NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

AY2022/23 SPECIAL SEMESTER

CC0002 Navigating the Digital World

Title: Resale HDB Flat Price Prediction Model

Tutor: Dr Josephine Chong

Tutorial Class: PT1

Group: 4

Date of Submission: 09/06/2023

Group members	**Matriculation Number**
Amizzuddin Bin Md Amin	U2222177J
Ng Gim Long	U2220723J
Chen Xin	U2220834F
Muhammad Nurhidayat Bin Suderman	U2220413J
Lux Pang Jian Wen	U2220935E

Contents

Executive Summary

Singapore HDB resale market plays a significant role in the overall real estate sector of the country. It provides first-time home buyers with an alternative method of securing housing in Singapore, especially those urgently needing a new home. It also offers opportunities to homeowners to build up their wealth as the value of HDB flats grows with time. There are various channels where we can get transaction data, for example, brochures from property agents. The Singapore government has released such data yearly, providing insights to potential homebuyers.

However, the market is complicated and can be affected by many factors. Many potential buyers may need help understanding how location, floor, amenities, remaining leasing years and other factors affect house pricing. It is because historical transaction data does not carry enough contextual information, nor does it tell the trend of housing price movement.

Our quantitative research aims to generate insights and a prediction model based on reliable past-year transaction data. Linear regression is adopted to analyze and identify factors affecting the price. The weightage of each factor will be indicated in the output model. This model aims to help potential homebuyers to make more informed decisions by providing a clearer price trend analysis over the years, to predict housing price movement or even to identify undervalued flats in certain areas.

It is worth mentioning that there are certain limitations to this model. This model does not consider market factors such as government policies, cooling measures, or economic growth. Other factors, such as financial crisis or global disease spread like "Covid-19", are also not considered. It is because these events are unpredictable. The dataset we use here might be incomplete. All these factors cause inaccuracy in price analysis or prediction.

Introduction

Singapore's Housing and Development Board (HDB) flats were first introduced on 1 February 1960 to address Singapore's housing crisis and allow Singaporeans to buy affordable houses to start their family. The aim was to ensure that every Singaporean would have the option to purchase a house of their own and start a family. Over time, Singapore had one of the world's highest homeownership rates, which clearly shows how successful HDB's flat system is.

Initially, the government meant for HDB flats to be a housing means for all Singaporeans and not as a form of investment. Nevertheless, the fact that flats can be resold at a higher price was noticed by homeowners. As a result, it eventually evolved into a form of investment for Singaporeans to get their future dream home. An example would be the BTO project named "Hougang Meadow". According to Teoalida, this BTO was launched on 26 November 2013. During the launch, the 4-room flat's indicative price range was only between $306K and $362K. However, according to the SRX, units between 7 and 15 floors were transacted between $700k and $747K in March 2023. It represents an increase of between 93% and 107% in prices when compared to the max price of $362K at launch. In reality, the percentage increase is likely to be far higher when the actual price of each unit is used for the comparison. It makes it an enormous temptation for first-time homeowners to purchase HDB Built to Order (BTO) flats as it represents a high-profit margin for them and gives them the best chance to upgrade to their future dream home.

Coupled with the low price and high-profit margins, the demand for BTO flats has increased since its introduction in 2001. According to The Straits Times, in 2021, the demand for BTOs increased by over 70% between 2020 and 2021. With a steep increase in demand and a gradual increase in supply, there are bound to be more people that are unable to get a flat. As such, these people would have to turn to resale flats which causes the demand to rise. While there is an increase in demand for resale flats, not all homeowners wish to sell their units. Therefore, an insufficient supply of flats exists versus the considerable demand for flats.

Our project aims to build a linear regression model with a time series analysis of factors to help potential buyers in the resale market to buy a flat that suits their budget better for their interim use and minimize the losses over five years of Minimum Occupation Period (MOP) to allow them to upgrade to their future dream home.

Some limitations that we foresee for this model would include the fact that it does not take into account inflation over the years, market cooling measures that the government uses to artificially reduce the prices of resale prices and any other economic or natural factors such as recessions or the spread of contagious diseases that affects the market which in turn prices of flats. These factors are hard to predict over time as they cannot be foreseen. An example would be the outbreak of Covid-19 which caused a decrease in resale flat volume, which, in turn, caused the resale prices in 2020 to spike by 5%, the most significant yearly increase since 2012, according to HDB as reported by Today. New buyers may also not have extensive knowledge about the historical prices of the neighbourhood that they intend to buy the unit from, and this could result in the buyer purchasing the unit at an inflated price as the buyer might find that the flat is the best price at the time of viewing and research.

With all that being said, the project's main objective is to study the prices of HDB resale flats between 2017 and 2023. Study the relationship between factors such as flat size and location and determine which factors can heavily influence the long-term price of the unit. Coupled with this knowledge, a potential buyer can work out and make a well-informed decision on which flat to select based on the best potential long-term value to allow them to either minimize or earn a profit when they eventually decide that it is the right time to upgrade to their future dream home. A dream home for most Singaporeans comes in the form of a landed property or condominium.

Quantitative Reasoning Techniques

"What would be the price I need to pay if I am looking at a 4-room with 90 leasing tenure in Hougang?" "What is the growth rate for overall housing prices in Singapore in the last five years?" "Which is the biggest factor that drives HDB resale price?" "Which location is the best buy for a potential capital gain in 10 years?" Potential homebuyers usually will look at past transaction data and use similar transaction units for their estimations. For example, if a transaction record shows that a 4-room flat with 70 years of leasing was sold for $600,000, homebuyers would expect a resale unit with the same flat type, and 90 years of leasing tenure costs an additional $100,000. However, the HDB resale market is complicated and affected by many factors. Past year transaction data does not tell the price trend of HDB resale units in the long run, nor does it reflect the latest updates as it takes time for relevant authorities to release the data.

To understand the resale market better and provide more insights to home buyers, we need to perform a more comprehensive analysis of past year data sets. Many approaches can be adapted to do property valuation or price trend analysis. We have decided to take a comparative approach that compares flat attributes such as floor area, room type, and leasing year and cross-comparison between different years. By comparison, we can have a more precise and better understanding of how various variables, numeric or contextual, drive the flat price.

Our team has decided to build a regression model with numerous variables, including flat type, floor area, location, and remaining lease. We aim to provide a more flexible and objective alternative to potential homebuyers or investors. A linear regression model allows us to quantitatively understand the relationship between those x-axis variables mentioned earlier and the y-axis variable resale price. Linear regression is a good choice for extensive data set analysis. On the other hand, a linear regression model outlines each factor's significance in determining resale price. The regression coefficients help us differentiate the level of importance of different independent variables so that we can better manage those factors. We can quickly isolate the factors and determine what shall be included in the model. That can be done by analyzing the P-values of each factor; factors with low P-values shall be kept in the model; otherwise, they shall be removed. By doing so, we can achieve a better and clearer understanding of how different factors impact the price. Moreover, linear regression generates relevant metrics such as mean standard error, R-squared or adjusted R-squared that help us access and ensure our model's consistency and performance.

