



REAL ESTATE ANALYSIS

EXTRACT AND UNDERSTAND THE DRIVERS
OF HOUSING PRICES IN SINGAPORE

GUIDELINE

1. DATA EXPLORATION

- Data collection
- Final dataset

2. TIME SERIES ANALYSIS (from 1990-2020)

- Understand the evolution of the square meter price

3. DRIVERS EXTRACTION

- Feature selection methods

4. HOUSING PRICE PREDICTION

- Compare ML models to understand the role of each feature

1. Data Exploration

Data Collection



Final Dataset



| | month | town | floor_area_sqm | flat_type | flat_model | price | storey | top_floor | Postcode | LATITUDE | LONGITUDE | District | mrt_dist | primary_dist | secondary_dist | agency_dist | orchard_dist |
|---|--------|------|----------------|-----------|------------|-----------|--------|-----------|----------|----------|-----------|----------|----------|--------------|----------------|-------------|--------------|
| 0 | 736846 | 1 | 85.00 | 4 | 1 | 600000.00 | 11 | 0 | 310142 | 1.34 | 103.85 | 12 | 0.44 | 0.34 | 0.55 | 0.59 | 3.82 |
| 1 | 736846 | 1 | 101.00 | 4 | 1 | 700000.00 | 17 | 0 | 310121 | 1.34 | 103.85 | 12 | 0.16 | 0.28 | 0.70 | 0.82 | 4.21 |
| 2 | 736846 | 1 | 100.00 | 4 | 1 | 780000.00 | 14 | 0 | 310154 | 1.33 | 103.85 | 12 | 0.17 | 0.32 | 0.44 | 0.33 | 3.61 |



Month

Start Date : 1990-01

End Date : 2020-01

Town

```
# List of all towns in the dataset (Central Area, Bishan, Geylang...)  
df.town.unique()
```

```
array(['ANG MO KIO', 'BEDOK', 'BISHAN', 'BUKIT BATOK', 'BUKIT MERAH',  
      'BUKIT TIMAH', 'CENTRAL AREA', 'CHOA CHU KANG', 'CLEMENTI',  
      'GEYLANG', 'HOUGANG', 'JURONG EAST', 'JURONG WEST',  
      'KALLANG/WHAMPOA', 'MARINE PARADE', 'QUEENSTOWN', 'SENGKANG',  
      'SERANGOON', 'TAMPINES', 'TOA PAYOH', 'WOODLANDS', 'YISHUN',  
      'LIM CHU KANG', 'SEMBAWANG', 'BUKIT PANJANG', 'PASIR RIS',  
      'PUNGGOL'], dtype=object)
```

Flat_type

```
df.flat_type.value_counts()
```

| | |
|------------------|--------|
| 4 ROOM | 302897 |
| 3 ROOM | 268309 |
| 5 ROOM | 166614 |
| EXECUTIVE | 61474 |
| 2 ROOM | 9530 |
| 1 ROOM | 1234 |
| MULTI-GENERATION | 498 |

Name: flat_type, dtype: int64

Flat_model

```
df.flat_model.unique()
```

```
array(['improved', 'new generation', 'model a', 'standard', 'simplified',  
      'model a-maisonette', 'apartment', 'maisonette', 'terrace',  
      '2-room', 'improved-maisonette', 'multi generation',  
      'premium apartment', 'adjoined flat', 'premium maisonette',  
      'model a2', 'dbss', 'type s1', 'type s2', 'premium apartment loft'],  
      dtype=object)
```

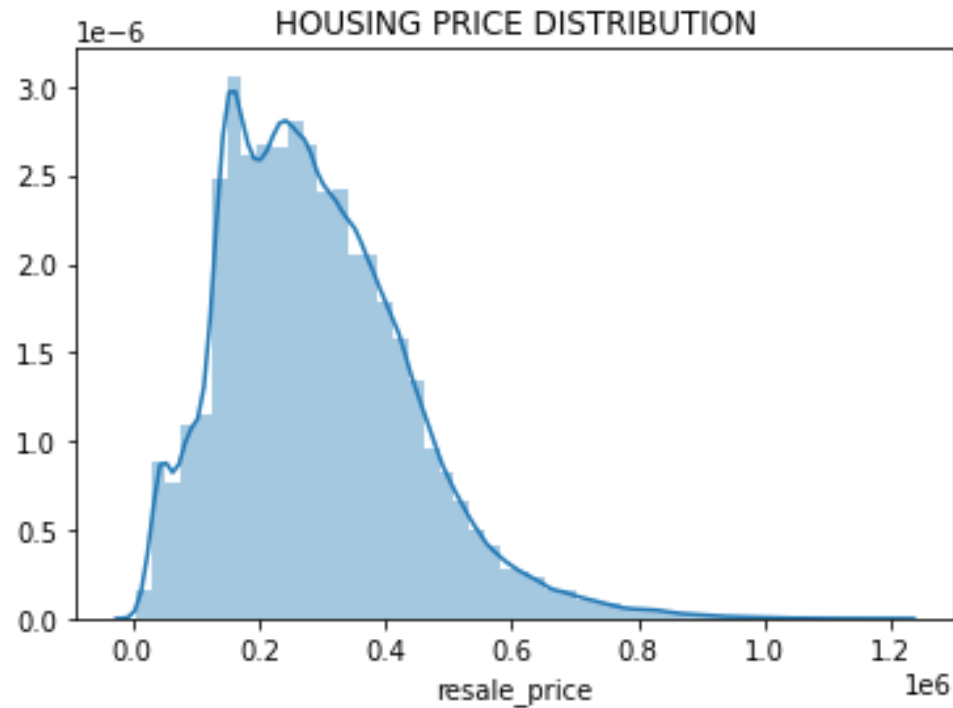
Storey

```
df.storey_range.unique()
```

```
array(['10 TO 12', '04 TO 06', '07 TO 09', '01 TO 03', '13 TO 15',  
      '19 TO 21', '16 TO 18', '25 TO 27', '22 TO 24', '28 TO 30',  
      '31 TO 33', '40 TO 42', '37 TO 39', '34 TO 36', '06 TO 10',  
      '01 TO 05', '11 TO 15', '16 TO 20', '21 TO 25', '26 TO 30',  
      '36 TO 40', '31 TO 35', '46 TO 48', '43 TO 45', '49 TO 51'],  
      dtype=object)
```

Price

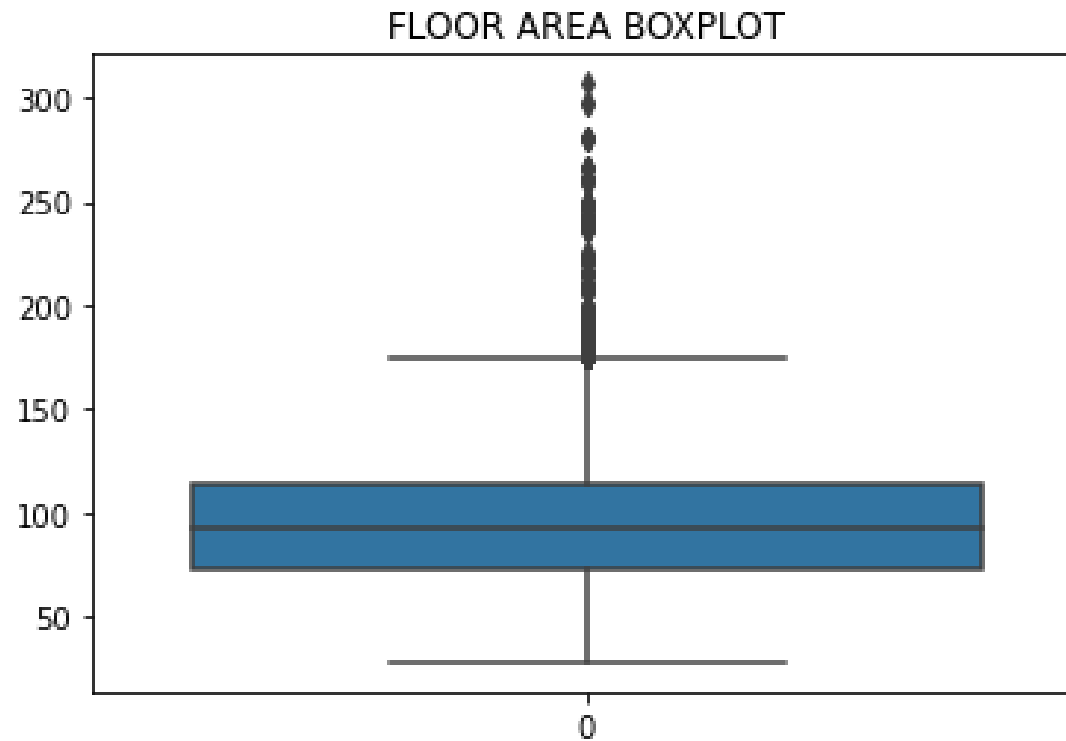
| | |
|-------|--------------|
| count | 810556 |
| mean | 291045 |
| std | 147336 |
| min | 5000 |
| 25% | 180000 |
| 50% | 272000 |
| 75% | 380000 |
| max | 1205000 |
| Name: | resale_price |



The housing price has a skew distribution. 50% of the houses have been sold between 180,000 and 380,000.

