Dataset Analysis of Auckland House Prices

Bridget Ong, July 2020

Executive Summary

The dataset being analysed contains 1050 records of Auckland house prices with each record having various variables that describe each house and the area, including the number of bathrooms and bedrooms, CV, land area and age groups. Using the variables in the base dataset, two extra columns were added to provide further insight and to determine whether correlations can be found with the deviation index and the population count in the area. The population count was sourced from Koordinates via an API call and the deprivation index was sourced from a dataset created by the University of Otago.

After exploring the data by calculating summary and descriptive statistics, and by creating visualisations of the correlation between each numerical variable, a few correlated variables were found. A linear regression model was tested using the training dataset to determine if the number of bedrooms can be accurately predicted based on other correlated variables.

Initial Data Analysis

The initial analysis of the data began with data cleaning. This involved checking for NaN values, duplicate rows and columns and any other unusual values. Firstly, rows with NaN values were found in the dataset, so they were subsequently dropped as our dataset is quite large and loss of a few rows is unlikely to have a substantial effect on the overall statistics. Secondly, duplicate rows were found and were dropped as this could negatively affect the summary statistics. Thirdly, redundant columns such as SA12018_code and the URPopnSA1_2018 were dropped as columns with the same value were already included in the dataset. Fourthly, the object type of land area was found to be a string which is unusual. Further investigation showed that some entries included a m^2. This was remedied by removing the m^2 and changing the land area column to a float object.

After data cleaning, summary and descriptive statistics were generated. Basic statistics were calculated for each numerical column, and the results were taken from 1040 observations which are shown here:

	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1
mean	3.782692	2.074038	850.772115	1.381557e+06	-36.894220	174.798720	7.006326e+06
std	1.171069	0.994353	1581.070983	1.163974e+06	0.128469	0.118222	2.583803e+03
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06
25%	3.000000	1.000000	323.000000	7.800000e+05	-36.950487	174.721131	7.004422e+06
50%	4.000000	2.000000	570.500000	1.080000e+06	-36.893455	174.797892	7.006334e+06
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.856094	174.880943	7.008383e+06
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06

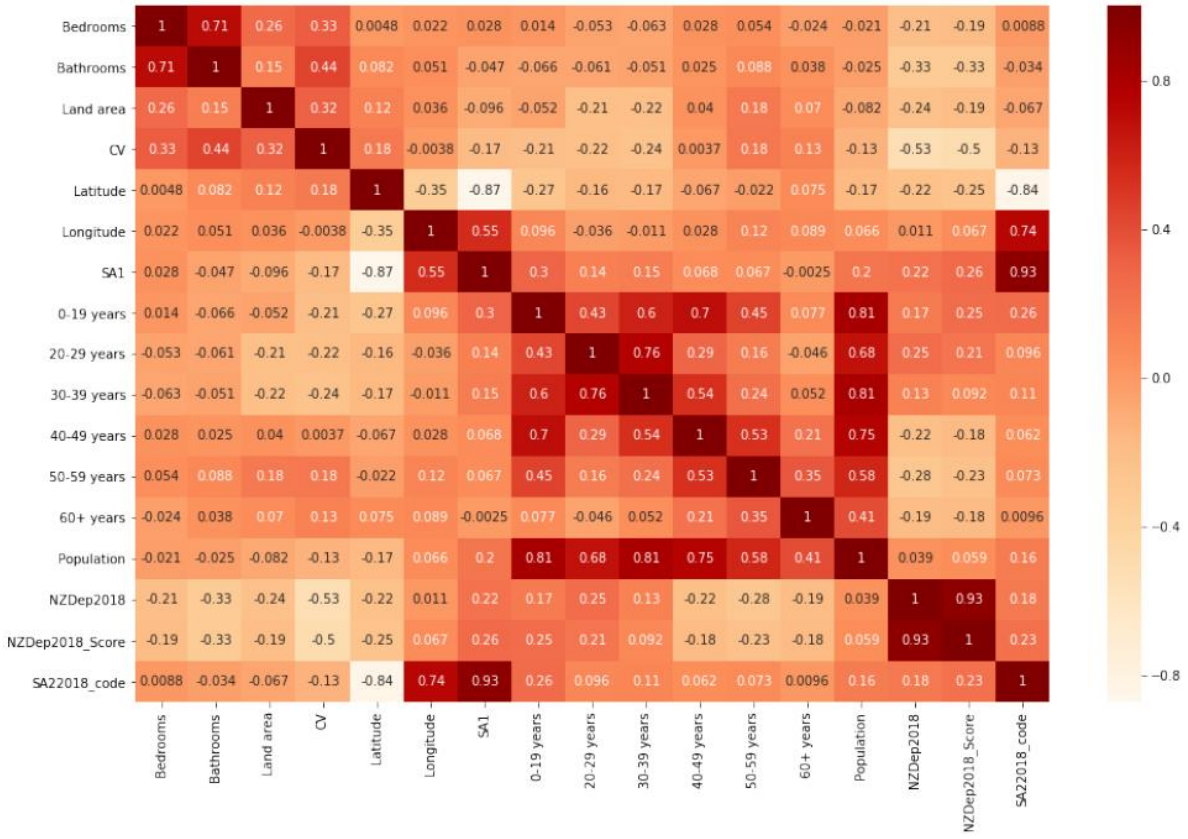
	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Population
mean	47.538462	28.952885	26.982692	24.124038	22.580769	29.313462	179.780769

