House market report of Birmingham with statistical analysis

Student ID: 210018653.

**Table of contents**: Page no

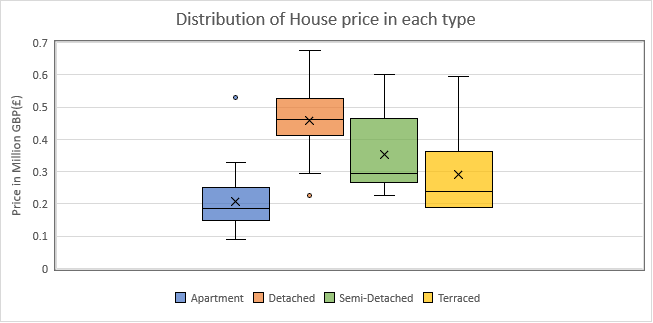
1. Introduction ………………………………………………………………………………… 1
2. Visualization of sample data ……………………………………………………..... 1
3. Overview of collected sample data ……………………………………………… 5
4. Confidence intervals of each house type …………………………………….. 6
5. Significance of average price of each house type ………………………… 6
6. Correlation between Price and other house characteristics …………. 7
7. Generation of predictive model ..…………………………………………………. 8
8. Appropriateness of the model ..……………………………………………………. 9
9. Conclusion ...…………………………………………………………………………………. 10

## **1.Introduction:**

We know that there should be a wide range of research work to done before buying a house in any area. Especially in the crowed area like Birmingham, B19 which lies in the city side should be analysed clearly to know the insights of the house type and its price ranges according to its characteristics that affects the value of the property. This report helps in understanding the house market of Birmingham, B19 and its area surrounding it which has detailed view of properties in sale that is derived from 100 sample data in each type of houses. The data for this analysis is collected from ‘Rightmove plc’ which is a UK-based company that runs ‘rightmove.co.uk’. In this analysis, I have collected the information of houses on sale in Birmingham with the postcode B19 and surrounding of 0.25 miles. As we see the data there are many apartment houses then the other type. This might be the case as this area is crowed enough. There were around 400 houses on sale in this area, from this I have gathered a sample of 100 houses that consist of almost all type of houses in it. The sampling method which I have used here is simple random sampling in which the samples are chosen randomly that have equal probability of types of house being selected. This gives us clear picture on average prices of each type of houses in this area. The prediction of house price is most reliable on the size and range of sample selected. This represents the whole population as the samples are selected randomly with equal probability.

# **2**. **Visualization of sample data:**

The appropriate way of visualization of this sample data is by graphs. There are principles and rules to be followed while building a graph like Tufte’s principle of graphical integrity, graphical excellence, and Gestalt’s Principle. These principles make sure that the data being conveyed is clear to the reader of the graphs.

**2.1) Distribution of House price:** 

* This box and whisker plot graph used here explains overall view and the distribution of price in each type of house. This is appropriate type of graph to show the details of maximum, minimum, mean, median, 1st and 3rd Quartile, outliers of the sample data.
* The box whisker plot is the appropriate way to show the details of maximum, minimum, mean, median, 1st and 3rd Quartile. The outliers of the sample data are shown
* Ambiguity of data is eliminated by providing clear and thorough labelling. This also shows that there is variation in the data and the use of design is avoided which may distract the audience.
* The Prices are standardized in Great Britain Pounds as the targeted audience are in UK. As this graph represents minimum, maximum, and average of the house price there no use of proportionality. These satisfy the Tufte’s principle of graphical integrity.
* The Gestalt’s principle of proximity and similarity in is followed here by using different colour to distinguish the different type of house with gap among them.
* Here the value of x-axis is excluded as there is no scope for it. The data in the graph is mutually exclusive collectively exhaustive that has no overlap. This graph is chosen to show a greater number of statistical data in shortest time with least use of ink in smallest place. Thus, it satisfies the principle of Tufte’s graphical excellence.

