Vol. 8 No. 2 (2020)

Correlation Analysis of Factors Affecting the Prediction of Price of Terrace Houses in Penang, Malaysia: A Case Study

Norhayati Yahya, Norziha Megat Mohd Zainuddin, Nilam Nur Amir Sjarif, Nurulhuda Firdaus Mohd Azmi

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Semarak, 54100 Kuala Lumpur, Malaysia
yat.yahya@yahoo.com, , norziha.kl@utm.my, nilamnur@utm.my, huda@utm.my

Article history

Received:
20 Nov 2020

Received in revised form:
27 Nov 2020

Accepted:
5 Dec 2020

Published online:
27 Dec 2020

*Corresponding author
yat.yahya@yahoo.com

Abstract

Buying a house with a competitive price is difficult for many individuals due to various reasons. One of the reasons is the value of the house is influenced by many factors. Hence, it is crucial to study the relationship between house features and house prices for the prediction of the house price. Different house features may influence house prices either increasing or decreasing. This research aims to identify the house attributes that affect the prediction of the price of terrace houses in Penang, Malaysia using 2,699 sub-sale terrace houses actual property transactions from January 2018 to December 2019. This data is provided by the Valuation and Property Service Department of Penang. These attributes are identified through the correlation analysis conducted. Five attributes are highly correlated with house price; the size of the house ('Built_Up'), the price per square feet ('Price_Psf'), the number of floors ('Floors'), the number of rooms ('Rooms'), and the location ('Location_NE'). Size ('Built_Up') has the strongest relationship with price. Furthermore, this study has found that house price is not associated with its type. It has also revealed that the terrace houses located in South Seberang Perai and Central Seberang Perai are cheaper than the ones in other parts of Penang. The findings would assist buyers, investors, or homeowners in making decisions and in strategizing for purchasing or selling properties.

Keywords: *House price prediction, house attributes, correlation analysis, Pearson correlation coefficient*

1. Introduction

As highlighted by [1], most of the individuals who live in Asia have a major goal to become homeowners. This goal is driven by the attractive and positive financial return of the house, unlike other assets. For example, in Malaysia, the transaction volume from 1990 to 2007 of the residential property has contributed more than 60 percent of the country's growth, which makes this industry the leading contributor. For homeowners or investors, there are two types of prospective returns from buying houses; rental payment and capital gain from the increasing value in the property [2]. Moreover, investing in housing commonly seemed to be attractive

and beneficial for owners due to their equity value that does not reduce vigorously [3]. Thus, this created a decent prospect to increase the wealth for them.

In the housing market, the initial prices are an important factor in the process of buying and selling houses. However, determining the initial selling price of a house usually depends on the seller, nevertheless, determining the right price in the sales process will affect the buyer's wish to bid and make selections. [3], [4], and [5] highlighted that the initial price for houses is different according to residential facilities, home geographical conditions, and various house features. Hence, the buyer, investor, or homeowner requires to obtain careful and correct information before making the final decision.

According to [6], Malaysia's wealth has been generated from the real estate market. Locally, house prices have increased exponentially every year from 2010 until 2019 among these four types of houses namely terrace house, high-rise, detached house, and semi-detach house [7]. This indicates that over the past ten years, the real estate market in Malaysia has experienced a significant price expansion The National Property Information Centre (NAPIC), the Valuation and Property Services Department of Malaysia (VPSD), has collected the volume and value of property transactions in Malaysia from the year 2001 until 2018 as shown in Figure 1.1. Based on the figure, the highest sales volume is mostly high after the year 2010 and has achieved the highest in 2012. In 2018, the overall domestic property market grew at a slower rate of five point nine percent compared to the year 2017 with seven-point one percent [8]. Hence, this signifies the uncertainty in the Malaysian real estate market conditions, and it requires further investigation.

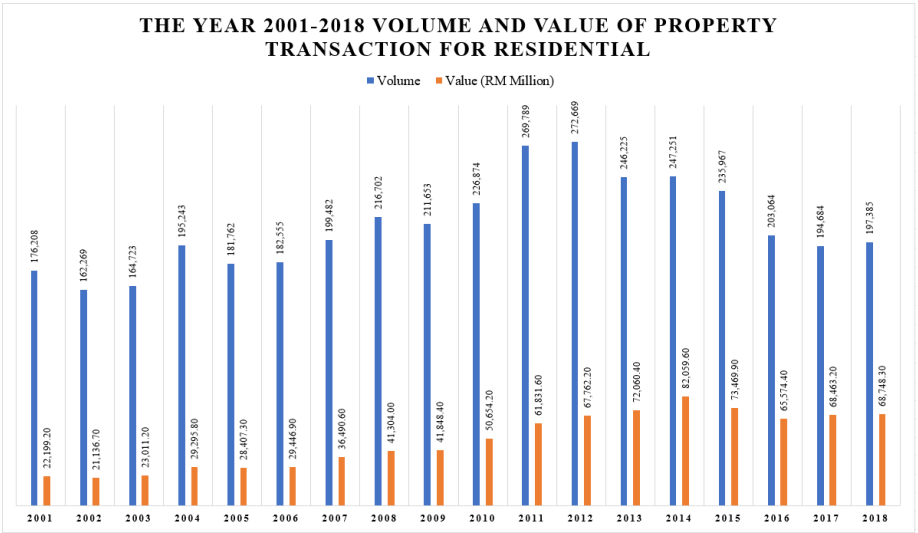


Figure 1.1 Volume and Value of property transaction for residential [7], 2001-2018

In a publication by Bank Negara Malaysia (BNM), "Risk Developments and Assessment of Financial Stability in 2018" [8], in Malaysia's real estate market, the overall property sales volume and values for the year 2018 grew at a slower rate compared to the year 2017, which signifies the uncertainty in the Malaysia real estate market. This uncertainty has raised questions among homeowners, investors, and buyers regarding what attributes or factors affecting the house prices and how

accurate the house prices can be predicted [5]. Besides, these questions are raised due to the investment in real estate is a long-term contract and required a huge sum of money. Therefore, homeowners, investors, and buyers require the most accurate information possible to assist them in making their decisions [9].

This research is to identify the house factors or attributes that affect the prediction of terrace house prices in Penang, Malaysia using the Valuation and Property Service Department (VPSD) dataset. Furthermore, the significance of this study can provide an idea in advising and assisting buyers to negotiate the price, especially for first-time buyers with relatively little experience, and advise purchasing strategies for buying properties. Hence, this article is divided into several sections; Section 2 presents the literature on factors that affecting house price prediction and provides a summary of five previous studies conducted in Malaysia. The methodology of the research is presented in Section 3. This will be followed by a discussion on the analysis of data and findings in Section 4, and finally, the conclusion is provided in Section 5.

2. Literature Review

As highlighted by [10], in the United States of America, the government has failed to produce data on house prices which cause 87 percent of house buyers there to rely on the internet on searching on factors relating house sales and house prices. For example, housing websites such as Trulia, Redfin, and Zillow provide an estimation of housing valuation based on the houses' features without charges [11]. Many buyers there resort to online search before contacting real estate agents.

In Malaysia, there are several popular property websites available to assist buyers in buying houses. Based on the portal ranking by traffic in the year 2018, there are four top property websites in Malaysia; PropertyGuru Malaysia, iProperty.com, PropSocial, and iBilik [12]. Generally, these websites are free. However, the houses listed on these websites are promoted by many real estate companies or agents, and the prices vary from one to another even for similar houses [9]. In contrast, the Brickz [13] is an independent property website which provides historical actual property transaction, collected from the Valuation and Property Service Department (VPSD) which officially records a property transaction once the stamp duty for the Sales and Purchase is paid in Malaysia. Moreover, Brickz has been compiling these officially recorded transactions since January 2014 and update the transacted data monthly. However, all the transactions recorded represent sub-sale transactions only [14].

