

# ABC Company Analysis



Dataset + Sheets: [Click here](#)




# **Business Understanding**

# Background

ABC Company is a **property listing** company in **Malaysia**. They generate revenue through a 20% joint-profit sharing model, but may face challenges in selling high-priced properties due to excessive room numbers or room size.

As a Data Analyst, I was expected to **deliver insight** for the company which can help the company to **maximize their profit** while helping users and tenant at the same time.



# Core Business Problem

With a **joint profit sharing of 20%**, it means that the more we can sell properties at the highest price, the higher the income or profit for the company will be. Even so, the property with the highest price is not always easy to sell. Coupled with the threat of a recession in 2023, it has made some people think again about investing in or buying property this year. For this reason, with the raw data we have, **we need to analyze it and extract insights to provide business recommendations for our company.**

# Problem Statement

How to increase **revenue 10%**  
gradually in **one year**?

# Objective

To find the right method to **increase revenue 10% in one year**



# Data Dictionary

Column Name	Definition
Location	The neighborhood in which the property is located
Price	The listed sales price in Malaysian Ringgit (RM).
Property Character	Whether it is sold by land or spaces of land the building is built upon / <u>reference</u>
Size	The total size of the property. Note that some properties are listed by different units of measurement
Rooms	The number of listed rooms.
Bathrooms	The number of bathrooms in the property.
Car Parks	The number of car parks on the property.
Property Type	Malaysian properties may fall in one of several property types depending on their characteristics that are defined by the listings/users.
Furnishing	Whether or not the property comes pre-furnished.

# **Data Cleaning**



# Data Cleaning Step

What to do	Column	Sheet	Reason
Remove duplicate	All	Raw Data Property Listing Dataset	To eliminate duplication of redundant data, because it can affect the analysis.
Check blank data in Location	Location	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Delete "Kuala Lumpur" form Location	Location	Raw Data Property Listing Dataset	To make the data simpler and easier to read the data.
Check blank data in Price	Price	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Delete blank data from Price	Price	Raw Data Property Listing Dataset	To facilitate analysis, because the 34 data cannot be filled in because the Price column is quite important, and we cannot estimate the data by looking at other columns. Then 34 data is also not too much when compared to the total existing data (0.6%).

# Data Cleaning Step

What to do	Column	Sheet	Reason
Delete "RM" form Price	Price	Raw Data Property Listing Dataset	To make the data simpler and easier to read the data.
Check blank data in Rooms	Rooms	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Calculate Rooms	Rooms	Raw Data Property Listing Dataset	To match the writing format of the Rooms column.
Replace "Studio" in Rooms with "1"	Rooms	Raw Data Property Listing Dataset	To match the writing format of the Rooms column.
Check blank data in Bathrooms	Bathrooms	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.

# Data Cleaning Step

What to do	Column	Sheet	Reason
Check blank data in Car Parks	Carparks	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Check blank data in Property Type	Property Type	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Delete Property Type Description (Corner, Intermediate, EndLot, etc)	Property Type	Raw Data Property Listing Dataset	To make data simpler and facilitate analysis
Check blank data in Property Character	Property Character	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Delete ":" in Property Character	Property Character	Raw Data Property Listing Dataset	To make data simpler and facilitate analysis.

# Data Cleaning Step

What to do	Column	Sheet	Reason
Check blank data in Size	Size	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.
Delete blank data from Size	Size	Raw Data Property Listing Dataset	To facilitate analysis. because the 83 data cannot be filled in because the Size column is quite important, and we cannot estimate the data by looking at other columns. In addition, the analysis that will be carried out is based on the characteristics of the existing properties, so we don't want data related to size to be empty.
Delete row with "Kuala Lumpur sqft" and "Malaysia sqft"	Size	Raw Data Property Listing Dataset	Data is not clear.
Delete row with "0 sqft"	Size	Raw Data Property Listing Dataset	Data is not clear because it is impossible for a property to have an of 0 sqft.
Split the Size with their Unit	Size	Raw Data Property Listing Dataset	To make data simpler and facilitate analysis.
Changing Units to sqft against other units (sqm and acres)	Unit	Raw Data Property Listing Dataset	To make data simpler and facilitate analysis.

# Data Cleaning Step

What to do	Column	Sheet	Reason
Calculating ambiguous size values (range, multiplication, symbol)	Size	Raw Data Property Listing Dataset	To equalize values and eliminate ambiguous data.
Changing the value of sq. to be sqft	unit	Raw Data Property Listing Dataset	To equalize values and eliminate ambiguous data.
Check blank data in Furnishing	Furnishing	Raw Data Property Listing Dataset	To see whether there is empty data or not, because it will affect the analysis.

# **Descriptive Statistics**

# Descriptive Statistics - by Prices

There is a very large gap between the mean and median.

Skewness that appears has a very positive value.

There is a very large gap between the minimum and maximum values.

These three things indicate the **potential of outliers**. For this reason, it is very good to use the **median** as a central tendency to represent the character of the Price column. By using the upper limit, we can find out that there are **370** potential outliers.

Price	
Mean	2166241.717
Standard Error	49764.45762
Median	1300000
Mode	1200000
Standard Deviation	3448141.886
Sample Variance	11889682464701
Kurtosis	444.8619057
Skewness	15.11601429
Range	129998850
Minimum	1150
Maximum	130000000
Sum	10400126481
Count	4801
Largest(1)	130000000
Smallest(1)	1150
Confidence Level(95%)	97561.14282
Q1	715000
Q3	2500000
IQR	1785000
Lower limit (Q1 - IQR)	-1962500
Upper limit (Q3 + IQR)	5177500

# Descriptive Statistics - by Total Rooms

There is not a very large gap between the mean and median.

The skewness that appears has a very neutral value.

There is not a large gap between the minimum and maximum values.

These three things indicate that there are not **potential of outliers**. For this reason, it is good to use the **mean/median** as a central tendency to represent the character of the Total Rooms column. But if we using the upper limit to detected potential outliers, we can find out that there are **17** potential outliers.

Price	
Mean	3.842844974
Standard Error	0.02264103028
Median	4
Mode	4
Standard Deviation	1.551530186
Sample Variance	2.407245919
Kurtosis	3.122071761
Skewness	0.6535122649
Range	19
Minimum	1
Maximum	20
Sum	18046
Count	4696
Largest(1)	20
Smallest(1)	1
Confidence Level(95%)	0.04438704557
Q1	3
Q3	5
IQR	2
Lower limit (Q1 - IQR)	0
Upper limit (Q3 + IQR)	8



# Descriptive Statistics - by Bathrooms

There is not a very large gap between the mean and median.