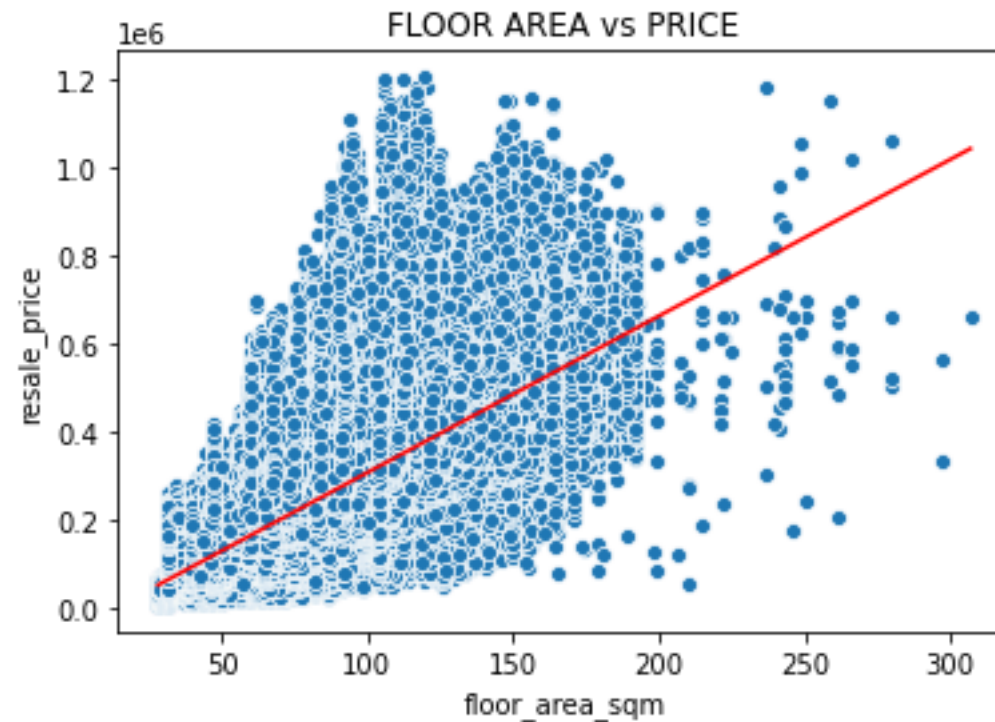
Floor_area_sqm

| | |
|-------|-----------------|
| count | 810556 |
| mean | 96 |
| std | 26 |
| min | 28 |
| 25% | 73 |
| 50% | 93 |
| 75% | 114 |
| max | 307 |
| Name: | floor_area_sqm. |



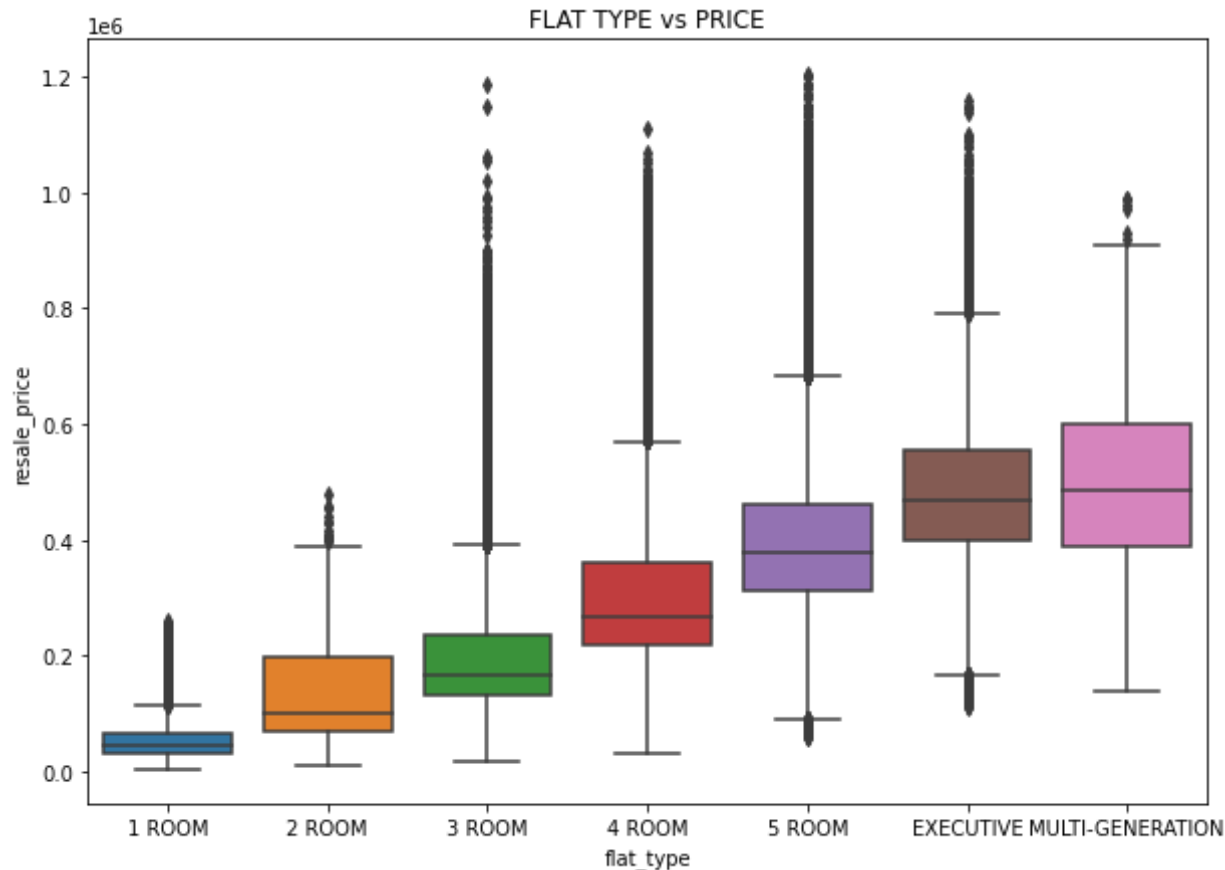
The floor area of the apartments is between 75 to 115 sqm, with an average of 95 sqm per apartment

Floor area vs Price



There is a positive correlation between the size of the apartment and its price.
Price increases as the floor area gets bigger.

