Statistics and spreadsheet study case

PROPERTY LISTING COMPANY IN MALAYSIA

Ismatul Maula

TABLE OVERVIEW

BUSINESS UNDERSTANDING

Core business problem, Data related to

the business problem, How statistics can help

DATA CLEANING AND REMOVAL OUTLIERS

Steps in analyzing the dataset

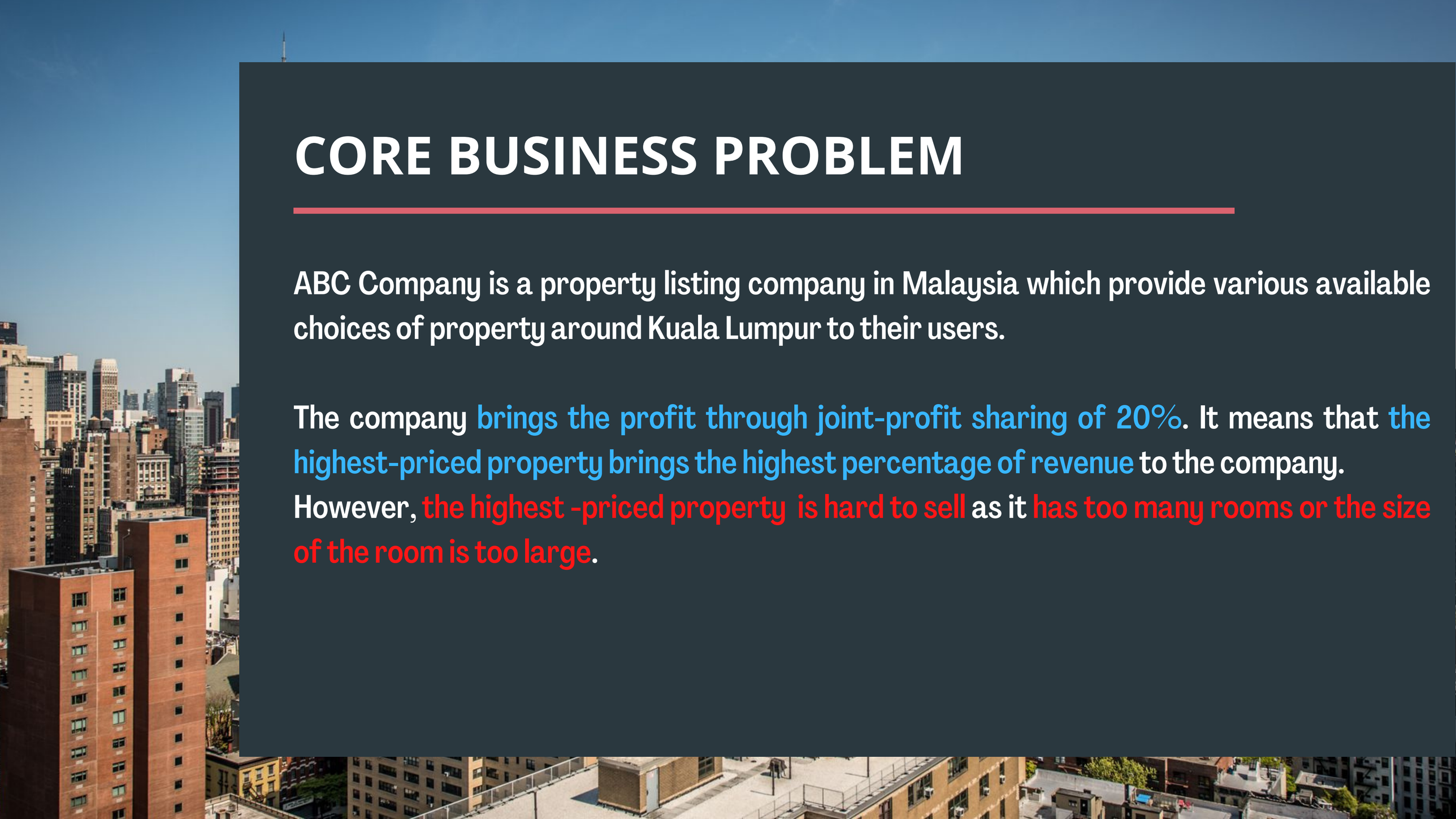
STATISTICS

Descriptive statistics, Exploratory Data Analysis,

Correlation and Regression

INSIGHT AND RECOMMENDATION

Giving recommendation after doing statistical measurement



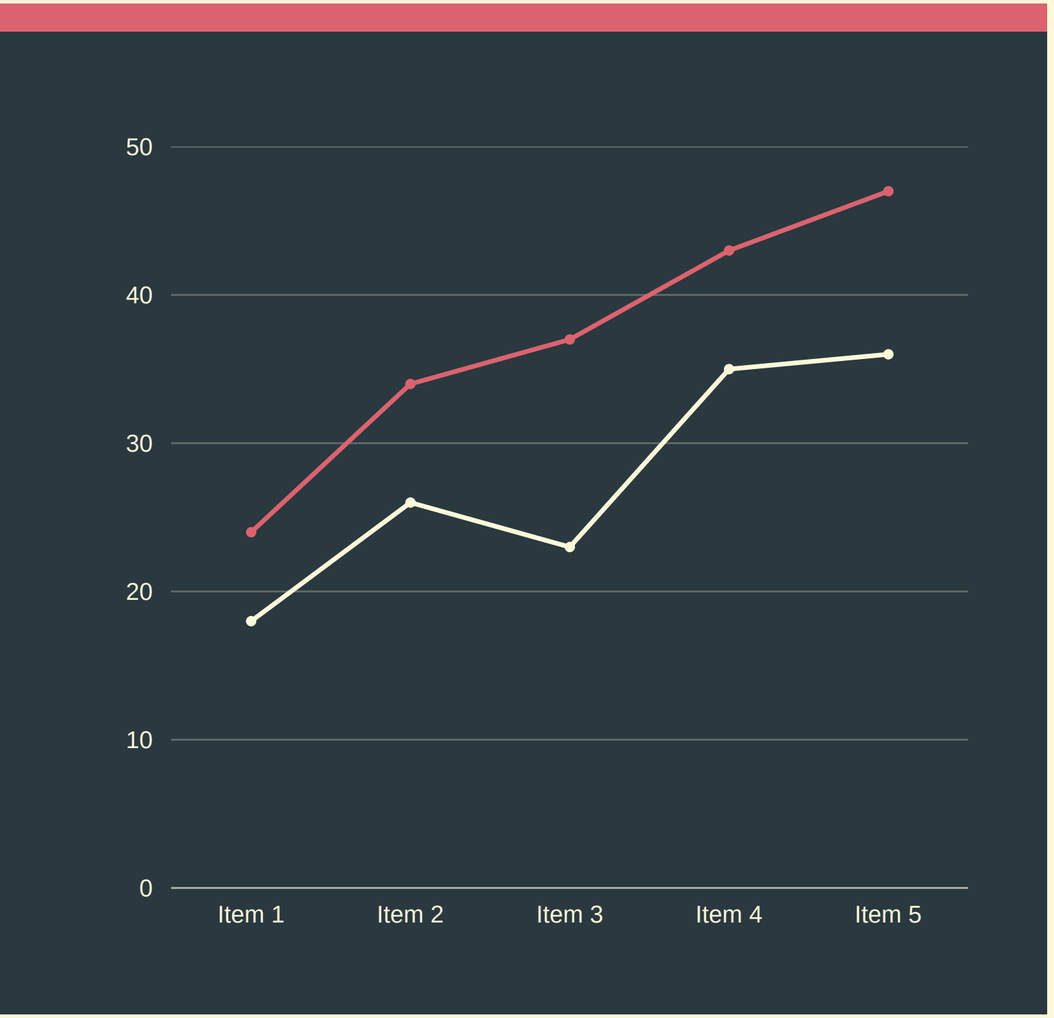
CORE BUSINESS PROBLEM

ABC Company is a property listing company in Malaysia which provide various available choices of property around Kuala Lumpur to their users.

The company brings the profit through joint-profit sharing of 20%. It means that the highest-priced property brings the highest percentage of revenue to the company. However, the highest -priced property is hard to sell as it has too many rooms or the size of the room is too large.

DATASET OVERVIEW

>>> Link to Dataset <<<



The data set is mainly consist of 5000 rows property listing company. The variables are:

- Location
- Price
- Room
- Bathroom
- Car Park
- Property Type
- Property Character
- Size
- Unit Area
- Furnishing

In order to answer the business problem, we analyze some variables that consist of numerical data which are **Price, Room, Bathroom, Car Park and Size** to get insight. Furthermore we also need to analyze categorical data type such as **location, property character and furnishing** to give another recommendation.

