HDB Resale
Analysis

Our homes

Vision

An outstanding organisation creating endearing homes all are proud of

STRATEGY

Mission

We provide **affordable, quality housing** and a **great living environment** where communities thrive

Services

Provide a **smooth** service for homebuyers to facilitate the best allocation

Tools

Establish internal **data tools** within HDB to make operations efficient

Inform

Enable **data-driven** decision making quickly

TABLE OF CONTENTS

01
Data
Data Wrangling

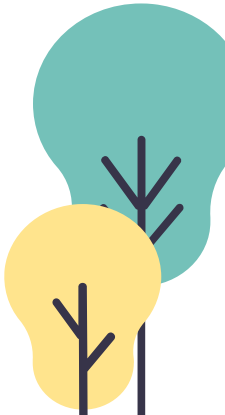
02
Dashboard
Dashboards to serve the homebuyer

03
Models
Get an estimate for your house simply. Understand the data in a new dimension

04
Analysis
Make data-driven decisions

05
Conclusion
Final comments

06
Q&A
Questions?



Data Sources

DATA PIPELINE Raw Sources

data.gov.sg

Data is taken from data.gov

OneMap, MapQuest, geopy

Location and distance data

COE

Vehicle COE data, https://coe.sgcharts.com

DATA PIPELINE conditioning

Data APIs

1990 – 1999

2000 – Feb 2012

Mar 2012 – Dec 2014

Jan 2015 onwards

GoogleV3, Nominatim

Compile

Combine like columns, fill missing data, reformat data

Now for some analysis

Exploratory Data Analysis

Data types

- Dates are case to pandas time for easy manipulation
- Average storey is instead of a range
- Transaction month and year are split to columns
- Address feature is made for easy identification in future
- _id is dropped

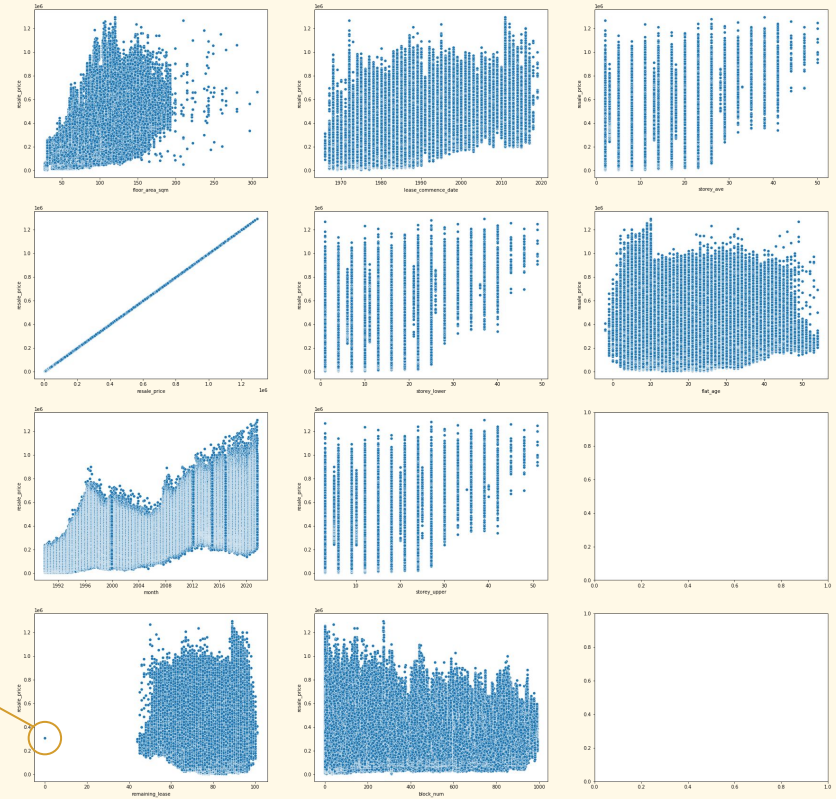
Missing Values

Many of the remaining_lease had to be managed to turn to years and for those that are missing it is computed based on below

Remaining_lease = 99 + commence_date - transaction_year

Outliers

Only one outlier is noted in the remaining_lease but is imputed based on the above formula



Dashboards

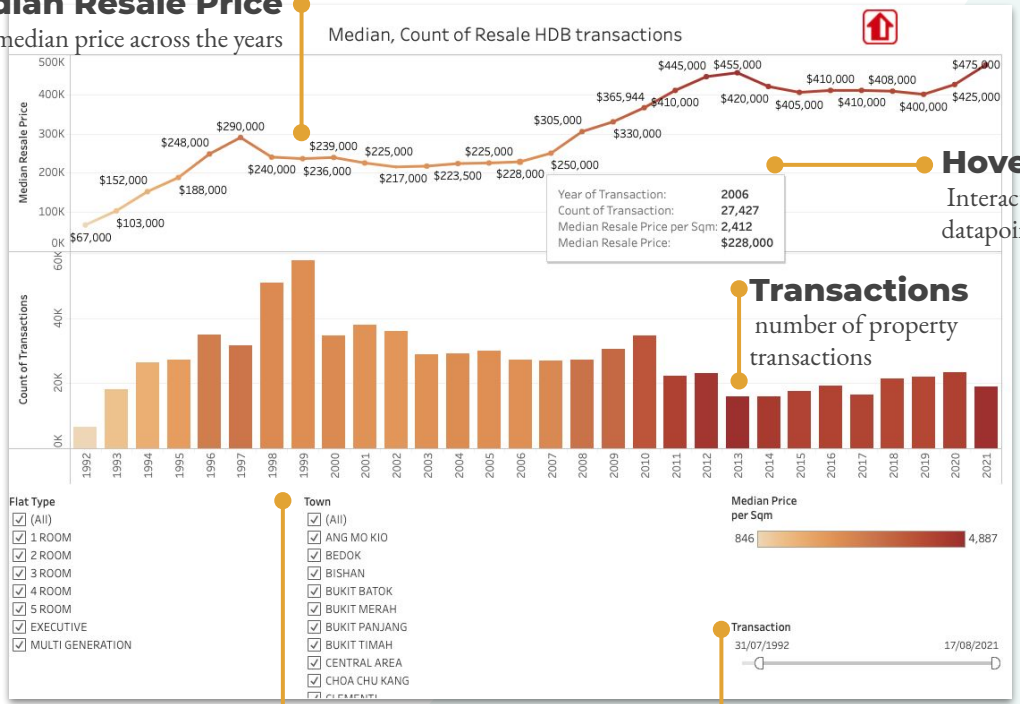
Created are targeted dashboards for resale HDB buyers

Please go to this site to see the dashboards and feel free to interact with them yourself
https://public.tableau.com/app/profile/julian.chang/viz/FirstDashboard_16315416189430
(go to 'First Dashboard')

UNDERSTAND National Trend of HDB Resale Prices

In tab `**HDB_Res_Nat**`, the trend over the years for the national median resale prices is plotted along side the number of transactions.

