JULY 2020

ANALYSIS OF HOUSE PRICING IN AUCKLAND

DATA SCIENCE PATHWAY - MSA

Prepared by: Chloe Haigh

EXECUTIVE SUMMARY

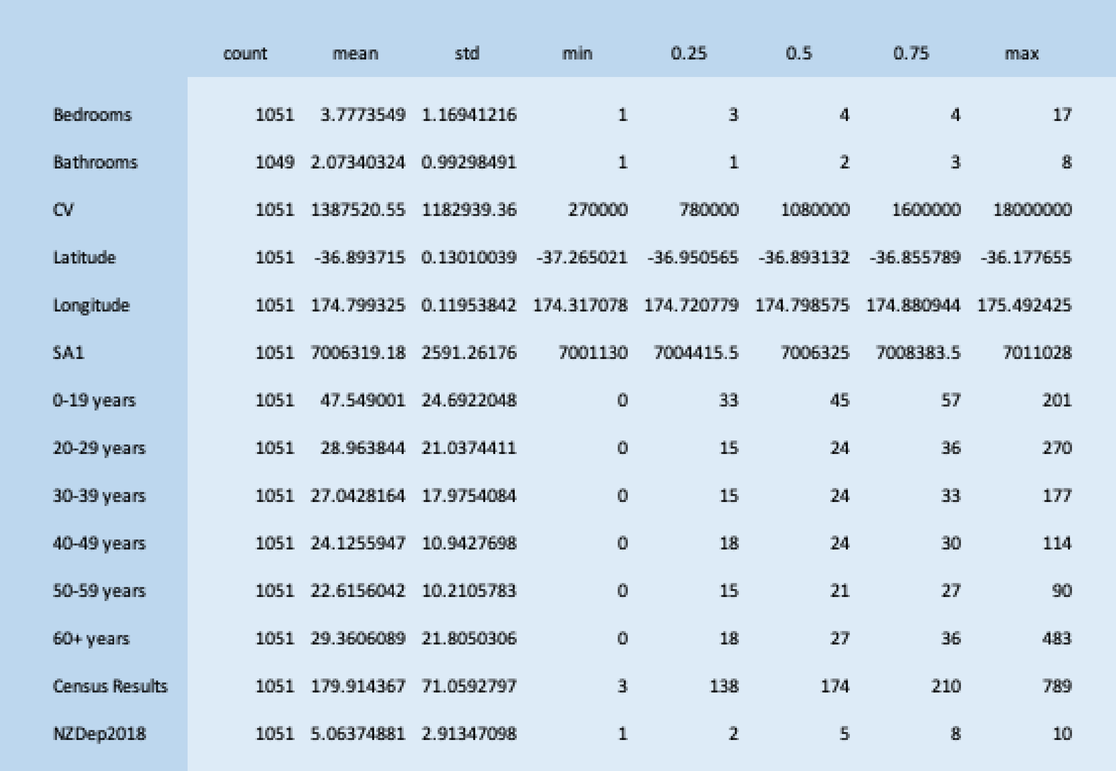
The dataset used in this analysis is the Auckland house price dataset which was provided in the git repo. It contains assorted property information for different addresses across Auckland suburbs, as well as information about the inhabitants and the location.

This analysis was based on 1051 properties across Auckland, with each property having 16 attributes, which includes housing specific information such as the number of bedrooms and bathrooms, the address and suburb, land area, capital value, the latitude and longitude of the house and the statistical area (SA1). The deprivation index (the decile score, out of 10) and census population from 2018 were attributes manually added into the original dataset. Other attributes were related to inhabitant information (meaning the number of people from different age groups who are living in the SA1 unit taken from the 2018 census).

The data was cleaned, converted to appropriate formats and then analysed. The correlation and pairwise relationships between the attributes were explored, and the summary statistics were also calculated. Finally, a linear regression model was developed from the data set which tried to predict the capital value of a property based on the attributes. The accuracy of this model is quite low and more work is required to improve the results.

