MY DREAM HOUSE

Business Report: Statistical Analysis of a Real World Data Set- SW3 postcode

TABLE OF CONTENTS

SUMMARY 3STATISTICAL

ANALYSIS 4 DATA COLLECTION

4DATA PREPARATION AND MANIPULATION---------- 4

SAMPLING METHOD 5 VISUALIZATION

6 DESCRIPTIVE STATISTICS 8

95% CONFIDENCE LEVEL INTERVAL----------9

LAVENE'S TEST 10

ONE SAMPLE T-TEST 10

CORRELATION ANALYSIS 11

REGRESSION ANALYSIS 11

RESIDUAL ANALYSIS 12

DERIVED STATISTICAL MODEL 14

APPENDIX 1 15

APPENDIX 2 16

APPENDIX 3 18

MY DREAM HOUSE

SUMMARY

The model focuses on studying the house market for **SW3** postcode of **City of London**. It is considering the properties which are on Sale. This summary focuses on predicting the house price of SW3 postcode based on various factors considering like No. of Bedrooms, No. of Bathrooms, No. of Stations, House types.

The project aims to draw the statistical model to depict the house price and identify the predictors which have significant effect on outcome of Price.

However, on looking at the parsimonious model, it is evident that only three predictor variables i.e No. of bedrooms, house type flat and house type studio plays significant role in depicting the house price.

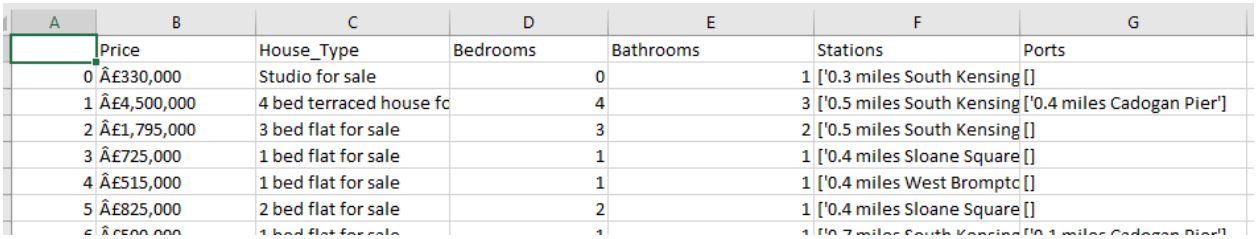
As per my conclusion, Flat should be considered as the best choice as compared to other house types for selecting house in SW3 postcode but it should not be termed as a best model since the independent variables explains only 67.6% of the variation for the dependent variable i.e Price.

STATISTICAL ANALYSIS

The model aims to study the house market data of SW3 postcode of City of London and find the predictors which plays an important role in determining the selling price of the typical house of that region. This model and all the necessary analysis would help the manager of the real estate agency to form a decision and would also indeed to help the clients to opt for particular house depending on their affordability, house type choice, preferable no. of bedrooms and no. of stations near the house.

DATA COLLECTION:

The data has been collected from Zoopla which is one of the websites for viewing the properties for rent/sale in United Kingdom. The data has been gathered by web data scrapping using BeautifulSoup library in Python. Through web scrapping, we could extract the HTML elements by inspecting the classes containing useful variables to perform our analysis.



	A	B	C	D	E	F	G
		Price	House_Type	Bedrooms	Bathrooms	Stations	Ports
0		Â£330,000	Studio for sale	0	1	['0.3 miles South Kensing	[]
1		Â£4,500,000	4 bed terraced house fc	4	3	['0.5 miles South Kensing	['0.4 miles Cadogan Pier']
2		Â£1,795,000	3 bed flat for sale	3	2	['0.5 miles South Kensing	[]
3		Â£725,000	1 bed flat for sale	1	1	['0.4 miles Sloane Square	[]
4		Â£515,000	1 bed flat for sale	1	1	['0.4 miles West Bromptc	[]
5		Â£825,000	2 bed flat for sale	2	1	['0.4 miles Sloane Square	[]

DATA PREPARATION AND MANUPLATION

The data obtained through web scrapping was not structured, so data preparation and manipulation was performed using R software in order to clean the data and make it suitable for analysis.

The steps taken for the same are as follows:

1. Price- Removed Â£ appended before the numeric values.
2. House type- Created a separate array to store all the possible house types by looking at the data. The house types were found to be Terraced, Property, End Terrace, Houseboat, Mews, Semi-detached, Flat, Town and Garage.

