Harvardx Data Science - HK Properties Price Prediction Project

William Tao

2022-11

Contents

1 Introduction

Property investment is one of the most important investment decisions in modern human history. No matter people need houses for livings or for asset investments, they always want to ensure that the houses they are going to purchase are properly priced. Identifying the true and fair market value of a house, therefore, will be beneficial for people to make purchasing decision. In general, the market price of a house depends on various factors including property size, location and others. We want to study how different factors affect the transaction price of a property and try to predict this transaction price based on these factors.

As one of the citizens in Hong Kong, in this HarvardX Capstone project, I am going to develop a property price prediction model based on the dataset of the "Hong Kong Property transactions Midland until 2018" cited from Kaggle [1]. This dataset is provided by a property agency in Hong Kong called Midland Realty [2]. According to the Kaggle page, this dataset records the real estate transactions in Hong Kong from 2003 to 2018. It contains information about the real estate transactions obtained from the Land Registry Department of the Government of the Hong Kong SAR and also from Midland Realty's own sources. It includes various attributes of each property transaction including the property price, property size, property address, as well as the date of transaction being perpared by Midland.

In this project, the objective is to development property price prediction models for the downtown area "Kowloon" of Hong Kong using the property transaction data from 2009 to 2018, which is the data in 10 years period. Multiple machine learning algorithms will be used in this project to examine the speed and performance of the property price models being developed based on these algorithms

In the following report, we explored and analyzed the selected variables in each data attribute types (factors) in the Hong Kong Properties Dataset. Data wrangling was performed to clean the property records. The variable types that were relevant for price prediction were selected for further data exploration and data visualization. Then, the cleaned dataset was split into training and test sets for model developments and testings. Four machine learning models were selected for the real estate price prediction model development. The prediction results of these four models will then be presented and compared. Finally, the conclusion will be presented to summarize this work and to discuss the possibility of further prediction accuracy improvement.

2 Data Source

The “Hong Kong Property transactions Midland until 2018” dataset consists of 16 different variables with in total 1262519 rows of data entries. The dataset was stored in a CSV file and there was NO variable names in the columns. The column names were only explained in the dataset’s Kaggle page [1]. These 16 columns of variables represent the following property’s data types:

1. Building Unit Type
2. Building Location
3. Building Estate
4. District
5. Longitude
6. Latitude
7. Building Floor
8. Building Unit
9. Building Net Area in square feet
10. Building Gross Area in saqure fee
11. Building Price in Million HK Dollar
12. Price per net area
13. Price per gross area
14. Transaction date
15. Percentage of change from previous transaction
16. Date of previous transaction

In this dataset, the first 4 columns (Building uint type, Building location, Building Estate) are in character types. The “Transaction date” and “Date of previous transaction” are with data in year-month-day format. “% change from previous transaction” is a percentage with plus and minus there. All the rest of are numerical data.

The following code shows the column names of the raw dataset and the first few rows of the dataset:

```
##  [1] "type"           "location"       "estate"         "district"
##  [5] "long"           "lat"            "floor"          "unit"
##  [9] "netArea"        "grossArea"      "totalPrice"     "
    netfeetPrice"
## [13] "grossfeetPrice" "salesDate"      "priceChange"    "lastDate"
```

type	location	estate	district	long	lat	floor	unit	netArea	grossArea	totalPrice	netfeetPrice	grossfeetPrice	salesDate	priceChange	lastDate
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	A	787	-	1165	14,806	-	2016-08-25	+86.4%	1996-11-26
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	A	787	-	625	7,943	-	1996-11-26		
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	B	624	-	22	352	-	2016-11-28	-96.6%	2011-08-22
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	B	624	-	647	10,360	-	2011-08-22	+7.8%	1997-04-14
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	B	624	-	600	9,611	-	1997-04-14	+25.0%	1996-11-26
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	B	624	-	480	7,688	-	1996-11-26		
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	C	624	-	455	7,288	-	1996-06-04		
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	D	708	876	530	7,483	6,050	1997-01-13		
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	E	766	-	938	12,239	-	2011-06-21	+45.7%	2009-06-23
B000000014	Taikoo Shing	Fu Tien Mansion	EST	114.218	22.2871	26	E	766	-	644	8,400	-	2009-06-23	+29.3%	2006-10-05

3 Data Preparation

3.1 Variables Explanations

In this project, we will select the useful factors from the 16 columns to build property price prediction models.

In these 16 types of factors, the “Building Unit Type” is a property type’s code. There is no further information from the dataset’s Kaggle page, or from Midland Realty’s website, or from the Hong Kong Government’s Department about the definition of this “Building Unit Type”. It may probably be an internal building type’s identifier being used by Midland Realty. So, we do not include this factor in our analysis.

For “Building Location”, “Building Estate” and “District”, they are the names of the real estate’s location, building estate and district respectively in Hong Kong. These factors are important to determine the property pricing.

The “Longitude” and “Latitude” consist of the geolocation information of the property. These two factors can be considered as duplicated locational information. They are similar to the detailed property’s address which also presents in the previous 3 factors. They will be used for data visualization but will not be used for model training.

The “Building Floor” mentions about the level of this property located in a building tower. In Hong Kong, most of the properties, e.g. residential flats, are located in different floors of a high rising building. The residential buildings usually have 5 to 60 levels. The level/floor where a property locates have significant effect to the house price. This factor shall be included

The “Building Unit” describes the location of a property (e.g. a flat in a building tower) in a floor of a building tower. In Hong Kong, it is common to find several property units share the same floor (e.g 2 to 8 residential flat share same building floor). Properties locate in the different floor but in the same unit number usually share the same orientation and field of view. Some properties with the same unit number of building tower may have much better view than those with different building unit number. As better field of view means higher price in Hong KOng, this factor is actually important for house price modeling. However, the definition and orientation of a “Building Unit” (e.g. flat number) vary across different estates and buildings. This means flat “A” in building X having a very good view does not mean the flat “A” in building “Y”also have a nice view. The “Building Unit” here, therefore, may not contain any information about the field of view and orientation of each building tower in this dataset. Without further information gathering and preparation, it may, tnot be so meaningful to include this “Building Unit” factor in our data analysis. In this capstone project, “Building Unit” will not be used for model building.

For “Building Net Area” and “Building Gross Area”, they refer to the sizes of a property. The “Building Net Area” only counts the actual area of the property itself. Whereas the “Building Gross Area”, on the other hand, includes the shared common areas. In Hong Kong, the Government has revised the rule that mainly the net building area should be used to define the per square feet price of a property. So, we will only include “Building Net Area” in this project.

The 3 factors “Building Price hkd”, “Price per net area” and “Price per gross area” state the price of a property in Hong Kong Dollar. In our project, we mainly include “Price per net area” in our analysis and model training.

The “Transaction Date” records the time of sale of a property. This field includes, the transaction year, month and day there. To simplify our analysis, we will extract the transaction Year of a property for our model development.

For "% change from previous transaction" and "date of previous transaction", these 2 factors contain information about the previous transaction and change of property change there. We will ignore these 2 factors in this project.

With the above consideration, the following factors will further be explored

1. Building Location
2. Building Estate
3. District
4. Building Floor
5. Building Net Area in square feet
6. Price per net area (in HKD)
7. Transaction Year

3.2 Data Wrangling

3.2.1 Data Cleaning

After picking the 7 useful factors for property price model development, the data in each variables are converted into right data types (e.g. from character to numeric, from character to date format…etc). There are empty or NA fields in the entries of the dataset across the selected factors. These entries with empty or NA fields have all been removed. Moreover, we also find that there are 28840 duplicated data entries (rows) in the dataset. These duplicated entries will also been removed to avoid double counts.

The transaction year of a property is extracted from the "Transaction Date" in the original dataset. To simplify our analysis, only transaction year will be used in data evaluation model training.

After data cleaning, there are in total 836711 row of data entries left. The whole dataset is still sizable.

3.2.2 Data Selection

In this part, the first thing we are going to do is to check the integrity of property pricing in the dataset and to remove those entries which have abnormally low and high per square feet price. According to reference [3], the average property price in 2003 (the lowest after 1997) is around HKD2,000 per net square feet. As we are going to analyze pricing data after 2009, we assume that the lowest per net square feet price is HKD2,000 and remove those entries lower than this price. According to [4], the highest per square feet price in Hong Kong history is HKD137,872. We will also remove those entries with the per net square feet price higher than this number.

As defined in the Introduction section, our scope is to build a prediction model for the Kowloon area in Hong Kong using the property transaction data from 2009 to 2018. The data entries that are out of this range are removed from the dataset. Moreover, we are targeting to study the property prices in the Kowloon downtown area of Hong Kong. We would only select the data entry with district codes that belong to the Kowloon downtown area. According to reference [5], The following 5 districts are in Kowloon downtown area:

1. Kowloon City (HASC: HK.KC)
2. Kwun Tong (HASC: HK.KU)
3. Sham Shui Po (HASC: HK.SS)
4. Wong Tai Sin (HASC: HK.WT)
5. Yau Tsim Mong (HASC: HK.YT)

However, the district codes in the dataset are not the same as HASC being used by the Hong Kong Government. By studying the ‘location’ and ‘estate’ variables, we can do the matching of the code with the following mappings:

1. KC = HK.KC
2. KT = HK.KU
3. SSP = HK.SS
4. WTS = HK.WT
5. YTM = HK.YT

Then, we only select the ‘district’ with the codes of “KC”, “KT”, “SSP”, “WTS” and “YTM” for our data analysis and modeling.

With all these data selection work, the final dataset size is reduced to 32140

The following shows the structure of the latest hkproperty dataset. It consists of 6 numerical variables and 3 categorical variables.

```
## ’data.frame’:   32140 obs. of  9 variables:
##  $ netfeetPrice: num  21089 18733 17013 10684 15259 ...
##  $ location    : Factor w/ 127 levels "26 Kimberley Road",..: 98 98
    98 98 98 98 98 98 98 98 ...
##  $ estate      : Factor w/ 556 levels "26 Kimberley Road",..: 169
    169 169 169 169 169 169 169 169 169 ...
##  $ district    : Factor w/ 5 levels "KC","KT","SSP",..: 3 3 3 3 3 3
    3 3 3 3 ...
##  $ long        : num  114 114 114 114 114 ...
##  $ lat         : num  22.3 22.3 22.3 22.3 22.3 ...
##  $ floor       : num  28 28 28 27 26 26 25 24 23 22 ...
##  $ netArea     : num  1180 1046 1046 897 1114 ...
##  $ salesYear   : num  2013 2016 2012 2010 2012 ...
```

