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) Replacing ‘storey_range’ categorical variable with ‘storey’ range which outputs the mean of the level range.
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.
4) ‘flat_model’ is being encoded based on the number of categories. (binary variable)

	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

storey	floor_area_sqm	remaining_lease
11	44	61
2	67	61
2	67	63
5	68	62
2	67	63
2	68	63
5	68	62
5	67	58
5	68	62
2	67	61
2	68	62
11	67	60
5	67	60
8	67	60
8	68	62
5	67	60
11	67	61
5	68	63
8	67	61
5	68	63
8	67	61
5	73	60
11	67	61
2	67	59
5	67	62
8	74	60
8	68	63
11	73	60

'storey range' variable is replaced with 'storey' variable that averages the minimum storey and maximum storey given in the range. The values are now quantitative instead of categorical.

2) Replacing '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 are more expensive as compared to 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

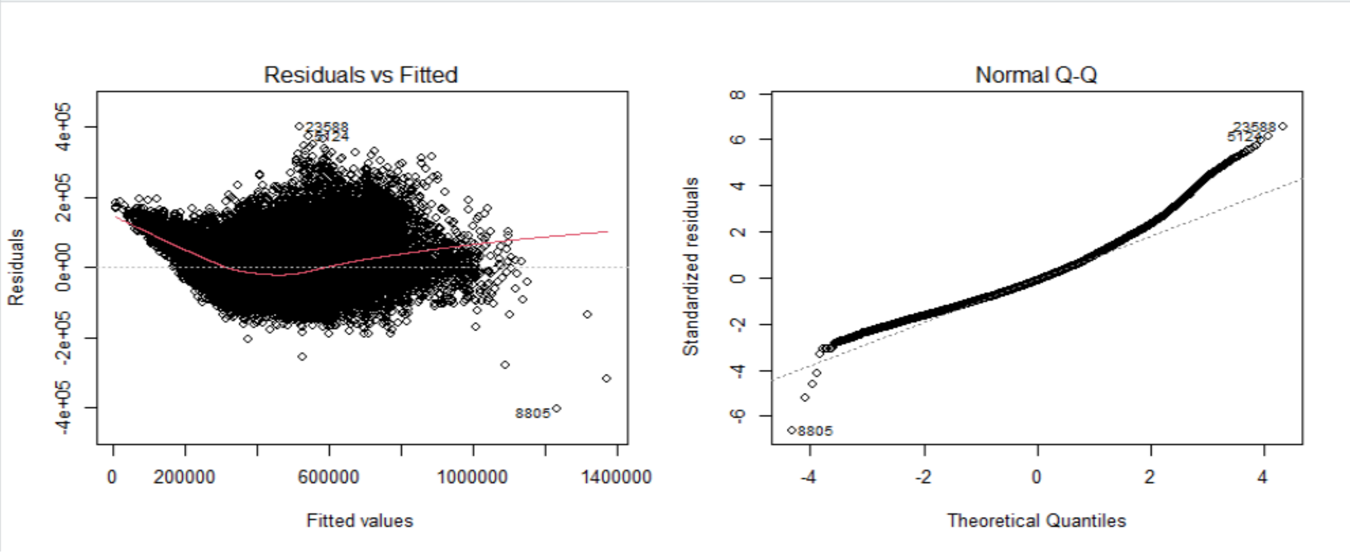
Columns in final dataset

```
dt_use.columns
```

```
Index(['month', 'town', 'flat_type', 'block', 'street_name', 'storey ‘      ,
       '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')
```

Addition of 27 variables

2 Exploratory Analysis



Based on the Residual vs Fitted values plot, we see that there is non-linearity in the data. Thus, it seems that models other than the multiple linear regression model will perform better.
From the normality plot, we can infer that a non-parametric model will perform better for this data as the data does not follow the normality QQ line.

