Effect of some characteristics on sale price of houses in Bristol during the years 2019 and 2020

Tran Thi Hanh, Nguyen Thi Ngoc Linh, Do Thuy Trang and Tran Kha Uyen

15/6/2022

Abstract

The study investigates the interrelations between the sale price of the house and other factors such as the size of the house, number of rooms, and property type,... using property sales in Bristol during the years 2019 and 2020. It is found that the size of the house and the current energy rating of the property have positive relationships while potential energy efficiency rating of the property and number of rooms have negative relationships with the house prices.

1 INTRODUCTION

House prices have become a major concern for many and are a frequent topic of political debate. The Bristol property market is on a serious upward price trajectory and that the boom is showing no signs of slowing down for a while. House prices, in particular, have soared over the last couple of decades with property prices in Britain equating to around four times as much as 20 years prior on average. Bristol, in particular, has seen the fifth largest house price increase UK-wide over the last 20 years according to new research by Plumbnation.

While the lack of housing supply is often cited as a major reason for rising prices, this study will look into a factor that is often ignored in these debates but which may be having a significant impact on house prices: some characteristics of the house. In this study, we will investigate whether or not the characteristics such as size, number of rooms,... of houses have influences on the sale prices.

2 DATA AND PRELIMINARY ANALYSIS

Data is available for 9151 property sales in Bristol during the years 2019 and 2020. From the given dataset, we make descriptive statistics for crucial factors such as sale price, total size, number of rooms, current energy rating of the property, potential energy efficiency rating of the property

Table 1: Descriptive Statistics

Variables	**Min**	**Mean**	**Max**	**Standard Deviation**
Price	60000	333527	3000000	193388.3
Size	18.00	93.08	558.00	39.23431
NR	1.000	4.508	14.000	1.506097
CEE	1.00	61.86	95.00	11.34474
PEE	17.00	79.94	106.00	7.531547

By looking at the correlation coefficient of the independent variables CEE and PEE with the target variable $Price$ (-0.18 and -0.12 respectively), we can see that they are weak correlations, we suppose these two variables are not significant in the regression model. The correlation coefficient between the independent variables $Size$ and NR are very strong. This will be taken into consideration as it might cause multicollinearity in the model.

Table 2: Correlation Matrix

	Price	**Size**	**NR**	**PEE**	**PEE**
Price	1.0000000	0.83494569	0.656485321	-0.1838663	-0.121235567
Size	0.8349457	1.00000000	0.813420303	-0.2209990	-0.087521975
NR	0.6564853	0.81342030	1.00000000	-0.2366453	-0.004890259
CEE	-0.1838663	-0.22099900	-0.236645271	1.00000000	0.384880514
PEE	-0.1212356	-0.08752197	-0.004890259	0.3848805	1.00000000

Moreover, we can see the difference in the mean of the house price between detached and flat types. According to statistics in Table 3, the mean of the house price for detached type is about two points five times as much as that for flat type. It shows the trend that detached houses tended to be preferable with higher prices in Bristol during the years 2019 and 2020 and vice versa.

Table 3: Mean of sale price of house for each property

Property	**Mean of sale price**
Detached	612231.4
Flat	245174.9

Figure 1 presents the two variables employed in the analysis: the sale price of house and the total floor area. In Figure 1, the size of the house displays an increasing trend and thus might be expected to have a positive influence on house prices. The larger the house is, the more expensive it is. Moreover, the houses with sale prices which are below 1 million pound was the most popular.

The relationship is approximately linear but curves up somewhat for the higher-priced homes. Because the relationship is approximately linear and we expect total size to be an important explanatory variable