Median Resale Price
median price across the years



Hover tooltip
Interactive tooltip for each datapoint

Transactions
number of property transactions

Filter Type
filter based on Flat Type

Filter Town
filter based on town

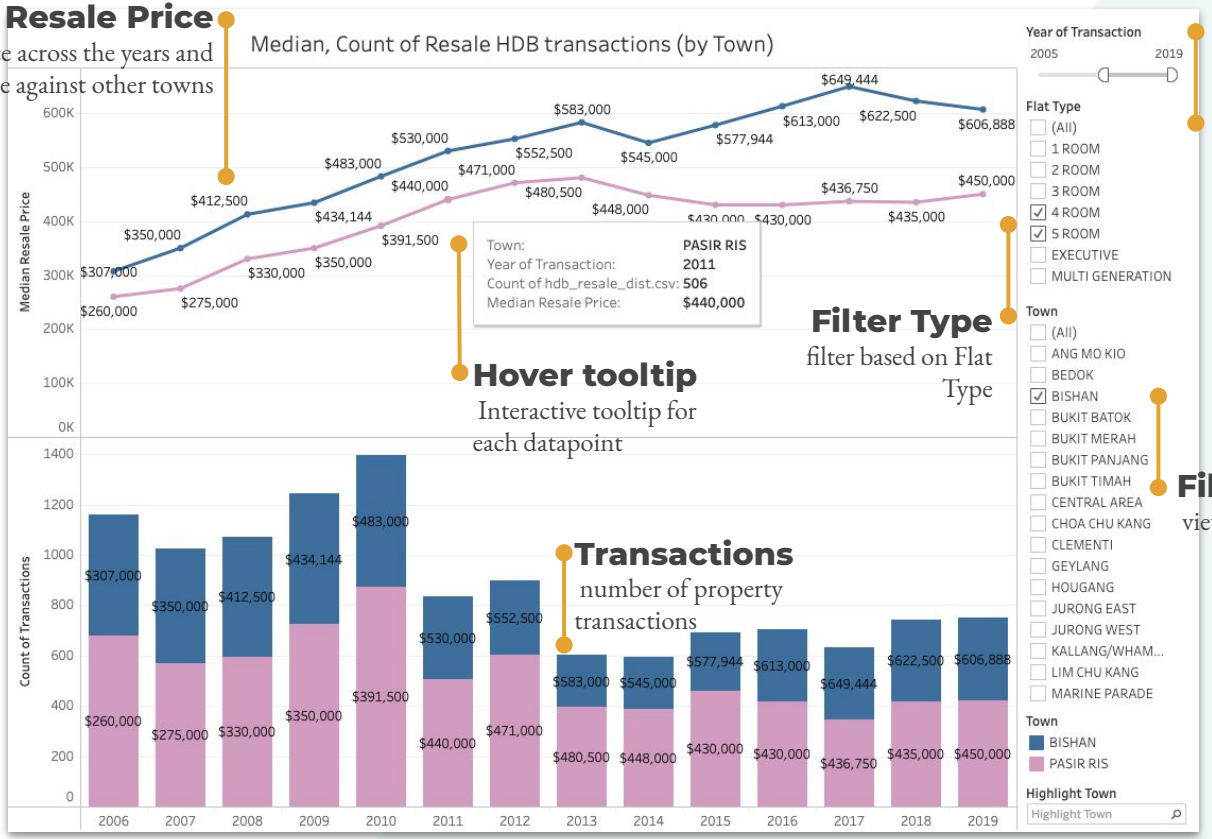
Filter Period
filter historical transaction date

FAMILIARISE with the local Trend of HDB Resale Prices

Then in tab `**HDB_Res_Town**` another tab shows the trend for each town.

Both allow for a slider to choose the the view with the year slider.

Median Resale Price
median price across the years and compare against other towns



Filter Period
filter historical transaction date

Filter Type
filter based on Flat Type

Hover tooltip
Interactive tooltip for each datapoint

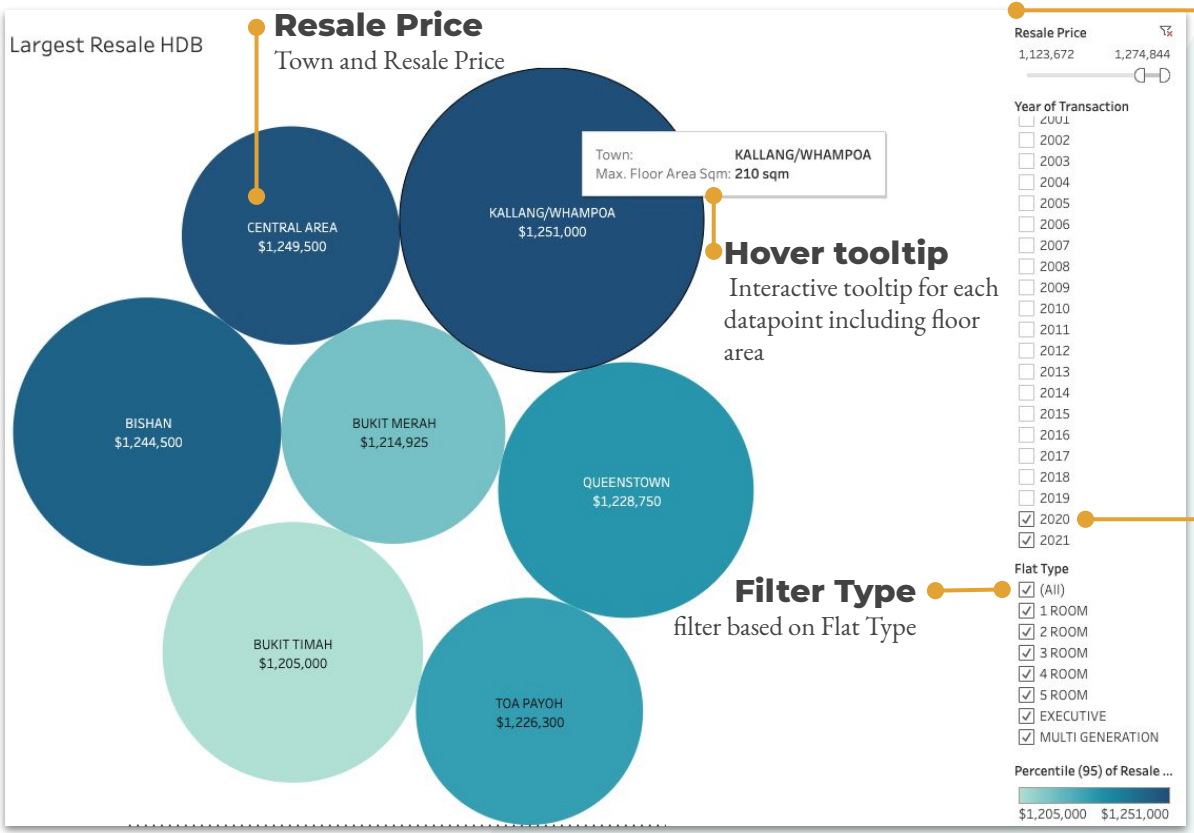
Filter Town
view by HDB towns

Transactions
number of property transactions

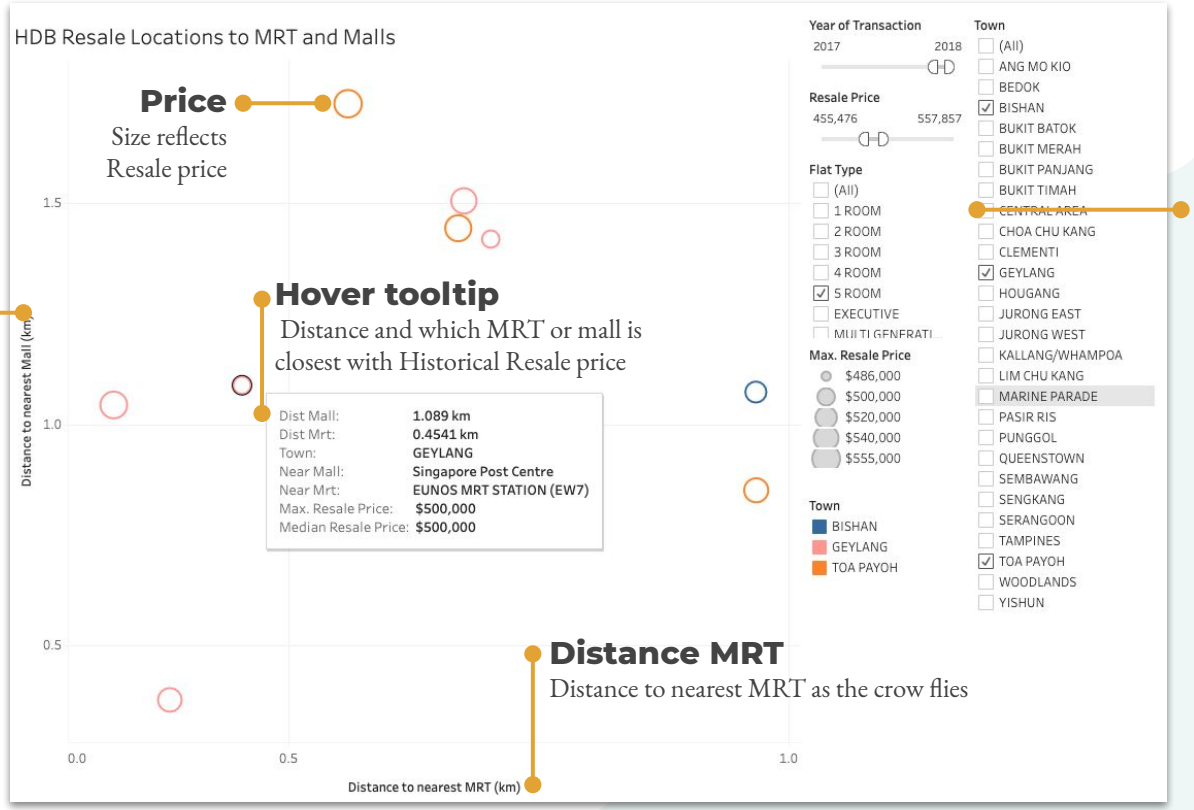
BUDGET for the Biggest Flat

In tab **`HDB_Res_Budget`** recommends the largest sold resale HDB.

largest apartment is largest bubble within the budget.



