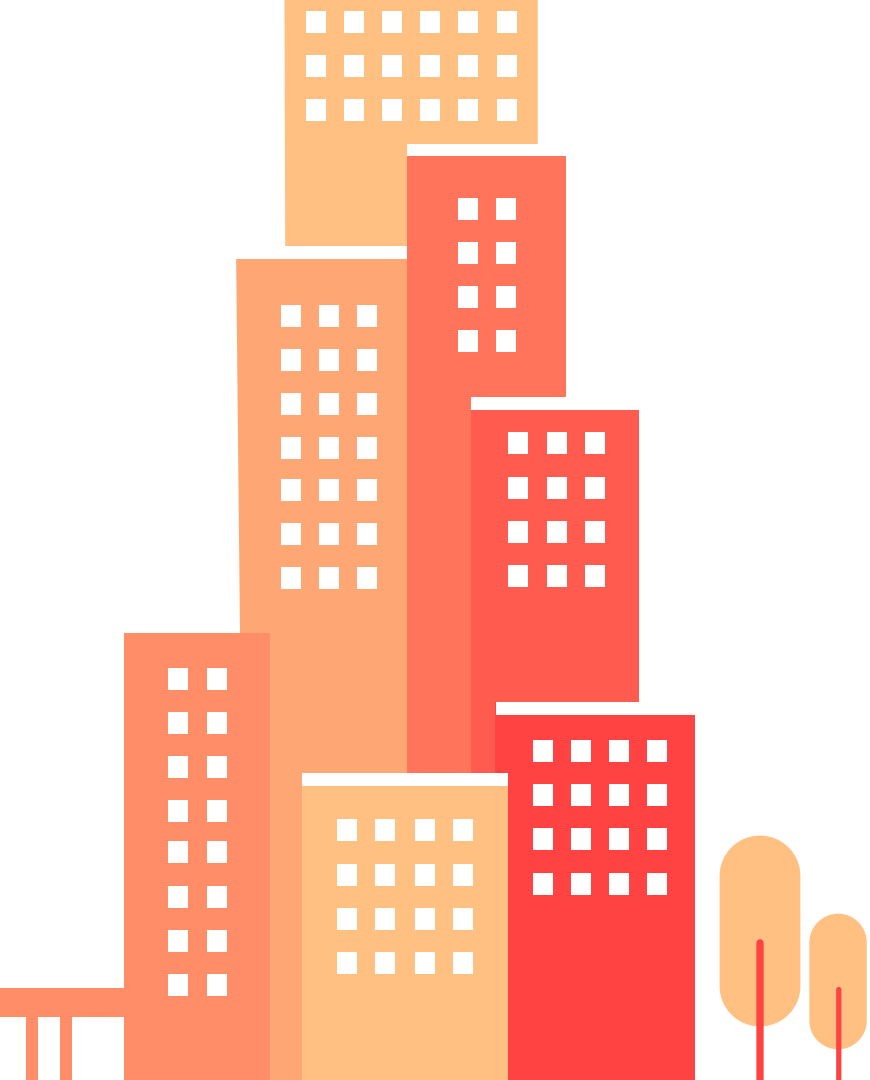


House Hunting with Data

Predicting HDB Resale Flat Prices

Group 1: Edmond, Joanne, Geok

Content

1. **Introduction**:
   - Problem statement, methodology, data dictionary

2. **Data Exploration & Modeling**:
   - Data cleaning, data transformation, model fitting and evaluation

3. **Conclusions & Limitations**:
   - Kaggle scores, key takeaways, limitations

Introduction

Problem Statement

An entrepreneur wanted to set up a new property agency in Singapore. She collected a list of flat-related data, but did not know how to use the data to predict HDB resale flat prices nor how to quantitatively understand how the data impact prices.

Objectives

- Develop a predictive model for the entrepreneur
- Show the relationship between key features and the price

Singapore

HDB resale flat prices up 10.3% in 2022, slower than 12.7% increase in 2021

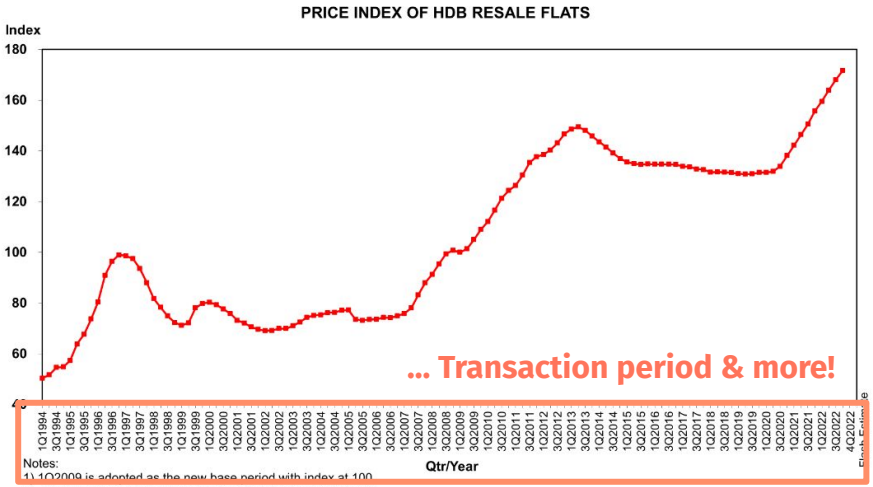
THE STRAITS TIMES

Price growth of HDB resale flats slows in December, analysts expect prices to stabilise in 2023

Locations...

