MSA

Microsoft Student Accelerator

Linear Regression Analysis of House Price

MSA PHASE 1

Jennifer Siahaan | 26 July, 2020

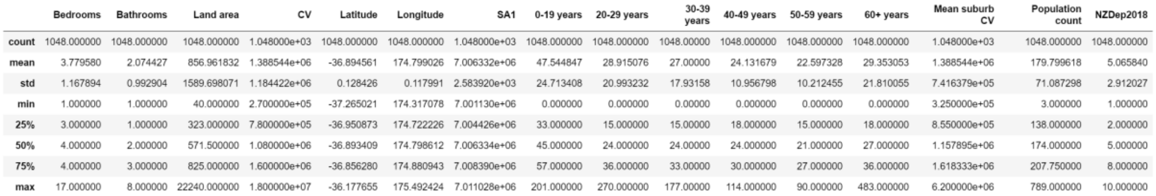
Executive Summary

The dataset was the house prices in Auckland, containing information on the characteristics of houses in Auckland and the area they were in, including the address, suburb, latitude and longitude, number of bedrooms and bathrooms, land area, CV, SA1 (an area unit classification), and the number of people within age groups living in the SA1 unit area based on the 2018 census. Two new columns were added to the dataset from other datasets: population count from the 2018 NZ census, and deprivation index from University of Otago's research on NZ indexes of deprivation.

Originally, the dataset contained 1051 observations for each of the 17 variables. However, after data cleaning, 1048 observations were left. The analysis was based on these 1048 observations and 10 variables, after taking out 7 variables due to negative correlations with the response variable. The response variable, the variable we are going to predict, was the CV. The rest of the variables were explanatory variables that help 'explain' the CV value.

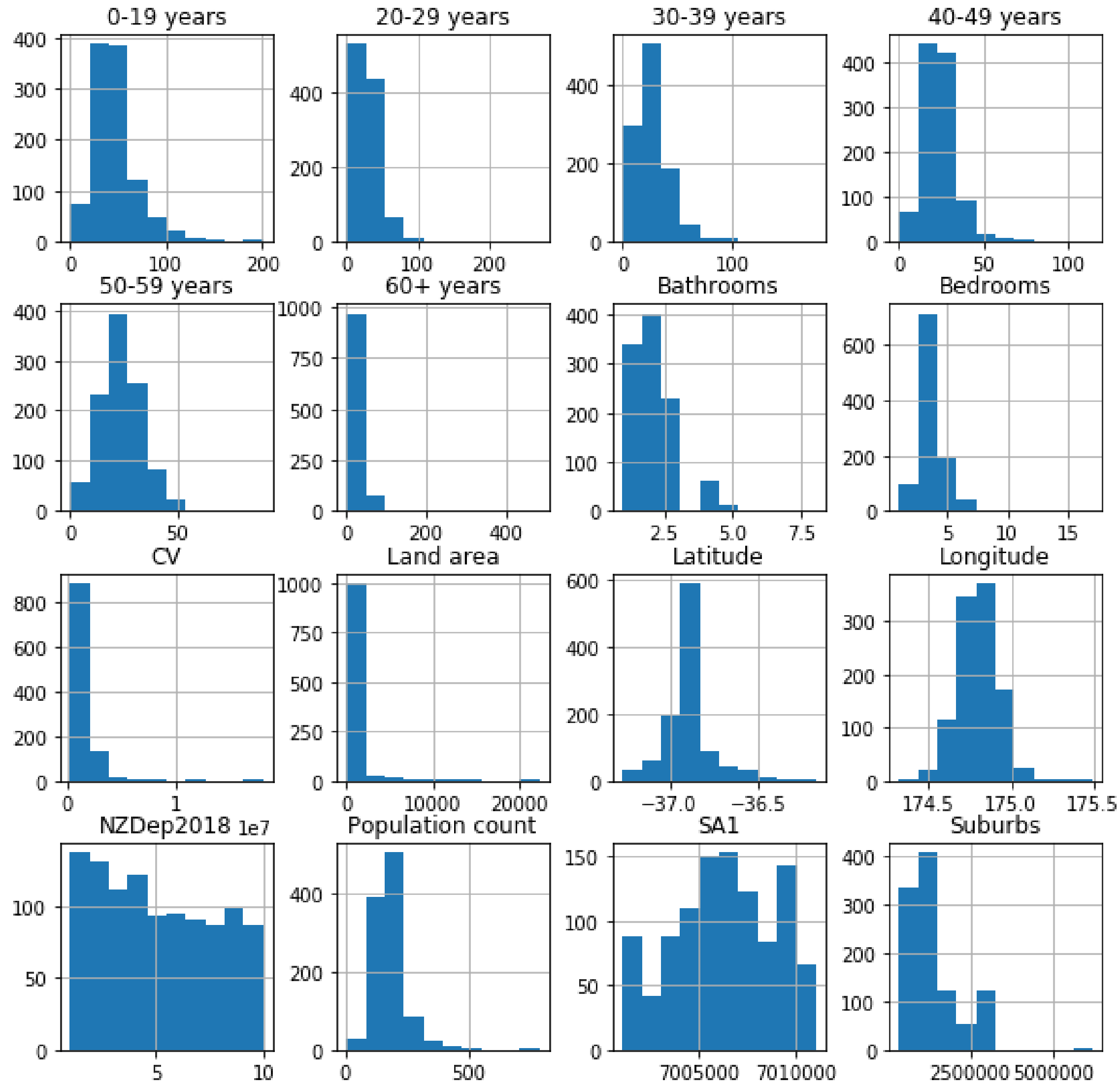
After analysing the data, its correlations and its patterns, the dataset was split into training and testing datasets, as part of the linear regression model creation to predict the CV. Because the model score was too low the first time, the categorical variable 'Suburbs' was replaced with the numerical variable 'mean suburb CV' to slightly improve the score of the linear regression model. The model was then fitted again for the second time.

