Resale House Prices

By: Carel Ong Shi Ting

1 Dataset

The original dataset consists of 66,497 observations from 2017 onwards. There are 11 features in the dataset. 8 of which are categorical variables and the remaining are quantitative variables. Below is the explanation of some of the variables.

‘floor_area_sqm’ : Numerical variable that gives us the floor area of a flat in square metres.

‘remaining_lease’ : Numerical variable, tells us the amount of time left before the housing lease expires. A typical housing lease lasts for 99 years.

‘storey_range’ : Categorical variable that gives us the range of levels where a particular flat can be found in.

2 Pre-processing

Pre-processing

1) Addition and renaming of remaining_lease variable
2) Splitting of storey_range to min_storey and max_storey
3) Generation of new variables: Distance to nearest MRT stations, Primary Schools, Shopping Malls and to CBD District (Raffles Place MRT), mature_estates, flat_premium and different levels for flat model.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price
2	2012-03	ANG MO KIO	2 ROOM	172	ANG MO KIO AVE 4	06 TO 10	45	Improved	1986	250000
3	2012-03	ANG MO KIO	2 ROOM	510	ANG MO KIO AVE 8	01 TO 05	44	Improved	1980	265000
4	2012-03	ANG MO KIO	3 ROOM	610	ANG MO KIO AVE 4	06 TO 10	68	New Generation	1980	315000
5	2012-03	ANG MO KIO	3 ROOM	474	ANG MO KIO AVE 10	01 TO 05	67	New Generation	1984	320000
6	2012-03	ANG MO KIO	3 ROOM	604	ANG MO KIO AVE 5	06 TO 10	67	New Generation	1980	321000
7	2012-03	ANG MO KIO	3 ROOM	154	ANG MO KIO AVE 5	01 TO 05	68	New Generation	1981	321000

E.g. Dataset with missing 'remaining_lease' variable.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Generation	1978	60 years 07 months	250000

E.g. Dataset with 'remaining_lease' variable specified in years and months.

However, for the dataset used (from 2017 onwards), 'remaining_lease' variable is already present. We will then rename the variable to change it to be in years, instead of years and months.
This step can be applied if we were to include more data for our analysis (from 1990-1999 or from 2012-2015 data etc.) to ensure consistency.

1) Addition and renaming of 'remaining_lease' variable

	month	town	flat_type	block	street_name	storey_range	min_storey	max_storey	floor_area_sqm	flat_model	lease_commence_date	remaining_
0	2017-01-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	10	12	44.0	Improved	1979	
1	2017-01-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	01	03	67.0	New Generation	1978	
2	2017-01-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	01	03	67.0	New Generation	1980	
3	2017-01-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	04	06	68.0	New Generation	1980	
4	2017-01-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	01	03	67.0	New Generation	1980	
...	...	...	...	...	...	...	...	...	...	...	...	
66492	2020-01-01	YISHUN	EXECUTIVE	355A	YISHUN RING RD	01 TO 03	01	03	145.0	Maisonette	1988	
66493	2020-01-01	YISHUN	EXECUTIVE	604	YISHUN ST 61	01 TO 03	01	03	164.0	Apartment	1992	
66494	2020-01-01	YISHUN	EXECUTIVE	606	YISHUN ST 61	01 TO 03	01	03	146.0	Maisonette	1987	
66495	2020-01-01	YISHUN	EXECUTIVE	611	YISHUN ST 61	01 TO 03	01	03	146.0	Maisonette	1987	
66496	2020-01-01	YISHUN	EXECUTIVE	824	YISHUN ST 81	01 TO 03	01	03	145.0	Apartment	1987	

66497 rows × 13 columns

'storey_range' variable is being split into 2 variables: 'min_storey' and 'max_storey', which gives the lowest and highest floor that the flat can be found in respectively.

2) Splitting of 'storey_range' (categorical) variable

- Longitudes and Latitudes are extracted using OneMap API and manually keyed in for those that are not found in the API.
- MRT Stations' longitude and latitude are obtained from a csv file – 'mrtdata', found on public GitHub repository.
- List of Primary Schools and List of Shopping Malls in Singapore are extracted from Wikipedia. Then, the respective longitudes and latitudes are obtain by searching these names using the OneMap API.