PREPARE and accommodate your lifestyle



Distance Mall
Distance to the mall as the crow flies

Filters
filter based on Year of Transaction, Resale Price, Flat Type, Town

Compiled locations with OneMap, MapQuest API the list.

Distance data is retrieved from geopy

Further work is possible to get driving distance with MapQuest

Models

Train and predict based on historical resale prices

© Darren Soh / Courtesy of Singapore Tourism Board

DATA SCIENCE APPROACH

DATA MINING AND CONDITIONING

Fix inconsistencies, formats and deal with missing values

EXPLORATORY DATA ANALYSIS

Form hypothesis, constraints and assumptions

FEATURE ENGINEERING

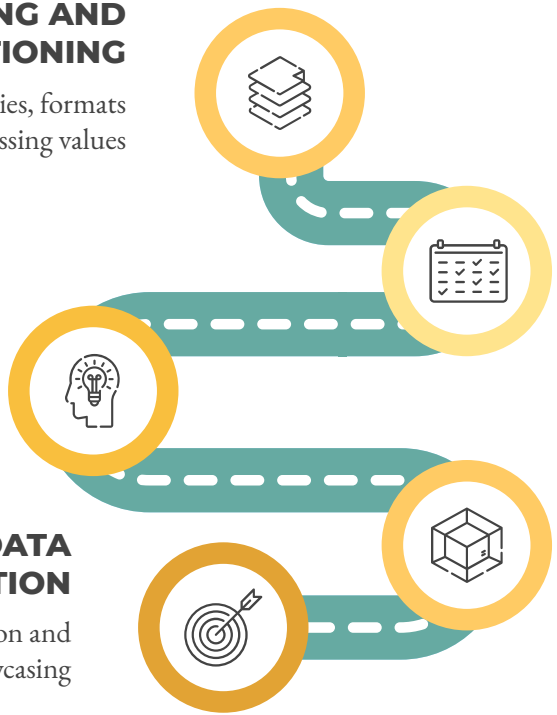
Adjust and condition data to answer the problem

PREDICTIVE MODELING

Train, evaluate and test data

DATA VISUALISATION

Explanation and showcasing



Predictive model 1

2014 Resale HDB Prices

Below features are used to make a simple and fast mode to predict 2014 HDB pricesl

- flat type
- flat age
- town.

Predictive Model 1 Feature Engineering

Best model to **predict resale HDB prices** accurately for 2014

- Features used: Flat type, Remaining Lease and town
  - remaining_lease is used instead of flat_age because it has a positive correlation so it is slightly easier to explain.
- 2014 transactions

Predictive Model 1 Training

Trialled models:

RMSE	Cross Validation Score	Test Score
OLS	$ 51,917	$ 53,436
Lasso	$ 52,265	$ 53,706
Random Forest	**$ 34,787**	$ 38,133

Best model chosen based on test Cross-Validation root mean squared error as **Random Forest**

root mean square error of $38,133

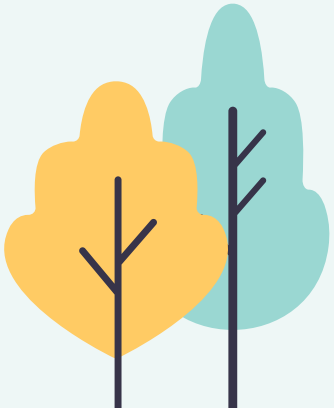
Predictive Model 1 Conclusion

Reasons for Random Forest

- **bagging aggregates** a close estimate of the relationships between coefficients
- **random subset selection**. although there are only 3 main features, there are dummified categorical features
- **Feature importance** allows for some model explanation

Improvements

- OLS was already underfit so Lasso did not improve the model,
- alternative more complex model may have even better results e.g. Random Forest with boosting



Predictive model 2

Unreasonable transaction price?

- Flat type: 4 ROOM

- Town: Yishun

- Flat Model: New Generation

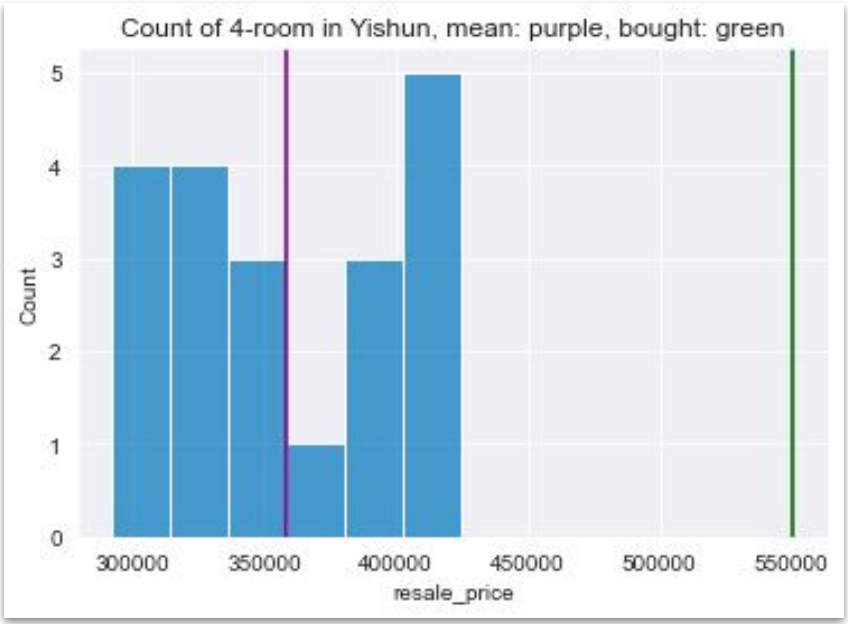
- Storey Range: 10 to 12

- Floor Area (sqm): 91

- Lease Commence Date: 1984

- Resale Price: 550,800

Predictive model 2 Exploratory Data Analysis



Plot historical transactions:

- 4-Room, Yishun, sold in 2017

Purple line is the mean historical transactions

Green line shows the proposed transaction ($ 550,800)

Predictive model 2 Feature Engineering

#1

Scaling

Remaining_lease, resale_price is scaled

#2

Features

Resale_price, Floor_area_sqm, Remaining_lease, Lease_commence_date, Block_num, Storey_ave, Transact_month, Transact_year, Town, Flat_type, Flat_model, street_name,