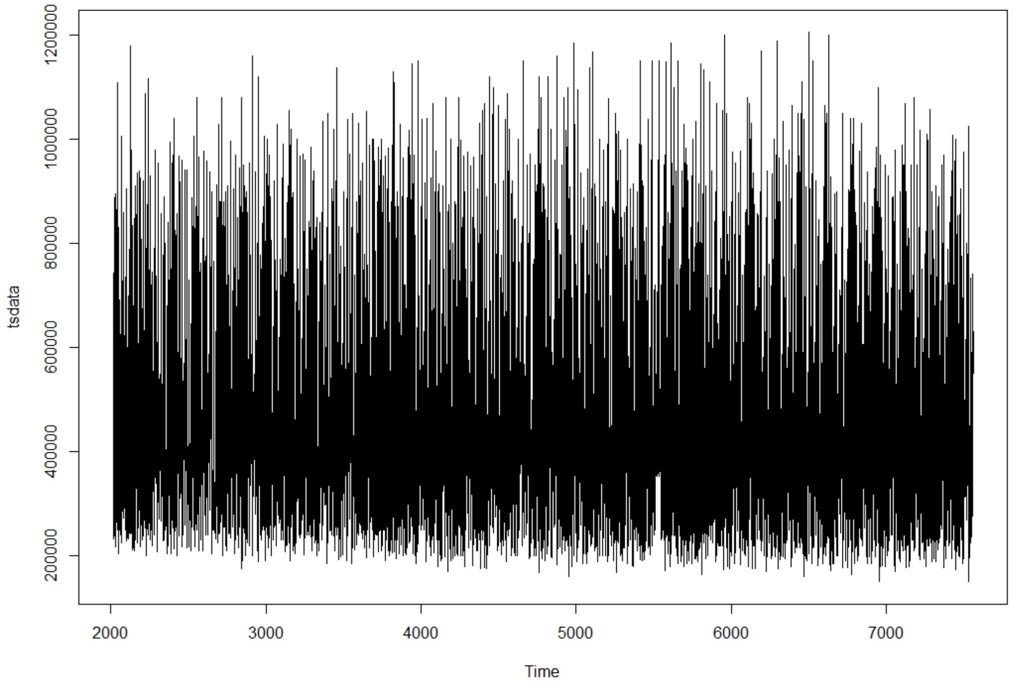
```
> vif(model)
                        storey                   floor_area_sqm                   remaining_lease
                      1.211482                         1.969859                          2.261614
 Distance.to.nearest.MRT.station Distance.to.nearest.Primary.School Distance.to.nearest.Shopping.Mall
                      1.210405                         1.119372                          1.232159
               Distance.to.CBD                     mature_estate                      type_premium
                      2.623927                         2.582868                          1.005086
                 Adjoined.flat                         Apartment                              DBSS
                     15.258561                       957.997900                        110.737222
                      Improved                        Maisonette                           Model.A
                   1415.148024                       231.045623                       1688.306075
            Model.A.Maisonette                          Model.A2                  Multi.Generation
                     13.443753                         1.042112                          4.739681
                New.Generation                 Premium.Apartment                        Simplified
                    907.911005                         3.995947                        311.204272
                      Standard                           Terrace                           Type.S1
                    206.786164                         5.541365                         14.363250
                       Type.S2                            Others
                      8.140022                         1.585377
```

Variance inflation factor values are calculated as shown above. We then drop the higher values amongst features, such as Adjoined flat, Apartment, Standard, Model A and Simplified.