In real estate, the property valuation is required to provide a quantitative measure of the benefits and liabilities of the ownership of the real estate. Valuations are required and often carried out, by several players in the real estate market such as real estate agents, appraisers, assessors, mortgage lenders, brokers, property developers, investors, and fund managers, lenders, market researchers, and analysts, as well as specialists and consultants. Market value is estimated through the application of valuation methods and procedures that reflect the nature of the

property and the circumstances under which the given property would most likely trade in the open market [15]. The appreciation of house values has created a decent opportunity for investors to purchase a property as a form of investment [2].

In a prior study by [4], in the real estate market, the initial price for each house and land are varied according to residential facilities and geographical conditions. Based on prior studies conducted, this research has found that the price of the house is influenced by several factors such as location (near to major highways, accessible by public transport), house features (building type, square feet, number of bedroom and bathroom), and neighborhood attributes (population density, nearest school, clinics, and shopping location) [5]. Supported by [3], in which the author has provided 40 features of the houses in studying the prediction of house prices. But the factors are only the area in square meter, location, year built, total bedroom, hall, kitchen, garage area, swimming pool area, and selling year. In the case of double-story sale transaction price prediction in Johor Bahru, Malaysia, the analysis has been carried out using two house features, which are the land area and main floor area [16]. Provided by [17], the prediction of house prices also affected by the land ownership type, size of land area, house type, and building qualities.

From a wider perspective, factors such as demographics, economy, and politics are associated with the housing price. Based on the economic point of view, the Malaysian house price movements are characterized by the regional demographics and regional finances, such as population, gross domestic product (GDP), housing finance, inflation rate, real property gains tax (RPGT), and cost of construction [18]. As emphasized by [19], other aspects such as interest rate, monetary liquidity, exchange rate, monetary policy are some of the factors that have little impact on the house prices in Malaysia.

Table 2.1 illustrates the comparison of five studies conducted on house price prediction in Malaysia. According to Table 2.1, several studies were conducted based on a dataset taken from the Valuation and Property Service Department (VPSD). The VSPD officially records a property transaction once the stamp duty for the Sales and Purchase is paid.

Table 2.1 The comparison of five previous studies on house price prediction in Malaysia

Article, Author, and Year	Sample Data	Variables	Methodology	Result
Neural Network Modelling to Predict House Prices Performance [17]	300 terrace houses for the year 1994 to the year 1996 in Kuala Lumpur and Selangor. Collected from the Valuation and Property Services Department, Ministry of Finance Malaysia (INSPEN).	Year the data was collected, land size, built upsize, type of land ownership, type of the house, age of the house, distance from town, environment quality, and building quality.	● To use a binary pattern for data representation with output signal in the magnitude of 0 to 1. ● To develop an Artificial Neural Network (ANN) model using a multilayer perceptron (MLP) to predict the terrace house prices in Kuala Lumpur and Selangor.	● The variables that have been represented are the type of terrace houses, type of land ownership, as well as area and building qualities.
Factors Affecting the Price of Housing in Malaysia [18]	120 housing prices from the year 2001 to the year 2010 in all states of Malaysia. Collected from Bursa Malaysia, the Department of Statistics of Malaysia and DataStream.	Gross domestic product (GDP), population, interest rate, inflation, cost of construction (COC), real property, and gains tax (RPGT).	● To develop a Multiple Linear Regression (MLR) model to predict the housing price in Malaysia.	● Only three macroeconomic variables (GDP, population, and RPGT) were found to be positively and significantly correlated with the housing price.
Analysis of housing prices in Petaling district, Malaysia using functional relationship model [20]	41,750 terrace house transacted records from November 2008 to February 2016 in the Petaling district. Collected from Jordan Lee and Jaafar (S) Sdn. Bhd and additional information from Google Maps.	Lot size, tenure type, years to the expiry of lease term, terrace type, number of bedrooms, main building size, distance to nearest shopping mall, and transaction date	● To identify the house attributes that significantly contributed to housing prices using p-value for Petaling district and six sub-regions. ● To develop seven models (six for the sub-region and one for all Petaling Jaya district) using Multiple Un-replicated Linear Functional Relationship (MPULFR) Model.	● The terrace type is not significant under the MR model while it reveals that the housing prices increased if there are a fewer number of bedrooms.

Article, Author, and Year	Sample Data	Variables	Methodology	Result
Multiple Regressions in Analysing House Price Variations [21]	1,500 of the double-story terraces for the years 2000 and 2007 in Kuala Lumpur, Malaysia. Collected from the National Property Information Centre (NAPIC) and some are observed through the site visits.	Building area, land area, age of the house, numbers of rooms, distance, the quality of amenities, neighborhood quality, type of holding, and locality.	● To examine the correlation between variables and the price variation using a standard cross-sectional hedonic model. ● To develop two Multiple Regression Analysis (MRA) models for year 2000 and 2007 to predict the double terrace houses.	● The result shows that Locality has been the most influential factor in price with 50.3% and 63% of price variation for year 2000 and 2007 accordingly. ● The second factor that is significant in price variation is the Building Area with 16.3% of price variation in the year 2000 and 14.1% price variation in the year 2007. ● A higher price is paid for land area and building area. ● Kuala Lumpur's housing is dominated by the preference for location and location-related factors.
Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia [22]	Housing selling prices for the year 2016 in the Petaling Jaya area, Selangor.	Buying price, floor, the green certificate, main floor area, number of bedrooms, distance, building category, ownership, category area, area classification, building classification, age of the building, buyers, and seller	● To apply the Pearson correlation coefficient, r value for feature selection. ● To utilize and compare the performance of five machine learning algorithms namely Random Forest, Decision Tree, Multiple Linear Regression, Lasso Regression, and Ridge Regression to predict the house prices.	● The buying price has the strongest correlation with the Selling price with a 0.73 coefficient correlation value.

Table 2.1 states that studies conducted using the Valuation and Property Service Department (VSPD) dataset was between the year 1994 and 2007 [17], [21]. Both studies were conducted in Selangor and Kuala Lumpur, respectively. The studies revealed that the determinant factors for the house price prediction in Selangor and Kuala Lumpur are the type of terrace house, type of land ownership, building quality [17], Gross domestic product (GDP), population, real property, and gains tax (RPGT) [18], the selling price [22], land area, building area and location-related factors [21]. Moreover, from all the five studies, only [17] and [20] specifically studied the determinant factors for terrace houses which were located in the Petaling District, in Selangor and Kuala Lumpur. Therefore, there is an opportunity to extend the study using the VPSD dataset to conduct the correlation analysis for terrace houses in the state of Penang.

3. Methodology

Figure 3.1 shows the operational framework of this study. The framework includes four main phases which are: Phase one; the project planning, Phase two; data acquisition and data pre-processing, Phase three; data analysis, and finally, Phase four; conclusion.

