STAT3011 – Project I

Topic Concerned: Department Transaction Dataset

Project Members:	Tse Kit Ying	1155092382
	Chan King Fung	1155094725
	Suen Tsz Ki	1155095081
	Mo Jackie Ryan	1155095511
	Chan King Yeung	1155119394

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Background

- Working for a renowned estate developer in the territories
- Analyzing the apartment transactions in the district of an Asian city from the past 10 years
- Identifying potency of developing estate projects for the company
- Attracting capital deposit of investors

Mr. Li is planning to build new buildings

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Modification of Some Variables

We restructured the following explanatory variables as part of data cleaning

- Building_Age ≔ Year_Sold – Year_Built
- N_Parking ≔ N_Parking_G + N_Parking_B
- N_School ≔ N_Elementary + N_Middle + N_ High + N_University
- AvgTime ≔ merge TimeToSubway and TimeToBusStop to the average time
    - Converted from categorical variables to numerical variable

Reasons of Modification:

- Some details of original data set are unnecessary
- Try to make variables more reasonable and user-friendly

Types of variables

- Response Variable: Price
- Explanatory Variables:
  - Internal Factors of the Apartment
    - Building_Age, Month_Sold , Size , Floor, N_Parking, N_FacilitiesInApt
  - External Factors of the Department
    - AvgTime (average time to the nearest transportation)
    - No. of Nearby Infrastructure (Nearby buildings, Public office, Hospital, etc.)
    - No. of Educational Facilities (Elementary, Middle, High, University)

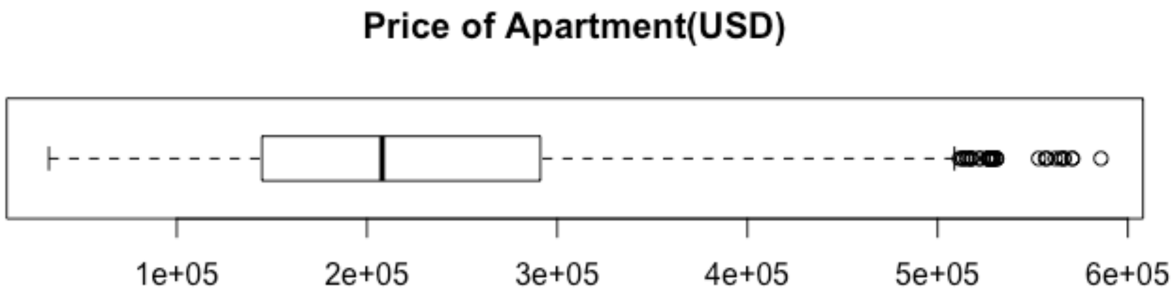
Response Variable: Price of Apartment

Basic statistics about the distribution of apartment prices:

- Mean: 221416.5
- Median: 207964

- S.D.: 106328.8
- IQR: 146398

- Min: 32743
- Max: 585840
- Q1: 144752
- Q3: 291150
- Min. Non-out: 32743*
- Max. Non-out: 508849^
- No. of Outliers: 35



* Min. Non-out = minimum Non-outlining value
^ Max. Non-out = maximum Non-outlining value

Explanatory Variable: Internal Factors (Building Age and Area)

Basic statistics about the distribution of the building age:

- Mean: 9.715890
- Median: 7

- S.D.: 8.545582
- IQR: 13

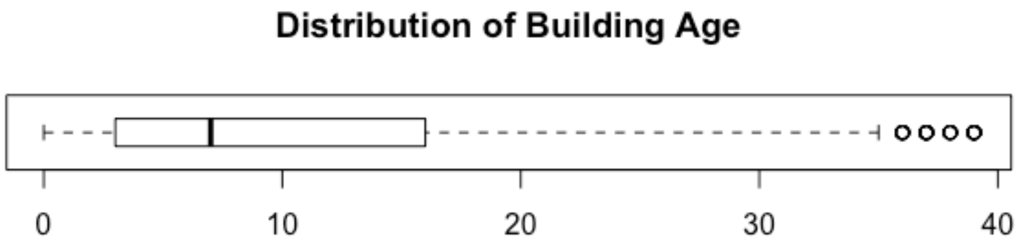
- Min: 0
- Max: 39
- Q1: 3
- Q3: 16
- Min. Non-out: 0
- Max. Non-out: 35
- No. of Outliers: 40

