Aston Business School
Birmingham

Descriptive Analytics:

Housing Market Analysis in Birmingham B74

Done by Walid Raad – 210191462

DUE DATE: 17 DECEMBER 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This housing market analysis was conduct for a real estate firm that is looking to invest in area B74. The data was collected from Rightmove, and the attributes collected from the dataset that affects the price are Type of the house, number of bedrooms, number of bathrooms, and distance to nearest station. Number of bedrooms and bathrooms were good attribute in affecting the price of the house. As the number of bedrooms or bathrooms increase the price increased as well. Flat houses are the most common type of houses in B74 with an average price of £198000. It has seen as well that detached houses are the most expensive type of houses amongst all. We have used confidence interval to estimate our mean population and ran a hypothesis testing to see if the average housing price in our sample is in line with the average price in the Birmingham and they all were in line with our hypothesis.

INTRODUCTION

Area B74 postcode is located in Birmingham UK, and it is the largest city in the westmidlands, were B74 is considered one of the wealthiest areas with an average house price found at £497,000. This report aims to study the housing market in area B74 and see what factors are contributing to its price. The data was collected from Rightmove which is a website that displays homes for sale. I have used a simple random sample method to collect my data which is an effective sampling technique that makes sure you select houses randomly for the study that gives a you representative of the population. Some of the limitations encountered were the limited number of variables available when collecting the sampling data. For instance, the area for houses couldn't be found which played a huge factor in predicting a better and more accurate model. Moreover, the house prices listed on the property website were only houses that are for sale, rather than the entire housing market in B74, which can have some inaccuracy when comparing our finding to the whole area.

For the remainder of this report, the terms *Bedrooms* and *Bathrooms* will refer to the number of bedrooms and numbers of bathrooms respectively. Price is measured in 1000s of GBP. Distance from Station is measured in miles.

VISUALIZATION

The bar chart below illustrates the average house price for each type of house.

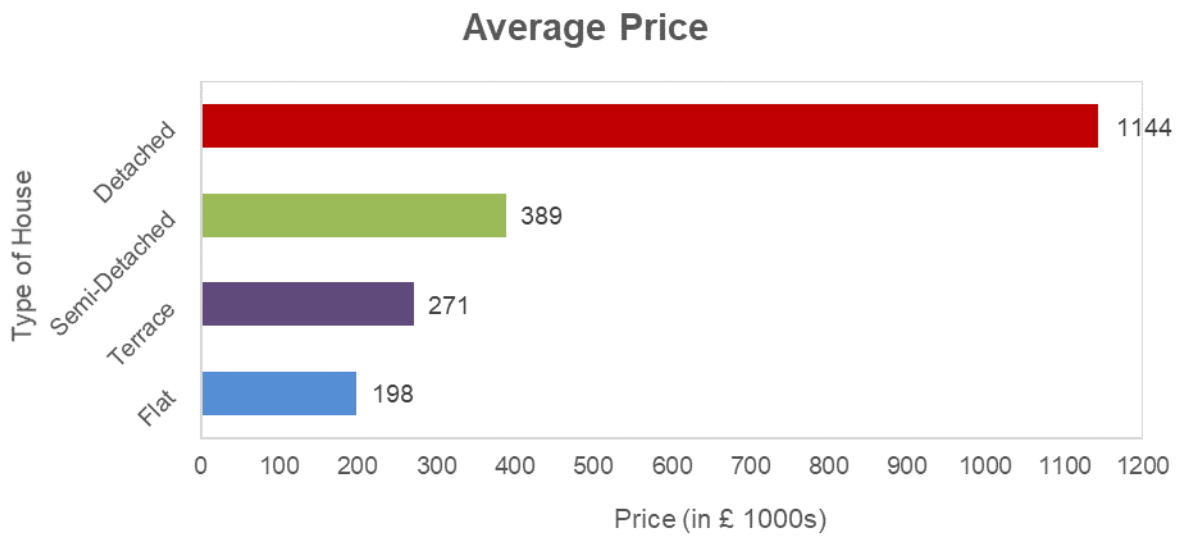

Figure 1: *Bar chart displaying average price per house type*

A quick glance at the graph shows that detached houses are the most expensive compared to others with an average price of £1,144,000. It also shows that Flat houses are amongst the cheapest with an average price of £198,000. In the graph the use of distinct colors for different house types and a white background conforms to gestalt principles of focal point which immediately draws attention to the house type, it also follows Tufte's principle of graphical excellence as its data-driven and multivariate.