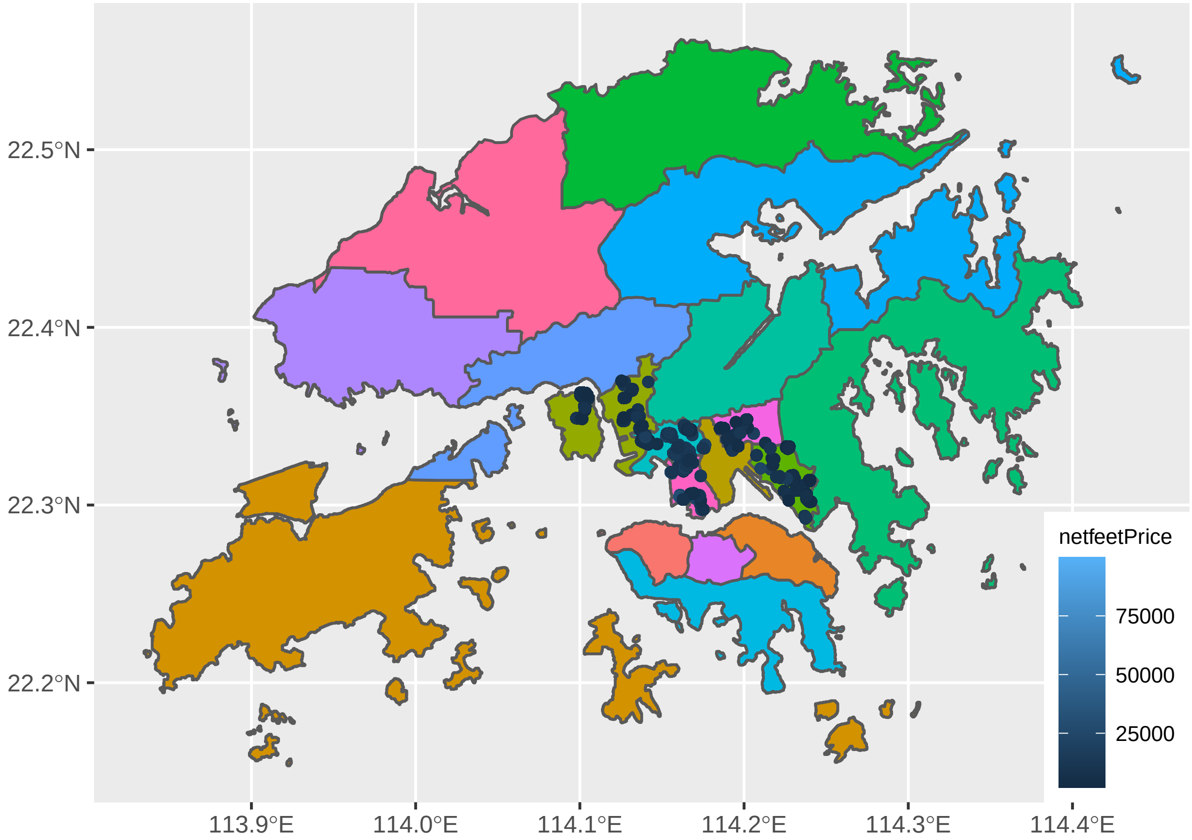
For the categorical variable, “estate”, it consists of 556 distinct factors. In our project, we will use decision tree type of regression algorithms to build prediction models. If we include “estate” variable in model training, it will take a very lengthy computational time. Since the geolocation information is also included in ‘location’ and ‘district’ variables, we choose to remove this variable from model predictions. And as mentioned earlier, the “long” and “lat” variables will be used for data visualization. They will also be removed from model training.

4 Data Exploration and Visualization

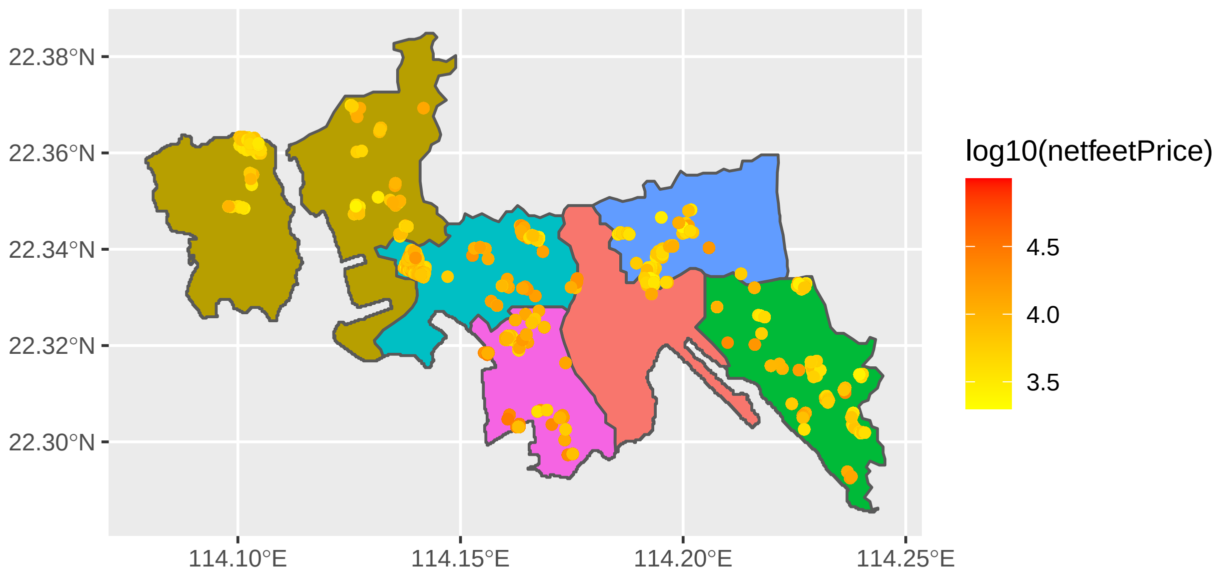
This section explores different types of selected variables in the Hong Kong property transaction dataset with the following paragraphs.

4.1 Transcation Location

The following diagram exhibits the map of Hong Kong with 18 districts in different colors with spatial data [7]. The blue color points are the locations of the transactions in the selected 5 downtown districts in the Kowloon area of Hong Kong.



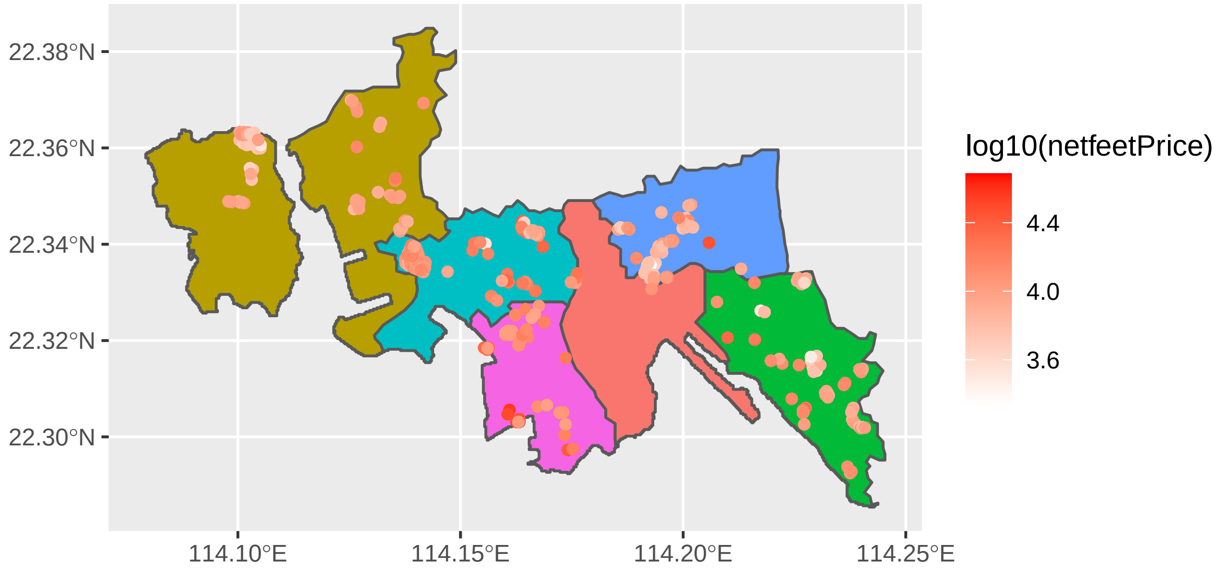
The following graph shows the zoom in view of the districts with property transactions. Surprisingly speaking, we can find that these property transaction records covers 6 districts instead of 5. This is because the dataset creator, Midland, accidentally combined 2 districts into one and included the district called “Kwai Tsing”, which belongs to New Territories region, with “Sham Shui Po” District”. As “Kwai Tsing” district is also one of the highly populated area in Hong Kong, it is decided to also include this district in our data analysis and model training.



Besides the issue being identified, we can also see that the property transaction records (there are in total 32140 records in 10 years period) are highly concentrated into some areas. Moreover, the district in RED color, called Wong Tai Sin (WTS), had not too many transaction records. This can be explained by:

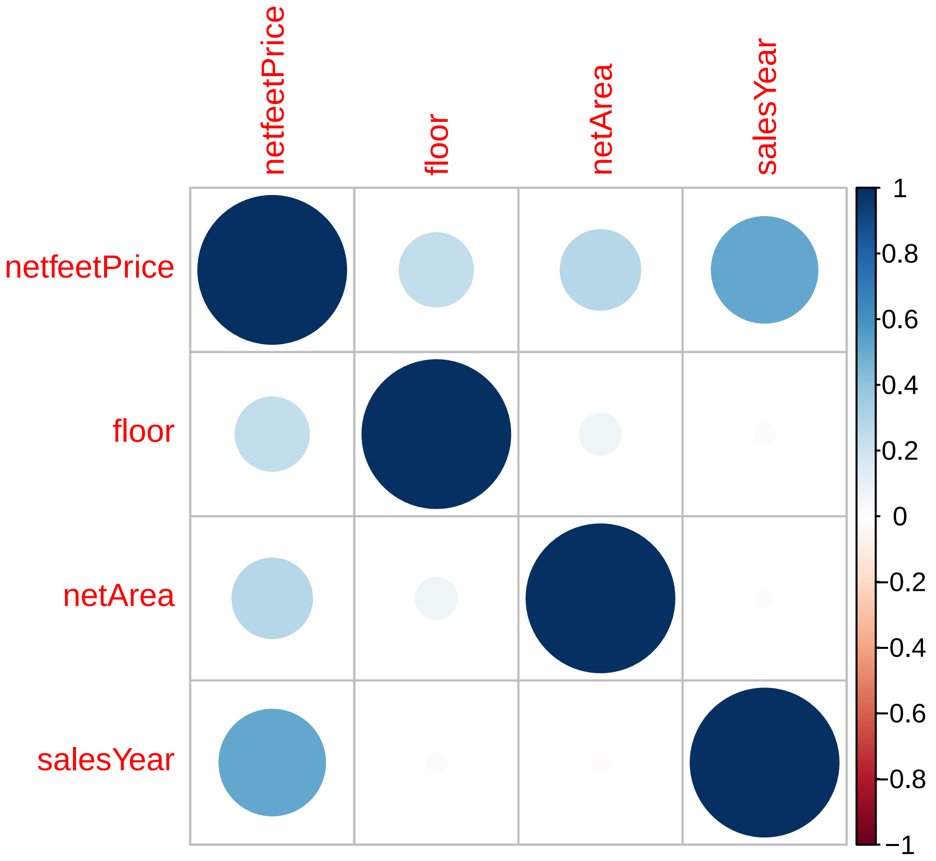
1. Property transactions usually happened in the key estates, garden and buildings
2. The footprint of the real estate agencies were mainly in the area with property transactions
3. The Wong Tai Sin District mainly has public properties and redevelopment areas

In term of property price, want to compare the property price different of the selected districts. We only pick one year of property transaction to illustrate this different. The following district graph shows the logarithmic transformed net per square feet price of the property in year 2017. It shows each district has both property transaction of relatively high and low prices except that the "Kwai Tsing" districts has a bit lower transaction prices in general. Moreover, the properties located close to the harbor have bit higher per net square feet price in general. Besides location, it can been that the transaction price of a property depends on other factors as well.



