**2022**

A Descriptive Analytics Project

Aishwarya Singh

7/1/2022



OIKOS- Model for Homestead Appraisal

Table of Contents

EXECUTIVE SUMMARY…………………………………………………………2

INTRODUCTION…………………………………………………….…………….3

SAMPLING METHOD…………………………………….……………………….3

VISUALISATION…………………………………………………………………..4

DESCRIPTIVE STATISTICS………………………………………………………7

HYPOTHESIS TESTING…………………………………………………………...9

CORRELATION ANALYSIS………………………………………………………11

REGRESSION ANALYSIS…………………………………………………………12

RESIDUAL ANALYSIS…………………………………………………………….14

**Executive Summary**

This report is to conduct a research on the Real Estate market in Birmingham to propose a house calculator. The Estate agency needs a report on the pilot visualization of the housing market. The visualization is being carried out with the samples taken from the location B74 which will further help in calculating the prices of other houses in the market of Birmingham, UK. The data for the research work has been acquired from the property website [www.rightmove.co.uk](http://www.rightmove.co.uk) with a +-1 mile radius around B74, Sutton Coldfield.

Visualization on the collected data has been plotted using Microsoft excel with Tufte Principles and other guidelines as the basis. The guidelines have been applied to the graphs and mentioned alongside them.

Sampling on the dataset was carried out using sampling techniques to create a sample representative of the population

Descriptive analysis has been performed on the dataset with a confidence interval of 95% to determine the mean population. An independent sample t test and a chi-square test have been carried out. All the analysis was performed using SPSS software.

The independent sample test has been conducted to compare the average price of houses in the sample to that of mean prices of houses in Birmingham, UK.

A chi-test was performed between the categorical variables to understand the association between the variables.

The correlation analysis demonstrates which features of the sample data are most important to determine the price of the house.

A model has been formulated using linear regression analysis based on the influential dependent variables and then used to build a price calculator.

Model's adequacy has been confirmed by residual analysis and all the assumptions were satisfied.

**1. Introduction**

This model aims at providing an insight into the perpetual real estate market of Birmingham and its prospective buyers and their behaviour. As per the assigned postcode of B74- the Royal town of Sutton Coldfield, a total of 110 houses’ data were collected as the sample through mixed sampling techniques of quota and stratified. The sample was collected with the filter of +-1 mile radius from a population of 424 houses of mixed .

Eight different variables were chosen to describe the houses viz., price which is the dependent variable used in the model. Categorical variables such as types of houses and energy efficient rating were selected. Other independent variables are nearest station (miles), floor area (sq. m), number of Bed rooms, bath rooms and receptions.

All these variables were common among all the different types of houses and were deemed parameters that people look into before buying a house.

Other variables such as network bandwidth, distance of nearest restaurant, etc. were not common for all the houses.

2. **Sampling**

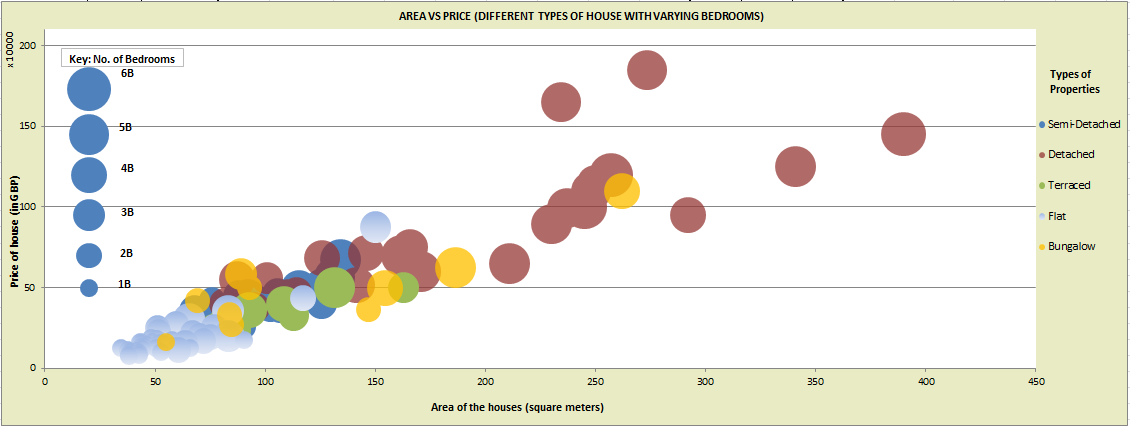
The sampling technique used here is a mix of quota and stratified sampling in order to create a sample of data that best represents the population. In order to deploy this, following steps were followed-  
1. The total number of houses (population) in B74 was recorded and the sample size was selected  
2. The ratio of the population to the sample size was taken  
3. The population was further divided into strata based on the variable- house type  
4. The ratio further helped in finding the periodic interval from where the next data had to be picked thus covering the whole target population.



