Linear Regression Analysis of Auckland House Prices

Sakyawira Nanda Ruslim, July 2020.

Executive Summary

The dataset is housing details in Auckland. The 2018 Census population is downloaded from koordinates.com, the 2018 Deprivation Index is downloaded from University of Otago's website. The rest of the data is provided by Microsoft Student Accelerator. The dataset contains the number of bedrooms, number of bathrooms, the address, the land area, CV (Capital Value), latitude, longitude,number of populations in certain ages, population number, and deprivation index. There were only 3 null values in the 1051 rows dataset. After exploring the data by visualizing the correlation between each numerical variable, no highly correlated variables are found. Linear Regression algorithms have been tested for this train dataset and have a score of 0.167.

Initial Data Exploration

The initial exploration of the data began with some summary for each attribute's minimum, maximum, mean, median, standard deviation, and distinct count. The results were taken from a cleaned dataset with 1048 entries, as shown here:

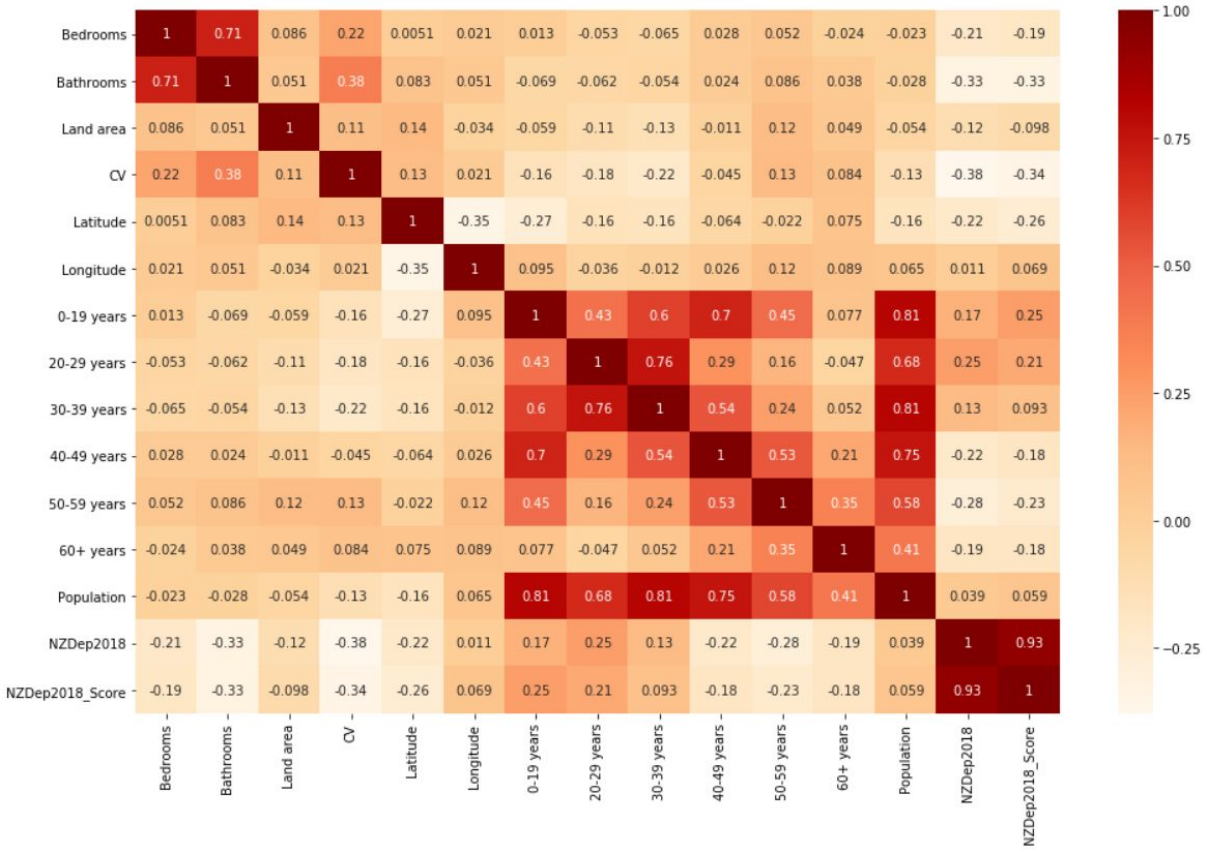
	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude
count	1048.000000	1048.000000	1048.000000	1.048000e+03	1048.000000	1048.000000
mean	3.779580	2.074427	856.961832	1.388544e+06	-36.894561	174.799026
std	1.167894	0.992904	1589.698071	1.184422e+06	0.128426	0.117991
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078
25%	3.000000	1.000000	323.000000	7.800000e+05	-36.950873	174.722226
50%	4.000000	2.000000	571.500000	1.080000e+06	-36.893409	174.798612
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.856280	174.880943
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424

0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Population	NZDep2018	NZDep2018_Score
1048.000000	1048.000000	1048.00000	1048.000000	1048.000000	1048.000000	1048.000000	1048.000000	1048.000000
47.544847	28.915076	27.00000	24.131679	22.597328	29.353053	179.799618	5.065840	986.518130
24.713408	20.993232	17.93158	10.956798	10.212455	21.810055	71.087298	2.912027	94.271599
0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	3.000000	1.000000	849.000000
33.000000	15.000000	15.00000	18.000000	15.000000	18.000000	138.000000	2.000000	918.000000
45.000000	24.000000	24.00000	24.000000	21.000000	27.000000	174.000000	5.000000	959.000000
57.000000	36.000000	33.00000	30.000000	27.000000	36.000000	207.750000	8.000000	1031.000000
201.000000	270.000000	177.00000	114.000000	90.000000	483.000000	789.000000	10.000000	1380.000000

It should be noted that some attributes have standard deviations of 3 or more, in particular: the land area, CV (Capital Value), number of populations in certain ages, population number, and deprivation index (Score format).

Correlation and Relationships

The correlation between the numeric columns were calculated and observed in the below correlation plot. (The right color bar indicated the correlation values. For example, dark red means correlation value is 1 and light beige means correlation value is negative 0.25)



The graph shows that *bedrooms number, bathrooms number, land Area, latitude,* and *number of population of 50-59 years old* have strong positive correlation with the *Capital Value* of the houses. On the other hand *total number of population*, *Deprivation Index*, and *number of population of 0-49 years old* have strong negative correlation with the *Capital Value* of the houses.

Analysis

Three null values were found in the dataset with 1051 entries, therefore dropping the rows that contain them is considered to have no significant impact in the model fitting. The *Land area* attribute has 'm^2' concatenated inside the values, therefore conversion from string to float was done.

Previously there were also attributes such as *Address* and *Suburb* but they were dropped because Linear Regression only works with numerical attributes. Numerical attributes that represent IDs have also been dropped.

In this analysis, a Linear Regression algorithm was used. These algorithms were trained with 70% of the data. Testing the model with the remaining 30% of the data yielded 0.16.

Conclusion

This analysis has shown that the Capital Value of a house can not be confidently predicted from the number of bedrooms, number of bathrooms, the land area, latitude, longitude,number of populations in certain ages, population number, and deprivation index. In particular, the Linear Regression algorithm only has a score of 16%.