

BUSI 650: Business Analytics Predictive Model

Task Requirement: (Question)

Selling property to buy cryptocurrency and tech stocks

Grading: 15% of overall grade

Submission: 9AM, Aug 12th on Turnitin

Due to the increasing property prices in the Seattle Area and the bear market in tech stocks and cryptocurrencies, Leo is looking to sell his property, transfer the funds into more volatile assets, and take advantage of the discounted prices. While realtors are convenient, they take 3% of the profit, which is significant when considering the potential value of the property. Thus, Leo is deciding to perform the market research himself and conduct the sale privately. Please use the data from Housing Prices to create a regression model with a 5% level of significance and propose a listing price. You are free to make assumptions as necessary, but please make note of it in the report. There is no need to separate the data into training and testing data, please use all of it (barring outliers, etc) in your regression analysis. This report should be no longer than 3 pages, with your supporting tables and figures included in the appendix (max 2 pages). Your report should have the following sections: Executive Summary or Introduction, Methods, Regression Analysis, Discussion, Conclusion Rubric (total of 15 points): (1) Executive Summary or Introduction with recommendation(s) (3) Methods (1) Assumptions (2) Identifying and eliminating problematic data (incomplete, outliers, etc) (6) Regression Analysis (1) Performing regression analyses and eliminating irrelevant factors (1) Performing regression analyses and incorporating relevant factors (2) Generating a multivariate regression formula for housing price (2) Pricing the property in question (4) Discussion (3) Insight into the relevant and irrelevant factors (1) Additional insights that may affect property pricing (1) Conclusion Property Information: Zip Code: 98078 (47.5601, -122.2139) Age: 51 years, unrenovated Rooms: 3 bedrooms, 1.5 bathroom, 2 floors Size (square feet): 5300 Living, 15420 Lot, 3800 Above Ground, 1500 Basement Ratings: View (2), Grade (9), Condition (4), No waterfront (0) Days on Market: 0 (so far) Price: ????

Answer

Introduction

The business analytics report deals with the various factors listed along with property prices in Seattle. Based on the case study, Leo is looking to sell his property to make his assets more volatile in the market as his assets are blocked in his property. Although Leo was trying to sell his property through a realtor with a profit of 3% on the property sold, he decided to do market research. Leo has collected data from Housing prices in Seattle, which have variables such as Price, Listing ID, Listing Date, Bedrooms, Bathrooms, Floors, Waterfront, View, Condition, Grade, Living (sqft), Lot (sqft), Above Ground (sqft), Basement (sqft), Year Built, Year Renovated, Zip Code, Latitude and Longitude. Before conducting the regression, certain assumptions would be made, and some irrelevant data would either be removed or converted into proper data. Regression analysis would be done to analyse relevant factors for the dependent variable "listing price". Further, the results calculated would result in insights. The recommendations given for the analysis are:

- Leo should draw on the coefficients to calculate the listing price.
- The basement variable did not add any relevance to the data.
- Other variables like neighbourhood group and no. of occupancy in a house could have added value to the model.
- Other software handling big data could have been used.

Methods

The methods used to analyse the given data set would be using regression analysis in excel. The regression analysis would be taking price as a dependent variable and other relevant factors as independent variables.

Assumptions

The assumptions taken after eliminating the irrelevant factors in the data are:

- To calculate days on the market, 13/8/2022 as present date would be considered today's date.
- The year of renovation would take two values as 0 – not renovated and 1 – renovated.
- The age of the property would be in year Today – Year Built.

Identifying and Eliminating Problematic data

To identify the problematic data, it is important descriptive, and box plots are calculated so that the rows with extreme values are removed to give better results. After regression on the

raw data, it was seen that Basement and Year Built presented with "#NUM" in their p-value. To rectify, basement and lot (sq ft) outliers were taken out, such as basement ≥ 1275 and lot ≥ 4030 were removed (boxplots in appendix). The regression still showed basement as an inconsistent variable. As a result, the basement was removed as a whole variable. The sample size came out as 20310.

The data removed – listing id, year built, basement and listing date (in final regression). The changed data – age and days on market.

Regression Analysis

Eliminating Irrelevant factors in Regression

The irrelevant factors in regression after the first regression with all factors were considered, the factors that seem irrelevant at their p-values were floors, the lot (sqft), basement and year built. The year built was changed to "age", and the listing date was considered "days on the market."

Proper Regression

After outliers removal with all outliers through box plots, all variables were significant (all p values < 0.05), but the basement variable still showed as p-value as "#DIV/0!." However, "basement" as a variable was removed for proper variables.