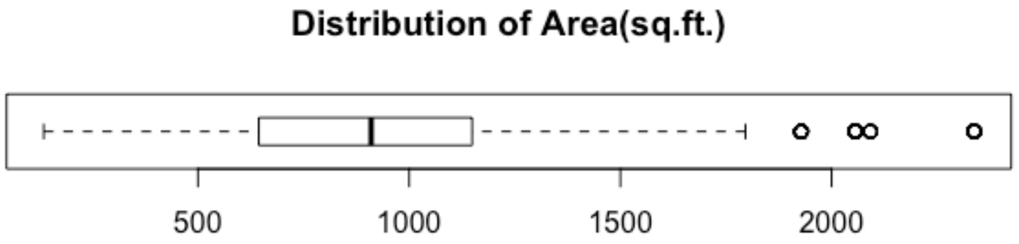
Basic statistics about the distribution of the area(sq.ft.):

- Mean: 955.6589
- Median: 910

- S.D.: 382.2002
- IQR: 505

- Min: 135
- Max: 2337
- Q1: 644
- Q3: 1149
- Min. Non-out: 135
- Max. Non-out: 1796
- No. of Outliers: 138





Explanatory Variable: Internal Factors (Floors and Parking Spaces)

Basic statistics about the distribution of floor:

- Mean: 12.036917
- Median: 11

- S.D.: 7.550668
- IQR: 11

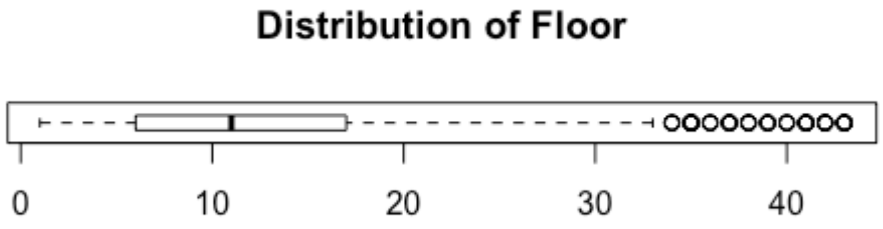
- Min: 1
- Max: 43
- Q1: 6
- Q3: 17
- Min. Non-out: 1
- Max. Non-out: 33
- No. of Outliers: 60

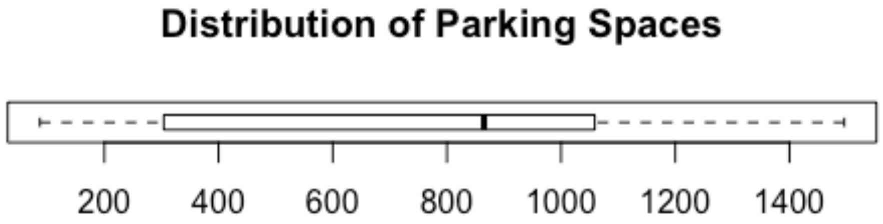
Basic statistics about the distribution of parking spaces:

- Mean: 766.9956
- Median: 865

- S.D.: 381.5948
- IQR: 755

- Min: 87
- Max: 1496
- Q1: 304
- Q3: 1059
- Min. Non-out: 87
- Max. Non-out: 1496
- No. Of Outliers: 0





Explanatory Variable: Internal Factors (N_APT and N_FacilitiesInApt)

Frequency Table for No. of Apartment Buildings in the Apartment Complex

N_APT	Frequency	Percentage Frequency
1	618	10.514%
2	368	6.261%
3	1139	19.377%
4	0	0%
5	64	1.089%
6	592	10.071%
7	1375	23.392%
8	1408	23.954%
9	0	0%
10	203	3.454%
11	0	0%
12	0	0%
13	111	1.888%
Total	5878	100%

Basic statistics about the distribution of N_APT

- Mean: 5.615005
- Median: 7

- S.D.: 2.812130
- IQR: 5

- Q1: 3
- Q3: 8
- Min. Non-out: 1
- Max. Non-out: 13
- No. Of Outliers: 0

Frequency Table for the No. of Facilities in the Apartment Complex

N_Facilities InApt	Frequency	Percentage Frequency
1	55	0.936%
2	69	1.174%
3	672	11.432%
4	1442	24.532%
5	1158	19.701%
6	0	0.000%
7	1225	20.840%
8	270	4.593%
9	203	3.454%
10	784	13.338%
Total	5878	100.000%