Flat types...

TOWNS	1-ROOM	2-ROOM	3-ROOM	4-ROOM	5-ROOM	EXECUTIVE
ANG MO KIO	-	*	$365,500	$516,500	$800,000	*
BEDOK	-	*	$355,000	$475,000	$680,000	$820,000
BISHAN	-	-	*	$640,000	$855,000	$1,045,000
BUKIT BATOK	-	*	$353,000	$500,000	$720,000	$790,900
BUKIT MERAH	*	*	$368,000	$765,000	$875,000	-
BUKIT PANJANG	-	*	$386,500	$471,900	$610,000	$750,000
BUKIT TIMAH	-	-	*	*	*	*
CENTRAL	-	*	$460,000	$680,000	*	-



Notes:
1) 1Q2009 is adopted as the new base period with index at 100.
2) The index from 1Q1990 to 3Q2014 is rebased to the new base period at 1Q2009.
3) The index from 4Q2014 onwards is computed using the stratified hedonic regression method.
4) The 4Q2022 index is a flash estimate.

Sources: 1. CNA, 2. ST, 3. HDB stats

Methodology

Data Cleaning

Prepare the data for model prediction

Exploratory Data Analysis

Understand the characteristics of each feature

Compare Results

Pick model & dataset that perform the best

Phase 1 Modelling

(using Baseline Dataset)

Using Linear Regression, Lasso, Ridge, Elastic Net

Phase 2 Modelling

(using Modified Dataset)

Using Linear Regression, Lasso, Ridge, Elastic Net

Data Dictionary

77 Data Features

Location

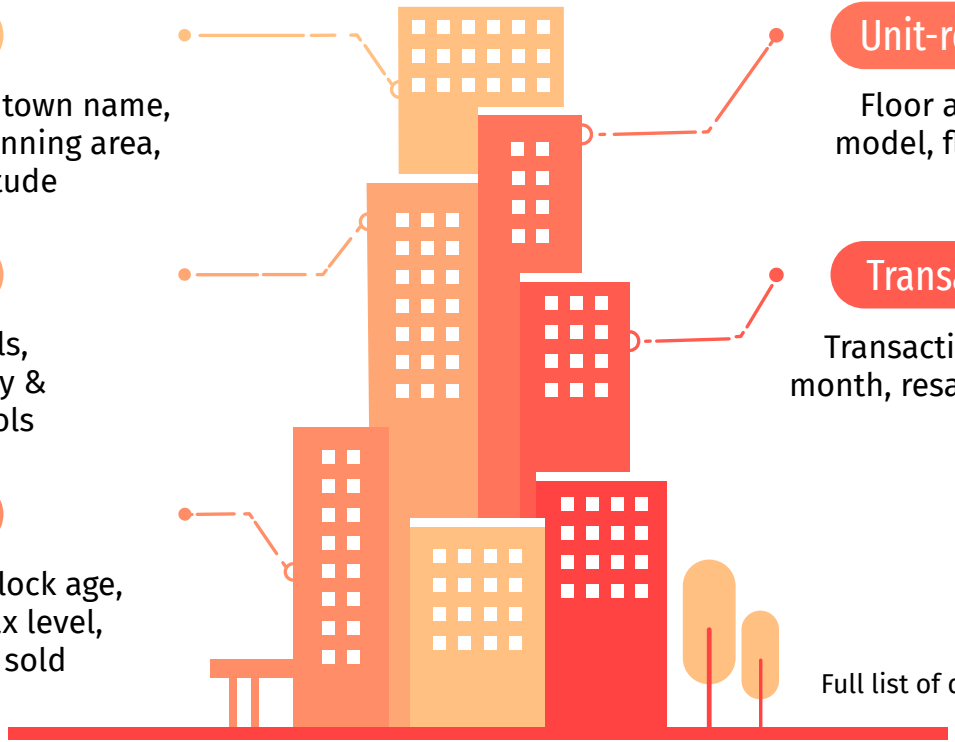
Address, postal, town name, street name, planning area, longitude & latitude