The pie chart below shows the proportion distribution of different house type.

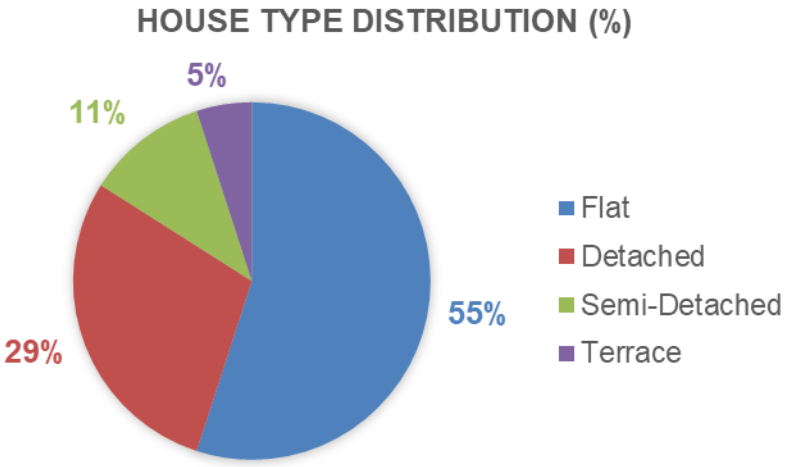

Figure 2: *House type by distribution (%)*

Flats are the most common type of house among all four, followed by detached, semi-detached and lastly terrace houses with only 5% categorized under terrace in B74. The consistency in color for different house types with the previous bar chart conforms to the principle of similarity.

The box-whisker plot highlights the price distribution of the house market by No. of bedrooms.

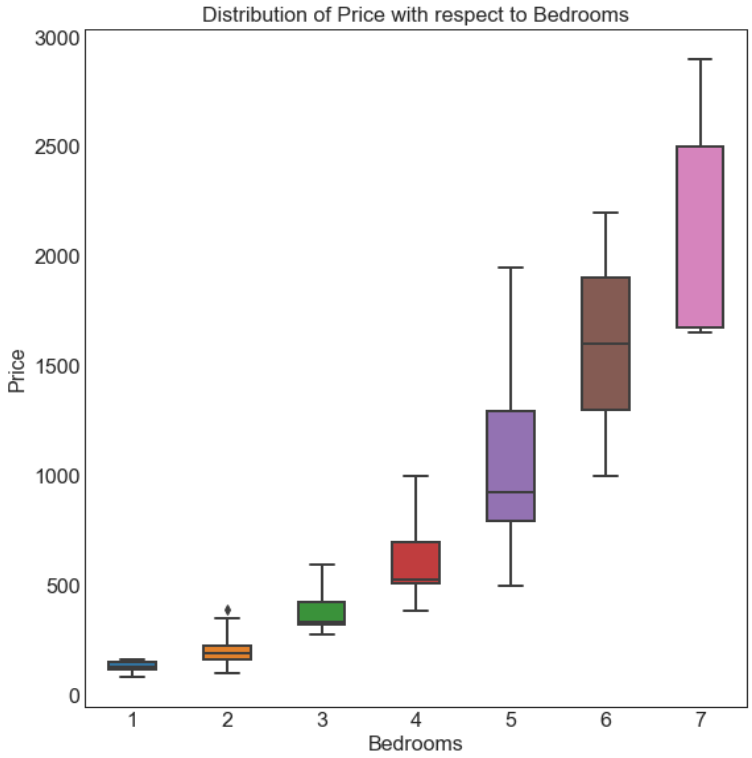

Figure 3: *Box whisker showing distribution of price per number of bedrooms*

Box whisker plot is used to show you the distribution of your numerical data and its symmetry by displaying them into quartiles (points divided into 3 quarters). The middle line in the box is the median (mid-point of your data), upper end and lower end of the box are your upper (75^{th}) and lower (25^{th}) quartile. The upper end of the stick is your maximum price, and the lower end of the stick is your lowest price. We can see from the median for houses with 1, 3, 4, and 7 bedrooms are headed toward the 25^{th} quartile which means it is positively skewed, and for houses with 2 and 6 bedrooms, the median is in the middle which means the points are spread evenly. Lastly, from the look of the plot we can see that the maximum and minimum price in the box plot increases as the number of bedrooms increases, this shows that there is a positive relationship between them. In the graph the use of distinct colors for different house types and a white background conforms to gestalt principles of focal point which immediately draws attention to the different number of bedrooms.