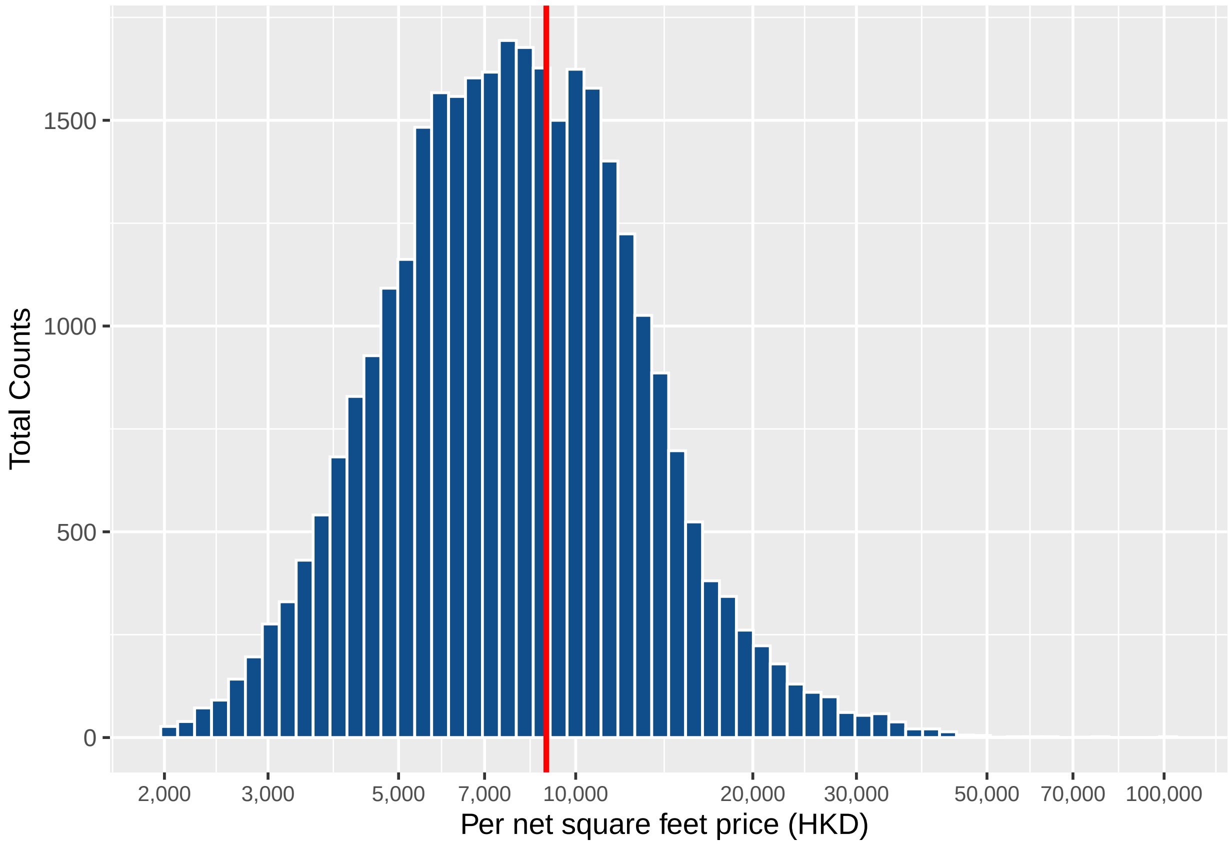
##Correlation

We try to study the correlation of the numerical variables in the selected districts of the Hong Kong property transaction dataset. The following table shows the correlation of the per net square feet price of property with the numerical variables. It can be seen that all of the numerical variables have positive correlations with the transaction price. The numerical factor which affect the property price most is the year of the property transaction.



4.2 Net Square Feet Price

The per net square feet price is the prediction target in this project. The following diagram shows the histogram of the per net square feet price of the property transactions in the selected districts in Hong Kong. Since the pricing data is ranged from 2004 to 9.9913×10^4, we have used the logarithmic scale, log10, in the x-axis to show the distribution of the 10 years property transaction prices. With the logarithmic transformation, the distribution has a “close to normal” distribution shape. The averaged transaction price is 8914.12 Hong Kong dollar.

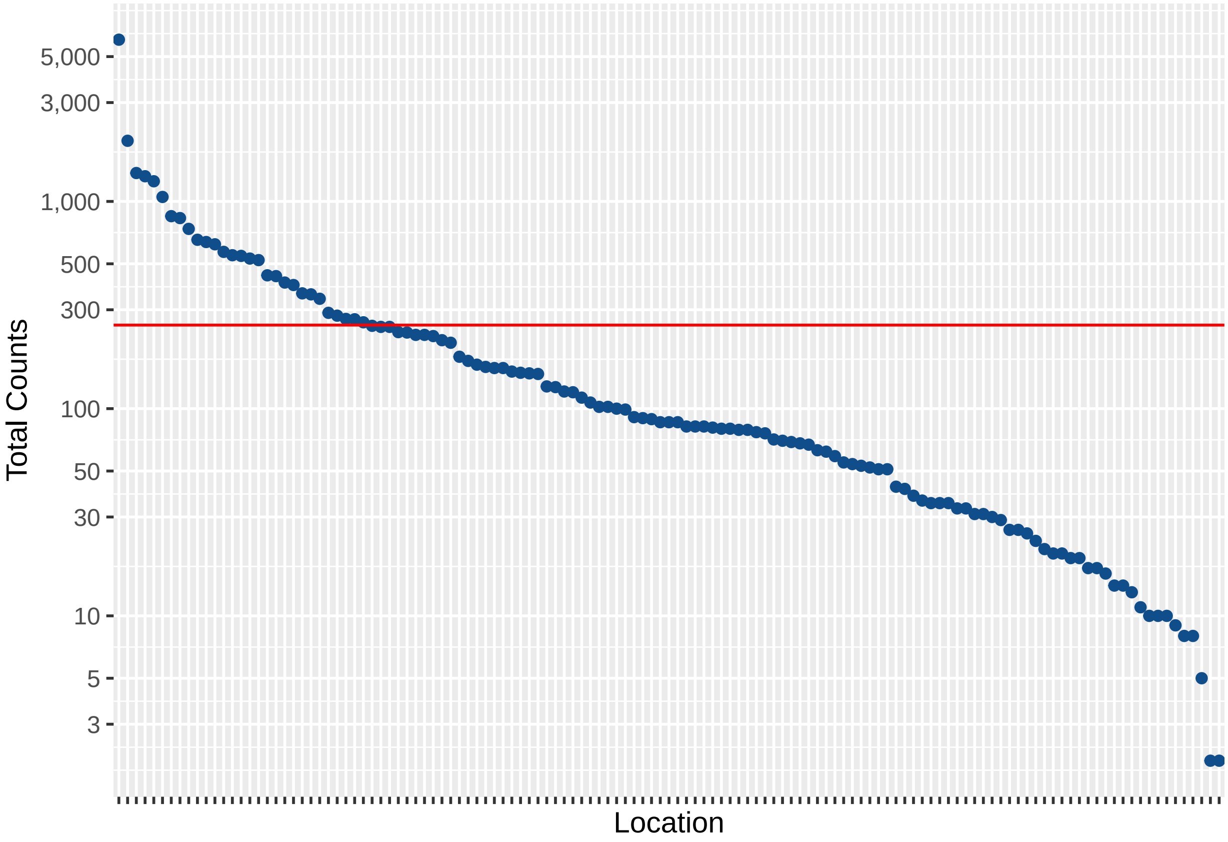


4.3 Location

The "Location" variable in the selected Hong Kong Property Transaction dataset describes the location of a property in a given district. There are in total 127 locations in the 5 selected districts. The following shows the list of these locations. In fact, it can be found from the names of these location that they are the names of residential estates and/or gardens in Hong Kong. This "location" factor actually contains part of the information about the "estate" factor which we have removed in the beginning of our studies.

```
##   [1] Taikoo Shing                     The Avenue
##   [3] The Legend At Jardine's Lookout Wah To Building
##   [5] Park Towers                      Yee On Building
##   [7] Centre Place                     Residence Bel-Air
##   [9] Redhill Peninsula                Metro Harbour View
##  [11] Cosmopolitan Estate              The Cullinan
##  [13] The Arch                         Imperial Cullinan
##  [15] Lime Stardom                     Fu Tor Loy Sun Chuen
##  [17] Shining Heights                  The Masterpiece
##  [19] Bijou Apartments                 I-Home
##  [21] Park Ivy                         Upper West
##  [23] 26 Kimberley Road                Bedford 28
##  [25] Parkes Residence                 Carmen's Garden
##  [27] City 18                          Nelson Court
##  [29] The Lodge                 Mei Foo Sun Chuen
##  [31] The Pacifica              Lai Yan Court
##  [33] High One                  Gardenia
##  [35] One New York              Beacon Heights
##  [37] The Prominence            Maison Rose
##  [39] High Point                Dynasty Heights
##  [41] High One Grand            Beacon Lodge
##  [43] Heya Delight              Court Regence
##  [45] Emerald 28                Kent Place
##  [47] Greenwood Villas          Southern Lodge
##  [49] Whampoa Garden            Laguna Verde
##  [51] Tung Tau (Ii) Estate      Le Billionnaire
##  [53] Kellet Court              Tin Ma Court
##  [55] Lions Rise                The Latitude
##  [57] Fung Tak Estate           Lower Wong Tai Sin (1) Estate
##  [59] The Forest Hills          Tsz On Court
##  [61] Kai Tak Garden            Choi Fung Court
##  [63] Ying Fuk Court            Hong Wah Court
##  [65] Hong Pak Court            Shun Chi Court
##  [67] Lei On Court              Hong Nga Court
##  [69] Tsui Ping (North) Estate  Choi Ha Estate
##  [71] No.8 Clear Water Bay Road Park Metropolitan
##  [73] The Spectacle             Canaryside
##  [75] Ocean One                 Chun Wah Court
##  [77] Lohas Park                Metro City
##  [79] Park Central              The Wings
##  [81] Festival City             Double Cove
##  [83] Shatin Centre             Shatin New Town (Lucky Plaza)
##  [85] Pictorial Garden          Oceanaire
##  [87] Marbella                  Peak One
##  [89] World-Wide Gardens        Serenity Park
##  [91] Mont Vert                 Tai Po Plaza
##  [93] Parc Versailles           Mayfair By The Sea I
##  [95] Cheung Wah Estate         Sereno Verde
##  [97] Riva                      Meadowlands
##  [99] Imperial Villas           Tai Hing Gardens
## [101] Century Gateway           Parkland Villas
## [103] Chelsea Heights           Hong Kong Gold Coast
## [105] Siu Shan Court            The Hillgrove
## [107] Park Island               Riviera Gardens
## [109] Belvedere Garden          Hongkong Garden
## [111] Chelsea Court             Ching Wah Court
## [113] Wonderland Villas         Cheung On Estate
## [115] Ching Tai Court           Yuet Lai Court
## [117] Tsing Yi Estate           Tsui Yiu Court
## [119] Kwai Chun Court           Yin Lai Court
## [121] Wah Yuen Chuen            Cheung Fat Estate
## [123] Kwai Yin Court            Ching Wang Court
## [125] Hang King Garden                    Discovery Bay
## [127] Caribbean Coast
## 127 Levels: 26 Kimberley Road Beacon Heights Beacon Lodge ... Yuet
   Lai Court
```

The following plot shows the total number of transactions in each location which sorted in descending order of transaction number.



There are significant differences among of transactions in property locations, which ranged from 2 to 6034. When we take a closer look into the top 10 and bottom 10 number of transactions in the 10 years period of the selected districts in Hong Kong, it can be seen that the bottom 10 locations had less than 11 transactions whereas the mostly traded property location had 6034 transaction in 10 years period. Therefore, the needs of regularization for this categorical variable may need to be considered.