Initial Data Analysis



	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Mean suburb CV	Population count	NZDep2018
count	1048.000000	1048.000000	1048.000000	1.048000e+03	1048.000000	1048.000000	1.048000e+03	1048.000000	1048.000000	1048.00000	1048.000000	1048.000000	1048.000000	1.048000e+03	1048.000000	1048.000000
mean	3.779580	2.074427	856.961832	1.388544e+06	-36.894561	174.799026	7.006332e+06	47.544847	28.915076	27.00000	24.131679	22.597328	29.353053	1.388544e+06	179.799618	5.065840
std	1.167894	0.992904	1589.698071	1.184422e+06	0.128426	0.117991	2.583920e+03	24.713408	20.993232	17.93158	10.956798	10.212455	21.810055	7.416379e+05	71.087298	2.912027
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	3.250000e+05	3.000000	1.000000
25%	3.000000	1.000000	323.000000	7.800000e+05	-36.950873	174.722226	7.004426e+06	33.000000	15.000000	15.00000	18.000000	15.000000	18.000000	8.550000e+05	138.000000	2.000000
50%	4.000000	2.000000	571.500000	1.080000e+06	-36.893409	174.798612	7.006334e+06	45.000000	24.000000	24.00000	24.000000	21.000000	27.000000	1.157895e+06	174.000000	5.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.856280	174.880943	7.008390e+06	57.000000	36.000000	33.00000	30.000000	27.000000	36.000000	1.618333e+06	207.750000	8.000000
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.00000	114.000000	90.000000	483.000000	6.200000e+06	789.000000	10.000000

