MSA Phase 1 Assignment

By Alfa Lee

Executive Summary

The dataset is a collection of various house prices in Auckland, with other notable attributes. The aim of this report is to examine how some of these attributes affect the house price (capital value).

First, I added two attributes: 2018 census population, and the NZ deprivation index, to my dataset by calling an API and subsequently merging two datasets, respectively. I then began exploring the data by calculating summary and descriptive statistics, and also visualising the data in a way to highlight correlations between variables. Afterwards, I applied a machine learning model using linear regression to the dataset to try and predict the house price as best as I can.

Initial Data Analysis

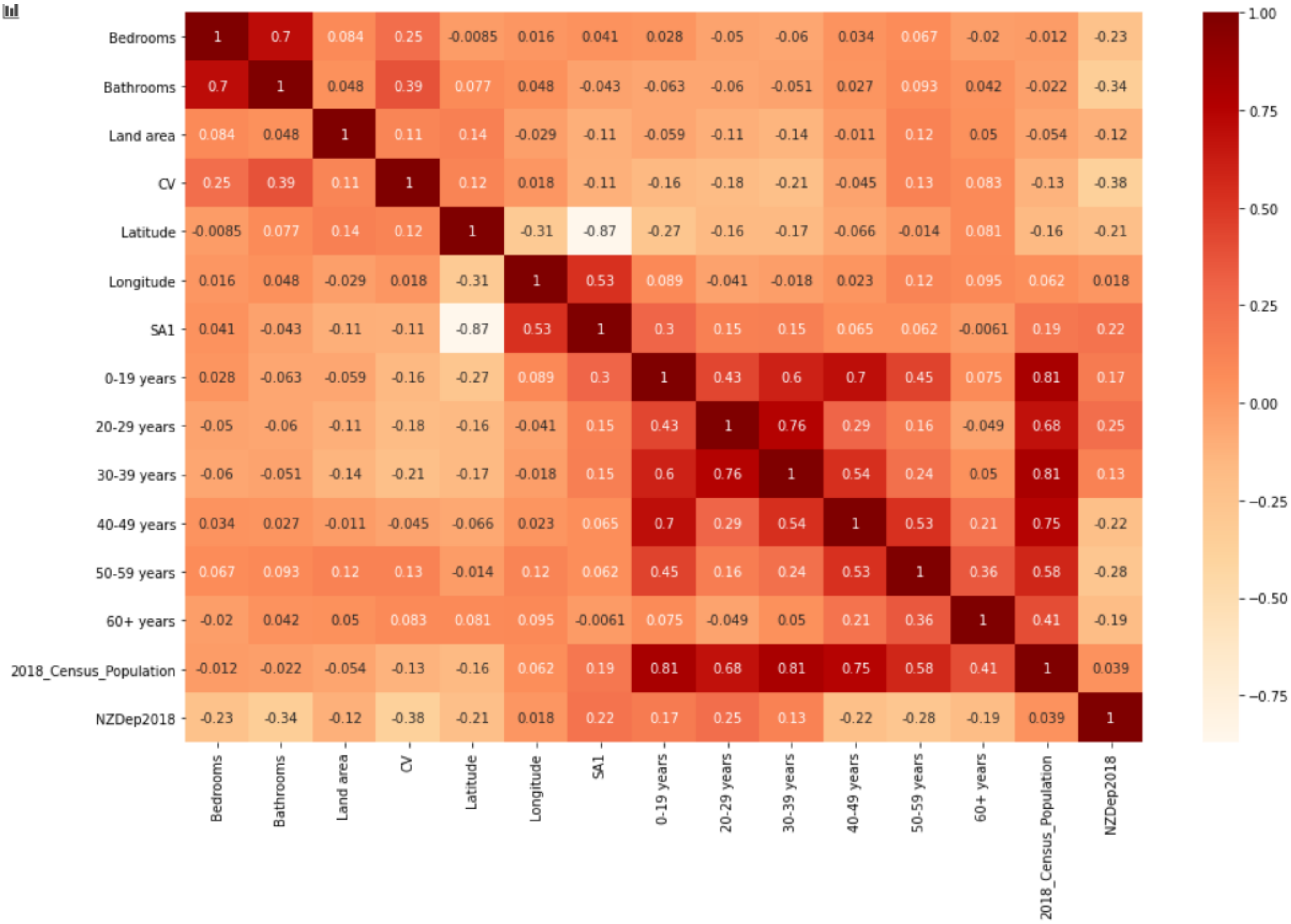
I cleaned the dataset by altering certain datatypes of the columns, removed / updated NaN values, and removed outliers.