The skewness that appears has a very neutral value.

There is not a large gap between the minimum and maximum values.

These three things indicate that there are not **potential of outliers**. For this reason, it is good to use the **mean/median** as a central tendency to represent the character of the Bathrooms column. But if we using the upper limit to detected potential outliers, we can find out that there are **12** potential outliers.

Price	
Mean	3.378343676
Standard Error	0.02500726194
Median	3
Mode	2
Standard Deviation	1.709480038
Sample Variance	2.922321999
Kurtosis	2.901588077
Skewness	1.074692354
Range	19
Minimum	1
Maximum	20
Sum	15787
Count	4673
Largest(1)	20
Smallest(1)	1
Confidence Level(95%)	0.04902603243
Q1	2
Q3	5
IQR	3
Lower limit (Q1 - IQR)	-2.5
Upper limit (Q3 + IQR)	9.5

# Descriptive Statistics - by Car Parks

- There is not a very large gap between the mean and median.
- The skewness that appears has a very neutral value.
- There is not a large gap between the minimum and maximum values.

These three things indicate that there are not **potential of outliers**. For this reason, it is good to use the **mean/median** as a central tendency to represent the character of the Car Parks column. But if we using the upper limit to detected potential outliers, we can find out that there are **27** potential outliers.

Price	
Mean	2.198937426
Standard Error	0.02394224487
Median	2
Mode	2
Standard Deviation	1.393594967
Sample Variance	1.942106931
Kurtosis	43.67412798
Skewness	3.992074124
Range	27
Minimum	1
Maximum	28
Sum	7450
Count	3388
Largest(1)	28
Smallest(1)	1
Confidence Level(95%)	0.04694271155
Q1	1
Q3	3
IQR	2
Lower limit (Q1 - IQR)	-2
Upper limit (Q3 + IQR)	6

# Descriptive Statistics - by Size

- There is a very large gap between the mean and median.
- Skewness that appears has a very positive value.
- There is a very large gap between the minimum and maximum values.

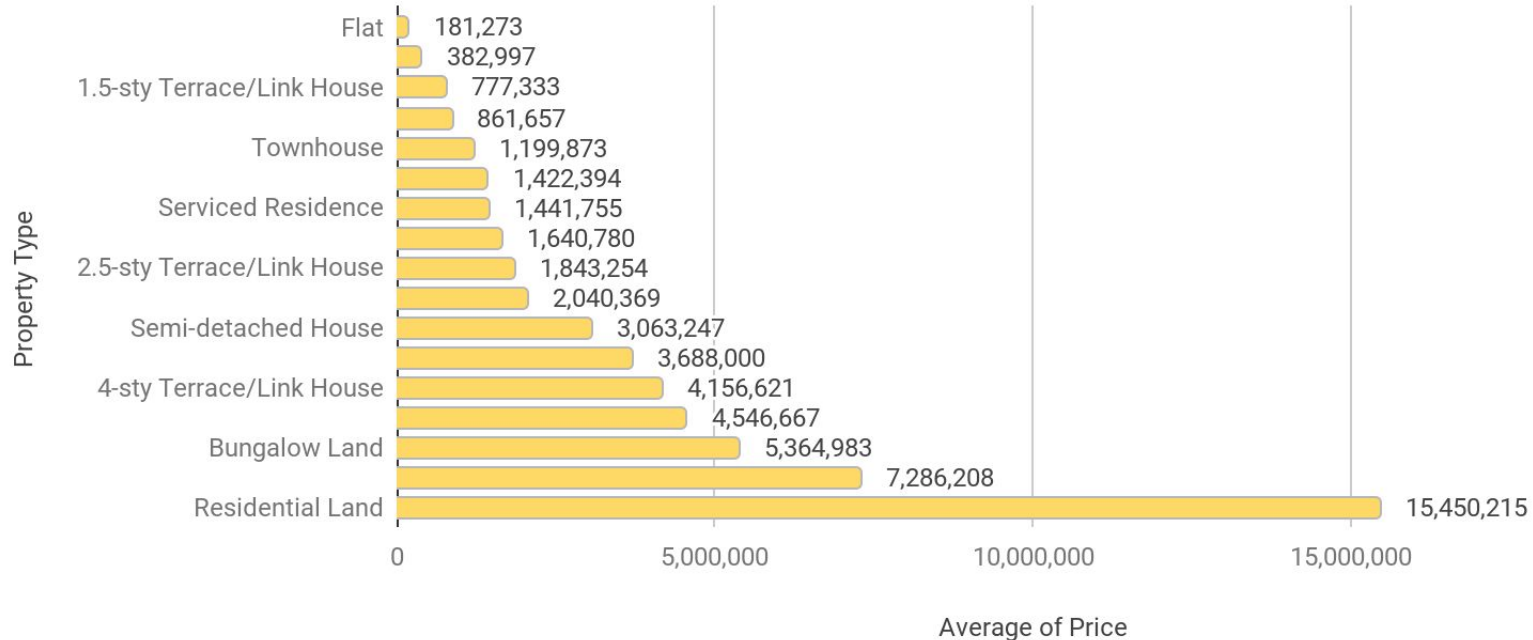
These three things indicate the **potential of outliers**. For this reason, it is very good to use the **median** as a central tendency to represent the character of the Size column. By using the upper limit, we can find out that there are **398** potential outliers.