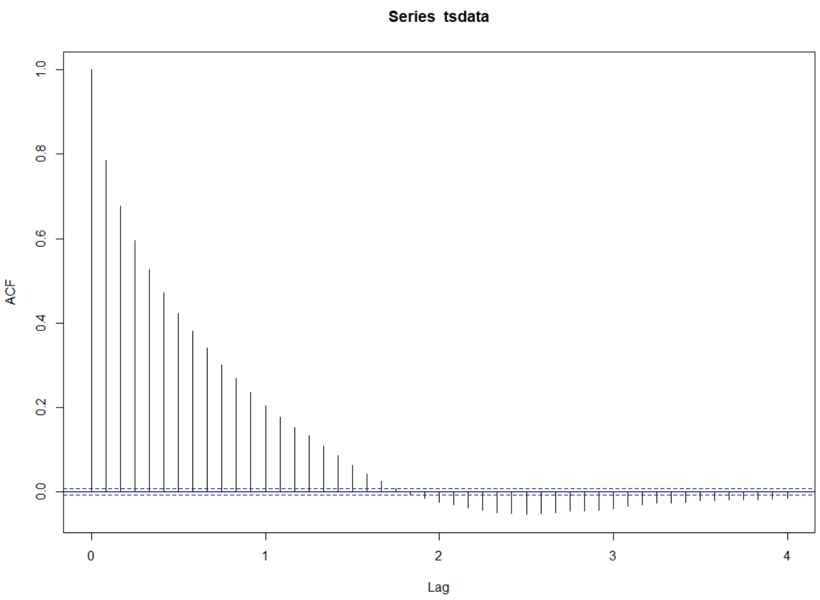
```
> vif(model1)
                        storey                   floor_area_sqm                   remaining_lease
                      1.204540                         1.474418                          1.866476
 Distance.to.nearest.MRT.station Distance.to.nearest.Primary.School Distance.to.nearest.Shopping.Mall
                      1.209131                         1.110929                          1.227755
               Distance.to.CBD                     mature_estate                      type_premium
                      2.579157                         2.571332                          1.004391
                          DBSS                          Improved                        Maisonette
                      1.105056                         1.273067                          1.333732
            Model.A.Maisonette                          Model.A2                  Multi.Generation
                      1.057139                         1.034263                          1.007143
                New.Generation                 Premium.Apartment                           Terrace
                      1.428346                         1.262616                          1.008168
                       Type.S1                           Type.S2                            Others
                      1.034836                         1.019120                          1.008053
>
```

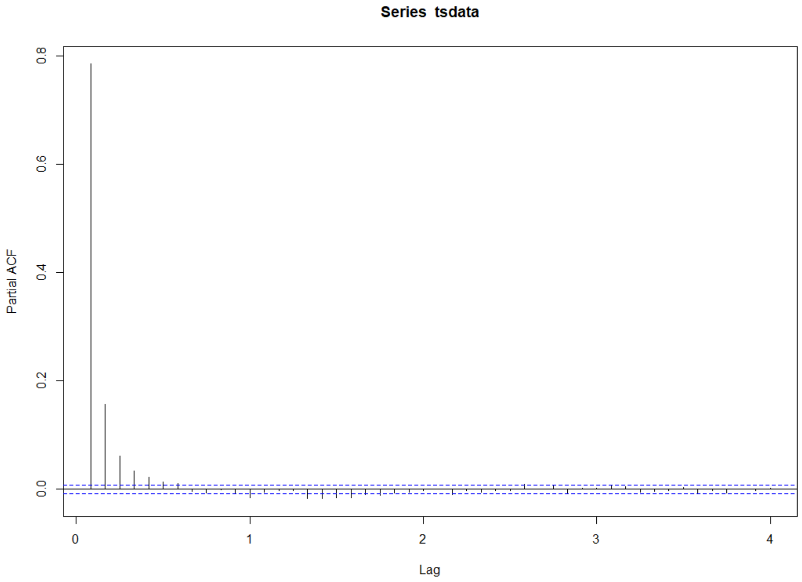
After dropping these variables, we see that there are no features that have a very high VIF value (>10) anymore. These will then be the features used for the analysis.

Checking for multicollinearity



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





- Geometrically decaying of autocorrelation values.
- No seasonality present in time series.

There are significant correlation at lag = 1, then followed by non-significant correlations.

This suggests that AR(1) – autoregressive term of order 1 will be a suitable prediction model for the dataset.