Facilities

Presence of malls, hawkers, primary & secondary schools

Block-related

Block number, block age, building age, max level, number of units sold



Unit-related

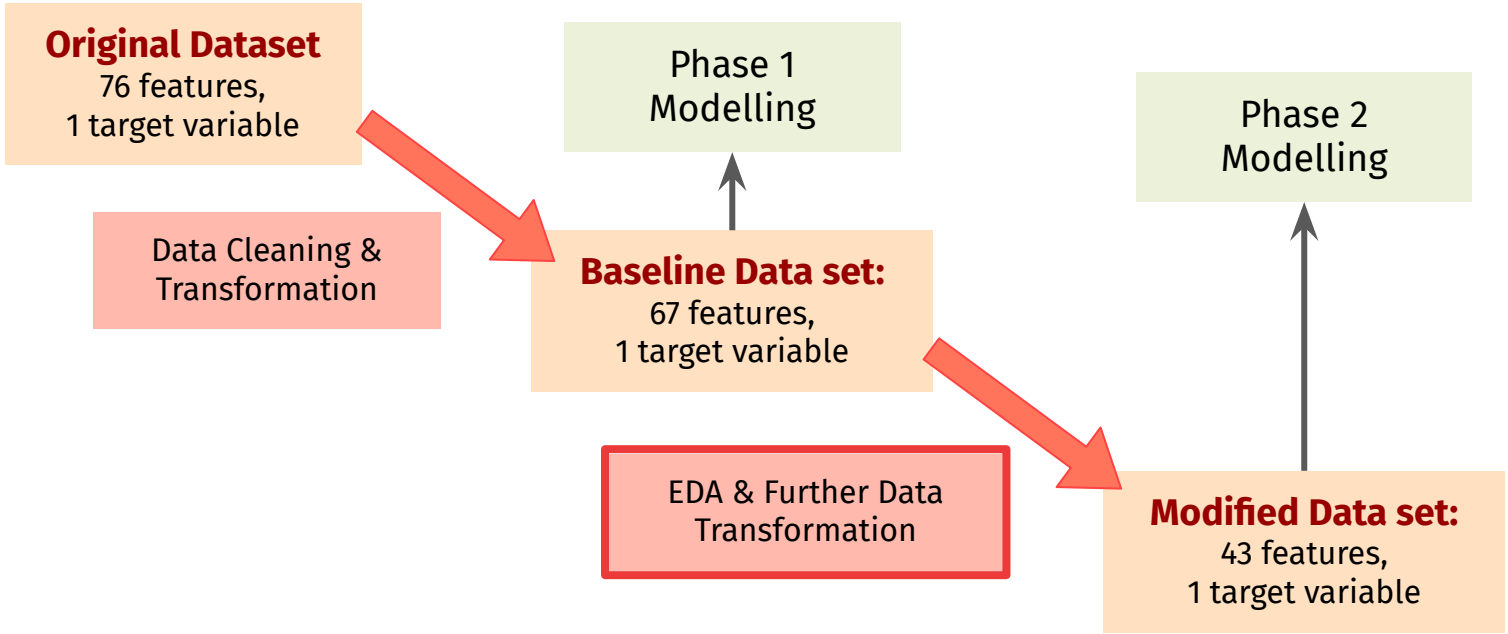
Floor area, flat model, flat type, storey

Transaction

Transaction year, month, resale price

Full list of data in annex A

EDA and Data Transformation Process

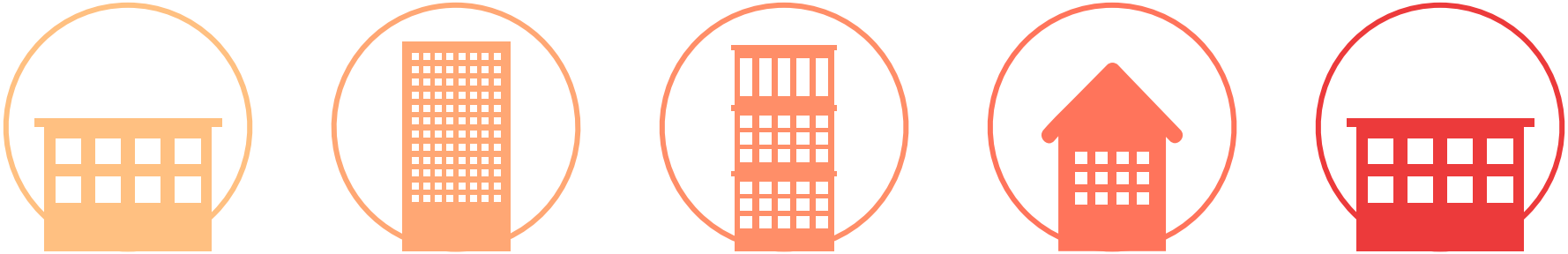


EDA - Looking at dataset as a whole

Overall EDA

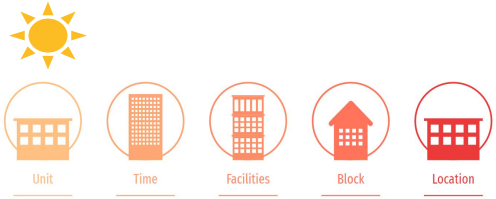
- **Correlation**
  - floor_area_sqft VS resale_price
    r = 0.65
  - Dropped 11 columns
- **Outliers**
  (Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3)
  - Outliers were found in about two-third of the features