DESCRIPTIVE STATISTICS

The table below summarizes the key results obtained throughout the study.

Table 1: *Summary of Housing Market Key Outputs*

All Type	Bedrooms	Bathrooms	Price (£ 1000s)	Distance from station (miles)
Count	100	100	100	100
Mean	-	-	497.03	1.132
Standard Error	0.16	0.12	55	0.09
Median	2	1	290	0.6
Mode	2	1	-	-
Standard Deviation	1.63	1.18	550.58	0.89
Variance	2.67	1.386	303140.47	0.792
Range	6	5	2810	2.7
Min	1	1	85	0.1
25th Q	2	1	160	0.4
75th Q	4	2	525	2.125
Max	7	6	2895	2.8

	Detached		Semi-Detached		Flat		Terrace	
Key Outputs	*Bedrooms*	*Price(£1000s)*	*Bedrooms*	*Price(£1000s)*	*Bedrooms*	*Price(1000s£)*	*Bedrooms*	*Price(1000s£)*
Count	29	29	11	11	55	55	5	5
Mean	5	1143.97	3.27	388.64	1.78	198.15	2.4	271
Standard Error	0.22	120.1529	0.12	27.41	0.086	12.78	0.40	42.35
Median	5	900	3	360	2	165	3	275
Mode	5	-	3	-	2	-	3	-
Standard Deviation	1.16	653.90	0.47	90.91	0.63	94.75	0.89	94.70
Variance	1.36	427586.39	0.22	8265.45	0.4	8977.3	0.80	8967.5
Range	4	2470	1	245	2	510	2	245
Min	3	425	3	280	1	85	1	120
25th Q	4	675	3	325	1	138.75	2	260
75th Q	5	1650	3.5	460	2	222.5	3	335
Max	7	2895	4	525	3	595	3	365

The summary table shows that the average house price in area B74 is £ 497,000, and the median house price is £ 290,000. The average price is larger than the median, this shows that the data is right skewed, which means that most of the houses in area B74 are below the average price. Most houses contain 2 bedrooms and 1 bathroom with an average distance of 1.13 miles to the nearest station. There is a large dispersion in how the prices are spread in the area, this is because it depends on the property type the person is looking for. In area B74 the most popular house is a flat where prices range from £ 85,000 to £ 595,000 this huge price range accommodates people from different classes to come and invest in a house. It can be seen as well that detached houses are the most lavish property type having an average price of £ 1,144,000 ranging from 3-7 bedrooms while semi-detached houses range from 4-3 bedrooms which indicated that detached houses come in bigger sizes than semi-detached. Finally, terrace houses are the least popular in the area having an average price of £ 271,000 ranging from 1-3 bedrooms same as flat, even though the size of both property types could be similar, terrace houses are more expensive on average.

INFERENTIAL STATISTICS

Confidence Interval

To check whether our average house price per house type we collected from our sample is a representation of the whole house market in B74 we will have to do inferential statistics. We can estimate that by using the confidence limit theorem while being 95% confident that my average house price per house type will lie within an interval. Z-Statistics was used for Flat since the number of observations for this house type is >30 and t-statistics was used for the other three house type since the number of observations for them are < 30.

Table 2: *Key outputs taken for specified confidence interval*

PROPERTY TYPE	*MEAN*	*COUNT*	*STANDARD DEVIATION*	*HIGH C.I OF 95%*	*LOW C.I OF 95%*	*MARGIN OF ERROR*
Detached	1143.96	29	653.9	1382	906	238
Flat	198.14	55	94.7	223	173	25
Semi-Detached	388.63	11	90.9	442	335	53.72
Terrace	271	5	94.7	354	188	83

Interpretation: We can conclude from the above table with 95% certainty that the average house price for detached houses in area B74 (for the population) lies between £ 906,000 and £ 1,382,000. The same thing applies to the other 3 house types. The margin of error is higher for terrace compared to semi-detached, this is because the number of observations for the terrace is lower than semi-detached, the more sample data there is the less the margin of error will be.

Two-tail test

Based on Rightmove's data it is estimated that the average house price in B74 is £494,122. I have conducted a two-tail test to prove that with 95% certainty that the average house prices in B74 are in line with the average house price I have collected in my sample data. One sample t-test is a type of hypothesis testing that is used to compare whether the mean of two groups is significant or not. The table below shows the statistical outputs obtained for a 95 % CI.