Table 1: The Top 10 Locations with the most transaction

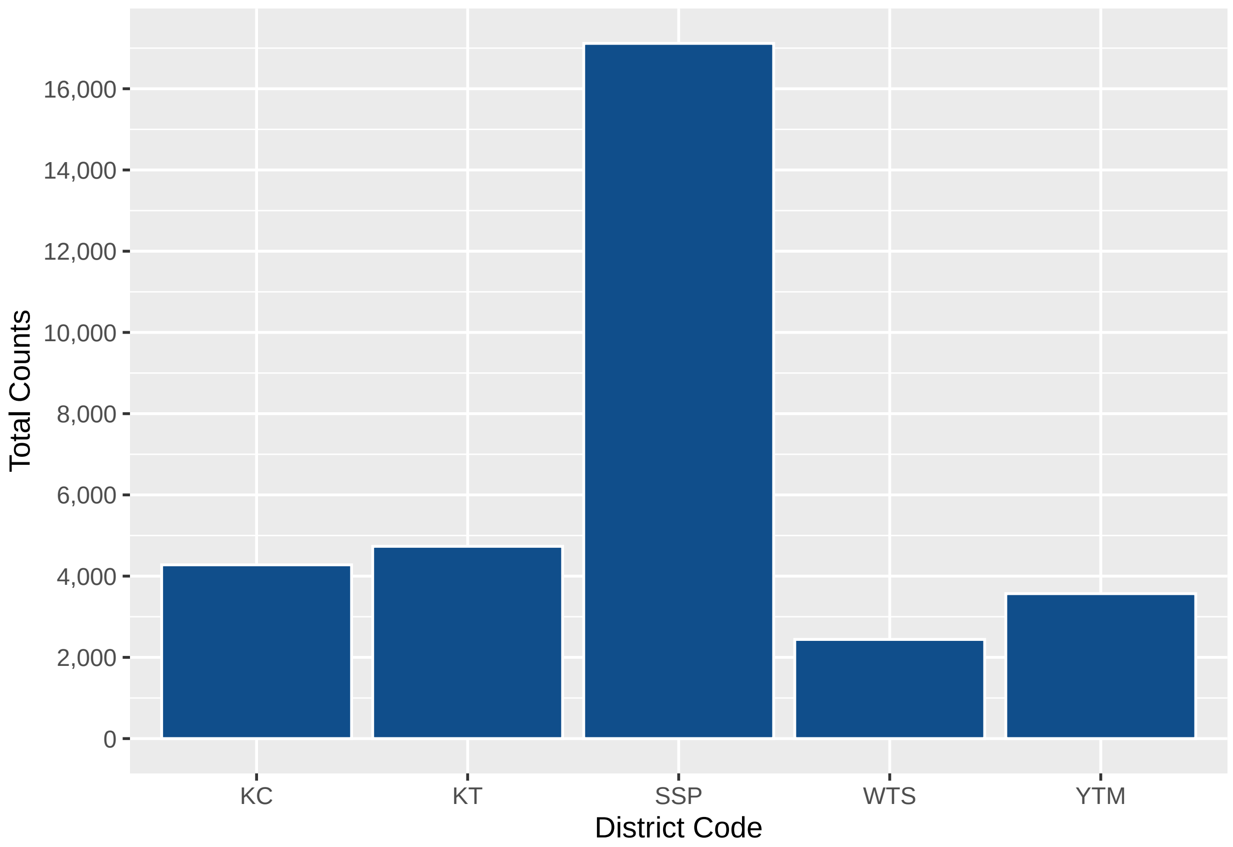
location	Count
Mei Foo Sun Chuen	6034
Park Island	1962
Caribbean Coast	1371
Whampoa Garden	1323
Laguna Verde	1251
Metro City	1051
Chelsea Court	849
Hong Wah Court	831
Park Central	737
The Arch	653

Table 2: The Bottom 10 Locations with the least transaction

location	Count
Heya Delight	11
High One Grand	10
Ocean One	10
Redhill Peninsula	10
Parc Versailles	9
Mayfair By The Sea I	8
World-Wide Gardens	8
Parkes Residence	5
Discovery Bay	2
The Lodge	2

4.4 District

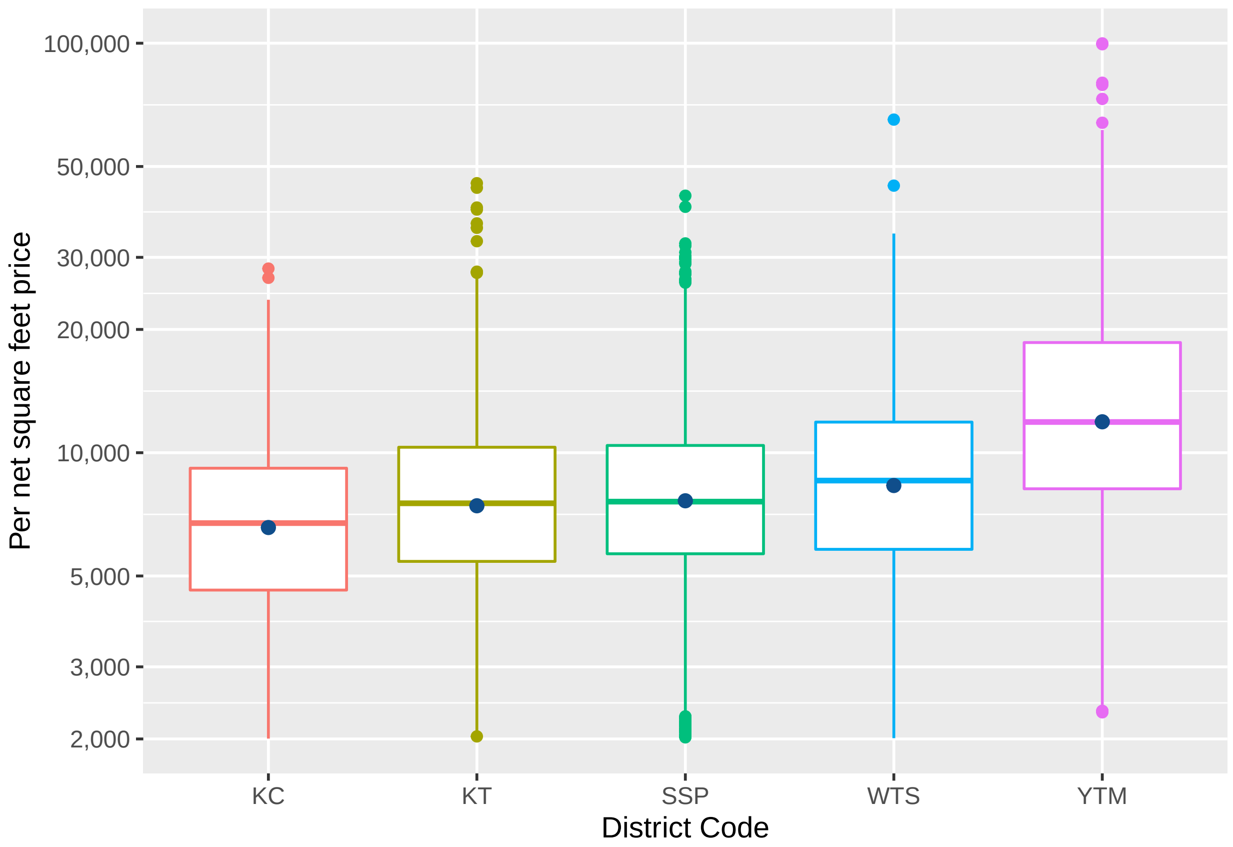
The “District” variable is another categorical factor in the selected Hong Kong property transaction dataset. The following diagram shows the number of transactions in each of the 5 selected districts in the period of 10 years.



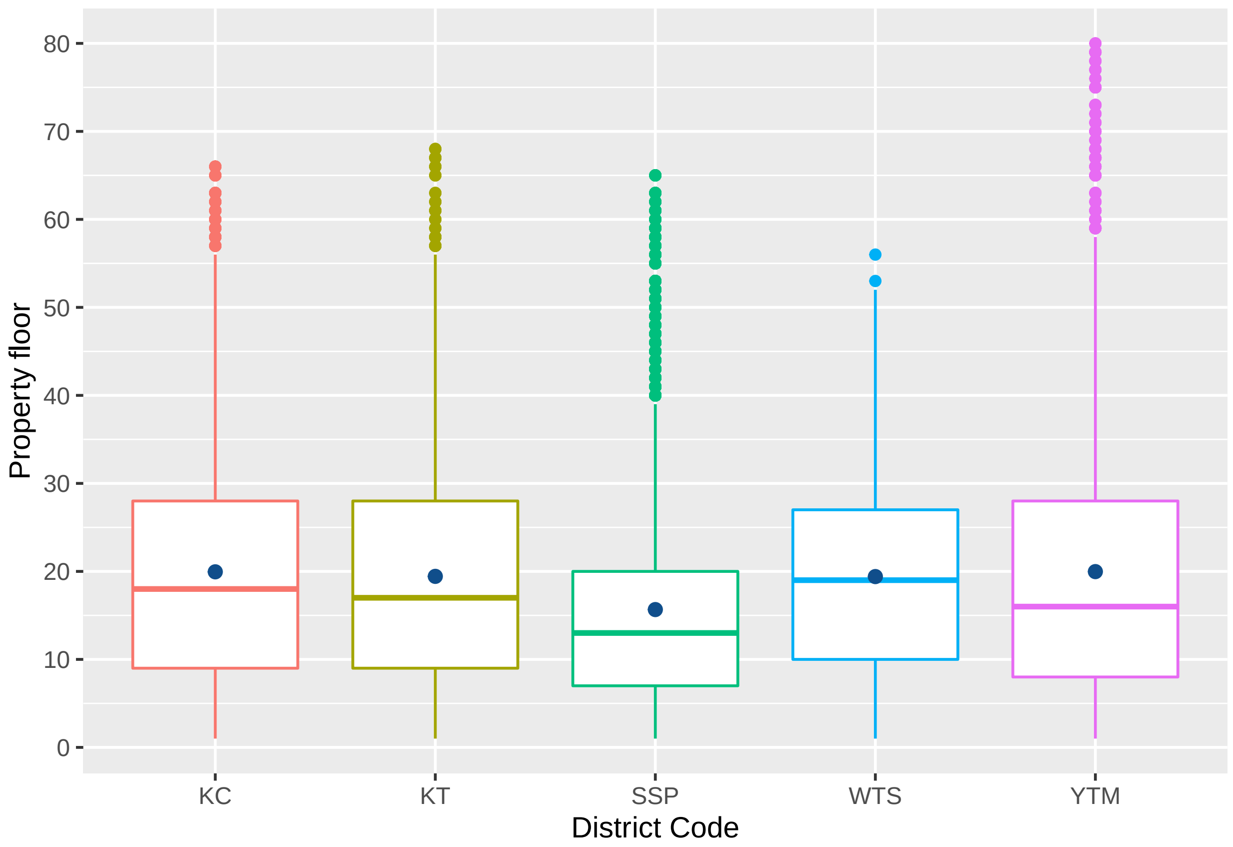
In the diagram above, the transaction count is in logarithmic scale. It can be seen that the "SSP" district has much more transactions compared to the rest of districts. This maybe due to the fact that the dataset creator accidentally combined the property transactions of SSP ("Sham Shui Po") district with another district "Kwai Tsing" as explained in the previous section. Another possbility is that the dataset creator, Midland property agency, has more retail outlets in SSP and Kwai Tsing districts so that Midland could obtain more property transaction there.

The following diagram shows the box plot of the per net square feet price of a property in each selected district.

In general, it can be identified that the KC (Kowloon City) district has the lowerest averaged property price. Where as the YTM (Yau Tsim Mong) district has the highest averaged and absolute property pricings. For KC district, it can be explained that there were many old buildings and buildings waiting for redevelopment. Therefore, the averaged property price is lower compared to other districts. For YTM district, there are new property development projects in the western part of this district and the properties were located closer to the habour area. Therefore, the properties there were usually with high values.