Price	
Mean	2856.260192
Standard Error	184.898567
Median	1650
Mode	1650
Standard Deviation	12811.48281
Sample Variance	164134091.8
Kurtosis	3002.58137
Skewness	50.52146735
Range	789983
Minimum	17
Maximum	790000
Sum	13712905.18
Count	4801
Largest(1)	790000
Smallest(1)	17
Confidence Level(95%)	362.4859259
Q1	1087
Q3	2850
IQR	1763
Lower limit (Q1 - IQR)	-1557.5
Upper limit (Q3 + IQR)	5494.5

# **Exploratory Data Analysis and Insights**

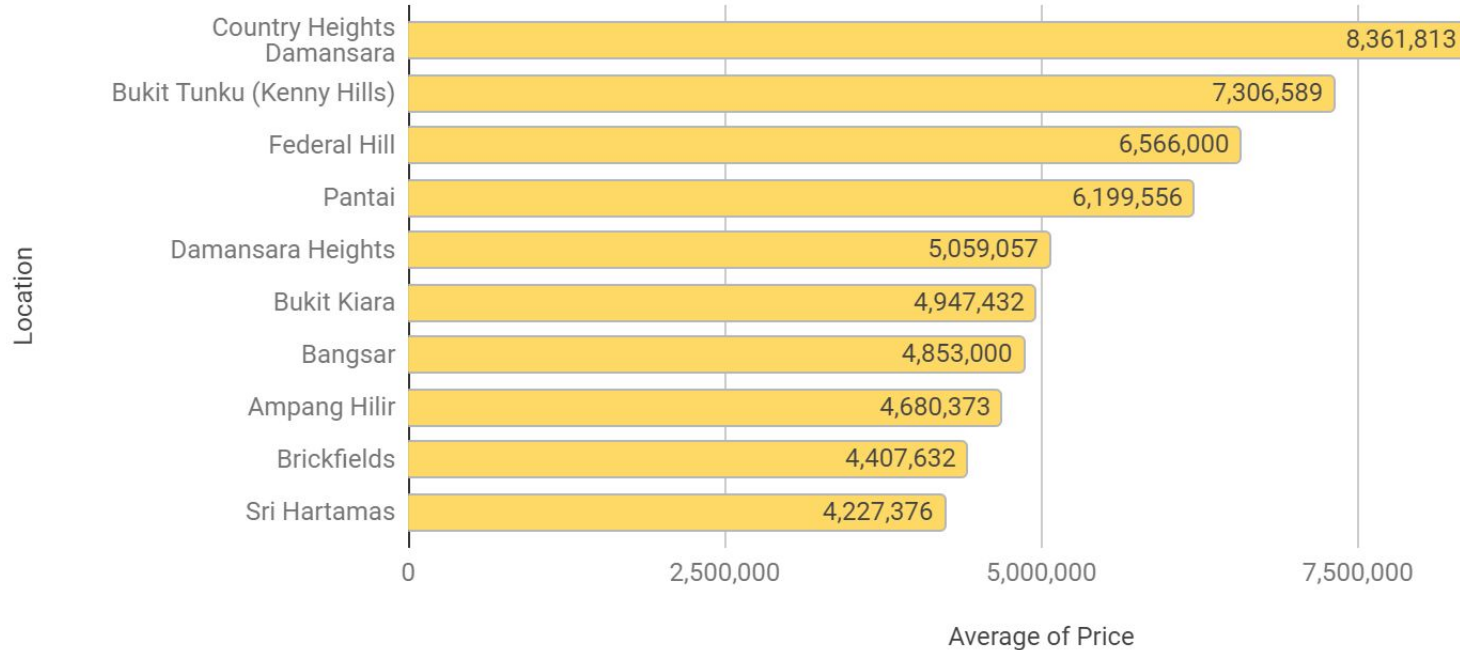
# Average Property Type Per Price

Average Property Type Per Price



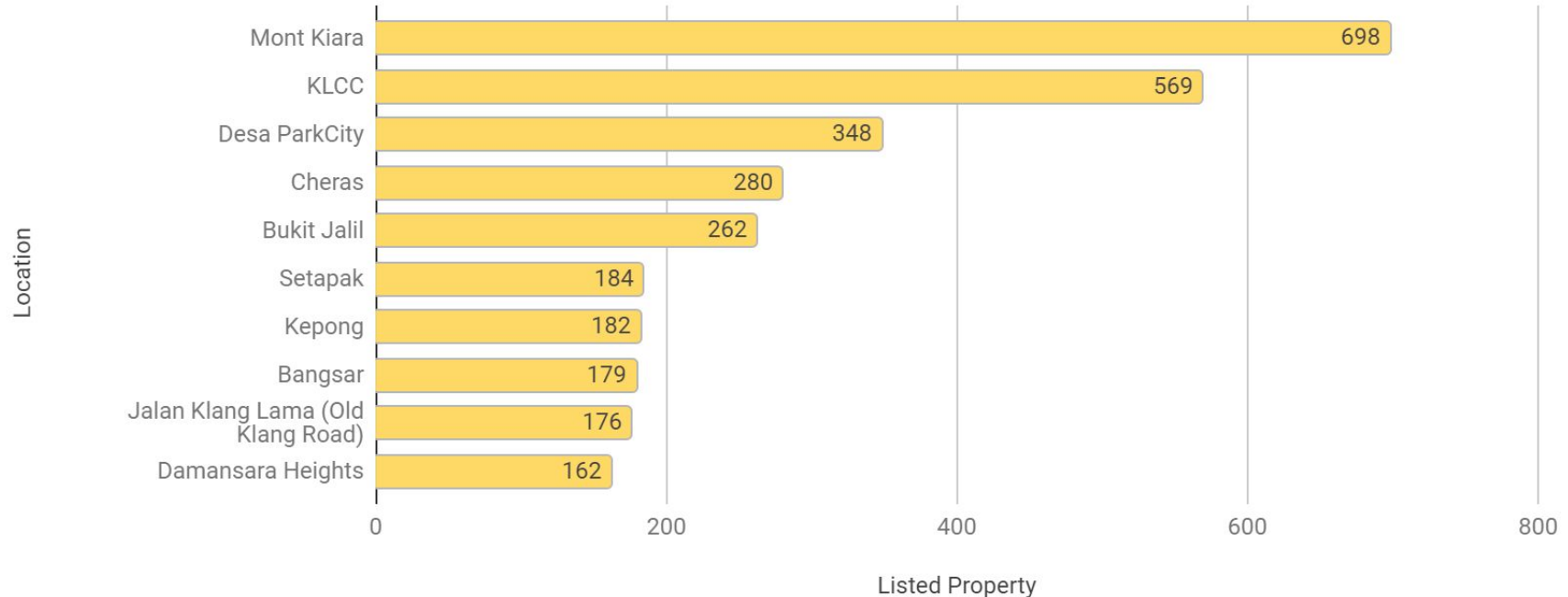
# Top 10 Average Property Price by Location