#3

Train-Test-Split

1:1 train to test set size with random state: 42

Predictive model 2 Random Forest Model

Model: Random forest on train dataset

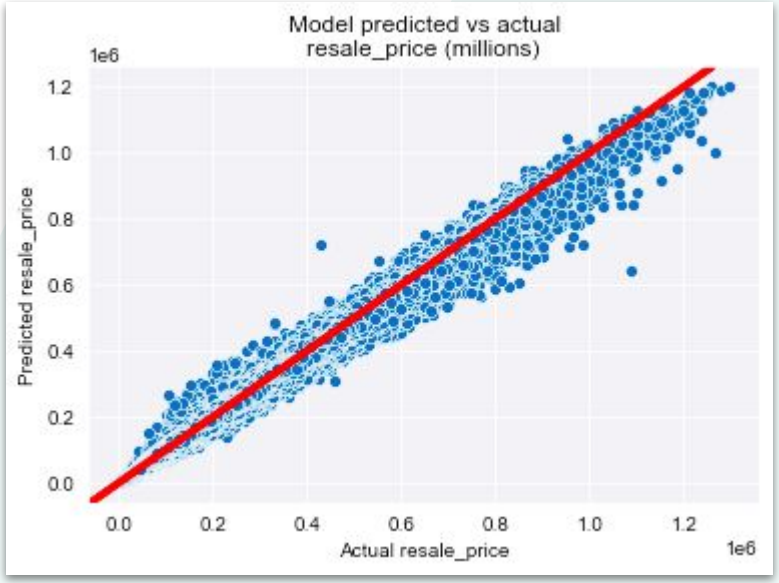
Initial train RMSE **$18,758**

Model re-trained with entire dataset

Final train RMSE **$16,975**

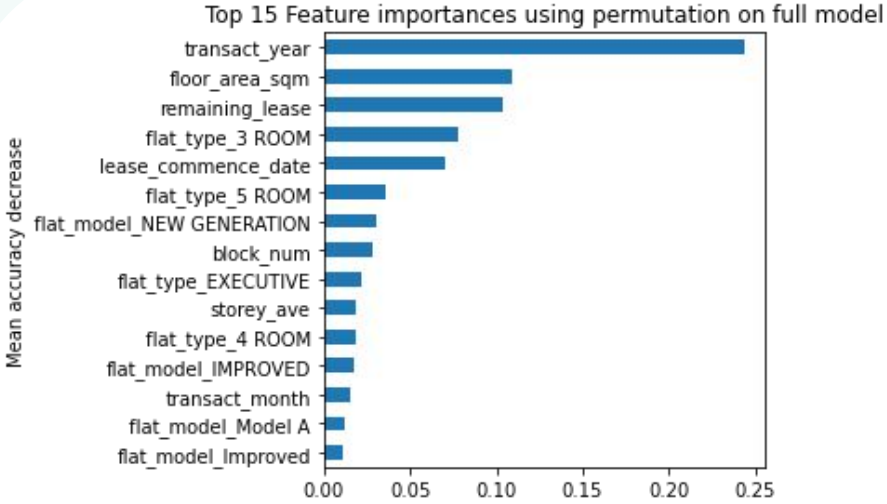
Based on the features given, model predicted **~ $458,000** and upper range of around **$491,950** with a high confidence,

Unusual to pay $550,800 for 4 ROOM HDB at Yishun in 2017



With regard to feature importance, we see that **transact year** plays a big role in resale price.

Based on $550,800 is in the population but not within the features given. Likely could be that the year it is transacted is actually in 2021 or later

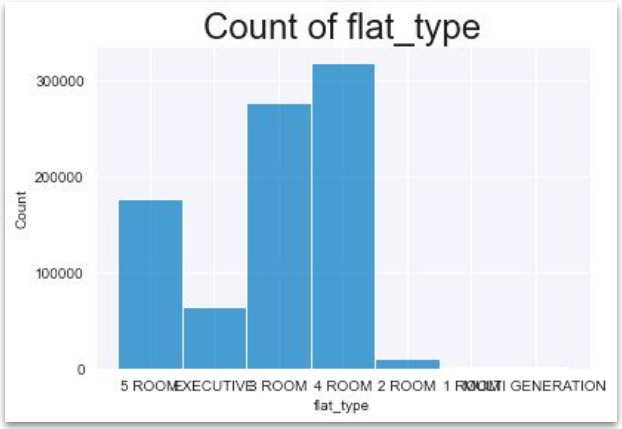


Predictive model 3

Supervised and Unsupervised Learning

predict flat type given a transaction's other characteristics

Predictive model 3 Exploratory Data Analysis

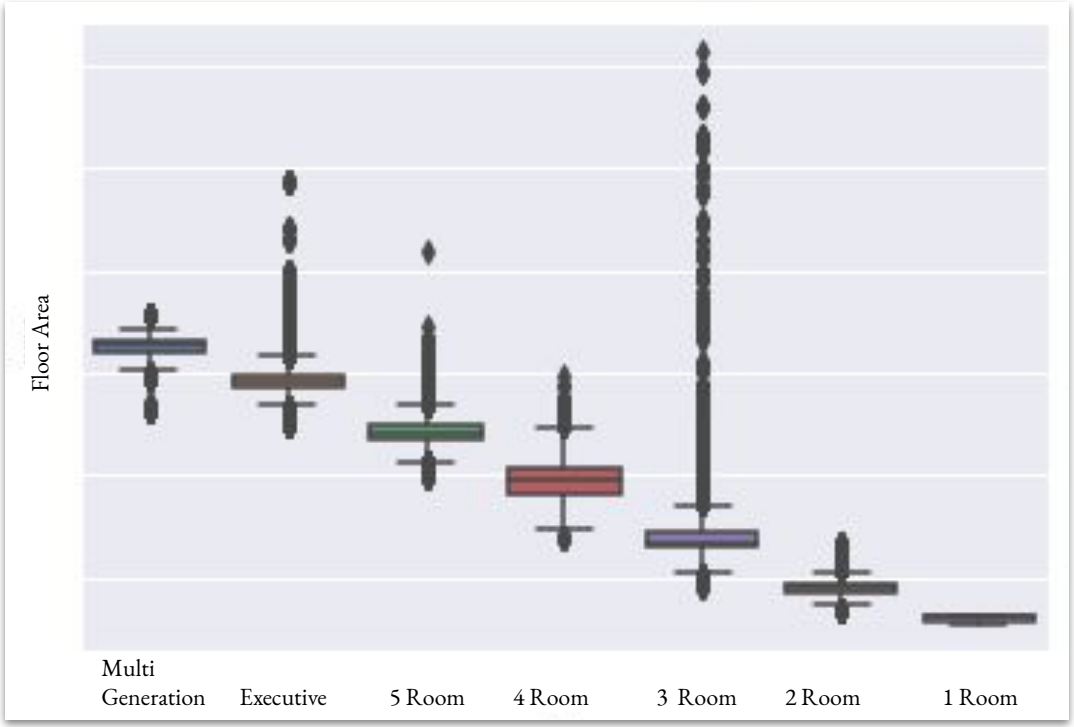


Data imbalance present

- 4 Room is the highest at 38%
- 3 room is the next highest at 33%
- 3 to 5 Room take up >90% of classifications