Stratified sampling technique was included in order to remove the possibility of bias. According to the calculations, every fourth data was collected once the population was divided on the basis of different types of houses.

**3. Visualisations**

1. **Bubble-Plot Multiseries**



* In alignment with Tufte's principles of graphical excellence-
* 1. the graph is **multivariate** as it shows the four different variables viz.,

Area on the x-axis  
number of bedrooms (size of the bubbles)  
types of houses (bubble colour)  
dependent variable, price on the y-axis

* 2. It uses the **smallest ink** possible, and **tells the truth**.
* The chart follows the Tufte’s principles of graphical integrity as-

1. The bubbles are **proportional** to number of bedroom they are displaying.

2. The graph is **data-driven, clear** and **contextualized** with respect to our model

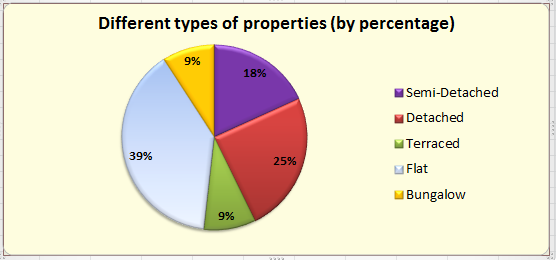
* Gestalt’s Principles have been taken into consideration as-

1. The chart conforms to the principle of **similarity** as different colours display different types of houses.  
2. **Figure and ground:** There is a strong contrast between the text and background colours. Also, the gradient, transparent and solid colour help in differentiating the houses.

* As per IBCS guidelines, the **title**, **labels** and **keys** are present for the chart.
* Therefore, it can be concluded that with increase in area, the number of bedrooms increase along with the price. This shows that there is collinearity between area and number of bedroom. Also, detached houses are big in size, with more area, hence more expensive than flats which have low floor area and hence the price.

2. **Pie Chart**

Pie chart, here is used to show the constituent of different types of houses in our sample which is representative of the population of B74 in consideration.