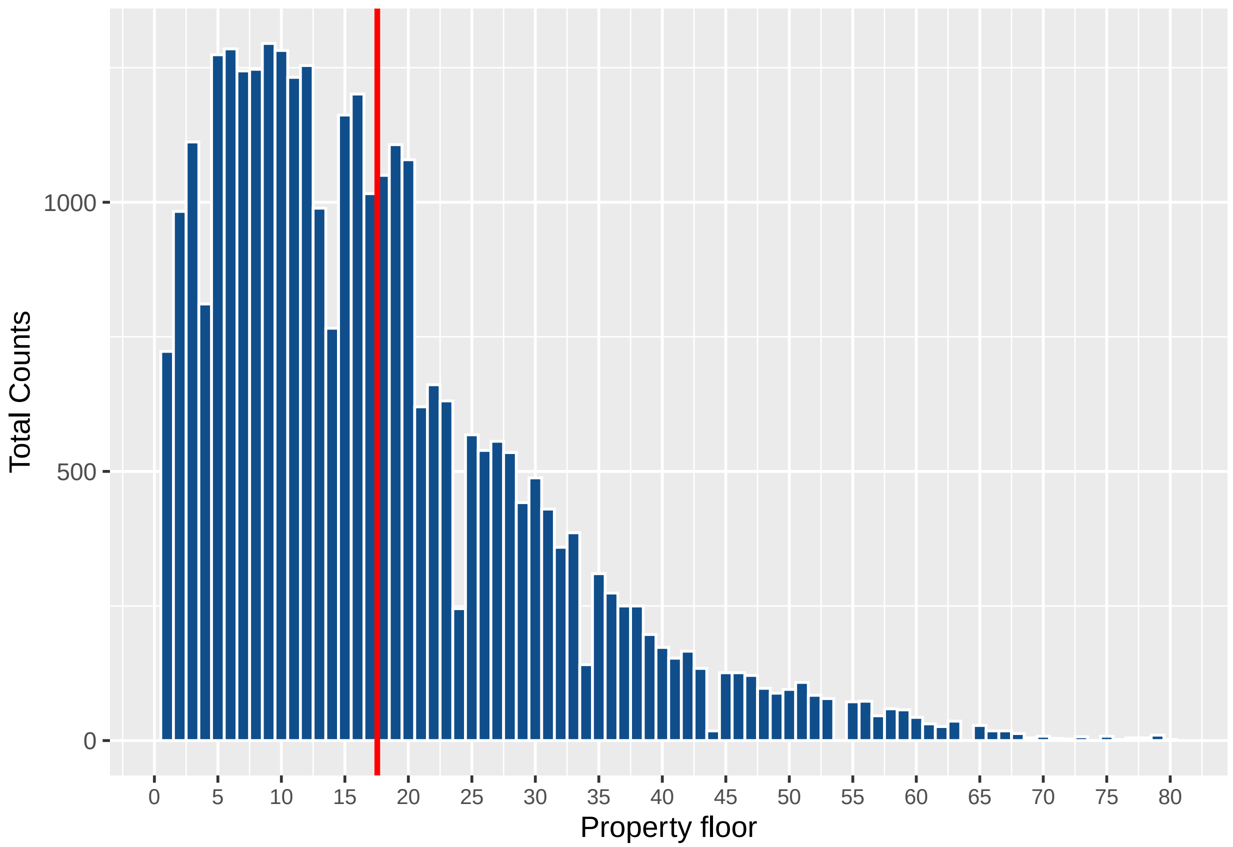


When we also look into the number of floor of building transactions in each district, we can find the averaged floor level of the transactions were around 15 to 20 floor in the selected districts. However, there are more high level/floor transactions happened in YTM district. This information also reflects that there are some very high raising buildings in YTM district in which properties usually priced higher. The following diagram shows the boxplot of the floor level of a property transaction versus the selected districts.



4.5 Property Floor

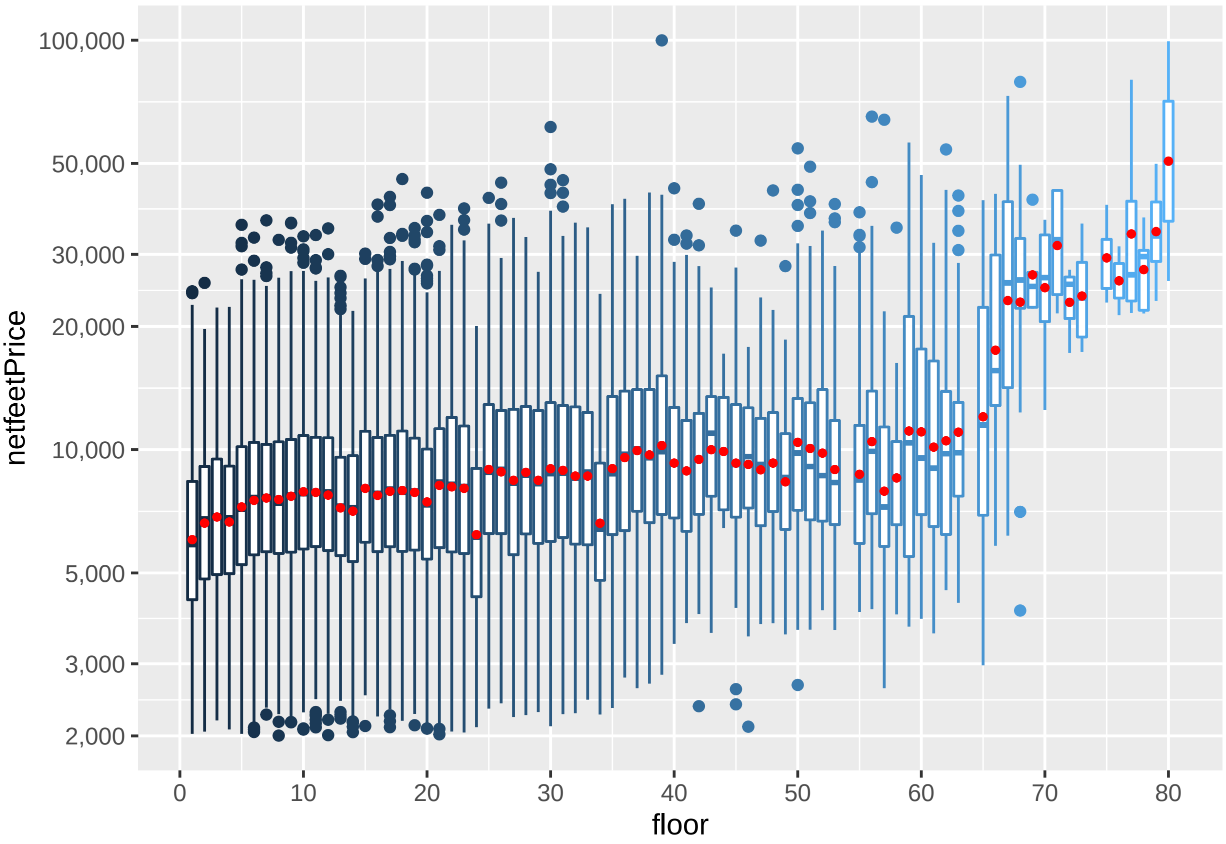
The "floor" variable states the floor level of a property transaction. The following diagram shows the total number of a property transaction located in the same floor/level of different buildings in the selected districts and time period. The distribution of the property floor skewed towards the left. The red vertical line shows the averaged level of a property located. In this dataset, the averaged floor level is 17.55. However, there is residential building in Hong Kong with property at the level of 80.



One “interesting observation” can be obtained from the property floor histogram above. It can be seen that the number of property transactions at the level/floor of 4, 14, 24, 34, 45, 54... were much less compared to the adjoin floors. This is due to the culture issue in Hong Kong. In Hong Kong, the pronunciation of ‘4’ in Cantanese is similar to that of ‘die’ [8]. Therefore, people in general do not want to buy property at the floor with the ending of ‘4’. Moreover, there were quite some property estates and gardens which intentionally skip the level of the ending of ‘4’. Therefore, the transaction counts of property with the ending ‘4’ in floor are much less and there are no floor with ending number ‘4’ after 50 floors.

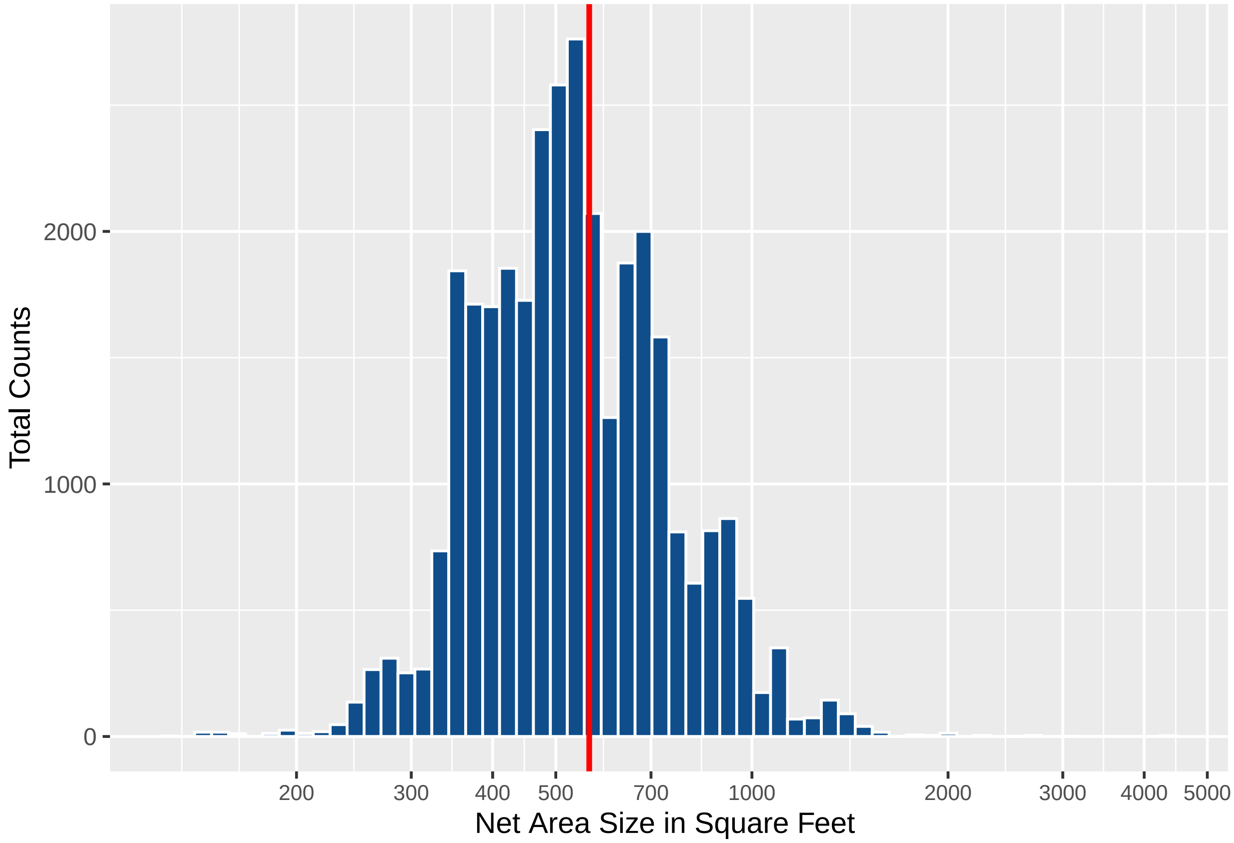
In term of per net square feet price, the following diagram show the boxplot of the floor number of the properties against their transaction price. The red dots show the averaged transaction price in each floor. As expected, the higher the floor of the property located, the higher the per net square feet price of this property will be. In general, the property enjoys much better and open view if it is located at the higher level of a residential building.

Similar “interesting observation” can also be found in the property price versus floor boxplot, the property floors with the ending number ‘4’ have lower transaction prices compared to the prices of the adjoin floors.