Getting data from reliable sources is vital in building an accurate model. We decided to use data from government resources, from https://data.gov.sg/, as we believe these datasets include all targeted independent variables and are complete and accurate. Based on simple correlation analysis, we have decided on a list of critical variables: floor area, flat type, flat model, storey range, remaining leasing and town. These variables are believed to have strong relationships with the resale price.

We must highlight some of the limitations of our model and data sets. We assume there is always a linear relationship between the input variables we will propose in the later part and the output variable, which is the resale price. However, according to the law of diminishing returns, it is not really linear. Although we did our correlation analysis, we did not attempt to identify possible autocorrelation issues. As these data sets were collected over specific periods, autocorrelation can happen as the current resale price can be correlated to the previous period's resale price, for example, due to inflation. It is also worth noting possible multicollinearity between proposed independent factors. For example, floor area is likely to be strongly correlated to other factors, such as flat type, as a 4-room flat is bigger than a 3-room flat.

We applied the linear regression model below to study the relationship between our selected independent and dependent variables.

Data Visualization

Total HDB flats in the resale market

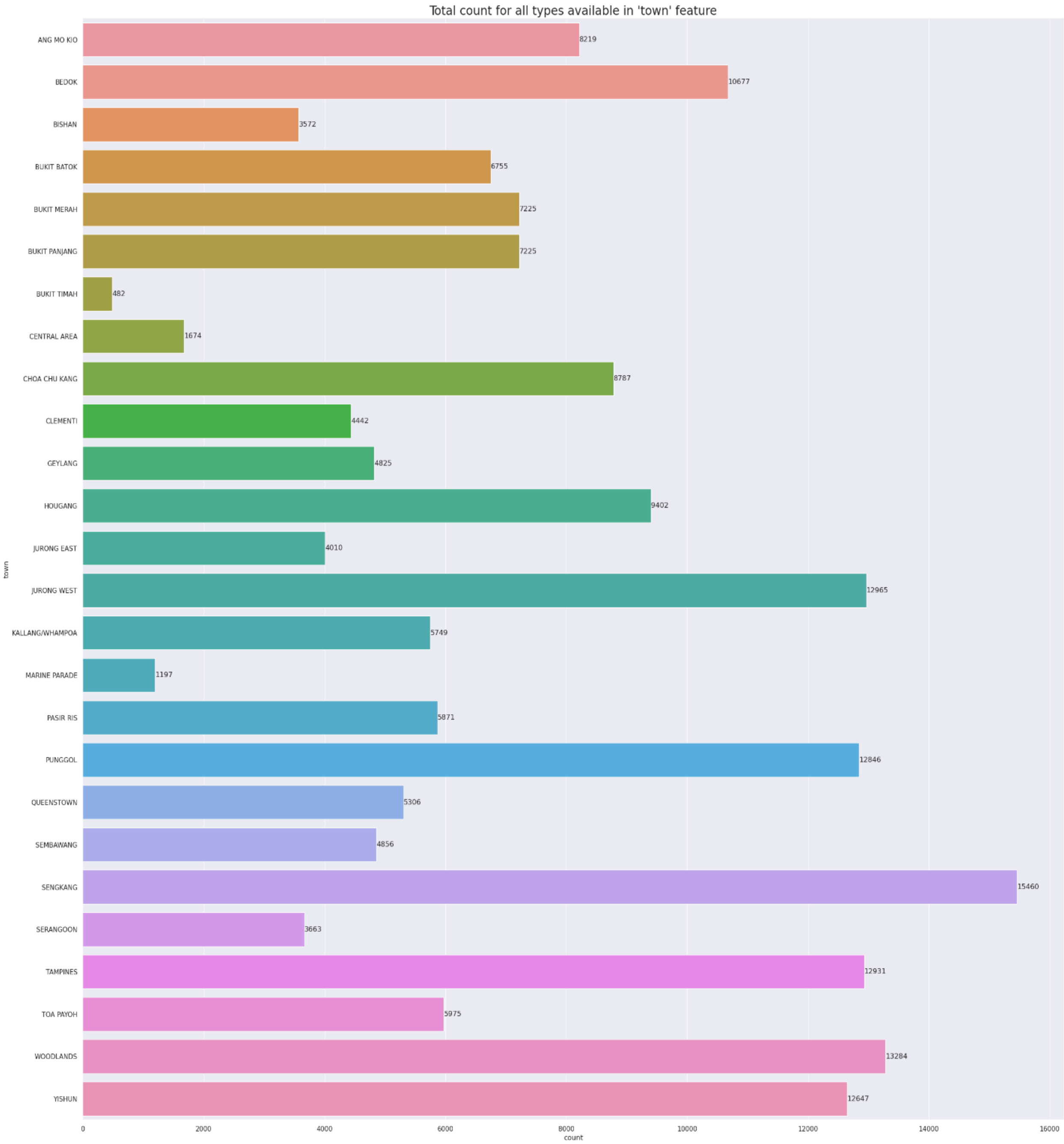

Figure 1 Total HDB flats in resale market from 2015 to 2023 group by town

Correlation matrix heatmap

Original features

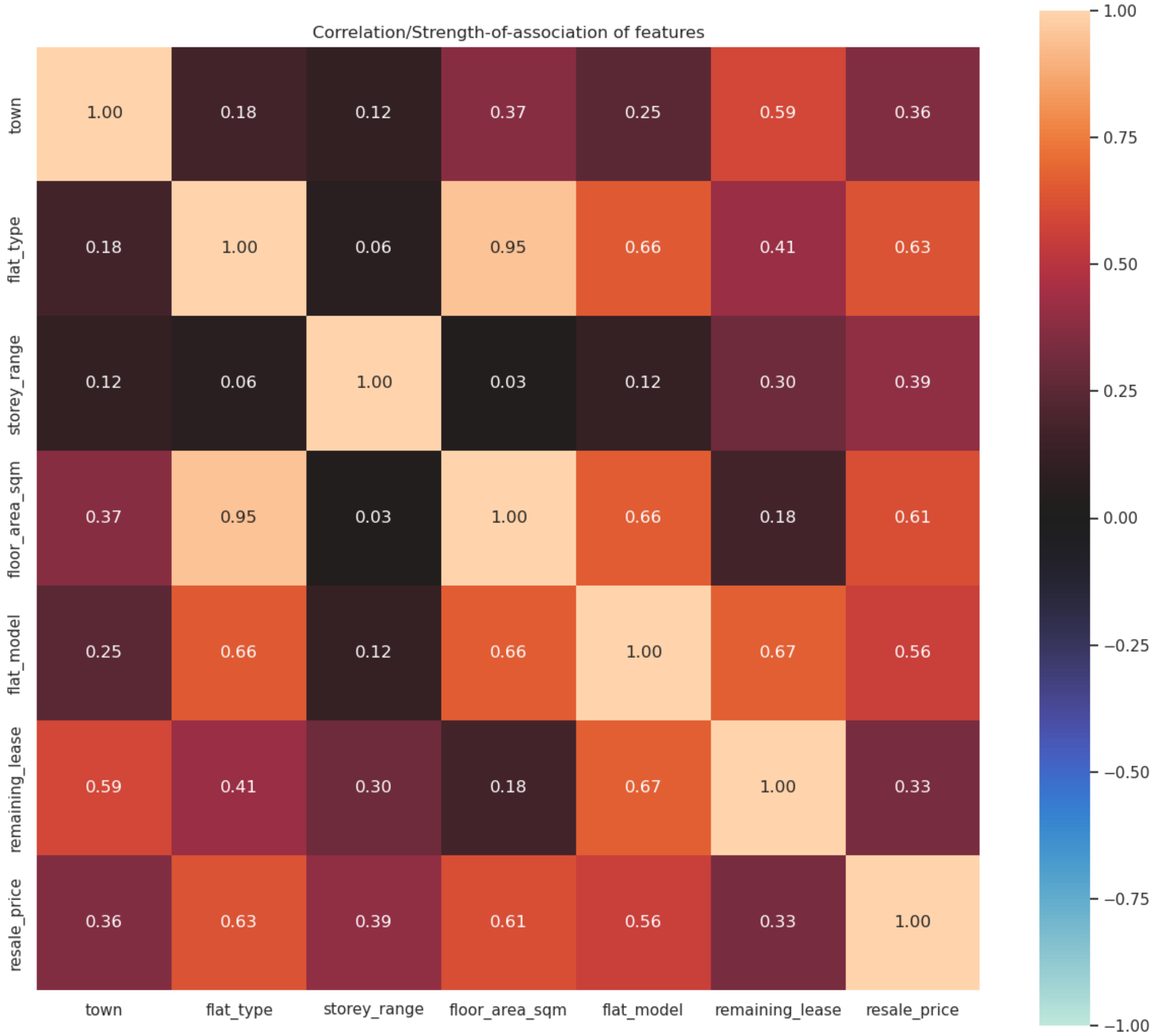

Figure 2 Correlation matrix based on the features available in the dataset

Addition of new feature

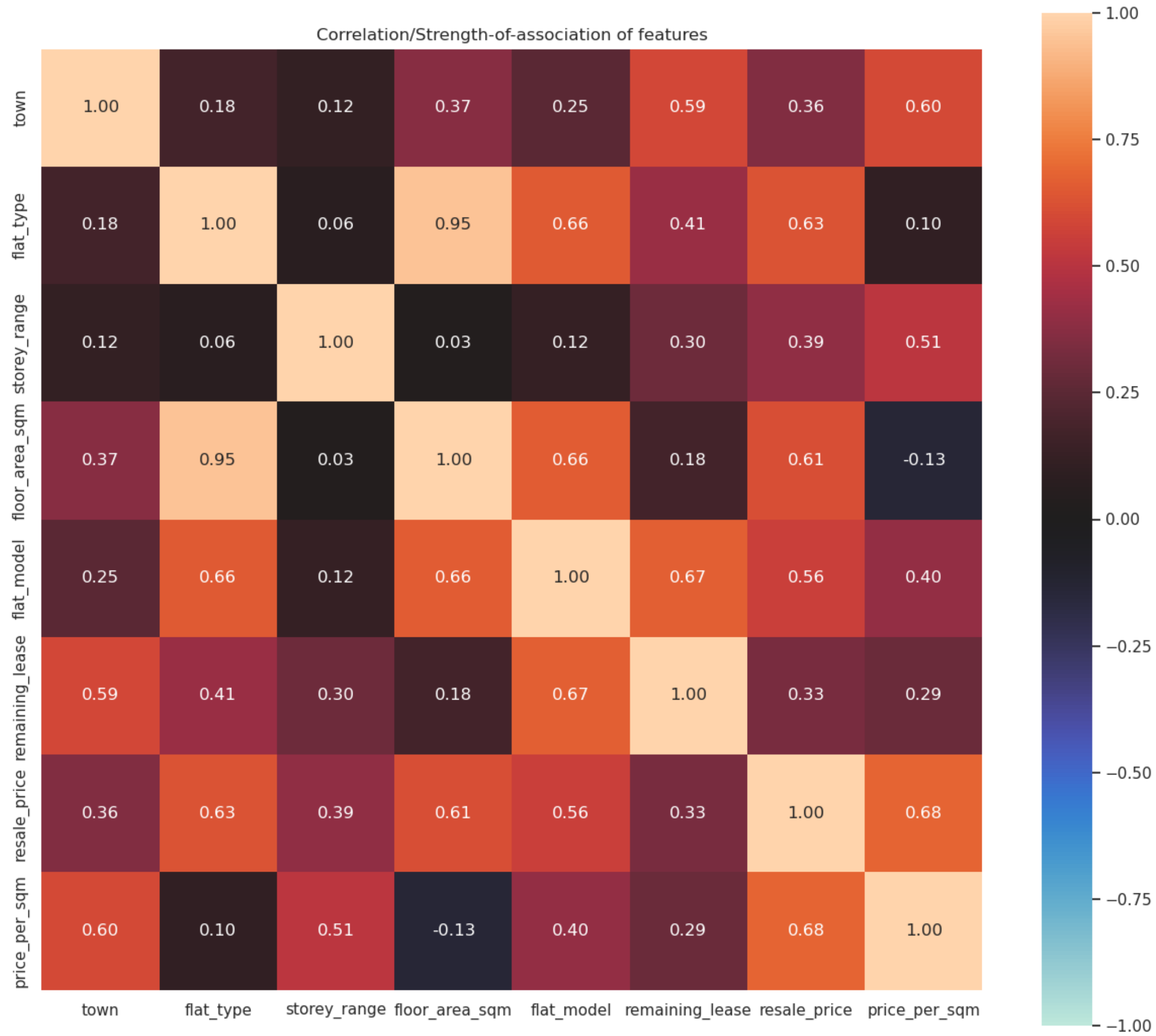

Figure 3 Correlation matrix with new feature 'price_per_sqm'