Top 10 Average Property Price by Location



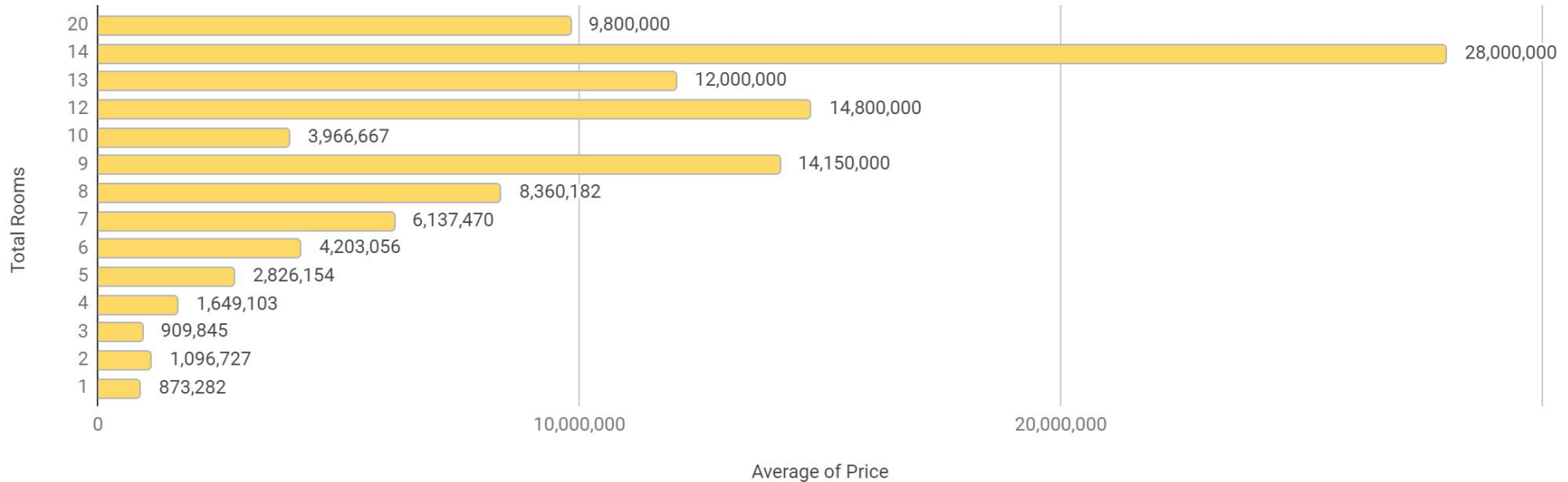
# Top 10 Location with Most Listed Property

Top 10 Location with Most Listed Property



# Average Property Price per Total Rooms

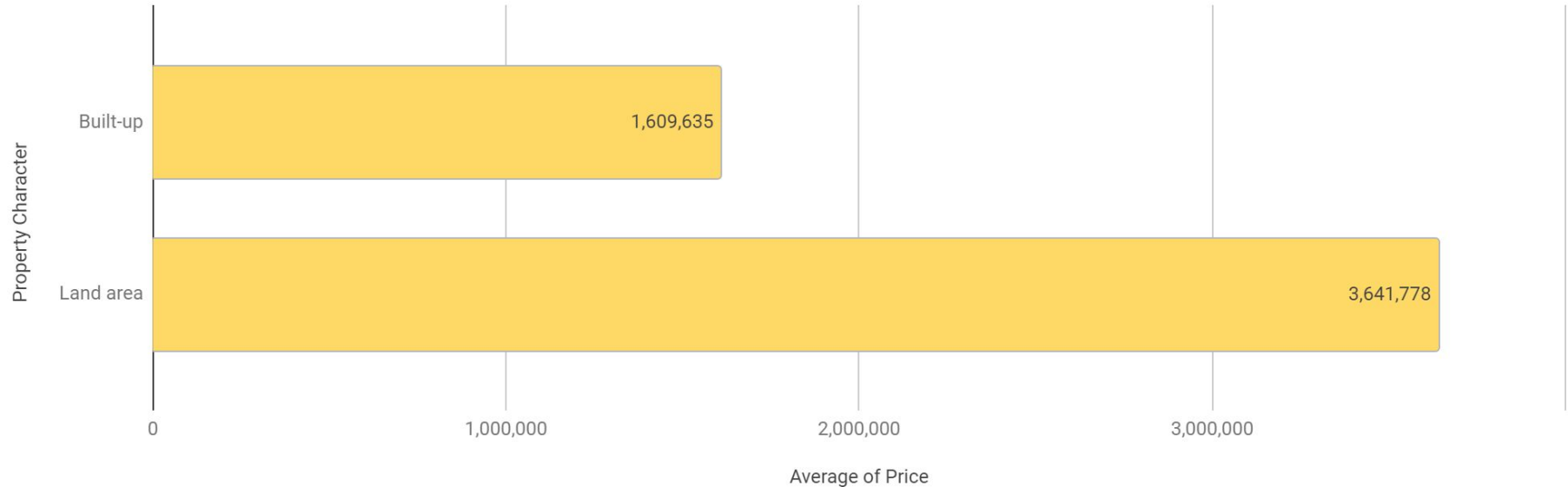
Average Property Price per Total Rooms





# Average Property Price per Property Character

Average Property Price per Property Character



# Average Property Price per Property Furnishing

Average Property Price per Property Furnishing



After we carry out exploratory data analysis on listed properties in general, we can conduct exploratory data analysis on listed properties more specifically. By using the Descriptive Statistics of the price variable, we can use quartiles as a grouping of property categories based on price. For example, we can classify property price ranges from **minimum - Q2 as affordable property and Q3 - maximum as luxury property**. After grouping it is known that there are **3599** property units which are included in luxury properties and there are **1202** property units which are included in affordable properties.



