Individual Coursework Submission Form

Specialist Masters Programme

Surname: Rawat	**First Name: Nikhil**
MSc in: Business Analytics	**Student ID number: 230044335**
Module Code: SMM634(PRD1 A 2023/24)	
Module Title: Analytics Methods for Business	
Lecturer: Professor Rosalba Radice	**Submission Date: 10/27/2023**

Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** **%**

HOUSE PRICES ANALYSIS

Regression Model

NIKHIL RAWAT
230044335

Aim: The aim of the analysis is to find the best model for the response variable (prices as a dependent variable) as a function of the "Eleven selected" independent variables, which are lot size, bedrooms, bathrooms, stories, driveway, recreation, full base, gas heat, aircon, number of garage places, and prefer. I have selected these variables on the basis of their importance while purchasing a house in practical life. The data used for the regression analysis is "House Prices", which contains information on 546 sales prices of houses sold in the city of "Windsor, Canada" during July-September, 1987.

There were many variables with binary options that have been taken for the regression analysis, and the binary code that is used for yes is '1' and for no is '0'. Even though variables like gas heat, prefer, full base, recreation, aircon and driveway are binary variables mostly containing 'no' as a binary option, which would rarely impact house prices, I have still taken them for our regression analysis because they can capture categorical information, making it possible to include qualitative variables in your model.

Therefore, in this analysis, my aim is to model the price of houses (variable: "price") as a function of various independent variables. To accomplish this, I have chosen to perform a multiple linear regression analysis.

Justification: I have chosen multiple linear regression because it allows us to model the relationship between the response variable (house price) and multiple predictor variables (lot size, number of bedrooms, bathrooms, etc.). This is appropriate, as I have several predictors that can potentially influence house prices.

Model Assumptions

Linearity:	Assuming a linear relationship between the predictors and the house price. I can verify this by plotting the scatterplots of each predictor against the sale price.
Independence:	I assume that the observations in the dataset are independent.
Homoscedasticity:	The variance of the error terms is constant across all levels of the predictors. I can check this assumption by examining a plot of the residuals against the fitted values.
Normality of Residuals:	Assuming that the residuals are normally distributed, I can validate this assumption using a scatter plot or statistical tests.
Level of Significance:	I have considered the variable a significant variable for predicting the price if the P value is lower than 0.05 or 5%.

Maximum Model:

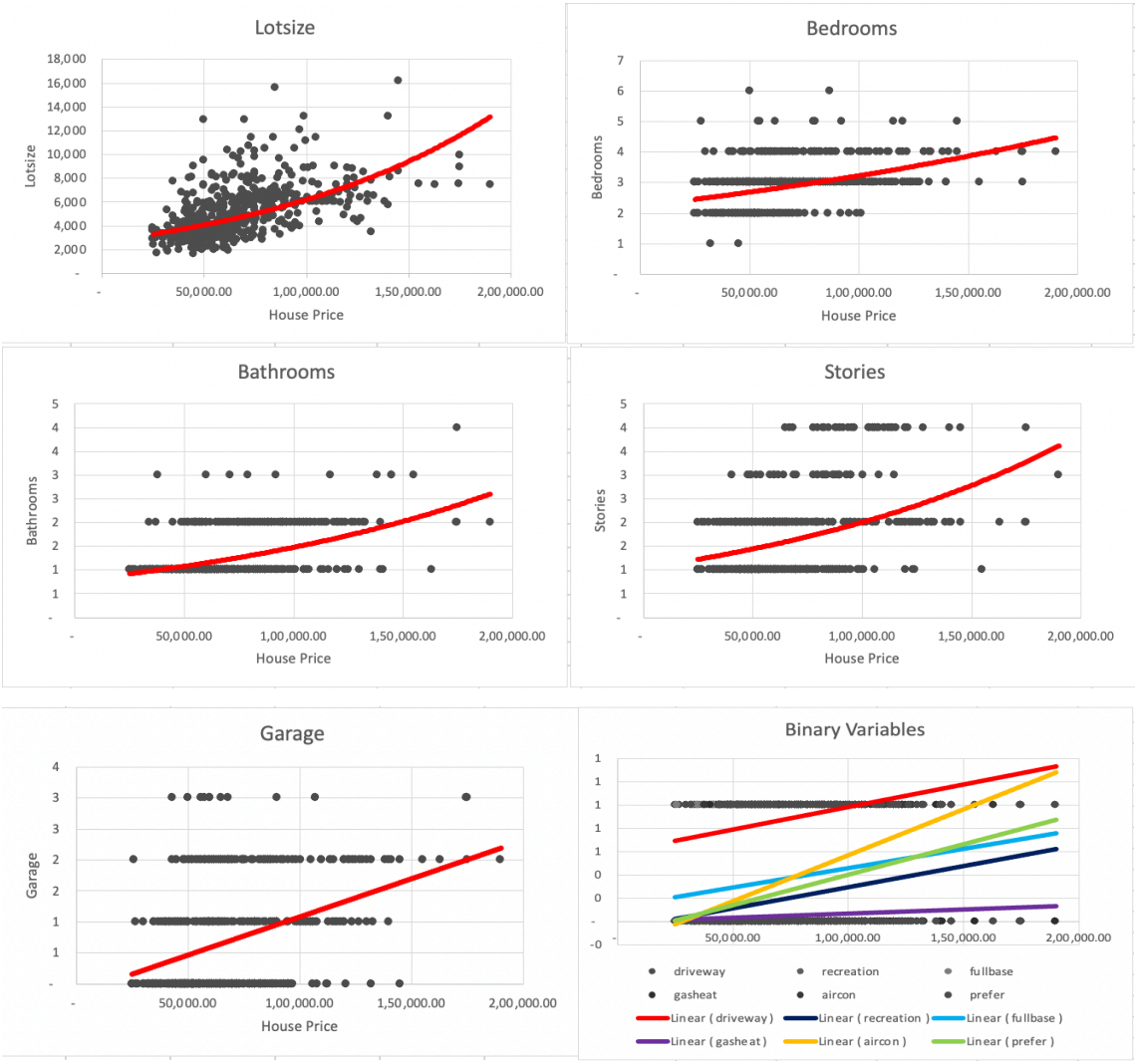
Since the value of K is 11 (the number of predictors), the maximum model can be:

$$\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + E$$

Where X1 = lot size, X2 = number of bedrooms, X3 = number of bathrooms, X4 = number of stories, X5 = recreation, X6 = number of garages, X7 = aircon, X8 = full base, X9 = gas heat, X10 = driveway, and X11 = prefer

Scatter plot and trend line of each predictor against house prices

Scatterplots of the predictors against the house prices visually demonstrate the relationships.



As I can assess from the above scatter plot, each predictor has a linear relationship with house prices. As per the scattered plot, lot size, number of bedrooms, and number of stories

have a more linear relationship with house prices as compared to bathrooms, garages, recreation, aircon, prefer, gas heat, driveway, and full base.

Now let's analyse the multiple linear regression of the data, considering all 11 predictors.

Result Summary

	Coefficients	Standard Error	t Stat	P-value
Intercept	-4038.350425	3409.4713	-1.184450629	0.236761638
Lot size	3.54630297	0.350299955	10.1236181	0.000000
Bedrooms	1832.003466	1047.00022	1.749764165	0.080733
Bathrooms	14335.55847	1489.920852	9.621691279	0.000000
Stories	6556.945711	925.2899059	7.086369007	0.000000
Driveway	6687.77889	2045.245829	3.26991445	0.001145
Recreation	4511.283826	1899.957691	2.374412781	0.017929
Full base	5452.385539	1588.023899	3.43344048	0.000642
Gas heat	12831.40627	3217.597061	3.987884754	0.000076
Aircon	12632.8904	1555.021065	8.123935227	0.000000
Garage	4244.829004	840.544182	5.05009623	0.000001
Prefer	9369.513239	1669.09066	5.613543627	0.000000

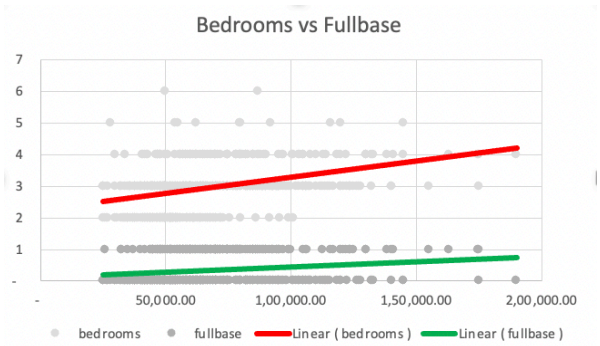
Regression Statistics	
Multiple R	0.82044111
R Square	0.67312362
Adjusted R Square	0.66639021
Standard Error	15423.186
Observations	546

	df	Significance F
Regression	11	0.000000
Residual	534	
Total	545	

After doing multiple linear regression analysis of the data, I found the intercept value to be -4038.350425. And if I go to check the individual P value of each predictor, which is less than 0.05 except in the case of the number of bedrooms as one of the predictors, that means all predictors except "bedrooms" have significance in relation to the house price.

But if I draw the conclusion above that the number of bedrooms does not have any significant relation to the house price, then it will be unfair to exclude it from the model selection.