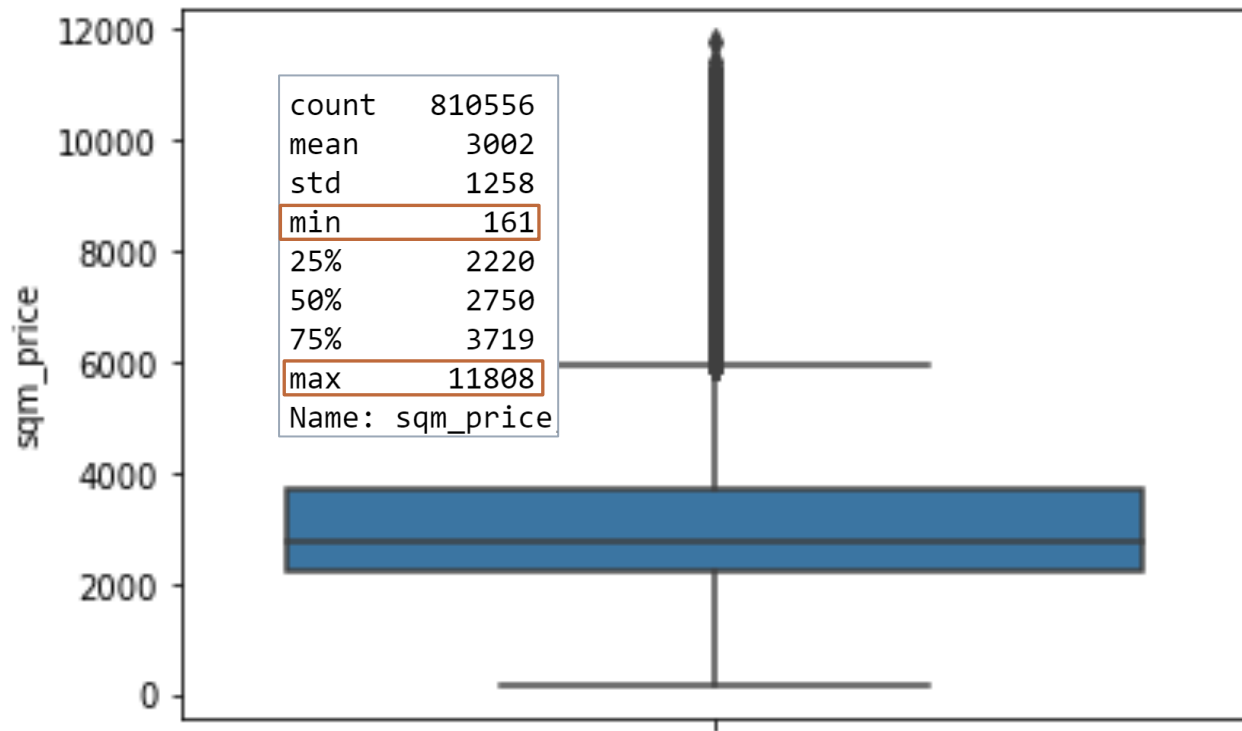
Flat_type vs Price



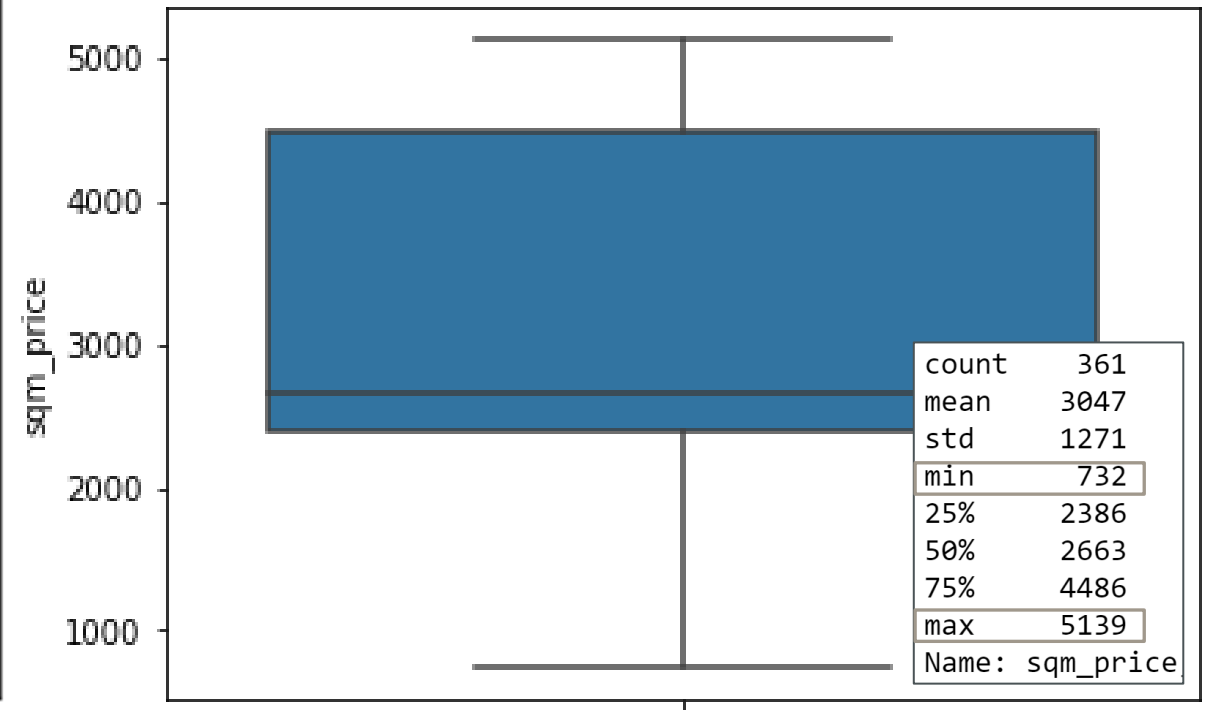
- There is a positive correlation between the number of rooms (size of the apartment) and its price.
- Executive and Multi-Generation are both premium flat types of respectively 3 and 4 rooms.