# Property Price Characteristics

Based on previous insights, we can classify property price into 2 groups:



## **Affordable Property**

(RM 1,150 - RM 1,300,000)

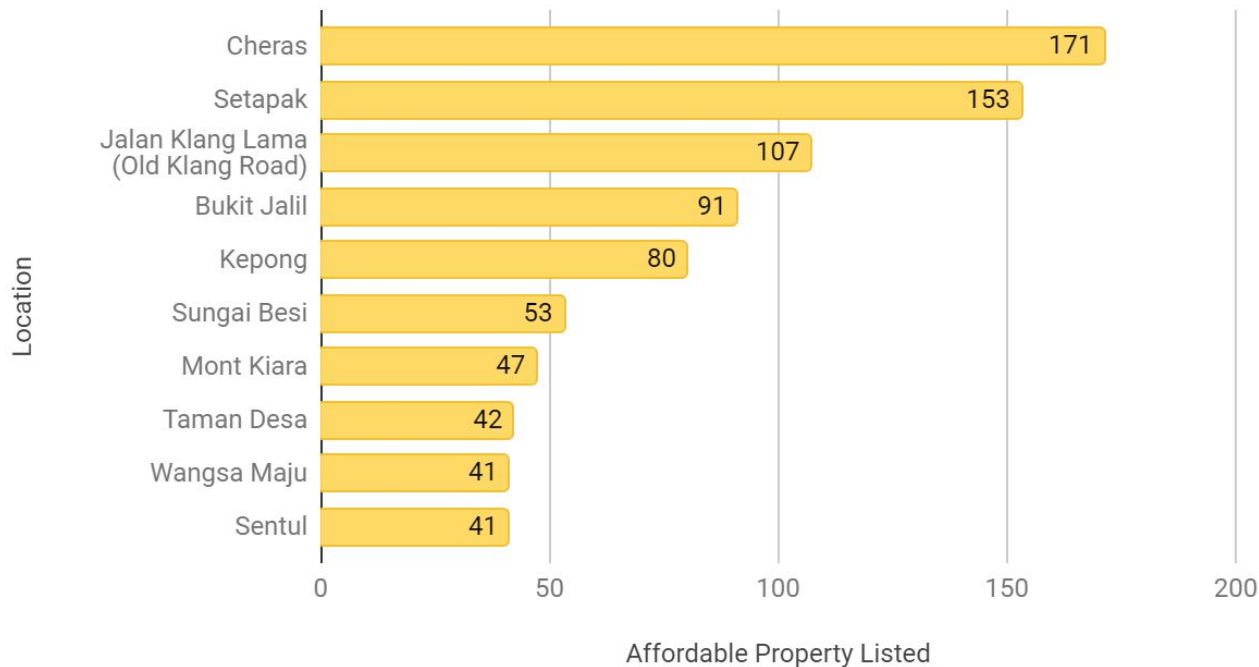


## **Luxury Property**

(>RM 1,300,000 - RM 130,000,000)

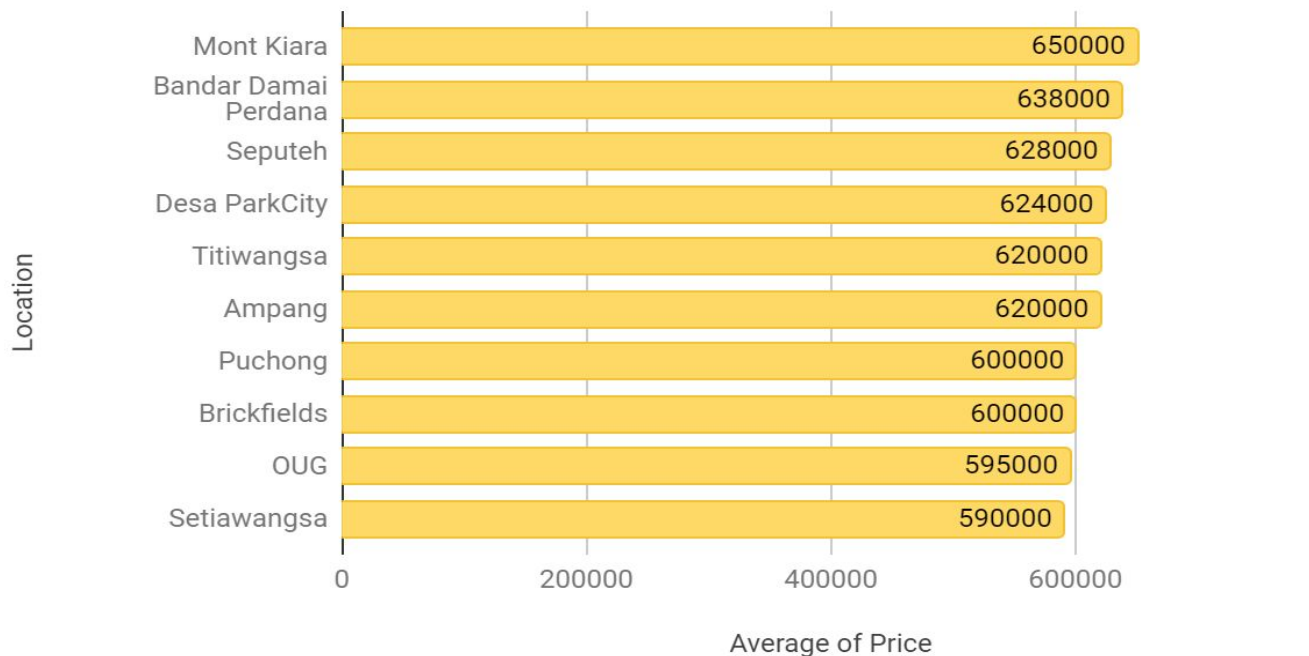
# Top 10 Location with Most Affordable Property Listed

Top 10 Location with Most Affordable Property Listed



# Top 10 Average Luxury Property Price by Location

Top 10 Average Affordable Property Price by Location



# Insights Gained

## Insights Gained of Affordable Property

The property room that have the highest percentage is property that have 3 room

The property bathroom that have the highest percentage is property that have 2 bathroom

The property car parks that have the highest percentage is property that have 2 car parks

The affordable property type that have the highest percentage in Kuala Lumpur is Condominium

The affordable property character that have the highest percentage in Kuala Lumpur is Built-up

The affordable property furnishing that have the highest percentage in Kuala Lumpur is Partly Furnished

The location that has the most number of affordable properties is Cheras with a total of 171 property units with an average price of RM 450,800. Affordable properties in Cheras have an average of 3 rooms, 2 bathrooms and 2 car parks with a property of around 960 sqm.

The location that has the highest average affordable luxury price is in Mont Klara with an average price of RM 650,000, while the that has the smallest average affordable property price is in Bandar Tasik Selatan with an average price of RM 180,000.

