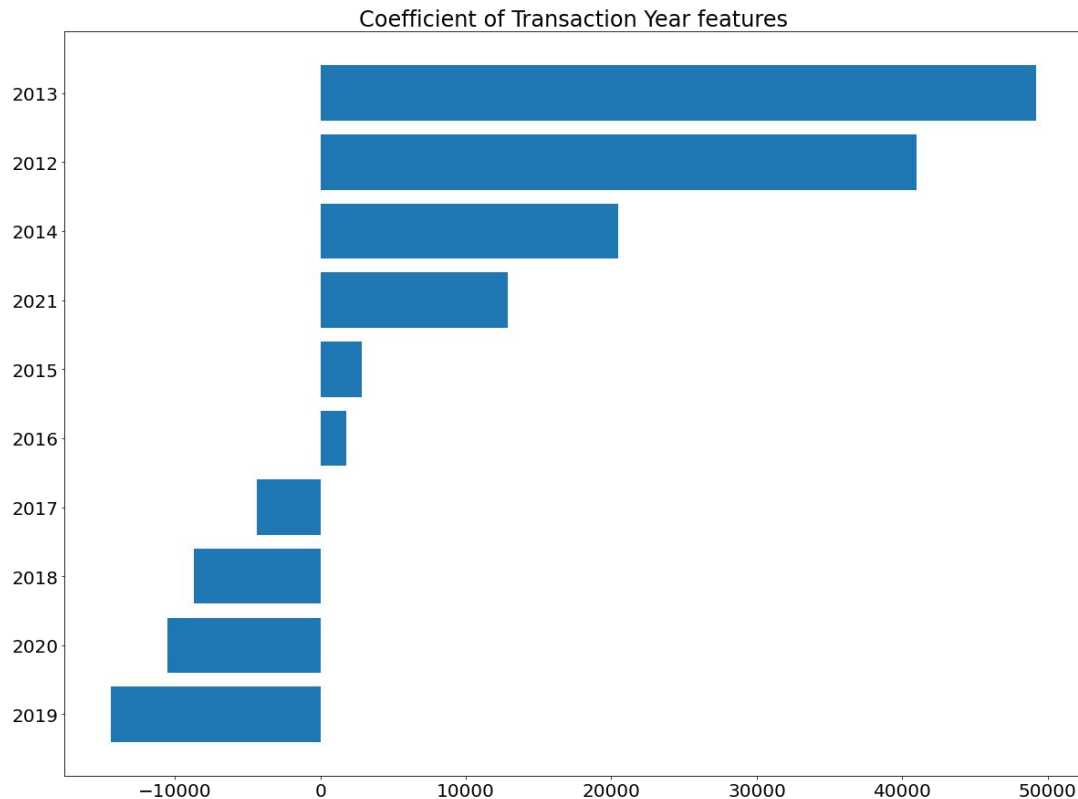
## 2nd Most Importance Feature: Flat Model



Having a **Terrace** flat **increases** the resale price by **\$350,000**.

On the contrary, having a **Standard** flat **penalises** the resale price by **\$24,000**.

## 3rd Most Important Feature: Transaction Year



As mentioned earlier, transaction **year** is treated as a **categorical variable** as it represents the **macro situation** in Singapore's housing market.

- **2012 - Low interest rates**
- **2013: Cooling Measures** were implemented throughout - observe a 2 year time-effect lag in **2014** and
- **2015 - US Fed started raising interest rates**
- **2020 - COVID** and continued to **2021**

# Limitation

## Accuracy vs Interpretation

- Include **Year** and **Month** into the model to **improve accuracy**
  - Not optimal for predicting transactions occurring beyond the training period
- Each **Town** has its own model to **improve accuracy**
  - Not optimal for predicting transactions occur in new town area
- Include **Address** into the model has the **2nd lowest score** but it is **not easy to interpret**
- Use **Standard Scaler** than Power Transform as it is **easier to interpret**
- Due to **time** and **GPU** constraints, we could not perform hyperparameter tuning on our models



# Conclusion

- **Best areas : Yishun, Choa Chu Kang and Bukit Batok**
  - Best return in the past 5 years
  - New amenities and new MRT stations
- **Central region and mature area** is not the best areas
- **Top 3 Factors**
  1. Floor Area
  2. Flat Model
  3. Transaction Year
- **Factors shaping the housing market**
  1. Government's measures, frameworks, and policies
  2. Inflation and loan rates
  3. Pandemic, such as COVID-19

**THE  
END**

A stylized graphic of a film strip, consisting of two rectangular frames stacked vertically, with short horizontal lines representing sprocket holes on the left and right sides. The graphic is positioned behind the text 'THE END'.