**2.2) Characteristics of houses:**

3

5

4

* The above bubble graph explains about the characteristics of houses in Birmingham(B19) and surrounding areas. The main reason to choose this type of bubble chart is to show three variables namely house size in x-axis, house price in y-axis, and the bubble size represents the number of bedrooms according to IBCS standards. This type of graph follows the Tufte’s graphical excellence principle as these are multi variant.
* Used different colour to distinguish between the house type which satisfies the Gestalt’s principle of similarity. The size of bubble increases proportionally with increase in number of bedrooms which satisfies the Proportionality principle of Tufte.
* The standardized in Great Britain Pounds as the targeted audience are in UK and it is in millions of units to reduce the use of ink which implies Tufte’s graphical excellence principle.
* The above graph depicts MECE. It is a graph with greatest number of insights of sample data consuming least time with minimal ink in the less space and tells only the truth.
* The gridlines of this graph are omitted as this may decrease the clarity of the graph which affects audience view.

**2.3) Average house price of B19 in line with average house price of surrounding areas:**

* Actual average housing price of the area B19 is compared with the surrounding areas. The bar chart above shows the true value of each area. The data was collected from <https://www.home.co.uk/>.
* The focus of audience here should be on the targeted postcode which is B19.So, the bar representing B19 is differentiated from others by using a different colour. The average line is also the focus thus it is differentiated by a different colour. This follows graphical excellence and Gestalt’s principle.
* The bar chart is the appropriate type to visualize the comparison of few categories.
* Proximity is satisfied by placing the columns close to one another, Similarity is followed by similar colour and shape, continuity is implied by arranging the average price of areas in increasing order. Thus, the Gestalt’s principle is followed.
* The usage of ink is reduced by making the price value in millions and the graph uses less space with more information that follows the graphical excellence of Tufte.

# **3. Overview of collected sample data:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive data** | **Price** | **Bedrooms** | **Bathroom** | **Nearest station** | **Size** |
| Count | 100 | 100 | 100 | 100 | 100 |
| Mean | 268332.7 | 2.68 | 1.86 | 0.36 | 1121.5 |
| Standard Deviation | 134647.4 | 1.427578 | 0.943023 | 0.215087485 | 762.8855 |
| Minimum | 90000 | 1 | 1 | 0.1 | 418 |
| First Quartile | 165000 | 2 | 1 | 0.2 | 617.5 |
| Second Quartile | 234150 | 2 | 2 | 0.3 | 947 |
| Third Quartile | 310275 | 3 | 2 | 0.5 | 1281 |
| Maximum | 675000 | 6 | 5 | 1.1 | 4400 |

The summarized features of the collected sample data are represented as a table below. This descriptive statistics on the dataset is obtained from python programming.

Table 3.1) Descriptive statistics of the sample data

The above table explains the descriptive statistics of the sample data. Here the count represents the number of data available in it. Mean is the average value of a category, that is the total sum of values divided by total number of values. Here the mean value of bedroom is 2.68 which is inappropriate in real-time that doesn’t make any sense. So, it can be rounded off to its nearest integer. From this it can be rounded off and said that the average number of bedrooms in the houses of this area is 3.

The Standard deviation (SD) measures the dispersion of a dataset relative to its mean. From the value of SD, it can be said that the higher standard deviation shows that the values are widely distributed, while a low SD shows that the values are closer to the mean value. According to this statement it can be said that the high SD value of price and size shows that the value of price and size of this sample is widely distributed, while the lower SD values of other variables indicates that the values of respective variables are closer to its respective mean value.

Minimum and Maximum values are specified here for each variable by which the range of values of each variable can be interpreted. In addition to it, when the dataset is arranged in an increasing order, the first quartile also known as lower quartile is the value under data points at 25%. Likewise, the second quartile refers to the value under data points at 50% and the third quartile or upper quartile refers to the value under data points at 75%. The observation of these values gives the overview of sample data that how it is distributed and at what range the values varies from.

# **4. Confidence intervals of each house type:**

A confidence interval is a range of values, bounded above and below the statistic's mean, that likely would contain an unknown population parameter. The probability that the population parameter is within the interval, is usually expressed as a percentage, which is called the confidence level. The confidence interval is obtained from the sample mean, standard deviation, and number of samples. This interval value of house price for each type of house shown here are calculated using the confidence interval formula with the 95% confidence level. The confidence level here is 95% which states that the unknown parameter of the data will come under the interval with the probability of 95%. On the other hand, there is a 5% of chance that the data can be outside the confidence interval.