Table 3: *Output of Test Statistics*

Significance Level taken as 5 %		
H_0: The average house price in B74 is in line with sample data – $\mu = 494{,}122$		
H_a: The average house price in B74 is not in line with sample data – $\mu \neq 494{,}122$		
Statistical Outputs	**Z**	**Critical Value $z_{\alpha/2}$**
	0.29	1.96

Interpretation: Test-stat (Z) < Critical-value. Therefore, we accept the null hypothesis (H0) at 5% L.O.S and conclude that there is significant evidence that the average house prices we have claimed from our sample data are in line with the average house prices in area B74.

CORRELATION MATRIX

The correlation matrix is a table that displays how strongly correlated our variables are with each other. The larger the coefficient between the two variables the stronger the relationship is between them.

Table 4: *Correlation Matrix*

Correlation Matrix				
	Bedrooms	**Bathrooms**	**Price**	**Distance from station**
Bedrooms	**1**	-	-	-
Bathrooms	0.84	**1**	-	-
Price (£ 1000s)	0.87	0.89	**1**	-
Distance from station	-0.18	-0.075	-0.17	**1**

We can conclude from the correlation matrix that there is a strong positive relationship between price and bedrooms and price and bathrooms. This means that the more the number of bedrooms or bathrooms there is in the house the higher the price of the house. However, the coefficient between price and distance to the closest station is low, which means there is no/weak relationship between them. This could be to the fact that most people in that area own a car, so it does not matter if there is a close station nearby or not.

REGRESSION ANALYSIS

In this section, we want to investigate whether independent variables (number of bedrooms, bathrooms, distance from the station, and the property type) affect predicting the dependent variable (price of the house). Python was used to estimate multiple regression model we will be able to explain the values of the Price by using more than one independent variable.

Initial Model

The figure below displays the results obtained for the initial model.

Dep. Variable:		Price		**R-squared:**		0.854
Model:		OLS		**Adj. R-squared:**		0.845
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-314.8357	69.195	-4.550	0.000	-452.244	-177.427
Bedrooms	141.9775	36.144	3.928	0.000	70.202	213.753
Bathrooms	261.0358	37.550	6.952	0.000	186.469	335.602
Distance_from_station	-34.9363	26.275	-1.330	0.187	-87.113	17.240
Type_Detached	-36.1056	101.451	-0.356	0.723	-237.567	165.356
Type_Semi	-70.1343	88.052	-0.797	0.428	-244.989	104.720
Type_Terrace	44.8430	107.171	0.418	0.677	-167.977	257.663

Figure 4: *Regression analysis results for Initial Model*

Results show that 84.5 % (Adj. R-squared) of the variance in predicting the price can be accounted for by using the 6 predictors collectively. Looking at the results of independent variables one by one we can conclude that Bedrooms (p-value= 0) and Bathrooms (p-value=0) both have a p-value< 0.05 showing they both are significant in predicting the price. However, the type of houses and distance from the station all have been insignificant in predicting the price of the house since their p-value is>0.05. Therefore, we begin by eliminating the highest p-value (Type detached) and re-run the model to see if the model has improved. We keep iterating until all our independent variables are significant (p-value<0.05), finding the most parsimonious model.

Modified Model

The figure below displays the results obtained for the modified model.

Dep. Variable:		Price		**R-squared:**		0.849
Model:		OLS		**Adj. R-squared:**		0.846
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-351.6400	44.313	-7.935	0.000	-439.588	-263.692
Bedrooms	136.1662	24.395	5.582	0.000	87.749	184.583
Bathrooms	260.0152	33.847	7.682	0.000	192.839	327.191

Figure 5: *Regression analysis results for Modified Model*

Results show that both bedrooms and bathrooms are significant, showing p<0.05 and adj. R-squared has improved slightly showing that 84.6% of the variance can be accounted for by using the 2 predictors collectively.

However, based on the results from the correlation matrix created earlier we can see that there is multicollinearity between the two independent variables. This can reduce the precision of the estimated coefficient, which reduces the predictive power of our model. Therefore, we will have to eliminate one of them and take the one with the highest Adj. - R squared value.

Final Model

The figure below displays the results obtained for the final model.