Exploratory Analysis of the time series

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'.

```
> names(dat)
 [1] "storey"                            "floor_area_sqm"                    "remaining_lease"
 [4] "resale_price"                      "Distance.to.nearest.MRT.station"   "Distance.to.nearest.Primary.School"
 [7] "Distance.to.nearest.Shopping.Mall" "Distance.to.CBD"                   "mature_estate"
[10] "type_premium"                      "Adjoined.flat"                     "Apartment"
[13] "DBSS"                              "Improved"                          "Maisonette"
[16] "Model.A"                           "Model.A.Maisonette"                "Model.A2"
[19] "Multi.Generation"                  "New.Generation"                    "Premium.Apartment"
[22] "Simplified"                        "Standard"                          "Terrace"
[25] "Type.S1"                           "Type.S2"                           "Others"
```

Features used to build model

3 Modelling

Including results obtained

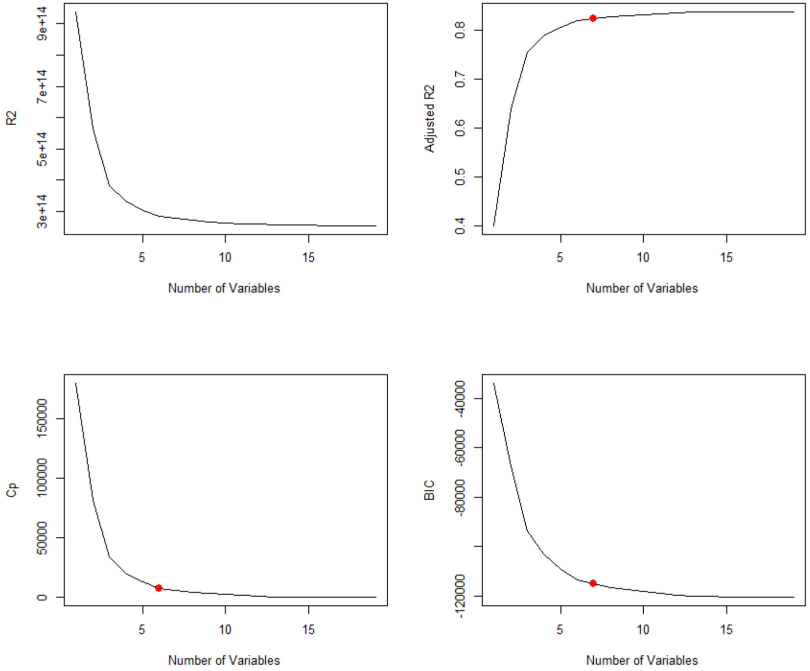
Models

Multiple Linear Regression

Random Forest (number of predictors to consider at each split is 9)

Boosting (Number of trees = 3300, interaction depth = 8)

ARIMA(1,1,1)(0,0,2) model



Forward selection is performed and 7 significant predictors are chosen based on the Cp, BIC and Adjusted R-squared plots.

```
> coef(regfit.fwd, 7)
         (Intercept)                      storey               floor_area_sqm
         -154386.773                    4902.566                     4370.108
     remaining_lease Distance.to.nearest.MRT.station          Distance.to.CBD
            3684.157                  -28245.000                   -13207.035
       mature_estate                        DBSS
           64699.436                  146103.007
```

ARIMA model

```
> model_ar
Series: tsdata
ARIMA(1,1,1)(0,0,2)[12]

Coefficients:
         ar1      ma1     sma1     sma2
      0.5399  -0.8244  -0.0306  -0.0272
s.e.  0.0113   0.0085   0.0039   0.0039
```

We obtained the following coefficients when fitting a ARIMA(1,1,1)(0,0,2) model.