* The chart follows the Tufte’s principles of graphical integrity as-

1. The pies are **proportional** to percentage value they are displaying.

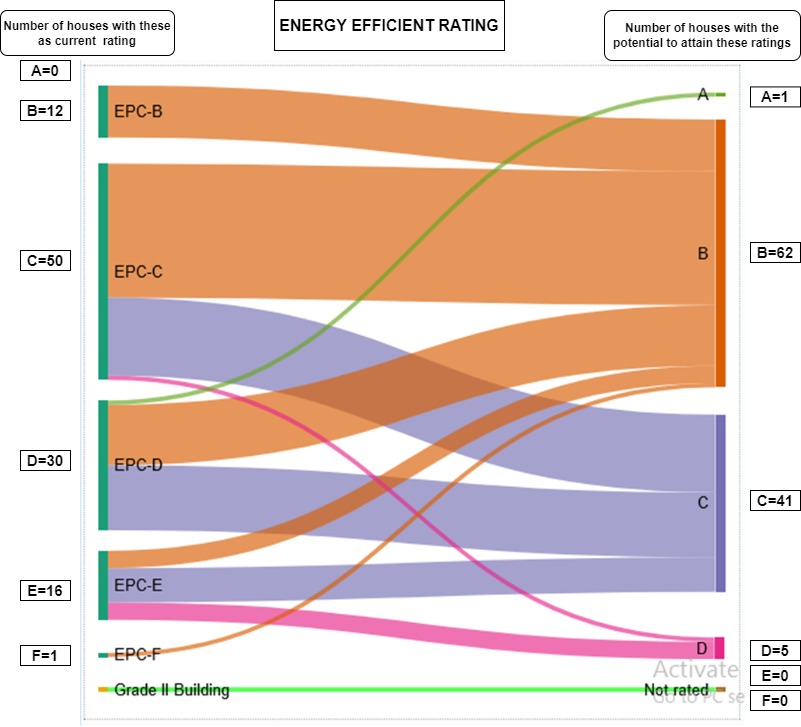
2. The percentage and colour clearly show that the data is leading (**data-driven and clarity**).

* In alignment with Tufte's principles of graphical excellence- the chart uses the **smallest ink** possible, and **tells the truth**.
* Gestalt’s Principles have been taken into consideration as-

1. The chart conforms to the principle of **similarity** as different colours display different types of houses.  
2. **Figure and ground:** There is a strong contrast between the text and background colours

* As per IBCS guidelines, the **title** is present for the chart.
* Therefore, it can be concluded that majority of the properties in the Sutton Coldfield are of type “Flat”.

3. **SANKEY CHART**   
Energy rating is evolving as an important criterion in the UK to decide the house prices, therefore, Sankey charts shows the current ratings of the houses and their potential to get upgraded to a higher level.



* The chart follows the Tufte’s principles of graphical integrity as-

1. The arrows depicting the flow are **proportional** to the number of houses they are displaying.

2. The and colour clearly show that the data is leading (**data-driven and clarity**).

* In alignment with Tufte's principles of graphical excellence- the chart uses the **smallest ink** possible, and **tells the truth**.
* It also follows MECE as the labels do not overlap.
* Gestalt’s Principles have been taken into consideration as-

1. The chart conforms to the principle of **similarity** as different colours display different types of energy ratings.  
2. **Figure and ground:** There is a strong contrast between the text and background colours

* As per IBCS guidelines, the **title** is present for the chart.
* Therefore, it can be concluded that majority of the properties in the Sutton Coldfield have the potential to attain a rating of ‘B’.

**4. Descriptive Statistics**

The data had 11 outliers which have been removed.

The measure of central tendency such as mean, median, mode and the measure of dispersion such as standard deviation, variance construct the descriptive statistics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Statistical summary table for the continuous variable | | | | | | | |
|  | Mean | Median | Mode | Variance | Standard deviation | Minimum | Maximum |
| Bedrooms | - | 2 | 2 | 1.76 | 1.32 | 1 | 6 |
| Bathrooms | - | 1 | 1 | 0.48 | 1.49 | 1 | 4 |
| Receptions | - | 1 | 1 | 0.55 | 0.74 | 0 | 3 |
| Area(sq.m) | 90.2814 | 81.67 | - | 2103.45 | 45.86 | 34.54 | 245 |
| Price(£) | 341307.5757 | 295000 | - | 45432960000 | 213150.08 | 75000 | 1000000 |
| Nearest station(miles) | 0.9191 | 0.5 | - | 0.68 | 0.82 | 0.82 | 3.1 |