	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Population
std	24.760576	21.038594	17.955181	10.978893	10.224770	21.878873	71.227962
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
25%	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000
50%	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000
75%	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	207.000000
max	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	789.000000

	NZDep2018	NZDep2018_Score	SA22018_code
mean	5.066346	986.227885	141556.346154
std	2.904714	93.536676	14627.900554
min	1.000000	849.000000	110400.000000
25%	2.000000	918.000000	132275.000000
50%	5.000000	959.000000	141850.000000
75%	8.000000	1030.250000	152575.000000
max	10.000000	1380.000000	170500.000000

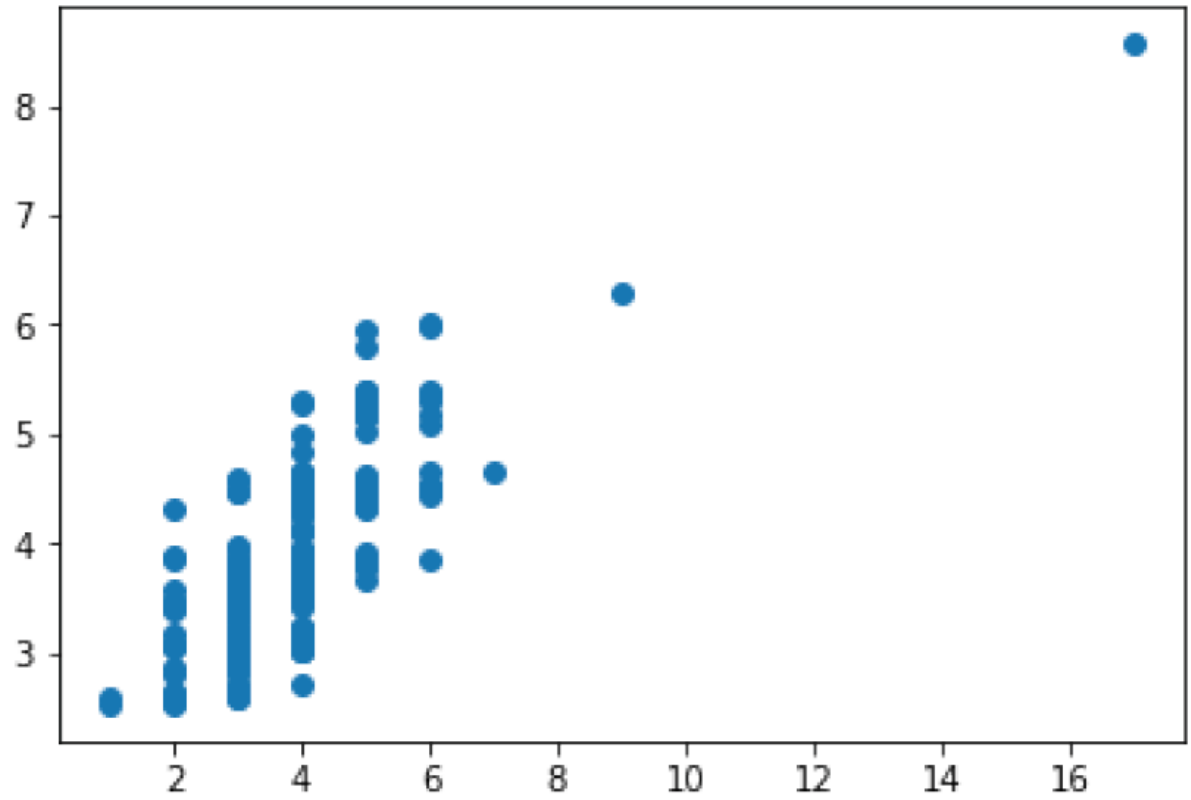
Analysis of correlations and patterns in the data

The correlation between the numerical columns was calculated and observed in the below correlation plot. (The colour bar indicates the correlation values. Dark red means correlation value is 1 and white means correlation value is negative 1).



The graph shows that bedrooms and bathrooms have a relatively strong positive correlation which makes sense logically (have enough bathrooms for the number of people living in the house). The CV of a house appears to have a moderate negative correlation with the deprivation index of an area, suggesting that houses in areas with a lower deprivation index tend to have a higher CV and vice versa. Population has a relatively strong positive correlations with some of the age groups, suggesting that the distribution of ages is similar in different areas.

Building a model



A linear regression model was chosen to predict the number of bedrooms in a house. Referring to the correlation matrix, it is noticed that age, population or location had no correlation with the number of bedrooms in a house, so those attributes were not included in the model. Above, we can see that the plot is relatively scattered, indicating that the model is not very accurate. This is likely because the attributes used in the model did not have a strong correlation with bedrooms with the sole exception being bathrooms. This resulted in a model that struggles to predict the number of bedrooms with the given attributes accurately.

Conclusions

This analysis has shown that the number of bedrooms in a house can not be confidently predicted from the variables of bathrooms, land area and CV. The linear regression model has an accuracy rate of 59%.