STAT3011
Project I Report

Suen Tsz Ki 1155095081
Tse Kit Ying 1155092382
Chan King Fung 1155095081
Mo Jackie Ryan 1155095511
Chan King Yeung 1155119394

Table of Contents

Executive Summary

This report is divided into three parts according to the major statistical methods used: Summary statistics, Factors affecting the price and Regression analysis of the whole picture. Based on the background introduction and conclusion, you will have a general idea of the report 's objectives. After further investigation into respective analysis methods, a more detailed picture of the report can be provided.

Firstly, based on the summary statistics, we have listed out the required data on the variables. For example, the mean, median and quantiles are calculated to give the trend concerning respective variables. Furthermore, standard error and outliners are listed to show the variation towards each category to reflect the deviations. In return, we can find the most favourable conditions for planning the constructions of buildings.

Secondly, based on factors affecting the price, we divided the variables into two major groups: Internal variables and External variables. From internal factors, we investigate the relationship of price versus building age, size, floor and facilities. While from external factors, we compare the trend of price against average time travelling to the transport station, nearby infrastructure and nearby education facilities. A detailed relationship can be found out in the related pages of this report. We decided to adopt Size, Floor, N_Parking, N_FacilitiesinApt, N_APT, N_Department, N_Mall, and N_Park as the potential factors during the analysis since they are positively correlated to the price.

Thirdly, based on regression analysis, we divided it into three parts. We have analyzed the selection of variables, the choice of regression models and the summary of the regression model. Eventually, we reach an agreement to use the 8 variables model in Table 13 by hypothesis testing of each beta against significance.

Lastly, we are going to provide recommendations for the new estate projects constructing based on the findings on the three statistical models and providing the reasons behind. Together with the Limitations and Improvements of the statistical models that we have found during our process of completing the project, we have pasted the result as a reference as if a further investigation would be carried out, a deeper and more comprehensive analytical method and statistical method can be done.

Background of Report

In this report, we have assumed that we are now working for a renowned estate developer, e.g. Cheung Kong Property Holdings Ltd as data analysts and we are now given a project to analyse the apartment transactions in the district of an Asian city from the past 10 years. After that, what we have to do is to identify the potency of developing estate projects in the territories and try to attract capital deposit of investors.

We believe that with the assumption of the background, it can help us set a specific goal that our analysis will be comparably target-oriented. The data and respective
graphs can be applied to explain the estate pricing phenomenon in this region and the habits of estate buyers. As a result, giving a thorough explanation of the usage of the variables in the dataset and eventually provide a favourable conclusion to our investigation.

All the materials (including graphs, hypothesis test) generated from the report is based on the dataset provided. We will not use any undefined or made-up information to ensure the origin and quality of it.

Summary Statistics

Basic statistics of the variables studied in this analysis will be covered here. According to the basic statistics, a rough background of the trend which how the data are distributed concerning different variables is shown.

a. Modification of Some Variables

Before going into the details of basic summary statistics, some of the variables given in the data set have been modified for better understanding of the variables. Some new variables are created by merging the original variables in the given data set. The regroupings of the variables are as follows:

- Building_Age = Year_Sold – Year_Build
- N_Parking = N_Parking_G + N_Parking_B
- N_School = N_Elementary + N_Middle + N_High + N_University
- AvgTime = Estimated time distance between the apartment and the nearest transportation, based on TimeToSubway and TimeToBus

Building_Age simply means the age of the apartment when being sold. The creation of this variable from Year_Sold and Year_Build is because we that neither of the old variables is considered when people are searching for apartments to buy. But the age of the apartment is a factor which is widely considered, because it is related to other critical problems, such as older apartments tend to need more money for maintenance and repairing, or they will be less comfortable to live in.

N_Parking is the total number of parking spaces near the apartment. Since people only care about whether there would be enough parking spaces for them to park their cars, the location of them (on the ground or in the basement) shouldn't be too much of a concern, N_Parking_G and N_Parking_B are added to form N_Parking.

N_School is the total number of schools or educational facilities around the apartments, gained by adding the numbers of elementary, middle and high schools, as well as universities. Again, we think that the overall number of educational facilities are more important than the number of each type. Therefore, they are added to form one new variable.