Here are the summary and descriptive statistics of the numeric columns of the cleaned dataset with the two added columns. I have left out numeric columns: “Latitude”, “Longitude” and “SA1”, as they are irrelevant. These results were taken from 1048 observations, or rows.

Feature	Bedrooms	Bathrooms	Land area	CV	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	2018_Census_Population	NZDep2018
mean	3.7643129	2.0677480	85.343511	1388524.905	47.55057252	28.9007634	26.9885496	24.1288167	22.6259542	29.398855	179.8425573	5.06870229
std	1.0973552	0.9764123	158.99556	1184432.772	24.70794681	21.0037446	17.9411609	10.9580083	10.2163499	21.823974	71.05928637	2.91441980
min	1	1	4	270000	0	0	0	0	0	0	3	1
25%	3	1	32	780000	33	15	15	18	15	18	138	2
50%	4	2	57	1080000	45	24	24	24	21	27	174	5
75%	4	3	82	1600000	57	36	33	30	27	36	207.75	8
max	9	7	2224	18000000	201	270	177	114	90	483	789	10

Correlation and Relationships

The correlation between the numeric columns were calculated and are displayed in the correlation plot below. The right colour bar indicates the correlation values: dark red means correlation coefficient is 1, and white means correlation coefficient is -1. I have only included the correlation plot, and left out the pairplot, as the latter is far harder



	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	2018_Census_Population	NZDep2018
Bedrooms	1	0.7	0.084	0.25	-0.0085	0.016	0.041	0.028	-0.05	-0.06	0.034	0.067	-0.02	-0.012	-0.23
Bathrooms	0.7	1	0.048	0.39	0.077	0.048	-0.043	-0.063	-0.06	-0.051	0.027	0.093	0.042	-0.022	-0.34
Land area	0.084	0.048	1	0.11	0.14	-0.029	-0.11	-0.059	-0.11	-0.14	-0.011	0.12	0.05	-0.054	-0.12
CV	0.25	0.39	0.11	1	0.12	0.018	-0.11	-0.16	-0.18	-0.21	-0.045	0.13	0.083	-0.13	-0.38
Latitude	-0.0085	0.077	0.14	0.12	1	-0.31	-0.87	-0.27	-0.16	-0.17	-0.066	-0.014	0.081	-0.16	-0.21
Longitude	0.016	0.048	-0.029	0.018	-0.31	1	0.53	0.089	-0.041	-0.018	0.023	0.12	0.095	0.062	0.018
SA1	0.041	-0.043	-0.11	-0.11	-0.87	0.53	1	0.3	0.15	0.15	0.065	0.062	-0.0061	0.19	0.22
0-19 years	0.028	-0.063	-0.059	-0.16	-0.27	0.089	0.3	1	0.43	0.6	0.7	0.45	0.075	0.81	0.17
20-29 years	-0.05	-0.06	-0.11	-0.18	-0.16	-0.041	0.15	0.43	1	0.76	0.29	0.16	-0.049	0.68	0.25
30-39 years	-0.06	-0.051	-0.14	-0.21	-0.17	-0.018	0.15	0.6	0.76	1	0.54	0.24	0.05	0.81	0.13
40-49 years	0.034	0.027	-0.011	-0.045	-0.066	0.023	0.065	0.7	0.29	0.54	1	0.53	0.21	0.75	-0.22
50-59 years	0.067	0.093	0.12	0.13	-0.014	0.12	0.062	0.45	0.16	0.24	0.53	1	0.36	0.58	-0.28
60+ years	-0.02	0.042	0.05	0.083	0.081	0.095	-0.0061	0.075	-0.049	0.05	0.21	0.36	1	0.41	-0.19
2018_Census_Population	-0.012	-0.022	-0.054	-0.13	-0.16	0.062	0.19	0.81	0.68	0.81	0.75	0.58	0.41	1	0.039
NZDep2018	-0.23	-0.34	-0.12	-0.38	-0.21	0.018	0.22	0.17	0.25	0.13	-0.22	-0.28	-0.19	0.039	1