Resale price market trend

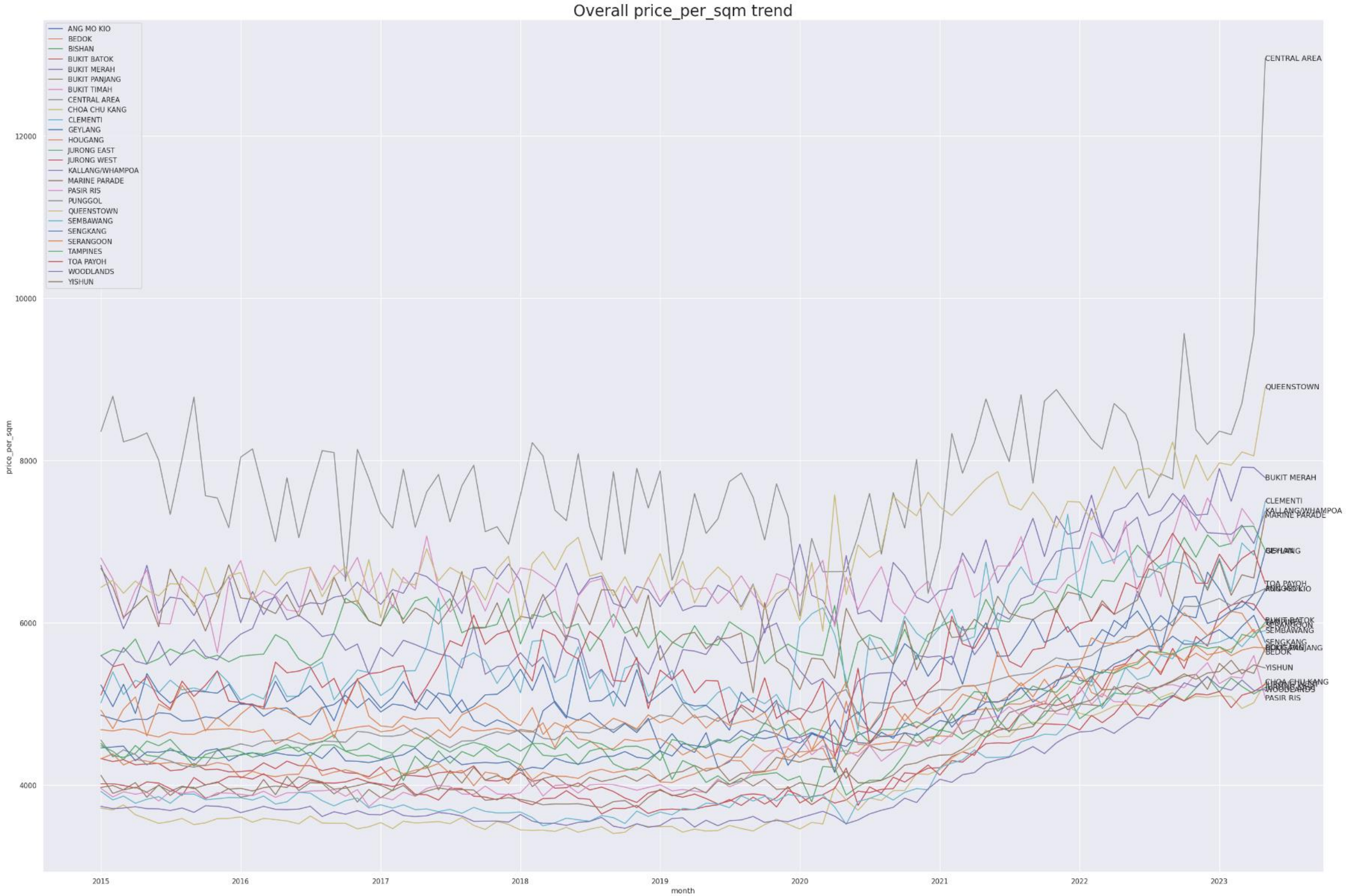

Figure 4 HDB price per square meter market trend from 2015 to 2023

Statistical model

```
Linear regression model for 'ANG MO KIO':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.296
Model:                            OLS   Adj. R-squared:                  0.289
Method:                 Least Squares   F-statistic:                     41.61
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           4.12e-09
Time:                        13:52:19   Log-Likelihood:                -759.22
No. Observations:                 101   AIC:                             1522.
Df Residuals:                      99   BIC:                             1528.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.346e+05   3.72e+04     -6.316      0.000   -3.08e+05   -1.61e+05
month_ordinal    0.3251      0.050      6.451      0.000       0.225       0.425
==============================================================================
Omnibus:                        2.438   Durbin-Watson:                   0.194
Prob(Omnibus):                  0.296   Jarque-Bera (JB):                2.144
Skew:                           0.249   Prob(JB):                        0.342
Kurtosis:                       2.489   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BEDOK':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.439
Model:                            OLS   Adj. R-squared:                  0.433
Method:                 Least Squares   F-statistic:                     77.34
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           4.64e-14
Time:                        13:52:19   Log-Likelihood:                -717.96
No. Observations:                 101   AIC:                             1440.
Df Residuals:                      99   BIC:                             1445.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.123e+05   2.47e+04     -8.600      0.000   -2.61e+05   -1.63e+05
month_ordinal    0.2946      0.033      8.794      0.000       0.228       0.361
==============================================================================
Omnibus:                        3.488   Durbin-Watson:                   0.165
Prob(Omnibus):                  0.175   Jarque-Bera (JB):                2.987
Skew:                          -0.320   Prob(JB):                        0.225
Kurtosis:                       2.452   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BISHAN':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.442
Model:                            OLS   Adj. R-squared:                  0.437
Method:                 Least Squares   F-statistic:                     78.56
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           3.28e-14
Time:                        13:52:19   Log-Likelihood:                -727.83
No. Observations:                 101   AIC:                             1460.
Df Residuals:                      99   BIC:                             1465.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.353e+05   2.72e+04     -8.642      0.000   -2.89e+05   -1.81e+05
month_ordinal    0.3274      0.037      8.863      0.000       0.254       0.401
==============================================================================
Omnibus:                        0.151   Durbin-Watson:                   0.518
Prob(Omnibus):                  0.927   Jarque-Bera (JB):                0.256
Skew:                          -0.086   Prob(JB):                        0.880
Kurtosis:                       2.823   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BUKIT BATOK':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.402
Model:                            OLS   Adj. R-squared:                  0.396
Method:                 Least Squares   F-statistic:                     66.54
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.10e-12
Time:                        13:52:19   Log-Likelihood:                -771.73
No. Observations:                 101   AIC:                             1547.
Df Residuals:                      99   BIC:                             1553.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -3.386e+05   4.21e+04     -8.052      0.000   -4.22e+05   -2.55e+05
month_ordinal    0.4653      0.057      8.157      0.000       0.352       0.579
==============================================================================
Omnibus:                        5.633   Durbin-Watson:                   0.050
Prob(Omnibus):                  0.060   Jarque-Bera (JB):                3.186
Skew:                           0.219   Prob(JB):                        0.203
Kurtosis:                       2.248   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BUKIT MERAH':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.463
Model:                            OLS   Adj. R-squared:                  0.458
Method:                 Least Squares   F-statistic:                     85.45
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           4.88e-15
Time:                        13:52:19   Log-Likelihood:                -733.34
No. Observations:                 101   AIC:                             1471.
Df Residuals:                      99   BIC:                             1476.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -2.592e+05   2.88e+04     -9.015      0.000   -3.16e+05   -2.02e+05
month_ordinal     0.3606      0.039      9.244      0.000       0.283       0.438
==============================================================================
Omnibus:                        0.003   Durbin-Watson:                   0.868
Prob(Omnibus):                  0.998   Jarque-Bera (JB):                0.057
Skew:                           0.006   Prob(JB):                        0.972
Kurtosis:                       2.884   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BUKIT PANJANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.810
Method:                 Least Squares   F-statistic:                     426.9
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.11e-37
Time:                        13:52:19   Log-Likelihood:                -685.56
No. Observations:                 101   AIC:                             1375.
Df Residuals:                      99   BIC:                             1380.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -3.658e+05   1.79e+04    -20.418      0.000   -4.01e+05    -3.3e+05
month_ordinal     0.5022      0.024     20.663      0.000       0.454       0.550
==============================================================================
Omnibus:                        4.115   Durbin-Watson:                   0.403
Prob(Omnibus):                  0.128   Jarque-Bera (JB):                2.314
Skew:                           0.104   Prob(JB):                        0.314
Kurtosis:                       2.288   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'BUKIT TIMAH':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.009
Method:                 Least Squares   F-statistic:                   0.09582
Date:                Sun, 11 Jun 2023   Prob (F-statistic):              0.758
Time:                        13:52:19   Log-Likelihood:                -838.22
No. Observations:                 101   AIC:                             1680.
Df Residuals:                      99   BIC:                             1686.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -1.875e+04   8.12e+04     -0.231      0.818    -1.8e+05    1.42e+05
month_ordinal     0.0341      0.110      0.310      0.758      -0.185       0.253
==============================================================================
Omnibus:                      152.213   Durbin-Watson:                   1.589
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5589.155
Skew:                          -5.620   Prob(JB):                         0.00
Kurtosis:                      37.667   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'CENTRAL AREA':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.034
Model:                            OLS   Adj. R-squared:                  0.025
Method:                 Least Squares   F-statistic:                     3.531
Date:                Sun, 11 Jun 2023   Prob (F-statistic):             0.0632
Time:                        13:52:19   Log-Likelihood:                -852.62
No. Observations:                 101   AIC:                             1709.
Df Residuals:                      99   BIC:                             1714.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -1.683e+05   9.37e+04     -1.797      0.075   -3.54e+05    1.76e+04
month_ordinal     0.2388      0.127      1.879      0.063      -0.013       0.491
==============================================================================
Omnibus:                       94.383   Durbin-Watson:                   1.147
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2573.512
Skew:                          -2.545   Prob(JB):                         0.00
Kurtosis:                      27.199   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'CHOA CHU KANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.656
Model:                            OLS   Adj. R-squared:                  0.652
Method:                 Least Squares   F-statistic:                     188.5
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.18e-24
Time:                        13:52:19   Log-Likelihood:                -737.57
No. Observations:                 101   AIC:                             1479.
Df Residuals:                      99   BIC:                             1484.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -4.077e+05      3e+04    -13.599      0.000   -4.67e+05   -3.48e+05
month_ordinal    0.5585      0.041     13.730      0.000       0.478       0.639
==============================================================================