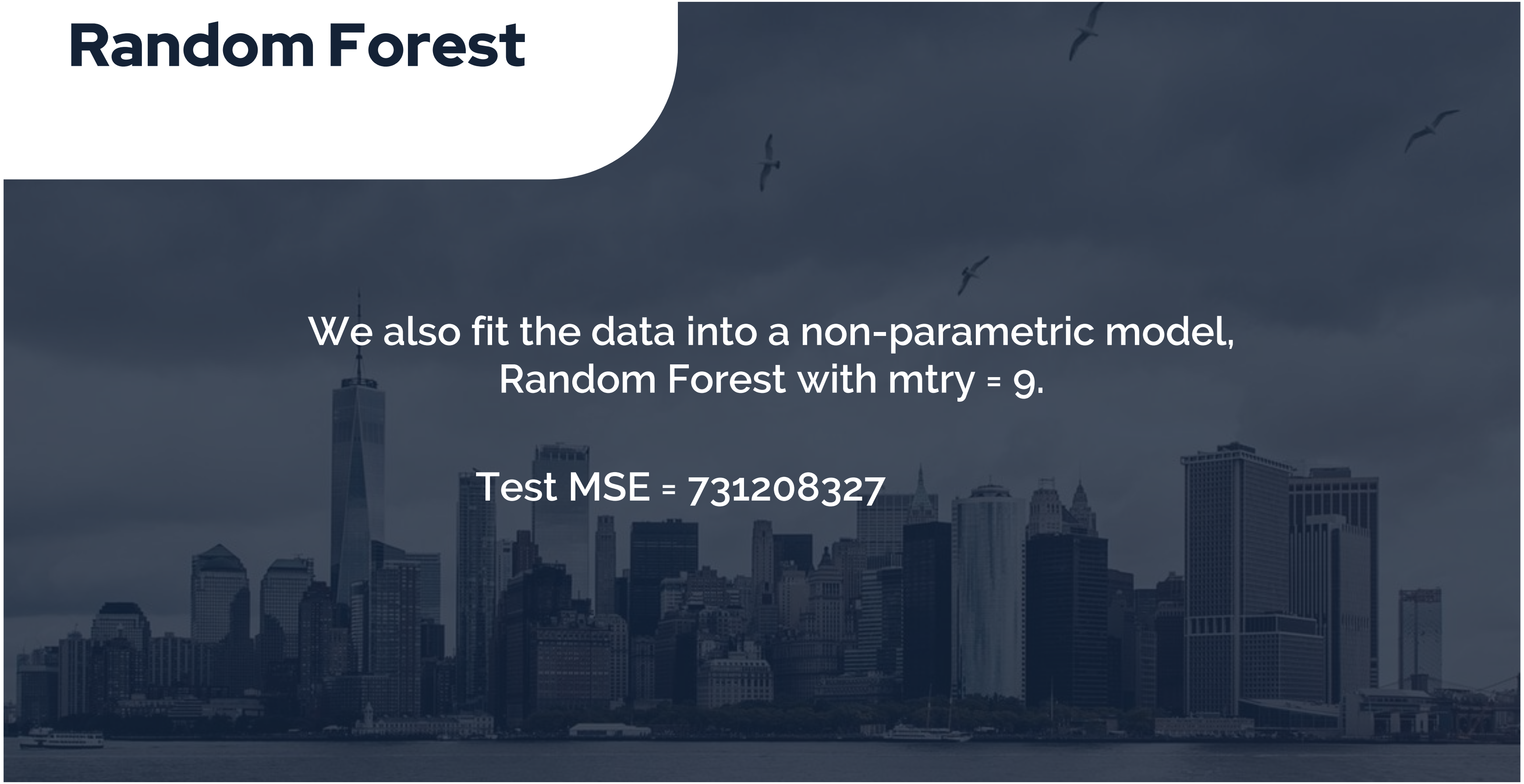
Multiple Linear Regression

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.

```
> summary(linearmodel)

Call:
lm(formula = resale_price ~ storey + floor_area_sqm + remaining_lease +
    Distance.to.CBD + mature_estate + DBSS + Distance.to.nearest.MRT.station,
    data = dat[train, ])

Residuals:
    Min      1Q  Median      3Q     Max
-251239  -44891   -7043   36950  497431

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -154970.65    2561.63  -60.50   <2e-16 ***
storey                            4931.60      52.66   93.64   <2e-16 ***
floor_area_sqm                    4365.48      12.10  360.93   <2e-16 ***
remaining_lease                   3673.61      27.15  135.33   <2e-16 ***
Distance.to.CBD                 -13097.89      98.13 -133.48   <2e-16 ***
mature_estate                    65180.91     897.78   72.60   <2e-16 ***
DBSS                            145954.24    2403.71   60.72   <2e-16 ***
Distance.to.nearest.MRT.station -28622.70     752.54  -38.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64390 on 53189 degrees of freedom
Multiple R-squared:  0.8239,    Adjusted R-squared:  0.8239
F-statistic: 3.556e+04 on 7 and 53189 DF,  p-value: < 2.2e-16
```