   Extracted the substring from House type column according to the values in array and stored it to new column

   For Example: For one of the entry for House_Type column is - 3 bed flat for sale

   As only ‘Flat’ is of our use here so it will match the newly created values of the array with substring ‘flat’ from string value ‘3 bed flat for sale’ and store the result in another column for this record as ‘flat’.

Frequency of the house types from the data file was recorded and only those house types were considered as a part of the population whose occurrence is more than 10.

Resultant house types along with frequencies:

flat	property	semi-detached	studio	terraced
393	16	11	71	78

Moreover, a separate column was designed which would assign numerical values for the above categorical house types.

3. Bathrooms – For several entries the no. of bathrooms was 0 where no of Bedrooms were more than 0. So for such cases no. of bathrooms were manipulated by dividing the number of bedrooms by 2 and applying the round off function to it.
4. Bedrooms - For several entries the no. of bedrooms was 0 where no of Bathrooms were more than 0. So, in such cases the similar value of no of bathrooms was assigned to bedrooms.
   Also, for the Studio property type the no of bedrooms and bathrooms were both 0 or either of them were 0. So, in such scenarios both the variables were assigned to value 1 as only 1 bed studio types houses existed in my population.
5. Stations- The sum of total no. of stations was calculated for each of the record from the original data population from the file.

   Ports was also being considered but as its frequency is less that Stations it was dropped before analysis.

	Price	Bedrooms	Bathrooms	stations_total	House_type_order
1	330000	1	1	2	4
2	4500000	4	3	1	5
3	1795000	3	2	2	1
4	725000	1	1	2	1
5	515000	1	1	1	1

So as a result, the cleaned data was ready for the analysis which was performed through SPSS.

Some of the key changes performed after importing the structured data file is as follows:

- The data type for all the variables is treated as continuous except the House_type_order which is treated as categorical type.

- Bathrooms variable is taken as continuous variable as the population has recorded maximum no. of bath rooms as 7. So, it is not feasible to create 6 dummy variables to depict individual categories for different no. of bathrooms as the predictor variables.
- For the five house types, four dummy variables was created as this would better explain the different ho use types which could explain the model.

SAMPLING METHOD

Random sampling technique is used in order to figure out the sample which would represent the data. This technique is common and preferable as each member of the population stands an equal chance of being selected as subject.

Statistics for Population

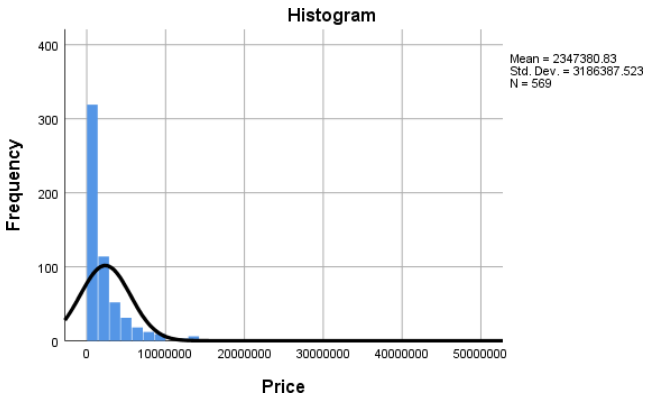
Price

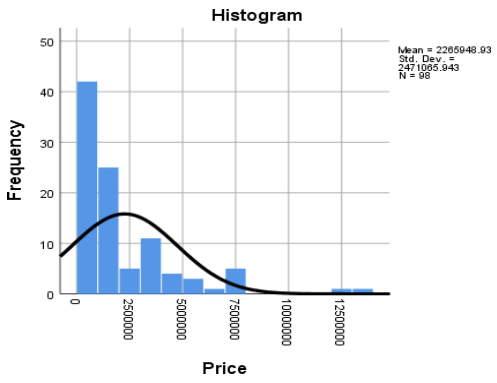
N	Valid	569
	Missing	0
Mean		2347380.83
Median		1295000.00
Std. Deviation		3186387.523

Statistics for Sample

Price

N	Valid	98
	Missing	0
Mean		2265948.93
Median		1312500.00
Std. Deviation		2471065.943



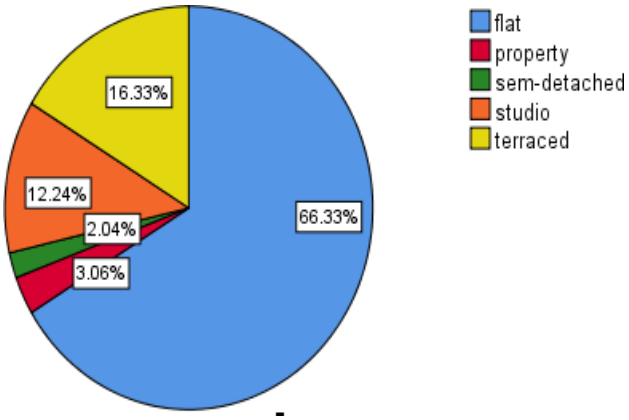


Sample of size 98 is selected to represent the population as its mean 2265948.93 is closer to the mean of population 2347380.83 of population of size 569 on assumption normal distribution of data.