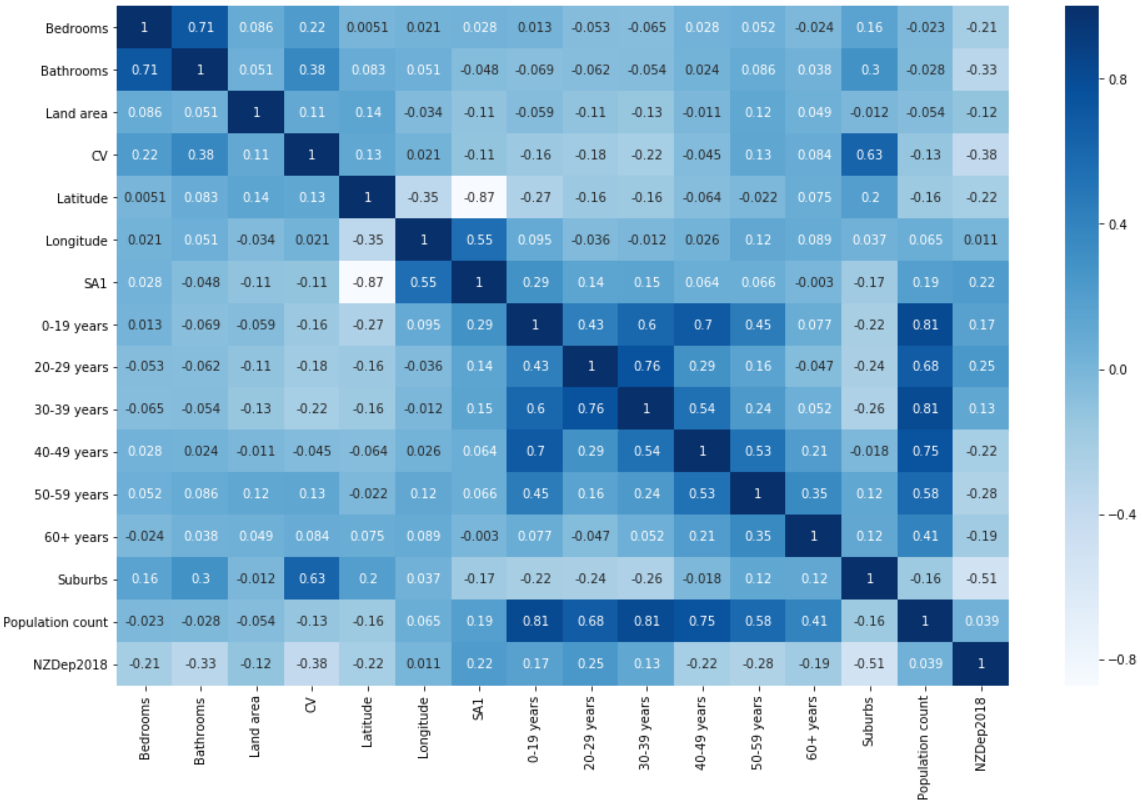
The data was analysed initially by calculating and observing the summary and descriptive statistics above. It showed the count, mean, standard deviation, minimum, maximum, and the lower, mid, and higher quartile of each of the variables in the dataset.



Histograms were also plotted above as part of the initial data analysis, to observe the distribution of the values of each of the variables in the dataset.

Data Correlations and Patterns Analysis

The correlation between the numeric columns were calculated and observed in the correlation plot below. Correlation ranges between -1 and 1, where 1 is perfectly positively correlated and -1 perfectly negatively correlated. The darker the colour, the higher the correlation value, and vice versa.



	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs	Population count	NZDep2018
Bedrooms	1	0.71	0.086	0.22	0.0051	0.021	0.028	0.013	-0.053	-0.065	0.028	0.052	-0.024	0.16	-0.023	-0.21
Bathrooms	0.71	1	0.051	0.38	0.083	0.051	-0.048	-0.069	-0.062	-0.054	0.024	0.086	0.038	0.3	-0.028	-0.33
Land area	0.086	0.051	1	0.11	0.14	-0.034	-0.11	-0.059	-0.11	-0.13	-0.011	0.12	0.049	-0.012	-0.054	-0.12
CV	0.22	0.38	0.11	1	0.13	0.021	-0.11	-0.16	-0.18	-0.22	-0.045	0.13	0.084	0.63	-0.13	-0.38
Latitude	0.0051	0.083	0.14	0.13	1	-0.35	-0.87	-0.27	-0.16	-0.16	-0.064	-0.022	0.075	0.2	-0.16	-0.22
Longitude	0.021	0.051	-0.034	0.021	-0.35	1	0.55	0.095	-0.036	-0.012	0.026	0.12	0.089	0.037	0.065	0.011
SA1	0.028	-0.048	-0.11	-0.11	-0.87	0.55	1	0.29	0.14	0.15	0.064	0.066	-0.003	-0.17	0.19	0.22
0-19 years	0.013	-0.069	-0.059	-0.16	-0.27	0.095	0.29	1	0.43	0.6	0.7	0.45	0.077	-0.22	0.81	0.17
20-29 years	-0.053	-0.062	-0.11	-0.18	-0.16	-0.036	0.14	0.43	1	0.76	0.29	0.16	-0.047	-0.24	0.68	0.25
30-39 years	-0.065	-0.054	-0.13	-0.22	-0.16	-0.012	0.15	0.6	0.76	1	0.54	0.24	0.052	-0.26	0.81	0.13
40-49 years	0.028	0.024	-0.011	-0.045	-0.064	0.026	0.064	0.7	0.29	0.54	1	0.53	0.21	-0.018	0.75	-0.22
50-59 years	0.052	0.086	0.12	0.13	-0.022	0.12	0.066	0.45	0.16	0.24	0.53	1	0.35	0.12	0.58	-0.28
60+ years	-0.024	0.038	0.049	0.084	0.075	0.089	-0.003	0.077	-0.047	0.052	0.21	0.35	1	0.12	0.41	-0.19
Suburbs	0.16	0.3	-0.012	0.63	0.2	0.037	-0.17	-0.22	-0.24	-0.26	-0.018	0.12	0.12	1	-0.16	-0.51
Population count	-0.023	-0.028	-0.054	-0.13	-0.16	0.065	0.19	0.81	0.68	0.81	0.75	0.58	0.41	-0.16	1	0.039
NZDep2018	-0.21	-0.33	-0.12	-0.38	-0.22	0.011	0.22	0.17	0.25	0.13	-0.22	-0.28	-0.19	-0.51	0.039	1