DATA CLEANING AND REMOVAL OUTLIERS

1

DATA CLEANING

The step of cleaning data that has been taken as follow:

1. Removing irrelevant values
2. Removing Duplicates
3. Convert Datatypes
4. Handling missing value (fill data with central tendency)

Before Data Cleaning : 5000 rows
After Data Cleaning : 4884 rows
(2.32% from initial data)

2

REMOVAL OUTLIERS

The step includes removing extreme outliers using tukey approach:

Before Data Cleaning : 4884 rows
After Data Cleaning : 4878 rows
(0.12% from initial data)

Total : Remove 122 data (2.44% of all data)

>>> Link to Dataset <<<

Descriptive Statistics

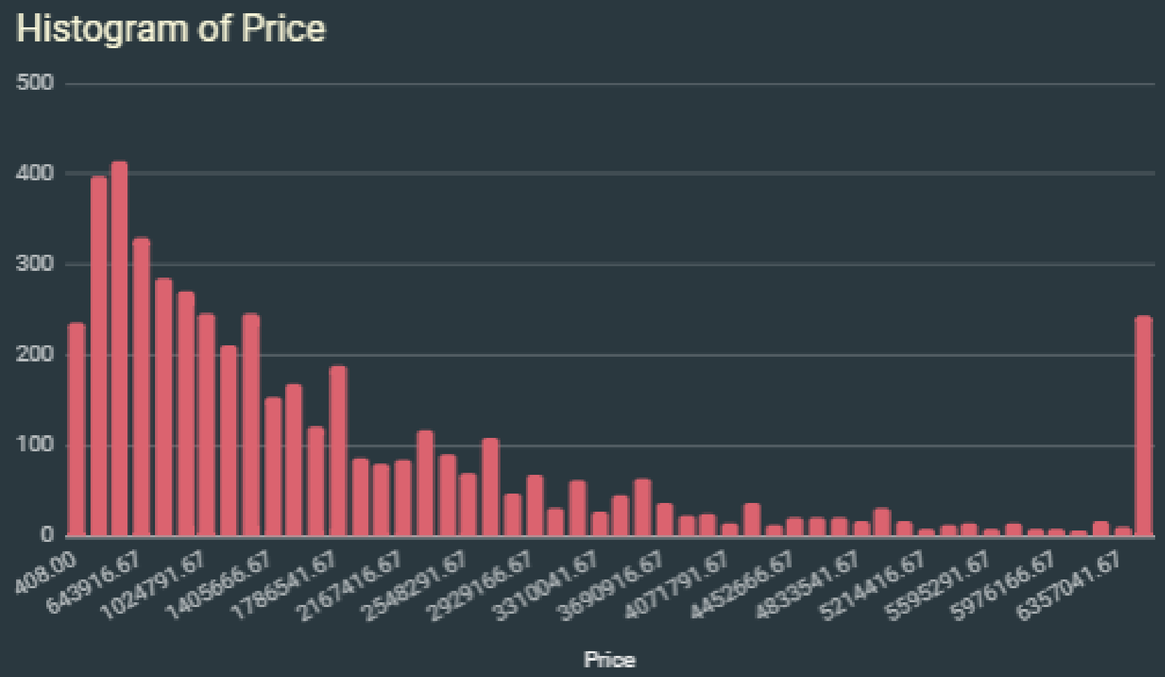
The columns will be analyzed are Price, Rooms, Bathroom, Car Park and Size because these columns are numerical and will give insight to answer business problem.

>>>link to dataset <<<

DESCRIPTIVE ANALYTICS

Price

Price	
Mean	2133665.712
Standard Error	48664.91683
Median	1300000
Mode	1200000
Standard Deviation	3398539.823
Sample Variance	11550072931232
Kurtosis	464.0454756
Skewness	15.46538412
Range	129999592
Minimum	408
Maximum	130000000
Sum	10405887677
Count	4877
Largest(1)	130000000
Smallest(1)	408
Confidence Level(95%)	95405.16391
Q1	700000
Q3	2480000
IQR	1780000
Lower limit	-1970000
Upper Limit	5150000
Coefficient of variation	1.592817377



Insight :

- The data has asymmetrical distribution or positive skewness where the mean value is bigger than median. It indicates column price has extreme outliers.