Sqm_price

SQM PRICE in Singapore since 1990

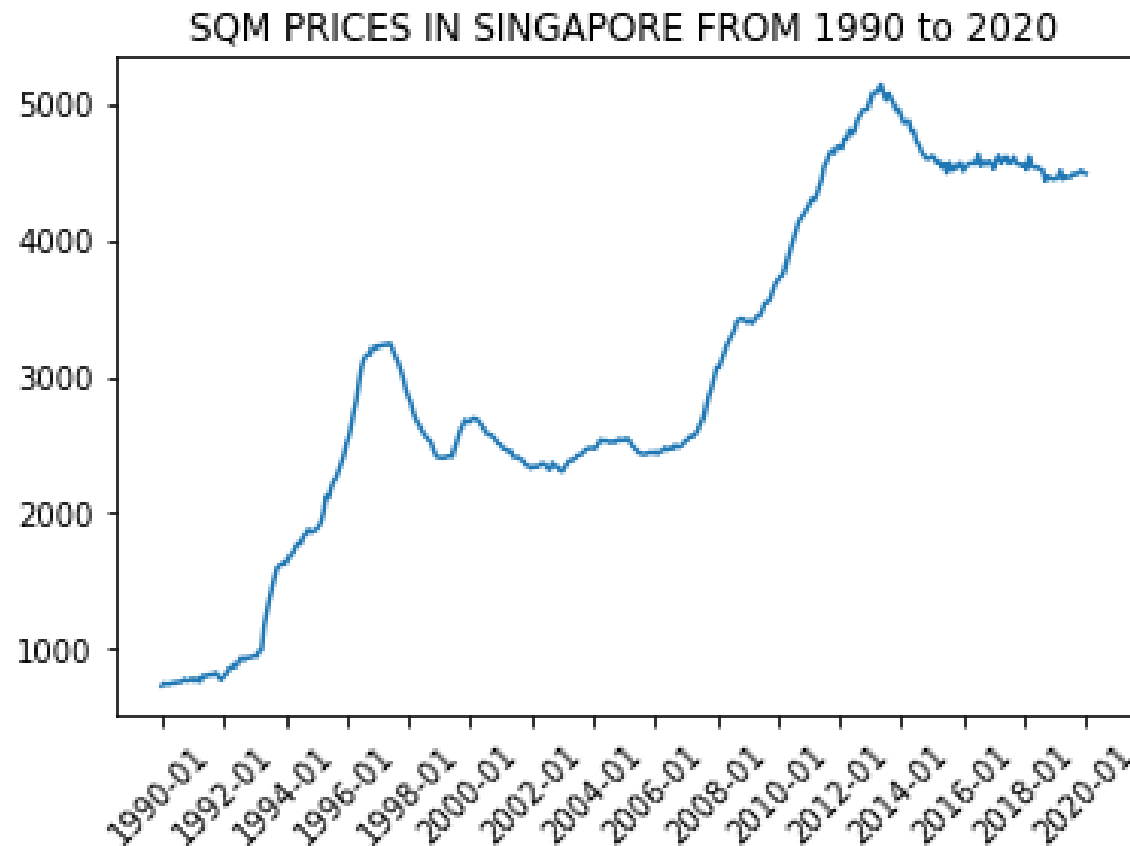


AVERAGE SQM PRICE per MONTH since 1990

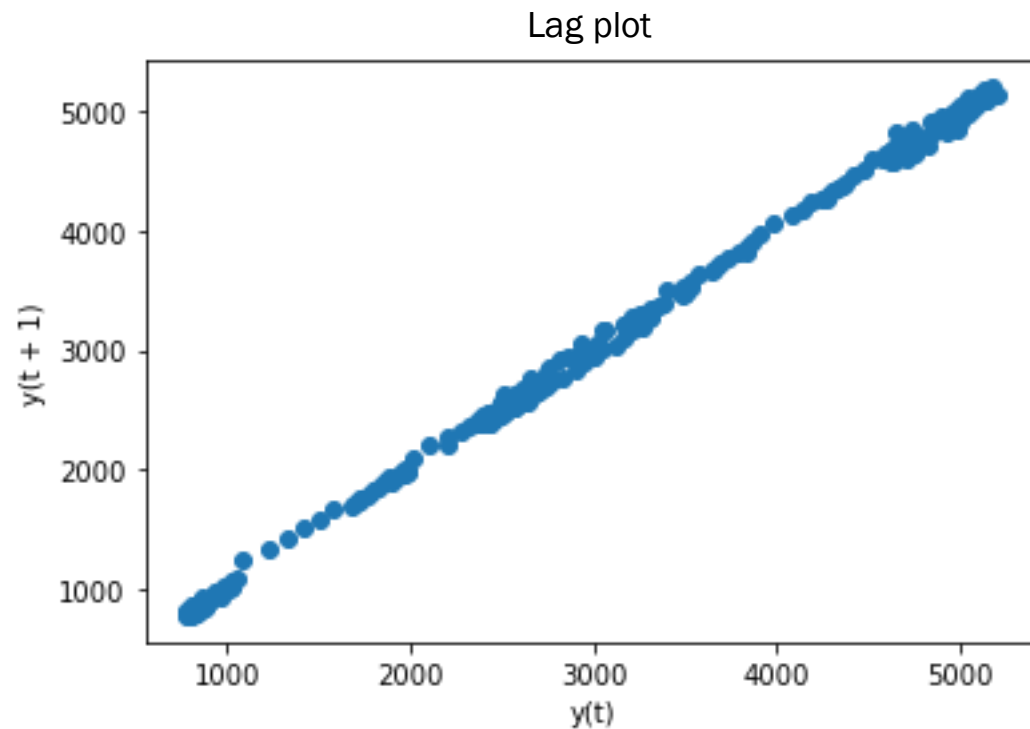


2. Time Series Analysis

Evolution of the housing price

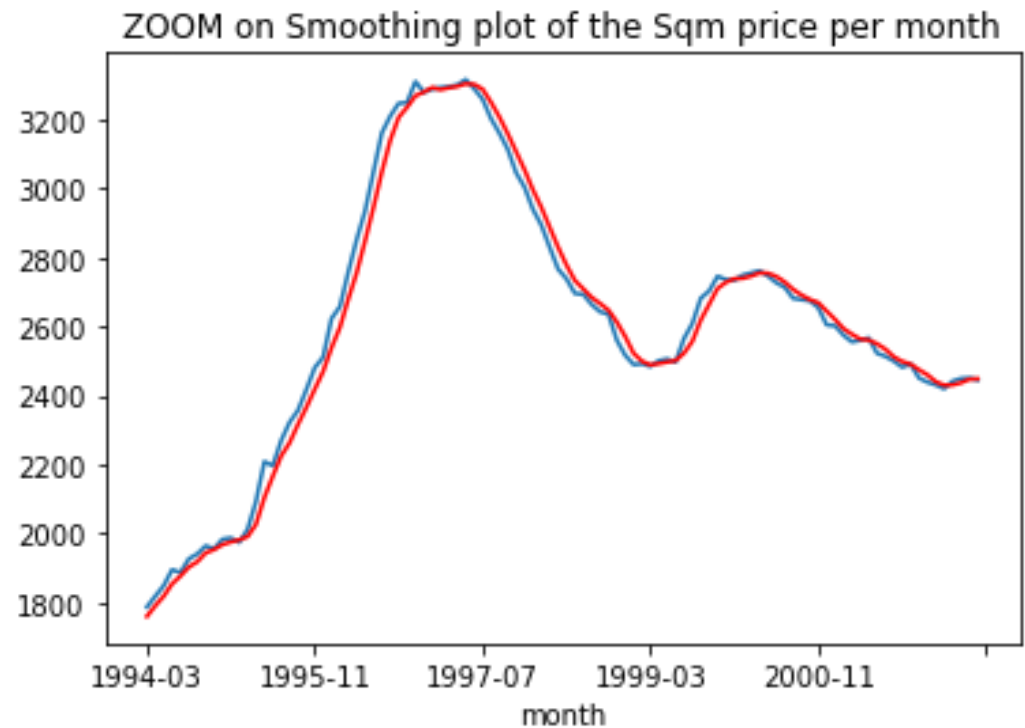
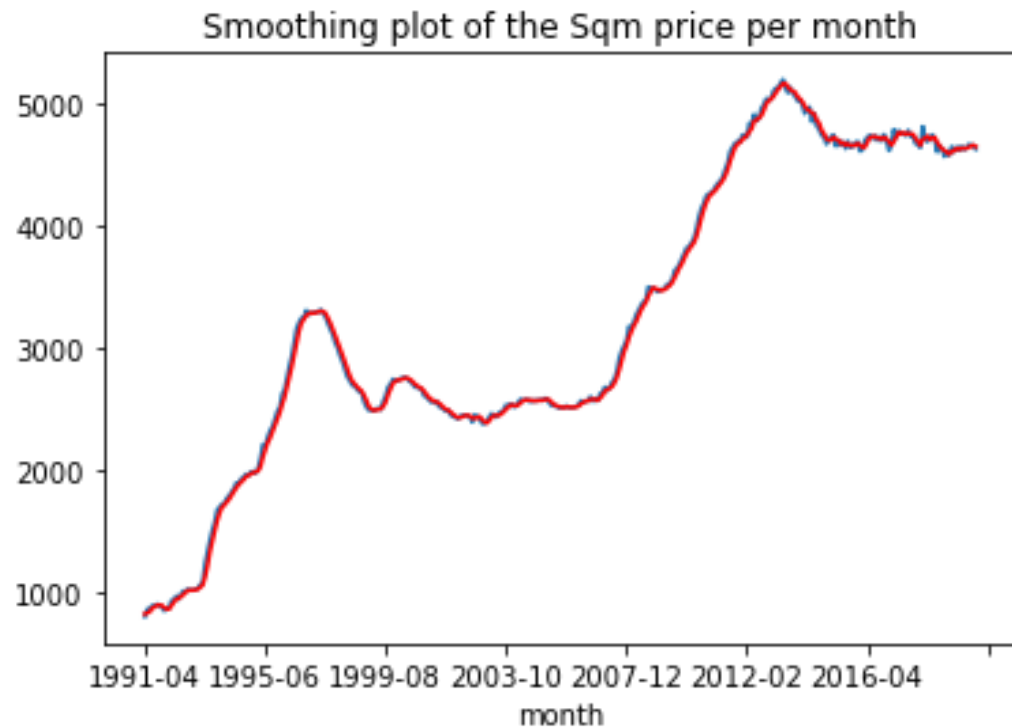


Lag plot



- Time series modeling assumes a relationship between an observation and the previous observation. Previous observations in a time series are called lags.
- We can see a strong positive correlation between the price of the sqm at month t and at month $t+1$

Moving Average Smoothing

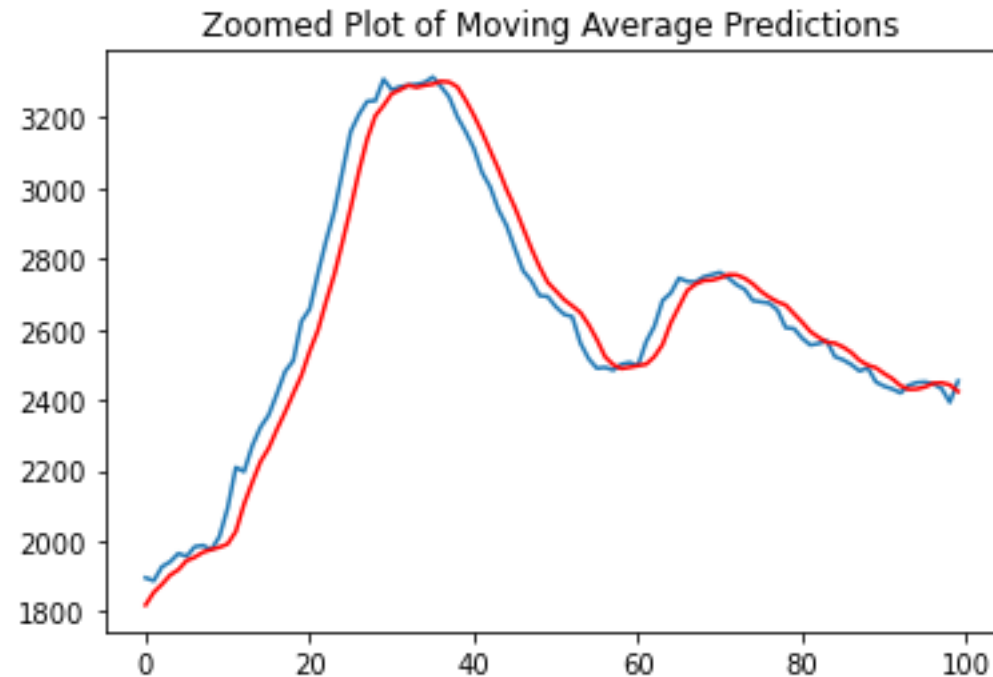


Smoothing is a technique applied to time series to remove the fine-grained variation between time steps.

Moving Average as Prediction

Predict the price for the next period (e.g. month)

- Compute the mean on the historical data and predict the value for the next period (next month)
- As new observations are made available (e.g. monthly), the model can be updated, and a prediction made for the next month



RMSE: 66.85

Abs error: 49.68

Stationarity in Time Series Data

- The observations in a stationary time series are not dependent on time.
- When a time series is stationary, it can be easier to model.
- Using Augmented Dickey-Fuller test, we accept the null hypothesis that says our real-estate data are not stationary. **THEY DEPEND ON TIME.**

ADF Statistic: -1.791865

p-value: 0.384508

Critical Values:

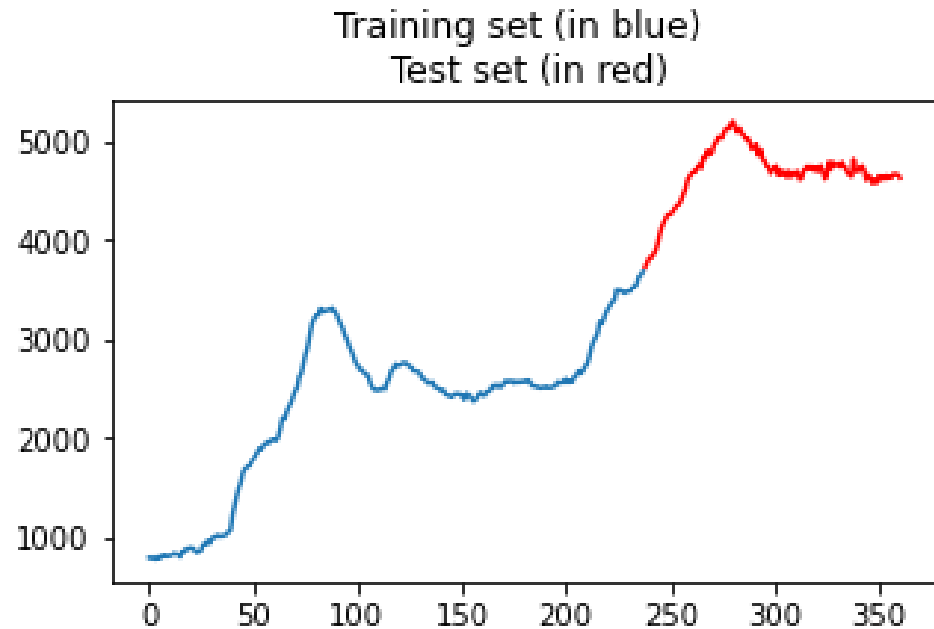
1%: -3.449

5%: -2.870

10%: -2.571

Train-Test Split

for Time Series Data



Observations: 361 (12 months x 30 years + Jan 2020 = 361)

Training Observations: 238

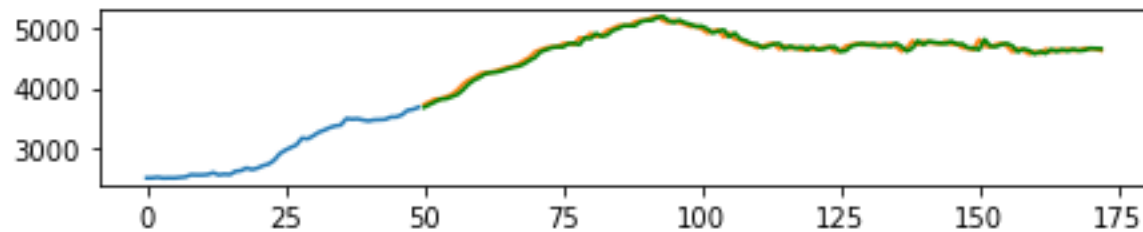
Testing Observations: 123

- When training a model we want to make sure that we don't use any information from the test set.
- In **blue** we can see the training set and in **red** the test set. Training set use 'historical' data, past data of the train set.