AvgTime is the estimated time required to go from the apartments to the nearest transportation. The value of this variable, which is numerical, is calculated from TimeToSubway and TimeToBusStop, which means the time required to go to the nearest subway and bus stop respectively. The values of these two values are ranges of time like “0-5 min” or “10-15 min” (for TimeToBusStop, it could be equal to “no bus stop nearby” and if so, it will be treated as a missing value) and we know that the actual exact times required to lie in these ranges. Assume the actual exact times for going the nearest subway station and bus stop follows a uniform distribution with respect to the range they belong to. For example, if a random data point of TimeToSubway is given by “0-5 min”, then the actual exact time follows a uniform distribution with parameters (0, 5), and we estimate that it is equal the mean of the parameters, i,e (0+5)/2 = 2.5 min, and if a random data point of TimeToBusStop is given by “15-20 min”, then the actual exact time follows a uniform distribution with parameters (15, 20), and we estimate that it is equal to the mean of the parameters, i,e (15+20)/2 = 17.5 min. Then, the AvgTime of the apartment is given by the mean of the estimated actual exact time for going to th nearest subway station and bus stop respectively. If the value of TimeToBusStop of an apartment is given by “no bus stops nearby”, the estimated actual time of going to the nearest bus stop will be treated as a missing value, and the AvgTime will be equal the estimated actual time of going to the nearest subway. This modification is to convert the two ordinal variables (TimeToSubway and TimeToBusStop) to a single numerical variable (AvgTime), for more convenient analysis of the data.

b. Types of Variables Involved

After the modification mentioned above, the 22 variables from the original data set are reconstructed to 16 variables, which we study in the analysis. And we have classified them into serval groups.

- Response Variable: Price (sale price of apartment in USD)
- Explanatory Variables:
  - Internal Factors of the Apartments:
    - Building_Age (age of the apartment when being sold)
    - Month_Sold (Month in which the apartment was sold)
    - Size (area of the apartment in square feet)
    - Floor (Floor on which the apartment is located)
    - N_Parking (No. of parking spaces around the apartment)
    - N_FacilitiesInApt (No. of facilities in the apartment complex)
    - N_APT (No. of Apartment Buildings in the Apartment Complex)
  - External Factors of the Apartments:
    - AvgTime: (Estimated time required to go to the nearest transportation)
    - N_PublicOffice (No of public offices nearby the apartment)
    - N_Hospital (No of hospitals nearby the apartment)

- N_Departmentstore (No of department stores nearby the apartment)
- N_Mall (No of shopping centres nearby the apartment)
- N_ETC (No of facilities like hotels, special schools nearby the apartment)
- N_Park (No of parks nearby the apartment)
- N_Schools (No of educational facilities nearby the apartment)

c. Basic Statistics

The summary statistics of some of the numerical variables are shown as follows:

Variable	Mean	Median	S.D	I.Q.R	Q1	Q3
Response Variable						
Price (USD)	221,416.5	207,964	106,328.8	146,398	144,752	291,150
Explanatory Variables: Internal Factors						
Building_Age	9.715890	7	8.545582	13	3	16
Area (sq. ft.)	955.6589	910	382.2002	505	644	1,149
Floor	12.036917	11	7.550668	11	6	17
N_Parking	766.9956	865	381.5948	755	304	1,059
N_APT	5.615005	7	2.812130	5	3	8
N_Facilities InApt	5.813032	5	2.330653	3	4	7
Explanatory Variables: External Factors						
AvgTime (min)	5.483158	5	3.288114	7.5	2.5	10
N_School	10.864069	10	4.437078	8	7	15
N_PublicOffice	4.140354	5	1.793642	2	3	5
N_ETC	1.940626	1	2.201917	5	0	5
N_Hosiptal	1.2961892	/	0.4798713	/	/	/
N_Department store	0.8962232	/	0.8111332	/	/	/
N_Mall	0.9416468	/	0.4014151	/	/	/
N_Park	0.6544743	/	0.6583500	/	/	/

Variable	Min	Max	Min. Non-Outlining value	Max. Non-Outlining value	Number of Outliners
Response Variable					
Price (USD)	32,743	585,840	32,743	508,849	35
Explanatory Variables: Internal Factors					
Building_Age	0	39	0	35	40
Area (sq. ft.)	135	2,337	135	1,796	138
Floor	1	43	1	33	60
N_Parking	87	1,496	87	1,496	0

N_APT	1	13	1	13	0
N_Facilities InApt	1	10	1	10	0
Explanatory Variables: External Factors					
AvgTime (min)	2.5	12.5	2.5	12.5	0
N_School	0	17	0	17	0
N_PublicOffice	0	7	0	7	0
N_ETC	0	5	0	5	0

Some other variables which are categorical variables, or numerical variables whose details cannot be displayed well enough in like the way above.