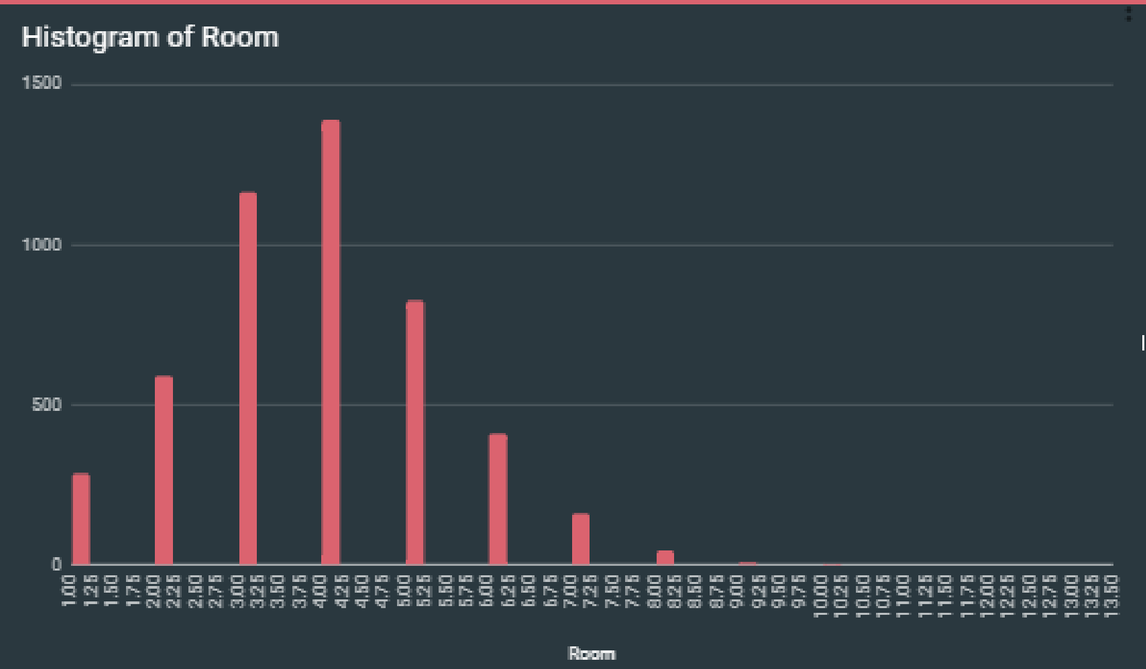
Recommendation :

- Since the data has positive skew it is recommended to use median as the central tendency.

DESCRIPTIVE ANALYTICS

Room

Room	
Mean	3.827
Standard Error	0.021
Median	4
Mode	4
Standard Deviation	1.496
Sample Variance	2.237
Kurtosis	0.333
Skewness	0.333
Range	12
Minimum	1
Maximum	13
Sum	18664
Count	4877
Largest(1)	13
Smallest(1)	1
Confidence Level(95%)	0.042
Q1	3
Q3	5
IQR	2
Lower limit	0
Upper limit	8
Coefficient of Variation	0.391



Insight :

- The data has symmetrical distribution or normal skewness. Where the value of mean,median and mode are slightly different. It indicates column room has mild outliers.

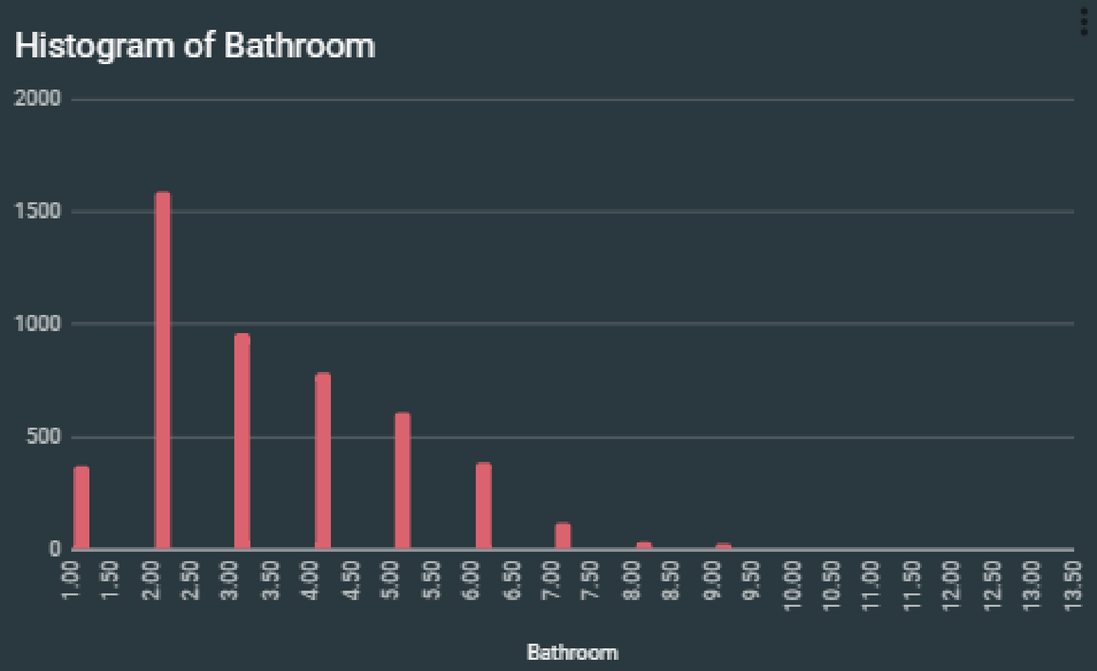
Recommendation :

- Since the data has normal skew it is recommended to use mean as the central tendency. Thus we can say that mostly property has 4 rooms.

DESCRIPTIVE ANALYTICS

Bathroom

Bathroom	
Mean	3.344
Standard Error	0.024
Median	3
Mode	2
Standard Deviation	1.651
Sample Variance	2.725
Kurtosis	0.675
Skewness	0.866
Range	12
Minimum	1
Maximum	13
Sum	16308
Count	4877
Largest(1)	13
Smallest(1)	1
Confidence Level(95%)	0.046
Q1	2
Q3	4
IQR	2
Lower Limit	-1
Upper Limit	7
Coefficient of Variation	0.494



Insight :

- The data has symmetrical distribution or normal skewness. Where the value of mean,median and mode are almost same. It indicates column bathroom has mild outliers.