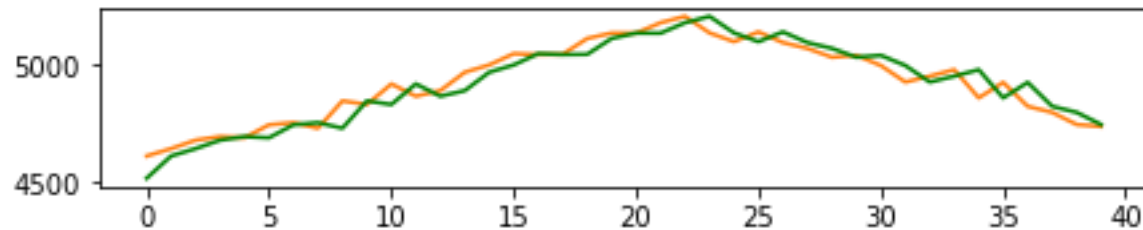
Persistence model for Forecasting

Predict the sqm price of the next month by taking the price of the previous month

Predictions of next month SQM price



Predictions vs Actual Values



Test RMSE: 51.44

Abs error: 42.06

| | t | t+1 |
|---|--------|--------|
| 0 | nan | 786.67 |
| 1 | 786.67 | 795.15 |
| 2 | 795.15 | 777.38 |
| 3 | 777.38 | 773.25 |
| 4 | 773.25 | 801.76 |

Regression Modeling for Time Series

Linear Regression using Lags as features

RMSE: 47

Abs error: 35.53

53

| | | | |
|---------|---------|---------|---------|
| 1990-04 | 1990-03 | 1990-02 | 1990-01 |
| 1990-05 | 1990-04 | 1990-03 | 1990-02 |
| 1990-06 | 1990-05 | 1990-04 | 1990-03 |

| | lag_1 | lag_2 | lag_3 | lag_4 | Y_pred | |
|---------|-------|--------|--------|--------|--------|--------|
| 1990-05 | 0 | 743.16 | 743.09 | 750.40 | 731.92 | 754.29 |
| 1990-06 | 1 | 754.29 | 743.16 | 743.09 | 750.40 | 748.00 |
| 1990-07 | 2 | 748.00 | 754.29 | 743.16 | 743.09 | ... |

We build a linear regression model to predict the price of the next month by using lags as a feature.

Time Series Results

- Using only the historical data, we can estimate the average price of the square meter in Singapore, for the next month, with an absolute error of **35 dollars**.

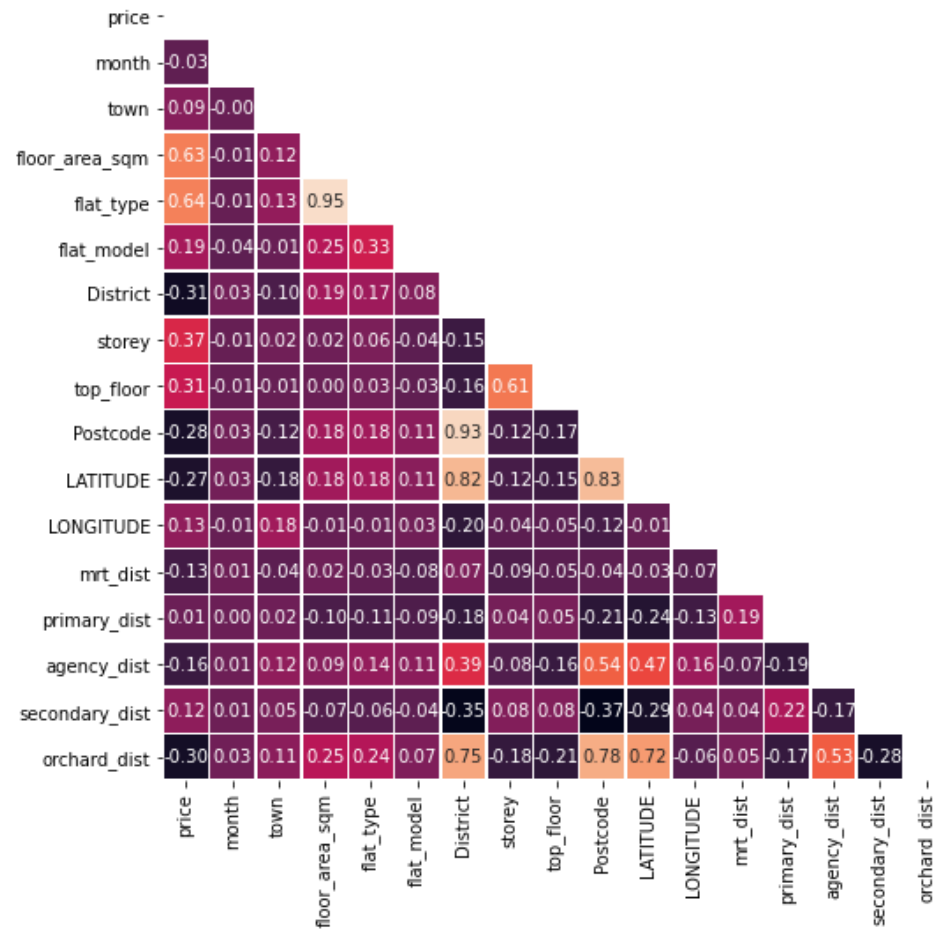
| MODELS | MOVING AVERAGE | PERSISTANCE MODEL | REGRESSION MODELING |
|--------|----------------|-------------------|---------------------|
| MAE | 49.68 | 42.06 | 35.53 |

- Now, let's try to understand what are the factors that drives the sqm price.

3. *Drivers Extraction*

What drives the prices of the houses ?

Correlation Matrix



- The first column of the Matrix shows the features correlation with our variable *Price*

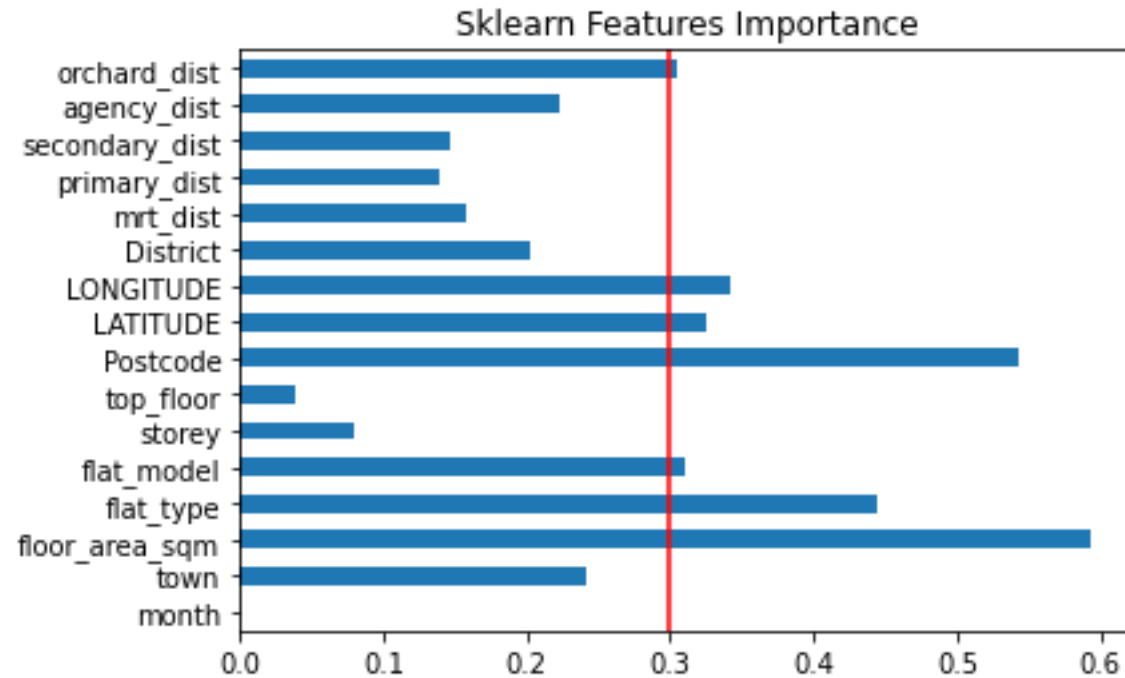
- floor_area_sqm
- flat_type
- District
- storey
- top_floor
- Orchard_dist

- We can spot some redundant features:

flat_type highly depends on *floor_area_sqm*
top_floor highly correlated to *storey*
postcode highly correlated to *latitude*