Table 1: Statistical summary table for continuous variables

We see that the average price of houses is £341307.57. The median for the houses priced is £295000 which suggests that half of the houses are priced below median price and the other half are more than £295000. The minimum and the maximum prices of the houses in the data is £75000 and £1000000. Standard deviation is the dispersion of values from the average. Around 95% of the data falls within two standard deviations of the mean which suggests that 95% of the houses lie in the price range -£84992.58 to £767607.57.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Statistical summary table for categorical variable based on price | | | | | | |
| Types of Properties | Mean | Median | Variance | Standard deviation | Minimum | Maximum |
| Detached houses | 628235.2941 | 600000 | 39862320000.00 | 199655.49 | 375000 | 1000000 |
| Bungalow | 417777.7777 | 420000 | 22581940000 | 150272.89 | 165000 | 625000 |
| Terraced House | 331000 | 312500 | 11882220000 | 109005.6 | 185000 | 500000 |
| Flats | 186498.8372 | 160000 | 17085880000 | 130712.98 | 75000 | 875000 |
| Semi-detached house | 401000 | 392500 | 13388420000 | 115708.34 | 200000 | 875000 |
| Energy efficient rating |  | | | | | |
| B | 179318.1818 | 170000 | 9300114000 | 96437.09 | 95000 | 435000 |
| C | 336043.3333 | 240000 | 61666220000 | 248326.83 | 75000 | 1000000 |
| D | 371428.5714 | 377500 | 23036770000 | 151778.69 | 85000 | 650000 |
| E | 422857.1428 | 412500 | 47275820000 | 217430.04 | 140000 | 895000 |
| F | 375000 | 375000 | 0 | 0 | 375000 | 375000 |

Table 2: Statistical summary table for categorical variables

The average price of bungalow is more compare to the means of other property types. The most priced house is a detached house and least price one is a flat. The median price of flat is £160000 which suggests that half of the houses priced in flats is less than the median price and the mean prices of other property types is greater than the median price of flat. The house with E, energy efficient rating is priced the most.

A confidence interval of 95% has been used. The upper and lower limit is £299320.47, £383294.7.

|  |  |  |
| --- | --- | --- |
| Types of Properties | Lower limit | Upper limit |
| Detached houses | 588906.3986 | 667564.19 |
| Bungalow | 388176.454 | 447379.101 |
| Terraced House | 309527.6636 | 352472.336 |
| Flats | 160750.4987 | 212247.176 |
| Semi-detached house | 378207.3324 | 423792.668 |

Table 3:

Since the average price of different property types falls within the lower and upper of limits. Its suggests that sample we have is a representative of the population with a with a 95% statistical probability.

**5. Hypothesis Testing**

Hypothesis testing is a sort of statistical reasoning in which data from the sample is used to make conclusions regarding a population parameter.

We will be testing the mean price of houses of the sample is a representative of average prices of houses in Birmingham.

Null hypothesis: Average price of houses of B72 is the same as the average price of houses in Birmingham.

Alternate hypothesis: Average price of houses of B72 is not the same as the average price of houses in Birmingham.

Price of houses in Birmingham is taken from the website <https://www.rightmove.co.uk/house-prices/birmingham.html> and is £238,581. We have carried out a one sample t test with a confidence interval of 95% in SPSS.

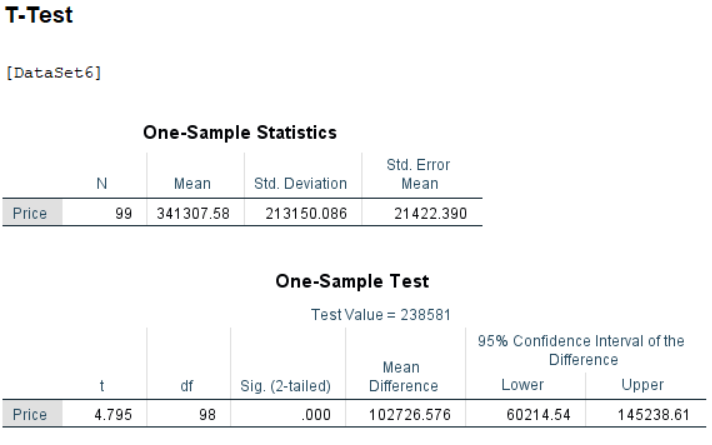


Table 4: one sample t test output

Since we see that the p value is less than 0.05, we reject the null hypothesis. Thus, average price of houses of B72 is not the same as the average price of houses in Birmingham.