No adjustment except for metric is not just accuracy

Predictive model 3 Feature Engineering

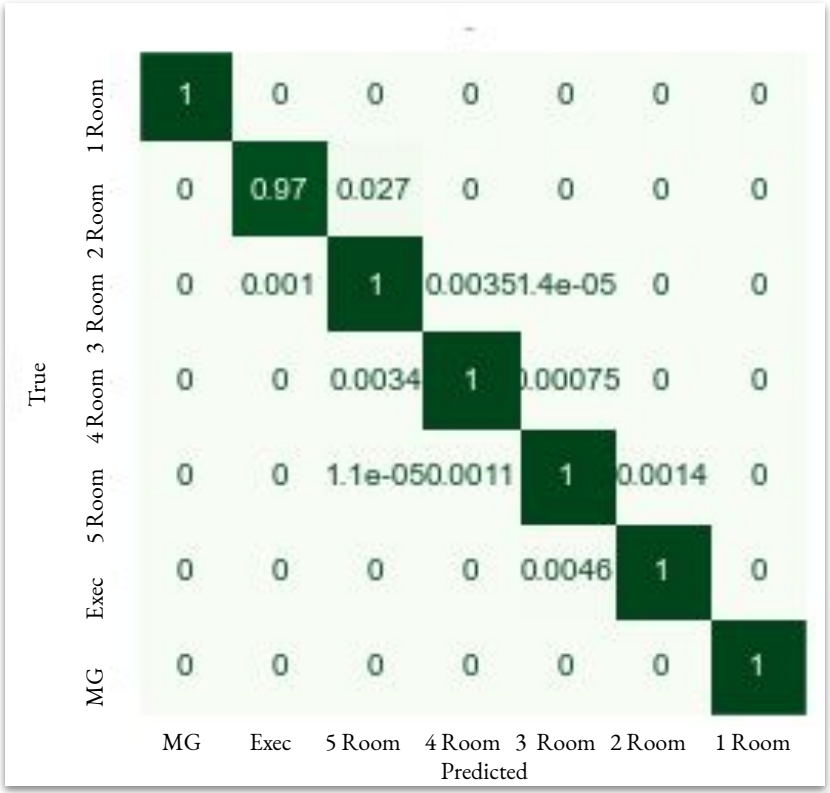


Floor Area would be a good metric to categorise the 7 flat_types

Features considered

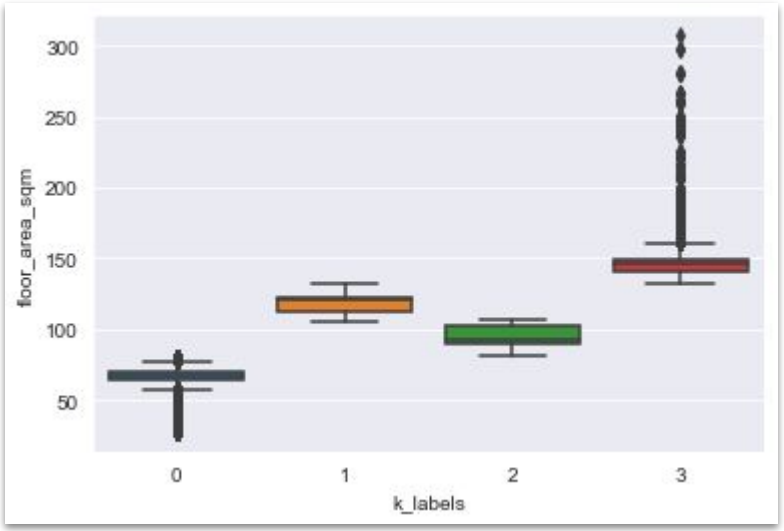
- flat_model
- resale_price
- flat_size_sqm
- storey_ave
- town

Predictive model 3 Supervised Classification



- **Supervised model** is created with Decision Tree Classifier
- Trained model has more than **99% high accuracy and specificity**
- But since all the data lost the model would cannot be trained

Predictive model 3 Unsupervised Classification



K mean cluster	Flat type classified	Sensitivity
0	1, 2, 3 Room	94%
1	5 Room	91%
2	4 Room	93%
3	5 Room, Exec, MG	93%

- **unsupervised model** is created with 4 clusters
- There is some success, as seen there has about **93% accuracy**
- But does not split to all flat types

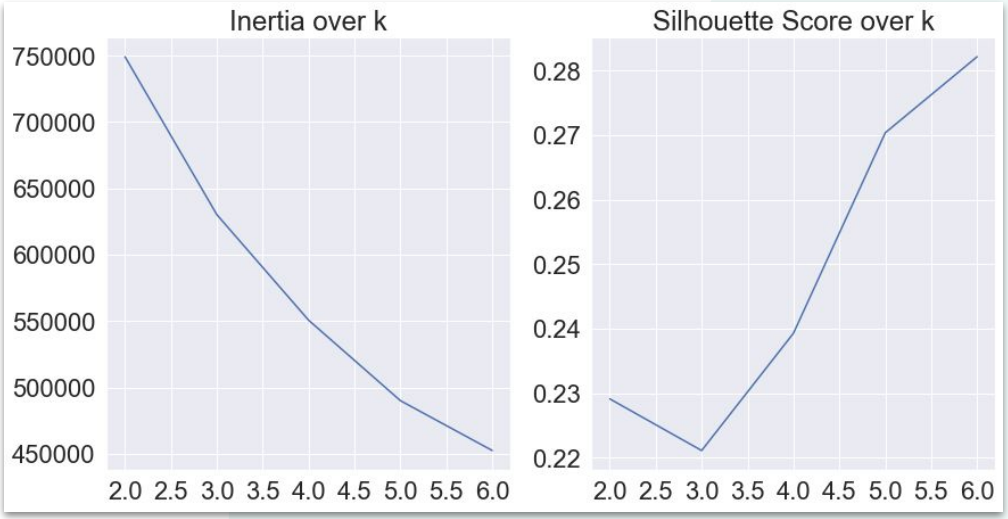
Predictive model 3 Model Tuning

- Low Silhouette and high Inertia is preferrable
- Trade-off between labeling time and predictive capability

3 groups would have the best accuracy

Improvements

- Compromise of 4 classes may be sufficient
- labeling a small sample and use semi-supervised model



Analysis

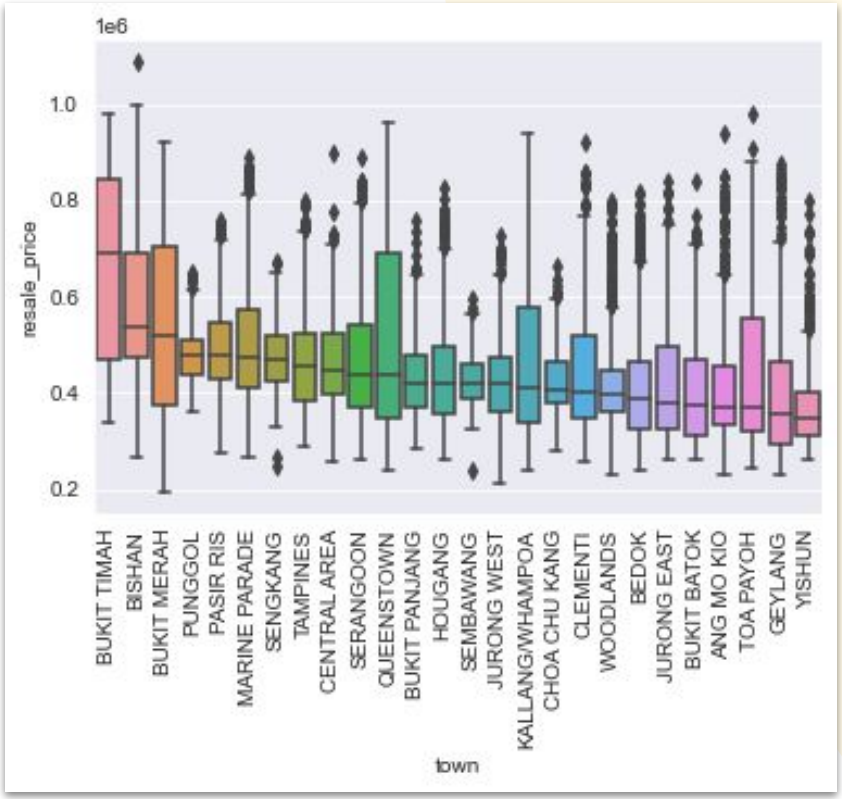
Data-driven decision making

Analysis 1

Is Yishun crazy because its cheap or cheap because residents are crazy?