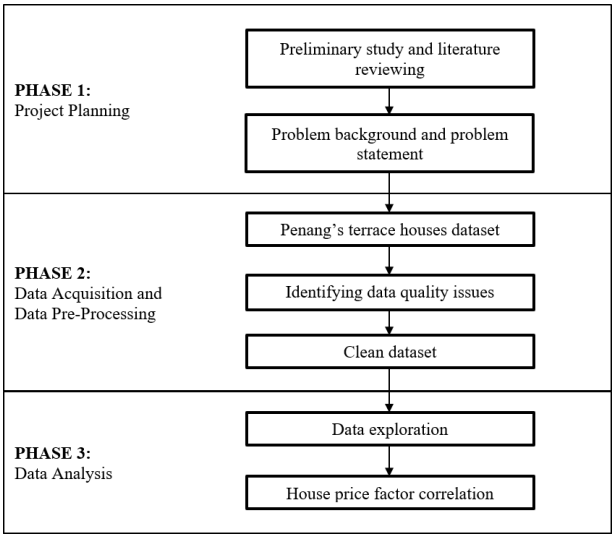


Figure 3.1 Operational framework

In phase one which is the project planning, the relevant literature review on house prices is done. This is to allow a thorough development of the research questions, research objectives, and research scope. Besides, it is done to gain a sound preliminary understanding of the house price determinant factors especially within the context of Malaysia.

The purpose of phase two is to gather Penang's terrace houses dataset for the year 2018 and 2019 and this data is collected from the Brickz. A pre-processing task is performed on this set of data. This dataset is collected from the free subscription which has a limited list of transaction data of the sold terrace houses in

Penang from January 2018 until December 2019. All the transactions represent sub-sale transactions only. The collected dataset consists of 2,699 sold terrace houses in Penang, Malaysia with 11 features or variables. The data collected by manually copying the list into the Excel sheet, because Brickz did not provide the data in downloadable files such as CSV or Excel format. The dataset contains ten features with two categorical data types and eight features are in numerical data types. For this study, the dataset will be transformed into a quantitative dataset to proceed with the data analysis. Table 3.1 presents a summary of the dataset features.

Table 3.1 Features description in the dataset

No	Feature Name	Description	Data Type
1	Transaction Date	The date of the property was sold *(Jan 2018 until Dec 2019)*	Numerical
2	Location	The property area or district *(consists of 33 locations)*	Categorical
3	Building_Type	The property building type *(intermediate, corner lot, or end lot)*	Categorical
4	Tenure	The property tenure type *(freehold – 1, least hold – 0)*	Numerical
5	Floors	The number of floors of the property *(1, 1.5, 2, 2.5, 3, 3.5)*	Numerical
6	Rooms	The number of rooms of the property *(0, 1, 2, 3, 4, 5, 6, 7)*	Numerical
7	Land_Area	The size of the property land area	Numerical
8	Build_Up	The size of the property	Numerical
9	Price_Psf	The property price per square feet	Numerical
10	Price	The property sold price	Numerical

Data cleaning purpose is to increase data quality by identifying and eliminating errors and inconsistencies from the dataset [5]. Hence, several tasks that are applied include removing missing or duplicate information, filtering meaningless data, and consolidating distinct data representations to have consistent and accurate data. The dataset undergoes the pre-processing process which includes renaming the heading, creating dummy data for categorical features via one-hot encoding, regrouping the data, removing uncommon features, and changing data type. These processes are to get the optimal number of features that describe all the important information in the dataset [23]. This is followed by a process of detecting and dealing with missing values, wrong spelling, and outliers, either to remove or replace them. All these are to ensure that the dataset is complete and accurate for the analysis of data and development of the predictive model.

The process involved in phase three is data analysis. This phase comprises two main activities which are data exploration and house price factor correlation. In data exploration, the dataset undergoes descriptive analysis (also called the explanatory data analysis). The data exploration or graphical exploration's goals are to fully understand and provide an in-depth preliminary investigation on the characteristics of the dataset and to determine any data quality issues that exist in the dataset. In particular, this process is to identify outliers, examine, and descriptive statistics of all the features [5]. The analysis will be displayed in statistical and graphical illustrations. As for the house price factor correlation, the dataset undergoes the correlation analysis.

The correlation analysis is one way to measure the strength of the relationship between two continuous or numerical features. Descriptive variables (house features) that correlate strongly with the house price (target feature) would be a good place to start building a predictive model. By determining which input features are associated with the house price, it will ensure that only relevant features are included in the model. Consequently, this is to produce a fitted house price prediction model. This analysis is done on all features using the Pearson Product Moment Correlation (also called the Pearson correlation coefficient), represented by r value. The Pearson correlation coefficient measurement is used to compute the correlation coefficient between ratio or interval features. Table 3.2 illustrates the interpretation of the Pearson correlation coefficient, r value adapted from [22]. Thus, this correlation analysis is presented in a correlation matrix table and correlation heatmap.

Table 3.2 Interpretation of the Pearson correlation coefficient, r value

r value	Interpretation
0.51 – 1.00	Strong
0.30 – 0.50	Moderate
0.20 – 0.29	Weak
0.10 – 0.19	Very weak
0.00	No association

Finally, in phase four, the house price factor correlation is identified through correlation analysis to identify the association with the house prices. In doing this, the highest Pearson correlation coefficient that is represented by the r value indicates the strongest relationship with the terrace house price.

4. Result and Discussion

4.1 Data Pre-processing

The pre-processing task is done for Penang's terrace houses dataset in Jupyter Notebook and using Seaborn as well as matplotlib libraries. The following section provides a detailed explanation of data pre-processing procedures that are set for the data.

4.1.1 Incomplete and Missing Data

In data preparation, problems arise when there are missing or empty values. A missing value in a variable occurs when an actual value exists but has been omitted during data entering. Referring to Figure 4.1, the dataset of Penang's terrace houses is free from incomplete and missing data. Hence, all 2,699 data will be used in the next pre-processing stage.

```
df = pd.read_excel(r'C:\Users\Fathy\Desktop\MASTER PROJECT 1\Master_Project_1_Norhayati\Penang_Terrace.xlsx')

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2699 entries, 0 to 2698
Data columns (total 10 columns):
Transaction Date    2699 non-null datetime64[ns]
Location            2699 non-null object
Building_Type       2699 non-null object
Tenure              2699 non-null int64
Floors              2699 non-null float64
Rooms               2699 non-null int64
Land_Area           2699 non-null int64
Built_Up            2699 non-null int64
Price_Psf           2699 non-null int64
Price               2699 non-null int64
dtypes: datetime64[ns](1), float64(1), int64(6), object(2)
memory usage: 211.0+ KB
```

Figure 4.1 Incomplete and missing data detection.

4.1.2 Handling Categorical Features

The dataset contains two features of the categorical data type and they are 'Location', and 'Building Type'. The categorical features need to be converted into an integer [24]. This process used to get the optimal number of features that describes all the important information in the dataset [23]. One of the approaches is by creating a dummy variable using the one-hot encoding method, in which each of the categories for the categorical feature is converted into a separate binary variable that has a value of '1' and '0' [25]. As described in Table 3.1, the 'Location' feature consists of 33 locations, and the 'Building Type' feature contains 3 building types.

