Aston University
BIRMINGHAM UK

DESCRIPTIVE ANALYTICS ASSIGNMENT

(Housing Market, Jaipur India)

DECEMBER 17, 2021

YOGENDRA SINGH
210216123

Table of Contents

Executive summary

While purchasing a house one looks for an area of the property, the number of bedrooms it has, location, bathrooms, hall, facing, etc. But in the real world what are the factors that decide the price of the house that can be understood by designing a model which could predict the significance of these house traits on the cost of the property. In this report, I have collected a sample of 100 out of a population of 140 which includes price of the house, area in sqft, Number of bedrooms, number of bathrooms, balconies, parking, facing of the house, and type of the house. In the first section, I visualized the relationship of factors on price with the help of graphs. To understand basic statistics of the data such as mean, median, mode, standard deviation, etc because then only we can analyse our data. To understand certainty with what our sample resembles population I used the 95% interval method. Performed two-sample t-test which helped me to know whether there is an association of mean price of sample to mean price of population. Designed a co-relation matrix with which I got to know which independent variables (area, bedrooms, bathrooms, parking, etc) have a major impact on dependent variable price and with what strength they are associated with each other. To make predictions about population and to check with what value our price gets altered if we increase or decrease the value of independent significant variables, I performed regression analysis, which not only helped me to know the power of our model in predicting values but also introduced me with the errors or residuals we can expect in our model by following major 5 residual assumptions. In conclusion, I checked the prediction of our model while inputting values from the sample to it and made a prediction.

Introduction

Aim: To give prediction about the population of Housing market in Jaipur, collected a sample of 100 and to know whether this model best describes the purchasing behaviour of Customers and what are its usage in real world to what extent. Also, can it be further improved?

Source: Magicbricks.com

Limitations: On some data area was mentioned in terms of carpet area, super area and so on, so it was creating confusion while collecting data.

Sampling method: Random sampling

To understand the purchasing behaviour in detail i.e., does a number of bedrooms play a vital role in the price of the property, or is it the Area of the property that decides the cost. To find answers to such questions I have collected a sample of 100 properties data from a population of Jaipur, India housing Market. Why have I taken a sample of 100? Because several samples average gives an idea about population average, as a mean of samples follows central distribution and that gives an idea about the population mean. I preferred random sampling for the sampling method because every element of the population incur an equal chance of being selected in the sample also eliminates biases in data. My sample looks like this

Price	Bedrooms	Bathrooms	Balconies	Type	Area(sqft)	location	Facing	Parking
16500000	4	4	3	Unfurnished	4500	Vaishali Nagar, J	East	1
11000000	3	3	1	emi-Furnishe	1800	alviya Nagar, Jaip	South	1
30000000	4	5	3	Furnished	4500	ck, Amrapali Circl	South	2
21100000	4	4	2	Furnished	4500	Malviya Nagar, Ja	West	1

Some Important information:

Independent variables: Area, Bedrooms, Bathrooms, Balconies, Type, Facing, location and Parking

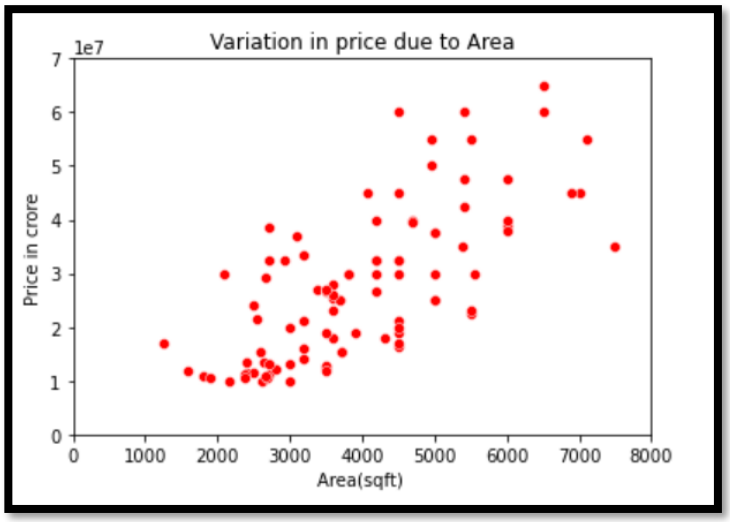
Dependent Variable: Price

Significance level: The significance level is the probability of rejecting our null hypothesis when it is true.