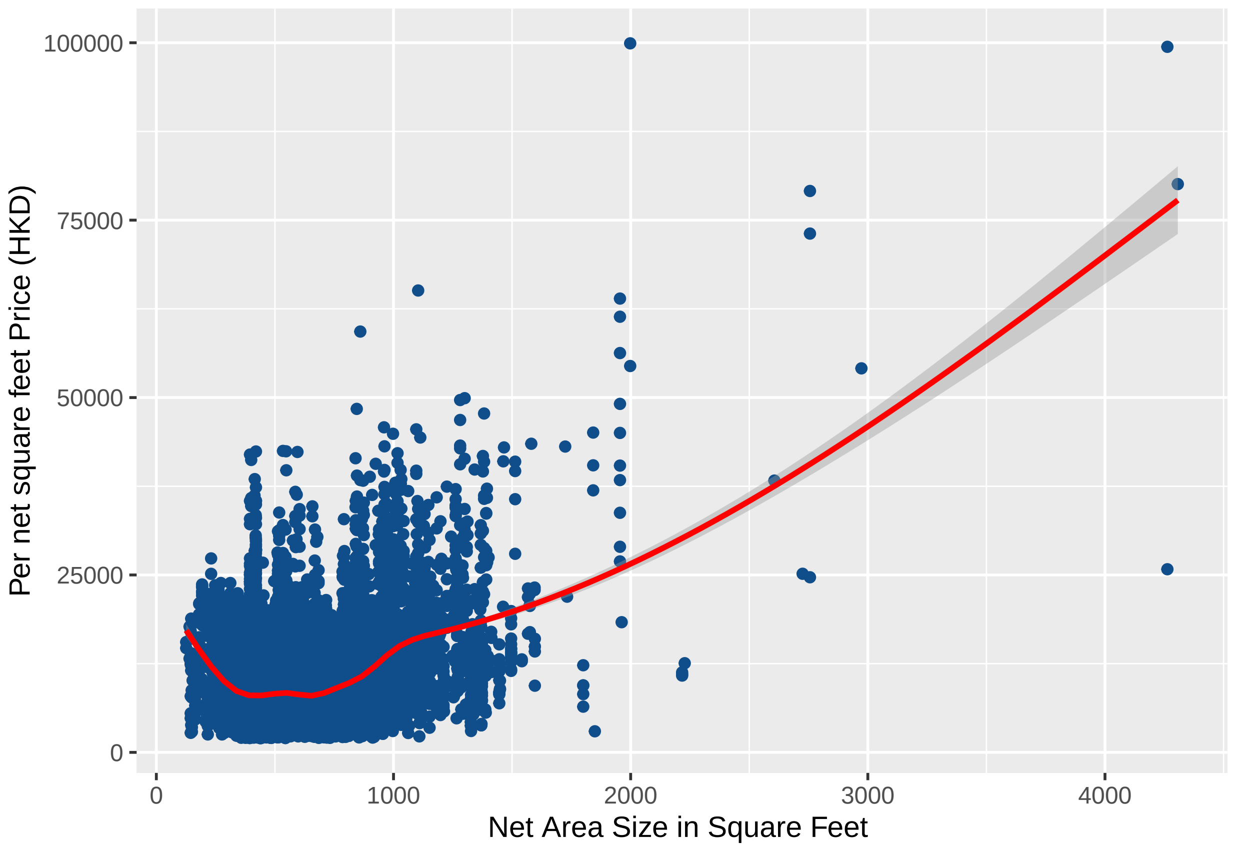


4.6 Net Area Size

The "Net Area Size" records the net total area in square feet of a property in the transaction record. The following diagram shows the histogram of property area size in the Hong Kong property transaction dataset. As there are less sizable properties compared to the majority, we logarithmically transformed the net area size (x axis) to obtain a better visualization. The red vertical line shows the averaged total area size, 562.69 sq feet, of the Hong Kong property.

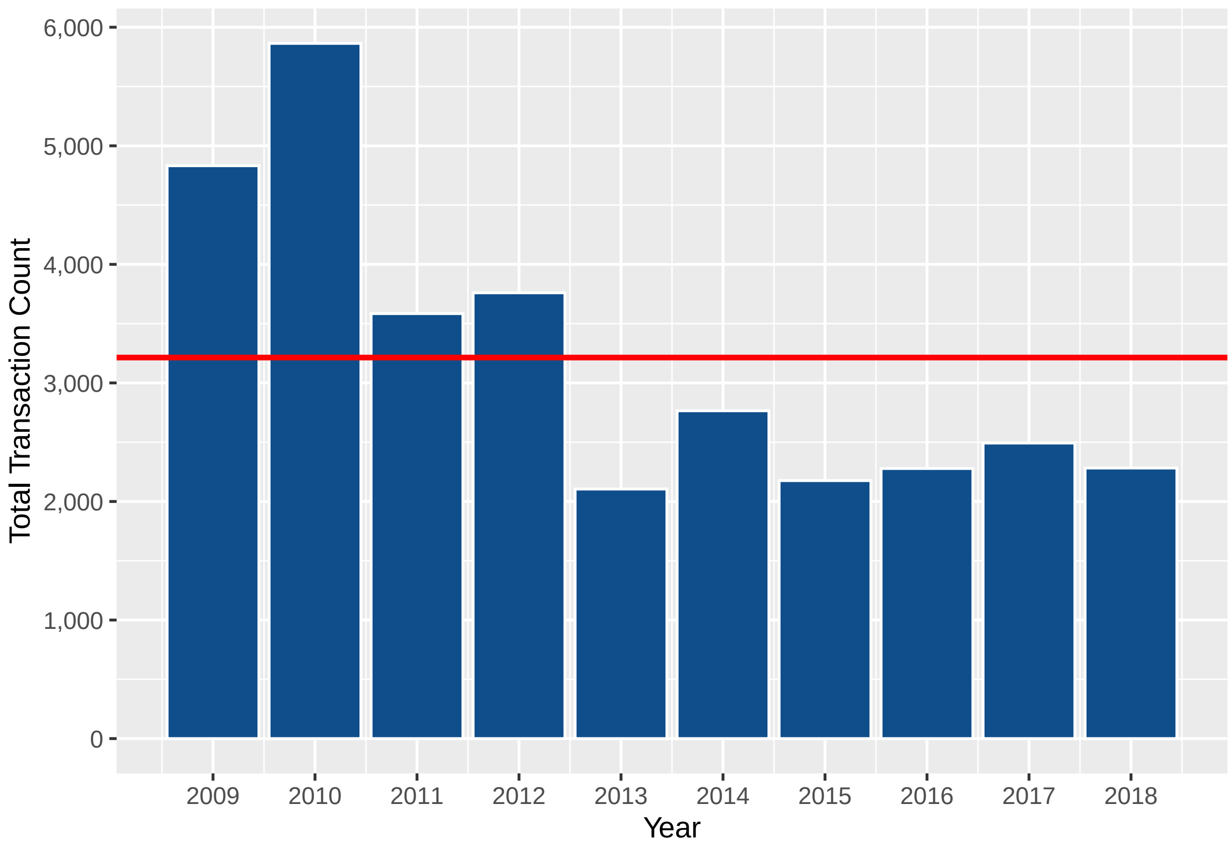


The following diagram shows the dot plot of the per net square price against the net total area size of a property. The red line shows a general trend, and relationship, between the property price and the net toal area size. In general, it can be observed that the area size of the properties in Hong Kong are smaller than 1,500 square feet with net per square price less than 37,500 Hong Kong dollar. Moreover, the bigger the total area size of a property is, the higher the net per square feet price of a property will be.



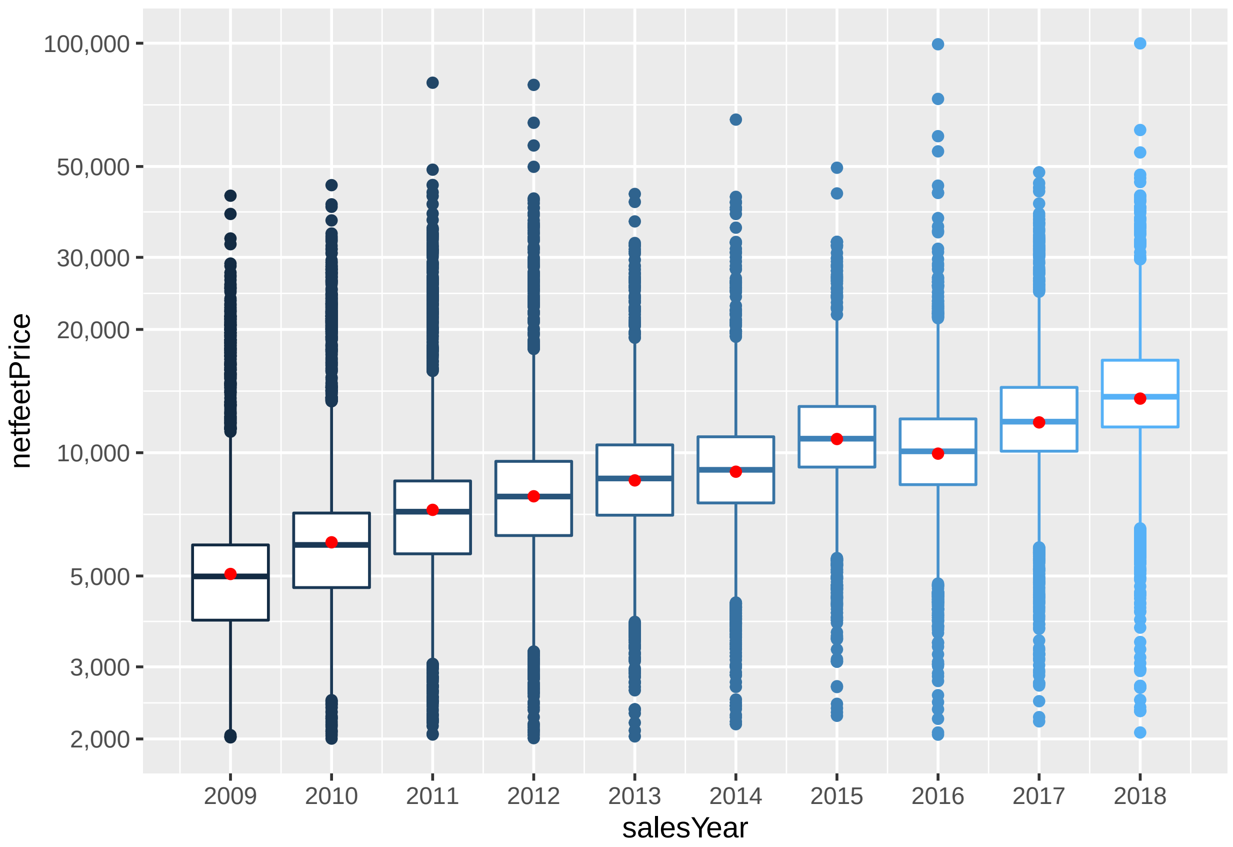
4.7 Year of Transaction

The "Year of sales" variable contains the information about the year of a property transaction happened. In general, the number of transaction in the period of 2009 to 2018 ranges from 2105 to 5864 with averaged number of transcation 3214.00.



The diagram above shows the total property transaction in each year in the period of 2009 to 2018. In general, the number of transaction decreases from 2009/2010 to 2013 and then keep a relatively stable range between 2,000 to 2,500 transaction afterwards. The reason behind this decline is due to the introduction of new property stamp duty rate by the Hong Kong Government started from 2010 in Hong Kong [9] and following by other new duties in the following years [10].

The following diagram shows the general property transaction price between 2009 and 2018 in the 5 selected districts. The red dots show the averaged per net square price of the property transaction. Even there was a decrease in the number of property transaction started from 2010, the averaged per net square feet prices are still increasing with time and the it has not been affected much by the introduction of new duties [9] [10].


netfeetPrice
100,000
50,000
30,000
20,000
10,000
5,000
3,000
2,000
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
salesYear

5 Modeling Methods

This section describes the data preparation and the selection of algorithms for prediction model training. The details of the selected algorithms will also be explained in the following paragraphs.