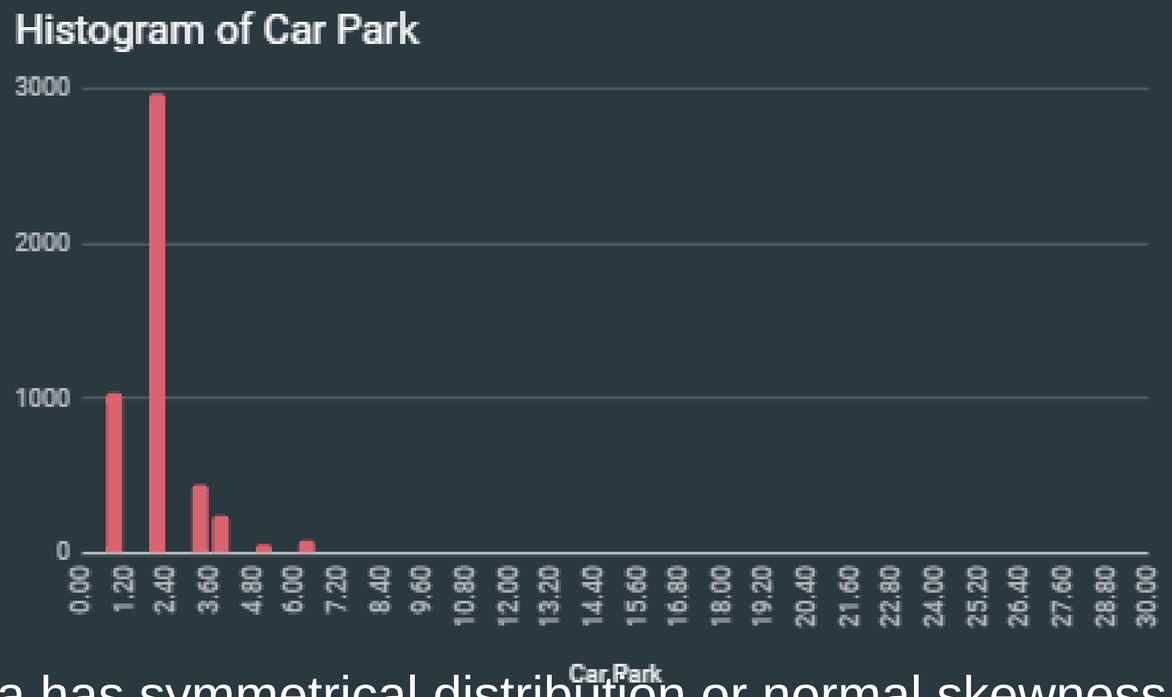
Recommendation :

- Since the data has normal skew it is recommended to use mean as the central tendency. Thus we can say that mostly property has 3 bathroom.

DESCRIPTIVE ANALYTICS

Car Park

Car Park	
Mean	2.138
Standard Error	0.017
Median	2
Mode	2
Standard Deviation	1.169
Sample Variance	1.367
Kurtosis	63.420
Skewness	4.856
Range	27
Minimum	1
Maximum	28
Sum	10428
Count	4877
Largest(1)	28
Smallest(1)	1
Confidence Level(95%)	0.033
Q1	2
Q3	2
IQR	0
Lower Limit	2
Upper Limit	2
Coefficient of Variation	0.547



Insight :

- The data has symmetrical distribution or normal skewness. Where the value of mean,median and mode are almost same. It indicates column Car Park has mild outliers.

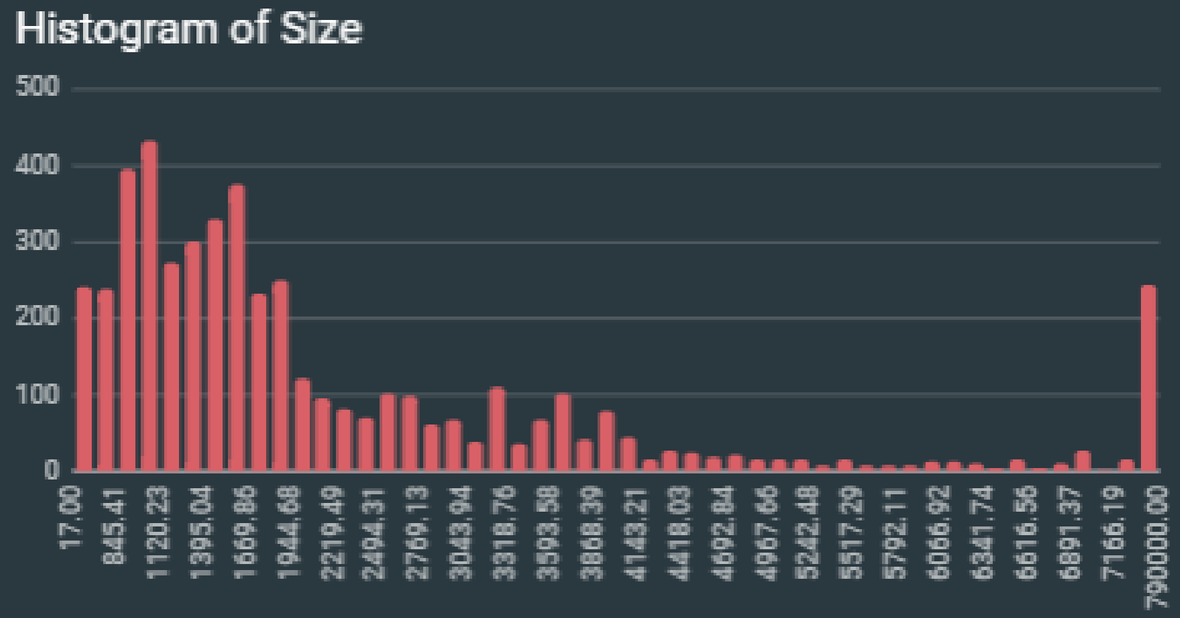
Recommendation :

- Since the data has normal skew it is recommended to use mean as the central tendency. Thus we can say that mostly property has 2 Car Park.

DESCRIPTIVE ANALYTICS

Size

Size	
Mean	2828.22
Standard Error	181.96
Median	1650
Mode	1650
Standard Deviation	12707.346
Sample Variance	161476838.554
Kurtosis	3054.310
Skewness	50.972
Range	789983
Minimum	17
Maximum	790000
Sum	13793240.47
Count	4877
Largest(1)	790000
Smallest(1)	17
Confidence Level(95%)	356.726
Q1	1094
Q3	2800
IQR	1706
Lower limit	-1465
Upper Limit	5359
Coefficient of Variation	4.493



Insight :

- The data has asymmetrical distribution or positive skewness where the mean value is bigger than median. It indicates column Car Park has extreme outliers.

Recommendation :

- Since the data has normal skew it is recommended to use median as the central tendency.

Exploratory Data Analysis (EDA)

We must calculate Quartile 1 - 4 for price column and use it as a filter to find the characteristics of the property.