Omnibus:                       52.292   Durbin-Watson:                   0.058
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.546
Skew:                          -0.188   Prob(JB):                       0.0230
Kurtosis:                       1.715   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'CLEMENTI':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.591
Model:                            OLS   Adj. R-squared:                  0.587
Method:                 Least Squares   F-statistic:                     142.9
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           6.42e-21
Time:                        13:52:19   Log-Likelihood:                -753.34
No. Observations:                 101   AIC:                             1511.
Df Residuals:                      99   BIC:                             1516.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -4.133e+05    3.5e+04    -11.792      0.000   -4.83e+05   -3.44e+05
month_ordinal    0.5684      0.048     11.954      0.000       0.474       0.663
==============================================================================
Omnibus:                        4.188   Durbin-Watson:                   0.846
Prob(Omnibus):                  0.123   Jarque-Bera (JB):                3.514
Skew:                          -0.412   Prob(JB):                        0.173
Kurtosis:                       3.394   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'GEYLANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.405
Model:                            OLS   Adj. R-squared:                  0.399
Method:                 Least Squares   F-statistic:                     67.43
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           8.43e-13
Time:                        13:52:19   Log-Likelihood:                -740.93
No. Observations:                 101   AIC:                             1486.
Df Residuals:                      99   BIC:                             1491.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.492e+05    3.1e+04     -8.040      0.000   -3.11e+05   -1.88e+05
month_ordinal    0.3453      0.042      8.212      0.000       0.262       0.429
==============================================================================
Omnibus:                       13.214   Durbin-Watson:                   0.479
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               14.868
Skew:                          -0.758   Prob(JB):                     0.000591
Kurtosis:                       4.112   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'HOUGANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.674
Model:                            OLS   Adj. R-squared:                  0.670
Method:                 Least Squares   F-statistic:                     204.5
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           8.04e-26
Time:                        13:52:19   Log-Likelihood:                -720.79
No. Observations:                 101   AIC:                             1446.
Df Residuals:                      99   BIC:                             1451.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -3.586e+05   2.54e+04    -14.120      0.000   -4.09e+05   -3.08e+05
month_ordinal    0.4926      0.034     14.300      0.000       0.424       0.561
==============================================================================
Omnibus:                       54.420   Durbin-Watson:                   0.186
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.313
Skew:                           0.125   Prob(JB):                       0.0258
Kurtosis:                       1.706   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'JURONG EAST':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.320
Model:                            OLS   Adj. R-squared:                  0.313
Method:                 Least Squares   F-statistic:                     46.54
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           7.22e-10
Time:                        13:52:19   Log-Likelihood:                -717.94
No. Observations:                 101   AIC:                             1440.
Df Residuals:                      99   BIC:                             1445.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -1.639e+05   2.47e+04     -6.640      0.000   -2.13e+05   -1.15e+05
month_ordinal    0.2285      0.033      6.822      0.000       0.162       0.295
==============================================================================
Omnibus:                        6.050   Durbin-Watson:                   0.247
Prob(Omnibus):                  0.049   Jarque-Bera (JB):                5.968
Skew:                          -0.550   Prob(JB):                       0.0506
Kurtosis:                       2.542   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'JURONG WEST':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.486
Model:                            OLS   Adj. R-squared:                  0.481
Method:                 Least Squares   F-statistic:                     93.56
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           5.66e-16
Time:                        13:52:20   Log-Likelihood:                -725.47
No. Observations:                 101   AIC:                             1455.
Df Residuals:                      99   BIC:                             1460.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.531e+05   2.66e+04     -9.516      0.000   -3.06e+05      -2e+05
month_ordinal    0.3490      0.036      9.673      0.000       0.277       0.421
==============================================================================
Omnibus:                       73.291   Durbin-Watson:                   0.107
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.954
Skew:                          -0.133   Prob(JB):                       0.0187
Kurtosis:                       1.651   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'KALLANG/WHAMPOA':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.368
Model:                            OLS   Adj. R-squared:                  0.361
Method:                 Least Squares   F-statistic:                     57.59
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.80e-11
Time:                        13:52:20   Log-Likelihood:                -778.95
No. Observations:                 101   AIC:                             1562.
Df Residuals:                      99   BIC:                             1567.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -3.368e+05   4.52e+04     -7.457      0.000   -4.26e+05   -2.47e+05
month_ordinal    0.4650      0.061      7.589      0.000       0.343       0.587
==============================================================================
Omnibus:                        5.865   Durbin-Watson:                   0.199
Prob(Omnibus):                  0.053   Jarque-Bera (JB):                2.690
Skew:                          -0.025   Prob(JB):                        0.261
Kurtosis:                       2.202   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'MARINE PARADE':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.010
Method:                 Least Squares   F-statistic:                   0.01155
Date:                Sun, 11 Jun 2023   Prob (F-statistic):              0.915
Time:                        13:52:20   Log-Likelihood:                -739.98
No. Observations:                 101   AIC:                             1484.
Df Residuals:                      99   BIC:                             1489.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         2812.0445   3.07e+04      0.092      0.927   -5.81e+04    6.37e+04
month_ordinal    0.0045      0.042      0.107      0.915      -0.078       0.087
==============================================================================
Omnibus:                        2.413   Durbin-Watson:                   0.916
Prob(Omnibus):                  0.299   Jarque-Bera (JB):                2.126
Skew:                          -0.042   Prob(JB):                        0.345
Kurtosis:                       3.706   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'PASIR RIS':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.822
Model:                            OLS   Adj. R-squared:                  0.820
Method:                 Least Squares   F-statistic:                     456.0
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           7.70e-39
Time:                        13:52:20   Log-Likelihood:                -687.61
No. Observations:                 101   AIC:                             1379.
Df Residuals:                      99   BIC:                             1384.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -3.861e+05   1.83e+04    -21.117      0.000   -4.22e+05    -3.5e+05
month_ordinal     0.5297      0.025     21.354      0.000       0.480       0.579
==============================================================================
Omnibus:                        6.007   Durbin-Watson:                   0.294
Prob(Omnibus):                  0.050   Jarque-Bera (JB):                2.722
Skew:                          -0.020   Prob(JB):                        0.256
Kurtosis:                       2.197   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'PUNGGOL':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.850
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                     560.8
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.45e-42
Time:                        13:52:20   Log-Likelihood:                -687.00
No. Observations:                 101   AIC:                             1378.
Df Residuals:                      99   BIC:                             1383.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -4.254e+05   1.82e+04    -23.407      0.000   -4.61e+05   -3.89e+05