EDA



Unit | Time | Facilities | Block | Location

Unit - Flat area, Flat Model, Flat types, Flat Storey



Floor Area VS $
Bigger floor area,
Higher Resale price

Flat Model VS $
Flat models with bigger floor area,
Higher Resale price

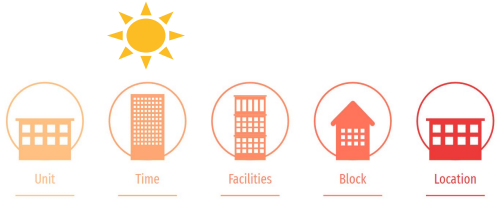
Flat Type VS $
Flat types with bigger floor area,
Higher Resale price

Flat Model VS $/sqft
Some flat models have higher range of price per sqft (e.g. Premium Apartment Loft)

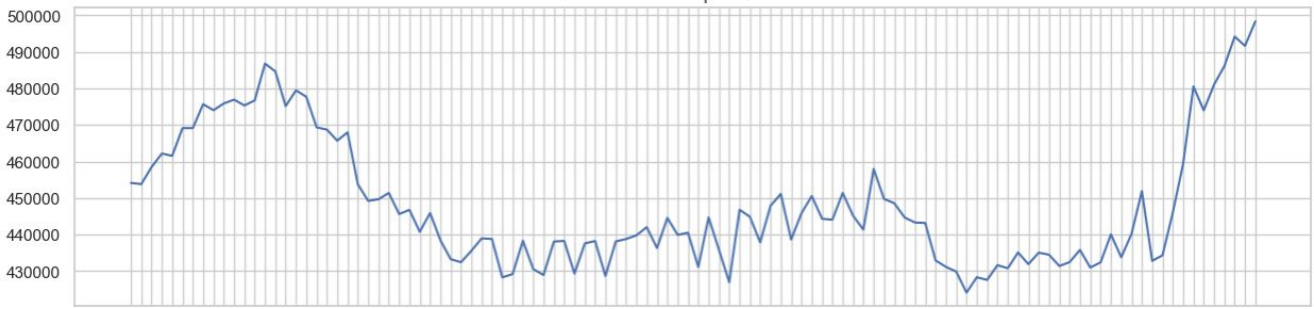
Flat Type VS $/sqft
1room and 2room flat types have higher Resale price.

Time - Transc Year, Month, YearMonth

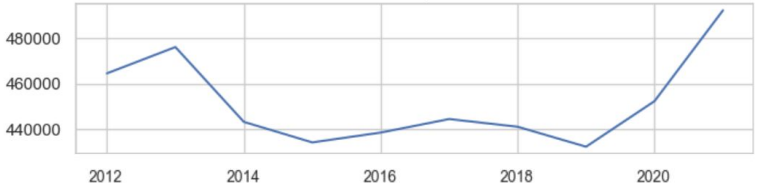
|r|<0.05



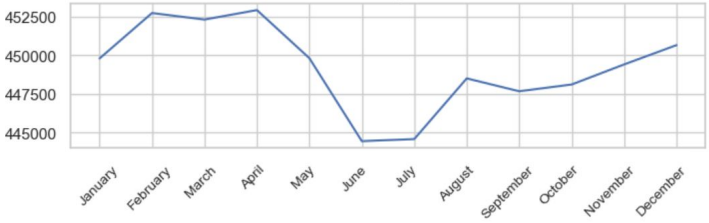
Mean Price Across the months
(Jan2012-Dec2021)