The following table explains the confidence interval of price of each type of house. This shows that the actual population mean of Birmingham B19 and around which was obtained from this source <https://www.home.co.uk/guides/house_prices.htm?location=birmingham> is within the derived confidence interval with 95% confident level.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **House Type** | **Sample Mean** | **SD** | **Number of samples** | **Confidence interval** | | **Actual Population Mean** |
| **From** | **to** |
| Apartment | 207241.5 | 76203.34 | 68 | 189129.12 | 225353.9 | 209689 |
| Detached | 459467.65 | 104814.18 | 17 | 405574.76 | 513360.5 | 508406 |
| Semi-Detached | 354438.89 | 143343.19 | 9 | 244064.63 | 464813.1 | 283282 |
| Terraced | 289991.67 | 155730.54 | 6 | 126599.48 | 453383.9 | 244825 |

Table 4.1) Confidence interval of each type of houses

From the above table, it can be concluded that the actual population mean of each type of house is within the confidence interval that was derived from the sample. It can be noticed that the gap of interval for house type with high sample size is less and vice-versa. So, the confidence interval will be more precise if the sample size is more.

# **5. Significance of average price of each house type:**

As we see the average price of each house type is different from its actual mean price. But the significance in difference between them can be explained by a statistical test called one-sample t-test. This test is chosen to predict the significance of average price of each house because we must compare sample of mean with its corresponding actual mean. As our test relies with one attribute, one sample t-test is the appropriate way to test the significance.

As per this one sample t-test, a null hypothesis and an alternate hypothesis should be set. According to the result any of these hypotheses will be accepted. The output of the test is the p-value. This result is interpreted by the p-value with the significance value(α) which 0.05. If the p-value is greater than the significance level(α), then the null hypothesis is accepted and the alternate hypothesis is rejected. On the flip side, if the p-value is less than the significance level(α) then the null hypothesis is rejected and the alternate hypothesis is accepted.

Formulating the null and alternate hypothesis:

* Null Hypothesis (H0) = Average price of sample is not significantly different from average price of actual population.
* Alternate Hypothesis(H1) = Average price of sample is significantly different from average price of actual population.

The test was conducting in python programming and the p-value of the test are populated in the below table.

|  |  |  |  |
| --- | --- | --- | --- |
| **House Type** | **P-value** | **Sample Mean** | **Actual Population Mean** |
|  |
| Apartment | 0.7919 | 207241.5 | 209689 |  |
| Detached | 0.0721 | 459467.7 | 508406 |  |
| Semi-Detached | 0.1747 | 354438.9 | 283282 |  |
| Terraced | 0.5091 | 289991.7 | 244825 |  |

Table 5.1) P-value of each type of houses

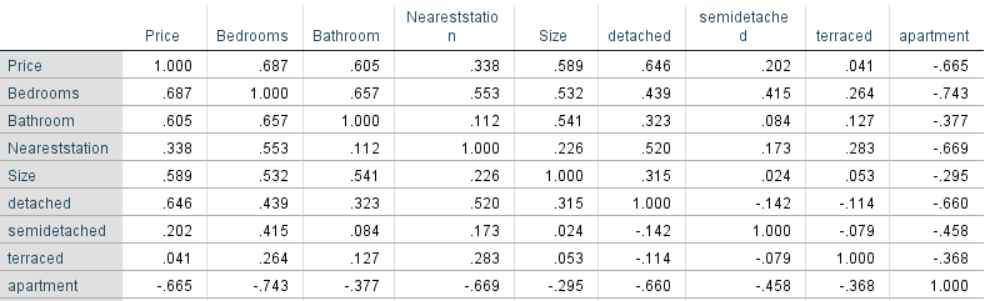
From the above table, it is shown that the p-value of all house type is greater than significance level(α=0.05) which tends to accept the null hypothesis.

Thus, average price of sample is not significantly different from average price of actual population. From this its inferred that the sample data chosen replicates the actual population.

# **6**.**Correlation between Price and other house characteristics:**

The price of a house is affected by many factors but the characteristics of a house that affects the most can be determined by correlation matrix. Here the price of house is dependant variable that is depend upon other variables like number of bedrooms, number of bathrooms, distance from house to nearest transport station, size of the house, etc. The characteristics that is most correlated with the dependent variable i.e., price is explained in the correlation matrix.

The correlation matrix is obtained from the SPSS, from the collected sample data.

Table 6.1) Correlation matrix of characteristics of house with respect to price

From the table, it can be said that bedrooms, bathroom, and size are highly correlated with the price of the house.