Features importance

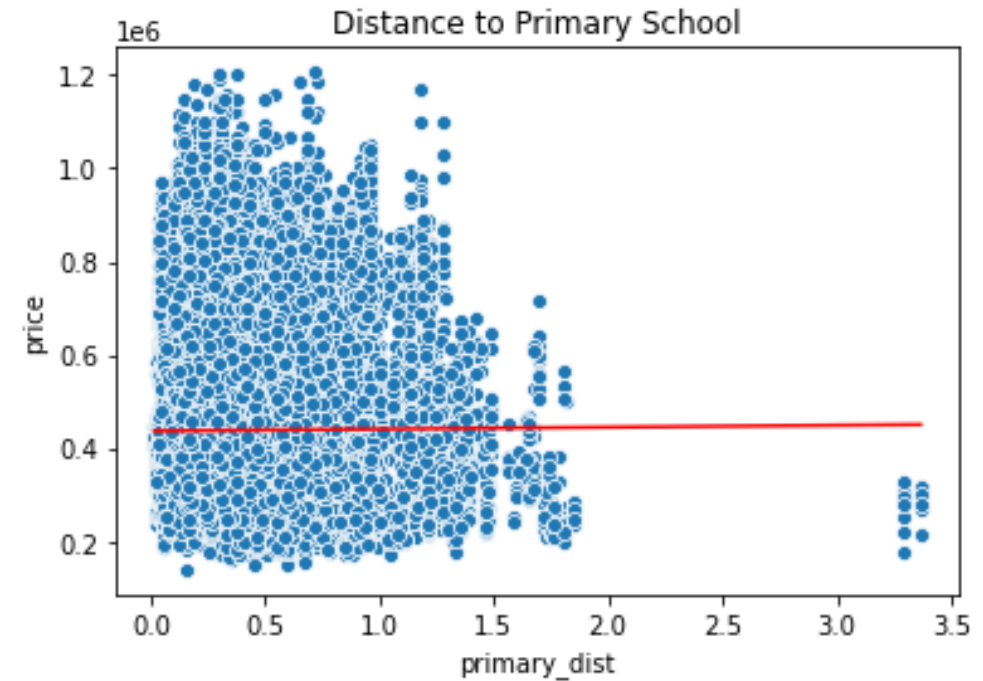
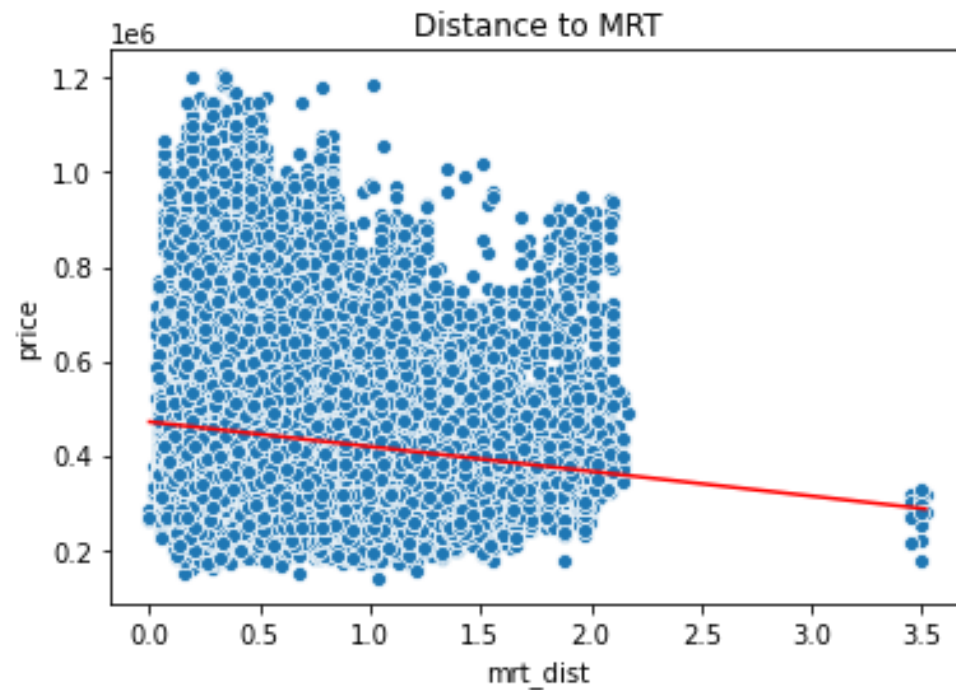
using `sklearn.feature_selection`



This method estimates the mutual information (entropy) for our target variable *Price*

NOTE

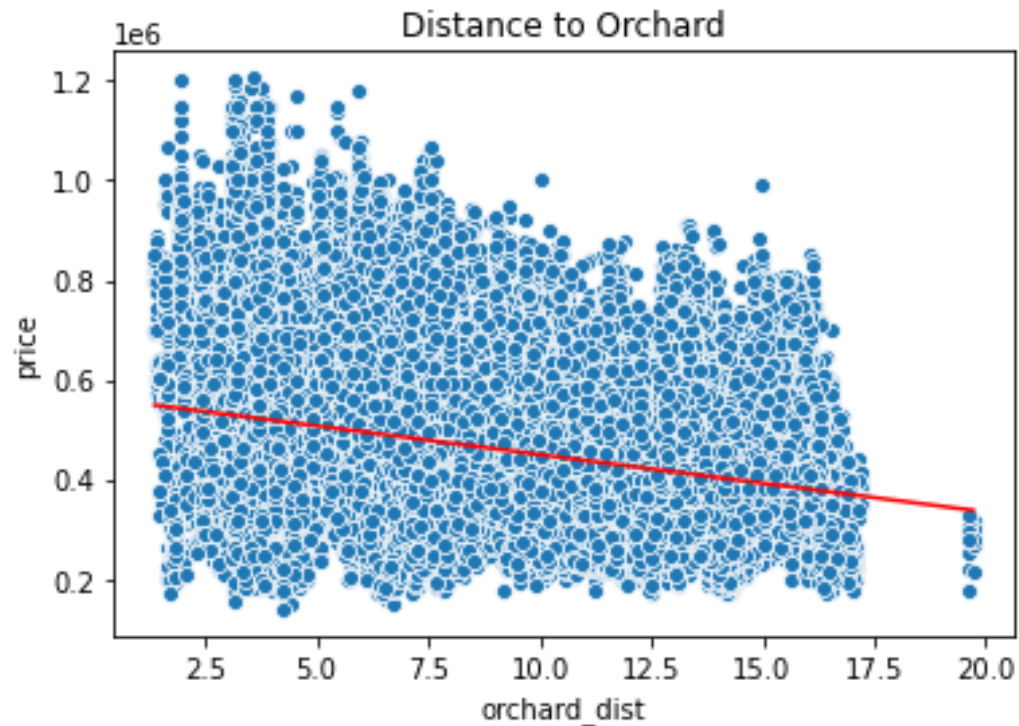
Understand the low influence of distance to POI



Most of the HDB have direct access to MRT, and schools.
That is why, we can't estimate the influence of the distance to the POIs on *Price*

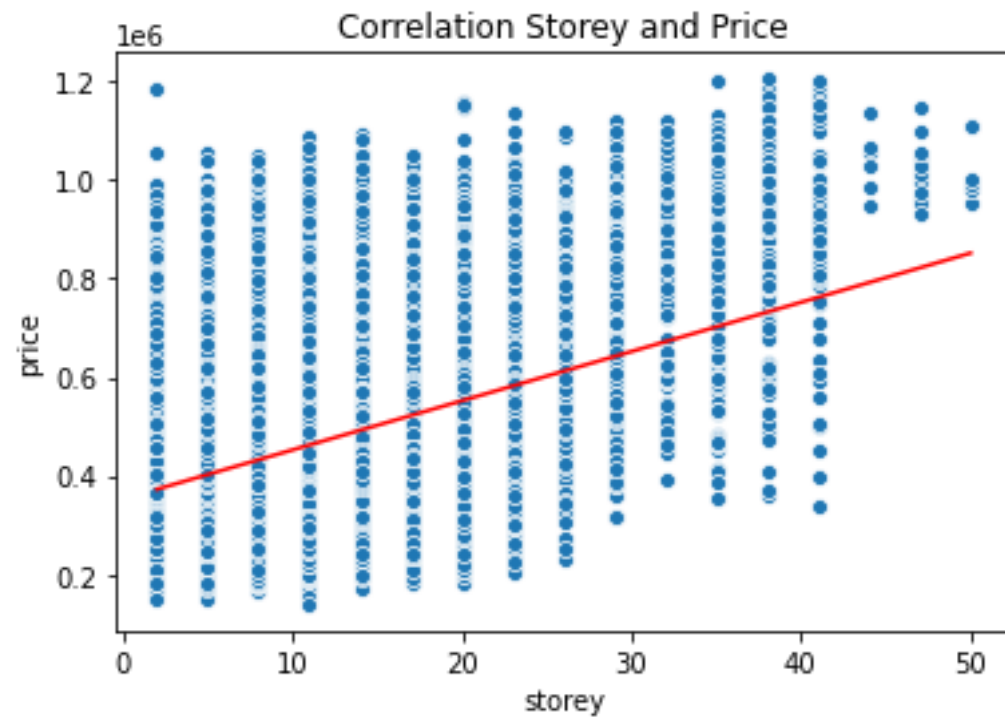
NOTE

Understand the low influence of distance to POI



In contrast, the orchard_dist feature have a wider range and can provide more insight on the impact of the distance to Orchard on the *Price*