So, I compared the bedroom variable with other variables and observed that only one variable that is affecting the bedroom variable immensely is full base, due to multicollinearity. As you can see in the graph, I decided to exclude full base because bedrooms are or should be more significant than full base while analysing the sale price of houses sold. Now if I exclude the full base and do analysis on 10 predictors from the model, I get the following result: The P value corresponding to bedrooms is well below 0.5, which means bedrooms are now significantly relatable to the price as per the new model, which includes all variables except the full base. And when I



compare R square in both tested models, there isn't any significant fall in the value of R square if I remove the full base from the model. But I get the P value of bedrooms as 0.02, which is well below the assumption value of significance (i.e., 0.05). F-statistics (significance F) is also well below 1, which means the model has a relationship with house prices.

	Coefficients	Standard Error	t Stat	P-value
Intercept	-3127.9595	3433.24674	-0.9110791	0.362664
Lot size	3.45250249	0.35273665	9.78776223	0.000000
Bedrooms	2341.88694	1046.81255	2.23715978	0.025687
Bathrooms	14819.3431	1498.12365	9.89193588	0.000000
Stories	5674.82223	897.823322	6.32064471	0.000000
Driveway	6886.52508	2064.93702	3.33498069	0.000912
Recreation	6793.14098	1797.78643	3.77861399	0.000175
Gas heat	13016.0322	3249.42341	4.00564363	0.000071
Aircon	12855.2812	1569.25854	8.19194595	0.000000
Garage	4287.96033	848.882037	5.05130294	0.000001
Prefer	10460.8891	1654.98192	6.32084798	0.000000

Regression Statistics	
Multiple R	0.81603158
R Square	0.66590754
Adjusted R Square	0.65966282
Standard Error	15577.9175
Observations	546

	df	Significance F
Regression	10	0.000000
Residual	535	
Total	545	

So, I can conclude that the best model that can predict house prices considering standard error and P-values is lot size, bedrooms, bathrooms, stories, driveway, recreation, gas heat, aircon, number of garage places, and prefer Out of a total of 11 predictors, the best model equation is: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + E$

$$\hat{Y} = -3127.9595 + 3.4525*(\text{Lot size}) + 2341.8869*(\text{Bedrooms}) + 14819.3431*(\text{Bathrooms}) + 5674.8222*(\text{Stories}) + 6886.52508*(\text{Driveway}) + 6793.1409*(\text{Recreation}) + 13016.0322*(\text{Gas heat}) + 12855.2812*(\text{Aircon}) + 4287.9603*(\text{Garage}) + 10460.8891*(\text{Prefer}) + 15577.9175$$

Where β_0(Intercept Value) = -3127.9595, which predicts that if all values become nil, then house prices will have an impact of -3127.9595, which is a constant value.

Limitations of the selected model

It's essential to acknowledge the limitations of the selected model and analysis.

Data Quality: If there is any change, there are potential issues with data quality, such as missing values, outliers, or biases in the data collection process, and how they might have impacted the results.

Non-linearity of the Data: The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that I draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.

Model Assumptions: Address any deviations from the regression model assumptions and how they might affect the accuracy of the analysis. If any of the assumptions turn out to be false or different, then the result would be wrong as well.

Limited Variables: There are many relevant variables that are not included in the dataset for regression analysis that could affect house prices. But in the real world, it could affect house prices, things like locality, crime rates, and facilities like gyms, shops, parks, etc.

Subject to deviation: The above data has been analysed through statistical tools and is subject to deviation from accurate house prices. These deviations are represented by residuals, and the model aims to minimise these deviations but cannot eliminate them.

Improvements to the Analysis

Feature Engineering: We could include some more new features or improve the existing ones in order to get a better relationship with the response variable. Features like time related as house price changes with time, distance to facilities (park, schools, local markets), etc.

Data source: We have to make sure that the source of the data is reliable and can provide the best quality on the basis of which analysis can be done. And if we can expand the data set with external sources and gather more additional information, we might improve model accuracy. External sources like the inflation rate, interest rates, real estate market data, etc.

Outlier Detection: To locate and deal with outliers that might be affecting the model, use robust methods. Another strategy is to eliminate extreme outliers. Removing too many outliers could result in information loss, so we must proceed with caution.

Variable Selection: To see if a more parsimonious model can be produced, one can investigate different combinations of predictor variables. In order to avoid multicollinearity, remove one variable from pairs with a high correlation. Make use of various variable selection techniques, such as backward or forward selection.

Cross-validation: Use cross-validation methods to determine the generalizability of the model. Through multiple training sessions and evaluation of the model on different sets of data, cross-validation helps detect and prevent overfitting.

In conclusion, the chosen regression model provides valuable insights into the factors affecting house prices in Windsor, Canada. By addressing the limitations and implementing improvements, the analysis can be further refined and made more robust.

Citation: Data Source: http://qed.econ.queensu.ca/jae/1996-v11.6/anglin-gencay/