month_ordinal     0.5838      0.025     23.681      0.000       0.535       0.633
==============================================================================
Omnibus:                        7.442   Durbin-Watson:                   0.149
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                7.837
Skew:                           0.673   Prob(JB):                       0.0199
Kurtosis:                       2.775   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'QUEENSTOWN':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.611
Model:                            OLS   Adj. R-squared:                  0.607
Method:                 Least Squares   F-statistic:                     155.2
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           5.42e-22
Time:                        13:52:20   Log-Likelihood:                -745.22
No. Observations:                 101   AIC:                             1494.
Df Residuals:                      99   BIC:                             1500.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -3.96e+05   3.23e+04    -12.245      0.000    -4.6e+05   -3.32e+05
month_ordinal     0.5466      0.044     12.458      0.000       0.460       0.634
==============================================================================
Omnibus:                       10.445   Durbin-Watson:                   1.020
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               10.730
Skew:                          -0.677   Prob(JB):                      0.00468
Kurtosis:                       3.848   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'SEMBAWANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.556
Model:                            OLS   Adj. R-squared:                  0.552
Method:                 Least Squares   F-statistic:                     124.2
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           3.58e-19
Time:                        13:52:20   Log-Likelihood:                -761.42
No. Observations:                 101   AIC:                             1527.
Df Residuals:                      99   BIC:                             1532.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -4.19e+05    3.8e+04    -11.034      0.000   -4.94e+05   -3.44e+05
month_ordinal     0.5740      0.052     11.143      0.000       0.472       0.676
==============================================================================
Omnibus:                       20.653   Durbin-Watson:                   0.056
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.042
Skew:                           0.372   Prob(JB):                       0.0296
Kurtosis:                       1.942   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'SENGKANG':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.694
Model:                            OLS   Adj. R-squared:                  0.691
Method:                 Least Squares   F-statistic:                     224.9
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           3.17e-27
Time:                        13:52:20   Log-Likelihood:                -713.95
No. Observations:                 101   AIC:                             1432.
Df Residuals:                      99   BIC:                             1437.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -3.512e+05   2.37e+04    -14.797      0.000   -3.98e+05   -3.04e+05
month_ordinal      0.4828      0.032     14.996      0.000       0.419       0.547
==============================================================================
Omnibus:                       26.160   Durbin-Watson:                   0.086
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                7.651
Skew:                           0.376   Prob(JB):                       0.0218
Kurtosis:                       1.881   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'SERANGOON':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.478
Model:                            OLS   Adj. R-squared:                  0.472
Method:                 Least Squares   F-statistic:                     90.48
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           1.27e-15
Time:                        13:52:20   Log-Likelihood:                -721.40
No. Observations:                 101   AIC:                             1447.
Df Residuals:                      99   BIC:                             1452.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -2.379e+05   2.55e+04     -9.314      0.000   -2.89e+05   -1.87e+05
month_ordinal      0.3297      0.035      9.512      0.000       0.261       0.398
==============================================================================
Omnibus:                        5.662   Durbin-Watson:                   0.424
Prob(Omnibus):                  0.059   Jarque-Bera (JB):                2.812
Skew:                           0.124   Prob(JB):                        0.245
Kurtosis:                       2.221   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'TAMPINES':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.730
Model:                            OLS   Adj. R-squared:                  0.727
Method:                 Least Squares   F-statistic:                     267.7
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           6.57e-30
Time:                        13:52:20   Log-Likelihood:                -693.32
No. Observations:                 101   AIC:                             1391.
Df Residuals:                      99   BIC:                             1396.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -3.118e+05   1.93e+04    -16.117      0.000    -3.5e+05   -2.73e+05
month_ordinal      0.4295      0.026     16.363      0.000       0.377       0.482
==============================================================================
Omnibus:                        1.798   Durbin-Watson:                   0.184
Prob(Omnibus):                  0.407   Jarque-Bera (JB):                1.510
Skew:                           0.144   Prob(JB):                        0.470
Kurtosis:                       2.474   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'TOA PAYOH':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.320
Model:                            OLS   Adj. R-squared:                  0.313
Method:                 Least Squares   F-statistic:                     46.60
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           7.07e-10
Time:                        13:52:20   Log-Likelihood:                -766.22
No. Observations:                 101   AIC:                             1536.
Df Residuals:                      99   BIC:                             1542.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -2.663e+05   3.98e+04     -6.687      0.000   -3.45e+05   -1.87e+05
month_ordinal      0.3687      0.054      6.826      0.000       0.262       0.476
==============================================================================
Omnibus:                        3.702   Durbin-Watson:                   0.429
Prob(Omnibus):                  0.157   Jarque-Bera (JB):                3.371
Skew:                          -0.447   Prob(JB):                        0.185
Kurtosis:                       3.042   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'WOODLANDS':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.551
Model:                            OLS   Adj. R-squared:                  0.547
Method:                 Least Squares   F-statistic:                     121.7
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           6.24e-19
Time:                        13:52:20   Log-Likelihood:                -738.44
No. Observations:                 101   AIC:                             1481.
Df Residuals:                      99   BIC:                             1486.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -3.297e+05   3.02e+04    -10.901      0.000    -3.9e+05    -2.7e+05
month_ordinal     0.4526      0.041     11.032      0.000       0.371       0.534
==============================================================================
Omnibus:                       20.068   Durbin-Watson:                   0.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                5.321
Skew:                           0.164   Prob(JB):                       0.0699
Kurtosis:                       1.924   Cond. No.                     6.12e+08
==============================================================================
```

```
Linear regression model for 'YISHUN':
                            OLS Regression Results
==============================================================================
Dep. Variable:                   mean   R-squared:                       0.635
Model:                            OLS   Adj. R-squared:                  0.631
Method:                 Least Squares   F-statistic:                     172.1
Date:                Sun, 11 Jun 2023   Prob (F-statistic):           2.19e-23
Time:                        13:52:20   Log-Likelihood:                -725.58
No. Observations:                 101   AIC:                             1455.
Df Residuals:                      99   BIC:                             1460.
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -3.451e+05   2.66e+04    -12.960      0.000   -3.98e+05   -2.92e+05
month_ordinal     0.4739      0.036     13.120      0.000       0.402       0.546
==============================================================================
Omnibus:                      170.887   Durbin-Watson:                   0.046
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                9.128
Skew:                           0.025   Prob(JB):                       0.0104
Kurtosis:                       1.528   Cond. No.                     6.12e+08
==============================================================================
```