Standard deviation can also be used to provide a relationship between sample and population. However, it cannot be used in this case as the data is skewed and follows non-normal distribution which is evident from comparing the histogram of the sample and population. We can clearly see that the data is positively skewed as the frequency distribution of the price doesn't entirely overlap with my normal curve.

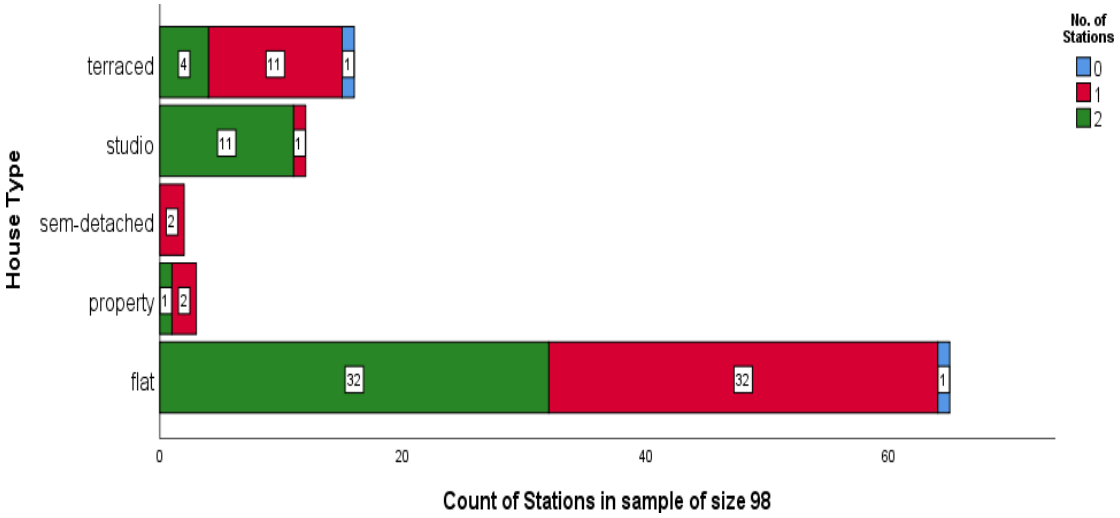
The limitation that was encountered while analysing the data was the non-normal distribution of the data which seemed to affect the test conducted and result while predicting the significant variables impacting the Price variable.

VISUALIZATION



The pie chart shows that the house types are dominated by the flats as they account to 66.33% percentage in the sample data followed by terraced (16.33%) and studio (12.24%) house types in my sample data.
Similarity principle is used because of representation of different House types in different color in a single chart.

The line graph shows the relationship between No. of Bedrooms and Price. It depicts the linear relationship between them which shows that the Price increases with increase in Bedrooms. The maximum no. of Bedrooms recorded in the sample is 6 and the maximum Price recorded is near 14million pounds where no. of Bedrooms has come out to be 4.
This graph abides Proximity as there are not miss connection to proceed with continuity and the graph shows relation between two entities.



The inverted stacked bar chart shows the count of no. of stations with respect to different house types in our sample data. It is evident that the flats has maximum no. of stations as compared to other house types and dominates the rest by having 32 records each for houses having 2 stations and 1 stations respectively. It follows Similarity principle because each color represents different values.

DESCRIPTIVE STATISTICS

Statistics

Price

N	Valid	98
	Missing	0
Mean		2265948.93
Median		1312500.00
Minimum		200000
Maximum		13950000
Percentiles	25	725000.00
	50	1312500.00
	75	3262500.00

From the figure it is clear that the maximum price for the particular house type is around 70 times than the least one and its nearly 6 times the average price of the house. The average of the house sale is about 11 times the least sale value of the house.
It also depicts the quantile distribution of price of the sample data where median (50%) is nearly 1.8 times than the lower quantile or 25% of the Price data and the upper quantile or 75% of the Price data is 2.4 times the median.
The quantile distribution is not even which explains the non-normal distribution of the data.