Viewing this correlation plot, we can see there is moderate positive linear correlation between Capital Value (CV) and the number of bathrooms in the house (c = 0.39), as well as a weak positive linear relationship between CV and number of bedrooms (c = 0.25).

The age of the people living in the house have a very small negative linear correlation to the CV of the house from ages 0-49, and have a very small positive linear correlation to the CV of the house from ages 50+. However, small correlation coefficients should be expected in this case as usually the age of the occupants, indeed also the census population (c = -0.13) do not have a large effect on house pricing.

Land area has a surprisingly weak positive linear correlation to the CV of the house (c = 0.11). Typically, one would expect land area to have a stronger linear correlation to the price of the house.

Notably, it seems that there is a moderate negative linear correlation between the house price and the deprivation index of the area (c = -0.38), suggesting that houses located in more socioeconomically deprived areas will be relatively cheaper.

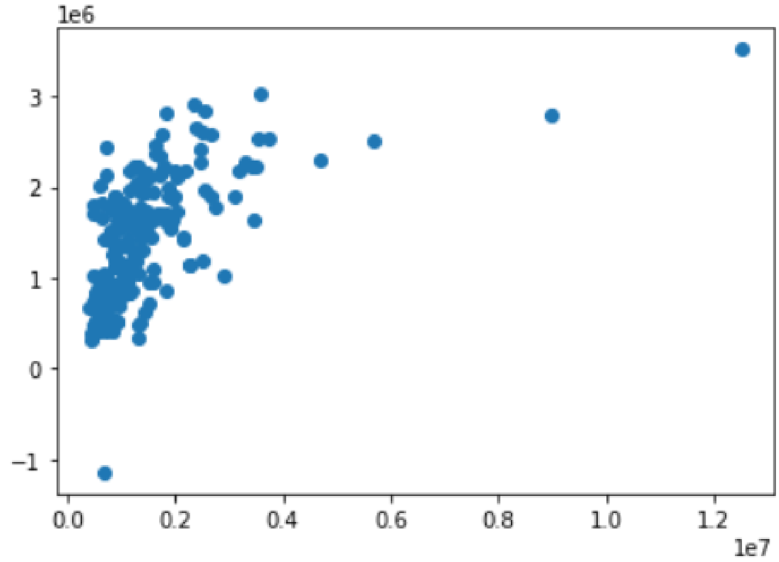
Machine Learning Model

In this analysis, I used linear regression in my machine learning model.

I split my dataset in two, using 20% for the training set, and 80% for the testing set.

In my machine learning model, I decided to drop the non-numeric, values of “Latitude”, “Longitude” and “SA1” because they do not affect the house price at all. I have also decided to leave out the categorical variable of “Suburbs”.

Below is the scatterplot, plotting the machine learning model’s predicted house prices against the actual house prices from the dataset.



My R^2 value (coefficient of determination) came out to be:

```
0.34273499686793674
```

This score could be improved possibly by including the categorical variable of “Suburbs” into the model, by representing each suburb with a set of numerical indicator variable columns, e.g. by utilising pandas “get_dummies” function.

Conclusion

The multiple handy python libraries such as scikit-learn, matplotlib, pandas and seaborn allowed me to conveniently and effectively process the housing dataset and to visualise it, so that correlations and relationships between the house price and other factors could be clearly identified and displayed.

From this dataset, the number of bedrooms and bathrooms have a weak to moderate positive correlation to the house price, while interestingly the land area has a very weak positive correlation to the house price. The dataset also shows that there is a moderate negative correlation between the house price and the deprivation score of that area.