CATÓLICA
CATÓLICA PORTO
BUSINESS SCHOOL

PORTO

Final Report

Regression and Data Analysis

João Nápoles – Nº351218121

Simão Magalhães – Nº355422067

Vasco Moutinho – Nº355422066

Universidade Católica Portuguesa – Porto

Católica Porto Business School

Porto, 22.05.2023

Introduction

Within the scope of the Regression and Data Analysis curricular unit, we were asked to select a dataset with a response variable subject to behaviour analysis. The analysis would be done using regression models, or other types of data analysis tools taught in the classroom.
As such, we decided to use the database available on BlackBoard "BatonRouge_HomeSales_data", a database that analyses the sales price of the house taking into account descriptive attributes among other factors such as the type of occupation, location and time on the market.
The aim of the report will provide an understanding of the data presented, referring to a descriptive analysis, followed by the construction, selection and approval of the chosen model.
Finally, a brief conclusion will be made where we will address our inferences and annotations of the results presented throughout the work.

1. Descriptive Analysis

This database, available on the BlackBoard, contains 1080 instances and 11 attributes, those being:

Price- Dependant Variable	sale price, dollars
sqft	total square feet
bedrooms	number of bedrooms
baths	number of full baths
age	age in years
occupancy	Owner = 1; Vacant = 2; Tenant = 3
pool	Yes = 1; No = 0
style	Traditional = 1; Townhouse = 2; Ranch = 3; New Orleans = 4; Mobile Home = 5; Garden = 6; French = 7; Cottage = 8; Contemporary = 9; Colonial = 10; Acadian = 11
fireplace	Yes =1 No = 0
waterfront	Yes =1 No = 0
dom	Days on the market

Table 1- Description of the attributes

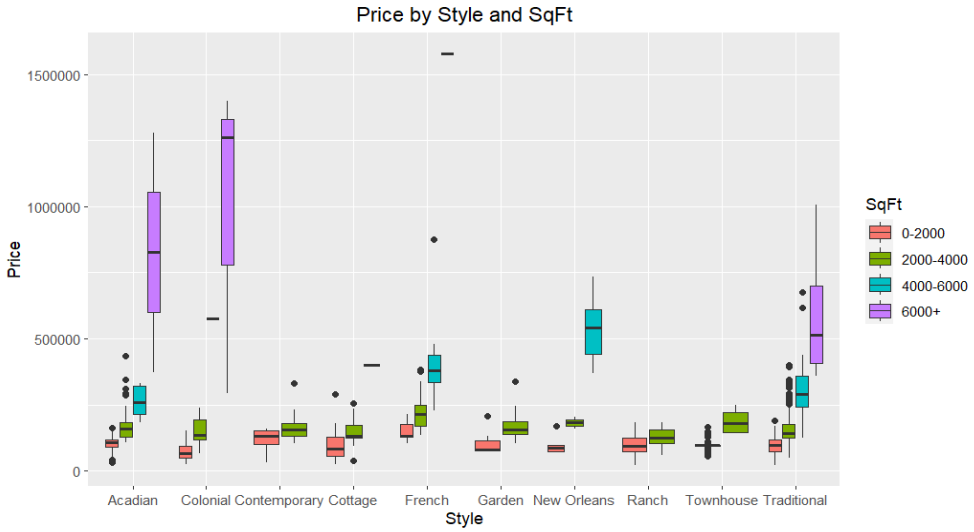
In the process of formatting the database so that it would present the desired format, we decided to change all observations that had numeric values associated with names. It was the case of the style and occupancy, that we replaced the numbers by their names, and the remaining variables that presented one as “No” and two as “Yes”.

We then proceeded to a statistical analysis of each of the numerical variables. The analysis comprised the standard deviation, mean, median, minimum and maximum, 1st quartile, 3rd quartile, interquartile range and coefficient of variation.