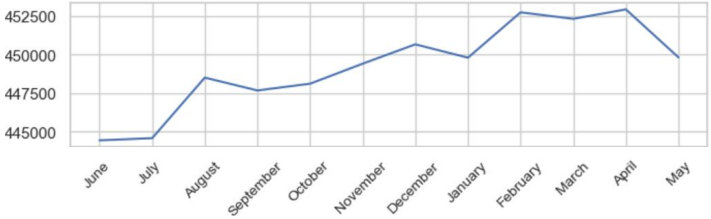


Mean Price Across the Years

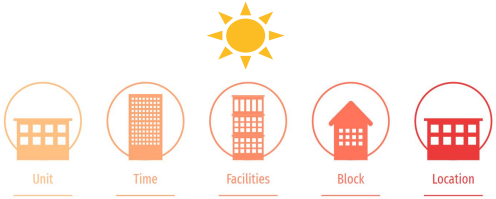


Mean Price Across the Months





Facilities - School, Transport, Mall, Hawker



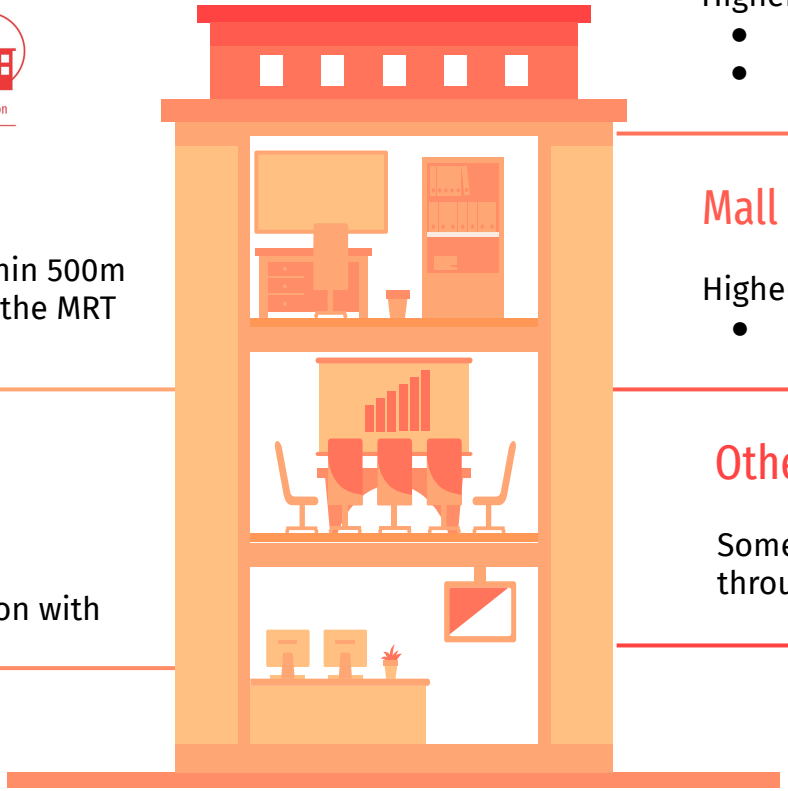
Transport

- All units have a bus stop within 500m
- Presence of MRT and Nearer the MRT station, Higher Resale Price

School

Higher resale price if:

- Have school within 2km
- Nearest school have affiliation with other school



Hawker

Higher resale price if:

- More Hawkers in nearby distances
- Shorter distance away from nearest hawker

Mall

Higher resale price if:

- More Malls in nearby distances

Others

Some columns carries same values throughout its rows (to be dropped)

Block- Rent, Dwelling Units,HDB Age, Hawker, Max Floor levels

Rent

Units in blocks that have presence of rental units appear to have higher resale price

Max Floor Level

Units in blocks with higher Max Floor Level appear to have higher resale price (r=0.496)

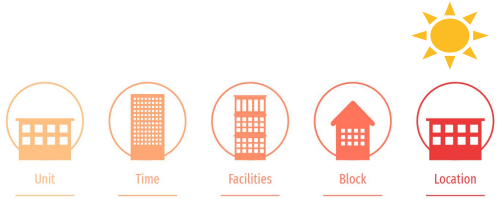
HDB Age

Units in older HDB Blocks appear to have lower resale price (r = -0.35)

Total dwelling units

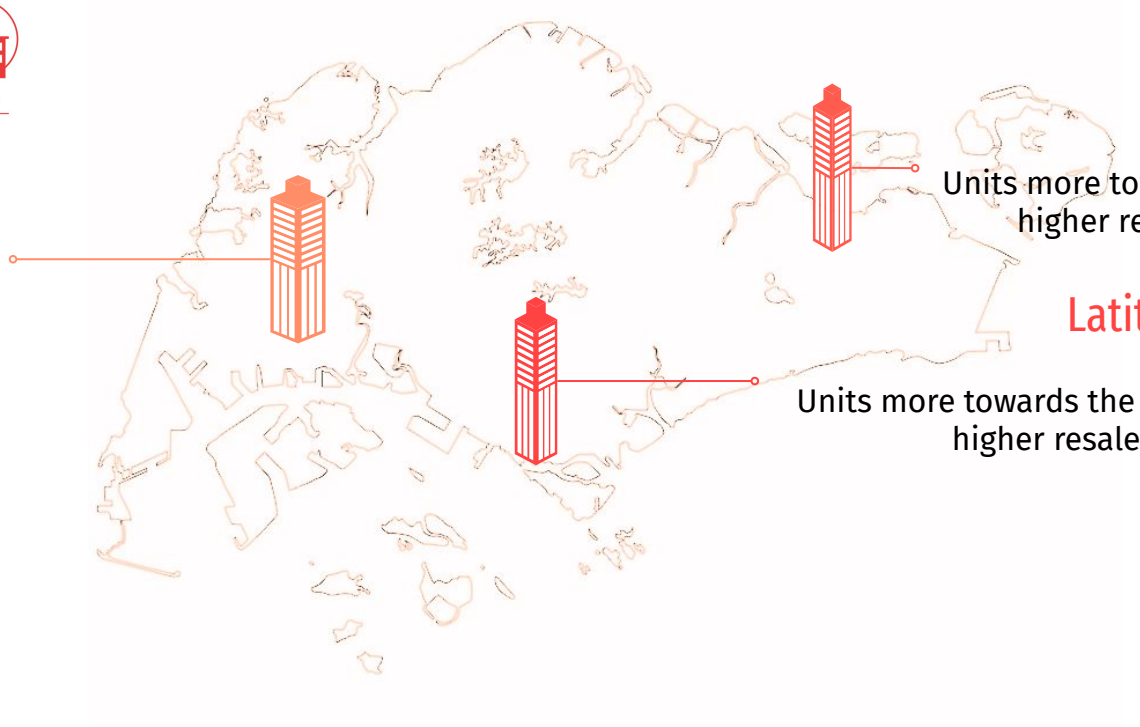
Units in blocks with more total dwelling units appear to have higher resale prices per sqft (r=0.27)