Outliers were detected for Price variable and removed. Detail information on Appendix 3.

Descriptive Statistics for Prices group by House Types and No. of bedrooms

House_type_all	Bedrooms		N	Minimum	Maximum	Mean
Flat	1	Price	25	485000	1295000	784380.00
	2	Price	30	699995	3950000	1679149.83
	3	Price	7	1525000	5500000	3017142.86
	4	Price	2	2550000	3500000	3025000.00
Property	1	Price	1	895000	895000	895000.00
	2	Price	1	1500000	1500000	1500000.00
	3	Price	1	4000000	4000000	4000000.00
Studio	1	Price	7	425000	895000	597142.86
Terraced	2	Price	1	3650000	3650000	3650000.00
	3	Price	6	2495000	6250000	3607500.00
	4	Price	4	4500000	7200000	5850000.00
	5	Price	2	3950000	5650000	4800000.00
	6	Price	1	5750000	5750000	5750000.00

From the table, the terraced house type has recorded maximum no of Bedrooms i.e 6 as opposed to other house types and it is found to be costliest house type with 4 no. of bedrooms. If a client is looking for a three-bedroom house then opting for flat house type would be the best choice as its average price is cheaper as compared to property and terraced house types with same no. of bedrooms. Moreover, opting for studio would be the optimal choice if one needs to go for a single bedroom option.

95% confidence interval of the average house price per house type

House_Types	N	Std. Deviation	95% Confidence Interval for Mean	
			Lower Bound	Upper Bound
Flat	64	1109308.741	1240934.09	1795128.25
Property	3	1646058.423	-1957369.14	6220702.47
Studio	7	190544.333	420918.74	773366.97
Terraced	14	1594932.437	3653756.28	5495529.43
Total	88	1652330.187	1601881.90	2302072.53

From the above table we can infer that we are 95% confident that the mean price of the any house type is around at least 16 lakh pounds which ranges upto 23lakhs.

Adding to it, we are 95% confident that mean price for the upper range for Property is 8 times to that of studio.

Moreover, the minimum mean price for the price of property is in negative which implies that there is a great variation between the prices of the property house type which could not be treated well even after the removal

of the outliers. But this is could be due to non-normal distribution of the sample data for this particular house type as well.

95 % confidence interval is taken into consideration as compared to 99% as it is more specific than 99% and would not cover broader range of price values to fall in that interval region.

Further comparison analysis done for 2 different sample size to calculate avg. price for different house types at 95% C.I on Appendix 1

In order to establish the equivalency of the average prices of the house types in SW3 region with city of London as a whole, first we will perform **Levene's test** which is used to test the variation of price mean among all the different types of house types.
It is concluded from the test that there is significant difference of the variance among the prices of the different types of house types.
Further detailed analysis given on Appendix 2

ONE SAMPLE TEST Test Value=721490

		T	Df	Sig. (2-tailed)	Mean Difference
Flat	Price	5.744	63	.000	796541.172
Property	Price	1.484	2	.276	1410176.667
Studio	Price	-1.727	6	.135	-124347.143
Terraced	Price	9.039	13	.000	3853152.857

One Sample T Test has been conducted to test if the average price of the different type of houses in our data sample is in line with the average price in the City of London(721490 pounds).

DECISION AND CONCLUSIONS

Since $p < 0.05$ for Sig., we reject the null hypothesis that the sample mean is equal to the hypothesized population mean of **City of London** and conclude that the mean price of the house type **Flat** and **Terraced** is significantly different than the average price of the overall price for houses in City of London.

While we fail to reject null hypotheses and state that sample mean of house type **Property** and **Studio** do not differ significantly than the average price of the overall price for houses in City of London.

Based on the results, we can state that the average price of the house type Flat is about 2.1 times costlier, Property is about 3 times costlier, Studio is about 1.2 times cheaper, Terraced is about 6.3 times costlier than the overall price for houses in city of London

COORELATION ANALYSIS

Correlations

		Price	Bedrooms	Bathrooms	Stations_Total	House_type_flat	House_type_property	House_type_studio
Pearson Correlation	Price	1.000	.804	.740	-.196	-.431	.021	-.242
	Bedrooms	.804	1.000	.844	-.123	-.378	-.006	-.278
	Bathrooms	.740	.844	1.000	-.006	-.343	.056	-.281
	Stations_Total	-.196	-.123	-.006	1.000	.056	-.046	.289
	House_type_flat	-.431	-.378	-.343	.056	1.000	-.307	-.480
	House_type_property	.021	-.006	.056	-.046	-.307	1.000	-.055
	House_type_studio	-.242	-.278	-.281	.289	-.480	-.055	1.000

Correlation analysis is performed to measure the strength of the linear relationships between the independent variables and the dependent variables. From the correlations table, it is evident that Price and Bedrooms are highly positively coorelated with each other(r=0.804) followed by positive Price and Bathrooms (0.740).