Basic statistics about the distribution of N_FacilitiesInApt:

- Mean: 5.813032
- Median: 5

- S.D.: 2.330653
- IQR: 3

- Q1: 4
- Q3: 7
- Min. Non-out: 1
- Max. Non-out: 10
- No. Of Outliers: 0

Explanatory Variable: Internal Factors (Month Sold) and External Factors (AvgTime)

Month which the apartments are sold:

Month Sold	Frequency	Percentage Frequency
January	623	10.599%
February	424	7.213%
March	576	9.799%
April	450	7.656%
May	606	10.310%
June	513	8.727%
July	550	9.357%
August	448	7.622%
September	387	6.584%
October	519	8.830%
November	412	7.009%
December	370	6.295%
Total	5878	100.000%

Frequency Table for the avearge time to the nearest transportation (min)

Average Time to the Nearest Transportation (min)	Frequency	Percentage Frequency
2.5	2564	43.620%
5	1398	23.784%
7.5	409	6.958%
10	1230	20.925%
12.5	277	4.712%
Total	5878	100.000%

Remarks for AvgTime:

1. The numerical values are taken from the class marks of the categorical variables. (e.g "0-5min" -> 2.5 min)
2. The numerical values above are the mean of the time to subway and bus stops.
3. If TimeToBusStop = "No_bus_stop_nearby", then only the time to the subway is considered.

Basic statistics about the distribution of AvgTime:

- Mean: 5.483158
- Median: 5

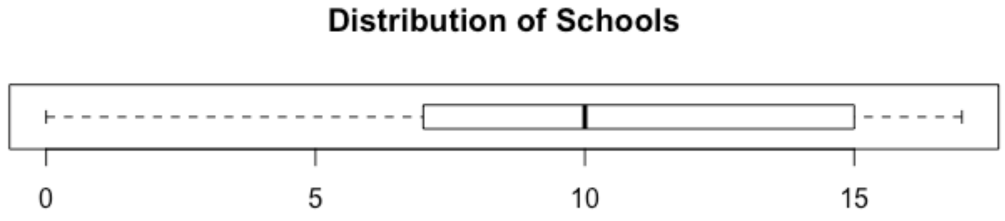
- S.D.: 3.288114
- IQR: 7.5

- Q1: 2.5
- Q3: 10
- Min. Non-out: 2.5
- Max. Non-out: 12.5
- No. Of Outliers: 0

Explanatory Variable: External Factors (No. of Educational Facilities)

Frequency Table for the no. of Educational Facilities

No. of Schools	Frequency	Percentage Frequency
0	64	1.089%
1	55	0.936%
2	0	0%
3	0	0%
4	63	1.072%
5	652	11.092%
6	298	5.070%
7	417	7.094%
8	381	6.482%
9	470	7.996%
10	609	10.361%
11	714	12.147%
12	0	0%
13	78	1.327%
14	162	2.756%
15	737	12.538%
16	0	0%
17	1178	20.041%
Total	5878	100.000%



Basic statistics about the distribution of no. of Educational Facilities

- Mean: 10.864069
- Median: 10
- S.D.: 4.437078
- IQR: 8

- Q1: 7
- Q3: 15
- Min. Non-out: 0
- Max. Non-out: 17
- No. Of Outliers: 0

Explanatory Variable: External Factors
(No. of Public Offices and No. of ETC)

Frequency Table for
No. of Public Offices:

No. of Public Offices	Frequency	Percentage Frequency
0	64	1.089%
1	413	7.026%
2	750	12.759%
3	1194	20.313%
4	358	6.091%
5	1806	30.725%
6	671	11.415%
7	622	10.582%
Total	5878	100.000%

Basic statistics about the distribution of No. of Public Offices:

- ◉ Mean: 4.140354
- ◉ Median: 5

- ◉ S.D.: 1.793642
- ◉ IQR: 2

- ◉ Q1: 3
- ◉ Q3: 5
- ◉ Min. Non-out: 0
- ◉ Max. Non-out: 7
- ◉ No. Of Outliers: 0

Frequency Table for the No. of
Facilities like Hotels and Special Schools:

No. of ETC	Frequency	Percentage Frequency
0	2593	44.114%
1	908	15.447%
2	462	7.860%
3	0	0%
4	0	0%
5	1915	32.579%
Total	5878	100.000%

