Intermediate Assignment: Spreadsheet & Statistics

By: Biyan Bahtiar Ramadhan

1. Business Understanding

- As a property listing company, we want to help property owner to sell as many property as they can to tenants.
- Although we gain most of our profit from selling high-priced property, high-priced property is hard to sell, due to high room counts or high room size.
- Thus, we want to know which is considered luxury property, high room counts and high room size; and affordable property.

1. Business Understanding

- We can use statistical method on the data to know which property is considered high-priced, high room counts and high room size.
- Data can be further categorized based on property type and property character

1. Business Understanding

- Statistical method to know which property is considered high-priced is mean, median or quartile 3.
- Choosing between mean or median depends on data distribution, if data distribution is skewed, I will use median instead of mean.

2. Data Cleaning: Handling Missing Data

Steps	Reason	Notes
Remove extra whitespace and remove duplicates	Each property listed should be unique.	Remove Extra Whitespace: Select all data>Data>Data cleanup>Trim Whitespace Remove Duplicates: Select all data>Data>Data Cleanup>Remove Duplicates
Add filter	Need filter to further clean the data and select subgroups	Select all data>Data>Filter

2. Data Cleaning: Handling Missing Data

Steps	Reason	Notes
Check missing data in price column	Price is the most important column based on our business problem. We can't have 0 or missing property price. Deleting missing price data also will ensure our data is clean of assumption in price.	Delete rows that contains missing data in Price column
Check missing data in Size column	Size is important variable to consider	Delete rows that has missing data in Size column plus either Property Type or Property Character, since without these information I won't be able to impute. I found that all missing data in Size also have missing data in Property Character.
Check missing data in Location, property Type, Property Character and Furnishing.	Before I can impute missing data in Rooms column, I have to make sure the data on these columns are filled	Delete rows that has no data in Location, Property Type and Property Character. No missing data in these columns. Missing data in Furnishing is imputed with Unknown.

2. Data Cleaning: Handling Missing Data

Steps	Reason	Notes
Check missing data in Rooms column	Rooms is important column	- Is Size column have no data? Yes: delete row - Is there any row with similar characteristics in location, price and size? Yes: impute Rooms with the value found there - Does the data make sense? No: delete

2. Data Cleaning: Handling Incorrect Data and Data Type

Steps	Reason	Notes
Remove ‘RM’ in Price column and change the data type to Malaysia Ringgit	Changing data type for easier analysis	Using Find and Replace. Change data type in Format>Number>Custom Currency>Malaysia Ringgit
Change all data in Rooms column to numeric	Changing data type for consistent and easier analysis	Studio change to 1 since it’s the same
Change the value in Property Type column to property type without additional attribute like corner,intermediate etc.	It is assumed that tenants are not interested whether the property is intermediate etc. and for easier analysis	
Remove “:” in Property Character	Convenience	
Change “sq.m” value in Property Character to Unknown	Easier analysis and there is already “Unknown” category	

2. Data Cleaning: Handling Incorrect Data and Data Type

Steps	Reason	Notes
Clean Size column by imputing it with correct data type	Changing data type for easier and consistent analysis	Most data in Size is numeric with sqft measurement, hence the data are cleaned to match that type • SPLIT function to split numbers and unit measurement • Convert acres and sq.m. to sqft • Change value 'Kuala Lumpur' to blank since numericvalue can't be imputed with Unknown • Range value in cell is imputed with its mean • Arithmetic equation is imputed with its solution • HTML character code is converted and its imputed with its arithmetic solution

2. Data Cleaning: Handling Incorrect Data and Data Type

Steps	Reason	Notes
Delete price below RM 100,000	Doesn't make sense to have property price 1,000 RM etc.	
Blank values in Size column is imputed with either median/mean based on data distribution	Numeric data can only be imputed with either mean or median	The decision is made later after we see the descriptive statistics

3. Descriptive Statistics: Variables Checked and The Reason

1. Price
    - Based on our business problem, company wants to separate between high-priced property and non high-priced property, since high-priced one are harder to sell.
2. Rooms
    - High-priced property have high room counts and high room size
3. Size