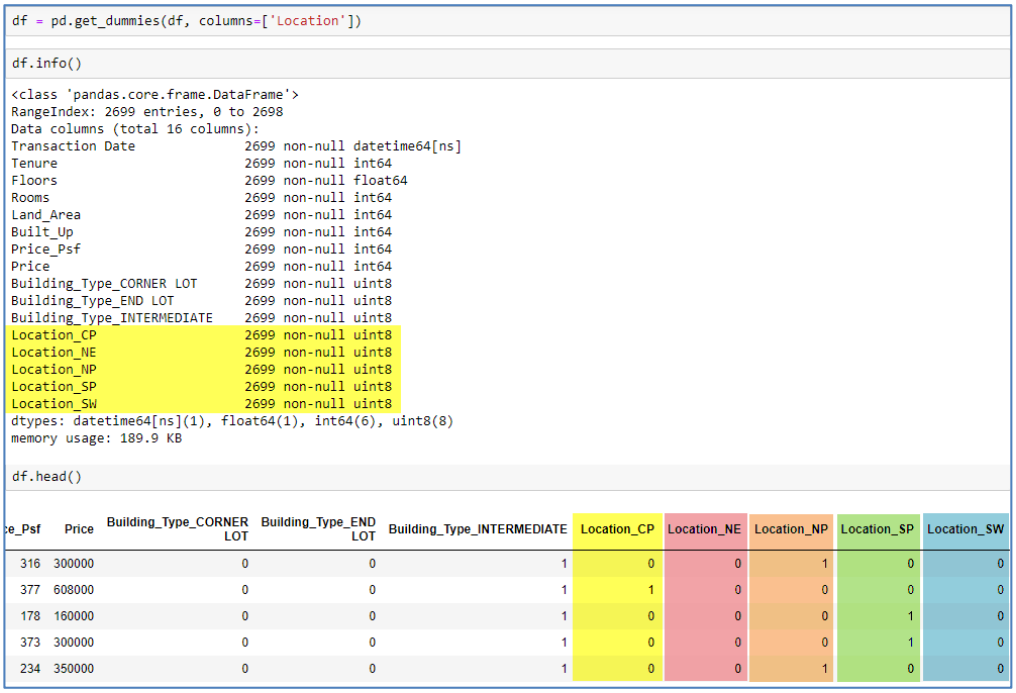
For the 'Building Type' feature, each category is replaced with three separate binary variables, which are 'Building_Type_CORNER LOT', 'Building_Type_END LOT', and 'Building_Type_INTERMEDIATE' as new features in the dataset. Each feature contains '1' and '0' value. Figure 4.2 shows the dummy variables conversion for 'Building_Type'.

```
df = pd.get_dummies(df, columns=['Building_Type'])

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2699 entries, 0 to 2698
Data columns (total 12 columns):
Transaction Date              2699 non-null datetime64[ns]
Location                      2699 non-null object
Tenure                        2699 non-null int64
Floors                        2699 non-null float64
Rooms                         2699 non-null int64
Land_Area                     2699 non-null int64
Built_Up                      2699 non-null int64
Price_Psf                     2699 non-null int64
Price                         2699 non-null int64
Building_Type_CORNER LOT      2699 non-null uint8
Building_Type_END LOT         2699 non-null uint8
Building_Type_INTERMEDIATE    2699 non-null uint8
dtypes: datetime64[ns](1), float64(1), int64(6), object(1), uint8(3)
memory usage: 197.8+ KB

df.head()
```

nsaction Date	Location	Tenure	Floors	Rooms	Land_Area	Built_Up	Price_Psf	Price	Building_Type_CORNER LOT	Building_Type_END LOT	Building_Type_INTERMEDIATE
18-01-02	NP	1	1.0	3	1540	949	316	300000	0	0	1
18-01-02	CP	1	2.0	3	1399	1613	377	608000	0	0	1
18-01-02	SP	1	1.0	4	1399	900	178	160000	0	0	1
18-01-04	SP	1	1.0	3	1320	805	373	300000	0	0	1
18-01-05	NP	1	2.0	4	1544	1497	234	350000	0	0	1

Figure 4.2 Dummy variable conversion for 'Building_Type' feature

For the ‘Location’ feature, it contains 33 location categories, which leads to the creation of 33 dummy or new variables in the dataset. Due to that, the current ‘Location’ feature is divided into five separate groups based on the districts in Penang. They are South Seberang Perai (SP), Central Seberang Perai (CP), North Seberang Perai (NP), Northeast Penang Island (NE), and Southwest Penang Island (SW) [26]. Thus, the new five features are introduced as ‘Location_SP’, ‘Location_CP’, ‘Location_NP’, ‘Location_NE’, and ‘Location_SW’ as the representation of five districts of Penang. It is a similar process to ‘Building_Type’ feature conversion, where the new five features contain ‘1’ and ‘0’ value. The conversion of the ‘Location’ feature is depicted in Figure 4.3.



```
df = pd.get_dummies(df, columns=['Location'])
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2699 entries, 0 to 2698
Data columns (total 16 columns):
Transaction Date              2699 non-null datetime64[ns]
Tenure                        2699 non-null int64
Floors                        2699 non-null float64
Rooms                         2699 non-null int64
Land_Area                     2699 non-null int64
Built_Up                      2699 non-null int64
Price_Psf                     2699 non-null int64
Price                         2699 non-null int64
Building_Type_CORNER LOT      2699 non-null uint8
Building_Type_END LOT         2699 non-null uint8
Building_Type_INTERMEDIATE    2699 non-null uint8
Location_CP                   2699 non-null uint8
Location_NE                   2699 non-null uint8
Location_NP                   2699 non-null uint8
Location_SP                   2699 non-null uint8
Location_SW                   2699 non-null uint8
dtypes: datetime64[ns](1), float64(1), int64(6), uint8(8)
memory usage: 189.9 KB
```

```
df.head()
```

e_Psf	Price	Building_Type_CORNER LOT	Building_Type_END LOT	Building_Type_INTERMEDIATE	Location_CP	Location_NE	Location_NP	Location_SP	Location_SW
316	300000	0	0	1	0	0	1	0	0
377	608000	0	0	1	1	0	0	0	0
178	160000	0	0	1	0	0	0	1	0
373	300000	0	0	1	0	0	0	1	0
234	350000	0	0	1	0	0	1	0	0

Figure 4.3 Conversion of ‘Location’ feature

With the creation of an additional eight new features as dummy variables via one-hot encoding, the total features that will be used for the correlation analysis are 16 features. Table 4.1 presents all the features post the data pre-processing procedure.

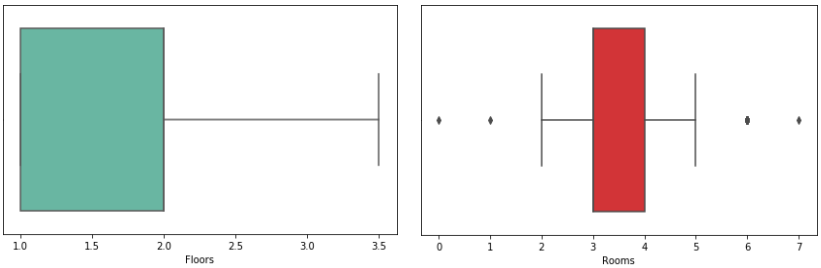
Table 4.1 Features description after data pre-processing task

No	Feature Name	Description	Data Type
1	Transaction Date	The date of the property was sold *(Jan 2018 until Dec 2019)*	Numerical
2	Tenure	The property tenure type *(freehold – 1, least hold – 0)*	Numerical
3	Floors	The number of floors of the property *(1, 1.5, 2, 2.5, 3, 3.5)*	Numerical
4	Rooms	The number of rooms of the property *(0, 1, 2, 3, 4, 5, 6, 7)*	Numerical
5	Land_Area	The size of the property land area	Numerical