Basic statistics about the distribution of No. of ETC:

- ❖ Mean: 1.940626
- ❖ Median: 1

- ❖ S.D.: 2.201917
- ❖ IQR: 5

- ❖ Q1: 0
- ❖ Q3: 5
- ❖ Min. Non-out: 0
- ❖ Max. Non-out: 5
- ❖ No. Of Outliers: 0

Explanatory Variable: External Factors
(No. of Hospitals and No. of Department Stores)

Frequency Table for No. of Hospitals:

No. of Hospitals	Frequency	Percentage Frequency
0	64	1.089%
1	4009	68.203%
2	1805	30.708%
Total	5878	100.000%

Basic statistics about the distribution of No. of Hospitals:

◉ Mean: 1.2961892
◉ S.D.: 0.4798713

Frequency Table for No. of Department Stores:

No. of Department Stores	Frequency	Percentage Frequency
0	2270	38.619%
1	1948	33.141%
2	1660	28.241%
Total	5878	100.000%

Basic statistics about the distribution of No. of Department Stores:

◉ Mean: 0.8962232
◉ S.D.: 0.8111332

Explanatory Variable: External Factors (No. of Shopping Malls and No. of Parks)

Frequency Table for No. of Shopping Malls:

No. of Shopping Malls	Frequency	Percentage Frequency
0	655	11.143%
1	4911	83.549%
2	312	5.308%
Total	5878	100.000%

Basic statistics about the distribution of No. of Shopping Malls:

- Mean: 0.9416468
- S.D.: 0.4014151

Frequency Table for No. of Parks :

No. of Parks	Frequency	Percentage Frequency
0	2640	44.913%
1	2629	44.726%
2	609	10.361%
Total	5878	100.000%

Basic statistics about the distribution of No. of Parks:

- Mean: 0.6544743
- S.D.: 0.6583500

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Internal Factors of the Apartment

Correlation between Price and Building_Age

Pearson Correlation Coefficients, N = 5878
Prob > |r| under H0: Rho=0

	Building_Age
Price Price	-0.33931 <.0001

- Based on the hypothesis testing of H0: ρ=0, we calculated the p-value is <0.0001.

- Building age have a negative correlation with price.(-0.33931)

- We suggest selling the apartment as soon as possible

Correlation between Price and Size

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

- We suggest building the apartment with a bigger size.

Correlation between Price and Floor

Pearson Correlation Coefficients, N = 5878
Prob > |r| under H0: Rho=0

	Floor
Price Price	0.33622 <.0001

- The higher floor it located, the higher price it has.

Correlation between Price and Facilities

Pearson Correlation Coefficients, N = 5878
Prob > |r| under H0: Rho=0

	N_Parking	N_FacilitiesinAPT
Price Price	0.43141 <.0001	0.50472 <.0001

•Number of parking spaces and facilities for residents like swimming pool, gym, playground have a positive correlation with price.

•We should build more facilities in order to raise the apartment's price.

External Factors of the Apartment

Correlation between Price and AvgTime

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

- The most negatively correlated variable with price.

- The closer the apartment is to the bus stop and subway the better.

Correlation between
Price and Nearby infrastructure with negative correlation

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

- Should consider building relatively less public offices and ETC nearby the apartment.

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

Correlation between Price and Nearby infrastructure with positive correlation

- Building malls have the weakest impact on price

- Should consider building mostly parks then any other facilities.

Pearson Correlation Coefficients, N = 5878
Prob > |r| under H0: Rho=0

	School
Price Price	-0.37857 <.0001

Correlation between Price and Schools

- It has a relatively strong negative correlation with price.
- We should prevent building any schools in the area

Potential Factors Affecting Price

- Things to remind for new buildings
  - Locate at somewhere near the subway station and bus stop (0~5 minutes)

- Things to remind for surrounding facilities
  - Build more parks, department store, malls and buildings
  - To obtain a higher sale price

