HDB Resale Price

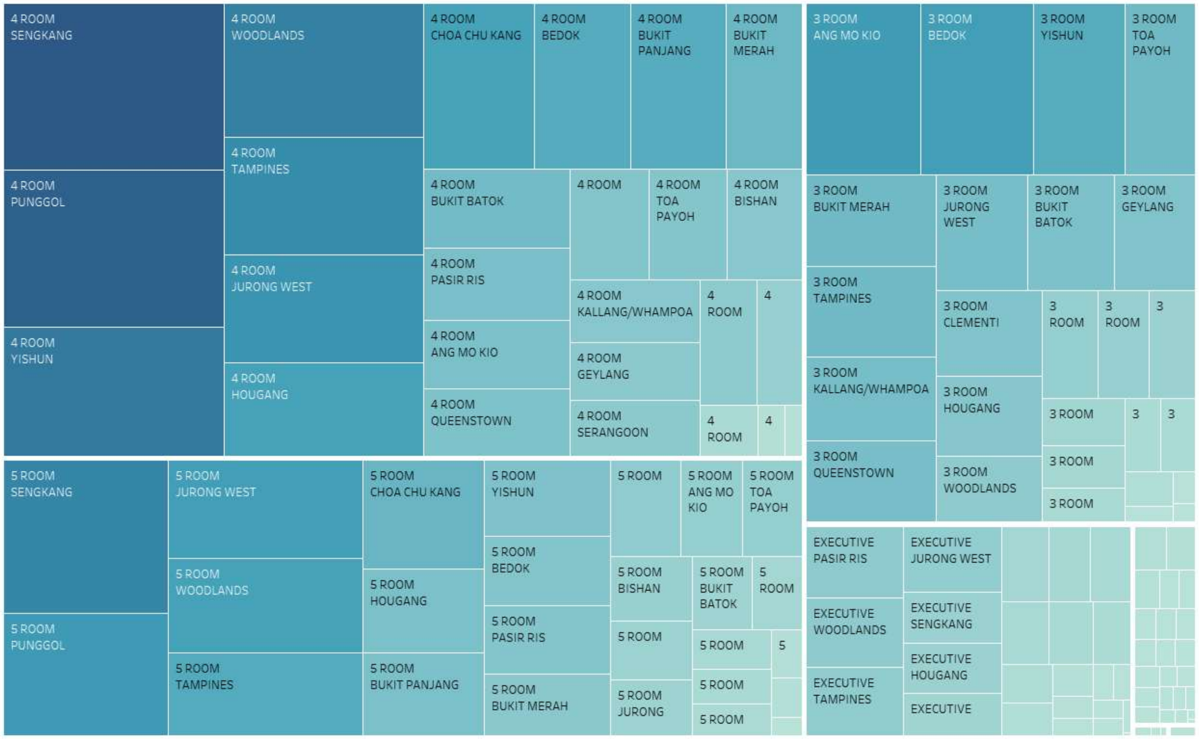
Machine Learning Linear Regression

Problem Statement

- With HDB resale price on the rise, there is a need to better predict sale price in order to budget for future purchase.
- Price prediction is based on several factors such as location, flat type and lease remaining.
- Dataset is obtained from singstat.gov.sg and data covers transaction from Jan 2017 to Jan 2022.