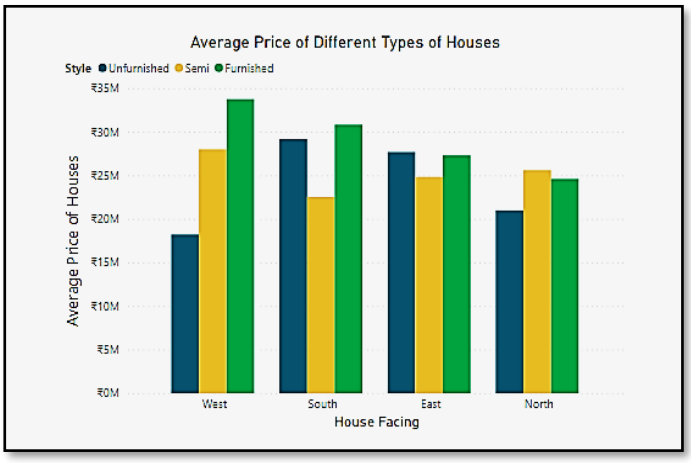
Normal Distribution: In simple words it's the symmetry around the centre

Data Visualization

Visualization of graphs depicts the relations and direction of the data. For the Housing data let's understand the relationship between price and area of the house. I used scatterplot because it best describes the relationship of two continuous variables.



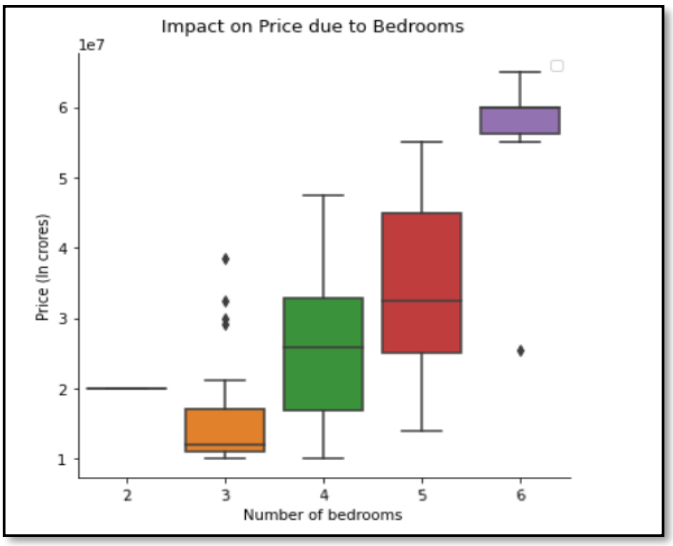
A) From the graph, it's clear that there are more houses available under 2500qft-4500sqft and price to these houses ranges from INR 2 crore-INR 4.5 crore. This graph also depicts that there are a smaller number of 6000sqft-7000sqft houses that means very few people willing to pay more than INR 6 crores also very few want house less than 2000sqft area. This graph shows positive relationship between price of the house to area of the house. Because of Tufte's principle of graphical integrity designed graph is proportional to numerical values as area is varying with respect to area, it's clear and detailed through labelling. We can see data variation with respect to Price.