Formula for calculating distance from flat to destination:

Difference in latitude = (Specific Flat's latitude– Place of Interest's Latitude)*110.574
Difference in longitude = (Specific Flat's latitude– Place of Interest's Latitude)*111.32
Distance = [(Difference in latitude)^2 + (Difference in longitude)^2]^0.5

3) Generation of new variables

	OBJECTID	STN_NAME	STN_NO	X	Y	Latitude	Longitude	COLOR
0	12	ADMIRALTY MRT STATION	NS10	24402.1063	46918.1131	1.440585	103.800998	RED
1	16	ALJUNIED MRT STATION	EW9	33518.6049	33190.0020	1.316433	103.882893	GREEN
2	33	ANG MO KIO MRT STATION	NS16	29807.2655	39105.7720	1.369933	103.849553	RED
3	81	BAKAU LRT STATION	SE3	36026.0821	41113.8766	1.388093	103.905418	OTHERS
4	80	BANGKIT LRT STATION	BP9	21248.2460	40220.9693	1.380018	103.772667	OTHERS
...	...	...	...	...	...	...	...	...
182	175	WOODLANDS SOUTH MRT STATION	TE3	23607.8309	45444.7113	1.427260	103.793863	OTHERS
183	146	WOODLEIGH MRT STATION	NE11	32173.3186	35706.3794	1.339190	103.870808	PURPLE
184	6	YEW TEE MRT STATION	NS5	18438.9791	42158.0124	1.397535	103.747431	RED
185	41	YIO CHU KANG MRT STATION	NS15	29294.1283	40413.0820	1.381756	103.844944	RED
186	13	YISHUN MRT STATION	NS13	28187.6787	45686.0701	1.429443	103.835005	RED

mrtdata dataset

```
['Admiralty Primary School',
 'Ahmad Ibrahim Primary School',
 'Ai Tong School',
 'Alexandra Primary School',
 'Anchor Green Primary School',
 'Anderson Primary School',
 'Anglo-Chinese School (Junior)',
 'Anglo-Chinese School (Primary)',
 'Angsana Primary School',
 'Ang Mo Kio Primary School',
 'Balestier Hill Primary School',
 'Beacon Primary School',
 'Bedok Green Primary School',
 'Bendemeer Primary School',
 'Blangah Rise Primary School',
 'Boon Lay Garden Primary School',
 'Bukit Panjang Primary School',
 'Bukit Timah Primary School',
 'Bukit View Primary School',
```

List of Primary School Names

```
['100 AM',
 '313@Somerset',
 'Aperia',
 'Balestier Hill Shopping Centre',
 'Bugis Cube',
 'Bugis Junction',
 'Bugis+',
 'Capitol Piazza',
 'Cathay Cineleisure Orchard',
 'Clarke Quay Central',
 'The Centrepoint',
 'City Square Mall',
 'City Gate Mall',
 'CityLink Mall',
 'Duo',
 'Far East Plaza',
 'Funan',
 'Great World City',
 'HDB Hub',
```

List of Shopping Malls

Distance to nearest MRT Station

Numerical variable; gives the distance from a flat to its nearest MRT station.

Distance to nearest Primary School

Numerical variable; gives the distance from a flat to its nearest Primary School

Distance to nearest Shopping Mall

Numerical variable; gives the distance from a flat to its nearest MRT station.

Nearest MRT Station

Qualitative variable; outputs names of the nearest MRT station, based on the location of the flat.

Nearest Primary School

Qualitative variable; outputs names of the nearest Primary School, based on the location of the flat.

Nearest Shopping Mall

Qualitative variable; outputs names of the nearest Shopping Mall, based on the location of the flat.

Distance to CBD

Numerical variable; gives the distance from a flat to Raffles Place MRT station.

flat_type_premium

Numerical variable; outputs the premium from purchasing a flat, based on the flat type.

A negative values means the buyer is able to save that specific amount when purchasing. A positive value suggests an additional cost incurred by the buyer.

Different levels for flat_model

Binary variable; 1 if the flat is of a particular flat model, say 'Apartment', and 0 otherwise. There are a total of 16 variables. Additionally, there is a binary variable – 'Others' where it returns 1 if the model is '2-room', 'Premium Apartment Loft', 'Improved-Maisonette' or 'Premium Maisonette', else 0.

Premium based on type of flat

flat_type	floor_area_sqm	lease_commence_date	remaining_lease	resale_price	flat_premium
1 ROOM	31.0	1975	56	180000.0	-222888.0
2 ROOM	46.0	2011	92	230000.0	-172888.0
3 ROOM	67.0	1982	63	292000.0	-110888.0
4 ROOM	93.0	1997	79	402888.0	0.0
5 ROOM	119.0	1999	80	480000.0	77112.0
EXECUTIVE	146.0	1994	75	600000.0	197112.0
MULTI-GENERATION	165.0	1987	68	798888.0	396000.0

Purchasing a 5-room flat will incur an additional cost of $77,112 while purchasing a 3-room flat allows buyer to save $110,888.