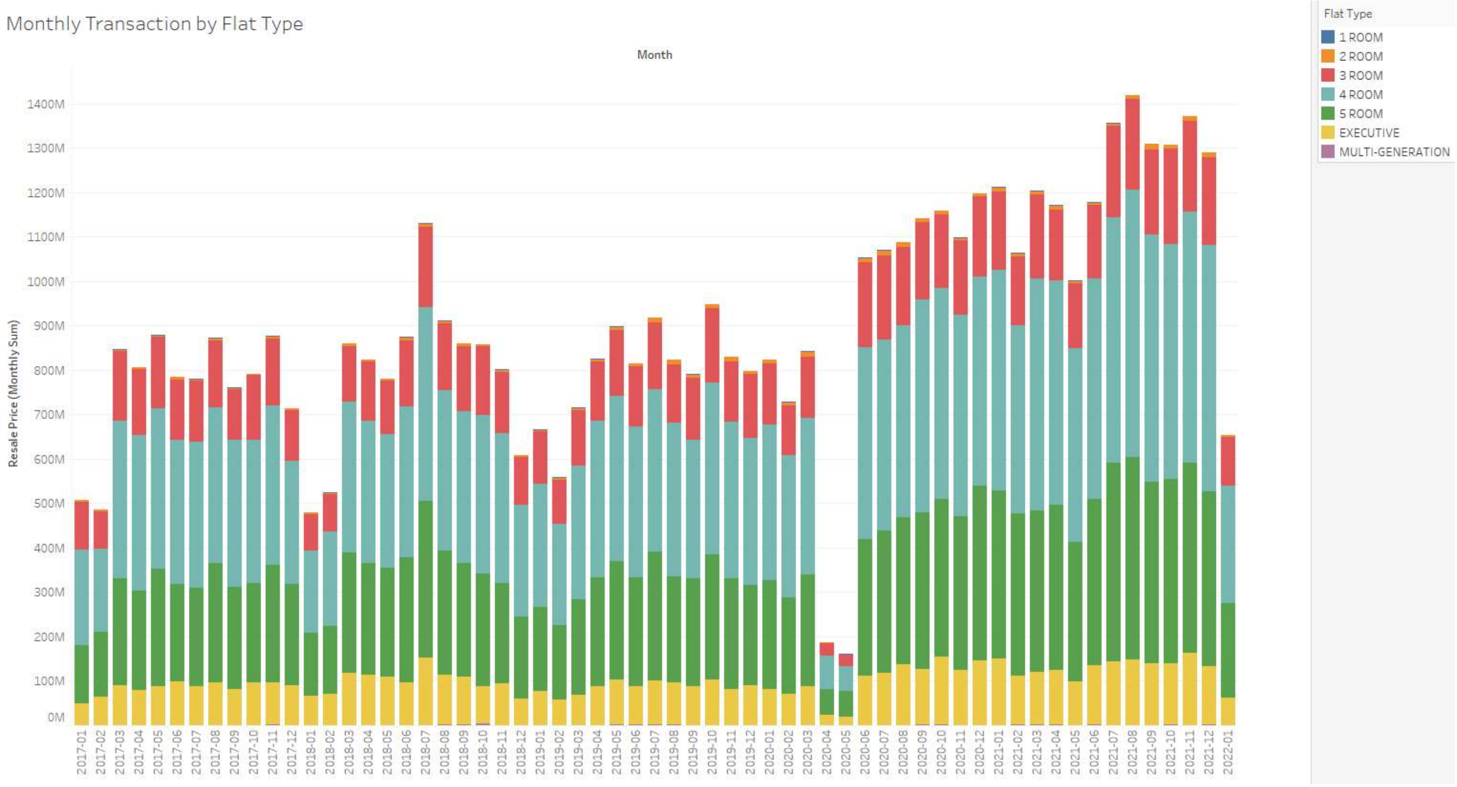
Data Observations

Flat Type and Location



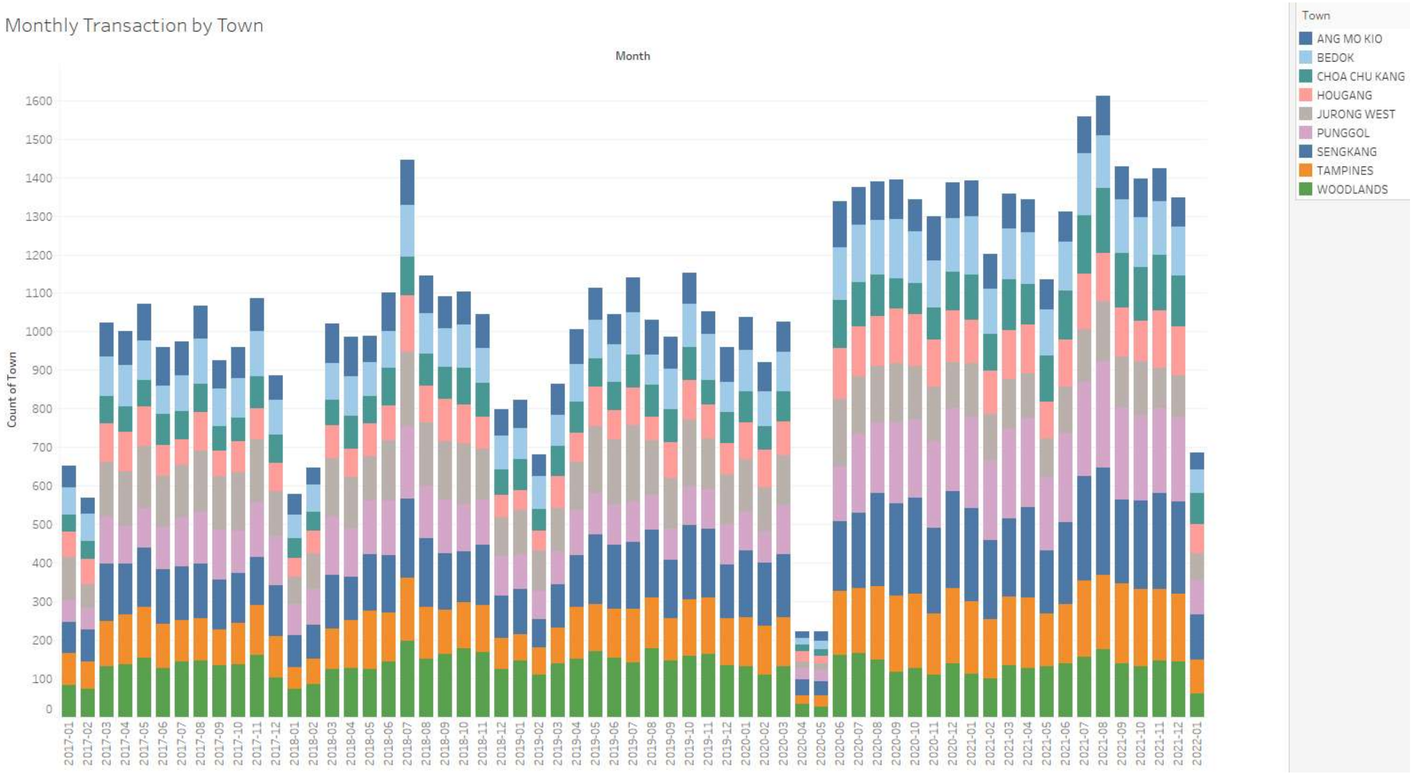
- Majority of resale flat type are 4-room and 5 room in Sengkang, Punggol, Yishun, Woodlands, Tampines.