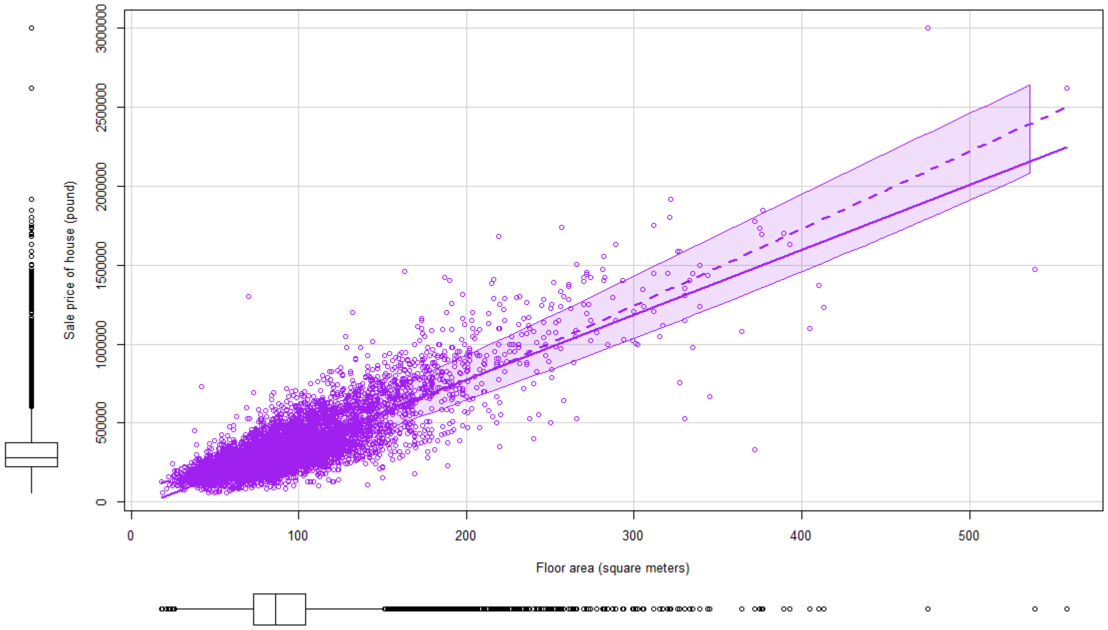


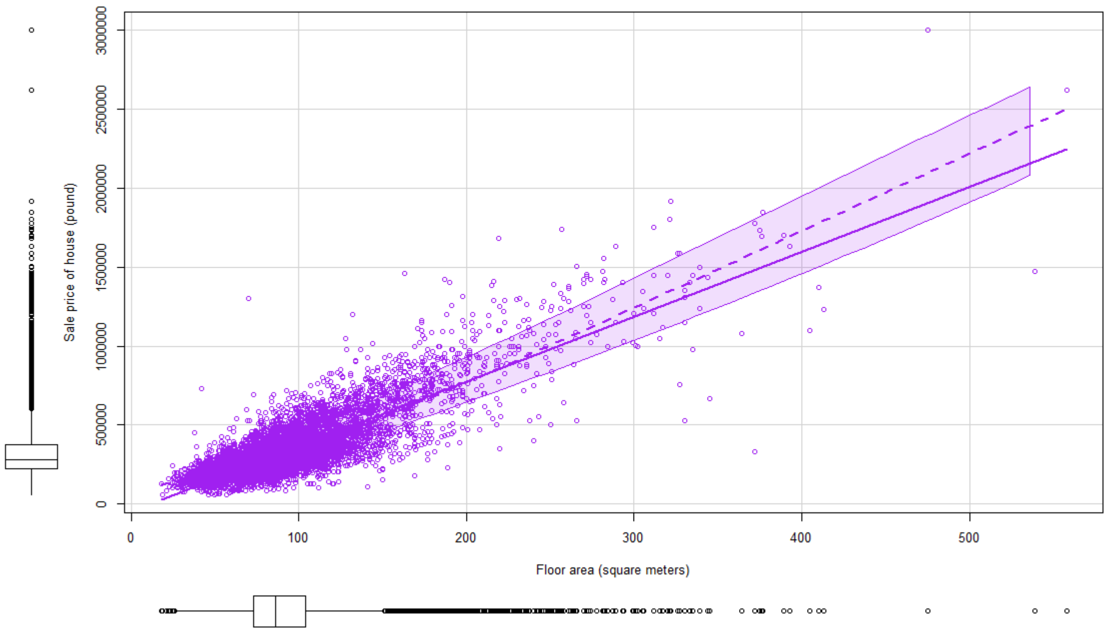
Figure 1. The chart shows relationship between the house price and the floor area in Bristol during the years 2019 and 2020

Figure 2 gives plot of house prices and number of rooms. The number of rooms displays an increasing trend and thus might be expected to have a positive influence on house prices. We can see that houses which have from 1 to 5 rooms had the relatively stable price (about 250000 pound). However, the sale prices of houses which have more than 5 rooms increased substantially from about 400000 to over 150000 pound.