- Ang Mo Kio
- Bedok
- Bishan
- Bukit Merah
- Bukit Timah
- Central
- Clementi
- Geylang
- Kallang/Whampoa
- Marine Parade
- Pasir Ris
- Queenstown
- Serangoon
- Tampines
- Toa Payoh

List of locations where Mature Estates are at in Singapore

After some research, it appears that the area in which the estates are located at have an impact on the resale house prices.

Specifically, these areas consist of estates that are more mature than other areas. This relationship is observed in our dataset as shown in the next slide.

Thus, we encode a binary variable, 'mature_estate' where 1 if the flat is a mature estate and 0 otherwise.

Premium based on area

Purchase of flats located in Central Area will incur additional cost of $295,888 while flats in non-Central area such as Sembawang will not. (in blue)

Flats situated in more mature areas (>20 years) such as Bishan, Bukit Timah incurs a much higher cost than flats in non-mature areas. (in green)

town	floor_area_sqm	lease_commence_date	remaining_lease	resale_price	Distance to nearest MRT station	D
ANG MO KIO	82.0	1980.0	61.0	345000.0	0.720505	0
BEDOK	84.0	1980.0	61.0	368000.0	0.606057	0
BISHAN	106.0	1988.0	69.0	628000.0	0.765247	0
BUKIT BATOK	92.0	1986.0	67.0	350400.0	0.620062	0
BUKIT MERAH	90.0	1986.0	68.0	583500.0	0.549554	0
BUKIT PANJANG	103.0	1999.0	80.0	417000.0	0.224331	0
BUKIT TIMAH	104.0	1988.0	69.0	716888.0	0.381359	0
CENTRAL AREA	82.0	1984.0	65.0	510000.0	0.297870	0
CHOA CHU KANG	108.0	1996.0	78.0	365000.0	0.494839	0
CLEMENTI	82.0	1980.0	61.0	405000.0	0.705524	0
GEYLANG	83.0	1981.0	62.0	375000.0	0.406267	0
HOUGANG	103.0	1989.0	70.0	401000.0	0.785793	0
JURONG EAST	94.0	1984.0	65.0	390000.0	0.825014	0
JURONG WEST	104.0	1997.0	78.0	385000.0	0.808901	0
KALLANG/WHAMPOA	86.0	1982.0	63.0	468000.0	0.438825	0
MARINE PARADE	76.0	1975.0	56.0	468000.0	1.900832	0
PASIR RIS	123.0	1993.0	75.0	470000.0	1.115484	0
PUNGGOL	93.0	2012.0	94.0	443000.0	0.231903	0
QUEENSTOWN	83.0	1986.0	67.5	550000.0	0.444014	0
SEMBAWANG	102.0	2001.0	82.0	370000.0	0.537483	0
SENGKANG	95.0	2004.0	86.0	425000.0	0.263405	0
SERANGOON	101.0	1986.0	67.0	470000.0	0.820323	0
TAMPINES	105.0	1988.0	69.0	450000.0	0.556134	0
TOA PAYOH	82.0	1984.0	64.0	425000.0	0.495828	0
WOODLANDS	103.0	1997.0	79.0	363000.0	0.610309	0
YISHUN	92.0	1987.0	68.0	337000.0	0.814945	0

