Linear Regression Analysis of Auckland House Prices

Nikesh Lala
Computer Systems Engineering
July 26 2020

Executive Summary

The dataset consists of information related to Auckland properties as well as extra data correlated to these properties which includes the population, which was collected from the NZ 2018 Census relating to the number of people residing in the area (SA1). Also, the deprivation index was included, a decile system which ranks the socio-economic measure of the area with a lower score associated with a better environment.

The additional information included in the dataset includes basic property information such as number of bathrooms, bedrooms. The land area (m^2), capital valuation (NZD), details on the location such as address, suburb SA1 which is a code for designated areas and latitude and longitude (used to get the population). Details were also included for number of people within age brackets from the area (SA1). The dataset contained 1043 distinct properties in Auckland which a total of 17 associated variables.

Through statistical analysis of the dataset the, by visualising the relationships between attributes and the individual distribution of their datapoints the optimal linear regression model was developed. Many variations were tested through trial and error by exploring multiple different options to transform the data as well as choosing which variables produced the best results. The models were evaluated on the R^2 value and the RMSE metrics which are advised for linear regression models on continuous data.

Initial Data Analysis

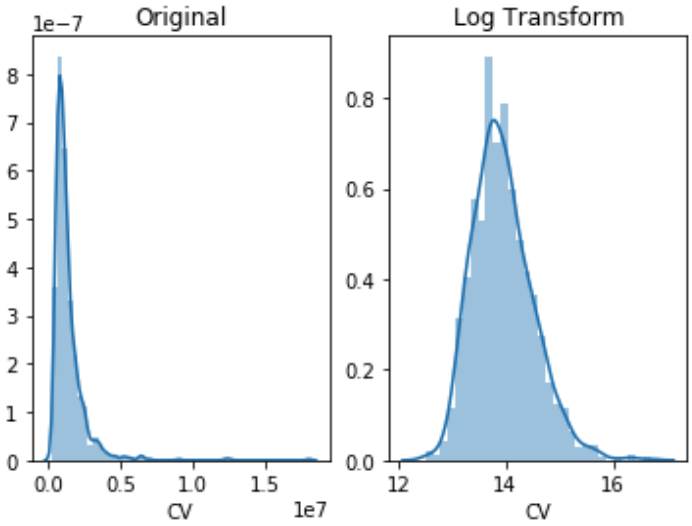
Data analysis was comprised of cleaning the data initially by filling in missing values from the data frame by manually through analysis of properties with similar attributes by manually adding data and using a model to predict the