# Insights Gained

## Insights Gained of Affordable Property

The location that has the highest average luxury property is in OUG with an average of 1,684 sqm, while the that has the smallest average luxury property is Pandan Perdana with an average of 450 sqm.

The location that has the most complete average number of facilities (rooms, bathrooms, parking) is in Bandai Damai Perdana with 4 rooms, 3 bathrooms, and 2 car parks.

The most sold property type in Cheras is Condominium with a total of 80 units sold, making Cheras the with the second most Condominium types sold in Kuala Lumpur after Setapak with 89 units. Cheras also has an affordable property character that has been built which is dominant compared to what is still a land .

The most sold property type in Mont Klara is Serviced Residence with a total of 31 units sold. Mont Klara also has an affordable property character that has been built which is dominant compared to what is still a land .

The most sold property type in Bandar Tasik Selatan is Flat with a total unit sold of 1 unit.

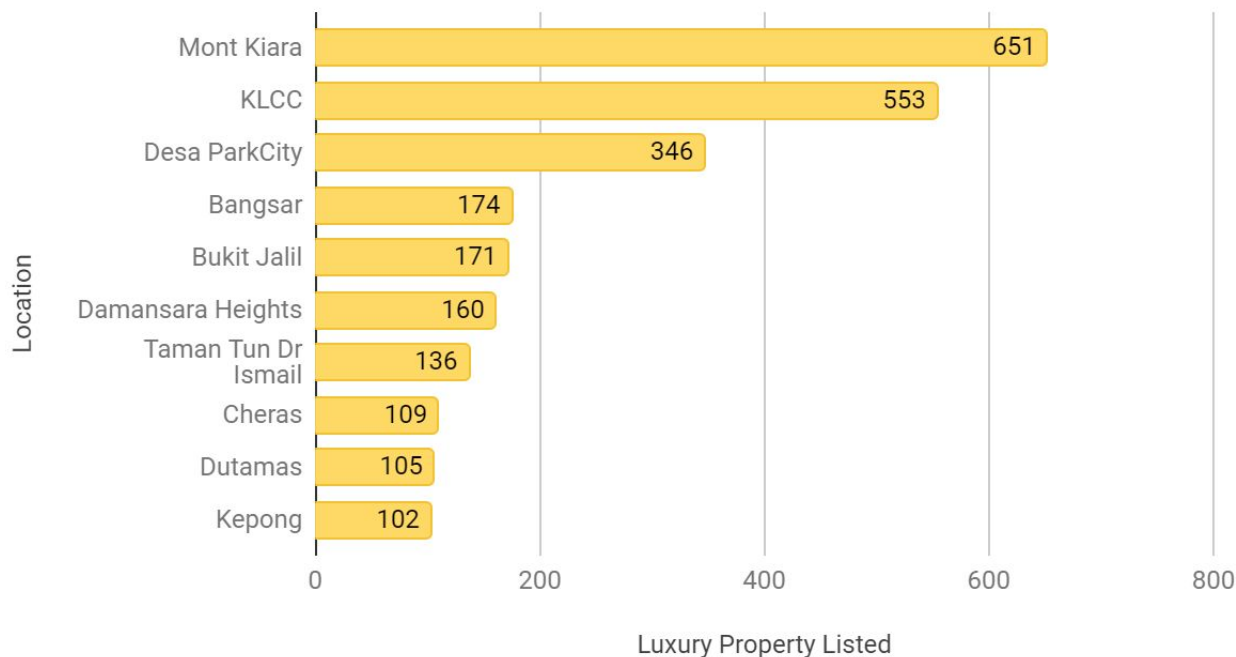
Mont Klara is the with the largest fully furnished affordable property, followed by Cheras in second place.

Mont Klara is the with the largest fully furnished affordable property, followed by Cheras in second place.



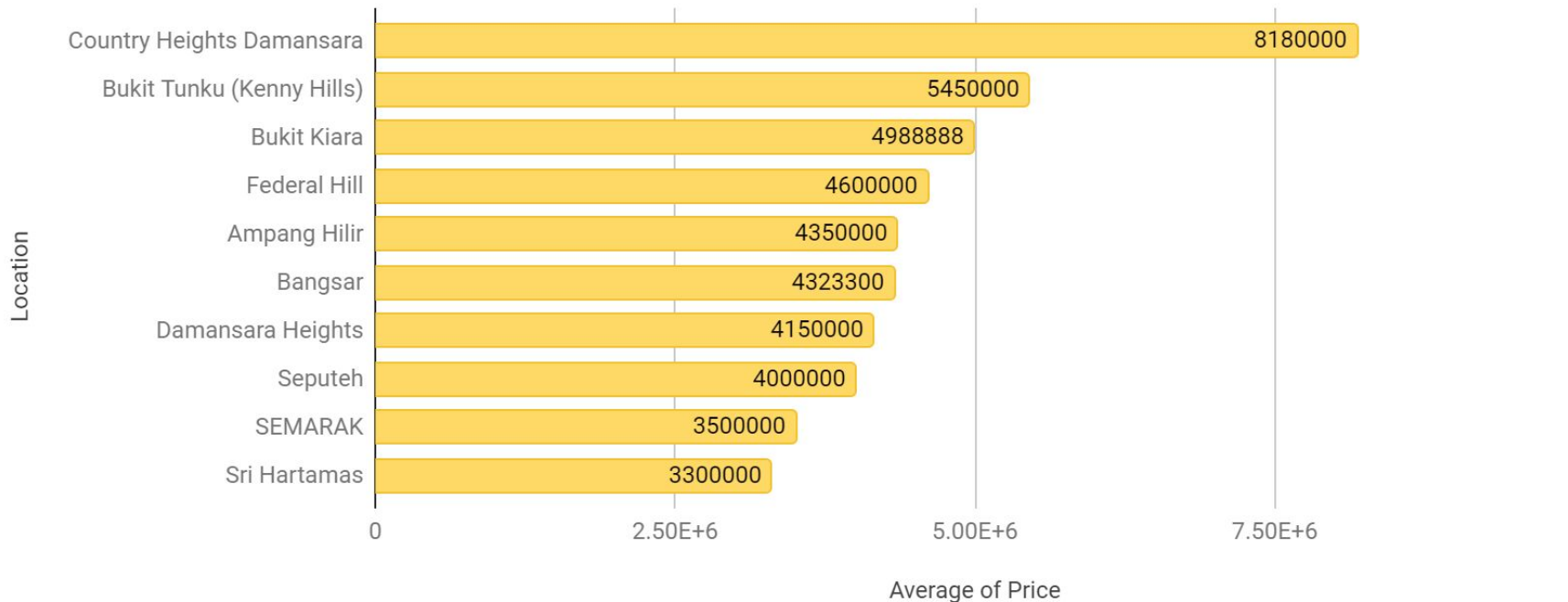
# Top 10 Location with Most Luxury Property Listed