Dep. Variable:		Price		**R-squared:**		0.801
Model:		OLS		**Adj. R-squared:**		0.799
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-231.0312	44.241	-5.222	0.000	-318.827	-143.236
Bathrooms	418.4260	21.091	19.839	0.000	376.571	460.281

Figure 6: *Regression analysis results for Final Model*

Results show that we have found the most parsimonious model (p-value<0.05) and have a good Adj.R squared (Goodness of fit statistics) with 79.9 % of the variance in predicting the price can be accounted for by using the bathroom variable alone.

DIAGNOSTIC ANALYSIS

For our final model to be ready for use we will need to test its adequacy and to do so we will have to run a diagnostic analysis. To see if a model is adequate, we will need to plot a scatter plot of the residuals and a histogram to investigate whether they are normally distributed, and they must adhere to the following 5 assumptions.

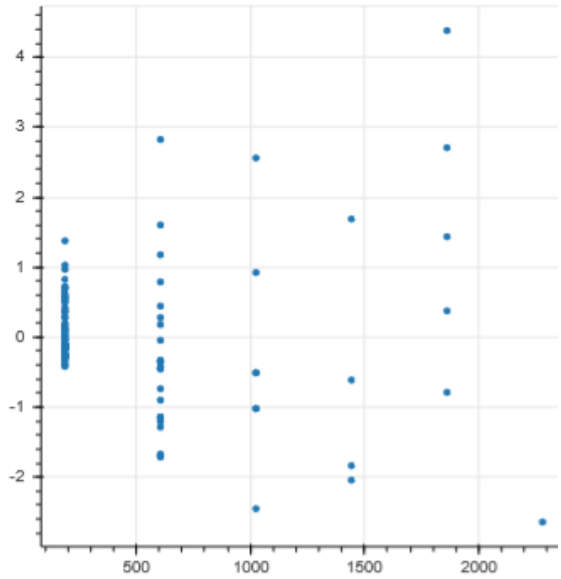

Figure 8: *Residual Plot*

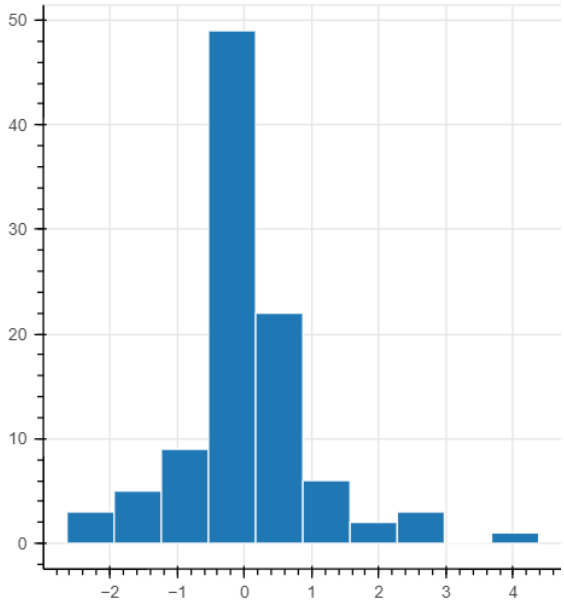

Figure 7: *Normality Plot*

The *assumptions* of linear regression are:

Linearity: We see that the residuals are randomly scattered around zero E(ε) = 0. There is no apparent pattern to the residuals, which suggests that a linear model is appropriate for this relationship.

Homoscedasticity: The variance of any residual point is the same, so it is not violated.

Independence of error: The residual plots are independent of each other so its not violated.

Normality: The residuals ε on the histogram is ***not normally distributed***, therefore the normality test is violated.

Multicollinearity: It has been dealt with in the linear regression and is not violated anymore.

Conclusion

My final model is inadequate because it has failed the normality test of residual, and this is because it contains only one number of significant independent variables. More variables should have been collected for more accuracy. However, the model will be tested to confirm that the model is not viable.

TESTING THE MODEL

The final predicative model used is shown below.

$$Price\ (£\ 1000s) = \ Intercept + Coefficient * Bathroom$$

$$=> \ \ Price = \ -231.0312 + 418.4260 * Bathroom$$

Table 5: *Prediction Summary*

Bathroom	Predicted Price (£ 1000s)	Actual Sample Price (£ 1000s)
3	£ 1024.24	£ 775
2	£ 605.8	£ 290

The table above indicates that there is a huge variation from predicted and the actual price. Therefore, the diagnostic analysis that has been done confirms that my model is unreliable to use.