The one-room home has the lowest price and the fourteen-room home has the highest price of all the homes in the data set. This observation may require special attention later.

Figure 2 suggests that the relationship between NR and $Price$ may be slightly curved. One simple kind of curved relationship is a quadratic function. The model with the NR variable might include the quadratic form of this variable

Figure 1. The chart shows relationship between the house price and the floor area in Bristol during the years 2019 and 2020



For the age variable, we can break the datas into 11 periods from 1 to 11 in order to analyze as following:

Then we can show the relationship between the house price and age as below:

Figure 3 presents the mean prices of house in each period of age. The mean price seemed to be decreased from the first period to the last period in the dataset. Houses in period 1 (before 1900) got the highest price while the mean price of houses in the eighth period was the lowest price during years 2019 and 2020 in Bristol. From the third period, the mean

Table 4: Mean of sale price of house for each property

Age (Original data)	**Period**
before 1900	1
1900-1929	2
1930-1949	3
1950-1966	4
1967-1975	5
1976-1982	6
1983-1990	7
1991-1995	8
1996-2002	9
2003-2006	10
2007 onwards	11

price went down sharply. However, there was a fluctuation of the mean price during the fifth period to the tenth period.

Our empirical analysis below will shed more light onto this matter.

3 VARIABLES

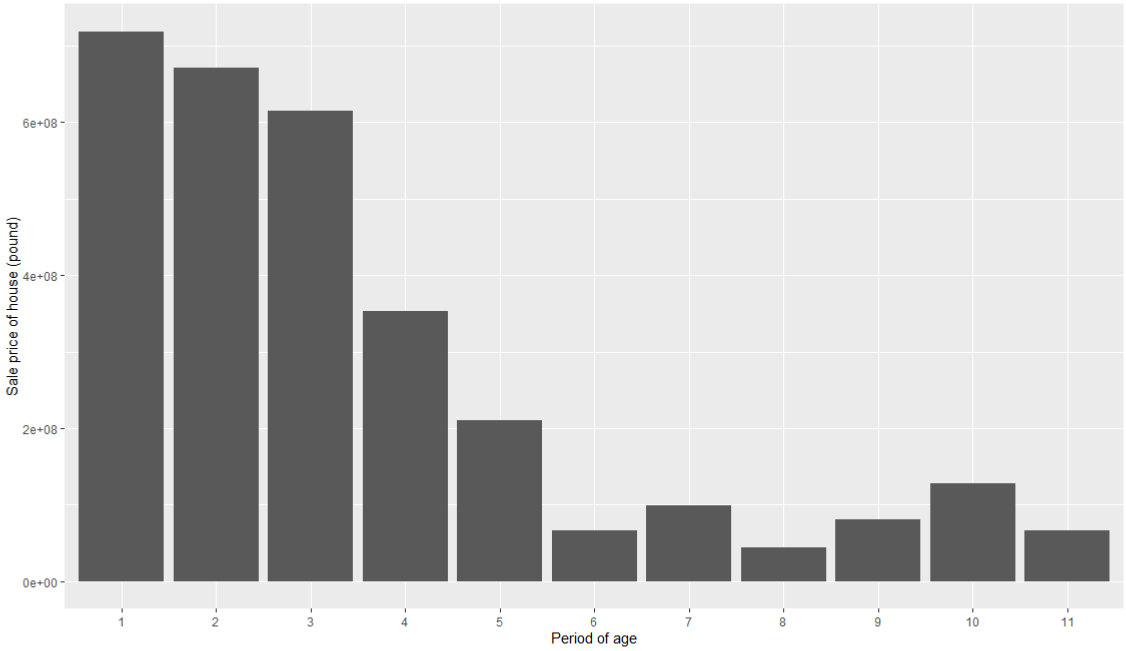
In general, we expect all explanatory variables to have a **positive relationship** with the response variable Price. Size and number of rooms represent the availability and affordability of the house, whereas other variables such as: types of property, energy efficiency have a substantial influence on the purchase of house and might make price goes up as they increase.