Positive coorelations means that if one entity increases then other entity increases as well.

REGRESSION ANALYSIS

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	859557.718	470096.987		1.828	.071
	Bedrooms	962543.428	120452.121	.640	7.991	.000
	House_type_flat	-1056057.026	323297.715	-.286	-3.267	.002
	House_type_studio	-1224958.288	512880.721	-.202	-2.388	.019

a. Dependent Variable: Price

Regression analysis was carried out in order to come up with all the independent variables which play a significant part in predicting output of price. This was done by looking at the Sig column of the above table and removing the variable which has the highest p value until we get the variables whose p value is less than 0.05.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.829[a]	.687	.676	941186.293

a. Predictors: (Constant), House_type_studio, Bedrooms, House_type_flat

b. Dependent Variable: Price

The **adjusted R square** of my most parsimonious model is 0.676 which means that my model is able to explain only **67.6%** of the variance for the dependent variable i.e. Price is explained by independent variables- Bedrooms, House Type Flat and House type Studio.

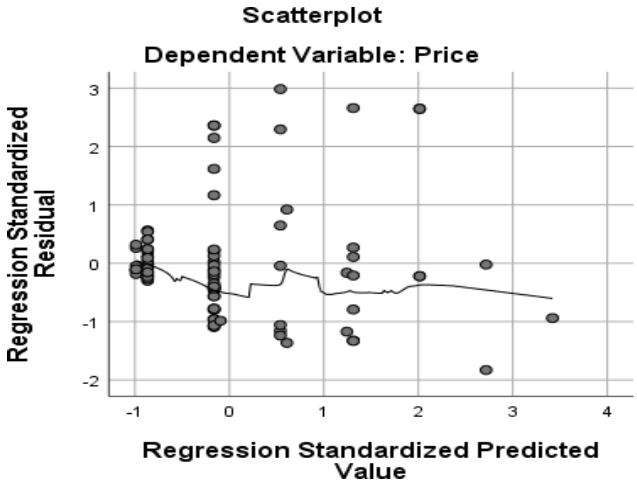
On analyzing the coefficients table further, it can be interpreted that:

1. If the no. of Bedrooms is increased with 1 then the Price will be increased by 962543.428 pounds keeping other variables constant.
2. House Type Flat and House Type Studio are inversely proportional to Price. If Price will increase then it will decrease and vice versa. For Instance if no. of Flat is increased by 1 then the Price would reduce by 1056057.026 pounds.
3. Bedrooms has most significant impact on Price followed by house type Flat and Studio.

After getting the parsimonius model, multicollinearity issue was assessed again by looking at the coorelations table and it found out to be that our model is free from multicollinearity issue.

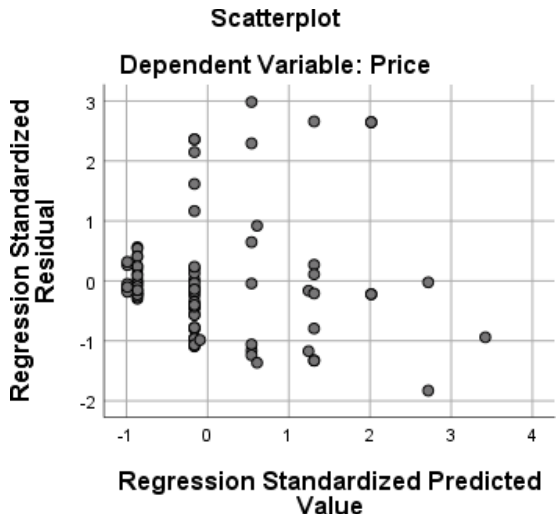
This factor is important to access as they try to explain the same portion of difference in the values of the dependent variable.