# **7.Generation of predictive model:**

A good predictive model is obtained by doing regression analysis of sample data. Conducting regression analysis, the independent variables are reduced as much as possible for those t-value is from -1.96 to 1.96. If the two highly correlated independent variables that if entered together in the model one or both appear to have no significant effect on the dependent variable but if entered individually, they have a significant effect on the dependent variable. And the main consideration while executing the regression analysis is observing the R-Squared and Adjusted R-Squared value of the model by which the quality of the model is determined. Here both the bedrooms and bathrooms are highly correlated. So, by entering bedrooms and eliminating bathroom and the variables which are insignificant to the prices the highest R-Squared and Adjusted R-Squared value is obtained. Thus, the parsimonious model.

* R Squared: 0.717
* Adjusted R Squared: 0.701

The regression should be done until a higher value of R-Squared and Adjusted R-Squared value is obtained. The quality of model will be improving with the increase of these value. These values will increase from 0 to 1. In general, if these values are greater than 0.7, then the model is said to good.

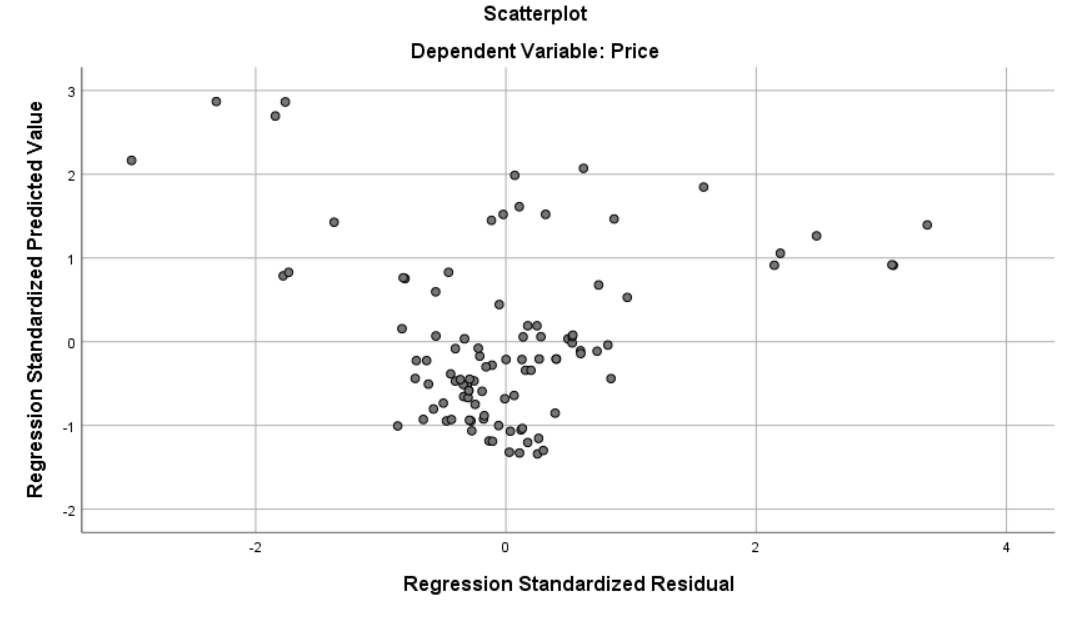
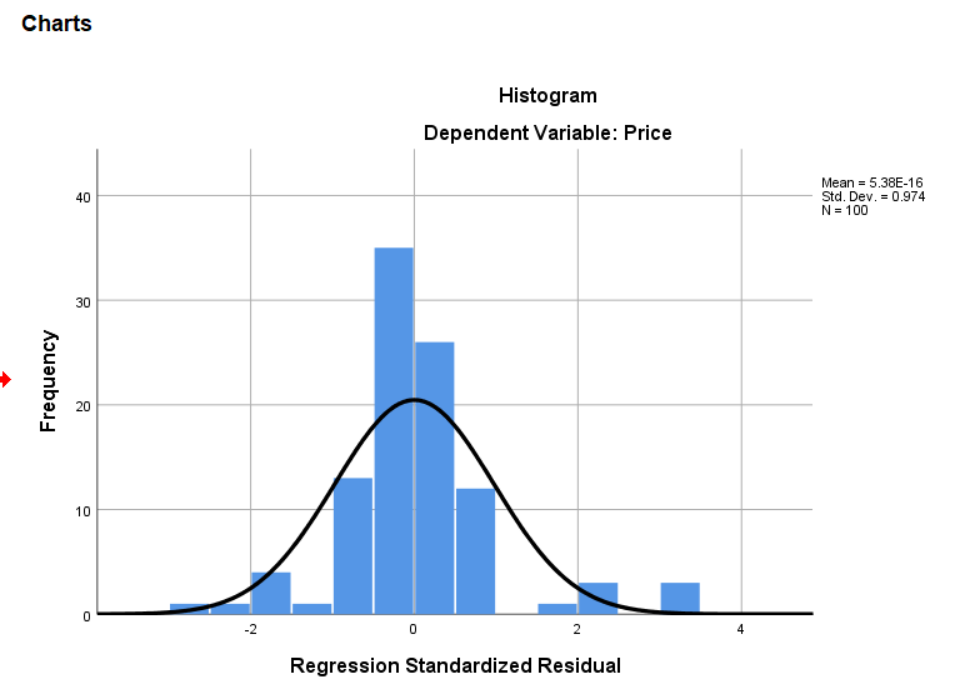
|  |  |  |  |
| --- | --- | --- | --- |
|  | **coefficient** | **t** | **P>|t|** |
| Intercept | 135643.742525 | 7.093095 | 0.000 |
| Detached | 194764.886590 | 7.598698 | 0.000 |
| Semi-Detached | 75856.957615 | 2.391931 | 0.019 |
| Bedrooms | 34641.606589 | 4.105827 | 0.000 |
| Nearest station | -149307.189981 | -3.319467 | 0.001 |
| Size | 47.849223 | 4.005100 | 0.000 |