Data Observations



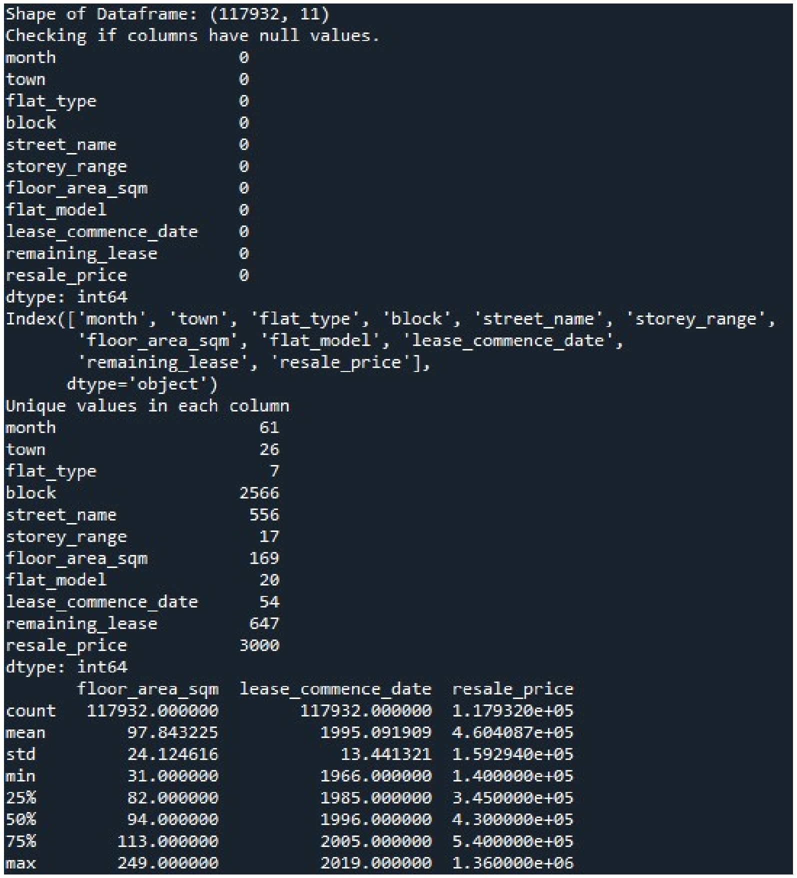
- 4 room and 5 room make up the bulk of sale.
- Sale were low in April/May 2020 due to circuit breaker
- Trend of rising prices over the years