The proper *multivariate regression* equation came out as:

$$-1451037.602 -16966.303*\text{Bedrooms} + 30612.63934*\text{Bathrooms} + 23253.6174*\text{Floors} + 361979.0263*\text{Waterfront} + 51209.38608*\text{View} + 26904.04591*\text{Condition} + 100300.1252*\text{Grade} + 108.7257544*\text{Living (sqft)} + 0.125918081*\text{Lot (sqft)} + 16.72167173*\text{Above Ground (sqft)} + 18.47010288*\text{Year Renovated} -405.7345469*\text{Zip Code} + 581843.833*\text{Latitude} - 105012.6893*\text{Longitude} -98.18943816*\text{Days on Market} + 2321.598084*\text{Age}$$

The *pricing in the property question* has been given in the table below:

	<i>Coefficient values</i>	<i>Property Info</i>	<i>Price</i>
Intercept	-1451037.602		-1451037.602
Bedrooms	-16966.303	3	-50898.90899
Bathrooms	30612.63934	1.5	45918.959
Floors	23253.6174	2	46507.2348
Waterfront	361979.0263	0	0
View	51209.38608	2	102418.7722
Condition	26904.04591	4	107616.1836

Grade	100300.1252	9	902701.1272
Living (sqft)	108.7257544	5300	576246.4984
Lot (sqft)	0.125918081	15420	1941.656804
Above Ground (sqft)	16.72167173	3800	63542.35258
Year Renovated	18.47010288	0	0
Zip Code	-405.7345469	98078	-39793632.89
Latitude	581843.833	47.5601	27672550.88
Longitude	-105012.6893	-122.2139	12834010.31
Days on Market	-98.18943816	0	0
Age	2321.598084	51	118401.5023
		Price =	\$ 1,176,286.08

Discussion

- The relevant data that had added value to the data are – Bedrooms, Bathrooms, Floors, Waterfront, View, Condition, Grade, Living (sqft), Lot (sqft), Above Ground (sqft), Latitude, Longitude, Days on Market, and Age.
- The data that could have been avoided is Year Renovated and Zip code. Zip code is not relevant data for property sales as latitude and longitude are enough to make an exact location. The basement had many outliers as the extreme values covered maximum data, which was better to be removed as a variable as a whole.
- The year renovated could have used a dummy variable as 0 = not renovated and 1 = renovated.
- The additional insight that could be added would be deletion of irrelevant factors and including other factors like neighbourhood group (areas which could have been categorised) and type of home whether can fill how many occupants, etc. (Seth, 2022)
- The data could have been analysed using software adept at handling big data like R, and SPSS, which could have dealt with outliers and regression analysis (Biju and Matthew, 2017).

Conclusion

The predictive model made for Leo took a lot of trial and error with and without outliers and independent variables to achieve the listing price. Moreover, the regression predicted can make an accurate view of the listing price on Leo's property, which will help him to sell in the market.

References

Biju, S.M. & Mathew, A. (2017). Comparative Analysis of Big Data Analytics Software in Assessing Sample Data. *Journal of International Technology and Information Management*, 26(2), pp.2-22.

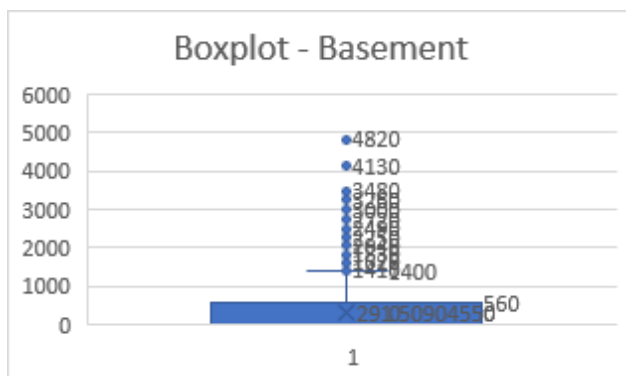
Seth, K. (2022). *U.S. Airbnb Open Data*. Retrieved on 12 August 2022 from <https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data>

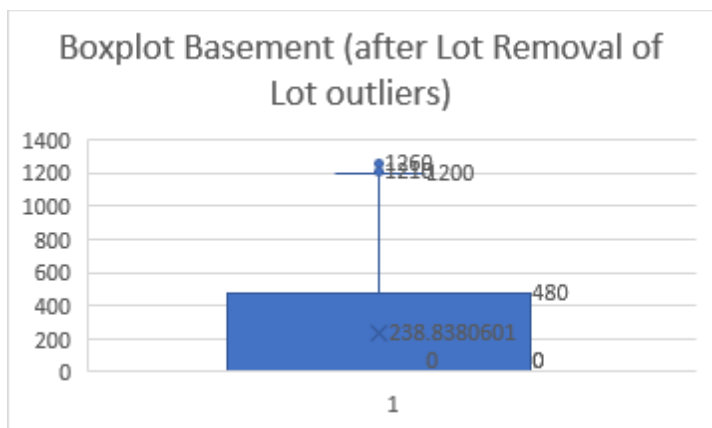
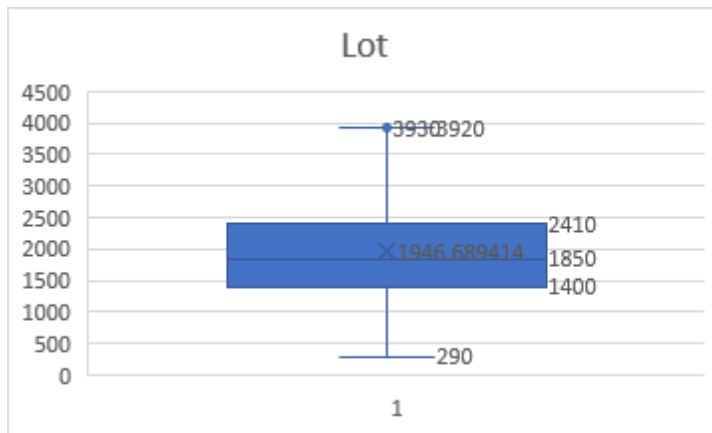
Appendices