The following is the frequency table for Month_Sold, which is a categorical variable.

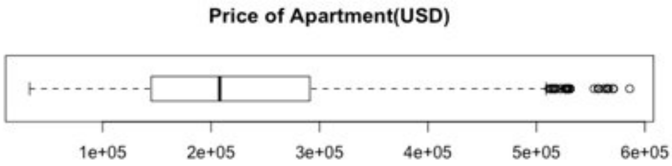
Month	January	February	March	April	May	June
Frequency	623	424	576	450	606	513
Percentage Frequency	10.599%	7.213%	9.799%	7.656%	10.310%	8.727%
July	August	September	October	November	December	Total
550	448	387	519	412	412	5,878
9.357%	7.622%	6.584%	8.830%	7.009%	6.295%	100.000%

The following is the frequency table for N_Hospital, N_Departmentstore, N_Mall, N_Park, which are numerical variables with too few possible values, and therefore presenting information like median, maximum value or outliners not useful. As a result, the frequency table is shown.

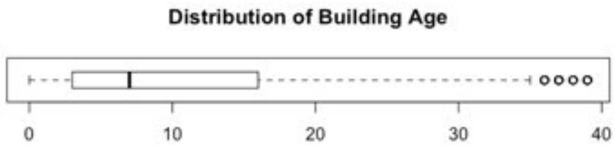
Variable	0	1	2	Total
N_Hospital	64 (1.089%)	4,009 (68.203%)	1,805 (30.708%)	5,878 (100.000%)
N_Departmentstore	2,270 (38.619%)	1,940 (33.141%)	1,660 (28.241%)	5,878 (100.000%)
N_Mall	655 (11.143%)	4911 (83.549%)	312 (5.308%)	5,878 (100.000%)
N_Park	2,640 (44.913%)	2,629 (44.726%)	609 (10.361%)	5,878 (100.000%)

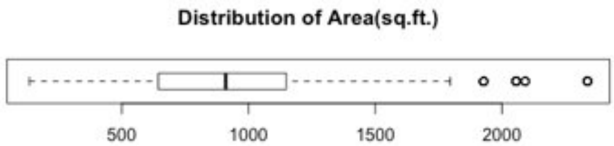
To provide and alternative perspective to overview the data, the followings are the box plots and frequency tables of some of the variables, which have been mentioned above.

The box plot of Price (USD):

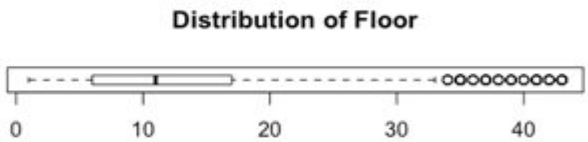


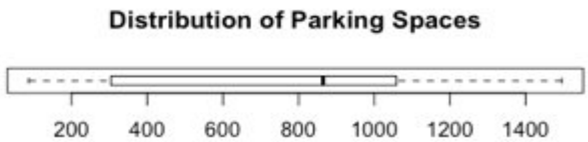
The box plots of Building_Age and Size:



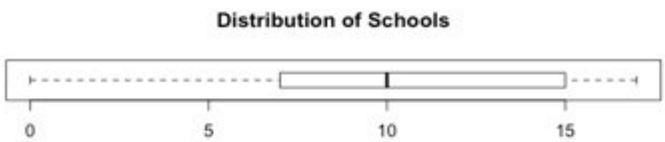


The box plots of Floor and N_Parking





The box plot of N_School:



Internal Factor Affecting Price

In this section, we are focusing on how the inner factors (e.g. size) are correlated with price. And based on the correlations, what should we consider building the type of apartment to achieve a high sales price.

a. Correlation between Price and Building age

From the correlation coefficient between Price and Building age in Table 1, we can notice that the Building age have a negative correlation with price. We suggest selling the apartment as soon as possible because the newest apartment has the highest price.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0	
	Building_Age
Price Price	-0.33931 <.0001

(Table 1)

b. Correlation between Price and Size

According to table 2, the variable of size is positively correlated with price. It shows that a bigger apartment sold with a higher price easier. As a result, it is strongly recommended that building the apartment with a bigger size.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0	
	Size
Price Price	0.69713 <.0001