Figure 5 Statistical models based on town

Market trend regression

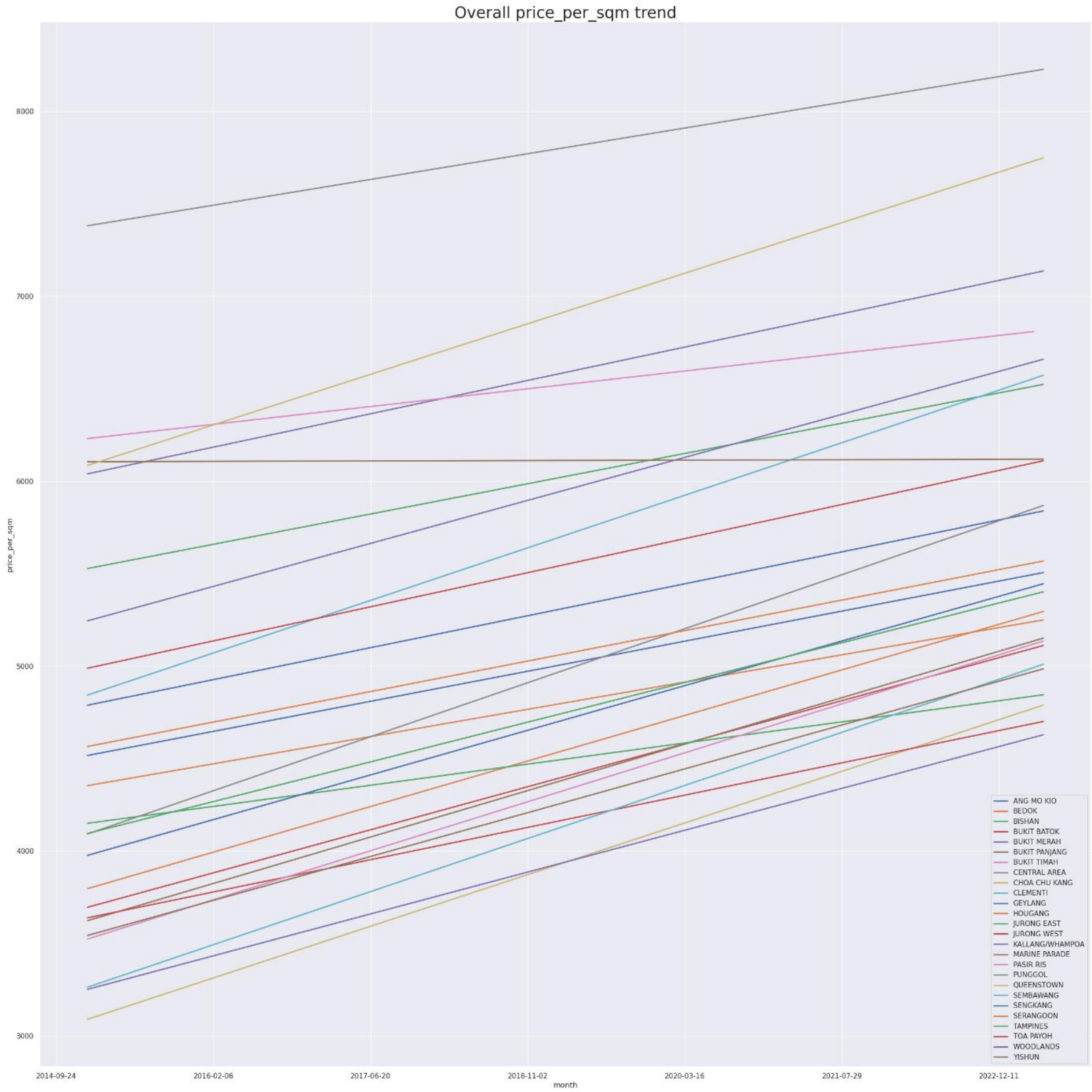

Figure 6 Linear regression model based on mean price of HDB flats catergorize by town

Data Analysis

Collection of data

The data that is used for the study is available for download at data.gov.sg. There are 5 datasets which are categorize based on year range as the following:

1. January 1990 to December 1999
2. January 2000 to February 2012
3. March 2012 to December 2014
4. January 2015 to December 2016
5. January 2017 to May 2023

It is observed that there is missing feature 'remaining lease' for dataset 1 to 3. Due to this reason, these datasets are not included into the study. The dataset comes with the following features:

Feature name	Description	Example	Data type
month	Date when flat registered in resale market	2012-03	datetime
town	Flat town location	ANG MO KIO	category
flat_type	Flat type	Improved	category
block	Flat block number	700A	category
street_name	Flat street name	ANG MO KIO AVE 6	category
storey_range	Flat storey range	11 to 15	category
floor_area_sqm	Flat size (in $metres^2$)	90	numeric
flat_model	Flat model	Model A	category
lease_commence_date	Date flat first release for lease by HDB	2003	datetime
remaining_lease	Total months left before surrender back to HDB	20	numeric
resale_price	Flat price (in SGD)	596000	numeric

Cleaning data

The features which are excluded are:

1. block

2. street_name
3. lease_commence_date

This is because these features not useful for the objective of the data analysis which is to observe the resale market trends over the years. Even though lease_commence_date is useful for the study to calculate the remaining lease, there is a feature remaining_lease which is available for the analysis.