B) To understand the impact on price due to Categorical value (Type of the house, House facing) I used the following bar graph to show the Average of different types of houses in four facing Majority of the furnished houses have "WEST" facing and least in "NORTH" facing because Indian follows the **science of dwelling**. That's why they purchase houses which are west faced, However, type of Houses Is not justified as there are few unfurnished houses at INR 3 crore also at INR 4 crore but furnished at INR 1 crore. Also, the second preference goes to the West facing wherein most of the houses are furnished. As per Tufte's principle of graphical integrity

designed graph is data driven and clear. we can see data varying with respect to Price also, information carried by variables not exceeding dimension. We can see proper ink ratio for the data. Object belonging to same group are placed close to each other so we can say its following Gestalt law of proximity and same colour to define type of houses so its also following gestalt law of similarity.

C) We can use a boxplot to understand the relation between the number of bedrooms with the Price of the House because boxplots provide an easy way to see the range and other statistics of a large group. we can observe that people who are willing to pay INR 2 crore to INR 4.5 crore can expect 4-5 bedrooms in their house. It is interesting to note that majority of the population want at least 3 bedrooms if they are paying more than INR 1 crore. The median of 4 and 5 bedrooms lies at INR 2.5 crore and INR 3.2 crore. There are some outliers (extreme values in data) in the graph which can be noticed by diamond dots. As per Tufte's principle we can observe provided data is proportional to Price range on y axis, it's clear and properly labelled and number of information is not exceeding number of dimensions. Moreover, this graph is following Gestalt law of similarity as one can understand that we have used varying colour to define relationship of number of bedrooms.



Descriptive Statistics

The statistic is a value that describes the dataset in some way. Information is presented in statistics in an easily analysed format, making conclusions easy to draw. Let's Understand some statistics on the housing market. From the following Statistic table, we have only considered a continuous variable

Price	
Mean	26992000
Standard Error	1394661
Median	25000000
Mode	30000000
Standard Deviation	13946606
Sample Variance	1.95E+14
Skewness	0.754812
Range	55000000
Minimum	10000000
Maximum	65000000
Sum	2.7E+09
Count	100
Confidence Level(95.0%)	2767309

Area	
Mean	3948.29
Standard Error	132.5677
Median	3700
Mode	4500
Standard Deviation	1325.677
Sample Variance	1757419
Skewness	0.5141
Range	6250
Minimum	1250
Maximum	7500
Sum	394829
Count	100
Confidence Level(95.0%)	263.0431

1)Mean price of houses is INR 2,69,92,000 but you can find variation (standard deviation: how data is dispersed around mean) and as per 1sigma rule 68% of our price lies in INR 1,30,45,394 to INR 4,09,38,606, and you can buy houses from INR 1,00,00,000 to INR 6,50,00,000 price range. We can observe that there are more houses available at INR 3cr from mode (what value repeated maximum time). Median of INR 2.5 crore represents the middle value when data is sorted in ascending order. Variance is nothing but square of standard deviation, Skewness in our statistics represents the symmetry of data here its 0.754 that means its normally spread and no ambiguity. Confidence level (95%) tell us that we are 95% certain that our mean of population price will lie in this Price range INR 2,42,24,691-INR 2,97,59,309

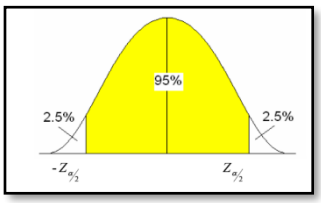
2) It's interesting to note that the Average area of the houses is 3948.29 sqft that means if you are paying an average of INR 2,69,92,000 you will get a house of 3948.29sqft area. However, you can find a deviation of 2622.61sqft to 5273. 97sqft.The area range of available houses is 1250-7500sqft. The middle value of the area is 3700sqft also there are 4500sqft houses available more. The skewness of 0.51 tells us that area is symmetrical distributed which means there are no huge gaps between values. The confidence level (95%) of the area tells that area of the population will lie in the Area range of 3685.296sqft-4211.33sqft

Association of Sample to the Population using 95% Confidence Interval

	Furnished	Semi-Furnished	Unfurnished
Lower Bound	25693099.02	22525730.82	21014810.7
Upper Bound	34551345.43	29711306.21	28385189.3