MSE = 3872241663

Random Forest



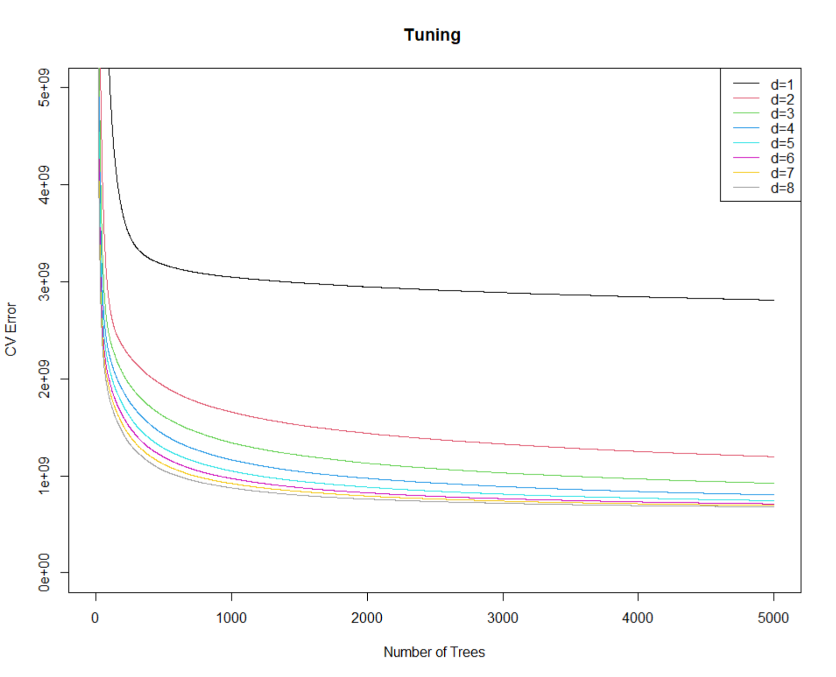
We also fit the data into a non-parametric model, Random Forest with mtry = 9.

Test MSE = 731208327

Boosting

Using a 10-fold cross validation, the optimal number of trees is 3300 with an interaction depth = 8 and shrinkage =0.1.

Test MSE = 725409220



4 Evaluation

What are the features affecting resale house prices?

Evaluation

Model	Mean Squared Error (MSE)
Multiple Linear Regression	3872241663
Random Forest	731208327
Boosting	725409220

Boosting is the best model yielding the lowest MSE of 725409220.

We can see that generally, more complex model performs better than the simple linear regression model.