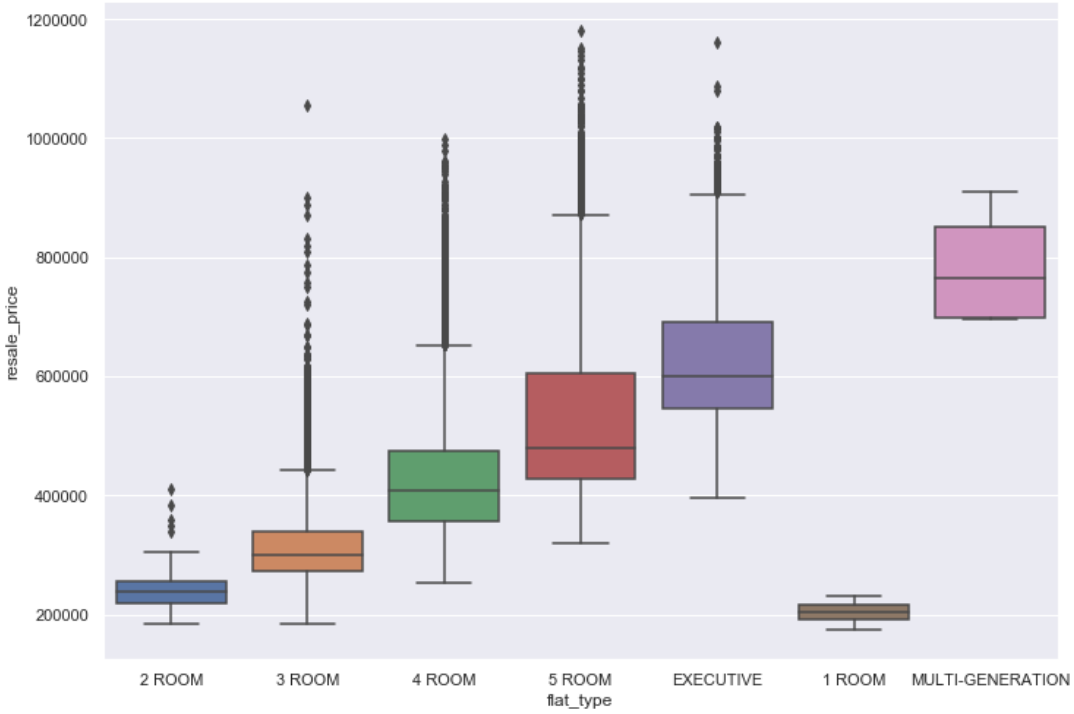
Columns in final dataset

Addition of 27 variables

```
dt_use.columns
```

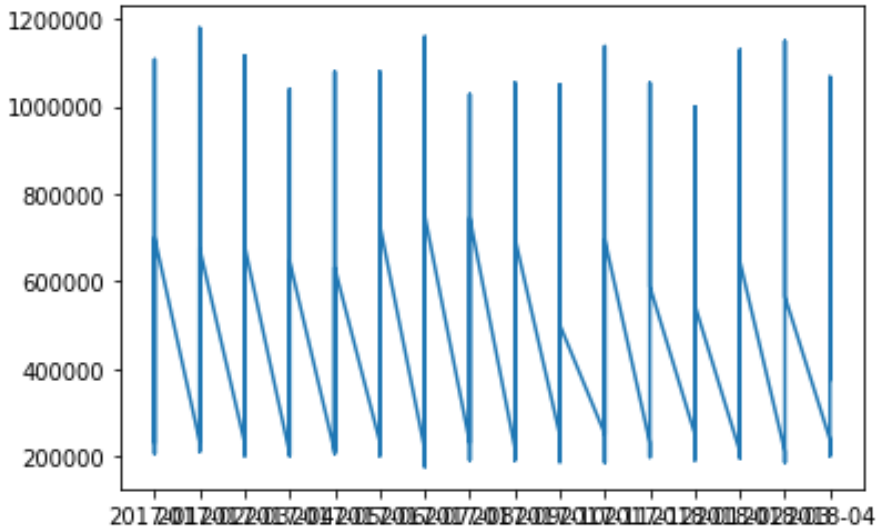
```
Index(['month', 'town', 'flat_type', 'block', 'street_name', 'storey_range',
       'min_storey', 'max_storey', 'floor_area_sqm', 'flat_model',
       'lease_commence_date', 'remaining_lease', 'resale_price',
       'Distance to nearest MRT station', 'Nearest MRT station',
       'Distance to nearest Primary School', 'Nearest Primary School',
       'Distance to nearest Shopping Mall', 'Nearest Shopping Mall',
       'Distance to CBD', 'mature_estate', 'Adjoined flat', 'Apartment',
       'DBSS', 'Improved', 'Maisonette', 'Model A', 'Model A-Maisonette',
       'Model A2', 'Multi Generation', 'New Generation', 'Premium Apartment',
       'Simplified', 'Standard', 'Terrace', 'Type S1', 'Type S2', 'Others'],
      dtype='object')
```