3. Descriptive Statistics

Spreadsheet:
docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013

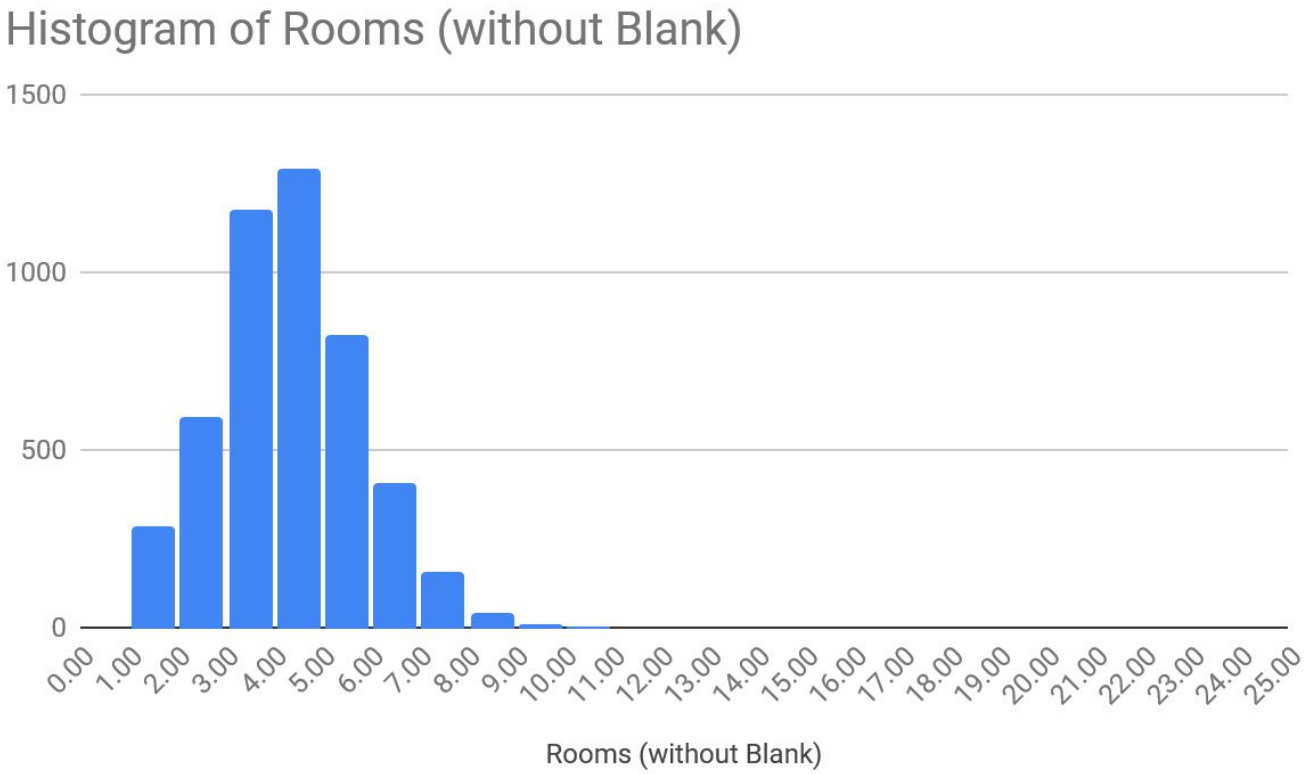
	Price	Rooms (No Blanks)	Size (No Blanks)
Mean	2152003.433	3.834650035	2835.3461
Standard Error	49148.31614	0.02228316979	184.4548843
Median	1300000	4	1650
Standard Deviation	3429834.178	1.545430601	12747.42069
Skewness	15.15755261	0.6559351313	51.45452712
Range	129895000	19	789696
Minimum	105000	1	304
Maximum	130000000	20	790000
Count	4870	4810	4776