RESIDUAL ANALYSIS



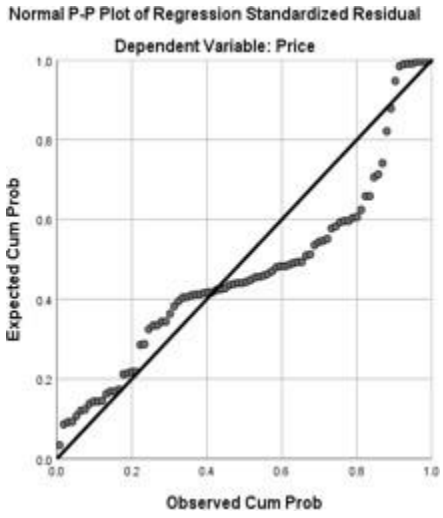
Testing Nonlinearity

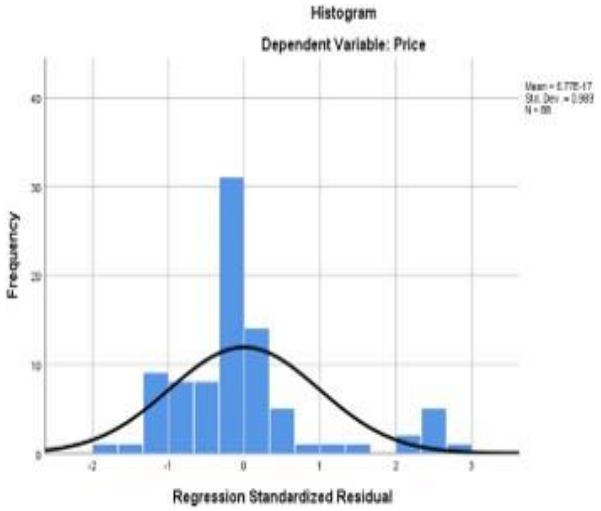
We can conclude that the relationship between the Dependent variable (price) and predictors is hardly linear around 0 zero since the residuals seem to be randomly scattered around zero.



Testing homogeneity of error variance

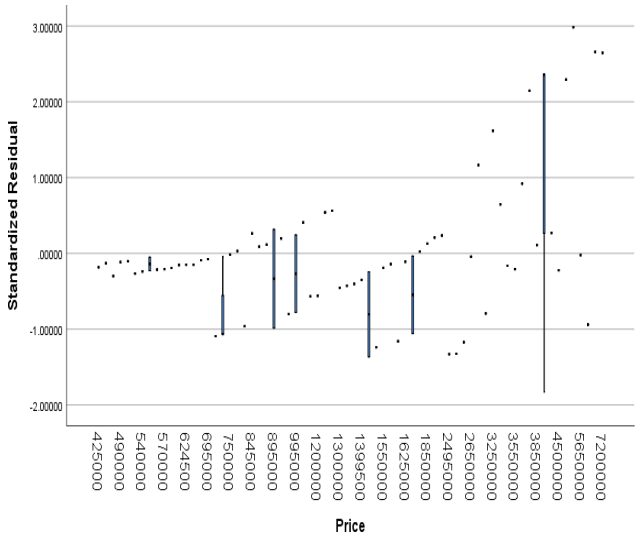
It violates the assumption as the points on the graph are deviating a lot and points above 0 and below 0 are different and not randomly scattered.





Tests on Normality of Residuals

The assumption is violated as the data points are deviated from the line and not following normal distribution.
Also, from the histogram it is evident that the Price data is not normal since it is positively skewed.



Issue of Independence

This suggests that the errors are not independent and there is a possibility that Price tends to have different mean residuals not centred at zero.

DERIVED STATISTICAL MODEL

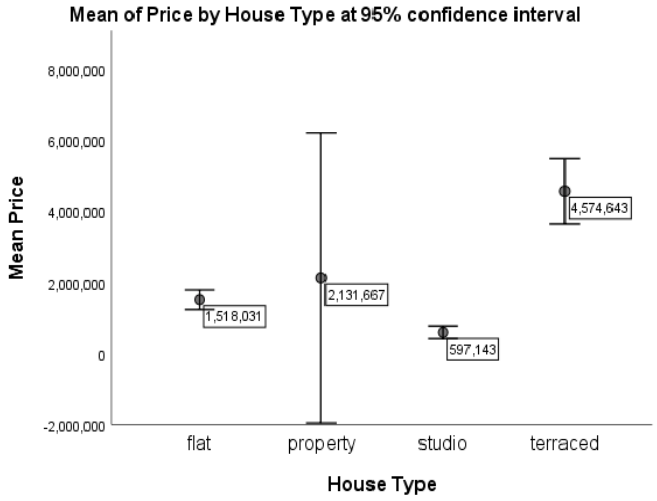
Price= 1095344.838 + 686295.168* Bedrooms + 411640.258*Bathrooms-215174.175*Stations_total-1136429.430 *House_type_flat-872650.124*House_type_property-1165789.058* House_type_studio