3.1 *REGRESSION OF PRICE ON FLOOR AREA*

The fitted model is:

$$Price = 49546.5 + 4115.5Size$$

Figure 3. The chart shows the relationship between the house price and its age in Bristol during the years 2019 and 2020



The coefficient for $Size$ is statistically significant ($t-value = 145.1$, $df = 9149$, $p-value < 2e-16$). Each additional square foot of area raises selling prices by \$4115.5 on the average. From the R^2, we see that 69.7% of the variation in the home prices is explained by a linear relationship with square feet. We hope that multiple regression will allow us to improve on this first attempt to explain selling price.
The coefficient totally alligns with our expectation as we believe that the larger the size of the house, the higher the price of it.

3.2 *REGRESSION OF PRICE ON NUMBER OF ROOMS*

The output for the regression of $Price$ on the variable number of rooms will be displayed in the Appendix. This model explains 43.1% of the variation in price. This is less than the 69.7% explained by $Size$, but it suggests that NR may be a useful explanatory variable ($t - value = 83.243$, $p - value < 2e - 16$). The fitted equation is

$$Price = 46451 + 84295NR$$

3.3 *REGRESSION OF PRICE ON AGE*

The coefficient for Age is statistically significant ($df = 9140$, $p - value < 2e - 16$). Each additional age of property may be make raises in selling prices. We see that R^2 is 0.08712 it mean that only 8.712% of the variation in the home prices is explained by a linear relationship with Age.

The coefficient dissimilar with our expectation as we believe that the latest age of the house, the higher the price of it but we can see that from the result after 1990 when the age is higher but the price of property is higher whereas before 1990 when the age get older will make the price of property increased. The people may be prefer the houses has old structure than modern houses.

3.4 *REGRESSION OF PRICE ON TYPE OF PROPERTY*

The output for the regression of $Price$ on the variable type of property will be displayed in the Appendix. This model explains 15.5% of the variation in price. This is less than the 69.7% explained by Size and 43.1% explained by type of property, but it suggests that NR may be a useful explanatory variable ($t - value = 83.243$, $p - value < 2e - 16$). The fitted equation is:

$$Price = 339302 + 272929D + -94127F$$

If we keep variable flat unchanged when the type of detached increased by 1 unit will make the price property increased by \$272929. While, when we keep type of detached unchanged, if we increase type flat by 1 unit will make the price property decreased by \$94127.

Estimate the regression of price on the variables size, number of rooms, age, detached and flat, year of sale completion. The model is:

$$Price =_1 +\beta_2 Size + \beta_3 NR + \beta_4 Year + \beta_5 Age + \beta_6 D + \beta_7 F$$

β_1 is minimum price when size, number of rooms, age, type of property equal to zero. About β_2 we expect that if the size is higher the price of property is higher and from the result we can see that the sign of coefficient is positive as expected. If we keep other variable unchanged, when the size increased by 1 unit will make the price of property increased by \$3.898e+03. About β_3, for number of rooms we expect that if the number of rooms is increased will make the price of property is higher and from the result we can see that the sign of coefficient is positive and that is suitable with think of us. If we keep other variable unchanged,when the number of rooms increased by 1 unit will make the price of property increased by \$4.482e+03. According the test we already implement, we has R^2 for Age is 0.08712. It means that just only 8.712% of the variation in the home prices is explained by a linear relationship with age. We can ignore age variable to model simple.

This model explains 72.98% of the variation in $Price$, little more than the 69.71%, 43.1%, 8.712%, 15.5% explained by simple linear regression of $Price$ on size, number of rooms, age, type of property respectively.

4 BUILDING MODEL

4.1 *CORRELATION*

The correlation between each independent variable with the target variable must not be weak. However, the correlation between two independent variables must not be too strong. Multicollinearity occurs when independent variables in a regression model are correlated. Therefore, we must keep the correlation between variable in the model at a acceptable degree.

As reference to the correlation result from the above, we can see that the correlation coefficient of the independent variables CEE and PEE with the target variable $Price$ (-0.18 and $-0.12 respectively$) are weak correlations.Besides, the t-test for the significance of both CEE and PEE ($R^2 = 3.38\%$, 1.47% respectively) once again indicates the weak relationship between these variables and the price.Therefore we can exclude these two independent variable from our model.