3. Descriptive Statistics

Spreadsheet:
docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013

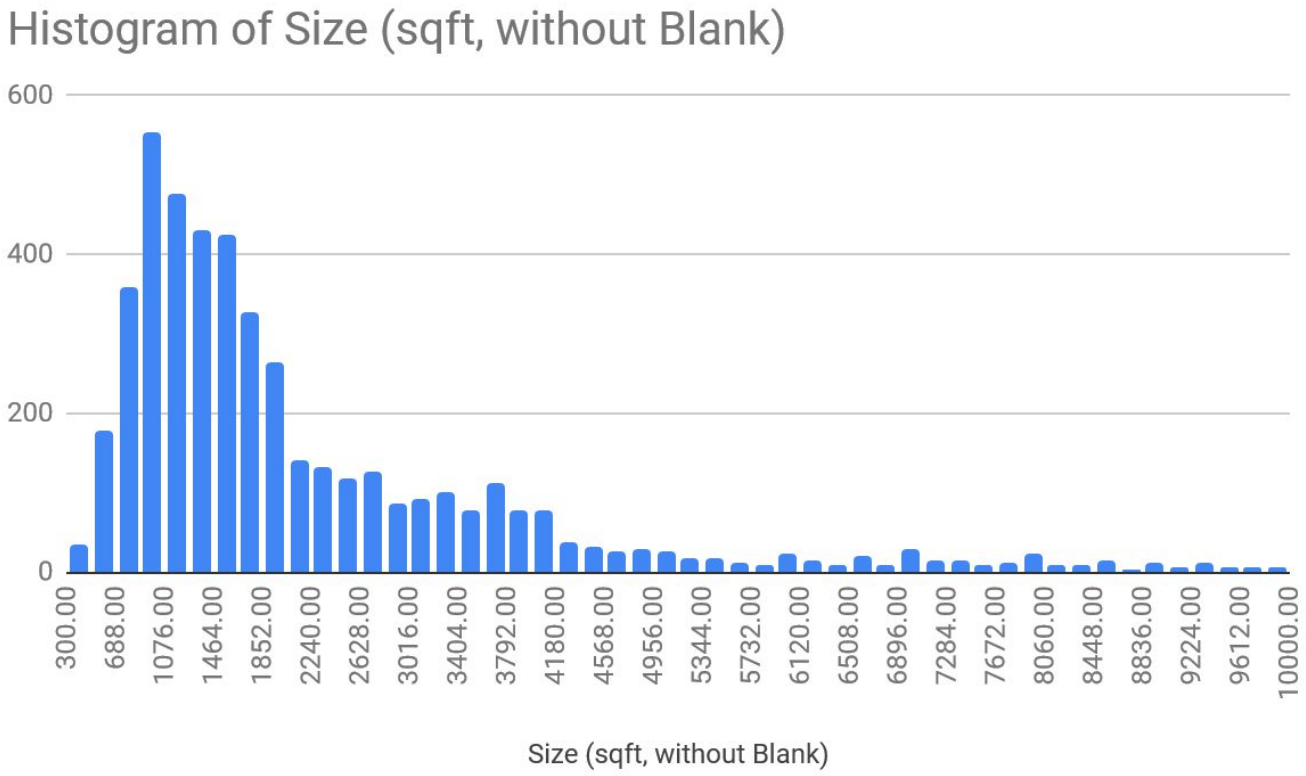
3. Descriptive Statistics

Spreadsheet:
docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013



3. Descriptive Statistics

Spreadsheet:
docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013



2. Data Cleaning: Handling Outliers and Impute Blank

Steps	Reason	Notes
Change blank value in Size Column to median/percentile 0.5 since the column has high	The column has high positive skew	
Determine whether Price, Room and Size column cell is an outlier	Outlier can interrupt our analysis	Use IF function and put the result column (contains Yes/No) next to respective column
Decision to remove an outlier is based on the proximity to the corresponding upper fence and based on the other variable	Outlier can also be valuable if it's meaningful to our business problem	• Bungalow Land and Residential Land seems to make the most outlier data in price and size and it's normal • On closer look, outlier price seems to be determined by property type, since many bungalow, serviced residence and 3-sty Terrace/Link house, *let's subset property type and see their descriptive statistics*

Supporting Table

	Price						
	Bungalow	Serviced Residence	3-sty Terrace/Link house	3.5-sty Terrace/Link house	Price(4-sty Terrace/Link house)	Price(Semi-Detached House)	Condominium
Percentile 0.25	RM4,200,000	RM590,000	RM1,280,000	RM3,600,000	RM3,741,781	RM2,325,000	RM696,500
Percentile 0.75	RM9,000,000	RM1,789,500	RM2,525,000	RM5,500,000	RM4,988,888	RM3,745,000	RM1,900,000
Lower fence	-RM3,000,000	-RM1,209,250	-RM587,500	RM750,000	RM1,871,119	RM195,000	-RM1,108,750
Upper fence	RM16,200,000	RM3,588,750	RM4,392,500	RM8,350,000	RM6,859,549	RM5,875,000	RM3,705,250