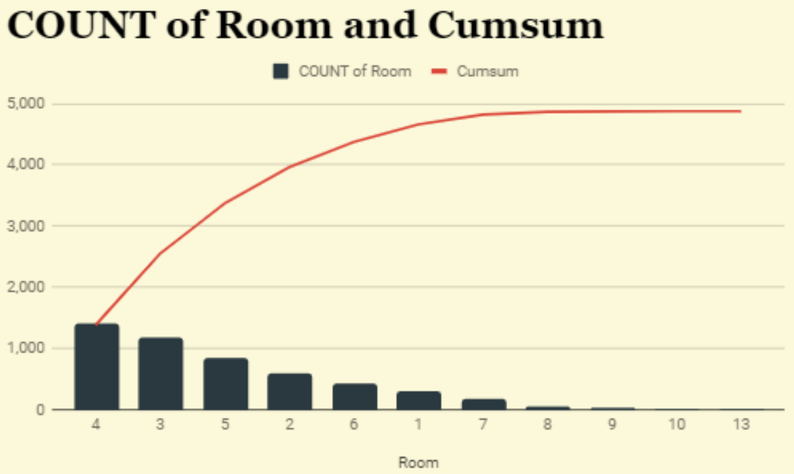
Quartile based on price result:

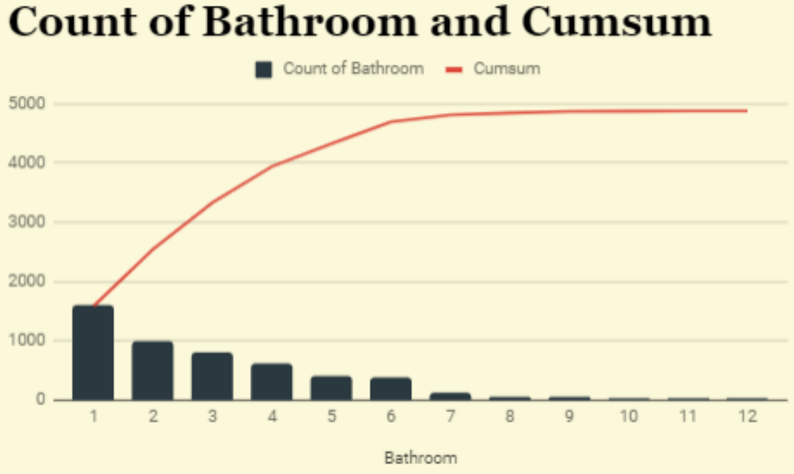
	Price	Type of Porperty
Q1	700000	AFFORDABLE PROPERTY
Q2	1300000	
Q3	2480000	LUXURY PROPERTY
Q4	130000000	

EDA-LUXURY PROPERTIES

>>> Link to Dataset <<<

Property Type	Median of Size	Median of Price
Residential Land	12141	6300000
Bungalow Land	9120	5878950
Bungalow	7500	5800000
3.5-sty Terrace/Link House	2805	4950000
4-sty Terrace/Link House	2270	4728005
4.5-sty Terrace/Link House	1367	3688000
Semi-detached House	3760	3050000
3-sty Terrace/Link House	1900	1785000
2.5-sty Terrace/Link House	1840	1680000
Townhouse	2262	1250000
2-sty Terrace/Link House	1760	1250000
Condominium	1620	1200000
Serviced Residence	1042	922500
1.5-sty Terrace/Link House	1760	680000
1-sty Terrace/Link House	1650	680000
Apartment	881	355000
Flat	650	170000
Grand Total	1650	1300000



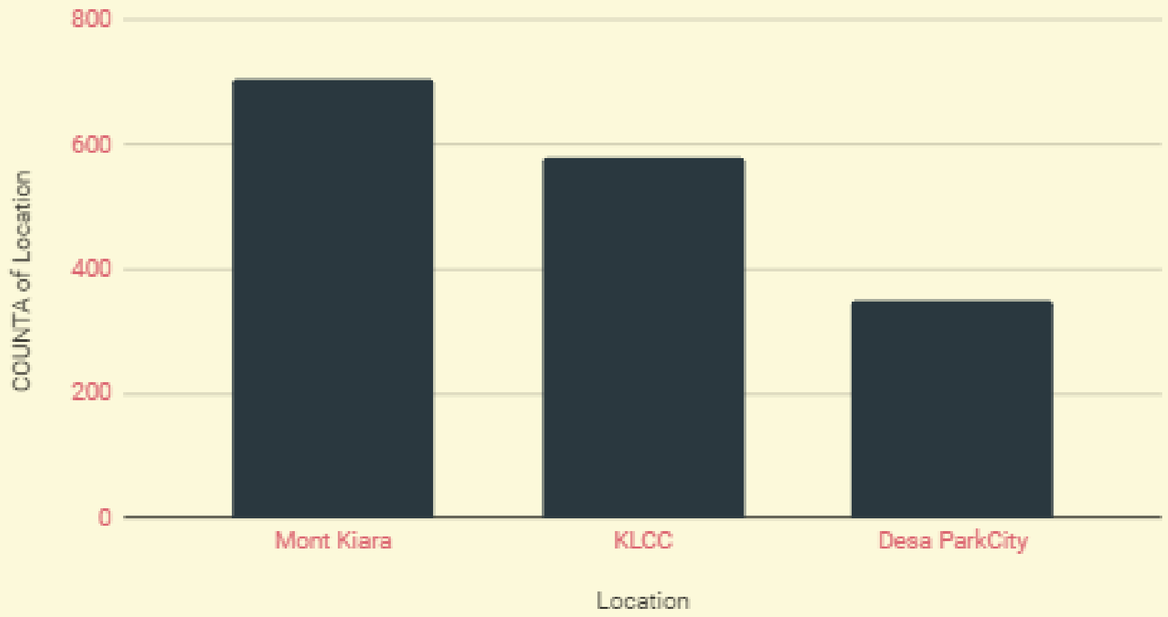


Insight :

- The range median size in luxury property is 12,141-650
- The range median prize in luxury property is RM 6,300,000-170,000
- 4-5 rooms amount is dominant compared to other amount in Luxury Property
- The dominant amount of bathroom in Luxury Property is 1-3