Additionally, the correlation coefficient between the independent variables $Size$ and NR are quite strong (0.813) and NR has a lower correlation with the dependent variable $Price$

$(0.65 < 0.83)$ than $Size$ does. We may decide to exclude the attribute NR from the model but we will do one more test to make sure this variable is not helpful.

Besides, regression of $Price$ on Age band which receives a very low R^2 (R^2 8%) may suggest that age does not contribute significantly to the house price. Therefore, we only utilize the below independent variables in our model: $Size$, NR, $Year$, D, F.

4.2 *FITTING MODEL*

The relationship between house price and other factors is demonstrated in the following form:

$$Price = \beta_0 + \beta_1 Size + \beta_2 NR + \beta_3 Year + \beta_4 D + \beta_5 F + u^{(1)}$$

where $Size$ represents the floor area in square meters; NR represents the number of rooms; $Year$ represents year that sale completed (taking values from 1, 2 and so on for periods starting from 'before 1900'); D is a dummy variable for type of house (1 if detached, 0 if not); F is a dummy variable for another type of house (1 if flat, 0 if not).

The fitted model is:

$$P = 59955.04 + 4139.4Size + 1215.6NR - 20082.9Year + 68126.7D + 48408.5F$$

By looking at the regression result, the model is significant ($F - stat = 4559$; $df = 5$ and 9145; $p - value < 2.2e - 16$) with $RSE = 103500$. All variables seem to contribute significantly except for NR ($t - value = 0.855$, $p - value = 0.392$). Hence, we can conclude that the variable NR is not significant to the fitting when having others presence and decide to exclude this variable from the model.

Then we try to fit the data again with the remaining atrributes which are Size, Year, Detached and Flat. The new equation looks like:

$$Price = \beta_0 + \beta_1 Size + \beta_2 Year + \beta_3 D + \beta_4 F + u^{(2)}$$

The regression result show that all the independent variables are highly significant. After excluding the insignificant variable NR, the F-statistic improved from 4599 to 5699 which is a good improvement. But there is no improvement on RSE and $adjustedRsquared$ value. We obtain the following regression result:

$$Price = -57173.3 + 4171.4Size - 20071.4Year + 68296.9D + 47019.26F$$

Keeping others fixed, each additional square foot of area raises selling prices by \$4171.4 on the average. The price house in 2019 is believed to be sold less than in 2020 by \$20071.4 when buying the same type of house with the same total square meter. The R^2 tells that 71.3% of the variation in the home prices is explained by a linear relationship with square feet, sale year and types of house, which is quite considerable.

We will try a different type of transformation on the dependent variables. In this model, we will use the natural logarithm to the Price and see the change in performance of the model:

$$\log(Price) = \beta_0 + \beta_1 Size + \beta_2 Year + \beta_3 D + \beta_4 F + u^{(3)}$$

All the independent variables estimated are significant with p-value ¡ 0 but the R^2 is smaller than this one in model (2). The regression result is:

$$\log(Price) = 11.84 + 0.0086Size - 0.058Year + 0.098D0.04F$$

Based on the regression, if we keep other variables fixed, we obtain an increase by 0.86% in the price if the house is one more meter larger. In the similar way, the house price is estimated to decrease by 5.8% if buying in 2019 instead of 2020. In the same sale year with the same size of house, if a customer look for a detached house then he will pay 9.8% more, while if it's a flat house, he will pay 4

To make sure our model can best represent the relationship between variables and draw precise inferences about the true coefficients, we need to assure the model is under the assumption of the OLS method. However, when we do the diagnostic tests for normality and serial correlation, both results rejects the null hypothesis, which means there exists heteroscedasticity and the error terms are not normally distributed. The result of all the tests can be found in the Appendix

5 CONCLUSION

From the previous part, we have two models that are considered to be the best model. For model (2), we notice that it has the highest F-statistic value for the overall significance test and all variables are statistically significant, it stands out to be the model we desire to use. The model technically represents the relationship of house price with other factors that are: total square meter, sale year and type of house (the data here contains 3 types of house: detached, flat or none of these). We will state it right here for a quick remind:

$$Price = -57173.3 + 4171.4Size - 20071.4Year + 68296.9D + 47019.26F$$

However, the existence of non-normal error terms and heteroscedasticity in all models indicates that the predicted values are likely to be far away from the true values. That is, even though the model has a high goodness of fit with the dataset and all the variables are significant, it is not helpful in predicting. One reason we found for this violation of assumption is to blame for the data itself. The price of house is not a stable variable and yet it differs from time to time, among types and among realtors. The transformation of response variable to the log form in model (3) could not help avoid heteroscedasticity or normality violation. We have tried out transforming both independent and dependent variables but it did not improve much. We suppose the homoscedasticity might be guaranteed if we use the Generalized Least Square (since the residuals seems to have linear relationship with the predictive variable) or other kind of model.