	Rooms
	Bungalow
Percentile 0.25	6
Percentile 0.75	7
Lower fence	5
Upper fence	9

	Size(sqft)					
	3-sty Terrace/Link House	3.5-sty Terrace/Link House	Bungalow	Condominium	Semi-detached House	Serviced Residence
Percentile 0.25	1650	2150	6000	1207	3200	838
Percentile 0.75	3211	5175	9500	2513	4296.5	1428
Lower fence	-692	-2388	750	-752	1555	-47
Upper fence	5553	9713	14750	4472	5941	2313

2. Data Cleaning: Handling Outliers and Impute Blank

Steps	Reason	Notes
(Cont'd) Decision to remove an outlier is based on the proximity to the corresponding upper fence and based on the other variable	Outlier can also be valuable if it's meaningful to our business problem	• For Bungalow, 3.5&4-sty Terrace/Link house and Semi-Detached House Price is higher than overall data, especially Bungalow and 3.5-sty terrace/Link house • Delete outlier in each subgroup
I ignore outliers and blank in land type	Land type can be built many buildings.	• Rooms outliers are comprised of mostly by land type and bungalow. • I decide to delete all of the outliers in Room column, except land type and Bungalow below 10 rooms • Delete outlier in each subgroup
I ignore outlier in Size column that has land property type	Bungalow Land and Residential Land Property Type typically have huge size	

End of Milestone 1

4. Exploratory Data Analysis

A. Characteristic of Luxury Property (Q3 - Q4)

	Price
Min	RM105,000
Q1	RM690,000
Q2	RM1,250,000
Q3	RM2,300,000
Max	RM130,000,000

Price (In Quartile)	MEDIAN of Price
Q0-Q1	RM500,000
Q1-Q2	RM950,000
Q2-Q3	RM1,680,000
Q3-Q4	RM3,500,000
Grand Total	**RM1,250,000**

Price (In Quartile)	MEDIAN of Rooms
Q0-Q1	3
Q1-Q2	3
Q2-Q3	4
Q3-Q4	5
Grand Total	**4**

Price (In Quartile)	MEDIAN of Size (sqft)
Q0-Q1	1,003.00
Q1-Q2	1,389.00
Q2-Q3	1,870.00
Q3-Q4	3,713.50
Grand Total	**1,620.00**

INSIGHTS

1.) Median room count is 5
2.) Median bathroom count is 5
3.) Median carpark count is 2
4.) Price range is RM2,300,000 - RM130,000,000, with median RM 3,500,000
5.) Top 3 most common property type are condominium, bungalow and serviced residence
6.) Median size is 3713.5 sqft

LINK

4. Exploratory Data Analysis

B. Characteristic of Affordable Property (Q3 - Q4)

	Price
Min	RM105,000
Q1	RM690,000
Q2	RM1,250,000
Q3	RM2,300,000
Max	RM130,000,000

Price (In Quartile)	MEDIAN of Price
Q0-Q1	RM500,000
Q1-Q2	RM950,000
Q2-Q3	RM1,680,000
Q3-Q4	RM3,500,000
Grand Total	**RM1,250,000**

Price (In Quartile)	MEDIAN of Rooms
Q0-Q1	3
Q1-Q2	3
Q2-Q3	4
Q3-Q4	5
Grand Total	**4**

Price (In Quartile)	MEDIAN of Size (sqft)
Q0-Q1	1,003.00
Q1-Q2	1,389.00
Q2-Q3	1,870.00
Q3-Q4	3,713.50
Grand Total	**1,620.00**

INSIGHTS:

1.) Median room count is 3
2.) Median bathroom count is 2
3.) Median carpark count is 2
4.) Price range is RM690,000 - 1,250,000, wih median RM950,000
5.) Top 3 most common property type are condominium, serviced residence and 2-sty Terrace/Link House
6.) Median size is 1389 sqft

LINK

4. Exploratory Data Analysis

D. Recommendation based on 4A and 4B

1.) The company should focus on marketing the affordable property, that is property within price range RM690,000 - RM1,250,000
2.) The target market for affordable property are family or group with size of 3-4.
3.) For busy working customer or customer who wants full service of their family and housing needs, they can be directed to choose either condominium or serviced residence.
4.) For target customer who is senior or want to invest in landed property, they can be directed to choose 2-sty Terrace/Link House

What to do next? (Based on the recommendation, what are the next plan of action that we can suggest to our company?)