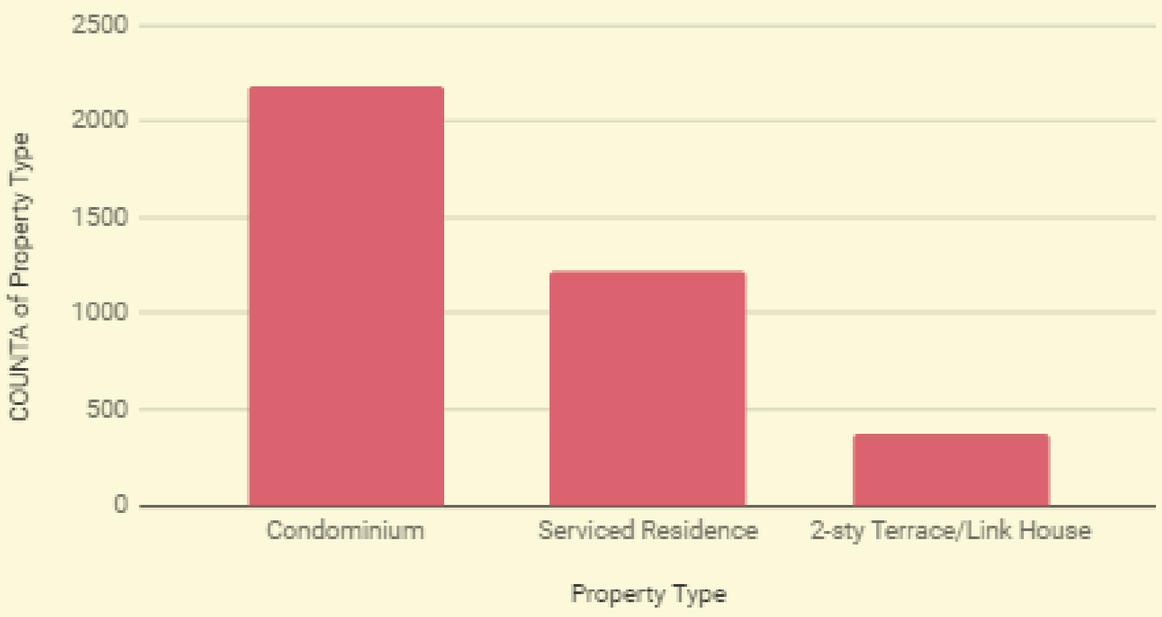
EDA-LUXURY PROPERTIES

Top 3 Location



We found that there are 3 location being top location in Luxury properties which are Mont Kiara followed by KLCC and Desa ParkCity.

Top 3 Property Type

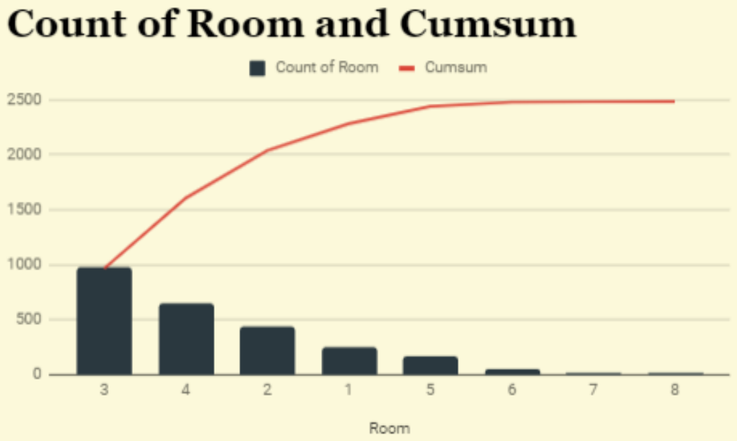


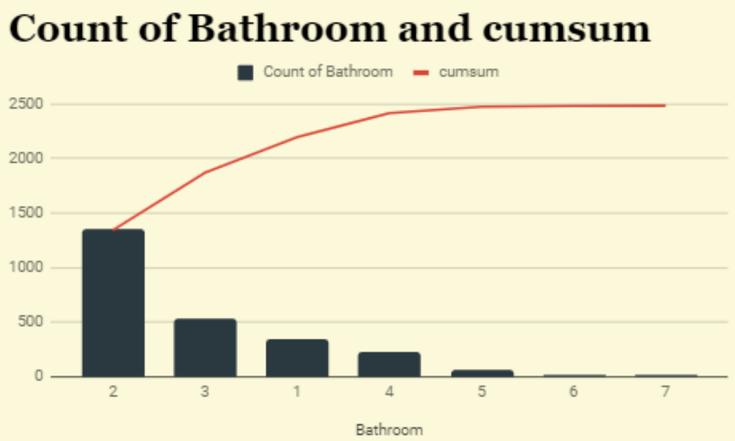
Based on the property type there are top 3 of it. The fisrt common amount is condonium and follwed with Serviced Residence and the third position is 2-sty Terrace/Link House

EDA-AFFORDABLE PROPERTIES

>>> Link to Dataset <<<

Property Type	Median of Size	Median of Price
Residential Land	17760	1238000
Bungalow Land	6400	1150000
Bungalow	5320	1100000
Semi-detached House	4002	1090000
1.5-sty Terrace/Link House	1705	1088000
3-sty Terrace/Link House	1650	800000
2-sty Terrace/Link House	1650	780000
1-sty Terrace/Link House	1600	750000
Townhouse	1500	710000
2.5-sty Terrace/Link House	1470	670000
Condominium	1273	664020
Serviced Residence	924	659000
Apartment	881	355000
Flat	650	170000
Grand Total	1170	710000



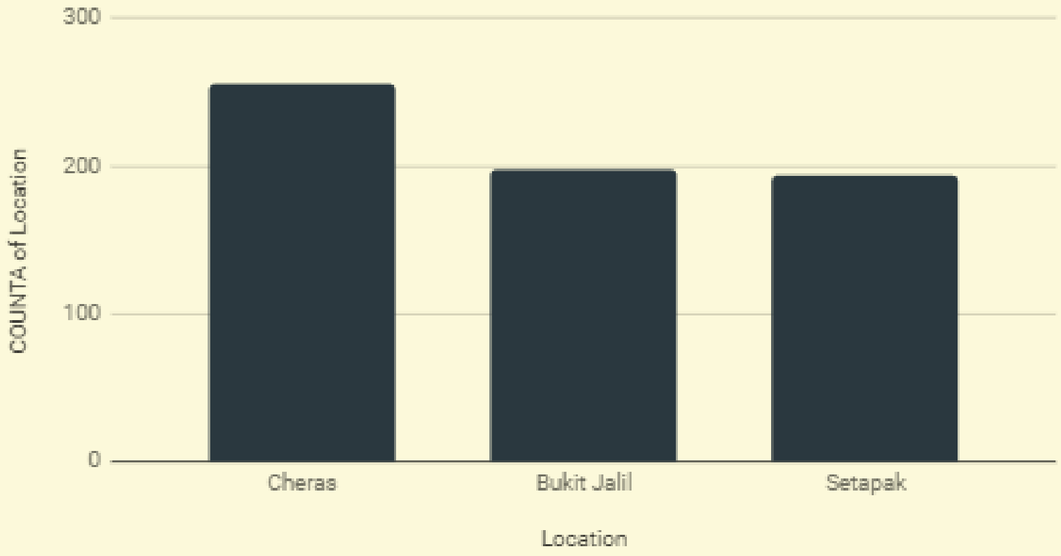


Insight :

- The range median size in affordable property is 17,760-650
- The range median prize in affordable property is RM 1,238,000-170,000
- Mostly the affordable property has 3 room and 2 bathroom

EDA-AFFORDABLE PROPERTIES

Top 3 Location



We found that there are 3 location being top location in Affordable properties which are Cheras followed by Bukit Jalil and Setapak.