DATA

The dataset required some cleaning as it contained some NaN values in the Bathrooms and Suburbs columns. The NaN values in the Bathroom column were filled with the median value for bathrooms, and for the Suburbs column, a K-Nearest Neighbours Classification model was built which predicted the relevant suburb based on the SA1, Latitude and Longitude of the house. The Land area column also was initially as the 'string' type, which was then converted to an 'int' type. After the dataset was cleaned, the statistics were collected and are displayed below.



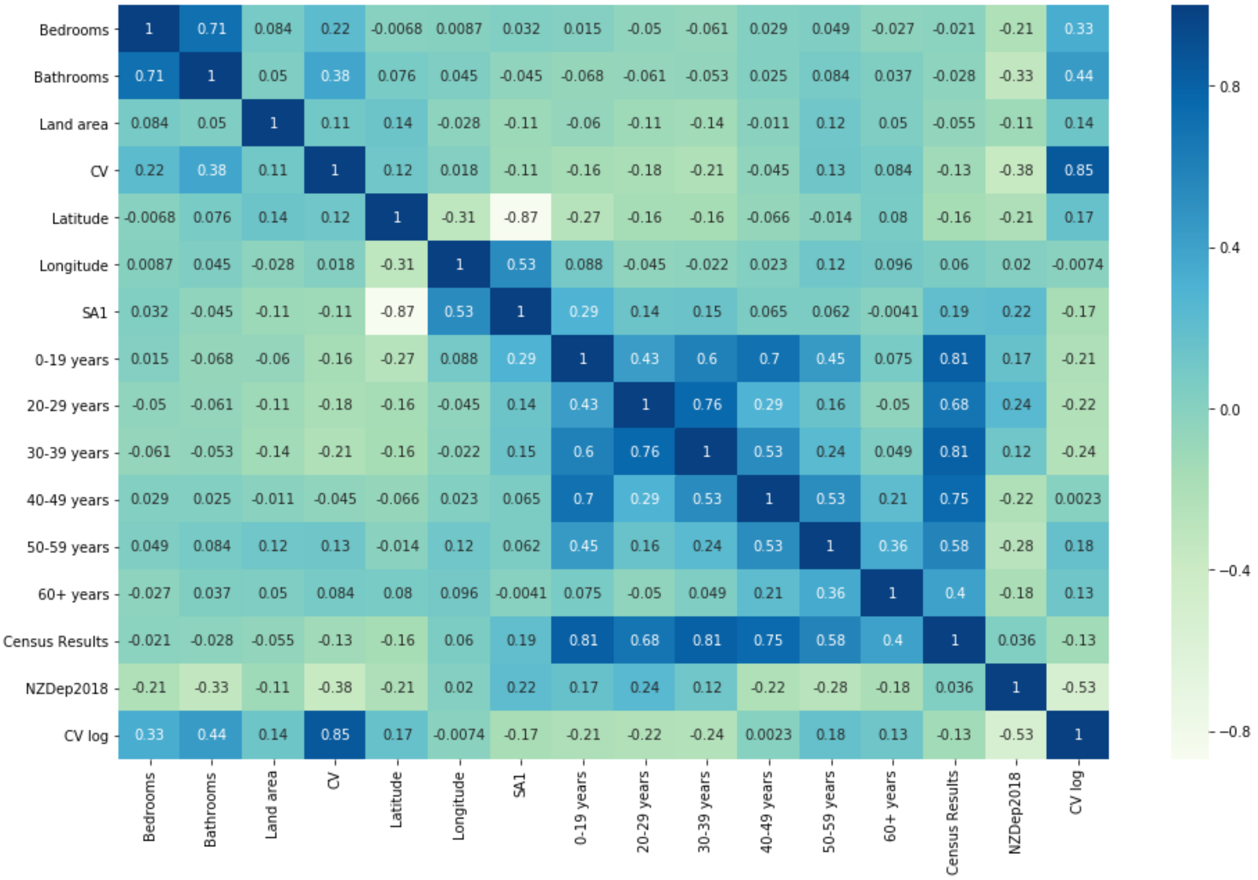
	count	mean	std	min	0.25	0.5	0.75	max
Bedrooms	1051	3.7773549	1.16941216	1	3	4	4	17
Bathrooms	1049	2.07340324	0.99298491	1	1	2	3	8
CV	1051	1387520.55	1182939.36	270000	780000	1080000	1600000	18000000
Latitude	1051	-36.893715	0.13010039	-37.265021	-36.950565	-36.893132	-36.855789	-36.177655
Longitude	1051	174.799325	0.11953842	174.317078	174.720779	174.798575	174.880944	175.492425
SA1	1051	7006319.18	2591.26176	7001130	7004415.5	7006325	7008383.5	7011028
0-19 years	1051	47.549001	24.6922048	0	33	45	57	201
20-29 years	1051	28.963844	21.0374411	0	15	24	36	270
30-39 years	1051	27.0428164	17.9754084	0	15	24	33	177
40-49 years	1051	24.1255947	10.9427698	0	18	24	30	114
50-59 years	1051	22.6156042	10.2105783	0	15	21	27	90
60+ years	1051	29.3606089	21.8050306	0	18	27	36	483
Census Results	1051	179.914367	71.0592797	3	138	174	210	789
NZDep2018	1051	5.06374881	2.91347098	1	2	5	8	10