Mothership.sg: Horse in Yishun Car Park

Analysis 1 Exploratory Data Analysis

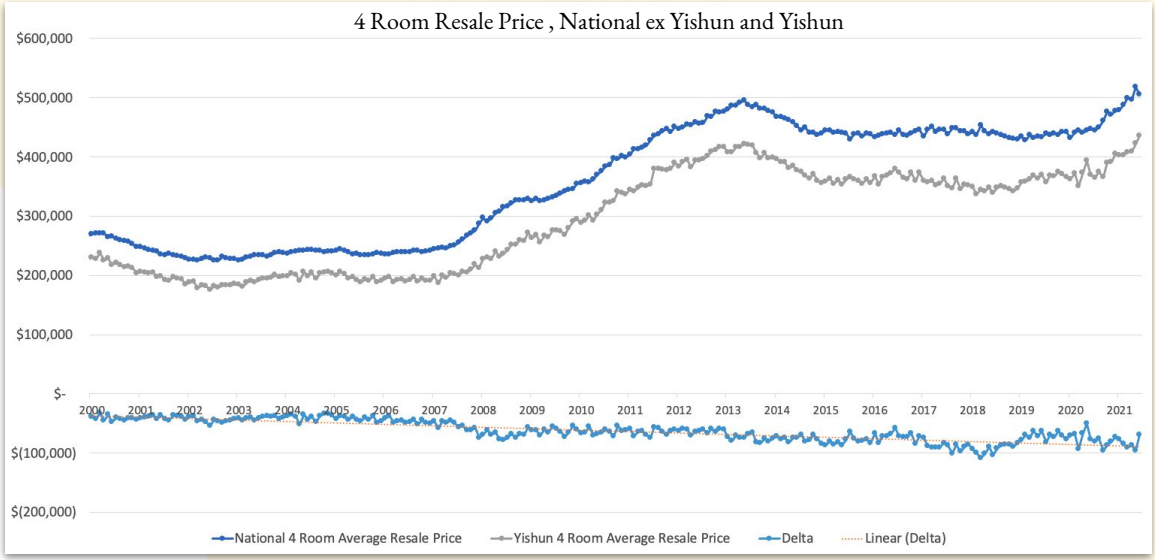


- Data is filtered to for only 4 ROOM flats as the base of comparison since it is the most popular flat type in yishun and nationwide
- Yishun has the lowest median income

Analysis 1 Deltas over time

- To inspect the change over time.
- 4 Room HDB mean resale price of Yishun and the rest of Singapore over time is plot
- Delta is computed (light blue)

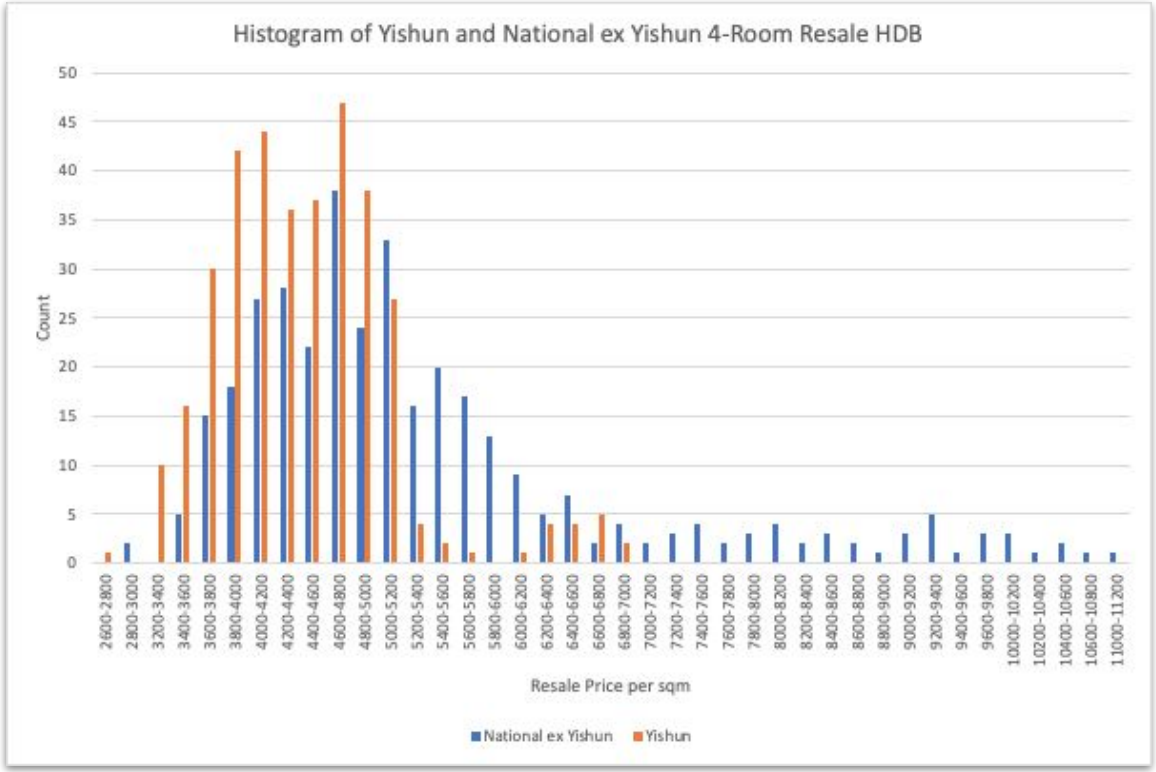
Yishun resale price is consistently less than its peers and decreasing. Difference is **doubling** every **10 years**



Analysis Hypothesis Test

As a snapshot of all **4 Room** resale HDB transactions back from start of **2021 to 13 Sep 2021**

National ex Yishun is resampled randomly to match the **Yishun** sample



Analysis 1 Yishun is less than the rest

t-test
Welches (non-paired sd) two-tailed performed

p value ~ 0.0001
p value is extremely significant

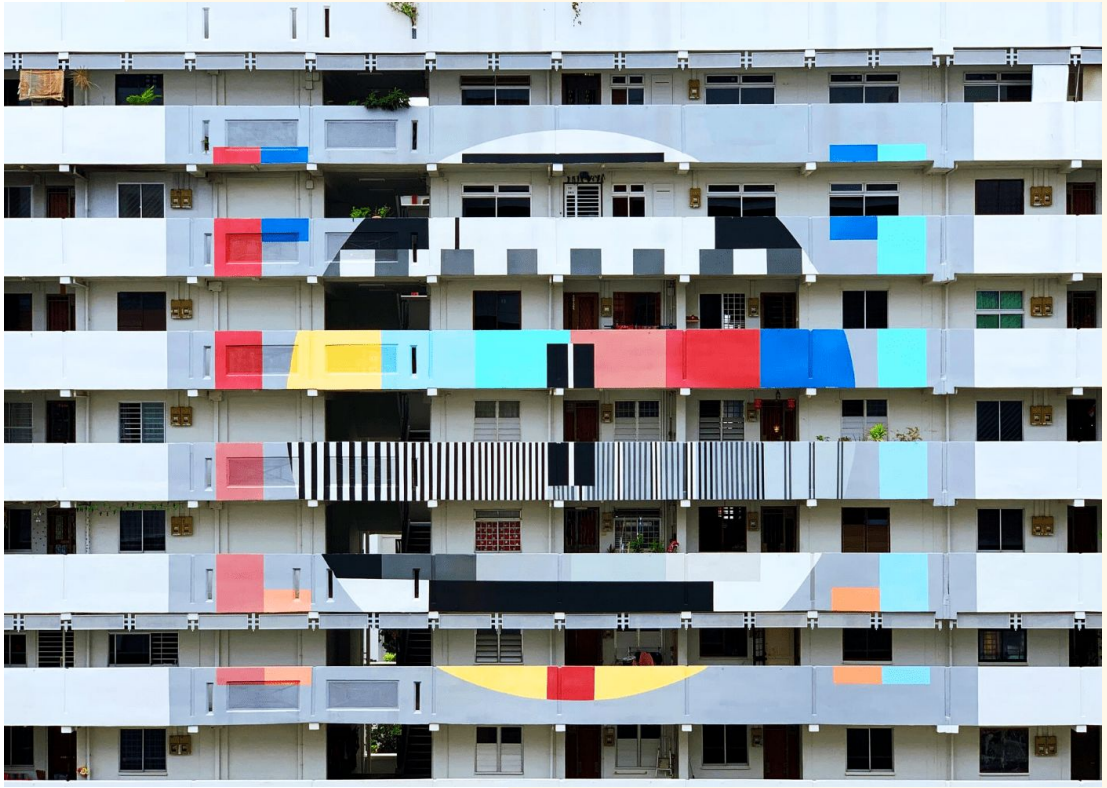
$1022 less
95% confidence Yishun HDB sell for **$837 to $1208** less than ret of SG in 2021