1.) Find ways to promote affordable property in our listings to our target customers
3.) Do A/B testing for each strategy

5. Statistical Measurement

docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013

A. What aspect has the strongest relation to the price of the property?

	Price	*Rooms*	*Bathrooms*	*Car Parks*	*Size*
Price	1				
Rooms	0.7231963489	1			
Bathrooms	0.7724361186	0.8240724917	1		
Car Parks	0.6321176156	0.6276742122	0.6477339345	1	
Size	0.7959366558	0.6676861953	0.7106245598	0.5295760378	1

t-critical	17.39280081
p-value	0
Conclusion	Very Strong +ve Correlation

INSIGHTS:

- Size has the strongest positive correlation to price and is statistically significant

5. Statistical Measurement

docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013

B. What aspect has the biggest impact to the price of the property?

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*	*Standardized Coefficients*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	
Bathrooms	190018.9719	36723.18234	5.174360167	0.0000006251547035	117538.7427	262499.201	117538.7427	262499.201	0.3229525899
Car Parks	139184.7749	45717.33181	3.044464088	0.002693119403	48952.87028	229416.6795	48952.87028	229416.6795	0.1340817948
Size	469.5544291	55.87038742	8.40435248	0	359.2835259	579.8253323	359.2835259	579.8253323	0.4570105067

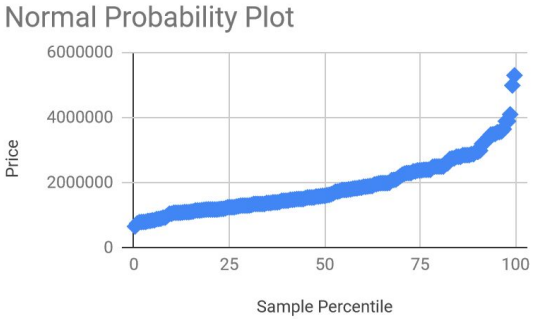
Adjusted R Square	0.954315528

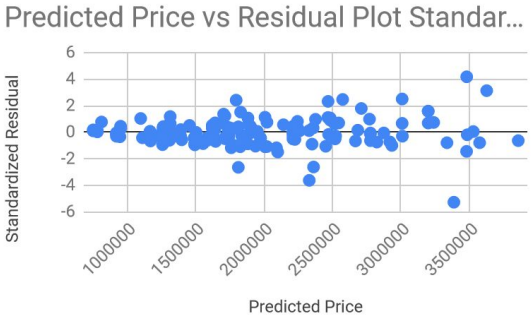
INSIGHTS:

- After 2 iterations, I found that size has the biggest impact on the price of property (1st iteration: remove Room column since p>0.05, 2nd iteration: remove intercept).
- The model fit is very good as it has r-squared 0.95

5. Statistical Measurement - Assumption Check

	Price	*Bathrooms*	*Car Parks*	*Size*
Price	1			
Bathrooms	0.7724361 186	1		
Car Parks	0.6321176 156	0.6477339 345	1	
Size	0.7959366 558	0.7106245 598	0.5295760 378	1





INSIGHTS

1.)No strong multicolinearity between independent variable (r<= -0.8 or r>= 0.8)

2.)Error distribution is not exactly normal, it seems to be right-tailed

3.)Auto-correlation is not checked because the data is cross-sectional, not time series

4.)The data does not fit the heteroscedasticity assumptions

docs.google.com/spreadsheets/d/1v4rkRcsCssJxSz5PzH9-g_rhOBSXfDpY7_WmoVxt3gU/edit#gid=1916302013

5. Price Offer for Customer and Recommendation

One user is looking for a property with 3 rooms, 4 bathroom, 3 car park and 2200 sqft.

Our price offer follows this formula:

*Price = 190018.971871943*Bathrooms+139184.7749*CarParks+469.554429085987*Size*

D. Recommendation

- From data available, company can offer **RM2,210,650**

End of Intermediate Assignment