95% Interval: From the collected sample data, we are making predictions about the Population keeping in mind that the average of samples means follows a normal distribution, here We are using a 95% Confidence Interval to predict the probability of parameter about the population. From the calculation, we can confirm with 95% certainty that means the price of a Furnished house will fall under cost INR 2,56,93,099.02 to INR 3,45,51,345.43 in Jaipur, India. Whereas mean of Semi-Furnished will fall under INR 2,25,25,730.82 to INR 2,97,11,306.21 and Unfurnished houses are cheaper, and their mean will lie between INR 2,10,14,810.7 - INR 2,83,85,189.3

However, because 5% of the mean is outside the 95% range of the sample mean distribution, there is a 5% chance that the specified cost interval will not include the population mean.



Two sample t-test: To make significance about the population we will make a conclusion using a one-sample t-test because it determines whether the mean of those two samples is significant or not. As per the website, https://www.propertywala.com average price of a property in Jaipur is INR 7,256 per sq. ft, and as per our collected data mean area of the sample is 3948 sqft so that implies an average price of the population should be INR **2,86,46,688.**

H0=**Average price of sample represent average price of the population (INR 2,86,46,688)**

Ha= **Average price of a sample doesn't represent the average price of the population (INR 2,86,46,688)**

We will check this from the significance level of 0.05. If the value is below 0.05, we will reject our Null hypothesis *H0* and if it's greater then we will accept our Null hypothesis *H0*. Why because the central limit theorem states if average of sample means follows a normal distribution, then the sum of their mean describes population mean. But if it falls in rejection area of 5%, however it also follows normal distribution but of different population mean that's why we accept Ha and reject H0. From the conducted test we found (p_value:0.2386)

greater than our significance level (0.05), so we have to say our average price of our sample represent average price Houses in Jaipur.

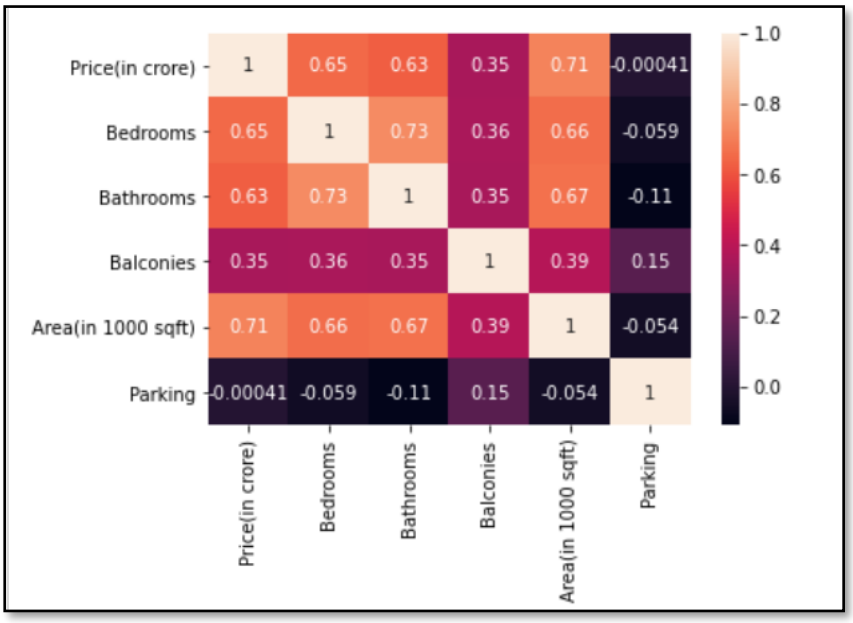
For reference we can follow this code which ran on python.

```
t_val,p_val=stats.ttest_1samp(df["Price(in crore)"],28646688)
if p_val<0.05:
    print(f"p value:{p_val} is less than 0.05 so we will reject null Hypothesis")
else:
    print(f"p value:{p_val} is greater  than 0.05 so we will accept  null Hypothesis")
```

```
p value:0.23828609156124958 is greater  than 0.05 so we will accept  null Hypothesis
```