2 Exploratory Analysis

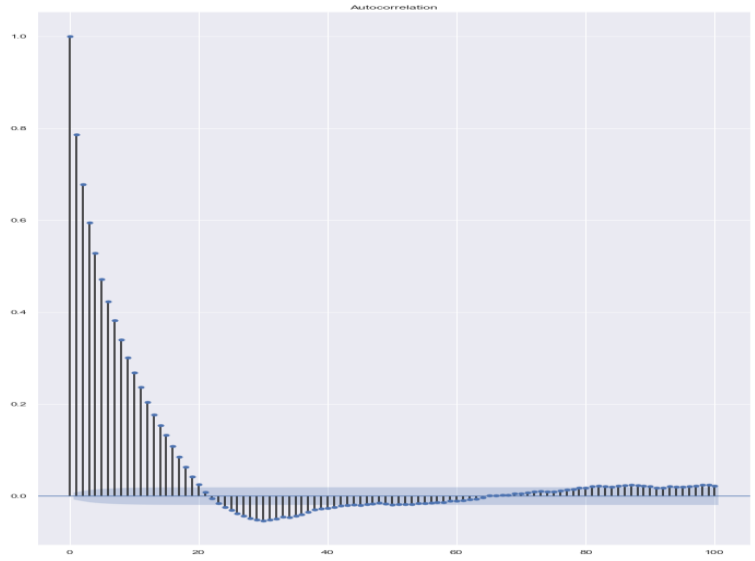


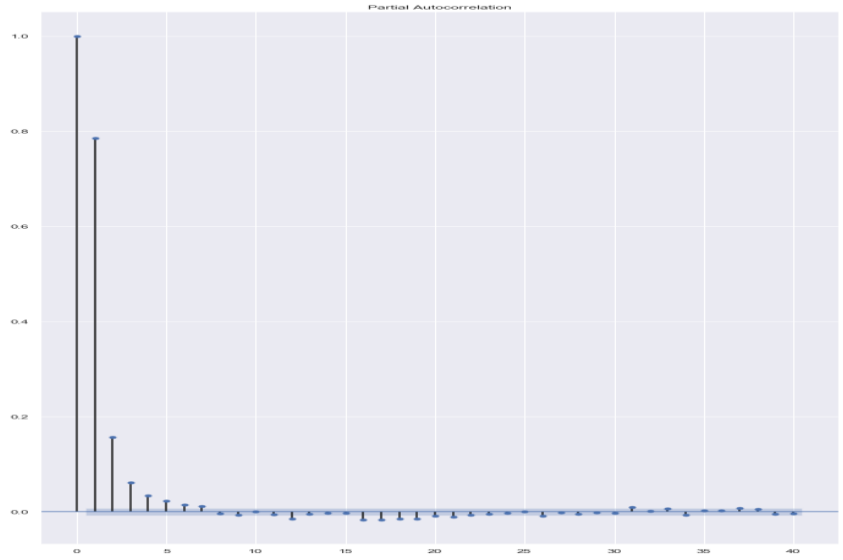
- Based on the box plot, we can see that the prices corresponding to 5 room flats have a right skewed distribution. This tells us that there are more observations with prices around $420,000 rather than being priced more than $600,000.
- The rest of the flat types generally have a symmetric distribution.
- There seems to be a number of outliers, based on the different flat types.

Box plots



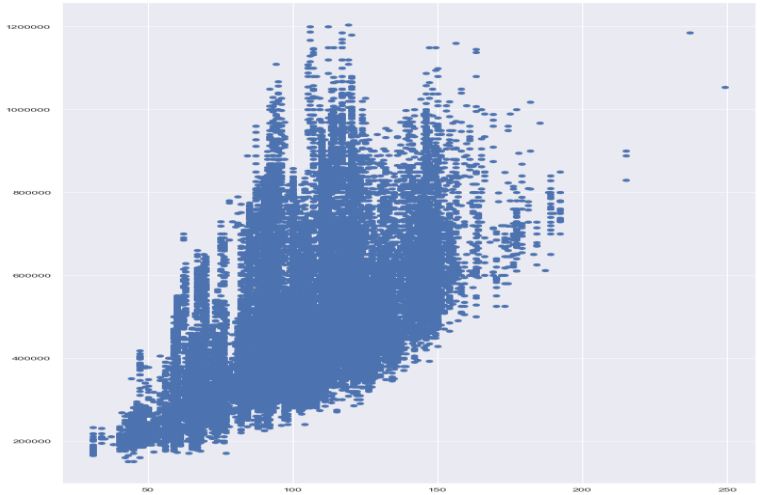
Plotting the time series, seasonality is not prevalent here. This is confirmed by the autocorrelation plot in the next slide.



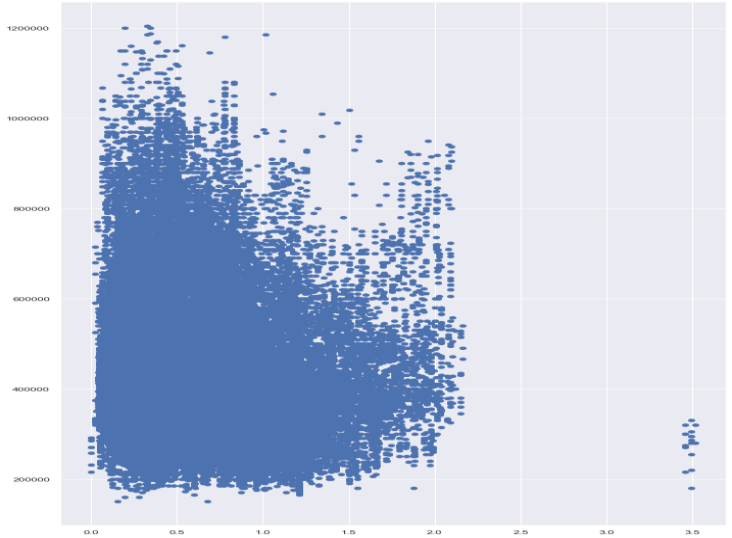


- Significant evidence of autocorrelation for lags >0 until lag 21. (total lags = 100)
- No seasonality present in time series.
- We can expect to observe an increasing trend in resale house prices once prices start to rise.