Potential Factors Affecting Price

Positive

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

Negative

- Building_Age
- N_PublicOffice
- N_Hospital
- N_ETC
- School
- AvgTime

Potential Factors Affecting Price

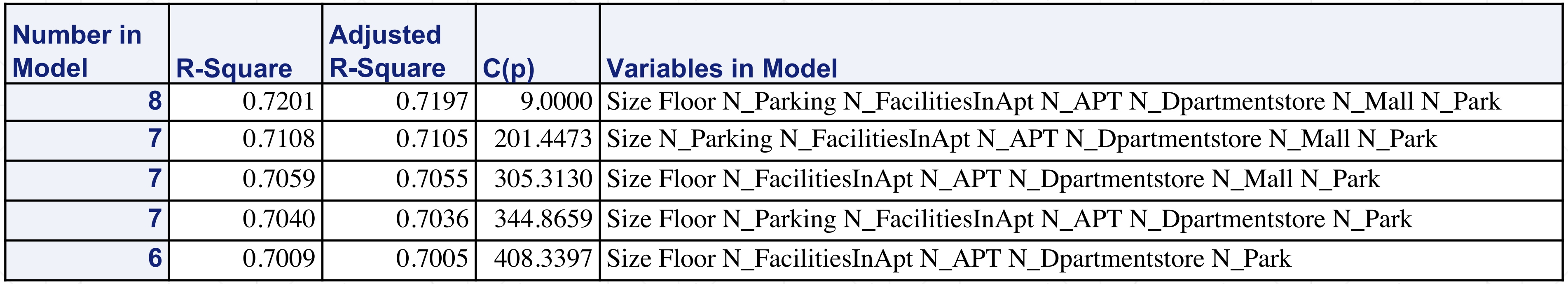
Variable Interested							
Size	Floor	N_Parking	N_FacilitiesinApt	N_APT	N_Department	N_Mall	N_Park

- We are only interested in tangible factors that positively correlated with Price

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Variable Selection

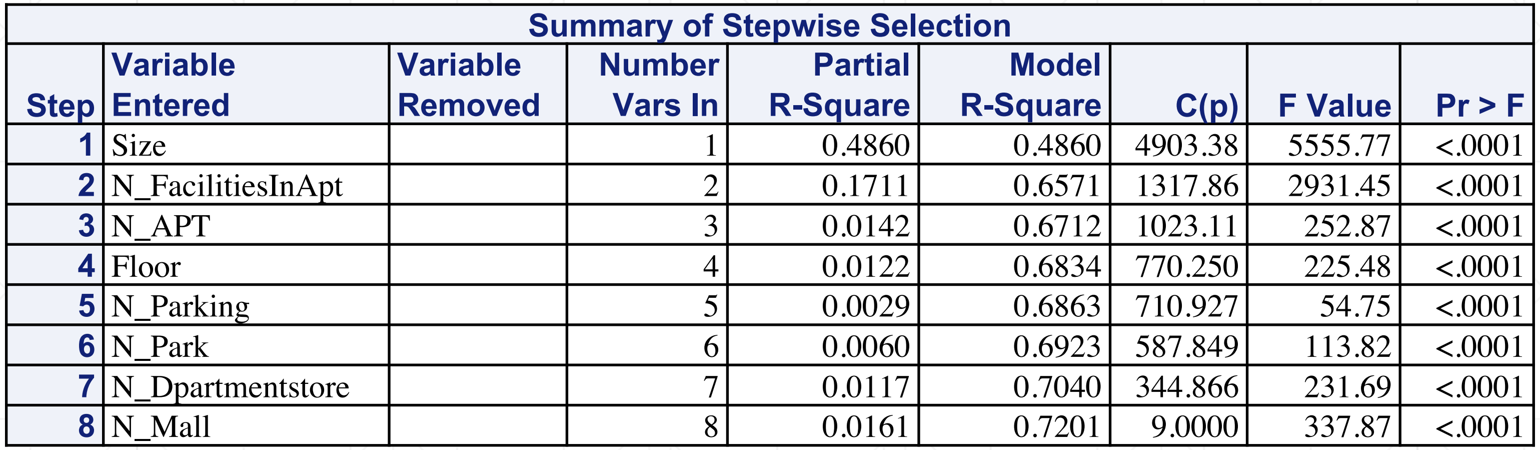


Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
8	0.7201	0.7197	9.0000	Size Floor N_Parking N_FacilitiesInApt N_APT N_Dpartmentstore N_Mall N_Park
7	0.7108	0.7105	201.4473	Size N_Parking N_FacilitiesInApt N_APT N_Dpartmentstore N_Mall N_Park
7	0.7059	0.7055	305.3130	Size Floor N_FacilitiesInApt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