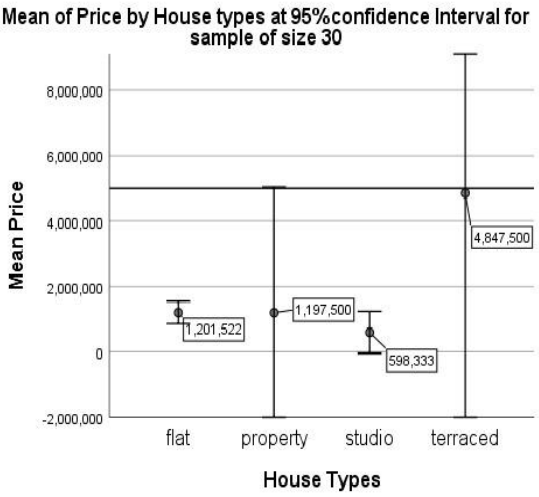
On computing one of the particular record from sample data with following values having Price 62000: Bedrooms=1, Bathrooms=1, Stations_total=2, House_type_flat=1, House_type_property=0, House_type_studio=0

We get Estimated Price by computing the above values in the derived equation as **626502.5** which is close to **62 0000**

APPENDIX 1

MEAN OF PRICE BY HOUSE TYPE AT 95% CONFIDENCE INTERVAL FOR SAMPLE OF SIZE 88 VS SAMPLE OF SIZE 30





Sample Size=88

The confidence intervals in the first plot are smaller than the ones in the second plot. With more data, the estimate of the means is more precise.

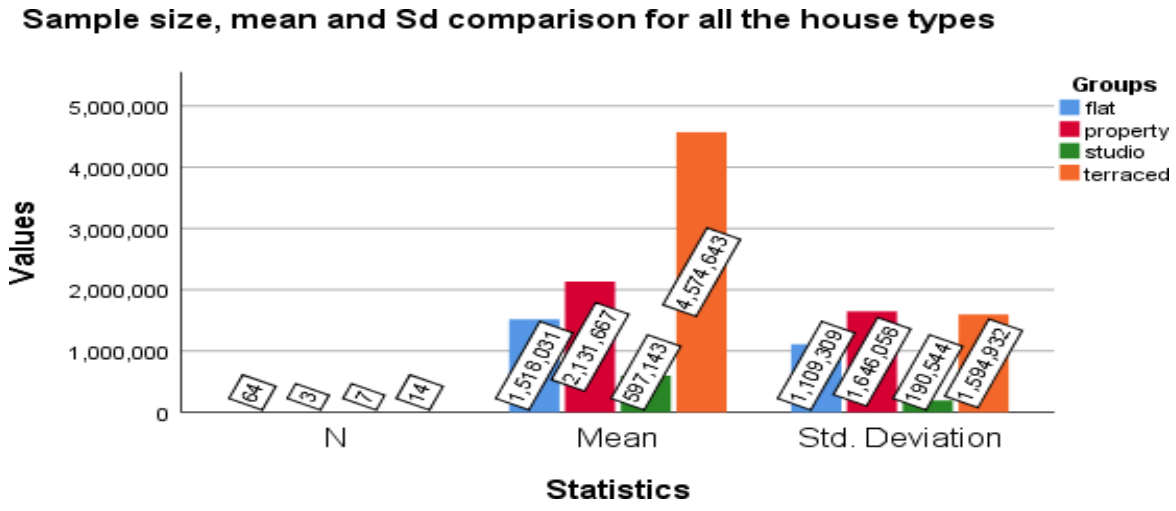
APPENDIX 2

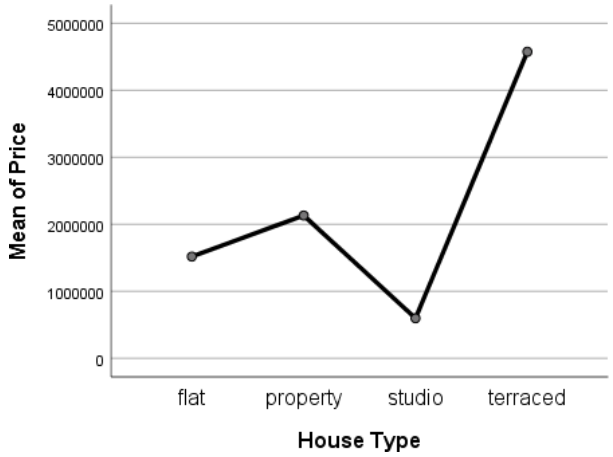
FURTHER ANALYSIS ON NON EQUALITY OF VARIANCES

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Price	Based on Mean	4.181	3	84	.008

Based on price means, the significance level is 0.008($p<0.05$) which means that the condition of homogeneity of variance is violated and there is significant difference of the variance among the prices of the different types of house types.