Top 3 Property Type

Both in Luxury and affordable properties Condonium is being the dominatn amount of property type then follwed by serviced and 2-sty Terrace/Link House.

EDA-RECOMMENDATION

Luxury Properties

Based on the characteristic mentioned in the previous deck , we recommend this luxury properties to the potential buyer who are looking for spacious and comfortable living space, have a lot of member family since this property offers many rooms and bathroom. th buyers can have premium property with a range of pricing options to suit their budget.

Affordable Properties

Based on the characteristic mentioned in the previous deck , we recommend this affordable properties to the potential buyer who are looking for budget-friendly options or first-time homebuyers and looking for reasonably sized homes that meet their needs without breaking the bank.

STATISTICAL MEASUREMENT-CORRELATION

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

	Price	*Rooms*	*Bathrooms*	*Car Parks*	*Size*
Price	1				
Rooms	0.723	1			
Bathrooms	0.772	0.824	1		
Car Parks	0.632	0.628	0.6477	1	
Size	0.796	0.668	0.7106	0.5296	1

Insight :

- **Size** has strong correlation with the price of property
- Bathroom and rooms also has strong positive correlation, it indicates multicollinearity
- Car parks has weak correlation with size

>>> Link to Dataset <<<

STATISTICAL MEASUREMENT-REGRESSION

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

1

Regression Statistics	
Multiple R	0.860
R Square	0.740
Adjusted R Square	0.734
Standard Error	435863.0007
Observations	177

2

Significance F
0

3

	Coefficients	*P-value*
Intercept	-211210.5351	0.069
Rooms	71340.88497	0.136
Bathrooms	151347.8424	0.001
Car Parks	160744.5183	0.004
Size	472.1989033	0

Insight :

1. The R square meaning that 0,73% variation of price data can be explained by 4 variables
2. Simultaneous Test : Overall effect is significant which means four independent variables have significant coefficients in explaining prices.
3. Partial Test : Bathrooms, Car Parks and Size have significant coefficient indicated by the P-value <= 5% unfortunately rooms doesn't have significant coefficient indicated by the P-value >= 5%. Thus we must ommit room variable.

STATISTICAL MEASUREMENT- BACKWARD REGRESSION

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

1

Regression Statistics	
Multiple R	0.858
R Square	0.736
Adjusted R Square	0.732
Standard Error	437426.406
Observations	177

2

Significance F
0

3

	Coefficients	*P-value*
Intercept	-141624	0.1838683151
Bathrooms	190889	0.000
Car Parks	176807	0.001
Size	488	0

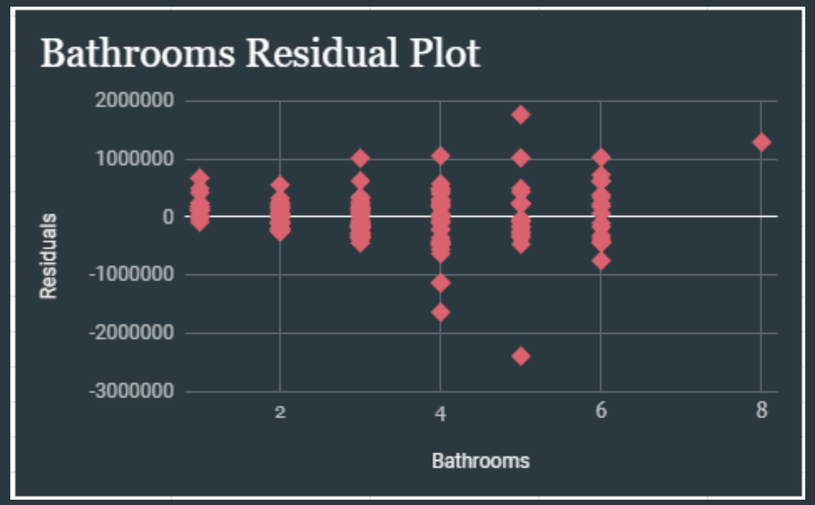
Insight :

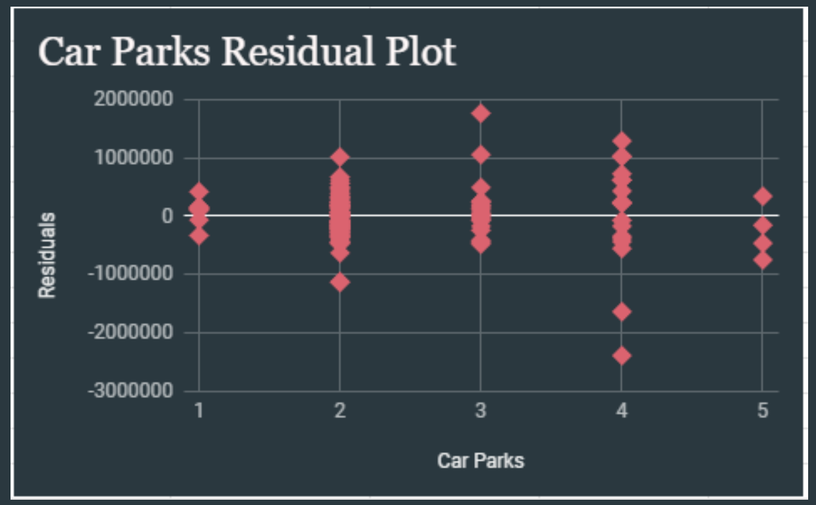
1. Model Evaluation : Deleting room does not imporve the adj R square. Adj R square is 0,73, meaning that 73% variation of price can be explained by 3 variables (Bathroom, Car Park and size)
2. Simultaneous Test : Overall effect is significant which means three independent variables have significant coefficients in explaining prices.
3. Partial Test : Bathrooms, Car Park and Size have significant coefficient indicated by the P-value <= 5% and Bathroom has the biggest impact to the price among others.