We have also compared mean of two different types of houses using independent sample t test.

Null hypothesis: The average price of semi-detached houses the same as the average prices of houses of flats

Alternate hypothesis: The average price of semi-detached houses is different from average prices of houses of flats

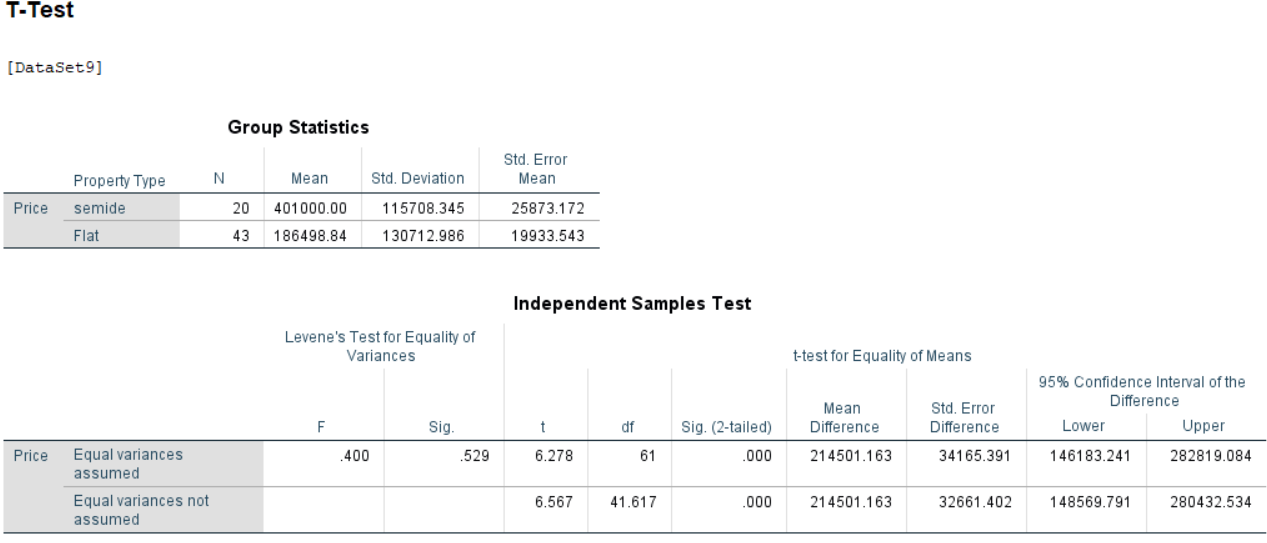


Table 5: Independent sample test output

Since the p value is less that the significance level, we reject the null hypothesis, Thus the average prices of houses of flats and semi-detached is different.

We have also carried out a chi-square test to check for the independence of variable.

Null hypothesis: Energy efficient rating is independent of property types.

Alternate hypothesis: Energy efficient rating is dependent on property types.