(Table 2)

c. Correlation between Price and Floor

Floor means which floor the apartment located. Based on table 3, we can see that the variable of Floor has a slightly positive correlation with price. The higher floor it located, the higher the price it has. Therefore, we should promote the higher floor more.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0	
	Floor
Price Price	0.33622 <.0001

(Table 3)

d. Correlation between Price and Facilities

The number of parking spaces and facilities for residents like swimming pool, gym, the playground has a positive correlation with price in table 4. We should build more facilities in order to raise the apartment's price.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0		
	N_Parking	N_FacilitiesinAPT
Price Price	0.43141 <.0001	0.50472 <.0001

(Table 4)

External Factor Affecting Price

In this section, we are focusing on how the outer factors (e.g. facilities around the department) are correlated with price. And based on the correlations, what should we consider building around the apartment to achieve a high sales price.

a. Correlation between Price and Average Time to the nearest apartment

From the correlation coefficient between Price and AvgTime in Table 5, we can notice that the Avgtime variable has a relatively strong negative correlation with price comparing to other variables. It is the most negatively correlated variable with the price. With that being said, it is strongly recommended to locate the apartment where the time takes from the apartment to subway station or bus stop should be around 0 to 5 minutes.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0	
	AvgTime
Price Price	-0.53233 <.0001

(Table 5)

b. Correlation between Price and Nearby Infrastructure

All the variables included in Table 6 are negatively correlated with price. Which is quite strange in this case, since the majority tend to value these facilities a lot. But with the data set given, it seems citizens in this city tend to have less of these facilities in the area. Therefore, we should prevent having facilities such as hospitals, ETC and public offices near the apartment.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0			
	N_Hospital	N_ETC	N_PublicOffice
Price Price	-0.25809 <.0001	-0.44245 <.0001	-0.46165 <.0001

(Table 6)

Based on Table 7, all variables are positively correlated with price. The variable park has the most notable positive impact on the sales price, followed by the variable department store. Therefore, by comparing their correlation coefficient, we should build relatively more parks and department stores near our apartment in order to achieve our goal.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0				
	N_Mall	N_APT	N_Dpartmentstore	N_Park
Price Price	0.08299 <.0001	0.16188 <.0001	0.29716 <.0001	0.31156 <.0001

(Table 7)

c. Correlation between Price and Schools

According to Table 8, the variable school has a relatively strong negative correlation with price. Since having more schools in the area will tend to drive the price down, therefore, it is recommended to not have any schools built near the apartment.

Pearson Correlation Coefficients, N = 5878 Prob > |r| under H0: Rho=0	
	School
Price Price	-0.37857 <.0001

(Table 8)

Overall, the apartment should be located somewhere near to the bus stop and subway station (0~5 minutes) and it is ideal to have more facilities such as department stores and parks nearby the department.

d. Potential Factors Affecting Price

We found that the following variables are positively correlated to the price.

Size, Floor, N_Parking, N_FacilitiesinApt, N_APT, N_Department, N_Mall, N_Park

As our research question is to investigate some potential factors that make an increment to the price, the above 8 variables are focused to be studied in the next sections.

Regression Analysis

SAS University Edition (SAS Institute) was adopted for data organization in our project, including data retrieving and cleaning. SAS has been used for statistical analyses in the project, such as preparing graphs and constructing hypothesis testing tables.

a. Variable Selection

To build a best-fit regression model, we consider the importance among the 8 variables selected by their correlation in the previous section. In the beginning, we adopted the adjusted coefficient of determination to select the relative "best" model among the 8 variables.

Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
8	0.7201	0.7197	9.0000	Size Floor N_Parking N_FacilitiesInApt N_APT N_Dpartmentstore N_Mall N_Park
7	0.7108	0.7105	201.4473	Size N_Parking N_FacilitiesInApt N_APT N_Dpartmentstore N_Mall N_Park
7	0.7059	0.7055	305.3130	Size Floor N_FacilitiesInApt N_APT N_Dpartmentstore N_Mall N_Park
7	0.7040	0.7036	344.8659	Size Floor N_Parking N_FacilitiesInApt N_APT N_Dpartmentstore N_Park
6	0.7009	0.7005	408.3397	Size Floor N_FacilitiesInApt N_APT N_Dpartmentstore N_Park

(Table 9)

We obtained Table 9 from SAS output about the variable selection. We sorted the table by adjusted R-Square of each possible model. The criteria of the method are to obtain the model with the highest value of R_a^2. From Table 1, we can notice that the full model, which contains all the variables, has the largest value of R_a^2 and we consider the full model can explain the variation of data in a good way. Even though other models mentioned in the above table have a similar R_a^2 value with the full model, the Mallow's statistic for those model does not close to the corresponding number of terms. Hence, we are not going to consider those models.