The coefficient shows how much the dependent variable is expected to increase and decrease when the independent variable increases by one.

The P>|t| value indicates the significance of the variable to the price of the house. If this value is nearer to 0, that variable is more significant.

# **8. Appropriateness of the model:**

* The goodness of fit is determined by Adjusted R-squared value which is obtained in the regression model should be greater than 0.7. In this model, the Adjusted R-Squared value is 0.701. So, it can be interpreted as goodness of fit.



* Residual analysis is given by five assumptions. Each of the assumption should hold then the model is adequate. Assumption **one** is linearity which is determined from the scatterplot of residuals. Here the assumption **holds** as the residuals are scattered equally above and below the zero.
* Assumption **two** is Normality which is determined from histogram. Here the assumption **holds** as the residuals are normally distributed.
* Assumption **three** is homoskedasticity which is determined from scatterplot. Here the assumption **holds** as the residual plot does not clear pattern.
* Assumption **four** is independence of error is determined from scatterplot. Here the assumption **holds** as the residual plots are randomly scattered.
* Assumption **Five** is Multicollinearity. It can be determined from the correlation matrix (refer table 6.1). Here this assumption **holds** as the multicollinearity between bedrooms and bathroom is omitted in regression model as they both make them insignificant in the model.

As all five assumption holds, this model can be said as adequate.

# **9. Conclusion:**

Derived statistical model or predictive model i.e., most parsimonious model is given by:

Price of house = 135643.742525 + 194764.886590(Detached) + 75856.957615(Semi-Detached) + 34641.606589(Bedrooms) - 149307.189981 (Nearest station) + 47.849223(Size) + e

**Example:**

The actual price of a detached house with 4-bedroom, 2 bathroom and size of 1475 sq. Ft which is 0.3 miles away from the railway is 499950 GBP.

Check whether the model can predict the actual price of the house

Price of house = 135643.742525 + 194764.886590(1) + 75856.957615(0) + 34641.606589(4) - 149307.189981 (0.3) + 47.849223(1475) + e

= 494760.50 + e

e is the error which is the difference between actual value and estimated value.

e = 499950 - 494760.50 = 5189.50

Thus, this model can predict the price of a house in the area B19 and surrounding. From the model it can be said that the house price of apartment and terraced type is constant whereas the price of detached and semi-detached is different from other by its respective value. The house price is positively affected by number of bedrooms in the house, size of the house and negatively affected by distance to the nearest station from the house which shows that if the house is nearer to the station then the price will be higher and vice-versa. Likewise, if the size of house and number of bedrooms increases the price of house also increases.