No	Feature Name	Description	Data Type
6	Build_Up	The size of the property	Numerical
7	Price_Psf	The property price per square feet	Numerical
8	Price	The property sold price	Numerical
9	Building_Type _CORNER LOT	The property with corner lot building type *(corner lot = 1, not corner lot = 0)*	Numerical
10	Building_Type _END LOT	The property with end lot building type *(end lot = 1, not end lot = 0)*	Numerical
11	Building_Type _INTERMEDIATE	The property with intermediate building type *(intermediate = 1, not intermediate = 0)*	Numerical
12	Location_CP	The property located at Central Seberang Perai *(if yes = 1, if no = 0)*	Numerical
13	Location_NE	The property located at Northeast Island *(if yes = 1, if no = 0)*	Numerical
14	Location_NP	The property located at North Seberang Perai *(if yes = 1, if no = 0)*	Numerical
15	Location_SP	The property located at South Seberang Perai *(if yes = 1, if no = 0)*	Numerical
16	Location_SW	The property located at Southwest Island *(if yes = 1, if no = 0)*	Numerical

4.1.3 Outlier Observation

Outliers can affect a prediction model accuracy by pulling the estimated prediction line further away from the true population line [27]. Moreover, an outlier is an observation or data which its presence is not consistent or isolated from the rest of the data [28]. A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. If the dataset is too large and the range of value is too big, then it shows some values as outliers based on an inter-quartile function [29]. Among the 15 features, five of the features are illustrated in the box plot, which is 'Floors', 'Rooms', 'Land_Area', 'Built_Up', 'Price_Psf' and 'Price' in Figure 4.4. The remaining eleven features; the 'Transaction Date', 'Price', 'Tenure', 'Building_Type_CORNER LOT', 'Building_Type_END LOT', 'Building_Type_INTERMEDIATE', 'Location_SP', 'Location_CP', 'Location_NP', 'Location_NE', and 'Location_SW' are not being processed for outlier detection because of the feature data type is a date, the features contain '1' and '0' value, and the target variable.



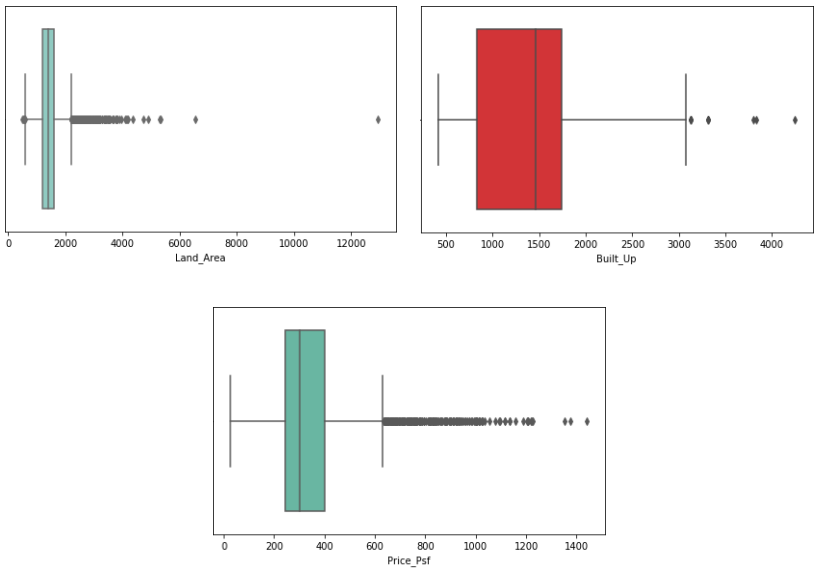


Figure 4.4 Box plot to identify outliers in 'Floors', 'Rooms', 'Land_Area', 'Built_Up', and 'Price_Psf' features

Table 4.2 presents a summary of the pre-processing procedure that has been done to the outlier. In this procedure, 543 records have been eliminated from the dataset. According to [30], one of the cases where the whole records can be removed is when the variables are related to the target variable. Since the 'Rooms', 'Land_Area', 'Built_Up', and 'Price_Psf' features are the determining variables for the target variable which is 'Price', therefore, removal of the outlier records is acceptable. Therefore, by removing all the outlier records, the total records left is 2,155 records. Figure 4.5 shows the outcome of the data pre-processing task.

Table 4.2 Pre-processing task to the outliers

Feature Name	Number of Outliers	Approached Used
Floor	0	No action is taken
Rooms	23	Remove the records
Land_Area	271	Remove the records
Built_Up	12	Remove the records
Price_Psf	237	Remove the records

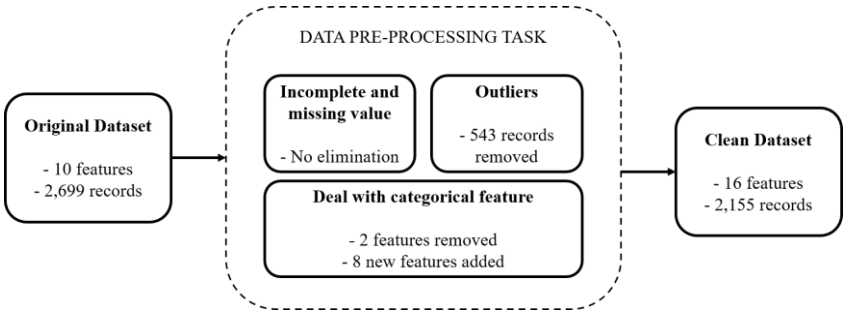


Figure 4.5 Outcome of the data pre-processing task

4.2 Data Analysis

After the data pre-processing task is conducted, a clean dataset is obtained. To gain a preliminary understanding of the dataset, the data analysis is conducted to determine the data pattern and distribution. All 2,155 records have undergone two types of data analysis, which are the descriptive statistic (also called the explanatory data analysis) and correlation analysis. The data analysis task is done for Penang's terrace houses dataset in Jupyter Notebook and using Seaborn as well as matplotlib libraries. The following section will provide a detailed explanation of the data analysis procedures.

4.2.1 Descriptive Statistics

Table 4.3 provides a summary of the descriptive statistic for all the dataset features excepts the transaction date, which is comprising of Mean, Standard Deviation, Minimum, Quartile 1 (25%), Quartile 2 (50%), Quartile 3 (75%), and Maximum.

Table 4.3 Descriptive statistic for the dataset

Features	Mean	SD	Min	25%	50%	75%	Max
Price	420,270.50	260,392.80	20,000.00	232,000.00	363,000.00	520,000.00	1,700,000.00
Tenure	0.94	0.23	0.00	1.00	1.00	1.00	1.00
Floors	1.72	0.63	1.00	1.00	2.00	2.00	3.50
Rooms	3.45	0.68	2.00	3.00	3.00	4.00	5.00
Land_Area	1,336.35	238.20	700.00	1,195.00	1,302.00	1,442.00	2,055.00
Built_Up	1,353.66	561.89	418.00	800.50	1,433.00	1,711.00	2,964.00
Price_Psf	303.34	106.12	27.00	235.00	286.00	356.00	632.00
Building_Type_ CORNER LOT	0.03	0.16	0.00	0.00	0.00	0.00	1.00
Building_Type_ END LOT	0.05	0.23	0.00	0.00	0.00	0.00	1.00
Building_Type_ INTERMEDIATE	0.92	0.27	0.00	1.00	1.00	1.00	1.00
Location_CP	0.34	0.47	0.00	0.00	0.00	1.00	1.00
Location_NE	0.06	0.25	0.00	0.00	0.00	0.00	1.00
Location_NP	0.23	0.42	0.00	0.00	0.00	0.00	1.00
Location_SP	0.28	0.45	0.00	0.00	0.00	1.00	1.00
Location_SW	0.09	0.28	0.00	0.00	0.00	0.00	1.00

In summary, the lowest terrace house price is RM 20,000.00 and the most expensive terrace house is sold at RM 1,700,000.00. The average price of the house is RM 420,270.50 which is quite high. As for the land area size, the average size of the land area that has been sold is 1,336.35 sqft and the maximum land size that has been sold is 2,055.00 sqft. Meanwhile, the average built-up area is 1,353.66 sqft and the maximum built-up size is 2,964.00 sqft which indicates that the terrace houses in Penang are huge.