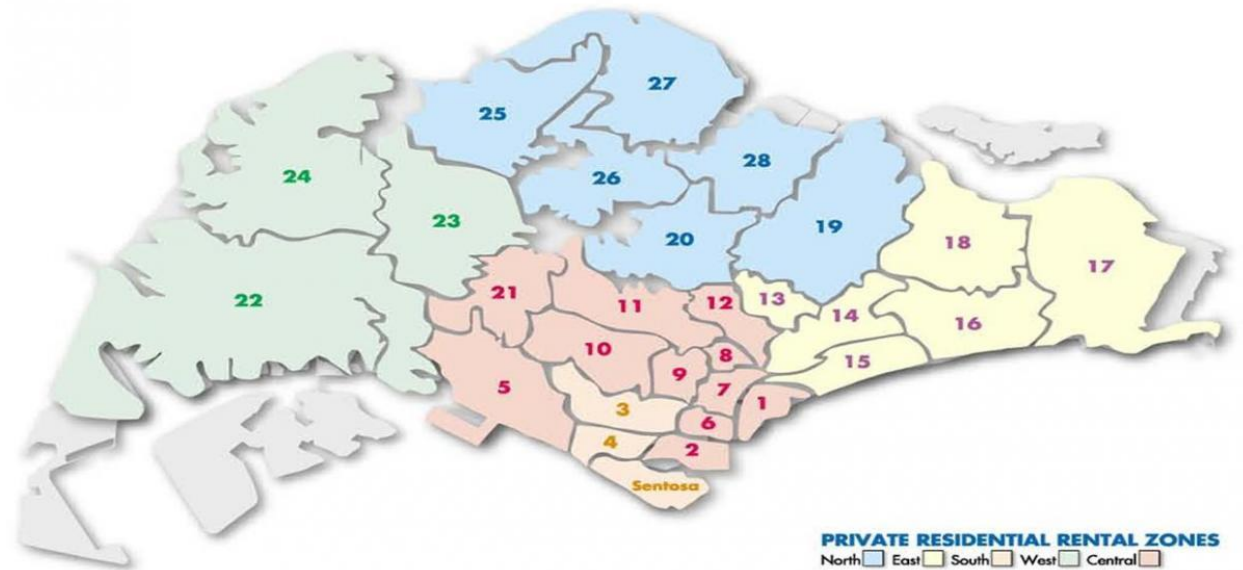
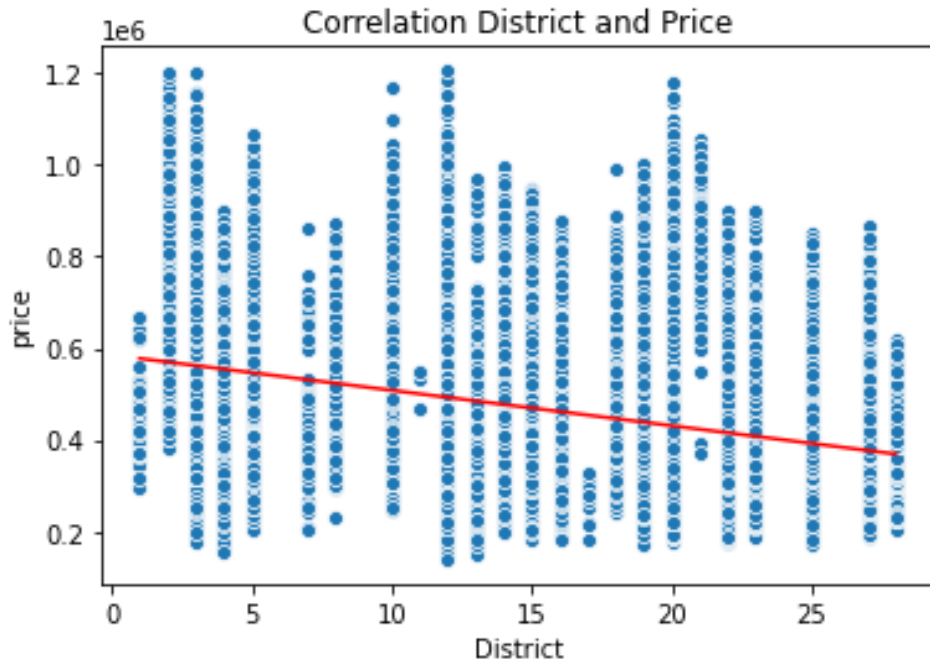
Correlation between Storey and Price



Being located on a higher floor, seems to give value to the apartment

NOTE

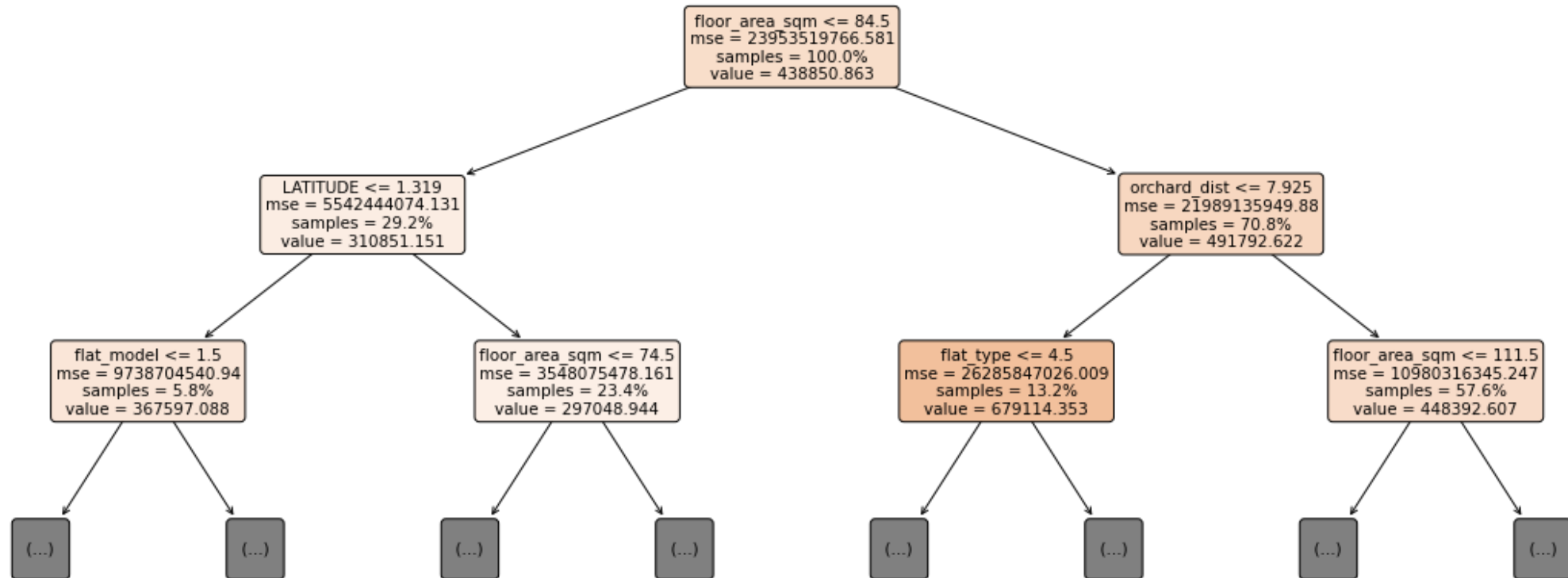
Understand the low influence of distance to POI



There is a negative correlation between the district number and the price. Higher the # of the district, lower the price is.

