Changes in second-hand housing transaction prices in Singapore from 2017 to 2018

(Proposal Report)

阎博(Student ID: 222025004)

陈庄(Student ID: 222025021)

林庆嘉(Student ID: 222025012)

邓哲涵(Student ID: 222025043)

刘佩佩(Student ID: 222025019)

# Background—introduction background and purpose of report

## 1.1 Background

Singapore is one of the best examples of the integration of urban economy, capital market and real estate development in the world. The introduction of an appropriate housing price policy can not only bring impetus to urban economic growth, but also effectively promote the accumulation of human capital and provide a good environment for industrial transformation and upgrading. Currently, there are five main types of property in Singapore, but not all of them are available for foreigners to buy. For example, Singapore stipulates that only domestic residents can buy HDB flats, and permanent residents can only buy second-hand HDB flats. In order to attract more foreigners to buy property in Singapore, Singapore has introduced many preferential policies about flats purchasing. Therefore, second-hand housing market plays a vital role in Singapore's real estate industry. According to the definition of the Urban Redevelopment Authority (URA), Singapore can be divided into three regions: Core Central Region (CCR), Rest of Central Region (RCR) and Outside Central Region (OCR). Different houses in different area may also have various prices.

## 1.2 Purpose of report

According to the price data of second-hand housing in Singapore from 2017 to 2018, we use multiple linear regression and hypothesis testing methods to explore what factors affect housing prices most. In the meanwhile, we also want to find out whether they are positively or negatively correlated with housing prices.

# 2.Preliminary data description

The data collected initially had 92270 cases of flat prices with their attributes (Figure 1).