Location-Longitude, Latitude, Town, Planning Area



Planning area and Town

- Resale prices range vary greatly among the areas.
- Meanwhile, some particular town/ planning areas observed distinctly low or high resale price



Phase 1 Modelling (67 features)

Models Performance

Dataset	Mode	Preprocessing	Train Score	Test Score	Cross Validation Score	RMSE Score	
Baseline Dataset	Linear Regression	• OneHotEncoder • StandardScaler	0.93901	0.93684	-1.36245e+17	35890.97	**Model A**
Baseline Dataset	LassoCV	• OneHotEncoder • StandardScaler	0.90602	0.90643	0.90553	43686.99	
Baseline Dataset	RidgeCV	• OneHotEncoder • StandardScaler	0.93892	0.93694	0.93607	35863.73	**Model B**
Baseline Dataset	ElasticNetCV	• OneHotEncoder • StandardScaler	0.05182	0.05216	Not conducted as Train/Test Score were poor	Not conducted as Train/Test Score were poor	
Modified Dataset	Linear Regression	• OneHotEncoder • StandardScaler	0.89723	0.89666	0.89640	45910.49	
Modified Dataset	LassoCV	• OneHotEncoder • StandardScaler	0.88242	0.88296	0.88208	48859.74	
Modified Dataset	RidgeCV	• OneHotEncoder • StandardScaler	0.89726	0.89665	0.89645	45913.38	**Model C**
Modified Dataset	ElasticNetCV	• OneHotEncoder • StandardScaler	0.02582	0.02600	Not conducted as Train/Test Score were poor	Not conducted as Train/Test Score were poor	

Phase 2 Modelling (43 features)

Coming up with Model D

Model C:
Modified Dataset
Ridge CV

- Bus Stop Name
- MRT Name
- Transc Year
- Transc Month

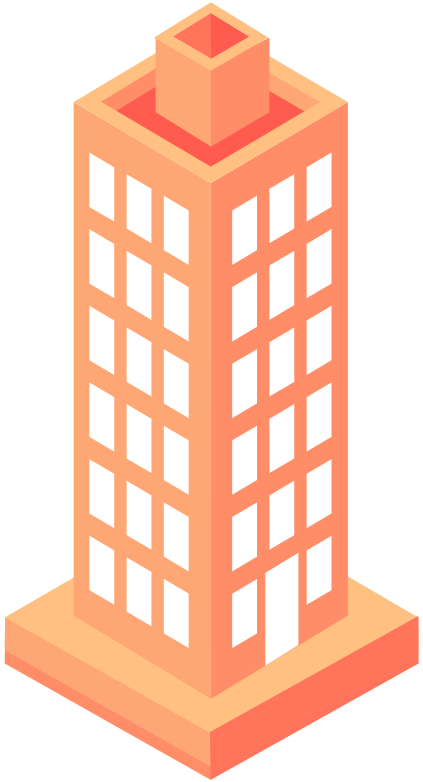
Model D:
Further Modified
Dataset
Ridge CV

Bus Stop Name & MRT Name to give more precise location of the unit

Transc Year and Month as categorical variable to account of unique events that occur at a particular time.

Kaggle Scores

Submission and Description	Private Score	Public Score
Model_D_pred.csv Complete (after deadline) · now	36625.44494	35918.82022
Model_C_pred.csv Complete (after deadline) · 1s ago	45958.69557	46490.91079
Model_B_pred.csv Complete (after deadline) · 1s ago	36097.93041	35539.92073
Model_A_pred.csv Complete (after deadline) · 1m ago	36171.07122	35572.08244



Key Takeaways from Model Performance

Model A

Baseline dataset with minimal data transformation from its original state gives a good linear regression model performance, however it has poor cross validation score (-1.4e+17).