Histograms are used to illustrate the frequency distribution of quantitative data. These plots are based on a single variable and show the frequency of the unique values of a given variable [29]. Figure 4.6 shows the histogram for 'Floors', 'Rooms', 'Land_Area', 'Built_Up', 'Price_Psf', and 'Price' features. By looking at the 'Price_Psf' histogram, it shows positive skew which indicates the tail on the right side of the distribution is longer or the mean and median will be greater than the mode. Similarly, the skewness for 'Price' is positive.

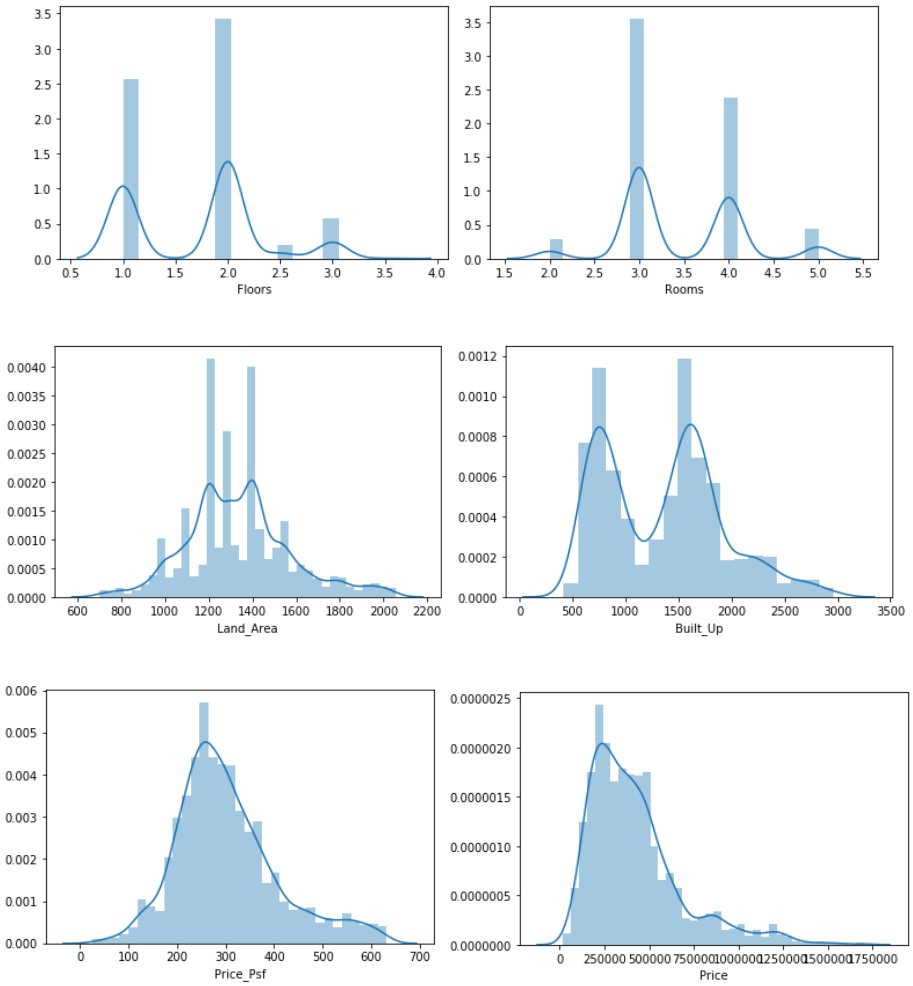


Figure 4.6 Histogram of the frequency distribution of Floors, Rooms, Land_Area, Built_Up, Price_Psf, and Price

The scatter plot is often used for visualizing relationships between two numerical variables. Thus, by plotting a scatter plot, the relationships between variables can be easily visualized and help to gain insights from the relationship patterns and correlation [29]. The pattern and relationship between 'Floors', 'Rooms', 'Land_Area', 'Built_Up', 'Price_Pst' with 'Price' are illustrated in Figure 4.7.

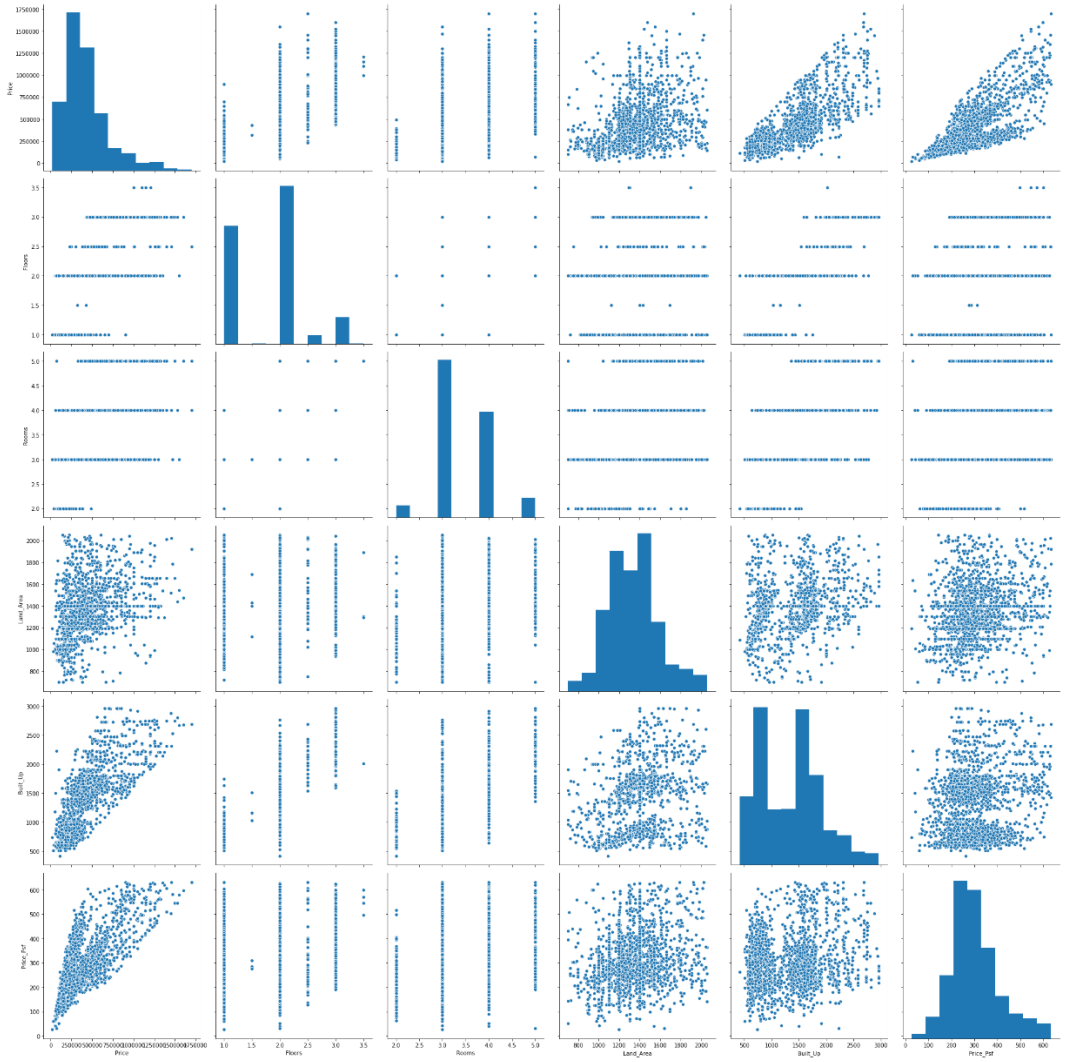


Figure 4.7 Scatter plot for 'Tenure', 'Floors', 'Rooms', 'Land_Area', 'Built_Up', 'Price_Pst' with 'Price'

In summary, there is a linear relationship between the 'Built_Up' and 'Price' as well as 'Price_Psf' and 'Price'. In other words, the larger the 'Built_Up' size, and the higher the 'Price_Psf' is, the higher the likelihood of 'Price' will be.