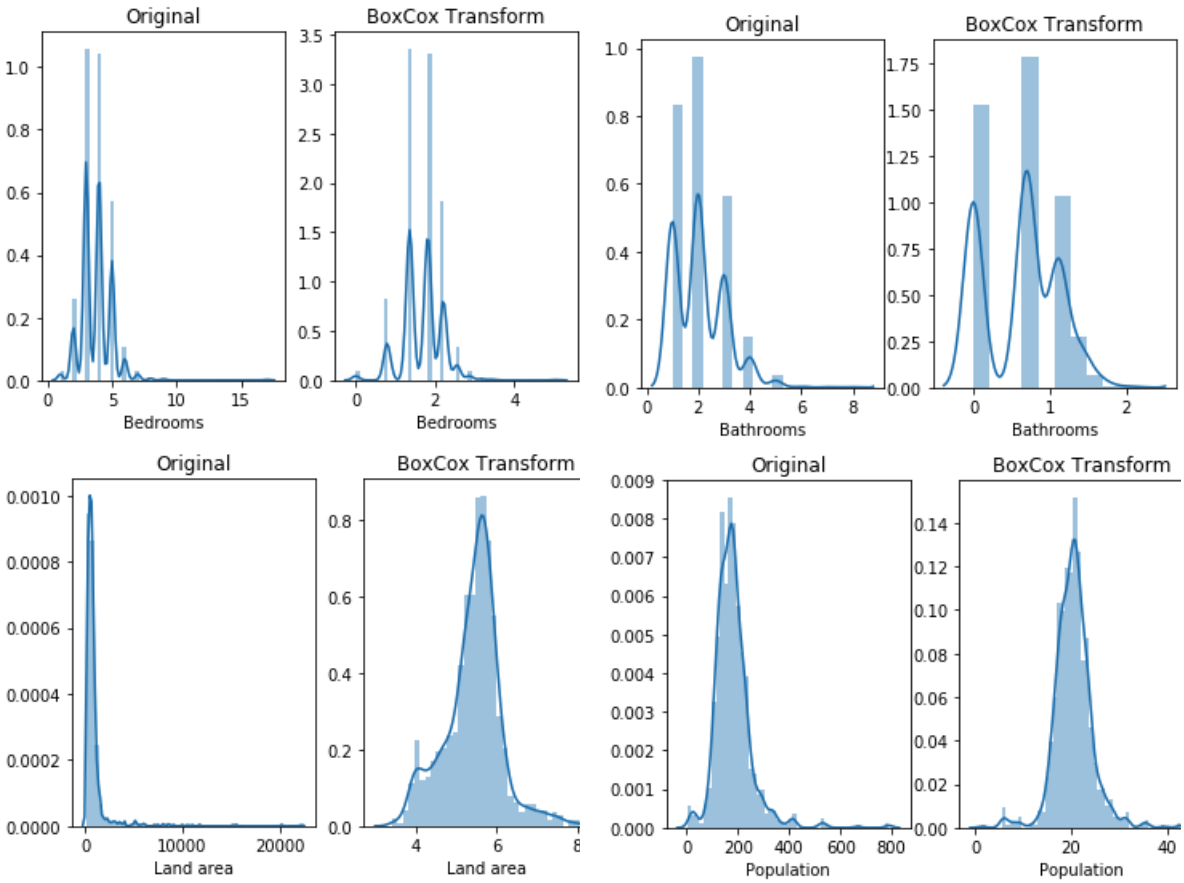
values. Duplicate rows were detected and removed as well as dropping erroneous rows that included implausible data. The below table is a statistical breakdown of the 1041 different properties included in the final data frame capturing all 17 different variables.

	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	Population	NZDep2018
count	1041.000000	1041.000000	1041.000000	1.041000e+03	1041.000000	1041.000000	1.041000e+03	1041.000000	1041.000000	1041.000000	1041.000000	1041.000000	1041.000000	1041.000000	1041.000000
mean	3.780019	2.071085	851.775216	1.380868e+06	-36.893405	174.799191	7.006316e+06	47.533141	28.976945	27.000000	24.118156	22.596542	29.322767	179.841499	5.066282
std	1.173679	0.992636	1580.893583	1.163590e+06	0.130233	0.119799	2.590188e+03	24.759752	21.069104	17.986533	10.975255	10.232515	21.889793	71.231238	2.907290
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000	1.000000
25%	3.000000	1.000000	323.000000	7.800000e+05	-36.950408	174.720076	7.004415e+06	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	2.000000
50%	4.000000	2.000000	571.000000	1.080000e+06	-36.893368	174.797892	7.006325e+06	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000	5.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.855774	174.880945	7.008382e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	207.000000	8.000000
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	789.000000	10.000000

Correlations and Patterns in data

Much of the data was positively skewed and very far from being normal, transformations were done on the attribute columns. Comparisons were made against the originals after altering the data to make the distribution more symmetrical and reduce the skew. Provided below are five plots which show the distribution of data that was used in the model showing the difference.





Below are two heatmaps (original and transformed) which plot the correlation between the variables, a dark purple colour strong positive relationship while a whiter colour show a strong negative relationship. We can see that the number of bedrooms, bathrooms and land area have the strongest positive relationship with the capital value of a property which is logical. The deprivation index showed a strong negative relationship since a lower number in that attribute would indicate a better socio-economic area.