Model B

RidgeCV on Baseline dataset performs the best

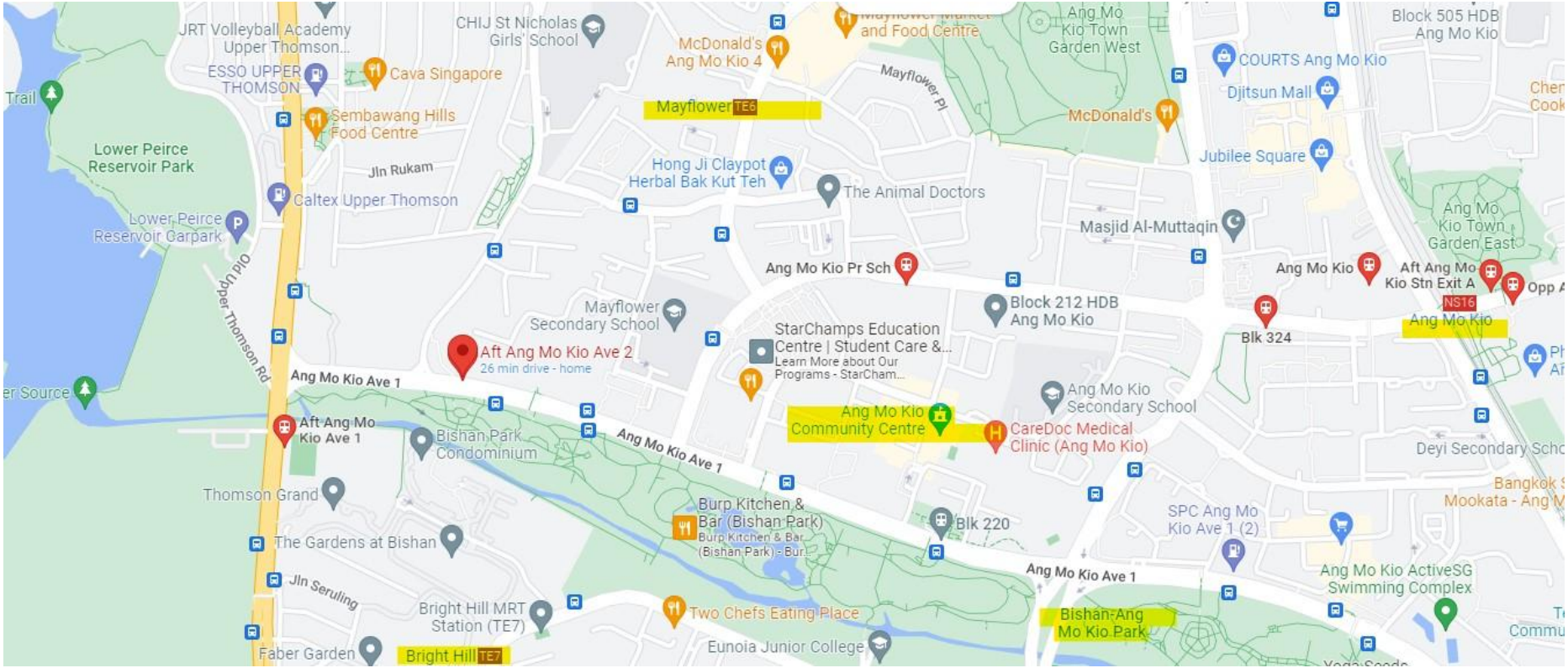
Model C

Dataset modifications based on earlier EDA on the features gives a poorer performance

Model D

Inclusion of 'Bus Stop Name', 'MRT Name', 'Transc Year' and 'Transc Month' gives a similar performance to Model B

To illustrate...



Conclusions on Key Features affecting Resale Price

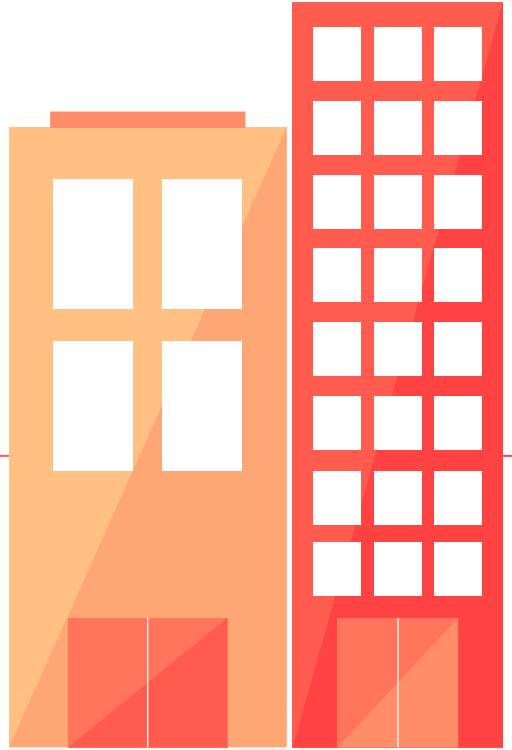
Proximity to popular schools

Events in specific year/month

Proximity to other facilities and amenities

Limitations

Incomplete information in dataset



Domain knowledge

Thank you!

Annex A - Original Dataset (1)

block	block number of the resale flat, e.g. 454	Block_characteristic
lease_commence_date	commencement year of the flat unit's 99-year lease	Block_characteristic
year_completed	year which construction was completed for resale flat	Block_characteristic
hdb_age	number of years from lease_commence_date to present year	Block_characteristic
total_dwelling_units	total number of residential dwelling units in the resale flat	Block_characteristic
1room_rental	number of 1-room rental residential units in the resale flat block	Block_characteristic
2room_rental	number of 2-room rental residential units in the resale flat block	Block_characteristic
3room_rental	number of 3-room rental residential units in the resale flat block	Block_characteristic
other_room_rental	number of "other" type rental residential units in the resale flat block	Block_characteristic
1room_sold	number of 1-room residential units in the resale flat	Block_characteristic
2room_sold	number of 2-room residential units in the resale flat	Block_characteristic
3room_sold	number of 3-room residential units in the resale flat	Block_characteristic
4room_sold	number of 4-room residential units in the resale flat	Block_characteristic
5room_sold	number of 5-room residential units in the resale flat	Block_characteristic
commercial	boolean value if resale flat has commercial units in the same block	Block_characteristic
exec_sold	number of executive type residential units in the resale flat block	Block_characteristic
market_hawker	boolean value if resale flat has a market or hawker centre in the same block	Block_characteristic
multigen_sold	number of multi-generational type residential units in the resale flat block	Block_characteristic
multistorey_carpark	boolean value if resale flat has a multistorey carpark in the same block	Block_characteristic
precinct_pavilion	boolean value if resale flat has a pavilion in the same block	Block_characteristic
residential	boolean value if resale flat has residential units in the same block	Block_characteristic
studio_apartment_sold	number of studio apartment type residential units in the resale flat block	Block_characteristic
max_floor_lvl	highest floor of the resale flat	Block_characteristic