Data Observations

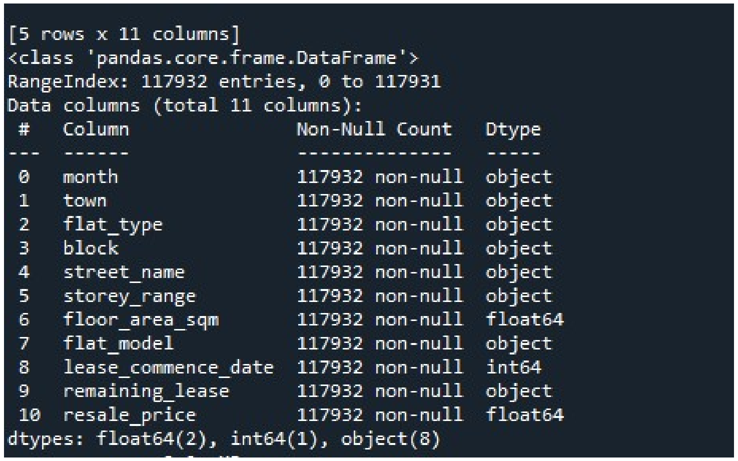


- Number of transaction in Singapore’s 10 most populous town.
- Sale were low in April/May 2020 due to circuit breaker
- Trend of rising prices over the years
- Sengkang and Punngol had the most number of sale

EDA



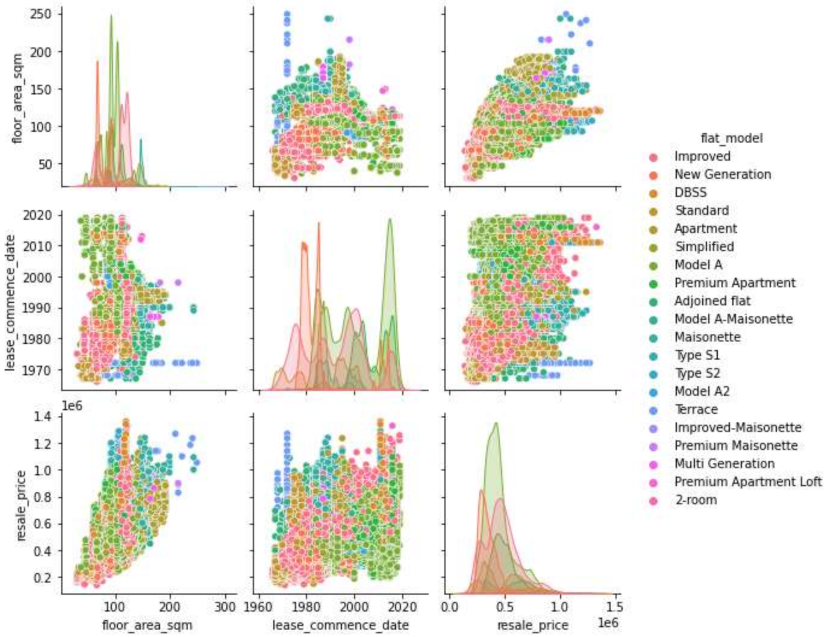
```
Shape of Dataframe: (117932, 11)
Checking if columns have null values.
month                  0
town                   0
flat_type              0
block                  0
street_name            0
storey_range           0
floor_area_sqm         0
flat_model             0
lease_commence_date    0
remaining_lease        0
resale_price           0
dtype: int64
Index(['month', 'town', 'flat_type', 'block', 'street_name', 'storey_range',
       'floor_area_sqm', 'flat_model', 'lease_commence_date',
       'remaining_lease', 'resale_price'],
      dtype='object')
Unique values in each column
month                    61
town                     26
flat_type                 7
block                  2566
street_name             556
storey_range             17
floor_area_sqm          169
flat_model               20
lease_commence_date      54
remaining_lease         647
resale_price           3000
dtype: int64
       floor_area_sqm  lease_commence_date  resale_price
count   117932.000000        117932.000000  1.179320e+05
mean        97.843225          1995.091909  4.604087e+05
std         24.124616            13.441321  1.592940e+05
min         31.000000          1966.000000  1.400000e+05
25%         82.000000          1985.000000  3.450000e+05
50%         94.000000          1996.000000  4.300000e+05
75%        113.000000          2005.000000  5.400000e+05
max        249.000000          2019.000000  1.360000e+06
```