Up to the current point of view, the full model seems to be the relative "best" model that we should use in our project. Nevertheless, given the fact that the adjusted coefficient of determination is not a good quantity to justify the model due to the dependence between regression sum of square and total sum of square, we cannot only use the above result to conclude the full model does a good job in fitting the data. Thus, we established the stepwise selection method to obtain the relative "best" model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Size		1	0.4860	0.4860	4903.38	5555.77	<.0001
2	N_FacilitiesInApt		2	0.1711	0.6571	1317.86	2931.45	<.0001
3	N_APT		3	0.0142	0.6712	1023.11	252.87	<.0001
4	Floor		4	0.0122	0.6834	770.250	225.48	<.0001
5	N_Parking		5	0.0029	0.6863	710.927	54.75	<.0001
6	N_Park		6	0.0060	0.6923	587.849	113.82	<.0001
7	N_Dpartmentstore		7	0.0117	0.7040	344.866	231.69	<.0001
8	N_Mall		8	0.0161	0.7201	9.0000	337.87	<.0001

(Table 10)

The concept of stepwise selection is using partial F-test repeatedly to select variables that are fitted to the data. Table 10 is the SAS summary table of stepwise selection. We can see that in step 8, there are altogether 8 variables added to the model. During the last step of the selection, all the variables have the test statistic that are less than the critical value. Therefore, there is no variable removed from the model. The above table carries out that the full model is the relative "best" model. Hence, we take the full model as the final model which including Size, N_FacilitiesInApt, N_Apt, Floor, N_Parking, N_Department, and N_Mall.

b. Final Regression Model

After we obtain the final model, we need to verify the fitness of the model to ensure the correctness of our recommendations.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	4.785E13	5.981E12	1887.3	<.0001
Error	5869	1.860E13	3168928325		
Corrected Total	5877	6.644E13			

(Table 11)

Table 11 is the ANOVA table generated from SAS. From the above table, the residual sum of square is significantly smaller than the total sum of square, which already gives us an insight into how the fit of the model. To make such a conclusion more formal in statistics, we constructed a hypothesis in testing the significance of the model.

We state the hypothesis as follow

$$H_0: \beta_i = 0, i = 1, \dots, 8. \quad vs \quad H_1: at\ least\ one\ \beta_i\ is\ not\ zero$$

From table 3, the p-value (<0.0001) which is less than 0.05 and we will reject H_0 at 5% level of significance. There is sufficient evidence to indicate that at least one β_i is not zero. In other words, we can say that the model is quite good at reflecting the behaviour of data.

c. Summary of Regression Model

The upcoming tables are the summary of the final regression model we adopted in our project.

Root MSE	56293	**R-Square**	0.7201
Dependent Mean	221416	**Adj R-Sq**	0.7197
Coeff Var	25.42414		

(Table 12)

Parameter Estimates								
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	1	-20616	3809.32566	-5.41	<.0001	0	-28083	-13148
Size	1	151.69431	2.26498	66.97	<.0001	1.38980	147.25412	156.13451
Floor	1	1459.74285	104.68281	13.94	<.0001	1.15868	1254.52600	1664.95971
N_Parking	1	92.94676	5.38144	17.27	<.0001	7.82068	82.39717	103.49636
N_FacilitiesInApt	1	23477	489.52387	47.96	<.0001	2.41405	22518	24437
N_APT	1	-13887	681.49858	-20.38	<.0001	6.81152	-15223	-12551
N_Dpartmentstore	1	45170	1935.06118	23.34	<.0001	4.56896	41377	48963
N_Mall	1	-51213	2786.14856	-18.38	<.0001	2.31974	-56674	-45751
N_Park	1	-65015	2508.50981	-25.92	<.0001	5.05811	-69932	-60097

(Table 13)

From the information of table 12, there is about 71.97% of the variation of data explained the final model, which is very good. We further construct the hypothesis testing for the importance of each parameter. We write the hypothesis below.

$$H_0: \beta_i = 0, i = 1, \dots, 8. \quad vs \quad H_1: \beta_i \neq 0$$

Clearly, we can see that for each parameter, their p-values are smaller than 0.05. It means that the null hypothesis will be rejected under a 5% level of significance. Simply put, each parameter is the significance of the final model.