The graph depicts greater differences in mean price of Flat, property, studio and terraced house types

As we have violated the test for homogeneity of variance, we are sceptical regarding the significance result of ANOVA test(used to test the dependence between different sample means)

ANOVA

Price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1212949356184 10.560	3	4043164520613 6.850	29.220	.000
Within Groups	1162320333559 32.380	84	1383714682808 .719		
Total	2375269689743 42.940	87			

As we have unequal sample size and homogeneity of variance is not satisfied then we use Welch's ANOVA test

Robust Tests of Equality of Means

Price

	Statistic[a]	df1	df2	Sig.
Welch	32.691	3	8.500	.000

a. Asymptotically F distributed.

We conclude that variance is statistically significant between level of independent variables

(I) House_type_all	(J) House_type_all	Mean Difference (I-J)	Std. Error	Sig.
Flat	Property	-613635.495	960415.033	.911
	Studio	920888.315*	156250.845	.000
	Terraced	-3056611.685*	448250.230	.000
Property	Flat	613635.495	960415.033	.911
	Studio	1534523.810	953077.216	.523
	Terraced	-2442976.190	1041570.987	.273
Studio	Flat	-920888.315*	156250.845	.000
	Property	-1534523.810	953077.216	.523
	Terraced	-3977500.000*	432304.767	.000
Terraced	Flat	3056611.685*	448250.230	.000
	Property	2442976.190	1041570.987	.273
	Studio	3977500.000*	432304.767	.000

On basis of Games-Howell Post-Hoc test conducted to study the significant individual groups considering non equality of variances, on the grounds of significant value of $p<0.05$ from above table, we can conclude that there is statistically significant difference is between Flat, studio and terraced house types or if we compare in groups of two then- flat and studio, flat and terraced, studio and terraced.

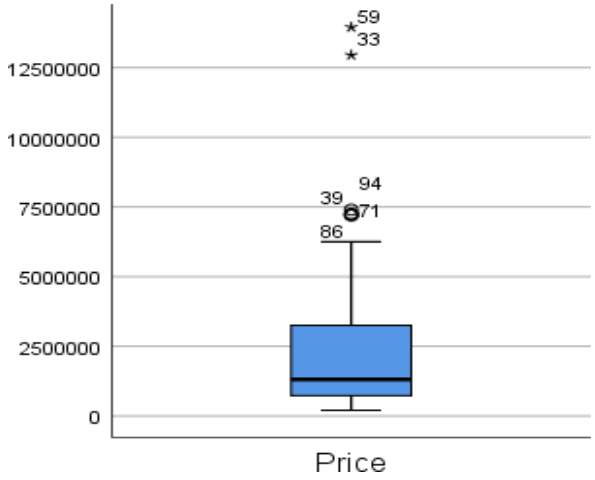
APPENDIX 3

OUTLIERS DETECTION

As per the project effective analysis, it is vital to remove the outliers as they are extreme values which do not follow the pattern in the data. So, we ought to remove from our sample data in order to carry out the regression analysis in future and derive the most parsimonious model

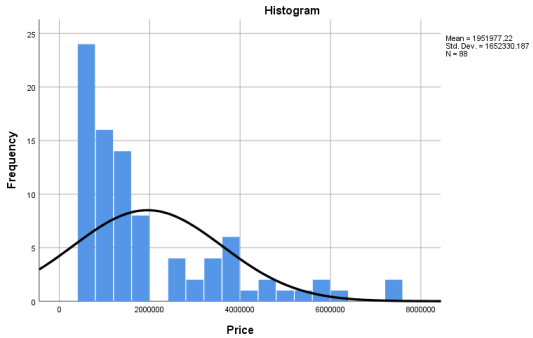
Extreme Values

			Case Number	Value
Price	Highest	1	59	13950000
		2	33	12950000
		3	85	7400000
		4	39	7250000
		5	94	7250000
	Lowest	1	72	200000
		2	41	260000
		3	25	300000
		4	1	330000
		5	61	399000



The numbers inside the graph denote the case numbers in the sample data which contains the extreme values of the price. The star denotes that those cases holds prices which are too extreme values in price data. So, after removing the mentioned case numbers from the above table, the data is moving towards the normal distribution

After removal of outliers the semi-detached house type got removed as those house types were not following the normal pattern of distribution of prices for other house types.



Price distribution after removal of outliers.