4.2.2 Correlation Analysis

The correlation analysis is one way to measure the strength of the relationship between two continuous or numerical features. A descriptive feature that correlates strongly with the house price (target feature) would be a good place to start building a predictive model. By determining which input features are associated with the house price will ensure that only relevant features are included in the model. Consequently, this is to produce a fitted house price prediction model. The analysis is done on all the features using the Pearson correlation coefficient represented by r value.

Table 4.4 shows the Pearson correlation coefficient, r value between all the variables, and the house price. The correlation coefficient quantifies the direction and strength of the relationship between two numeric variables, and lies between -1.0 and 1.0 [31]. A value of the correlation coefficient close to +1 indicates a strong positive linear relationship (i.e., one variable increases with the other). A value close to –1 indicates a strong negative linear relationship (i.e., one variable decreases as the other increases). A value close to 0 indicates no linear relationship [32]. In this research, the level of relationship is illustrated in Table 3.2.

Table 4.4 Correlation between features and 'Price' using the Pearson correlation coefficient

Features	Pearson correlation coefficient, r value
Built_Up	0.775
Price_Psf	0.705
Floors	0.703
Rooms	0.56
Location_NE	0.55
Land_Area	0.385
Location_SW	0.374
Tenure	0.135
Building_Type_INTERMEDIATE	0.02
Building_Type_END LOT	-0.006
Building_Type_CORNER LOT	-0.025
Location_NP	-0.06
Location_CP	-0.132
Location_SP	-0.341

A heatmap is used to graphical display denotes the values of numerical data. A heatmap is useful for two purposes, especially from a data mining perspective. It is useful in visualizing correlation tables and in visualizing missing values in the data [33]. In this research, the heatmap has been used to visualize the correlation between all the house features and the terrace house price as illustrated in Figure 4.8.

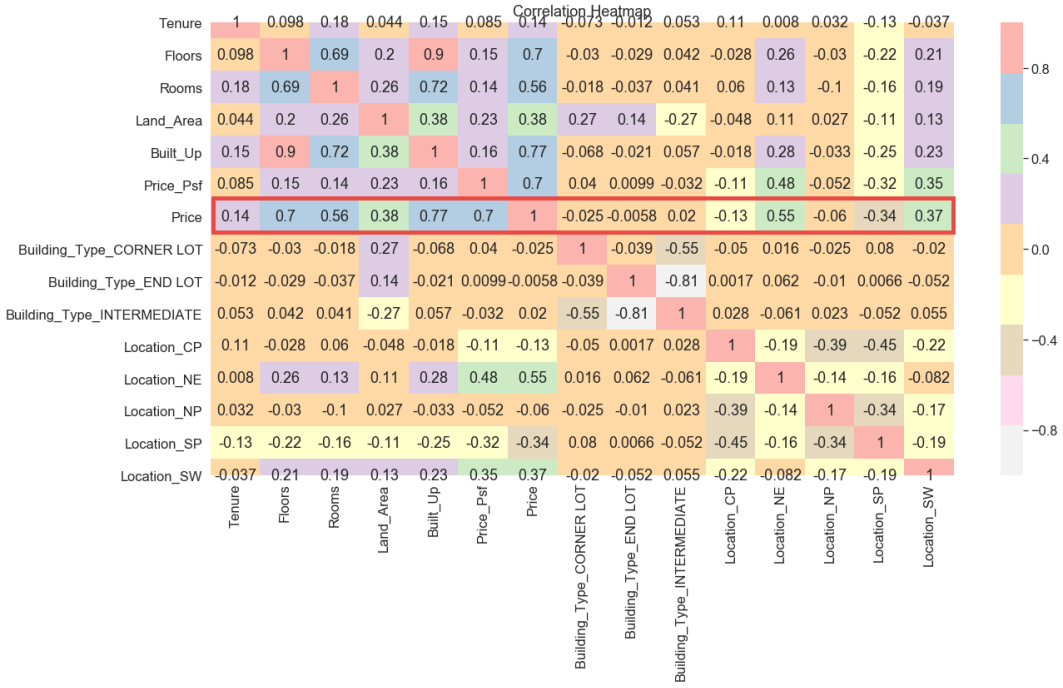


Figure 4.8 Correlation heatmap between house features and 'Price' using the Pearson correlation coefficient

4.3 Discussion

Table 4.5 shows the study results of the Pearson correlation coefficient between the house features and target variable, 'Price'. It also provides the direction and strength interpretation of the relationship between the house features and 'Price'. From the table, the correlation analysis has identified that the size of the property ('Built_Up') has the strongest relationship with the terrace house price with the correlation coefficient, r value of 0.775. This result can be interpreted as the bigger the size of the property is, the higher the house price will be.

Table 4.5 Correlation between variables and 'Price' using the Pearson correlation coefficient and the value interpretation

Variables	Pearson correlation coefficient, r value	Interpretation
Built_Up	0.775	Strong
Price_Psf	0.705	Strong
Floors	0.703	Strong
Rooms	0.56	Strong
Location_NE	0.55	Strong
Land_Area	0.385	Moderate
Location_SW	0.374	Moderate
Tenure	0.135	Very Weak

Variables	Pearson correlation coefficient, r value	Interpretation
Building_Type_INTERMEDIATE	0.02	No Association
Building_Type_END LOT	-0.006	No Association
Building_Type_CORNER LOT	-0.025	No Association
Location_NP	-0.06	No Association
Location_CP	-0.132	Very Weak (Negative)
Location_SP	-0.341	Moderate (Negative)

Besides, the results show that the other strong correlated terrace house features to the house price include the property price per square feet ('Price_Psf') with an r value of 0.705, the number of floors of the property ('Floors') with an r value of 0.703, and the number of rooms of the property ('Rooms') with an r value of 0.56. These signify that the higher the property price per square feet is, the higher number of floors, and rooms of a house, the higher the price will be.

Across all the house features, one location-based feature has a strong relationship with the price. It has been found that the property located at Northeast Island ('Location_NE') has a strong correlation with an r value of 0.55. The analysis shows that if the terrace house is located at Northeast Island, there will be an increase in the house price. It is not surprising, as Northeast Island is situated within the heart of George Town, which is also the Penang's capital city [34]. This leads to expensive house prices.

Additionally, it is also found that four house features have no association with the house price; the property with intermediate building type ('Building_Type_INTERMEDIATE'), the property with end lot building type ('Building_Type_END LOT'), the property with corner lot building type ('Building_Type_CORNER LOT'), and the property located at North Seberang Perai ('Location_NP'). This no association means that the house price is not affected by the property building type and if the terrace house is situated at North Seberang Perai, Penang. Interestingly, North Seberang Perai is largely covered by paddy fields and the largest town in the district is Butterworth [35].