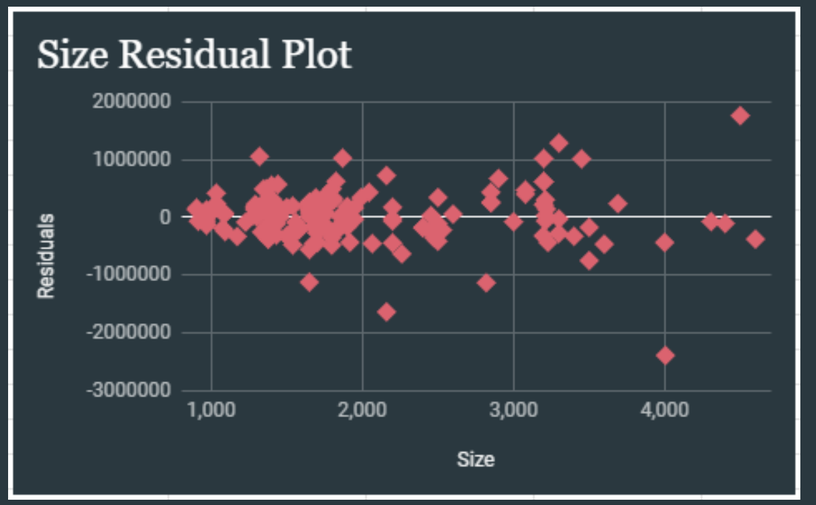
>>> Link to Dataset <<<

STATISTICAL MEASUREMENT- ASSUMPTION CHECK

- The data is autocorrelation because the residuals has specific pattern
- After removing one independent variable, We must check the homoscedasticity and Normal Probability plot (QQ plot)



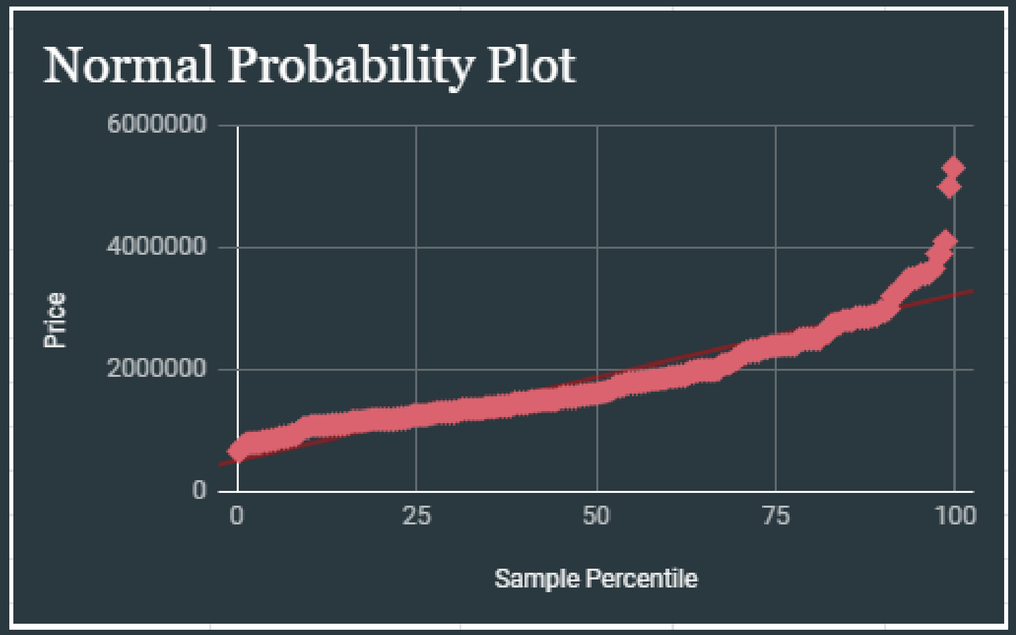




The residual plots seem like not scattered and do not approach a constant pattern for each observation. It violates the homoscedasticity assumption.

STATISTICAL MEASUREMENT- ASSUMPTION CHECK

QQ Plot



- The residual close to the line, so it is normality assumption.
- It means we have to ba careful when predictive expensive price.

PRICE OFFER

Based on the regression result, our regression equation will be:

Price = -141624 + (190889*Bathroom)+(176807*Car Parks)+ (488*Size)

Assume there is one loyal user who is looking for a property in Desa ParkCity with 3 rooms, 4 bathroom, 3 car park, and 2200 sq ft requirements. At what price of property can you offer the client?

The predicted price offer from regression formula is RM 2,225,953.

That is the predicted price offer from the equation, but it is not a reliable model, we must consult to the head of data about the model that we made before applying the regression.

RECOMMENDATION

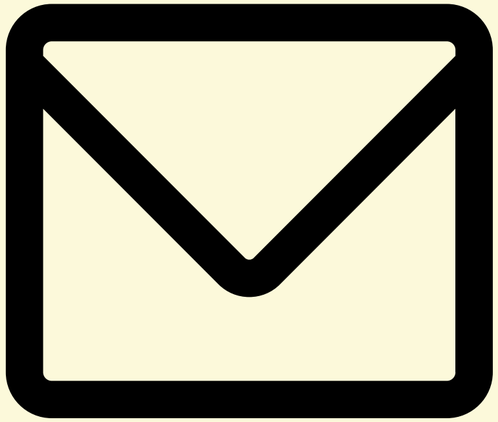
- From the result of correlation and regression size has the highest correlation to price, but bathroom has the biggest impact to the price. when the number of bathroom was added then add 190.889 to the price.

- Since the result of previous model of regression is homoscedasticity,We have to consult with the Head of data to develop better model in order to predict price recommendation

Connect with me

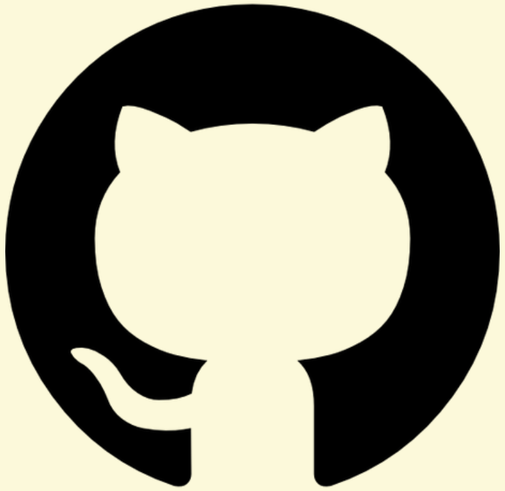
LINKEDIN

click here



EMAIL

maismaula01@gmail.com



GITHUB

click here