First Regression (with all factors)

SUMMARY OUTPUT									
<i>Regression Statistics</i>									
Multiple R	0.835979								
R Square	0.698861								
Adjusted R	0.698606								
Standard Error	201535								
Observations	21613								
<i>ANOVA</i>									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>				
Regression	16	2.04E+15	1.27E+14	3341.387	0				
Residual	21597	8.77E+14	4.06E+10						
Total	21613	2.91E+15							
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
Intercept	9319727	2884151	3.231359	0.001234	3666579	14972875	3666579	14972875	
Bedrooms	-35473.2	1892.408	-18.745	8.82E-78	-39182.4	-31763.9	-39182.4	-31763.9	
Bathroom	40998.75	3255.64	12.59314	3.09E-36	34617.45	47380.05	34617.45	47380.05	
Floors	4890.195	3575.607	1.367655	0.171434	-2118.26	11898.65	-2118.26	11898.65	
Waterfront	578959.8	17376.8	33.31797	2.1E-237	544900	613019.6	544900	613019.6	
View	54862.36	2116.216	25.92475	6E-146	50714.43	59010.3	50714.43	59010.3	
Condition	25626.76	2352.719	10.8924	1.48E-27	21015.25	30238.26	21015.25	30238.26	
Grade	99889.29	2065.839	48.3529	0	95840.09	103938.5	95840.09	103938.5	
Living (sqft)	154.582	4.30623	35.8973	3.7E-274	146.1415	163.0225	146.1415	163.0225	
Lot (sqft)	-0.05279	0.034764	-1.51865	0.128866	-0.12094	0.015346	-0.12094	0.015346	
Above Ground	34.55614	4.319981	7.999142	1.32E-15	26.08866	43.02362	26.08866	43.02362	
Basement	0	0	65535	#NUM!	0	0	0	0	
Year Built	-2638.2	72.726	-36.2759	#NUM!	-2780.75	-2495.65	-2780.75	-2495.65	
Year Renovation	18.37389	3.65531	5.02663	5.03E-07	11.20921	25.53857	11.20921	25.53857	
Zip Code	-602.585	32.88502	-18.324	1.95E-74	-667.042	-538.128	-667.042	-538.128	
Latitude	607247	10734.02	56.57216	0	586207.5	628286.5	586207.5	628286.5	
Longitude	-207930	12946.78	-16.0604	1.04E-57	-233307	-182553	-233307	-182553	

Box Plots for Outliers





SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.818038							
R Square	0.669187							
Adjusted R	0.668926							
Standard Error	152737.2							
Observations	20310							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	16	9.58E+14	5.99E+13	2565.611	0			
Residual	20293	4.73E+14	2.33E+10					
Total	20309	1.43E+15						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-1451038	2234321	-0.64943	0.516067	-5830487	2928411	-5830487	2928411
Bedrooms	-16966.3	1534.998	-11.053	2.55E-28	-19975	-13957.6	-19975	-13957.6
Bathroom:	30612.64	2650.846	11.54825	9.4E-31	25416.77	35808.51	25416.77	35808.51
Floors	23253.62	2823.159	8.236737	1.88E-16	17720	28787.24	17720	28787.24
Waterfront	361979	15949.72	22.695	1.3E-112	330716.3	393241.8	330716.3	393241.8
View	51209.39	1812.152	28.25888	2.4E-172	47657.42	54761.35	47657.42	54761.35
Condition	26904.05	1844.882	14.58307	6.29E-48	23287.93	30520.16	23287.93	30520.16
Grade	100300.1	1660.772	60.39367	0	97044.88	103555.4	97044.88	103555.4
Living (sqft)	108.7258	4.06128	26.7713	3.4E-155	100.7653	116.6862	100.7653	116.6862
Lot (sqft)	0.125918	0.028798	4.372477	1.23E-05	0.069472	0.182364	0.069472	0.182364
Above Ground	16.72167	3.845938	4.347879	1.38E-05	9.183322	24.26002	9.183322	24.26002
Year Renovation	18.4701	2.909152	6.348964	2.21E-10	12.76793	24.17228	12.76793	24.17228
Zip Code	-405.735	25.8757	-15.6801	4.36E-55	-456.453	-355.016	-456.453	-355.016
Latitude	581843.8	8329.897	69.85006	0	565516.6	598171.1	565516.6	598171.1
Longitude	-105013	10093.81	-10.4037	2.76E-25	-124797	-85228	-124797	-85228
Days on Market	-98.1894	9.512404	-10.3223	6.44E-25	-116.835	-79.5444	-116.835	-79.5444
Age	2321.598	57.02409	40.71259	0	2209.826	2433.37	2209.826	2433.37