There are significant correlation at lag = 2, then followed by non-significant correlations. This suggests that AR(2) – autoregressive term of order 1 will be a suitable prediction model for the dataset.



floor_area_sqm vs resale_price



Distance to nearest MRT station vs resale_price

We plotted the scatter plots and above are some of the plots obtained. We can see that there seems to be a positive linear relationship between floor_area_sqm and resale_price. On the other hand, the distance to nearest MRT station seem to have a weak correlation with the response variable. This weak relationship is generally seen amongst the rest of the predictors and thus may suggest that a more complex model will work better for prediction.

Dataset used for modelling

The dataset consists of 66,497 data with 27 features, taken from 2017 onwards.

'lease_commencement_date' and 'month' is not included for building the models as remaining_lease is calculated using these two features, similarly for 'flat_type' and 'flat_model' and 'town'.

	min_storey	max_storey	floor_area_sqm	remaining_lease	Distance to nearest MRT station	Distance to nearest Primary School	Distance to nearest Shopping Mall	Distance to CBD	mature_estate	type_premium	...	Model A2	Multi Generation	New Generation	Premium Apartment	Simplified	Standard	Terrace	Type S1	Type S2	Others
0	10	12	44.0	61	1.000279	0.184712	1.000041	8.615607	1	-222888.0	...	0	0	0	0	0	0	0	0	0	0
1	01	03	67.0	61	1.268809	0.227339	0.871785	9.715041	1	-222888.0	...	0	0	1	0	0	0	0	0	0	0
2	01	03	67.0	63	1.072235	0.780672	1.527983	10.828734	1	-222888.0	...	0	0	1	0	0	0	0	0	0	0
3	04	06	68.0	62	0.946066	0.695564	1.027995	9.097905	1	-222888.0	...	0	0	1	0	0	0	0	0	0	0
4	01	03	67.0	63	1.095144	0.789146	1.571708	10.869368	1	-222888.0	...	0	0	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
66492	01	03	145.0	68	1.139079	0.119200	0.846877	15.707783	0	396000.0	...	0	0	0	0	0	0	0	0	0	0
66493	01	03	164.0	72	0.557491	0.585917	0.701735	15.316410	0	396000.0	...	0	0	0	0	0	0	0	0	0	0
66494	01	03	146.0	67	0.573533	0.498131	0.660099	15.265012	0	396000.0	...	0	0	0	0	0	0	0	0	0	0
66495	01	03	146.0	67	0.470970	0.524051	0.616043	15.142529	0	396000.0	...	0	0	0	0	0	0	0	0	0	0
66496	01	03	145.0	67	0.403879	0.525211	1.000586	14.474343	0	396000.0	...	0	0	0	0	0	0	0	0	0	0

3 Modelling

Including results obtained

Models


Stacking multiple models
Multiple Linear Regression
Random Forest (mtry = , number of decision trees = 1000, interaction_depth = 5)
Boosting
Multi-Layer Perceptron (MLP)

Multiple Linear Regression

After a 80-20 train-test split on the dataset, we fit it into a simple regression model and obtained the following coefficients for the intercept and variables.

```
Intercept:
 -4.921391724425419e-16
Coefficients:
 [ 0.09151514  0.09151514  0.6566977   0.33410327 -0.07409992  0.04516793
 -0.04041537 -0.36272714  0.21838925 -0.00232437  0.01386493  0.02794807
  0.1057294  -0.0279386   0.06506468 -0.03595318  0.01284799  0.00097908
  0.01935869  0.01126629 -0.05751659  0.02838513  0.00773897  0.04627041
  0.0470471   0.04467463  0.01258824]
Mean Squared Error: 0.16143391521059922
R2: 0.8385660847894008
```

MSE = 0.16143

Random Forest

We also fit the data into a non-parametric model, Random Forest with 1000 trees and interaction depth of 5.