Variable:	Sd	Mean	Median	Min	Max	1st Quartile	3rd Quartile	IQR	CV
Price	122912,8	154863,2	130000	22000	1580000	99000	170162,5	71162,5	0,7936865
sqft	1008,098	2325,938	2186,50	662	7897	1604,5	28000	1195,5	0,4334157
bedrooms	0,7094959	3,17963	3	1	8	3	4	1	0,2231379
baths	0,6120669	1,973148	2	1	5	2	2	0	0,3101981
age	17,19425	19,57407	18	1	80	5	25	20	0,8784197
dom	94,89677	74,05648	40	0	728	14	100,25	86,25	1,2814107

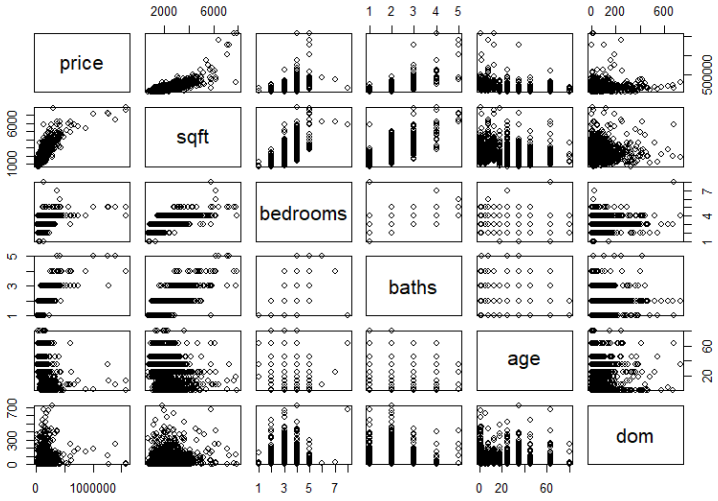
Table 2 - Descriptive Statistics of the Observations

There is a considerable variability in the price and square footage of homes sold, while the number of bedrooms and bathrooms is more consistent. The age of homes sold also varies considerably, with a mean of almost 20 years.
In the context of the provided data, we can see that the CV for price is 0.79, indicating that the variability of housing prices is relatively high compared to the mean. The IQR for price is 71,162.5, which means that the middle 50% of the housing prices falls within a range of $99,000 to $170,162.5.



Graph 1 - Price by Style and SqFt

From the graph we can see that the higher the squared foot, the higher the price, with the most relevant differences for Acadian, colonial and traditional style houses. Nevertheless, a French house is the most expensive, even if it is not one of the most square foot houses (between 4000 and 6000 sqft).



Graph 2 - Scatterplot Matrix of the Numeric Variables

Finally, and in order to analyse the distribution of the different variables taking into account each other we made a scatterplot matrix, presenting all the numerical variables, varying the graphs according to

the variable they are facing. We can see that, in general, 80-year-old homes spend fewer days on the market than newer homes.

2. Development of the Regression Model

2.1. Regression Model Selection

For this work, it is of high importance to understand the impact of the different explanatory variables on house prices and to select the model that best fits our analysis. To make this possible, a filtering of variables was performed starting from the initial model including all explanatory variables.
Then, three more models were created by successively excluding variables with no major impact on price change.

Thus, we consider our initial model as:

Model 1

Price = β0 + β1 sqrt + β2 bedrooms + β3 baths + β4 age + β5 home + β6 occupancy + β7 pool+ β8 style + β9 fireplace + β10 waterfront

By entering the summary command for model 1, we conclude that the occupancy variable is not statistically relevant in predicting house prices.
Through the same, we also obtain the values of the multiple R-Squared (0.658) and the adjusted R-Squared (0.6518), meaning that 65.8% of the variance of the dependent variable (price) is explained by the independent variables included in the model. The adjusted R-Squared value accounts for the number of independent variables included in the model and is adjusted in order to penalise the inclusion of unnecessary variables in the model, which can lead to overfitting.
We also verify that the initial model has no "NA" values, meaning that there is no multicollinearity (occurs when the explanatory variables are excessively correlated).