Annex A - Original Dataset (2)

Mall_Nearest_Distance	distance (in metres) to the nearest mall	Facilities_Hawker
hawker_food_stalls	number of hawker food stalls in the nearest hawker centre	Facilities_Hawker
hawker_market_stalls	number of hawker and market stalls in the nearest hawker centre	Facilities_Hawker
Hawker_Nearest_Distance	distance (in metres) to the nearest hawker centre	Facilities_Hawker
Hawker_Within_1km	number of hawker centres within 1 kilometre	Facilities_Hawker
Hawker_Within_2km	number of hawker centres within 2 kilometres	Facilities_Hawker
Hawker_Within_500m	number of hawker centres within 500 metres	Facilities_Hawker
Mall_Within_2km	number of malls within 2 kilometres	Facilities_Hawker
Mall_Within_1km	number of malls within 1 kilometre	Facilities_Hawker
Mall_Within_500m	number of malls within 500 metres	Facilities_Hawker
pri_sch_latitude	latitude (in decimal degrees) of the the nearest primary school	Facilities_School
pri_sch_longitude	longitude (in decimal degrees) of the nearest primary school	Facilities_School
sec_sch_latitude	latitude (in decimal degrees) of the the nearest secondary school	Facilities_School
sec_sch_longitude	longitude (in decimal degrees) of the nearest secondary school	Facilities_School
cutoff_point	PSLE cutoff point of the nearest secondary school	Facilities_School
vacancy	number of vacancies in the nearest primary school	Facilities_School
pri_sch_name	name of the nearest primary school	Facilities_School
sec_sch_name	name of the nearest secondary school	Facilities_School
affiliation	boolean value if the nearest secondary school has an primary school affiliation	Facilities_School
pri_sch_affiliation	boolean value if the nearest primary school has a secondary school affiliation	Facilities_School
pri_sch_nearest_distance	distance (in metres) to the nearest primary school	Facilities_School
sec_sch_nearest_dist	distance (in metres) to the nearest secondary school	Facilities_School
bus_stop_latitude	latitude (in decimal degrees) of the the nearest bus stop	Facilities_Transport
bus_stop_longitude	longitude (in decimal degrees) of the nearest bus stop	Facilities_Transport
mrt_latitude	latitude (in decimal degrees) of the the nearest MRT station	Facilities_Transport
mrt_longitude	longitude (in decimal degrees) of the nearest MRT station	Facilities_Transport
bus_stop_name	name of the nearest bus stop	Facilities_Transport
bus_stop_nearest_distance	distance (in metres) to the nearest bus stop	Facilities_Transport
mrt_name	name of the nearest MRT station	Facilities_Transport
bus_interchange	boolean value if the nearest MRT station is also a bus interchange	Facilities_Transport
mrt_interchange	boolean value if the nearest MRT station is a train interchange station	Facilities_Transport
mrt_nearest_distance	distance (in metres) to the nearest MRT station	Facilities_Transport

Annex A - Original Dataset (3)

address	combination of block and street_name	Location
postal	postal code of the resale flat block	Location
street_name	street name where the resale flat resides, e.g. TAMPINES ST 42	Location
planning_area	Government planning area that the flat is located	Location
town	HDB township where the flat is located, e.g. BUKIT MERAH	Location
Latitude	Latitude based on postal code	Location
Longitude	Longitude based on postal code	Location
Tranc_Month	month of resale transaction	Purchase
Tranc_Year	year of resale transaction	Purchase
Tranc_YearMonth	year and month of the resale transaction, e.g. 2015-02	Purchase
resale_price	the property's sale price in Singapore dollars. This is the target variable that you're trying to predict for this challenge.	Purchase
floor_area_sqm	floor area of the resale flat unit in square metres	Unit_characteristic
full_flat_type	combination of flat_type and flat_model	Unit_characteristic
lower	lower value of storey_range	Unit_characteristic
mid	middle value of storey_range	Unit_characteristic
upper	upper value of storey_range	Unit_characteristic
flat_model	HDB model of the resale flat, e.g. Multi Generation	Unit_characteristic
flat_type	type of the resale flat unit, e.g. 3 ROOM	Unit_characteristic
floor_area_sqft	floor area of the resale flat unit in square feet	Unit_characteristic
mid_storey	median value of storey_range	Unit_characteristic
storey_range	floor level (range) of the resale flat unit, e.g. 07 TO 09	Unit_characteristic