Top 10 Location with Most Luxury Property Listed



# Top 10 Average Luxury Property Price by Location

Top 10 Average Luxury Property Price by Location



# Insights Gained

## Insights Gained of Luxury Property

The property room that have the highest percentage is property that have 4 room

The property bathroom that have the highest percentage is property that have 4 bathroom

The property car parks that have the highest percentage is property that have 2 car parks

The luxury property type that have the highest percentage in Kuala Lumpur is Condominium

The luxury property character that have the highest percentage in Kuala Lumpur is Built-up

The luxury property furnishing that have the highest percentage in Kuala Lumpur is Partly Furnished

The location with the highest number of luxury properties is Mont Klara with a total of 651 property units with a average price of RM 1,900,000. Luxury Property in Mont Klara has an average of 4 rooms, 4 bathrooms and 2 car parks with a property of around 2,535 sqm.

The location that has the highest average luxury property price is Country Heights Damansara with a average price of RM 8,180,000, while the that has the smallest average luxury property price is Bandar Damai Perdana with a average price of RM 725,000

# Insights Gained

## Insights Gained of Luxury Property

The location that has the highest average luxury property is Country Heights Damansara with a average of 10,600 sqm, while the Region that has the smallest average luxury property is in KL City with a average of 861 sqm.

The location that has the most complete average number of facilities (rooms, bathrooms, parking) is Country Heights Damansara with 7 rooms, 8 bathrooms and 2 car parks.

The most sold property type in Mont Klara is Condominium with a total of 521 units sold, making Mont Klara the with the most Condominium types sold in Kuala Lumpur. Mont Klara also has a luxury property character that has been built which is dominant compared to those that are still in the form of land s.

The most selling property type in Country Heights Damansara is the Bungalow with a total of 18 units sold. However, the majority of existing luxury properties are still land s and have not yet been developed.

The most sold property type in Bandar Damai Perdana is the 2-sty Terrace/Link House with a total of 2 units sold.

KLCC is the location with the largest fully furnished luxury property, followed by Mont Klara in second place.

# Recommendation

# Recommendation for Affordable Property

We can focus our target market on prospective buyers with low income ranges or on millennials and new couples who want to own property at low prices in the South Bandar Tasik .

We can focus our target market on prospective buyers with high income ranges who want to own affordable properties in the Mont Klara .

We can add types of affordable properties that are fully furnished, because they have a lot of sales compared to partly furnished and unfurnished. Especially on condominium property types in the Cheras .

We can focus our sales on Serviced Residences in the Mont Klara, to increase sales.

We can focus our sales on flats in the South Bandar Tasik, to increase sales.

# Recommendation for Luxury Property

We can focus our target market on prospective buyers with high income ranges in the Country Heights Damansara by conducting campaigns for prospective customers with high incomes.

We can focus our target market on prospective buyers who have large families in the Country Heights Damansara because the properties there have a fairly large size and quite a lot of facilities.

We can focus our campaign on targeting potential buyers with modest incomes but who want to own luxury properties in the Bandar Damai Perdana.

We can focus on the condominium property type in our campaign because this property type sells the most, especially in the Mont Klara which has the most units sold in Kuala Lumpur.

We can focus on the Bungalow property type in the campaign on the Country Heights Damansara because that property type sells the best

We can increase the number of condominium property types that are fully furnished, because fully furnished condominium property types have more sales than partly furnished and unfurnished, especially in the Mont Klara.

# **Milestone 2:**

**Correlation, Regression,  
Hypothesis Testing**



# Background

Now we focus on analyzing property data in **Desa Park City**, to get interesting insights.



# Property in Desa Park City

## Characteristics



### **Average Price**

(± RM 1,600,000)



### **Average Rooms and Bathrooms**

(4 rooms with 3 bathrooms)



### **Average Car Parks**

(2 car can park)




### **Average Size**

(± 1,700 sqft)




### **Total Unit**

(177 unit)



Before making a decision based on our analysis, we need to fulfill the **assumption of linear regression** so that later our regression can be accepted.



# Correlation

	<i>Price</i>	<i>Rooms</i>	<i>Bathrooms</i>	<i>Car Parks</i>	<i>Size</i>
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

There are indications of multicollinearity, due to the strong correlation of several independent variables, which is **more than 0.6 (room with bath and bathroom with size)**

# Linear Regression

Regression Statistics	
Multiple R	0.8601625255
R Square	0.7398795703
Adjusted R Square	0.7338302579
Standard Error	435863.0007
Observations	177

From the results of the regression carried out, it can be seen from the Adjusted R Square value **which is more than 0.6**, it can be concluded that the regression carried out is quite good.

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4	92942645214849	23235661303712	122.3080461	0
Residual	172	32675967530914	189976555412		
Total	176	125618612745763			

If seen from the Significance F value is **less than 0.05**, it can be concluded that the regression is quite good and all variables collectively have a significant effect on property prices.

# Linear Regression

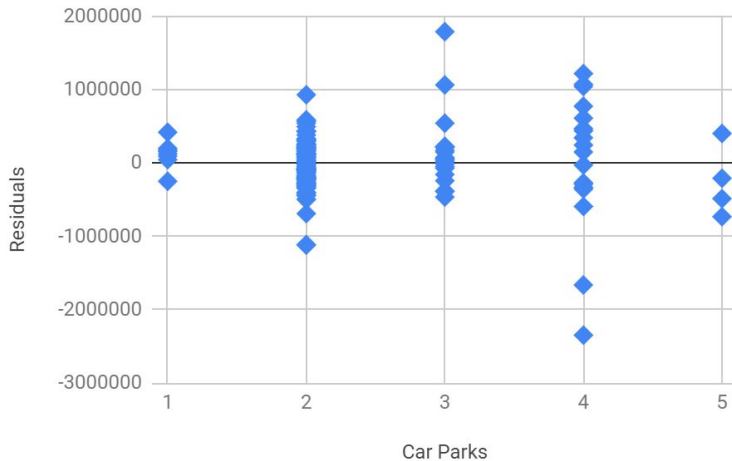
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-211210.5351	115518.5799	-1.828368521	0.06922654636	-439227.1302	16806.05996	-439227.1302	16806.05996
Rooms	71340.88497	47631.5609	1.497765003	0.1360272523	-22676.77298	165358.5429	-22676.77298	165358.5429
Bathrooms	151347.8424	45059.53782	3.358841429	0.0009638828928	62406.9777	240288.7071	62406.9777	240288.7071
Car Parks	160744.5183	54500.31337	2.949423743	0.003626963817	53168.95862	268320.0779	53168.95862	268320.0779
Size	472.1989033	58.26250062	8.104679653	0	357.1973423	587.2004643	357.1973423	587.2004643