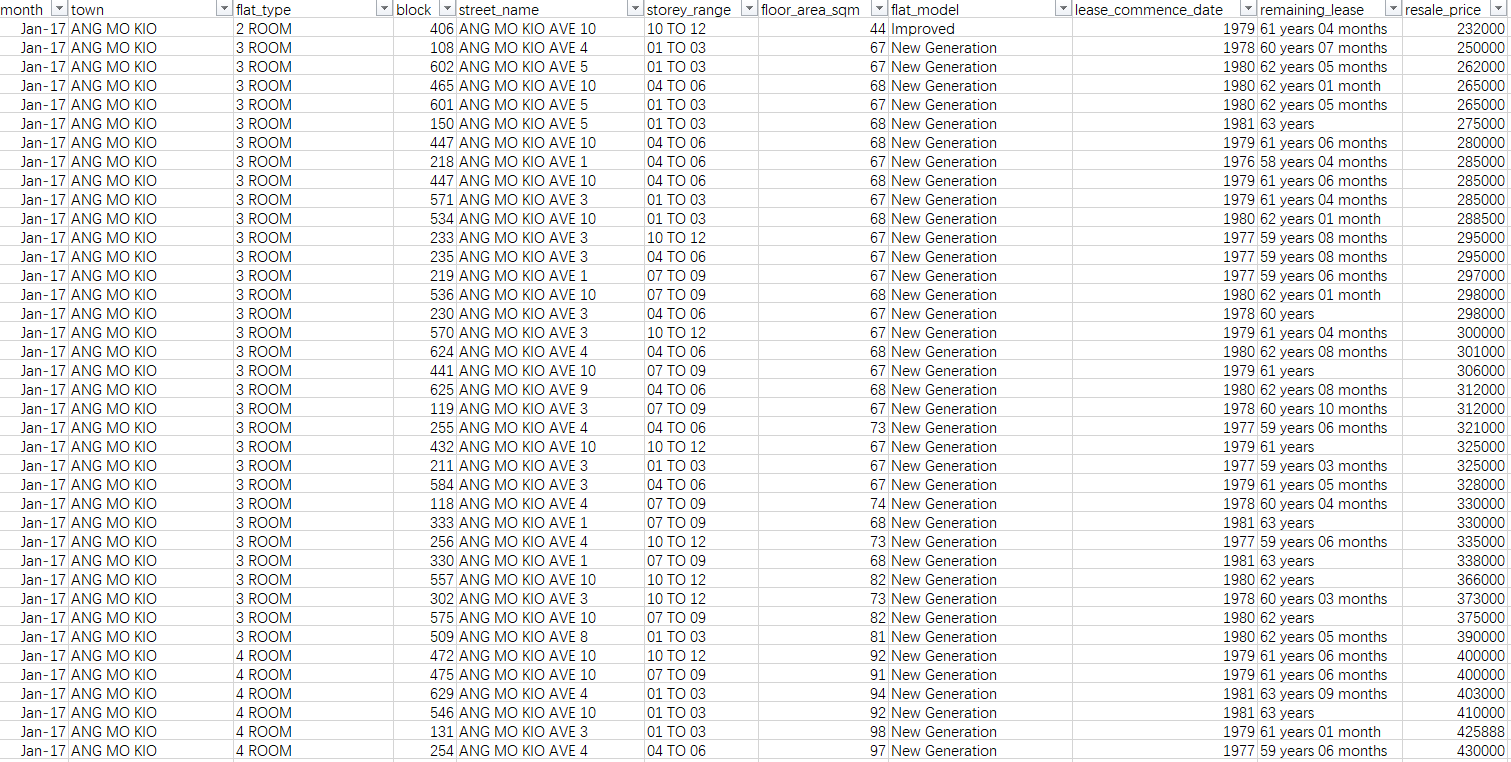


Figure 1 Part of original data

## 2.1 Steps of data pre-processing

（1）The data is cleaned by deleting those rows which have missing values

（2）The data is divided into two parts according to the row ‘month’, which are time of records of house prices. Here one part contains records before January 2019 for regression model building, the other part contains records after January 2019 for validation.

（3）Since the houses are specified into unit of town, ‘street\_name’ column is deleted since it is useless anymore, and the ‘block’ column is also deleted because under consideration, it is not meaningful in regression model

（4）For ‘town’, ‘flat\_type’, ‘flat\_model’ which are categorical variables, since they are needed to be included in regression model, they are converted from categorical/text variable into dummy variables through OrdinalEncoder module of sklearn.preprocessing package in Python. The dummy variables and their corresponding categorical variables of these three columns of variables are shown in Figure 2.1, Figure 2.2, and Figure 2.3 respectively.



Figure 2.1 Dummy variables and corresponding categorical variables of column ‘town’

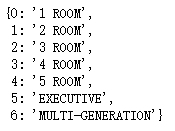


Figure 2.2 Dummy variables and corresponding categorical variables of column ‘flat\_type’

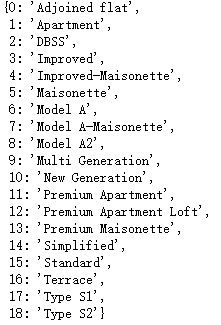


Figure 2.3 Dummy variables and corresponding categorical variables of column ‘flat\_model’

（5）In origin, the column ‘storey\_range’ represents approximate storey of the flats, which are intervals and in datatype of string. Therefore, in data pre-processing, it is converted into average value of its lower bound and upper bound.

（6）For column ‘remaining\_lease’, originally the values of this column are in format of ‘YY years MM months’, which are also in datatype of string, for convenience of later regression modeling and analysis, they are converted into datatype of integer and in unit of month.