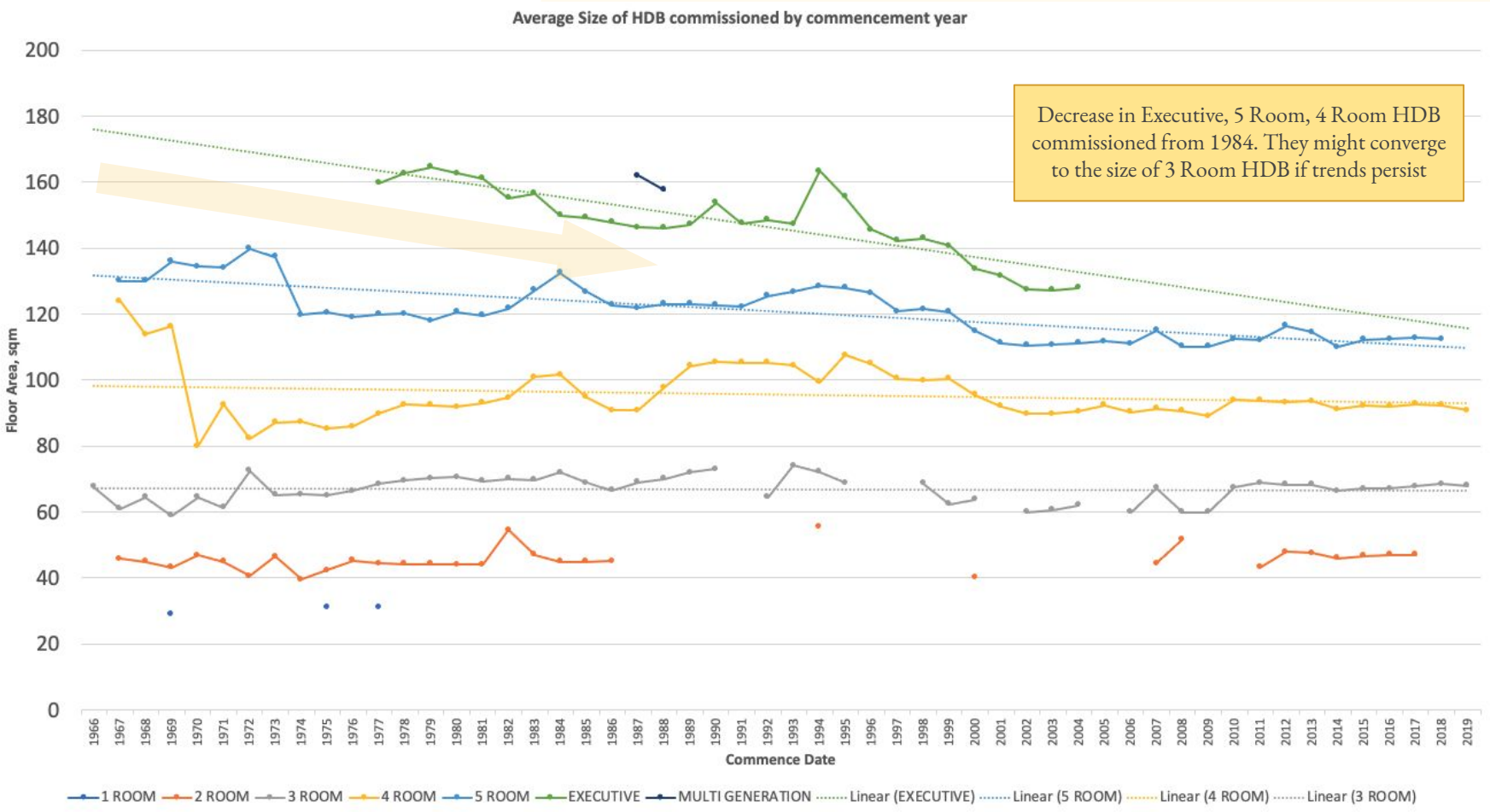
Yishun undervalued
Yishun is statistically cheaper than the rest

Causality
No causal link to 'crazy' town nor exclude time-series factors like inflation

Analysis 2

Shrinking HDB flats



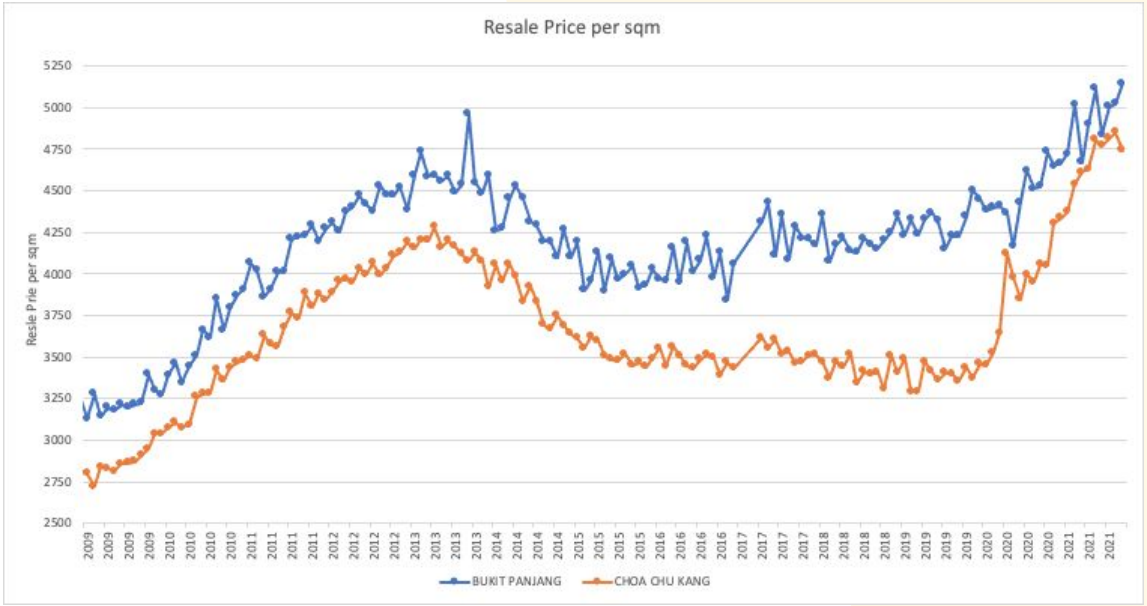

Average Size of HDB commissioned by commencement year
Floor Area, sqm
200
180
160
140
120
100
80
60
40
20
0
1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Commence Date
1 ROOM
2 ROOM
3 ROOM
4 ROOM
5 ROOM
EXECUTIVE
MULTI GENERATION
Linear (EXECUTIVE)
Linear (5 ROOM)
Linear (4 ROOM)
Linear (3 ROOM)
Decrease in Executive, 5 Room, 4 Room HDB commissioned from 1984. They might converge to the size of 3 Room HDB if trends persist

Analysis 3

Yes, access to transport is a factor of house prices

A view of the HDB blocks on Choa Chu Kang Loop (by: Albert Chua/EdgeProp Singapore)

Analysis 3 Exploratory Data Analysis



Resale prices of 4 Room HDB of adjacent towns Bukit Panjang (BP) and Choa Chu Kang (CCK) is compared

But Choa Chu Kang should not be significantly affected by the new MRT. To ensure this, only houses near CCK or BP MRT is chosen for analysis. Then their average per sqm resale price is compared

Analysis 3 Difference in Differences

Price per sqm of 4-Room HDB

Town	Before MRT Built		After MRT Built
Choa Chu Kang	$2,598	a	$3,921
Bukit Panjang	$2,836	b	$4,433

a.	BJ 4-Room average resale price difference after MRT built	$1,597
b.	CCK 4-Room average resale price difference after MRT built	$1,323
c.	delta of differences (a - b)	**$274**

Analysis 3 Difference in Differences OLS

$$px\_per\_sqm = BP + mrt\_built + BP * mrt\_built$$

Next Ordinary Least Squares was performed

average resale price increase of a Bukit Panjang 4-Room flat after the MRT is built is **$274**