MSE = 0.19478

```
Mean Squared Error: 0.1947754449750024
R2: 0.8052245550249976
                                  importance
floor_area_sqm                      0.567767
Distance to CBD                     0.339838
remaining_lease                     0.055932
DBSS                                0.011525
Model A                             0.011105
New Generation                      0.004086
mature_estate                       0.003304
Distance to nearest Shopping Mall   0.003131
min_storey                          0.001114
max_storey                          0.001057
Terrace                             0.000842
Distance to nearest MRT station     0.000114
Type S1                             0.000088
Distance to nearest Primary School  0.000058
Simplified                          0.000026
type_premium                        0.000009
Improved                            0.000004
Model A2                            0.000000
Multi Generation                    0.000000
Model A-Maisonette                  0.000000
Premium Apartment                   0.000000
Maisonette                          0.000000
Standard                            0.000000
Apartment                           0.000000
Adjoined flat                       0.000000
Type S2                             0.000000
Others                              0.000000
```

Boosting

We fit out dataset into the boosting model (400 trees) as well. Boosting improves the prediction accuracy by using and combining information from previously grown trees and building the new trees sequentially.

```
Mean Squared Error: 0.06221936457141065
R2: 0.9377806354285892
```

MSE = 0.062219

Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron is a feedforward neural network. We fit out model into a neural network containing 100 neurons in the dense layer and obtained the following results.

```
Mean Squared Error: 0.037354230941652 59
R2: 0.9626457690583474
```

MSE = 0.037354

Stacking multiple models

Stacked MLP, Boosting and RandomForest with the same hyperparameters as before for better prediction accuracy. The results obtained are as follows:

```
Mean Squared Error 0.0683789887696536
R2: 0.9316210112303464
```

MSE = 0.068379

4 Evaluation

What are the features affecting resale house prices?

Evaluation

Model	Mean Squared Error (MSE)
Multiple Linear Regression	0.16143
Random Forest	0.19478
Boosting	0.062219
MLP	0.037354
Stacking	0.068379

MLP is the best model yielding the lowest MSE of 0.037354.
We can see that generally, more complex model performs better than the simple linear regression model.

Variable Importance

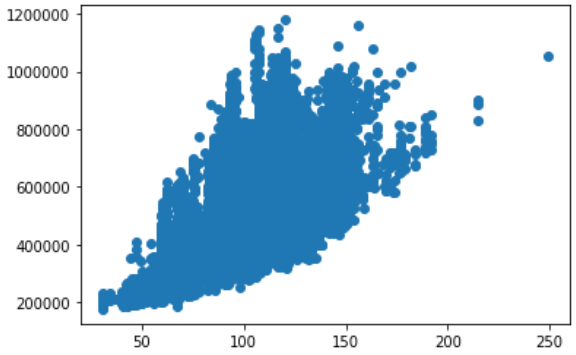
```
Intercept:
 -4.921391724425419e-16
Coefficients:
 [ 0.09151514  0.09151514  0.6566977   0.33410327 -0.07409992  0.04516793
 -0.04041537 -0.36272714  0.21838925 -0.00232437  0.01386493  0.02794807
  0.1057294  -0.0279386   0.06506468 -0.03595318  0.01284799  0.00097908
  0.01935869  0.01126629 -0.05751659  0.02838513  0.00773897  0.04627041
  0.0470471   0.04467463  0.01258824]
```

Coefficients from MLR

Taking FLOOR_AREA_SQM as an example, the coefficient tells us that for a unit change in this variable, the resale house price will increase by 0.6566977,keeping all other predictors constant. This tells us that it has the highest impact in affecting resale house price.

Random Forest

	importance
floor_area_sqm	0.567919
Distance to CBD	0.339795
remaining_lease	0.055935
DBSS	0.011488
Model A	0.011251
New Generation	0.003850
mature_estate	0.003320
Distance to nearest Shopping Mall	0.003103
min_storey	0.001118
max_storey	0.001080
Terrace	0.000858
Distance to nearest MRT station	0.000107
Type S1	0.000088
Distance to nearest Primary School	0.000057
Simplified	0.000019
type_premium	0.000009
Improved	0.000003
Model A2	0.000000
Multi Generation	0.000000
Model A-Maisonette	0.000000
Premium Apartment	0.000000
Maisonette	0.000000
Standard	0.000000
Apartment	0.000000
Adjoined flat	0.000000
Type S2	0.000000
Others	0.000000



floor_area_sqm vs resale_price

The top 4 features are floor_area_sqm, Distance to CBD and remaining_lease, DBSS.

This is in line with the scatter plots obtained which showed a moderate linear relationship between floor_area_sqm and resale_prices.

Additionally, DBSS, Model A or New Generation flats tend to affect the resale house prices more significantly than other flat models. We can also see that Distance to CBD area plays a huge part in affecting the prices as compared to other proximities.