Removing the occupancy variable, we create a second model having the following structure:

Model 2

Price = β0 + β1 sqrt + β2 bedrooms + β3 baths + β4 age + β5 dom + β6 pool + β7 style + β8 fireplace β9 waterfront

For this model, the R-squared values decreased slightly to 0.6567 and 0.6512 (adjusted R-squared).
Next, we performed a step-by-step search for the best model, based on Akaike's information criterion (AIC) using the step command, which allowed the creation of the following model:

Model 3

Price = β0 + β1 sqrt + β2 bedrooms + β3 baths + β4 age + β5 dom + β6 style + β7 waterfront

Through the given step, variables were removed (one at a time) taking into account their AIC (Akaike Information Criterion), analysed later, until there is no other variable that can be removed without the AIC increasing again.
That said, we verify that the pool and fireplace values were removed because they do not have a relevant statistical weight on the price variation. We also find that the value of the R-squared multiple decreases to 0.6564, despite an increase in the adjusted R-Squared increasing to 0.6515.

Since days on market (dom), do not show statistically significant values, we decided to create a last model to determine the difference of R-squared values:

Model 4

Price = β0 + β1 sqrt + β2 bedrooms + β3 baths + β4 age + β5 style + β6 waterfront

As expected, the R-squared multiple value decreased to 0.6554, as well as the adjusted R-squared value (0.6509), meaning that all variables in the model become statistically more significant.
Even presenting the lowest R-squared value, we must take into account the AIC value to decide which model is best . This value is quite important since it evaluates the quality of the model taking into account its complexity.
Computing the AIC for each model presented, we conclude that the best regression model 3, once presents the lowest value (27257.86), indicating a better fit of the model allied to a lower complexity.

In order to check the quality of model 3 against the most complete model (model 1), we ran the ANOVA command, considering the following possibilities:

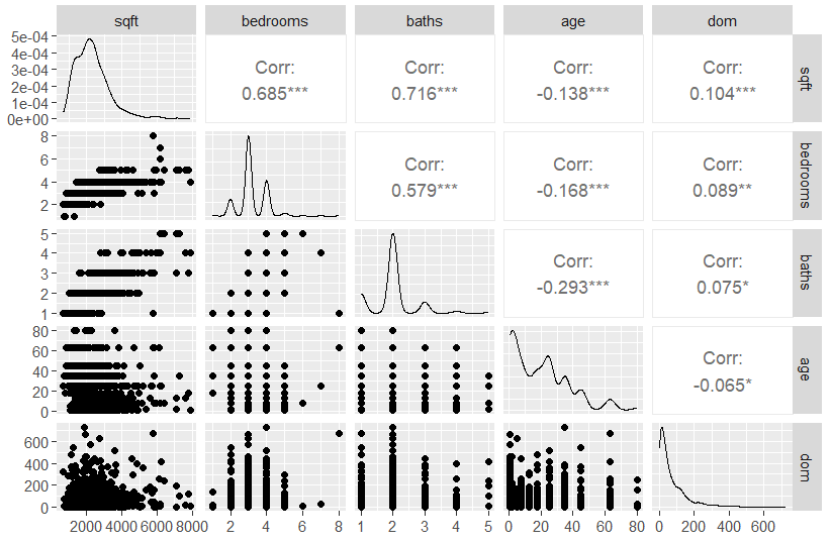
H0: Model 3 (reduced model)
H1: Model 1 (more complete)

Looking at the ANOVA table, the F-statistic for comparing Model 3 and Model 1 is 1.251 with a p-value of 0.2876. This suggests that the null hypothesis does not really compare the predictability of these models, but states instead that the additional variables in model 3 do not add any explanatory power beyond that already given by the variables included in Model 1.
Additionally, as seen below, when we compare the AIC values for the models, we see that Model 3 has the lowest AIC (27257.86), which is slightly lower than Model 1 (27260.77), but the difference is small.

Therefore, based on the F-test and AIC values, Model 3 seems to be the most satisfactory model as it has a simpler model specification and there is no significant improvement in fit by adding additional variables in Model 1.

2.2. Multicollinearity

In this section, we focused on the analysis of multicollinearity between variables. This phenomenon occurs when independent variables in a regression model are strongly correlated, a correlation that is strong is a problem because the independent variables must be independent, and should not have a strong tendency to vary together (and the higher the correlation, the worse). In the presence of multicollinearity, the independent variables tend to change in unison, and this will compromise the interpretation of the regression coefficient.