- Adjusted Coefficient of Determination
  - List all the possible variable selection
  - Criteria: highest adjected R-square
- Results: select the full model

$$Price = Size + N\_FacilitiesInApt + N\_Apt + Floor + N\_Parking + N\_Park + N\_Dpartment + N\_Mall$$

Variable Selection



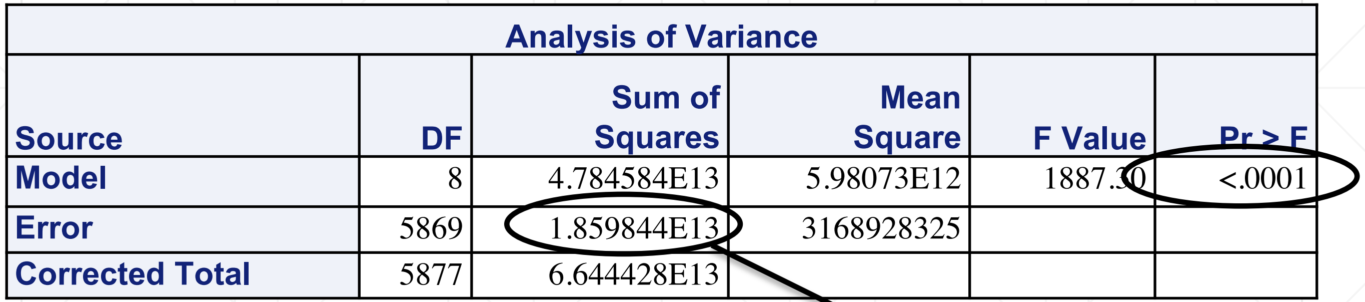
Summary of Stepwise Selection

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

- Stepwise Selection
  - Partial F-test to select variables
  - Select the relative ‘best’ regression model
- Results: select the full model

$$\boxed{Price = Size + N\_FacilitiesInApt + N\_Apt + Floor + N\_Parking + N\_Park + N\_Dpartment + N\_Mall}$$

ANOVA Analysis



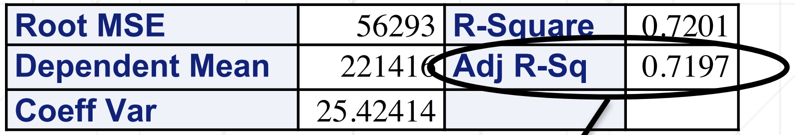
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	4.784584E13	5.98073E12	1887.30	<.0001
Error	5869	1.859844E13	3168928325		
Corrected Total	5877	6.644428E13			

GOOD IN EXPLAINING VARIATION

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

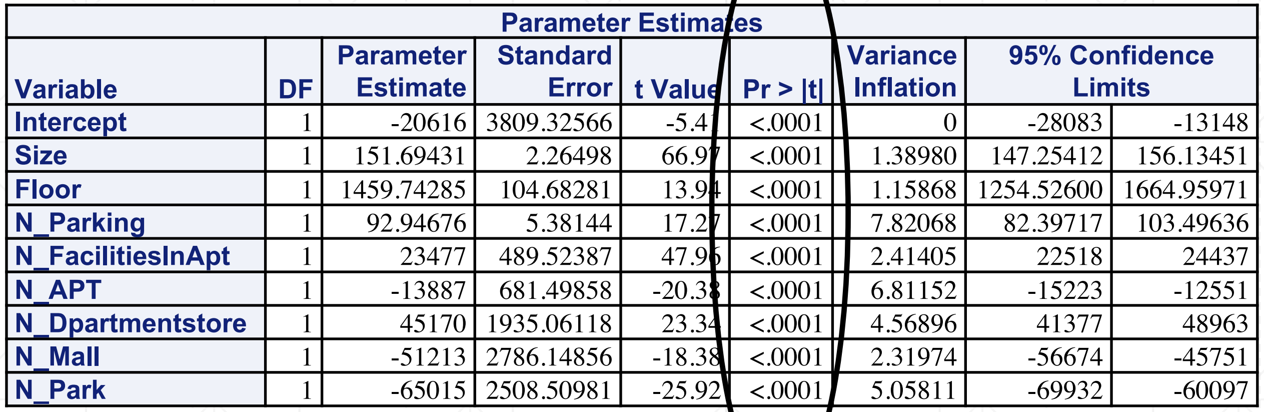
- The p-value is smaller than 0.05
- Reject H_0 at the 5% level of significance
- The regression model is significant