To conclude, the best model that represents the effect of other factors on house price we have regressed is in the following form:

$$Price = -57173.3 + 4171.4Size - 20071.4Year + 68296.9D + 47019.26F$$

Note that the above model does not satisfy the assumptions of Least Square method, it can represent the relationship of house price with other significant elements that influence it though.

6 APPENDIX

Table 5: Regression of Price on Number of rooms

	Estimate	**Std. Error**	**t-value**	**p-value**
(Intercept)	-46451	4813	-9.652	2e-16 ***
NR	84295	1013	83.243	2e-16 ***

Table 6: Regression of Price on Age

	Estimate	**Std. Error**
(Intercept)	361889 ***	4291
1930-1949	-15022 *	6139
1950-1966	-98363 ***	6630
1967-1975	-67551 ***	8142
1976-1982	-64059 ***	13103
1983-1990	-65823 ***	11016
1991-1995	-50764 **	16097
1996-2002	-63729 ***	12022
2003-2006	-114997 ***	9194
2007 onwards	-121069 ***	11946
before 1990	58651 ***	6197

Table 7: Regression Price on Type of Property

	Estimate	**Std. Error**	**t-value**	**p-value**
(Intercept)	339302	2171	156.29	2e-16 ***
D	272929	8367	3259	2e-16 ***
F	-94127	4563	-20.63	2e-16 ***

Table 8: Regression Of Price with Size, NR, Year, D and F

	Intercept	Size	Year	D	F	NR	F-stat
Estimate	-59955.04	4139.43	-20082.9	68126.7	48408.5	1215.61	
Std. Error	4827.6	48.86	2182.19	5119.13	3288.83	1421.25	
T-value	-12.419***	84.727***	-9.203***	13.308***	14.719***	0.855	
							4559***

Table 9: Regression Of Price with Size, Year, D and F

	Intercept	Size	Year	D	F	F-stat
Estimate	-57173.31	4171.45	-20071.40	68296.90	47019.26	
Std. Error	3567.61	31.39	2182.11	5115.18	2859.68	
t value	-16.026***	132.879***	-9.198***	13.352***	16.442***	5699***

Table 10: Regression Of Logarithm of Price with Size, Year, D and F

	Intercept	Size	Year	D	F	F-stat
Estimate	11.84	0.00861	-0.058	0.0979	-0.044	
Std. Error	0.0098	0.0000	0.006	0.014	0.0078	
t value	1204.343***	99.588***	-9.658***	6.950***	-5.615***	3549***

Table 11: Diagnostic test for Model (2)

Test	**Name**	**Result**	**Accepatable point/ Conclusion**
Test for normality	Jarque-Bera Test	$p-value < 2.2e-16$	Random errors are not normaly distributed
Multicolinearity	VIF	$VIF < 2$	No serious multicolinearity
Heteroscedascity	Breusch–Pagan test	$p-value < 2.2e-16$	Heteroscedasticity is present in the regression model
Serial Correlation	Durbin-Watson Test	$p-value = 0$	The residuals in this regression model are autocorrelated
Wrong functional form	RESET Test	$p-value < \ 2.2e-16$	The model is mis-specified

Table 12: Diagnostic test for Model (3)

Test	**Name**	**Result**	**Accepatable point/ Conclusion**
Test for normality	Jarque-Bera Test	p-value 2.2e-16	Random errors are not normaly distributed
Multicolinearity	VIF	VIF 2	No serious multicolinearity
Heteroscedascity	Breusch–Pagan test	p-value 2.2e-16	Heteroscedasticity is present in the regression model
Serial Correlation	Durbin-Watson Test	p-value = 0	The residuals in this regression model are autocorrelated
Wrong functional form	RESET Test	p-value < 2.2e-16	The model is mis-specified