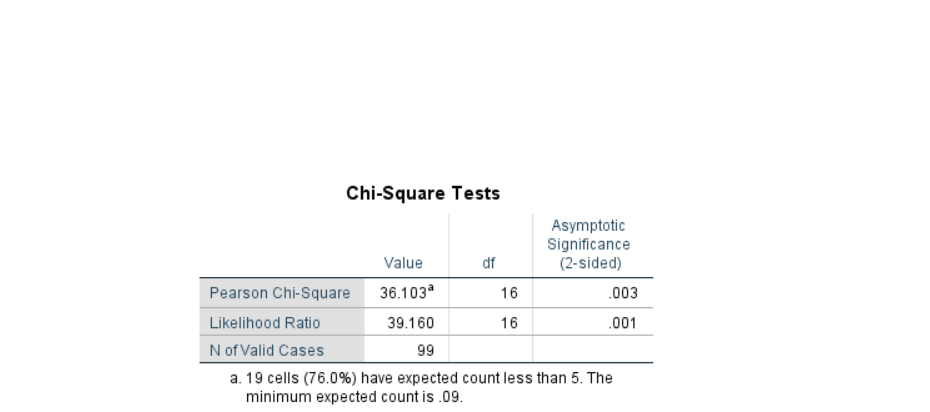


Table 6: chi-square output

As we see that the p value is less than 0.05, we reject null hypothesis. Hence, Energy efficient rating is dependent on property types.

6. **Correlation analysis**

Correlation analysis determines the degree of strength between variables. The independent and dependent variable will both increase or decrease if there is a strong relation between them.

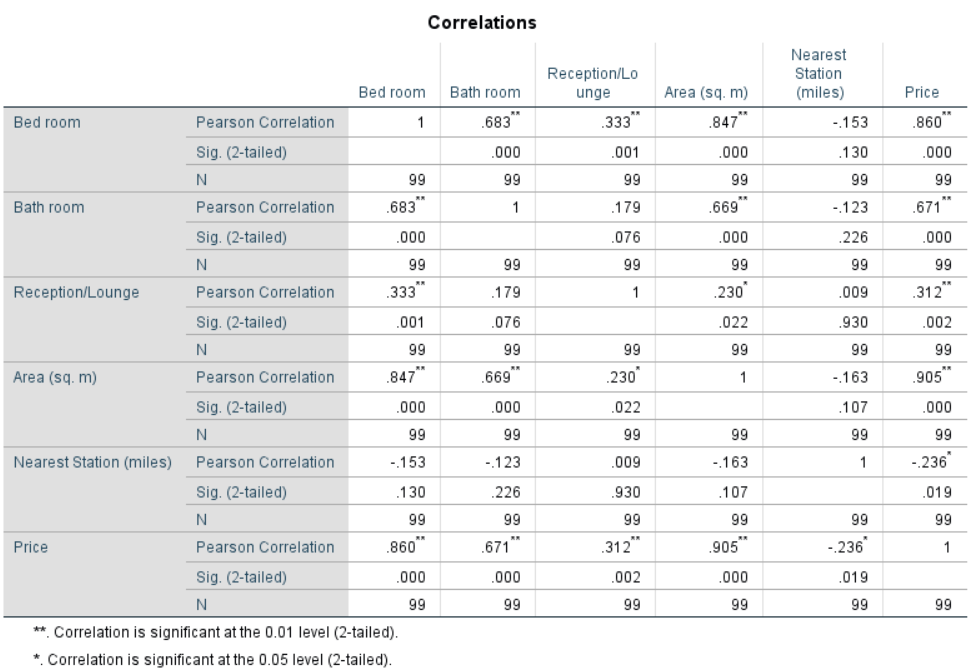


Table 7: correlation table

We see that there is a strong correlation between bedrooms and area, so there is multicollinearity in the data. To handle multicollinearity we will check for adjusted r-square with bedroom variable excluding area and then including area and excluding bedroom. The one with better adjusted r square will be included in the model.

7. **Regression Analysis**

To define the relation between independent and dependent variables and regression analysis was performed using SPSS.

Since the adjusted r-square was 0.864 when area was included and 0.806 when bedroom was included. So we will go with area as it gives a better adjusted r-square.