CO-relation Matrix

The Co-relation matrix is no doubt a useful matrix it shows what strength variables are co-related and its values remain between -1 to 1. If values are 1 or -1 that means two variables are highly co-related but if -1 it is also highly co-related but negatively which means they are inversely proportional. Let's understand with the help of co-relation matrix



As you can see Area (in sqft) is highly significant to Price (in crore) having value of 0.71 that means Area of house plays very crucial role in deciding the price of the house. Bedrooms, Bathrooms having p value:0.65 & 0.63 also significant but lesser than Area. Parking of the house is not showing major impact in price of the house having p value (-0.00041).

Balconies have very little impact on the price of the house because its value is not good enough (0.35). If we consider co-relation with the independent variable, we will notice Bedrooms and Bathrooms are highly co-related with (strength of 0.73) also Area showing good strength in co-relating with Bathrooms and Bedrooms. From the co-relation matrix, we can't see any variables which have a major negative impact on price. In conclusion, we can say the Area of the house, Number of bedrooms and number of bathrooms are the key factor in deciding the price of the house.

Regression Analysis

In step 6 we measured the significance on the dependent variable (Price) due to independent variables (Area, number of bathrooms, number of bedrooms, etc) also by what strength they are related to each other. But by how much value our dependent value gets impacted by adding or removing 1 unit from independent variables can be measured by a model. And the strength of our model can be measured by the coefficient of determination also known as R. square or adjusted R. square.

We are considering the assumption that all independent variables have a linear relationship because linear regression only accepts linear relations. Linear regression is nothing but a linear equation for a model which has **a) intercept**: Default value when values of all independent variables are zero, **b) Slope**: coefficient of independent variables, that describes by how much our dependent variable will be impacted if I unit added or removed, **e) residuals**: these are the remaining factors which could impact the value of the dependent variable but not considered.

Using equation Y=a+b1x1+b2x2+b3x3+……+e and following are the steps to get our model

1)We introduced dummy values 0 or 1 to our variables like facing, Type because they don't have numerical values so to calculate their impact, we assigned values 0 or 1. However, in model it will remove the most repeated categorical value like in Facing: East

2)We first added all independent variables into model and noticed their t value and Value of R. square and found that there are few variables which has no significance in deciding the model(p_val>0.05) Like "Parking"," Balconies"," Facing"," Bathrooms" ,"Type" so we removed one by one by eliminating the highest p_val variable first and so on and noticed strength of our model by R2 and adjusted R2 and at last re-evaluated the model.

From the following model summary, we can see highlighted values variables have no significance on dependent variable Price (In crore). Goodness of fit can be observed by Adj.R-Squared value, which is 0.542 it's an average model, if it would have been more than 0.7, we could say it's a good model.

Dep. Variable:	Price	**R-squared:**	0.584
Model:	OLS	**Adj. R-squared:**	0.542
Method:	Least Squares	**F-statistic:**	13.87
Date:	Sun, 12 Dec 2021	**Prob (F-statistic):**	1.10e-13
Time:	02:13:49	**Log-Likelihood:**	-1724.3
No. Observations:	99	**AIC:**	3469.
Df Residuals:	89	**BIC:**	3495.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.379e+07	5.13e+06	-2.690	0.009	-2.4e+07	-3.6e+06
Type[T.Semi-Furnished]	1.493e+04	2.66e+06	0.006	0.996	-5.27e+06	5.3e+06
Type[T.Unfurnished]	-2.741e+06	2.46e+06	-1.112	0.269	-7.64e+06	2.16e+06
Facing[T.North]	-2.785e+06	2.88e+06	-0.968	0.336	-8.5e+06	2.93e+06
Facing[T.South]	-9.305e+05	3.17e+06	-0.294	0.770	-7.23e+06	5.36e+06
Facing[T.West]	-1.144e+06	2.46e+06	-0.466	0.643	-6.03e+06	3.74e+06
Area	4615.9107	1059.329	4.357	0.000	2511.046	6720.776
Bedrooms	4.129e+06	1.72e+06	2.403	0.018	7.15e+05	7.54e+06
Bathrooms	1.537e+06	1.49e+06	1.034	0.304	-1.42e+06	4.49e+06
Parking	6.997e+05	1.22e+06	0.574	0.567	-1.72e+06	3.12e+06