Features importance

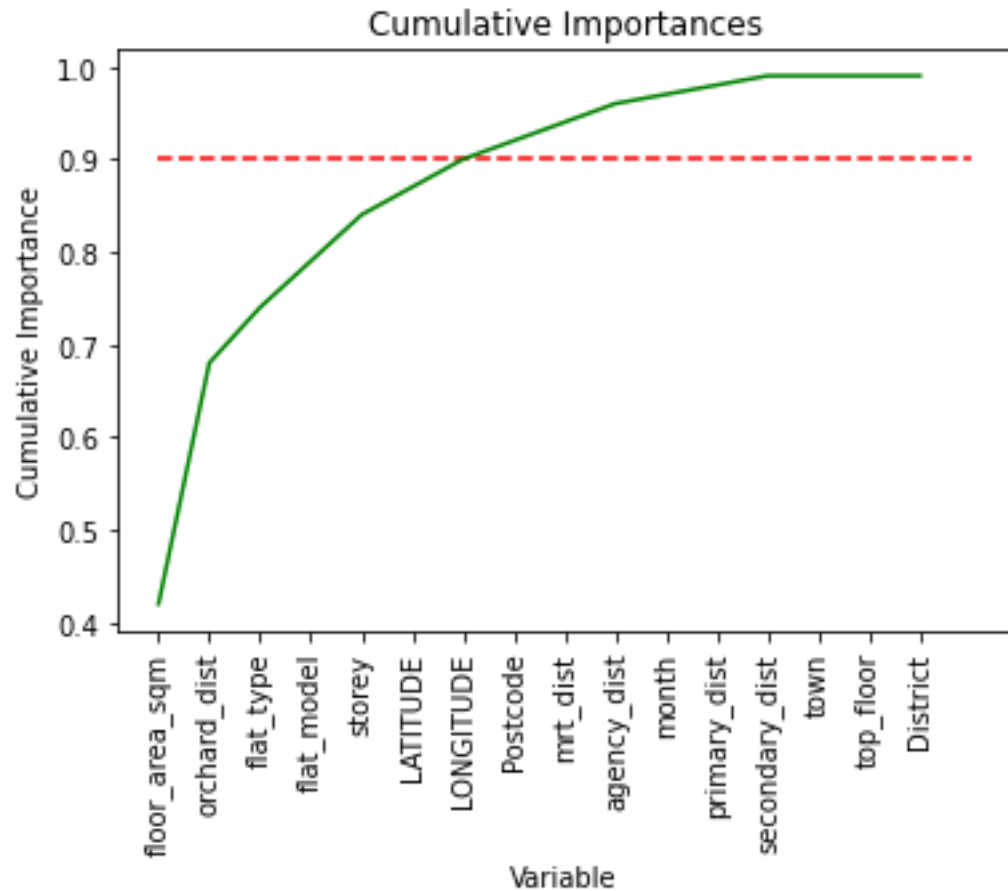
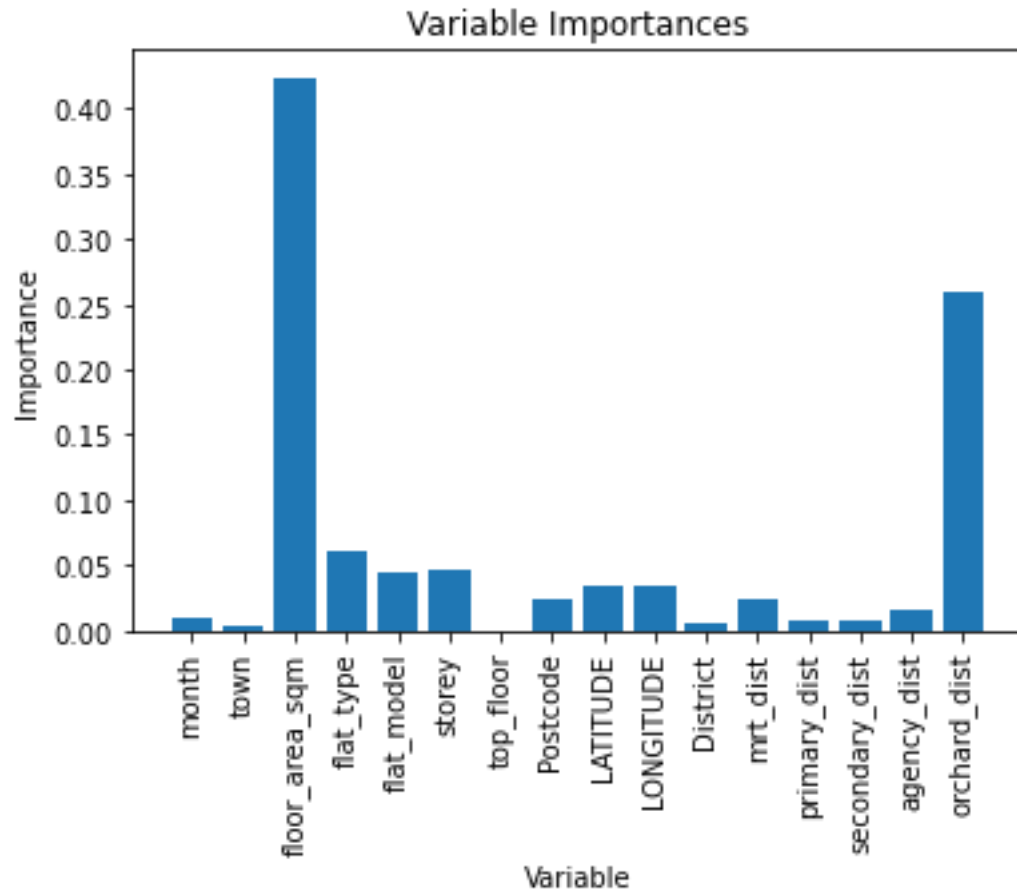
using Random Forest



Visualization of a tree from the Random Forest

Features importance

using Random Forest



Drivers Extraction Results

| METHODS | CORRELATION MATRIX | SKLEARN FEATURES IMPORTANCE | RANDOM FOREST FEATURES IMPORTANCE |
|-----------------------|--|---|--|
| IMPORTANT FEATURES | floor_area_sqm Orchard_dist storey District | floor_area_sqm Orchard_dist flat_model Latitude Longitude | floor_are_sqm Orchard_dist Storey flat_model Latitude Longitude |

4. *Housing Price Prediction*

Linear Regression

| FEATURES FROM | CORRELATION MATRIX | SKLEARN FEATURES IMPORTANCE | RANDOM FOREST FEATURES IMPORTANCE | SIMPLEST |
|---------------|--------------------|--------------------------------|--------------------------------------|----------|
| R-SCORE | 0.72 | 0.64 | 0.73 | 0.70 |
| MAE | 63,180 | 69,988 | 62,314 | 65,266 |

Linear Regression Results

$$y_{\text{pred}} = 101078 + 4685 \text{ floor_area_sqm} - 16852 \text{ orchard_dist} + 7547 * \text{storey}$$

| | |
|----------------|---------|
| Intercept | 101078 |
| floor_area_sqm | 4685 |
| storey | 7546 |
| orchard_dist | -16852 |
| dtype: | float64 |

We can interpret this equation and say:

The price of an apartment is composed of:

1. A base of \$101,078
2. Plus \$4,685 for each square meter
3. Minus \$16,852 for each km from Orchard MRT
4. Plus \$7,546 for each storey

Doing so, we will get the price of the apartment with an error of \$65,266.

Random Forest

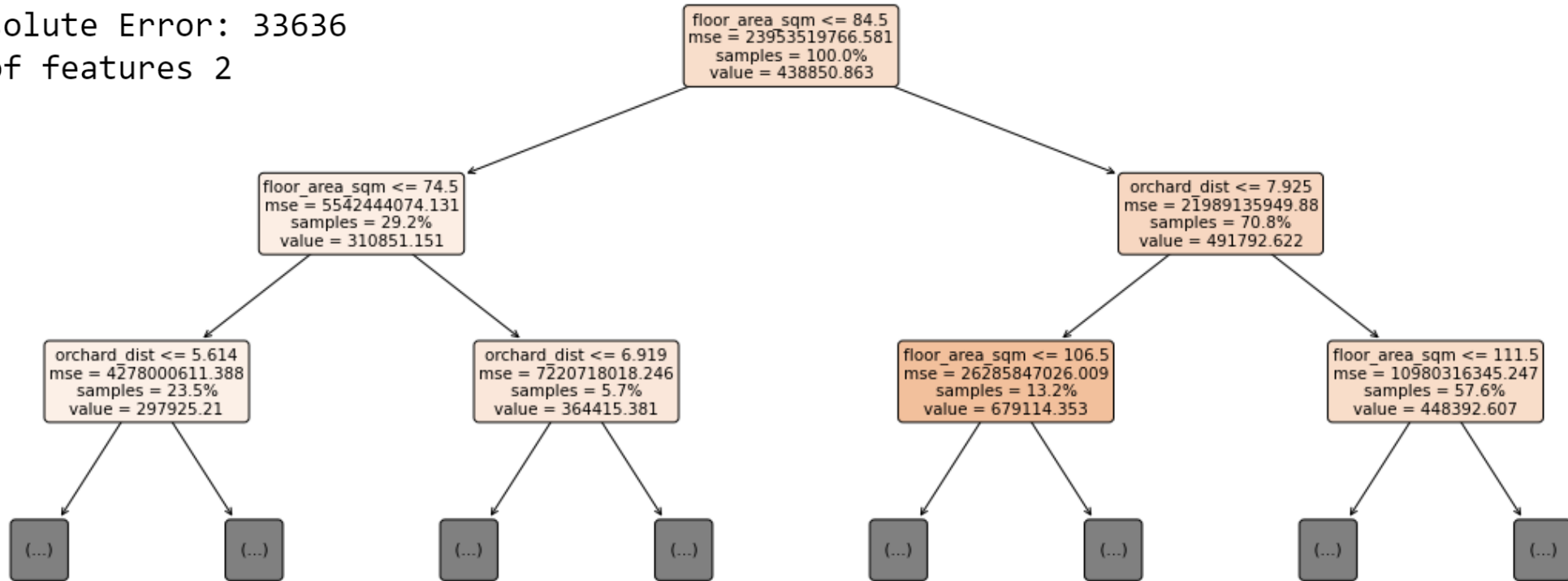
| FEATURES FROM | ALL FEATURES | RANDOM FOREST FEATURES | 3 MAIN FEATURES <i>(floor_area, orchard_dist, storey)</i> | SIMPLEST |
|---------------|--------------|------------------------|--|----------|
| # OF FEATURES | 16 | 6 | 3 | 2 |
| MAE | 20,820 | 23,776 | 34,620 | 33,636 |

Random Forest

Tree Visualization (*floor_area_sqm*, *orchard_dist*)

Mean Absolute Error: 33636

Number of features 2



CONCLUSION

The goal of this project was to understand the key drivers of the prices of the houses in Singapore.

By combining, time series analysis, feature selection techniques and Machine Learning, we estimate the key drivers of the housing price in Singapore to be, in order:

1. The size of the apartment
2. Its distance to the city center (how close to Orchard)
3. Its floor number (higher floor the better)
4. Its location (district, postcode, exact location)

Using these features we estimated the price of the apartments with an error of SGD 30,000.

IMPROVEMENTS

- By using the publicly available data of housing in Singapore, we collect data on HDB. The study could be different if mixing public and private housing would have been possible. Although, note that more than 80% of Singapore population live in HDB.
- A key ratio that I had difficulties to embedded in the study is the Supply and Demand ratio for time analysis. As the supply is possibly computable, I still haven't found a way to estimate the demand.
- As a main result of the study, one of the key element to determine the price of the house is its location. So what make a district attractive ? What are the drivers of the attractiveness of a district ? All these questions could lead to another project...