If seen from the P-value, it can be seen that the rooms variable has a P-value of more than 0.05, so it can be concluded that the rooms variable has no significant effect on property prices independently and is taken into account to be excluded from the analysis.

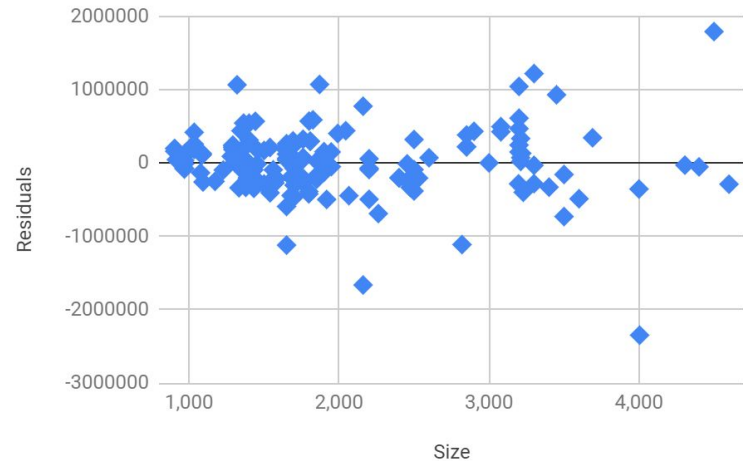
If we look at the Coefficients, it can be seen that the car parks variable has the greatest influence on property prices, but this cannot be used as a reference for determining property prices, there are still many other factors and variables that can affect property prices. So it is more appropriate that the Coefficients value of car parks makes property prices tend to rise.

# Linear Regression

Car Parks Residual Plot



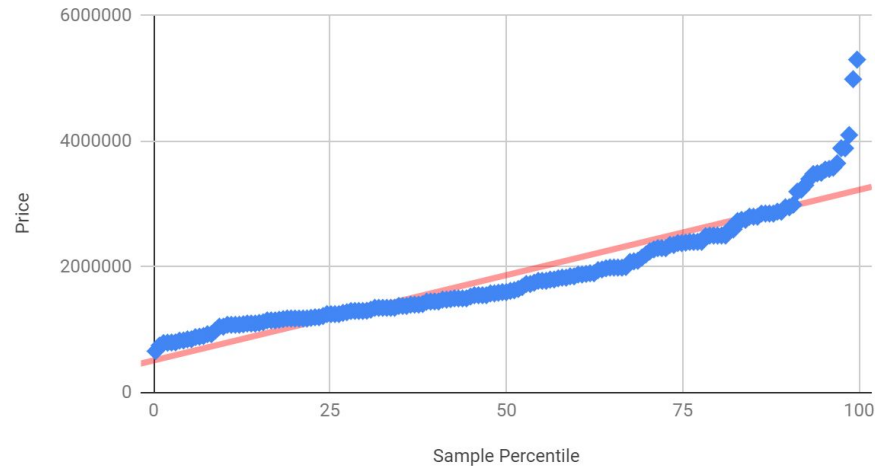
Size Residual Plot



There are two variables that are included in Heteroscedasticity, because the data collects in a certain number range.

# Linear Regression

Normal Probability Plot



If you look at the trend line of the normal probability plot, it can be concluded that the distribution is good (**Normal Distribution**). **With a note, at a high price prediction the error is likely to be greater**



# Price Prediction

It should be underlined that our data **fails to meet the classical assumptions**, but if the classical assumptions are ignored, the resulting price prediction model is:

**Before Remove Multicollinearity**

$$\text{Price (y)} = -211210.535112713 + 71340.8849710077(\text{rooms}) + 151347.842417535(\text{bathrooms}) + 160744.518274569(\text{car parks}) + 472.19890329509 (\text{size}) - 435863.000737948$$

So based on this equation, if the customer want the property that have **3 rooms, 4 bath rooms, 3 car parks, and an area of 2200 sqft**, the property price will be **RM 2.129.274.63** or we can also make a range of price predictions by adding/subtracting price predictions with Standard Error to **RM 1.693.411.63 - RM 2.565.137.63**.

# Price Prediction

It should be underlined that our data **fails to meet the classical assumptions**, but if the classical assumptions are ignored, the resulting price prediction model is:

**After Remove Multicollinearity**

$$\text{Price (y)} = -141623.842689144 + 190888.51418757(\text{bathrooms}) + 176806.815225346 (\text{Car Parks}) + 488.231517900416 (\text{Size}) - 437426.406287077$$

So based on this equation, if the customer want the property that have **4 bath rooms, 3 car parks, and an area of 2200 sqft**, the property price will be **RM 2.226.459,99** or we can also make a price prediction range by adding/subtracting price predictions with Standard Error to **RM 1.789.033,59 - RM 2.663.886,4.**

# Recommendation

If there are customers who are looking for a property with 3 bathrooms and 3 rooms with an area of more than 2200 sq ft and affordable prices. **Bungalow type is the best choice for them.** The bungalow type is also ranked 2nd out of the property type which has the highest average price in the luxury property category. By selling it to customers, **we can gain high revenue.**

However, because our regression model **fails to meet the assumption of linear regression**, we need to consult this matter with our head data so that we can make more informed decisions. The steps that can be taken are to do a tukey's test or examine our data further and verify it.

# Thanks

## Contact me:

rafiqnaufal97@gmail.com

+62 812 809 05778

linkedin.com/in/rafiqnaufal

Please keep this slide for attribution

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