**FIRST REGRESSION ANALYSIS:**

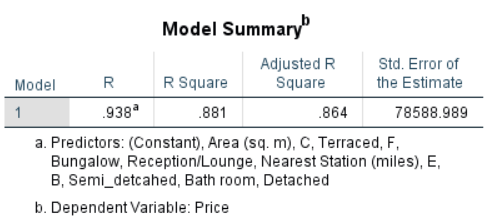


Table 8: model summary 1

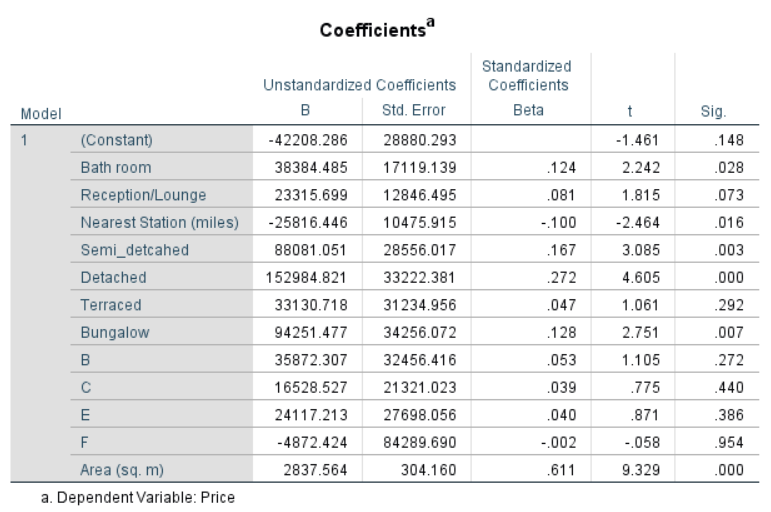


Table 9: coefficient table 1

After removing in the insignificant variables, we got a parsimonious model.

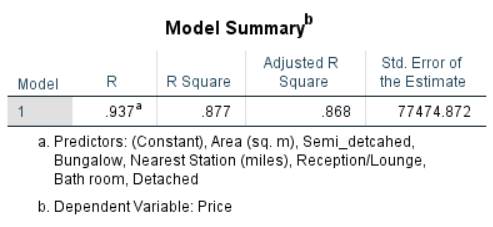


Table 10: summary table 2

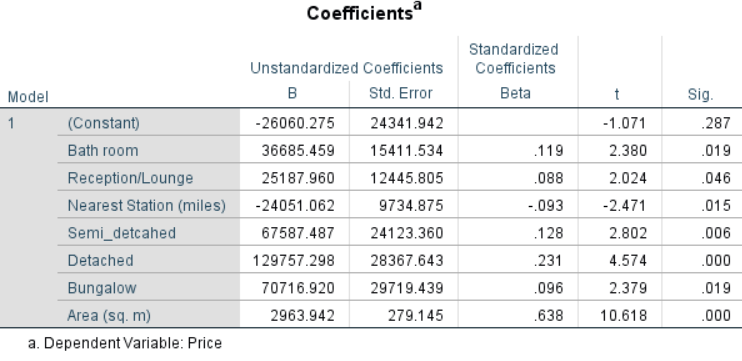


Table 11: Final Coefficient Table

The relation between independent and dependent variable can be described as

Price = (-26060.275) + (Bathroom)(36685.459) + (Reception/Lounge)(25187.96) + (Nearest Station(miles))(-24051.062) + (Semi\_detached)(67587.487) + (detached)(129757.298) + (Bungalow)(70716.920) + (Area(sq.m))(2963.942)

Predicting the price of a house using the above relation considering 2-bathrooms,1-reception/lounge, 0.5 miles nearest station, semi-detached house, area of 103 sq.m

Price = (-26060.275) + (2)(36685.459) + (1)(25187.96) + (0.5)(-24051.062) + (1)(67587.487) + (103)(2963.942)

Price is £433346.585.

**8. Adequacy Check**

To check for model’s adequacy, we will have to look into residuals. The difference between observed and the predicted value in the model is residual. We will check the adequacy of the model using the p-p plot and scatter plot which we got during the model output to validate few assumptions.