Recommendations for New Buildings

Based on the findings of the statistical analysis method that we have found, we recommend the coming estate projects in the territories should fulfil the requirements below as they are associated with higher selling price.

From the internal factors:

1. **Constructing apartments with large size**

Flats with larger size will generally create a better living environment as it enables the users to have a higher flexibility in decorating their home and hence having a higher living standard

2. **Building estates with higher floor**

Apartments at higher floors can provide a good vision for the users. Moreover, they are often accompanied with lower air pollution level

3. **Increasing the number of parking spaces**

Parking spaces are valuable in the territories and they are attracting to those who are the car owners in the district

4. **Enlarging the number of facilities**

Providing more facilities surrounding can increase the convenience of life of residents living in the estates.

From the external factors:

1. **Building more department stores and malls**

Shopping convenience can further increase the diversity of shopping choice and hence increasing the living standard of residents

2. **Having more parks in the area**

Recreational facilities can provide rooms for the children to extend their social circles. Meanwhile, it also favours the elderly in doing rehabilitation exercises

3. **Increasing the number of blocks**

It can increase the number of apartments that we are going to sell. On the other hand, with the recreational facilities surrounding, it can provide opportunities to broaden their social circle.

Limitations and Improvement of Analysis Model

Back to the regression model, there is an unsatisfied issue that we found in the final model which is the multicollinearity problem. From table 5, there is a column of variance inflation factor (VIF) for each term to diagnose the multicollinearity problem. We can see that the VIF's are in between 1 to 10, which describes the issue exists but not severe in the model. Multicollinearity problem will lead to unstable of the mean square of error and the least square estimates will not be accurate under such situation. Due to such matter, we attempted to remove some of the terms with the highest value of VIF. However, we do not have sufficient information to make such a decision. Since we only have the record of data, we do not know how those data were collected. If variables are overcounted during the progress of collecting data, the given variables are highly correlated. For example, N_Department and N_mall are 2 variables that easily to have such problem. By intuition, we know that some department stores are located inside the malls. What important is that we lack knowledge about the given data set. There may be some aliased terms from the given data set, which will lead to perfect multicollinearity problem. Such statistical phenomenon makes the least square estimates sensitive over the model or data. Thus, we cannot completely rely on the given data set. Since the research question of our project is to discover which types of a variable will lead to an increment in the price, the actual predicted value of the price is less important to our purpose. As a result, we may ignore the resulting value of least square estimates, and hence we remained the 8 variables as the final model.

Another limitation on our analysis is that the project only focuses on the tangible factors. We ignored the intangible factors that are affecting the price. For instance, the time effect is one of the implicit components that may make a change to the price. There may exist a seasonal component during a year that is controlling the price of the building. In reviewing the further studies of a similar project, test for seasonality using Kruskal-Wallis test is essential to discover such the seasonal factor. Multiplicative time series model is also a favourable method to understand how the seasonal factor that is affecting the price.

We have not fully utilised the given data set during the project, such as Sold_Month is not used in the whole statistical analysis. There may have more possible results from the given data set. For future studies, more precise details of data collection are needed for model fitting. Many more other methods are also interested to try out and explore the behaviour of data.

Conclusion of the Report

To conclude the whole report that we have done, we have conducted the analysis using three major statistical methods: Summary Statistics, Correlations and Regression Analysis.

We have listed out the major important findings in respective pages. We sincerely believe that with the help of the data, we are able to reach a conscious that in which extent the provided dataset is useful in helping us to characterise the favourable factors and models that suit to an upcoming proposed constructing estates projects.

With a view of the findings that we can get, we have given suggestions on the proposed construction ideas of the buildings. Provided with some brief ideas on the advantages of ideal construction styles, it is believed that it can increase the potential selling price of the apartment in the estates and hence yields a higher return of profit to the estate development companies. Not only with this, but an ideal construction project that caters the needs of residents living inside can also provide convenience and comfort to them. In return, generating a higher rate of living stability and living standard and thus increase fame, not only the project estates but also the fame of the company. In the long run, the company can be benefited both from investment and also promotion.

Although our ideas seem to be good to launch, we still find out some limitations and improvements that can be done by modifying the models. Here, we will not talk about the detailed of what to change, but we targeted on why we need to improve our investigation models and methods. With a deeper insight into the dataset and models, we believe that the whole analysis can be thoroughly tackled. As a result, a detailed explanation of the changes in variables and models can be obtained. Lastly, reaching satisfying results which can be more precious and objective-oriented.