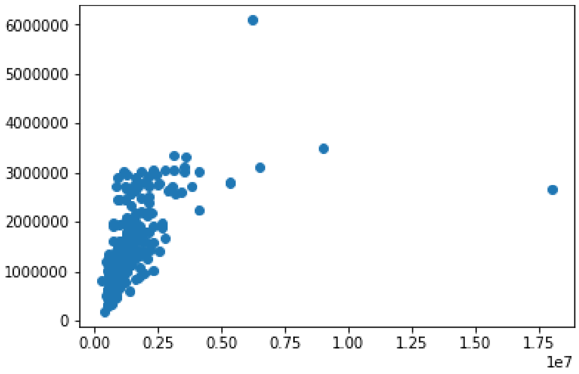
Correlation is a good way to remove variables irrelevant to the variable we want to predict when building a machine learning model. For example, the variables SA1, 0-19 years, 20-29 years, 30-39 years, 40-49 years, population count, and NZDep2018 are negatively correlated with the variable CV. Therefore, they will be excluded in the machine learning model predicting CVs. The variables 50-59 years and 60+ years are very similar to the variables 0-19 years, 20-29 years, 30-39 years, 40-49 years. Thus, they will also be excluded from the machine learning model predicting CVs, to maintain consistency.

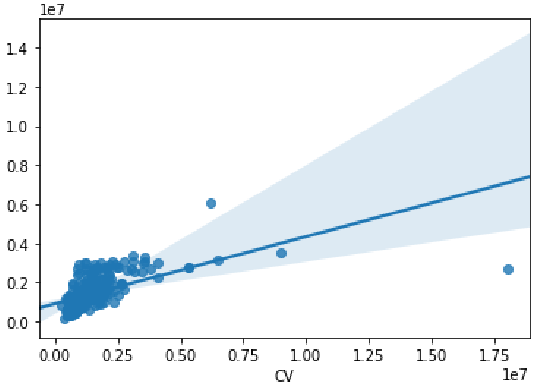
Model Building and Commentary

Linear regression is used to predict numerical values and trends based on available data. It is a type of regression model that is predictive and supervised. Predictive modelling takes a look at historical data to uncover patterns that can be used to predict future trends, while supervised learning is using a data set as the basis for predicting the regression of data through the use of machine learning models.

Linear regression examines multiple variables at the same time and plots points based on them. It draws a line through the center of the points. Analysis of that line will show how closely the explanatory variables affect the response variable. Linear regression is then proceeded by relevance analysis to identify significantly relevant attributes. Irrelevant attributes can be excluded from the analysis. In this case, the relevance analysis was done by analysing the correlations between CV and other variables, as explained in the previous section. Seven variables remained to be used in the linear regression model – bedrooms, bathrooms, land area, CV, latitude, longitude, and the mean suburb CV.

The linear regression machine learning model was trained with 70% of the dataset and tested with the remaining 30%, as these percentages, from trial and error, increased the model accuracy without risk of overfitting it. The dataset was split randomly.





The model yielded the results above (left – scatter plot, right – linear regression plot). The blue regression line indicates the model's predictions at varying variable values. The score for this model is 0.3402701355419463, or around 34% in accuracy.

Conclusion

This analysis of house price has shown that the house's capital value can be predicted from its number of bedrooms and bathrooms, land area, latitude, longitude, and mean suburb CV using a linear regression model. However, it does not have a high accuracy rate. More research and refining of this model, as well as exploring other alternatives, should be taken into consideration in the future.