1. The average of the residuals is zero which means the residuals should be scattered around zero. This can be seen in the p-p plot which satisfies the assumption.
2. The residuals are independent i.e. residuals should be scattered which is satisfied here.
3. Standard deviation of the residuals is the same for all values. There is no be any increasing or decreasing pattern which satisfies the assumption.
4. The residuals should be normally distributed which can be seen in the p-p plot. This is also satisfied.
5. There is no multi co-linearity in the data. Thus this assumption is also satisfied.

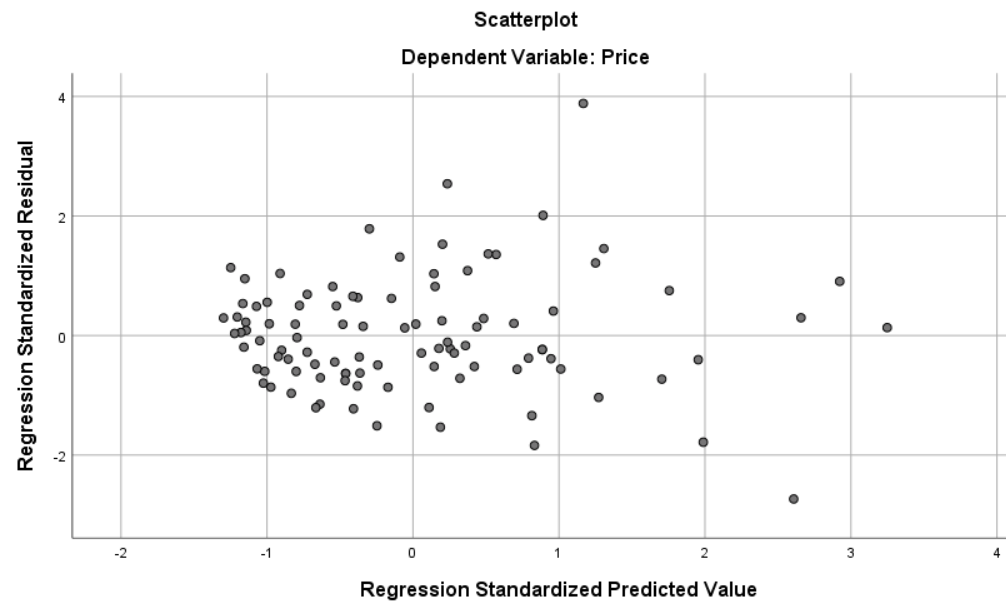


Figure 4: Residual plot

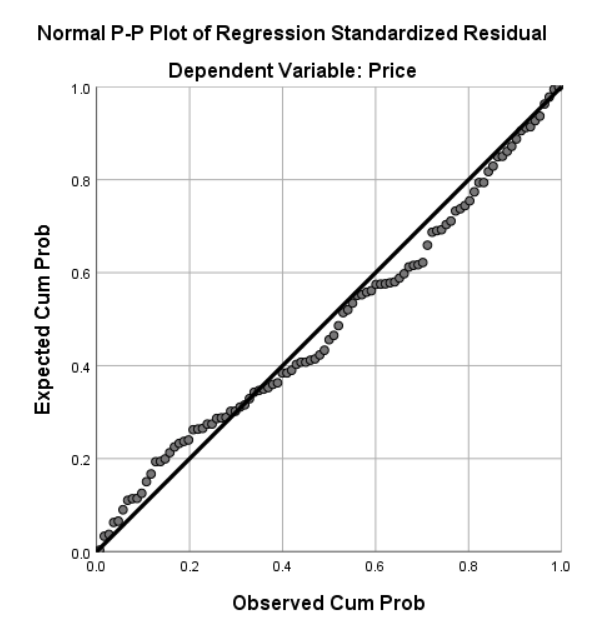


Figure 5: P-P plot

The adjusted r-square is 0.868 which suggests that the model is pretty good and since all the assumptions are satisfied, it can be used for prediction.