5.1 Data Preparation

In order to perform model training and testing, the “Hong Kong Property transaction Midland unit 2018” dataset is separated into training dataset and testing dataset with the radio of 80% to 20%. The data samples in these two data subsets are randomly picked from the Hong Kong property transaction dataset. The training data subset consists of 25710 data samples and testing data subset consists of 6430. The training data subset will be used for model training and cross validation. The testing data susbset will only be used to test and performance of the created models.

5.2 Model Selection

In the wrangled “Hong Kong Property transaction Midland until 2018” dataset, it consists of 4 numerical and 2 categorical variables. The 2 categorical variables have been factorized as different levels (“district” has 5 levels and “location” has 127 levels). We are going to use the regression based methods to build property price prediction models. This means 3 numerical plus 2 categorical variables will be used to predict the property transaction price. The following 4 regression based algorithms will be used for model developments. They are “Generalized Linear Model”, “K-Nearest Neighbors”, “Random Forest” and “Extreme Gradient Boosting”. Besides, the averaged per net square feet price, μ of the training set will also be used as the branchmark to predict a property transaction price Y with the following equation:

$$Y = \mu + \epsilon$$

where ϵ is the error term

The value μ of the training dataset is 8918.85 Hong Kong Dollar.

Method	Model_RMSE	Model_MAE	Model_Rsquare	Train_Time
Simple Average	4957.17	3457.17	NA	0

5.2.1 Generalized Linear Model

In the first model, the Generalized Linear Model will be used to predict the property transaction pricing. This is the simplest model with advantage of fast training speed. We will use the training dataset to fit a generalized linear model to predict the property price Y with the following equation:

$$Y = \beta_0 + \beta_l x_l + \beta_d x_d + \beta_f x_f + \beta_a x_a + \beta_y x_y + \epsilon$$

In the above equation:

$beta_0$ is the mean, μ, of the per square feet property price x_l is the location of the property for price prediction x_d is the district of the property for price prediction x_f is the level of the property for price prediction x_a is the net area size of the property for price prediction x_u is the year of the property transaction for price prediction ϵ is the error term

The performance and training speed of the generalized linear model will be compared with other machine learning model in our studies.

5.2.2 K-Nearest Neighbors

As we learned from the machine learning course, K-Nearest Neighbors (K-NN) is a type of machine learning algorithms. The principle behind K-NN algorithm is to find the K predefined number of training samples that are the closet in the Euclidean distance to a new point and predict the label for the new using these samples. The formula for Euclidean distance between the new point and training data point is listed below:

$$D(X_f, X) = \sqrt{\sum_{i=1}^{n}(X_{f,i} - X_i)^2}$$

where, X_f is the feature vector of the new data point and X is the feature vector of a single training data point.

K-NN has been used as one of the regression for both numerical and categorical result prediction. It has also been used for house price predictions as well (e.g. [11] and [12]). So, we used K-NN in this project for our property price prediction algorithm. In our dataset, there are 5 factors to predict the property price. Therefore, the number of features, n, is equal to 5.

In this project, we will also find the number of neighborhood points, the value of "k", which optimizes the prediction performance of the property price model.

5.2.3 Random Forest

Random forest is a decision tree based machine learning algorithm. It contains many decision trees representing a distinct instance of the classification of data input into the random forest. The random forest algorithm considers the instances individually and it takes the one with the majority of votes as the selected prediction.[13]

The samples in the dataset will be taken as input to each tree in the classification. Then, the features will be randomly selected and being used in growing the tree for each node. As every tree would only be pruned until the end of the exercise when the prediction decision is reached, the random forest method enables classifiers with weak correlation to generate stong classifiers.

In our project, Random Forest will be used for property price prediction. In the Hong Kong property dataset, the 'location' variable has 127 levels. To speed up the model training time, A fast implementation of random forest for high dimensional data called "ranger" will be used [14]. A tune length of 10 will be used to find the optimal model with different "mtry" value and "splitrule" type while keeping the default value, 5, of "min.node.size".

5.2.4 Extreme Gradient Boosting

Since the introduction in 2014 [15], extreme Gradient Boosting (XGBoost) is one of the most popular and powerful machine learning algorithms. XGBoost has been used in many machine learning competitions. As the core of XGBoost algorithm can be run in parallel, computer with multicore CPU or computer cluster can be used to significantly speed up the training process. Meanwhile, this model is also used for property price prediction and valuation [16]. Moreover, as explained in the previous section, regularization would be required to handle the out liners. XGBoost has this capability by default to regularize the data.

XGBoost is a decision tree-based algorithm sitting under the supervised branch of Machine Learning. It can be used for both classification and regression problems. The algorithm uses the 'Similarity score' and 'Gain' to determine the best node splits to build the trees. Besides in [15], [17] provides a brief introduction of this algorithm.

In this project, we mainly try to use this famous tree based algorithm to compare its training speed and performance with other machine learning algorithms. In this project, the default parameters of XGBoost will be used .

5.3 Cross-Validation

As introduced in the Harvardx Data Science program and textbook [6], we will use cross validation method to improve the performance of the regression models and to reduce the over fitting issue. In this project, 5-fold cross validation will be used to randomly split the training dataset equally into 5 groups. The 4 randomly selected groups will be used for model training and the remaining will be used for model validation. To reduce the processing time, the cross validation process will only repeat 3 times.

5.4 Machine Used

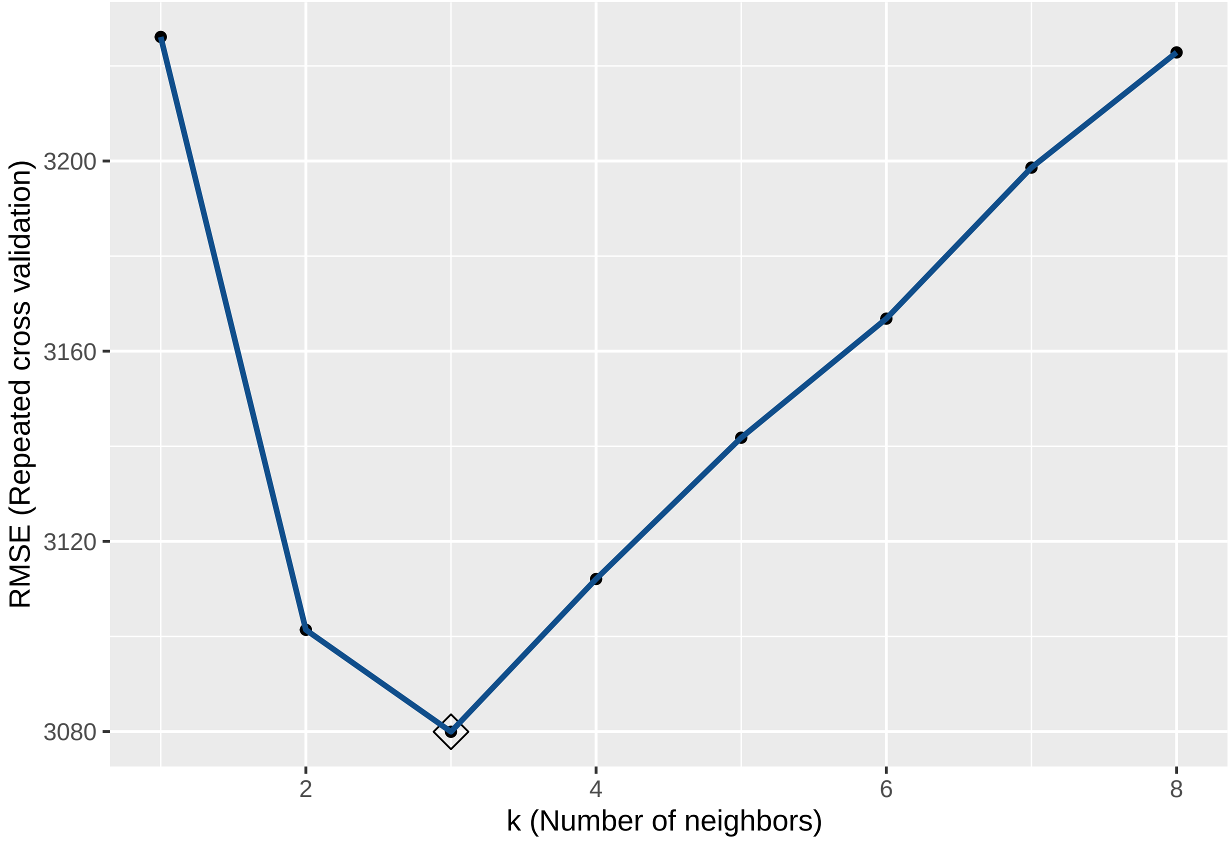
A personal computer with 32GB DDR RAM, an Intel Core i9-9900K CPU running at 3.6GHz, a nVidia RTX2080 super graphic card with a Microsoft Window 10 Operation System has been used in this project for model training and prediction.

6 Result Presentation

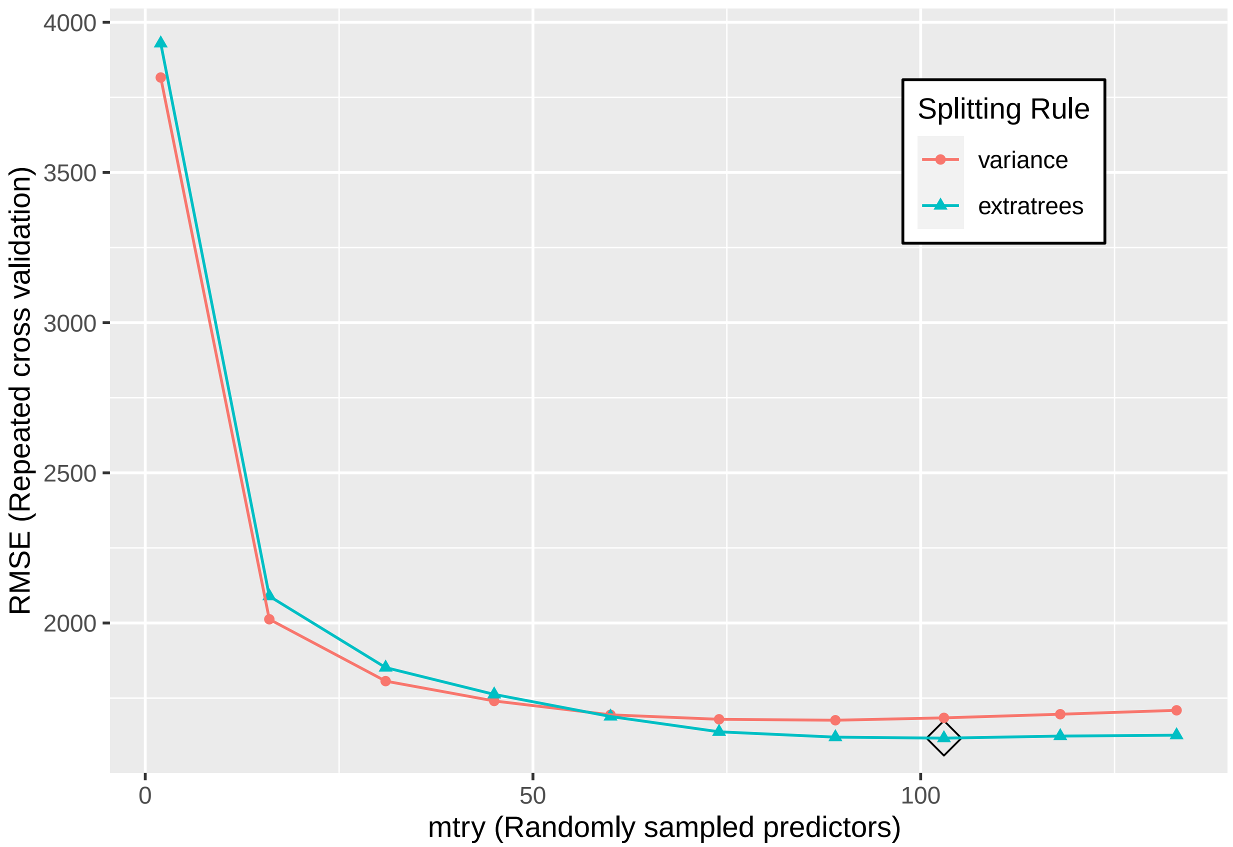
6.1 Model Tuning

During the model training, the performances of the K-NN and Random Fortress models were fine tuned.