Summary of Model



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

71.97% of variation explained by model



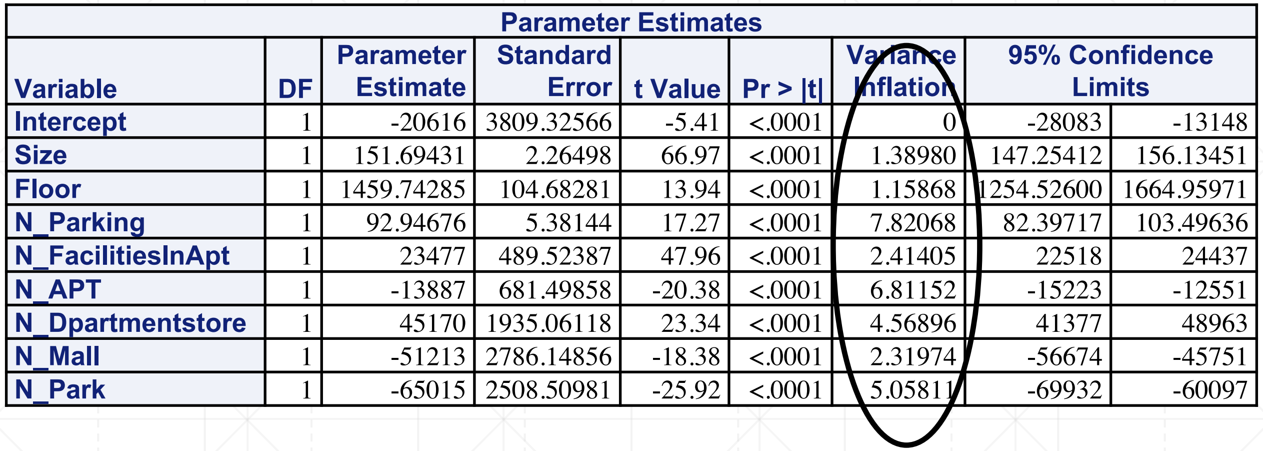
Parameter Estimates								
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	1	-20616	3809.32566	-5.41	<.0001	0	-28083	-13148
Size	1	151.69431	2.26498	66.97	<.0001	1.38980	147.25412	156.13451
Floor	1	1459.74285	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

Each β_i is significant (p-value<0.05)

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Issue in the Fitted Regression Model



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

- Multicollinearity problem exists
  - But does not severe
  - Try to reduce aliased terms
- Insufficient details about variables
  - i.e., department stores and mall

Limitation on Analysis

- Ignored intangible factors
  - i.e., time effects
  - May have seasonal component during a year

- Further study on the data set
  - Test for seasonality using Kruskal-Wallis test
  - Discover seasonality factor using multiplicative time series model

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Interpretation of the Regression model

- Regression model

$$y_{Price} = -20616 + 151.69x_{size} + 1459.75x_{floor} + 92.95x_{parking} + 23477x_{facilities} \\ -13887x_{apt} + 45170x_{department} - 51213x_{mall} - 65015x_{park}$$

- A building is not valuable if x_i's are defined as 0
  - No one will buy a house in desert island

- For each increase in unit of Size, Floor, Parking, Facilities, or department
  - Price will increase corresponding to its parameter
  - Holding other as a constant

Suggestion for new Buildings

Internal Factors

- Extend the size of each flat
  - A larger size generally provide a higher living standard for the buyers
- Build a taller building
  - A good vision from high-floor department can be valuable
- Increase number of parking space
  - Parking space is valuable in the district
- Enlarge the number of facilities
  - With increasing number of facilities surrounding, we can provide a comfortable living environment for the residents

External Factors

- Building more department stores and malls surround the building
  - Shopping convenience
- Having more parks in the area
  - Children's favorite place to have physical activity
- Increase the number of apartment buildings in the apartment complex
  - Provide opportunities to expand social circles

Process of the Presentation

1. Background
2. Summary of Variables
3. Potential Factors Affecting Price
4. Regression Analysis
5. Limitation in the Analysis and Improvement
6. Recommendations for New Buildings
7. Conclusion

Conclusion

- What we have discussed so far:
  - The final model we fitted from the data
  - Limitation in our models
  - Recommendation we made for new buildings
  - Further improvement to be made in the future

- We sincerely hope that this report and presentation can help investigate the characteristics of the district and provide additional ideas for the company

Thank you!