Lastly, we can infer that the remaining number of years left in a flat's housing lease affects the prices too. Flats will more likely be in demand if there are more years left to the 99-year housing lease.

Conclusion

Conclusion

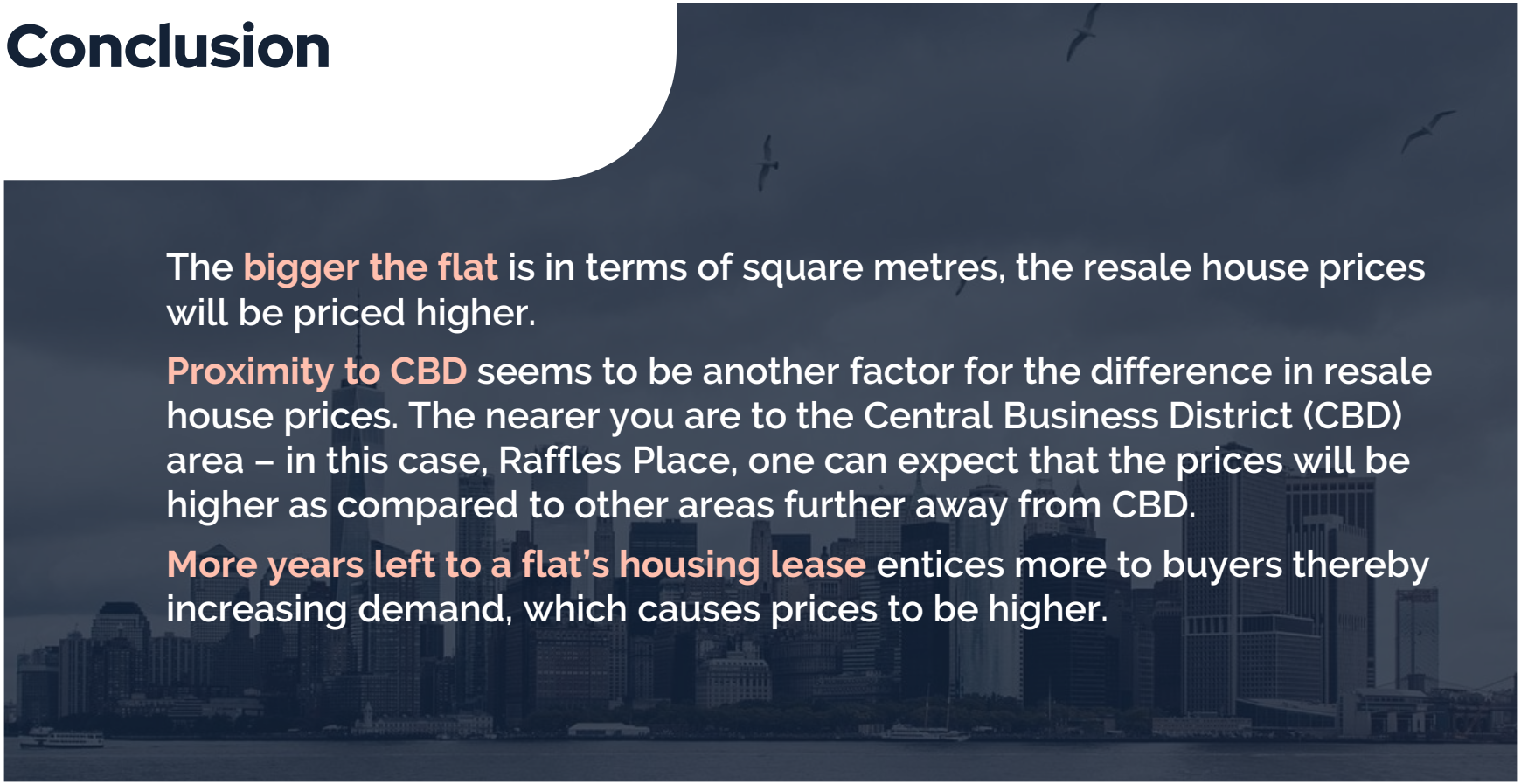
Important features affecting resale house prices: (top 3)

Flat area in square metres, Distance to MRT stations, Number of years left in housing lease.

Best Model

Multi-Layer Perceptron (MLP)

Conclusion



The **bigger the flat** is in terms of square metres, the resale house prices will be priced higher.

Proximity to CBD seems to be another factor for the difference in resale house prices. The nearer you are to the Central Business District (CBD) area – in this case, Raffles Place, one can expect that the prices will be higher as compared to other areas further away from CBD.

More years left to a flat's housing lease entices more to buyers thereby increasing demand, which causes prices to be higher.


Thank You!