3)After removing non-significant variables our R. square and Adjusted R squared values improved to little extent and we got our most **parsimonious model** as:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.751[a]	.565	.556	9297247.0581	.565	62.887	2	97	<.001

a. Predictors: (Constant), Area(in 1000 sqft), Bedrooms

b. Dependent Variable: Price(in crore)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14595591.429	4456526.311		-3.275	.001
	Bedrooms	5087675.880	1420898.042	.320	3.581	<.001
	Area(in 1000 sqft)	5262.784	940.279	.500	5.597	<.001

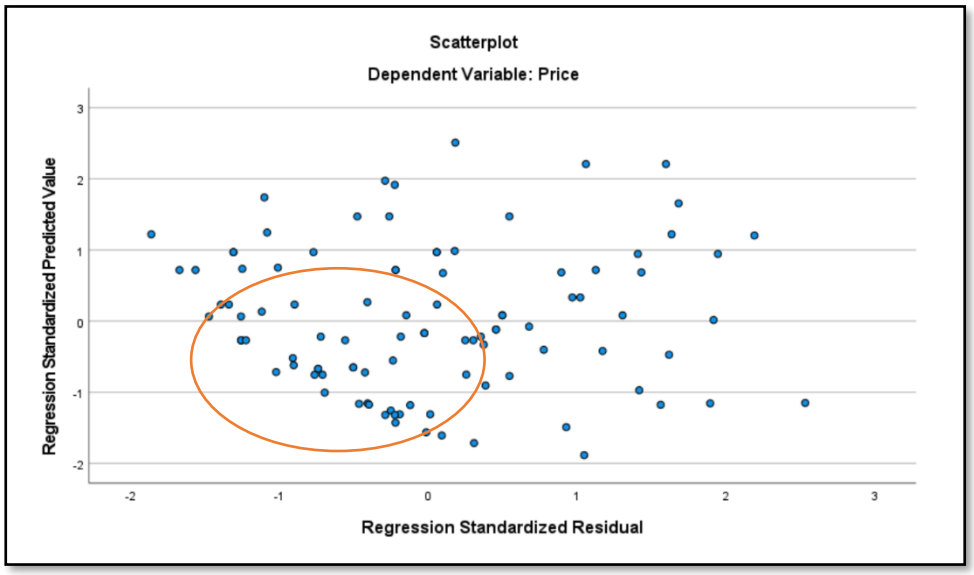
a. Dependent Variable: Price(in crore)

Price (in crore) = -1,45,95,591.429 + 5,087,675.880*Bedrooms + 5,262.784*Area (in 1000 sqft)

Which means if we add 1 bedroom, Price will be altered by INR 5,087,675.880 and 1 sqft of additional area will change price to INR 5,262.784.

Residual Analysis

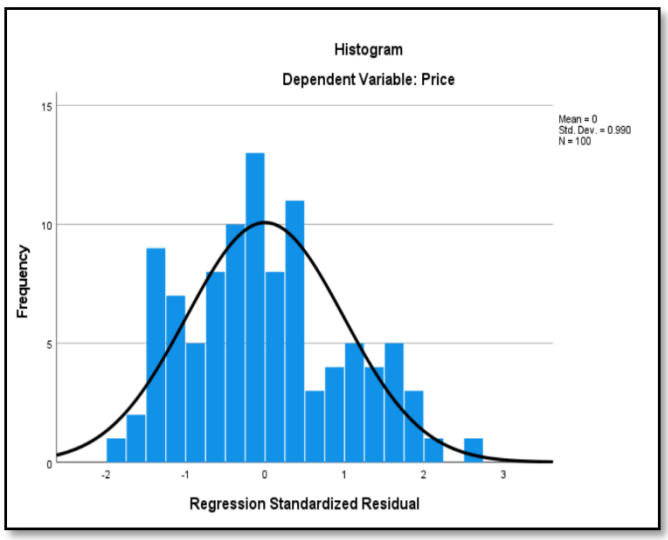
While deciding our model there were certain parameters or factors which we couldn't include but might have an impact on our most **parsimonious model** are known as residuals. And to understand their impact we will do residual analysis by following five assumptions:

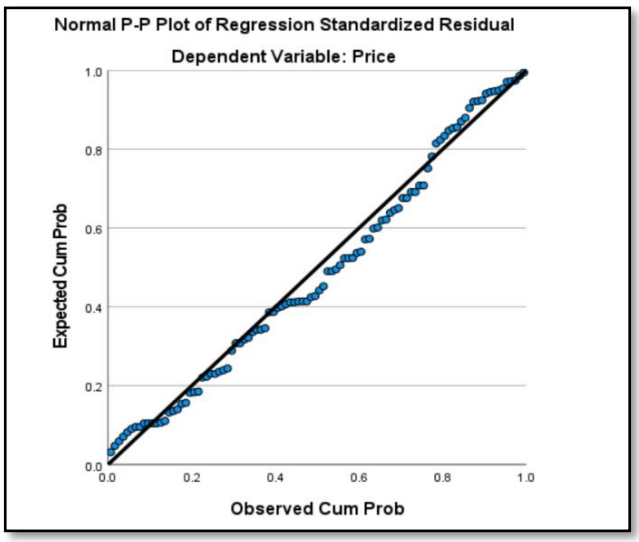


1) In the above graph of the Regression standardized residual, we can see that there are some joint residuals in a marked circle which raises the question on perfectibility of our model

2) There is no clear pattern that can be observed in the graph

3): Distance of residuals or errors around zero is the same.





4): From the above graph we can see that all the residuals are not following perfect normal distribution as values to the left have more peaks than the right which shows our model is not good enough.

5): In our derived model we have considered multicollinearity between area and Bedrooms because keeping them making our model strong i.e., Increasing R square value.

So, all our residual assumptions didn't fit well so we can say our model is not perfect enough from our obtained Coefficient of determination or R. squared value to be 0.556, it could have been better if it was above 0.7. Also, our residual analysis is not following all assumptions perfectly. Hence, we can say our model is not perfect to predict values.

Now the question comes how come can we improve our model. There are some instances in the data where we can see 6 bedrooms at 6cr and same number of bedrooms at 3cr which are outliers.

Other factors such as age of the property, main road in front of the house, Major market besides, near malls, near schools could have been considered while collecting the data which haven't been included. It can be improved by adding some more independent variables and removing the above exceptions.

Usage of Derived model

From the following derived statistical model now, we will now predict value by inputting values from the collected data.

Price (in crore) = -1,45,95,591.429 + 5,087,675.880*Bedrooms + 5,262.784*Area (in 1000 sqft)

As we can notice some price difference between our predicted and expected price. Hence, it's clear that our model doesn't justify the usage.

Predicted Price	Expected Price	Difference Price
34525315.97	45000000	10474684.03
24701134.49	18000000	-6701134.491
27683696.77	14000000	-13683696.77
39788099.97	22500000	-17288099.97
47682275.97	45000000	-2682275.971
11368112.33	20000000	8631887.669

References

https://www.propertywala.com

https://statisticsbyjim.com/glossary/significance-level/

https://excelcharts.com/data-visualization-excel-users/gestalt-laws/

https://www.magicbricks.com/property-for-sale-rent-in-Jaipur/residential-real-estate-Jaipur

https://community.mis.temple.edu/mis5208sp18/2017/04/22/edward-tufte-s-principles-of-graphical-integrity/