Graph 3- Correlations of independent variables

By analysing the graph above, we can see that there are some positive correlations between the variables in model 3, however, most of them are not very strong, since they are low values, with the exception of the correlations between Bedrooms and stf (0.685), Baths and Bedrooms (0.579), and between Baths and sqft (0.716), which may be the most significant value.

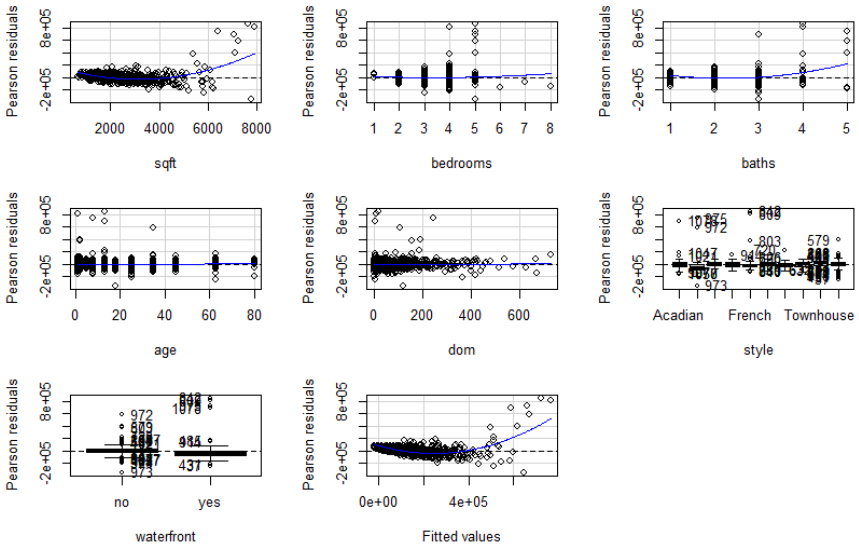
To empirically reinforce the previous inferences, we applied the diagnostic measure VIF to better assess the presence of multicollinearity, if any: it turns out that, according to the diagnostic measure, no variable was found in regression model 3 with the VIF above 10, so there are no indicators of severe multicollinearity, as we expected.

2.3. Autocorrelation

The Breusch-Godfrey test, to assess whether or not there is self-correlation between disturbances, was not performed since our observations are not ordered in time (time series).

2.4. Heteroscedasticity

Heteroscedasticity is characterised as a statistical phenomenon that occurs when the mathematical hypothesis model exhibits distinct variance for Y and X (X1, X2,..., Xn), which contradicts the assumption that the variance of errors is constant and equal for all individuals, i.e., $Var(Yi|Xi) = Var(\mu i) \neq \sigma 2$.



Graph 4 - Heteroscedasticity

Before moving on to the Breusch-Pagan test, by analysing Graph 4, we can visually deduce that there will be a heteroscedasticity problem associated with our final model, since the residuals are not equally distributed around the line where the residuals are zero (e.g., the Age variable).

However, we decided to resort to the Breusch-Pagan test on regression model 3 in order to test the following hypotheses:

$H0$: α1 = α 2 = α 2 = α 3 = α 3 = α 4 = α 5 = α 6 = α 7 = 0 (presence of homoscedasticity)
$H1$: H0 is not verified (presence of heteroscedasticity)

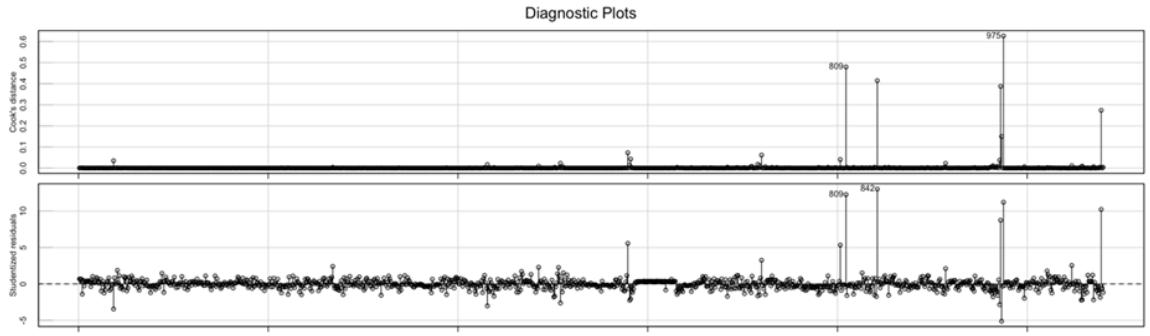
Since p-value is approximately zero, we reject the null hypothesis, so we have a heteroscedastic model, i.e., the OLS method is not producing the "best" estimators.