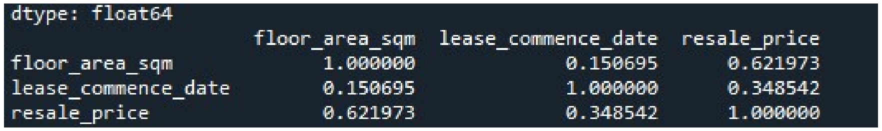


```
[5 rows x 11 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 117932 entries, 0 to 117931
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   month                117932 non-null  object 
 1   town                 117932 non-null  object 
 2   flat_type            117932 non-null  object 
 3   block                117932 non-null  object 
 4   street_name          117932 non-null  object 
 5   storey_range         117932 non-null  object 
 6   floor_area_sqm       117932 non-null  float64
 7   flat_model           117932 non-null  object 
 8   lease_commence_date  117932 non-null  int64  
 9   remaining_lease      117932 non-null  object 
 10  resale_price         117932 non-null  float64
dtypes: float64(2), int64(1), object(8)
```

- Dataframe has 117,932 entries with 11 columns
- No null values found.
- Datatype in Dataframe contains object, float and integer.

EDA

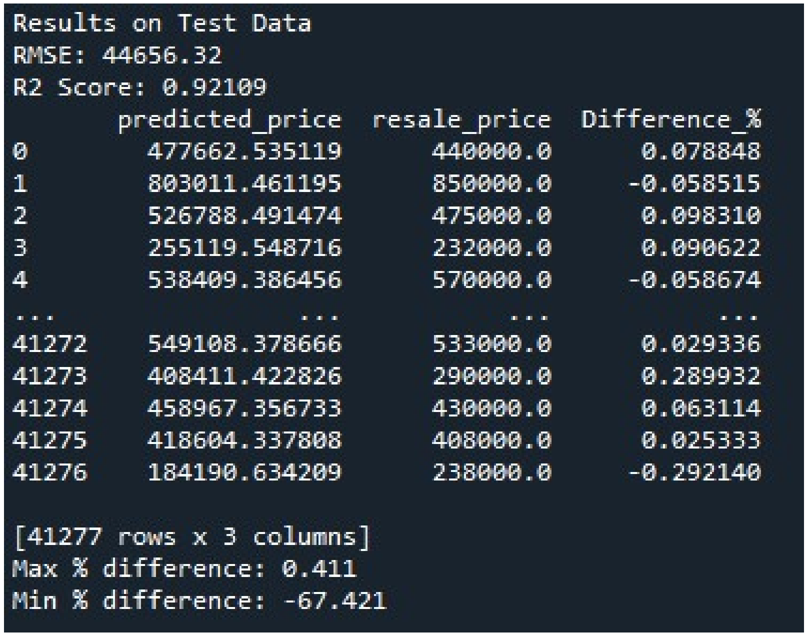




```
dtype: float64
                     floor_area_sqm  lease_commence_date  resale_price
floor_area_sqm             1.000000             0.150695      0.621973
lease_commence_date        0.150695             1.000000      0.348542
resale_price               0.621973             0.348542      1.000000
```

- Strong correlation of 0.621 between floor_area_sqm and resale_price. This suggests that a higher floor area will result in a higher resale price which is generally true.
- Lease commence date also has a correlation with resale price. Generally, a later lease commence date will fetch a higher resale price as there are more years remaining in the flat.
- Floor area and lease commence date do not seem to have a correlation given a score of 0.151.

Results and Suggestions



```
Results on Test Data
RMSE: 44656.32
R2 Score: 0.92109
       predicted_price  resale_price  Difference_%
0        477662.535119      440000.0      0.078848
1        803011.461195      850000.0     -0.058515
2        526788.491474      475000.0      0.098310
3        255119.548716      232000.0      0.090622
4        538409.386456      570000.0     -0.058674
...                ...           ...           ...
41272    549108.378666      533000.0      0.029336
41273    408411.422826      290000.0      0.289932
41274    458967.356733      430000.0      0.063114
41275    418604.337808      408000.0      0.025333
41276    184190.634209      238000.0     -0.292140

[41277 rows x 3 columns]
Max % difference: 0.411
Min % difference: -67.421
```

- LinearRegression model was used with accuracy of 92%
- Predicted price differs from -67% to 0.4%. (See attached excel file for full comparison)
- Random Forest model may be tested to see if it brings about greater accuracy.
- As it stands, current model with 92% accuracy is sufficient.
- Past sales data from before 2017 may be used to train model for greater accuracy.