Data after pre-processing is shown in Figure 2.4

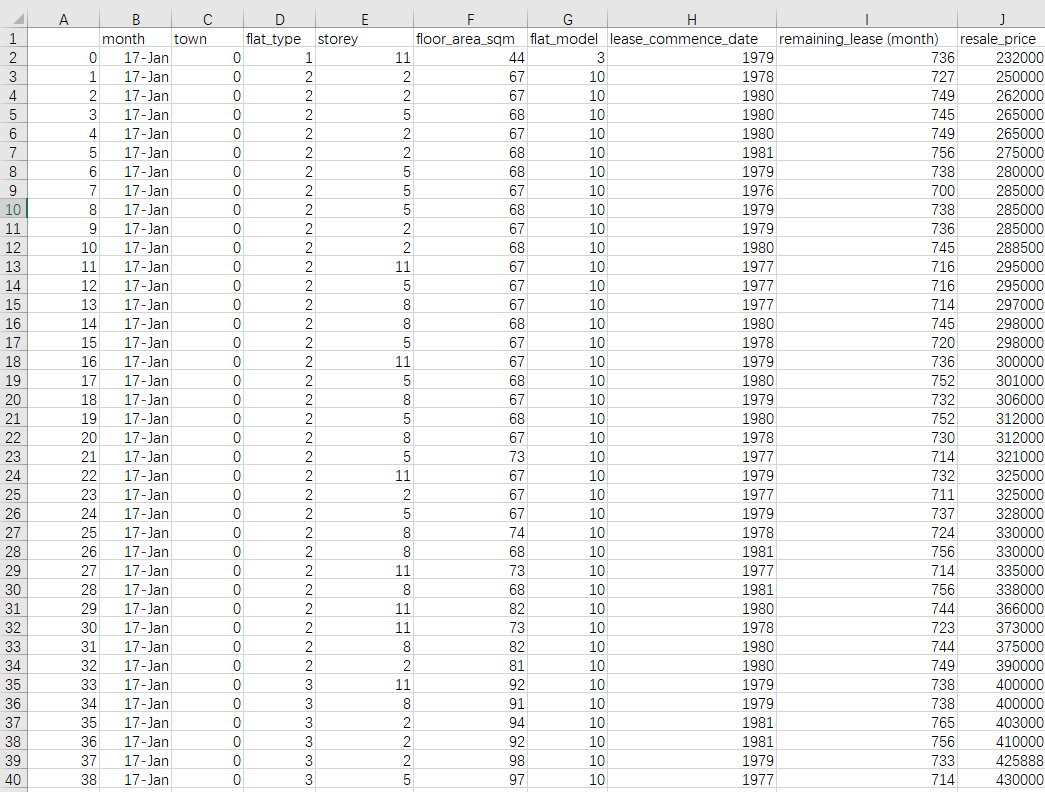


Figure 2.4 Part of data after pre-processing

## 2.2 Data description

|  |  |
| --- | --- |
| Variable Name | Description |
| month | Time of record of house price (in month) |
| town | Which of town the house is located in |
| flat\_type | Type of flat house |
| storey | The storey (floor) of flat house |
| floor\_area\_sqm | The area of flat house (in square meter) |
| flat\_model | Model of flat house |
| lease\_commence\_date | The start time of lease commencing |
| remaining\_lease (month) | Remaining time of lease (in month) |
| resale\_price | Price of second-hand house |

# 3.Research methodology and hypothesis

## 3.1.Multiple Linear Regression

Core question: What are the factors that affect housing prices? Are they positively or negatively correlated with house prices?

Factors : house area, type of flat house, the storey (floor) of flat house, the area of flat house (in square meter), remaining years of property rights, etc.

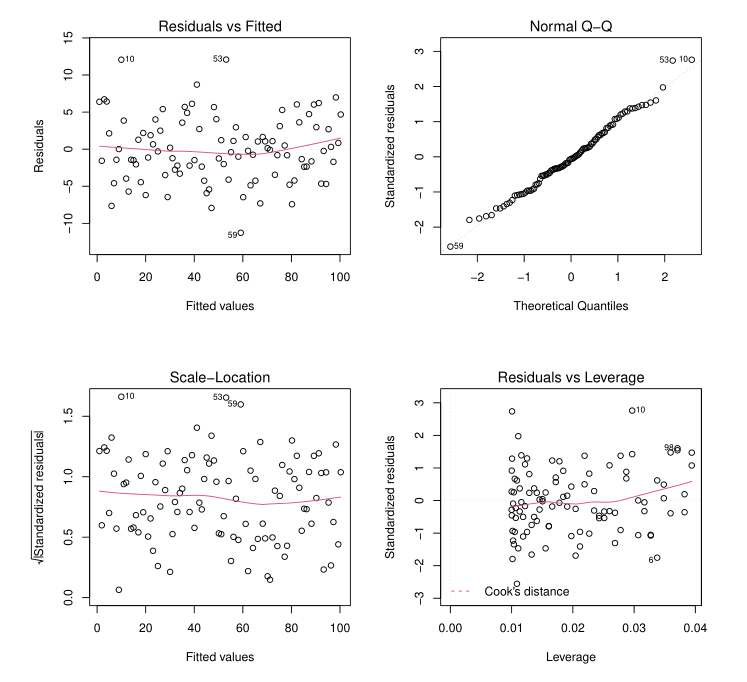
Including:

(1) Find correlation (scatter plot between two variables, Pearson correlation coefficient r)

(2) Regression (ANOVA table, R)

(3) Test of collinearity between independent variables

(4) Residual analysis



## 3.2 Hypothetical test

According to the definition of the Urban Redevelopment Authority (URA), Singapore can be divided into three regions: Core Central Region (CCR), Rest of Central Region (RCR) and Outside Central Region (OCR). The average house price between these administrative areas will be compared.

H0a: The average price of second-hand housing in the Central Region is not smaller than those in Outside Central Region.

H1a: The average price of second-hand housing in the Central Region is smaller than those in Outside Central Region.

H0b: The average prices of second-hand housing in the Core Central Region are not smaller than those in the Rest of Central Region.

H1b: The average price of second-hand housing in the Core Central Region are smaller than those in the Rest of Central Region.

## 3.2.1 T-test

data <- read.csv("h\_data.csv")

Filter out houses in Outside Central Region

CR <- subset(data,data$type != "OCR")

Filter out houses in Central Region

OCR <- subset(data,data$type == "OCR")

PCR=CR$price\_pre\_sqm

POCR=OCR$price\_pre\_sqm

x1=mean(PCR)

x1

## [1] 6001.862

x2=mean(POCR)

x2

## [1] 4187.315

s1=var(PCR)

s1

## [1] 2209270

s2=var(POCR)

s2

## [1] 609594.8

t=(x1-x2)/sqrt((s1/length(PCR))+(s2/length(POCR)))

t

## [1] 107.3671

t\_base=qt(0.05,length(PCR)+length(POCR)-2)

t\_base

## [1] -1.64489

Result: Because t>-t\_base=1.64489, we can not reject h0.

**Therefore, we have 95% confident that the average price of second-hand housing in the Central Region is not smaller than those in Outside Central Region.**

## Wilcoxon Rank Sum Test

A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations. Recall that the parametric test compares the means (H0: μ1=μ2) between independent groups.

In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

H0: The two populations are equal versus. (The average price of second-hand housing in the Central Region is equal to those in Outside Central Region.)

H1: The two populations are not equal. (The average price of second-hand housing in the Central Region is not equal to those in Outside Central Region.)

# help(wilcox.test)

wilcox.test(price\_pre\_sqm~CR\_OCR,data=data)

##

## Wilcoxon rank sum test with continuity correction

##

## data: price\_pre\_sqm by CR\_OCR

## W = 244739953, p-value < 2.2e-16

## alternative hypothesis: true location shift is not equal to 0

Since p-value is smaller than 0.01, we reject H0.

**We have 99% confident that the average price of second-hand housing in the Central Region is different from those in Outside Central Region.**

# 4.Summary and Recommendation

In this project, we have made a whole process of data mining and also have found some results out of it.

In the data processing part, we focus on selection of the data and the transformation of it. Through this, the data mining part can be benefit.

In the data mining part, we use multiple linear regression and hypothesis testing methods to explore what factors affect housing prices most, meanwhile also try to find out whether they are positively or negatively correlated with housing prices.

From result of T-test ,the results show that firstly we have 95% confident that the average price of second-hand housing in the Central Region is not smaller than those in Outside Central Region. In the meantime, the outcome of Wilcoxon Rank Sum Test shows that We have 99% confident that the average price of second-hand housing in the Central Region is different from those in Outside Central Region.

For Multiple Linear Regression….[待补充]

In the future, our team is focusing on optimization of this project. Due to the results, there are some helpful ideas that we want to give out. Firstly, we want to apply some methods like normal distribution check to see whether the basic assumptions are met in our hypothesis. Furthermore, we would like to introduce F-test to this, and concentrate on the value of variance. What’s more, the optimization of our data mining method is also the key to success. Last but not the least, we would like to digger more data of different years, in order to create the validation of our whole model, and also can be as the expansion of the datasets.