Thus, to correct the model, we keep the OLS estimates, but replace the classical standard errors by robust standard errors. To do this, we create a list where we include the dependent variable and the explanatory variables in Regression Model 3, and then create a column where the robust standard errors are included. To be able to analyse this effect, we perform a *coeftest* where we conclude that the explanatory variables continue to be explanatory.

2.5. Outliers

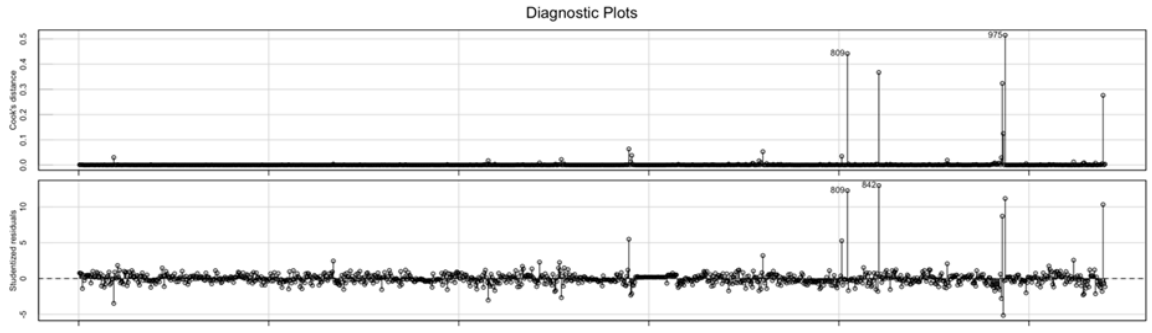
In order to complement the analysis, we proceeded to identify outliers, values that can cause irregularities in the accuracy of the models used.

As such, using the influenceIndexPlot(Mod3res) command, and through Cook's distance, it is possible to identify as the largest outliers, observations 809 and 975. Additionally, concerning the Studentized residuals, we have for example observations 809 and 842, as visible in the following graph.



Graph 5 - Outliers Mod1res

At the same time, a similarity is visible in the graphs for regression model 3 (graph 6), where the values of the outliers are shown to be the same.



Graph 6 - Outliers Mod3res

As such, by identifying the values of the outliers, we can proceed to evaluate their influence on the model 3. As seen in table 3, after removing the largest outliers identified through Cook's distance (observations 809 and 975), all the statistics of the model 3 remained the same. Thus, we can infer that the outliers did not have a significant impact on the model.

	Median	Min	Max	1º quartile	3º quartile
Before	-1969	-351079	858911	-28669	25509
After	-1969	-351079	858911	-28669	25509

Table 3 - Descriptive Statistics of the Model 3

3. Conclusions

In conclusion, the model 3 proved to be the most suitable, based both on the F-test and AIC values because, as we can see on the variables occupancy, swimming pool, fireplace and days on the market, these are not statistically relevant to predict changes in price.

Analysing the multicollinearity, as expected, we can infer that the model 3 has no VIF values above the expected, showing no severe multicollinearity. Also, regarding autocorrelation, the analysis was not possible since our database does not have a temporal spectrum.

During the process of developing model 3, a problem of heteroscedasticity was identified, since the residuals were not equally distributed around the line where the residuals are zero. In order to correct the problem, the OLS estimates were kept unchanged, but at the same time the classical standard errors were changed to robust standard errors.

Lastly, by analysing the model's outliers, two observations seemed to be highly relevant which could have meant a distortion of the normality of the model. However, those values turned out to be insignificant in changing the model's accuracy, as shown in it´s unchanged statistics.

As such, the previous tests verify the reasonableness of the multiple linear regression model, revealing that the explanatory factors included in the model can, in fact, justify the changes in home's price.