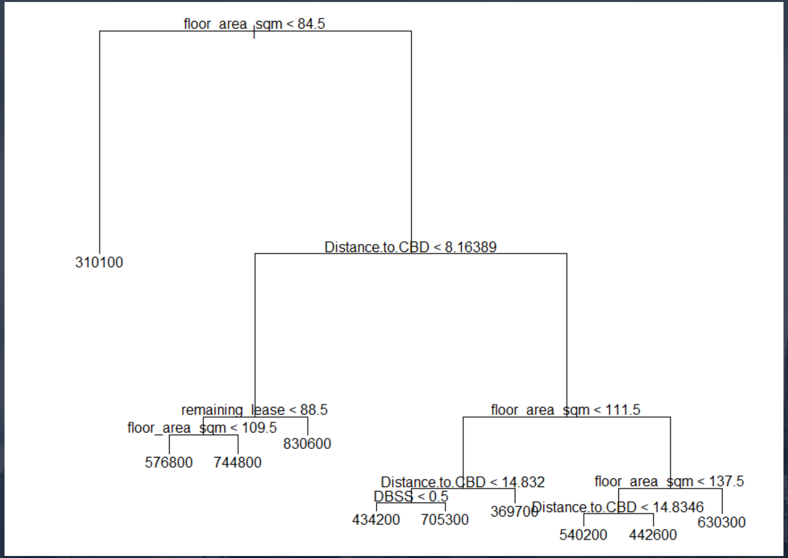
Variable Importance

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -154970.65    2561.63  -60.50   <2e-16 ***
storey                         4931.60      52.66   93.64   <2e-16 ***
floor_area_sqm                 4365.48      12.10  360.93   <2e-16 ***
remaining_lease                3673.61      27.15  135.33   <2e-16 ***
Distance.to.CBD              -13097.89      98.13 -133.48   <2e-16 ***
mature_estate                 65180.91     897.78   72.60   <2e-16 ***
DBSS                         145954.24    2403.71   60.72   <2e-16 ***
Distance.to.nearest.MRT.station  -28622.70     752.54  -38.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64390 on 53189 degrees of freedom
Multiple R-squared:  0.8239,	Adjusted R-squared:  0.8239 
F-statistic: 3.556e+04 on 7 and 53189 DF,  p-value: < 2.2e-16
```

Coefficients from MLR

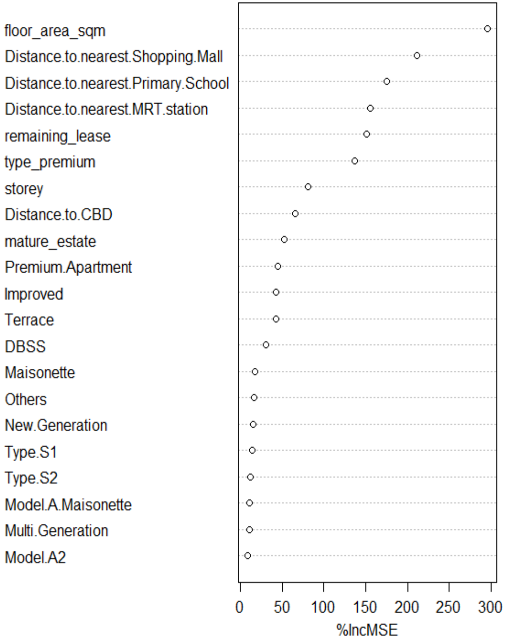
Based on the multiple linear model, DBSS’s coefficient tells us that for a unit change in this variable, the resale house price will increase by 145954.24, keeping all other predictors constant. This tells us that it has the highest impact in affecting resale house price. However, it seems that the standard error is the highest for this variable and thus this conclusion on variable importance should be view with caution.

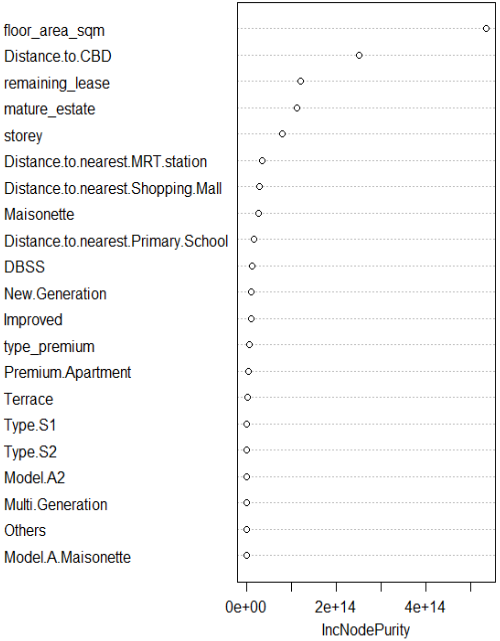
Decision Tree



We can see that the variable at the top of the decision tree is ***floor_area_sqm*** followed by ***Distance to CBD*** and then ***remaining_lease***. This tells us that ***floor_area_sqm*** is the most significant in affecting the resale flat prices.

Random Forest

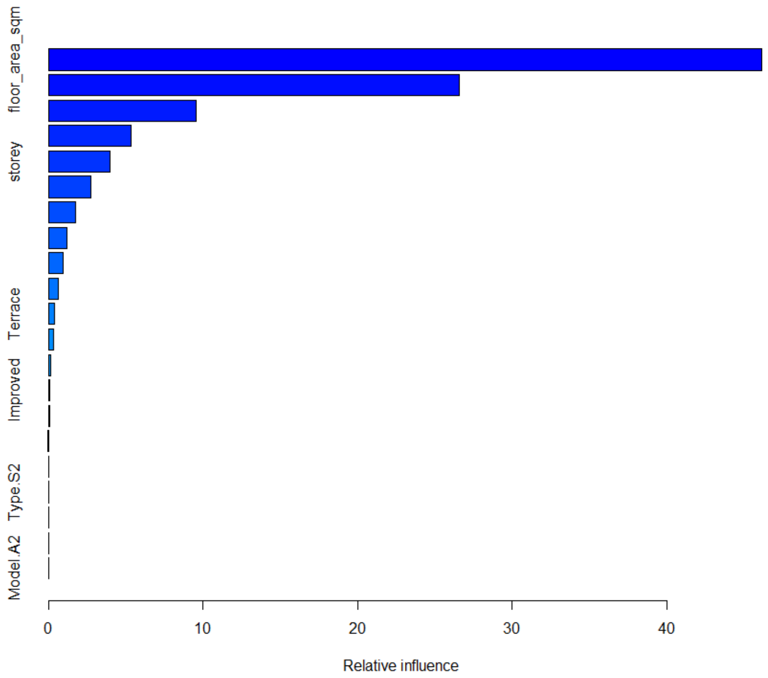