And is statistical significant with

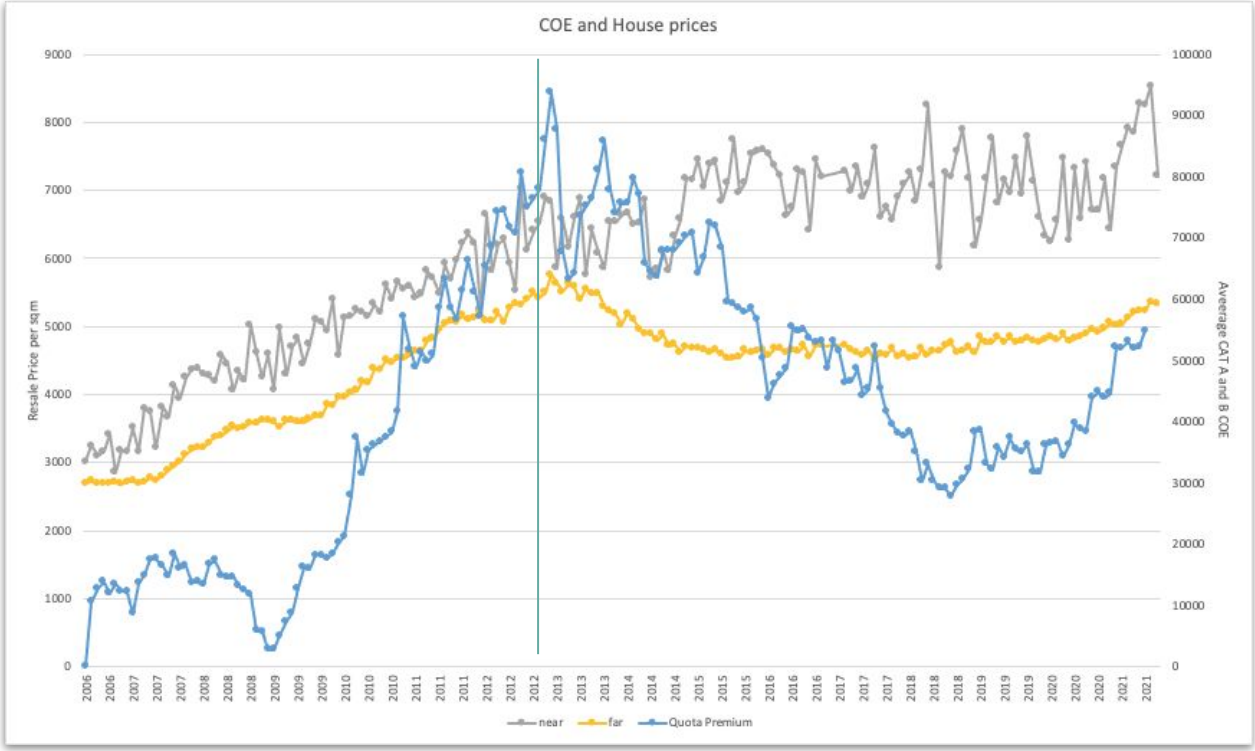
P-value << 0.05

Feature	coefficient	P-value H0: coefficient = 0
Bukit Panjang (BP)	$239	0
mrt_built	$1323	0
Bukit Panjang:mrt_built	**$274**	0

Analysis 4

COE do not make houses less affordable

Analysis 4 Exploratory Data Analysis



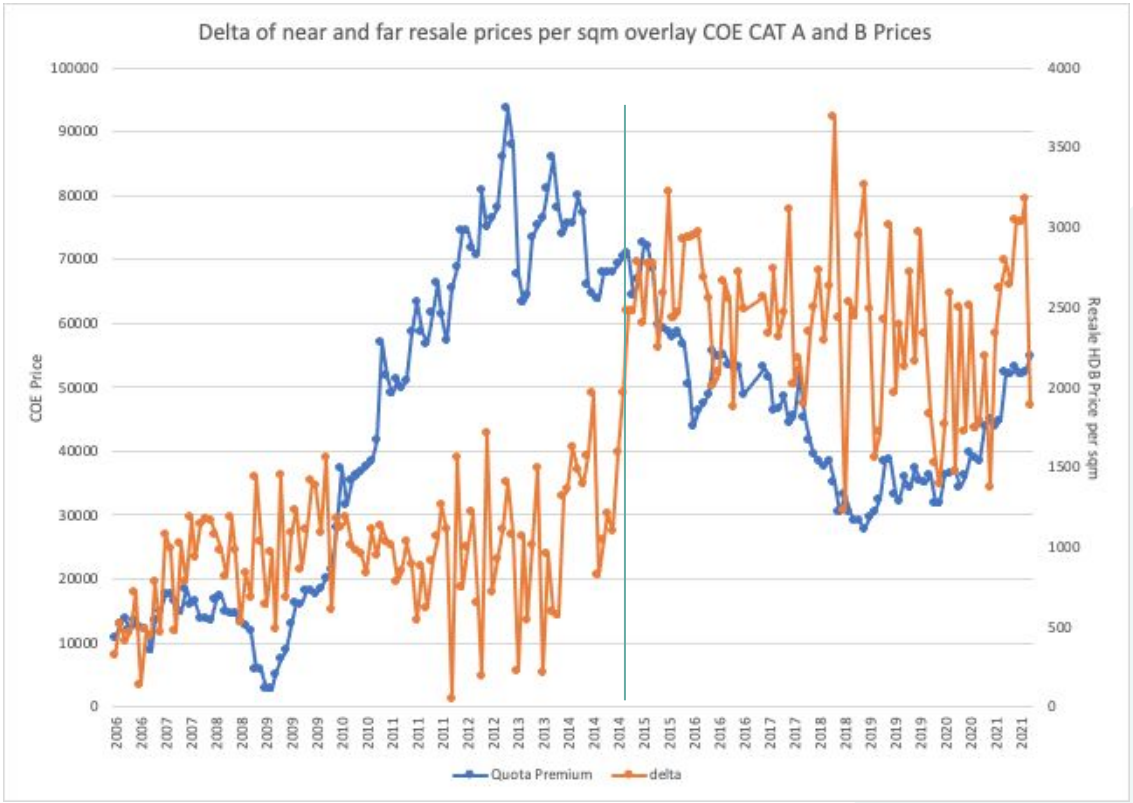
Two towns are grouped:

- Near:
  - Central Area, Kallang/Whampoa
- Far:
  - Sengkang, Punggol

Average resale price per sqm and overlaid with the **COE price**.

There is a shared local peak at 2013 but to find out more the differences is computed

Analysis 4 Resale prices unrelated to COE



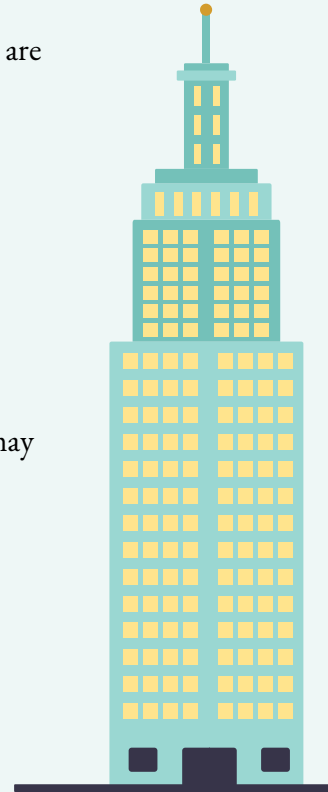
HDB **resale price difference (Delta) of 'near' and 'far'** towns are overlaid with **COE prices** again.

COE peaks at 2012

Delta oscillates with step in 2014

Correlation (COE, delta) = 0.16

There is little correlation. There may not be enough information to constrain and analyse this relationship.



Conclusion

House prices have long been understood as

Accessibility to city, transport and facilities. “Location, location, location”

This was proven in this presentation, but causality cannot be established. Based on stochastic method we can sometimes use these trends to make predictions on resale house prices or vice-versa.

Strongly held beliefs were explained and modelled with large datasets quickly and accurately. With that HDB has the data tools to measure events and make precise decisions.

THANK YOU

Does anyone have any questions?

changjulian17@gmail.com
github.com/changjulian17/
public.tableau.com/app/profile/julian.chang/
linkedin.com/in/julian-chang/