For the K-NN model, the optimal number of neighbor, “k”, being used in the regression is 3. The following diagram show the plot of different k value being tested to identify the best tuning value.



For the Random Forest model, “ranger”, the best number of features (randomly sampled predictors) to split at each node, “mtry”, is found to be 103. The following diagram shows the plot of the RMSE values against teh 10 “mtry” values of the random forest prediction model.



For the GLM model, tuning is not applicable. Whereas for the Extreme Gradient Boosting, we use default setting for model training and performance evaluation.

6.2 Model Performances

The following table shows the summary of the performance, in terms of RMSE, Mean Absolute Error (MAE) and R Square (Rsquare) values, as well as the training time of each property price models created by the 4 selected prediction regression algorithms.

Method	Model_RMSE	Model_MAE	Model_Rsquare	Train_Time
Simple Average	4957.17	3457.170	NA	0
GLM	1955.18	1109.827	0.846386	16.8157 secs
KNN	3079.95	1806.934	0.621822	40.9241 mins
Random Forest	1616.63	875.563	0.895393	28.9864 mins
XGBoost	1598.91	877.705	0.897033	11.4282 mins

From the above training result, it shows that the GLM model delivers satisfactory performance with relatively low RMSE and RAM, high R square value and the shortest training time.

For K-NN algorithm, the model performance is actually the worst compared to the others and it also take a very long time for model training. The model performance may be affected by the imbalanced data (e.g. more transaction in SSP district compared to the others in “district” factor) and out liners (There are properties will very high per net square price). For the processing time’s issue, K-NN is computationally intensive and its training speed decrease significantly with grow dataset size.

The random forest model has a relatively high performance compared to the previous 2 models. It yields a R square value close to 0.9 and with averaged RMSE of predicted result of 1616.63

In the 4 prediction models, the Extreme Gradient Boosting (XGBoost) gets the best overall performance with the lowest RMSE and the highest R square value. Its performance is slightly better then that of random forest model. However, the XGBoost takes much short time for model training.

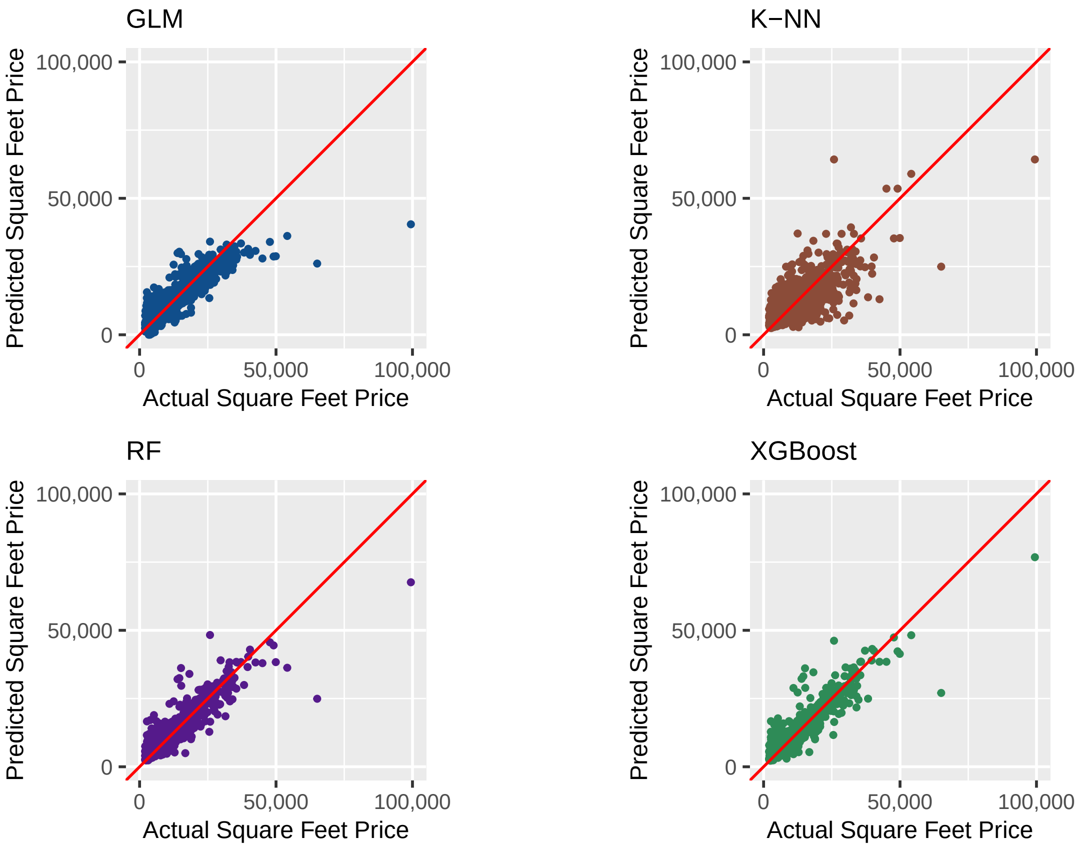
6.3 Test Data Results

The following show the prediction performance of the 4 regression models with testing data subset as input.

Method	TEST_RMSE	TEST_MAE	TEST_Rsquare	PRED_Time
GLM	1997.75	1118.383	0.837601	0.212566 secs
K-NN	2930.36	1723.471	0.656002	35.8504 secs
RF	1680.10	877.506	0.885172	0.554751 secs
XBGBoost	1639.09	878.057	0.837601	0.0183661 secs

The RMSE, MAE and the R square results of the 4 regression models are similar to those values obtained from the model training with cross validation. In term of prediction speed, however, it can be find that the XGBoost perform much better than the others. As stated in the previous section, the coding design of XGBoost takes average of parallel and multi-core CPU processing. Therefore, it can perform prediction much faster than the others.

The following plots show the prediction results versus the actual transaction price of the 4 prediction models.



It can be seen that the predicted property price using the K-NN model has the largest deviation from the actual property price. When the variations actually increase with the value of property price. For the plots of GLM model, the predicted property price has less deviation from the actual property price compared to those in K-NN result. However, it seems that the GLM tends to under estimate the property prices with larger transaction values.

For the plots of RF and XGBoost, they have more or less the same deviation and they both preform better than that of the GLM model. But visually speaking, the prediction performance of XGBoost looks better for the properties with high per net square feet price. The prediction results of properties with high per net square prices from XGBoost model seem to have less deviation from the actual pricing.

7 Conclusion

In this project, our objective is to development property price prediction models for the downtown area "Kowloon" of Hong Kong using the property transaction data from 2009 to 2018. The data samples from the 5 districts (plus 1 more district being combined in SSP district by the dataset creator) in the Kowloon downtown area have been used for model training and testing. Multiple machine learning algorithms, including generalize linear model (GLM), K nearest neigbhors (k-NN), Random Forest, and Extreme Gradient Boosting (XGBoost) have been used to examine the speed and performance of the property price models being developed based on these algorithms. From our studies in this project, we identified that the prediction model being developed by XGBoost algorithm gives a good prediction result while it has much faster prediction time comparing to other algorithm when using in a computer with multi-core CPU.

Further work can be performed with the current property price to improve the prediction performance. For example, data normalization can be performed to normalize numerical predictors (e.g. "floor" and "per net square feet price") with different ranges in the HK property transaction dataset [18] [19]. Secondly, we have only use the default XGBoost algorithm in this project. We can perform a further study to improve the performance of this model by tuning available parameters in this algorithm.

Owing to the limitation of time and processing power, we only studied and processed part of the available samples and variable types of the Hong Kong property transaction dataset in this project. If time and processing power are not the issues, we can take into account of data samples from all the available time period and all districts in Hong Kong. We can also include more variable types (e.g. the "estate", "latitude" and "longitude"). Further studies will be required to (1) speed up the training process with the large volume data samples; (2) identify a right regression algorithm which can handle a categorical variable with hundreds of levels, and (3) the way to make use of the rest of available variables. For example, the method being proposed in [20] to sensibly use latitude and longitude data in model training.

8 Reference

[1] Alexrebull Natur, 2018, “Hong Kong Property Transaction Miland until 2018”, https://www.kaggle.com/datasets/alexrebull/hong-kong-property-transactions-midland-until2018

[2] Midland Realty, 2022, “Midland Realty Homepage” https://www.midland.com.hk

[3] The Park Island Blogger, 2019, “Hong Kong property prices - back on the way up in 2019”, http://parkislandhongkong.blogspot.com/2019/03/hong-kong-property-prices-back-on-way.html

[4] Christopher Caillavet, 2022 “Home on Hong Kong’s Peak Sells for $74.8M as City’s Luxury Market Rages”, https://www.mingtiandi.com/real-estate/finance/mount-nicholson-unit-sells-on-hong-kongs-peak

[5] Dr. Holger Frommert, 2022, “Geo Map of the Hong Kong Special Administrative Region of the People’s Republic of China”, http://www.geo-ref.net/ph/hkg.htm

[6] Rafael A. Irizarry, 2019, “Introduction to Data Science - Data Analysis and Prediction Algorithms with R”, http://rafalab.dfci.harvard.edu/dsbook

[7] Dana Paige Seidel, 2018, “Introduction to working with Spatial Data in R”, https://www.danaseidel.com/MovEco-R-Workshop/Materials/Day2/Spatial_Data_in_R/

[8] Rachel LeungOctober, 2019, “Hong Kong Numerology: Why Is Four So Unlucky?”, https://zolimacitymag.com/hong-kong-numerology-why-is-four-so-unlucky/

[9] The Government of the Hong Kong SAR, 2010, “Special stamp duty to curb speculation in residential properties”, https://www.info.gov.hk/gia/general/201011/19/P201011190294.htm

[10] https://www.okay.com/en/property-guides/an-overview-of-special-stamp-duty-buyers-stamp-duty-and-ad-valorem-stamp-duty/152

[11] Hilary Cordell, 2016, “An Overview of Special Stamp Duty, Buyer’s Stamp Duty and Ad Valorem Stamp Duty”, https://garba.org/posts/2022/knn/

[12] Dr. Kishor Keshav, 2021, “Designing an optimal KNN regression model for predicting house price with Boston Housing Dataset”, https://medium.com/mlearning-ai/designing-a-optimal-knn-regression-model-for-predicting-house-price-with-boston-housing-dataset-faef377536e3

[13] Wikipedia, 2022, “Random forest”, https://en.wikipedia.org/wiki/Random_forest

[14] Marvin N. Wright, Stefan Wager and Philipp Probst, 2022 ,“Package ‘ranger’ ”, https://cran.r-project.org/web/packages/ranger/ranger.pdf

[15] Tianqi Chen and Carlos Guestrin, 2016, “XGBoost: A Scalable Tree Boosting System”, https://arxiv.org/pdf/1603.02754.pdf

[16] Yun Zhao, Girija Chetty, Dat Tran, 2019, “Deep Learning with XGBoost for Real Estate Appraisal”, https://ieeexplore.ieee.org/document/9002790

[17] Arthur Mello, 2020, “XGBoost: theory and practice”, https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6

[18] https://towardsdatascience.com/boost-machine-learning-performance-by-30-with-normalization-and-standardization-156adfbf215b

[19] GreekDataGuy, 2020, “Boost Machine Learning Performance by 30% with Normalization and Standardization”, https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

[20] Zimo Wang, Yicheng Wang, Sensen Wu, 2022, "House Price Valuation Model Based on Geographically Neural Network Weighted Regression: The Case Study of Shenzhen, China", https://arxiv.org/abs/2202.04358