```
> importance(rf)
                                      %IncMSE IncNodePurity
storey                              81.097955  7.842021e+13
floor_area_sqm                     295.284303  5.350555e+14
remaining_lease                    150.662933  1.209053e+14
Distance.to.nearest.MRT.station    155.585115  3.448441e+13
Distance.to.nearest.Primary.School 175.525537  1.662282e+13
Distance.to.nearest.Shopping.Mall  211.071170  2.759082e+13
Distance.to.CBD                     65.472582  2.506382e+14
mature_estate                       52.986737  1.126161e+14
type_premium                       137.017204  5.162826e+12
DBSS                                30.492567  1.189272e+13
Improved                            43.025798  8.861547e+12
Maisonette                          17.521713  2.665173e+13
Model.A.Maisonette                  11.132575  1.047987e+11
Model.A2                             9.109145  2.673125e+11
Multi.Generation                    11.126378  2.577878e+11
New.Generation                      15.199486  1.038075e+13
Premium.Apartment                   44.540655  3.344686e+12
Terrace                             42.852227  2.026481e+12
Type.S1                             14.699080  6.424866e+11
Type.S2                             12.002911  4.459013e+11
Others                              16.885249  2.470836e+11
```

(not in order)

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




```
> summary(boost)
                                                            var      rel.inf
floor_area_sqm                                   floor_area_sqm 46.149516723
Distance.to.CBD                                 Distance.to.CBD 26.575961457
remaining_lease                                 remaining_lease  9.542479439
mature_estate                                     mature_estate  5.352846351
storey                                                   storey  3.994869508
Distance.to.nearest.MRT.station     Distance.to.nearest.MRT.station  2.744062840
Distance.to.nearest.Shopping.Mall   Distance.to.nearest.Shopping.Mall  1.742328157
Distance.to.nearest.Primary.School Distance.to.nearest.Primary.School  1.177789758
New.Generation                                   New.Generation  0.975713513
DBSS                                                       DBSS  0.649715318
Terrace                                                 Terrace  0.364403763
type_premium                                       type_premium  0.310048960
Premium.Apartment                             Premium.Apartment  0.139056523
Improved                                               Improved  0.093160653
Maisonette                                           Maisonette  0.076946589
Others                                                   Others  0.045301002
Multi.Generation                               Multi.Generation  0.028337555
Type.S2                                                 Type.S2  0.023887146
Model.A.Maisonette                           Model.A.Maisonette  0.008015732
Type.S1                                                 Type.S1  0.003285789
Model.A2                                               Model.A2  0.002273226
```

Similar results on variable importance is obtained for Boosting.

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!