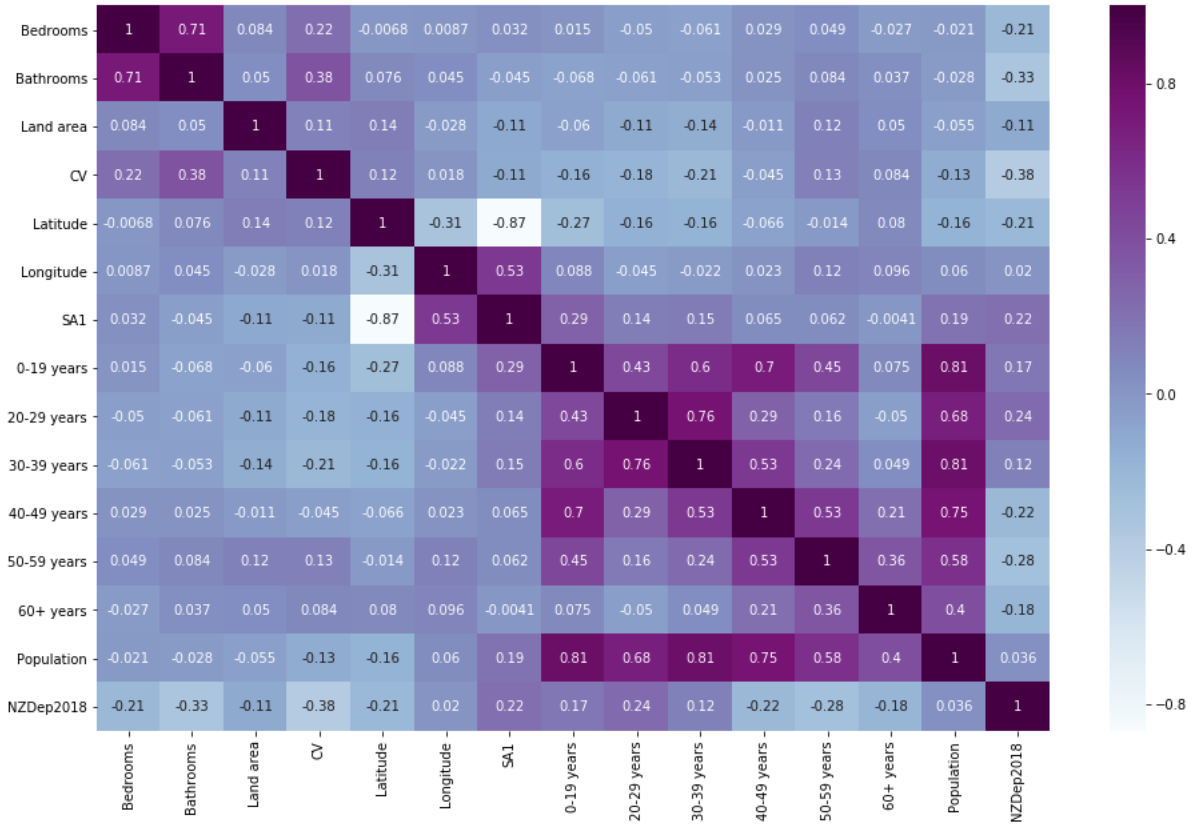


Figure 1: Original Variables

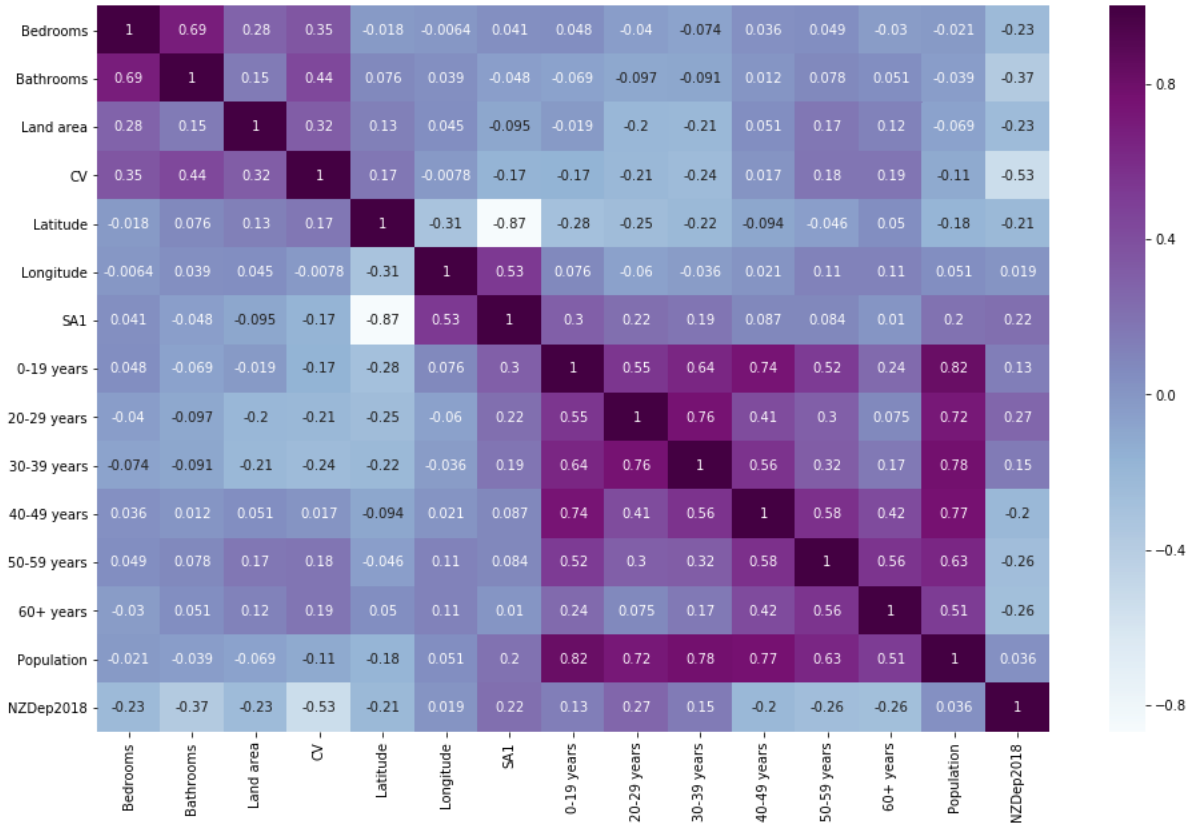


Figure 2: Transformed Variables

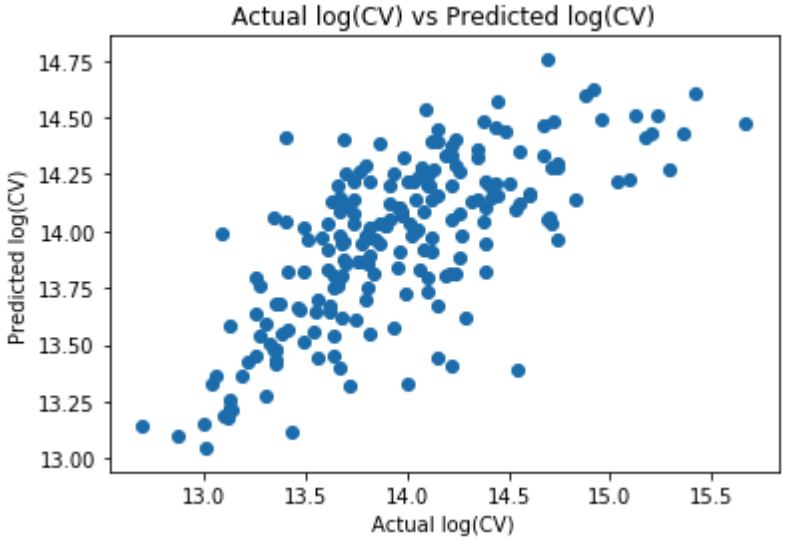
Model Analysis

A linear regression model was used to predict the outputs, it can only take in numeric data, so the address and suburb columns outright excluded. The visual plots of the heatmaps were used to show the relationship between a property's capital value (dependent variables) and how it was influenced by the other (independent) variables.

The final independent variables used in the model where the bedrooms, bathrooms, land area and deprivation index which all showed strong relationships in the heatmap. The land area which is precious commodity and becoming increasingly valuable in society and also the latitude where the model may have detected a relationship and possible trends in a properties positional data related to its capital value.

RMES	R^2 Score
0.37911	50.58%

The scatter plot below shows the models prediction of the properties logged capital value versus the actual logged capital value.



Conclusion

The exploration into developing a model to predict the house prices with the provided data shows that the model is an unreliable means for property valuation. The discrepancy between the predictions and actual values remain too large to trust the model as it can only account for 50% of the variation.

A suggestion would be to provide the model with a larger dataset to learn by expanding it on other regions. Or possibly more numeric variables that relate to a property's location or area demand as those two attributes have a large effect on a property's capital value.