Analyzing the data

Finding the correlation

Based on the original dataset correlation matrix, The following numerical features has strong correlation to its categorical features.

Numerical feature	Categorical features
resale_price	flat_type, flat_model
remaining_lease	flat_model, town
floor_area_sqm	flat_type

Given that linear regression model will be utilized to make prediction of the resale market trend. There is a need to combine these numerical features into a new feature as a predictor value for the regression model. By doing this, it will be simpler to generate the regression model.

Adding missing feature

Deeper observation, it is found that there is missing feature in the provided dataset which is 'price per sqm'. This was resolved by creating a new column in the dataset using the formula:

$$price\_per\_sqm = \frac{resale\_price}{floor\_area\_sqm \times remaining\_lease}$$

The formula is derived by simple division to get the price per square metres and knowledge of HDB lease decay. With the new added feature, price_per_sqm, the new correlation matrix with the shows that price_per_sqm has moderate positive correlation between town and resale_price which exactly fulfill the objective of the analysis.

Result

This model reveals that areas such as Punggol, Sembawang, and Queenstown are shown to have growing trends. In contrast, one would like to avoid places like Central Area, Bukit Timah and Marine Parade as these places are expensive and have slow property growth. These conclusions are based on the mean values fed into the linear regression model. However, note that these results may not be accurate for the slow-growth areas such as Orchard and Bukit Timah, and Marine Parade, as these areas have the least resale flats available in the dataset. For further information on the data analysis process, please visit the following link.

Conclusion

In conclusion, we have developed a linear regression model with a time series analysis of the resale flats trends that allows us to discern a quantitative relationship between the various variables available in the dataset. It allows us to understand better how flat pricing has been affected by its qualities over the years. It is essential as we intend for this to be used by potential home buyers to determine which flat qualities and which town or neighbourhood they should focus on, based on the last seven years of data, to allow them to lose the least amount or earn the most amount of money at the point that they decide to sell the flat to upgrade to their dream home. This prediction model is mainly meant for first-time homeowners who intend to purchase or upgrade to a condominium or landed property in the future but are in urgent need of housing and do not have the finances to support their goals of purchasing their dream home.

However, there is some room for improvement in the model that we have developed to tackle this problem. Firstly, we could include a more extensive dataset and split units based on the various flat types. One example is that there are different 4-room flat types, including 4I, 4A, 4S, and 4NG, to name a few. It will allow us to generate a more accurate model that can predict the price changes between the various flat types and this will allow the user to make a better judgement on what type of flats he or she should go for. At the moment, due to time constraints, we could only broadly classify all the flats under the common types of 1-room, 2-room, 3-room, 4-room, 5-room, executive and multi-generation flats.

Secondly, we could identify the various auto-correlation issues that may be present in the dataset because the data was collected from different periods and resolve those issues. It will, in turn, allow us to generate a more accurate prediction model and, therefore, increase the reliability and usability of the prediction model. It is vital as we intend for this model to be used by first-time homebuyers to make the best choice, and they may not have first-hand knowledge of the past transaction history of the estate they are considering. Therefore, it will have another benefit for first-time homebuyers as they can refer to the analysis before deciding on what they wish to purchase and, after that, do more research into the flat type, qualities and town or neighborhood they wish to have for their first HDB flat.

Finally, we discovered there was there missing information, such as unit price per sqm as mentioned earlier. Another example would be that the "remaining lease" column was missing in three of the five Excel files obtained from data.gov.sg. Therefore, it limits the amount of usable data for this analysis. Being able to include these three Excel files in our analysis could significantly affect it as there are other data missing from all five datasets. Additionally, if the dataset contains the duration the resale flat is placed in the market before it gets sold, this data may influence the buyer's choice during their flat selection.

Reflection

Amizzuddin

I would like to comment on the efforts made by every team member in the group. Everyone works in parallel so that every part of the group work can progress despite the need for more data clarity from other team members. It is essential because this way of working will allow continuous working and improvement. It is, in fact, the main idea of working agile. Everyone volunteers on the tasks where they excel, which shows excellent initiative. I am happy and look forward to working with the same team member again.

On a personal reflection, this project exposed me to different ways of generating linear regression on time series data. Finding the correct method to approach was not easy, but it helped me think of ways to manipulate the data such that the variable used is 'fair' to feed into the linear regression model.

Gim Long

I would like to state that I had a great time working with this fantastic team. Everyone did their utmost and gave good quality work, ensuring the project's progress was smooth sailing. It allowed us to complete our report and presentation slides around one week before the deadline and gave us enough time to finetune our work further. Everyone was cooperative, kept to a given timeline, and took pride in their work, which made the team efficient. Along this journey, I also learned more about QR techniques and how they can be applied in my daily life. I will work with everyone in this team again if the opportunity arises.

Chen Xin

This is my first learning and working on quantitative reasoning. It helps me to have a better understanding of how we shall conduct research. On the other hand, I gained a better understanding of linear regression models, such as what it is suitable for and the limitations of it. In this project, we did not prepare or clean up the data well, so the results would not be good enough. However, the process of making this report is more important. Along the way, we got to know each other better and helped each other out; we strengthened what we had learned. It is a fantastic team and experience.

Nurhidayat

The team plays to each other's strengths, making the group efficient. We completed most of our work by week 4, with minor changes on week 5. The team was also communicative, so we had no problems sharing ideas and minimal disagreement due to the common goal to finish the project as efficiently and to the best of our potential. I give 10/10 to everyone's motivation to finish this before the summer break.

Lux

This project taught me that there are many more aspects to a QR presentation than just the data itself, although the data is the crux of it. Overall, QR is an interesting and useful technique that can be used in numerous aspects of life and at varying levels, from personal to corporate levels. I am encouraged and motivated to see my teammates work so efficiently, complimenting each other's strengths while still being able to harmoniously sync the different parts of the project and reach a conclusion to the best of our efforts.