The last finding in this correlation analysis is the negative moderate and negative very weak relationship between the property located at South Seberang Perai ('Location_SP'), and the property located at Central Seberang Perai ('Location_CP') with the house price. The negative direction means there is a decrease in the house value for the terrace house situated in both South Seberang Perai and Central Seberang Perai. Although the correlation is moderate and very weak, it affects house prices. Hence, this shows that with the same terrace house features, the buyer can purchase the house cheaper here than other locations.

5. Conclusion

All in all, this study has found that five house attributes are highly associated with the price of a terrace house in Penang. These attributes are the size of the house, the house price per square feet, the number of floors of the house, the number of rooms of the house, and the location which is in the Northeast Island of Penang. Among these attributes, size has the strongest relationship with price. Besides, this study has also found that house price is not associated with the types of house, be it an end lot or a corner lot. It has also revealed that the terrace houses located in South Seberang Perai and Central Seberang Perai are cheaper than the ones with the same attributes in other locations in the state.

It is recommended that future work to include more house attributes such as location (near to major highways, accessible by public transport), and neighborhood characteristics (population density, nearest school, clinics, and shopping location) This would definitely add more insights to factors affecting prices of terrace houses and would eventually provide a better understanding on Malaysia's real estate market.

References

[1] T. Tech Hong, "Home owning motivation in Malaysia," *J. Accounting, Bus. Manag.*, vol. 1, no. 1, pp. 93–112, 2009.

[2] S. Lip Sean and T. Tech Hong, "Factors Affecting the Purchase Decision of Investors in the Residential Property Market in Malaysia," *J. Surv. Constr. Prop.*, vol. 5, no. 2, pp. 1–13, 2014.

[3] N. Vineeth, M. Ayyappa, and B. Bharathi, "House Price Prediction Using Various Machine Learning Algorithms," in *Communications in Computer and Information Science*, 2018, vol. 837, pp. 425–433.

[4] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, no. 1, pp. 171–176, 2018.

[5] A. Komagome-towne, "Models and Visualizations for Housing Price Prediction," Faculty of California State Polytechnic University, Pomona, 2016.

[6] Z. Zainuddin, "An Empirical Analysis of Malaysian Housing Market : Switching and Non-Switching Models," Lincoln University, 1994.

[7] "Key Statistics - NAPIC," *National Property Information Centre*, 2019. [Online]. Available: http://napic.jpph.gov.my/portal/web/guest/key-statistics. [Accessed: 19-Feb-2020].

[8] Bank Negara Malaysia, "Risk Developments and Assessment of Financial Stability in 2018," *Bnm*, pp. 11–34, 2018.

[9] R. Bafna, A. Dhole, A. Jagtap, A. Kazi, and A. Kazi, "Prediction of Residential Property Prices – A State of the Art," *Int. Adv. Res. J. Sci. Eng. Technol.*, vol. 5, no. May, pp. 2007–2010, 2018.

[10] A. S. Ravikumar, "Real Estate Price Prediction Using Machine Learning," National College of Ireland, 2018.

[11] A. Nguyen, C. Fernandes, N. Webb, and H. Holt, "Housing Price Prediction," Union College, 2018.

[12] J. Umbelina, "Top Malaysia property portals ranking by traffic in 2018 - Joseba Umbelina," *Joseba Umbelina*, 2019. [Online]. Available: https://www.josebaumbelina.com/real-estate-markets/top-malaysia-property-portals-ranking-by-traffic-in-2018/. [Accessed: 30-Apr-2020].

[13] "Discover Actual Transacted Property Prices In Malaysia," *Brickz Research Sdn Bhd*. [Online]. Available: https://www.brickz.my/. [Accessed: 19-Jun-2020].

[14] "Why brickz," *Brickz Research Sdn Bhd*, 2020. [Online]. Available: https://www.brickz.my/about/. [Accessed: 25-Jun-2020].

[15] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French, "Real estate appraisal: A review of valuation methods," *J. Prop. Invest. Financ.*, vol. 21, no. 4, pp. 383–401, 2003.

[16] S. N. A. Rahman, N. H. A. Maimun, M. N. Razali, and S. Ismail, "The artificial neural network model (ANN) for Malaysian housing market analysis," *Plan. Malaysia*, vol. 17, no. 1, pp. 1–9, 2019.

[17] K. R. Ku-mahamud, A. Abu Bakar, and N. Norwawi, "Multi Layer Perceptron Modelling in the Housing Market," *Malaysian Manag. J.*, vol. 3, no. 1, pp. 61–69, 1999.

[18] T. San Ong, "Factors Affecting the Price of Housing in Malaysia," *Financ. Bank. An Online Int. Mon. J.*, no. 5, p. 415, 2013.

[19] S. H. Kok, N. W. Ismail, and C. Lee, "The sources of house price changes in Malaysia," *Int. J. Hous. Mark. Anal.*, vol. 11, no. 2, pp. 335–355, 2018.

[20] Y. F. Chang, W. C. Choong, S. Y. Looi, W. Y. Pan, and H. L. Goh, "Analysis of housing prices in Petaling district, Malaysia using functional relationship model," *Int. J. Hous. Mark. Anal.*, vol. 12, no. 5, pp. 884–905, 2019.

[21] A. Yusof and S. Ismail, "Multiple Regressions in Analysing House Price Variations," *Commun. IBIMA*, vol. 2012, pp. 1–9, 2012.

[22] T. Mohd, S. Masrom, and N. Johari, "Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, pp. 542–546, 2019.

[23] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2017-Decem, pp. 319–323, 2018.

[24] S. Raschka and V. Mirjalili, *Python Machine Learning - Machine Learning and Deep Learning with Python, scikit-learn and TensorFlow*, Second Edi. Packt Publishing, 2017.

[25] H. Brink, J. W. Richards, and M. Fetherolf, *Real-World Machine Learning*. Manning Publication Co., 2017.

[26] "Penang - Wikipedia," *Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Penang#Local_governments. [Accessed: 26-Jun-2020].

[27] S. Abbasi, "Advanced Regression Techniques Based Housing Price Prediction Model," *13th Int. Conf. Iran. Oper. Res. Soc.*, no. February, pp. 1–10, 2020.

[28] N. Adnan, M. H. Ahmad, and R. Adnan, "A Comparative Study On Some Methods For Handling Multicollinearity Problems," *Matematika*, vol. 22, no. 2, pp. 109–119, 2006.

[29] S. Doshi, "Analyze the data through data visualization using Seaborn," *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/analyze-the-data-through-data-visualization-using-seaborn-255e1cd3948e. [Accessed: 08-Jul-2020].

[30] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015.

[31] GraphPad, "What is the difference between correlation and linear regression?," *GraphPad*, 2019. [Online]. Available: https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/. [Accessed: 31-Mar-2020].

[32] V. Bewick, L. Cheek, and J. Ball, "Statistics review 7: Correlation and regression," *Crit. Care*, vol. 7, no. 6, pp. 451–459, 2003.

[33] G. Shmueli, P. C. Bruce, P. Gedeck, and N. R. Patel, *Data Mining for Business Analytics. Concepts, Techniques and Applications in Python*, 1st ed. John Wiley & Sons, Inc., 2020.

[34] "Northeast Penang Island District - Wikipedia," *Wikimedia Foundation, Inc.*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Northeast_Penang_Island_District. [Accessed: 11-Nov-2020].

[35] "North Seberang Perai District - Wikipedia," *Wikimedia Foundation, Inc.*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/North_Seberang_Perai_District. [Accessed: 11-Nov-2020].