As the capital value is a price, it is likely to be skewed if left unchanged so it was converted to the log value as another attribute.

04

CORRELATION MATRIX

The correlation between the different attributes was calculated and is displayed in the heat map below. The scale on the right shows the strength of the correlation between the attributes, with dark blue being the most correlated and light green being the least correlated.



	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	Census Results	NZDep2018	CV log
Bedrooms	1	0.71	0.084	0.22	-0.0068	0.0087	0.032	0.015	-0.05	-0.061	0.029	0.049	-0.027	-0.021	-0.21	0.33
Bathrooms	0.71	1	0.05	0.38	0.076	0.045	-0.045	-0.068	-0.061	-0.053	0.025	0.084	0.037	-0.028	-0.33	0.44
Land area	0.084	0.05	1	0.11	0.14	-0.028	-0.11	-0.06	-0.11	-0.14	-0.011	0.12	0.05	-0.055	-0.11	0.14
CV	0.22	0.38	0.11	1	0.12	0.018	-0.11	-0.16	-0.18	-0.21	-0.045	0.13	0.084	-0.13	-0.38	0.85
Latitude	-0.0068	0.076	0.14	0.12	1	-0.31	-0.87	-0.27	-0.16	-0.16	-0.066	-0.014	0.08	-0.16	-0.21	0.17
Longitude	0.0087	0.045	-0.028	0.018	-0.31	1	0.53	0.088	-0.045	-0.022	0.023	0.12	0.096	0.06	0.02	-0.0074
SA1	0.032	-0.045	-0.11	-0.11	-0.87	0.53	1	0.29	0.14	0.15	0.065	0.062	-0.0041	0.19	0.22	-0.17
0-19 years	0.015	-0.068	-0.06	-0.16	-0.27	0.088	0.29	1	0.43	0.6	0.7	0.45	0.075	0.81	0.17	-0.21
20-29 years	-0.05	-0.061	-0.11	-0.18	-0.16	-0.045	0.14	0.43	1	0.76	0.29	0.16	-0.05	0.68	0.24	-0.22
30-39 years	-0.061	-0.053	-0.14	-0.21	-0.16	-0.022	0.15	0.6	0.76	1	0.53	0.24	0.049	0.81	0.12	-0.24
40-49 years	0.029	0.025	-0.011	-0.045	-0.066	0.023	0.065	0.7	0.29	0.53	1	0.53	0.21	0.75	-0.22	0.0023
50-59 years	0.049	0.084	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	0.18
60+ years	-0.027	0.037	0.05	0.084	0.08	0.096	-0.0041	0.075	-0.05	0.049	0.21	0.36	1	0.4	-0.18	0.13
Census Results	-0.021	-0.028	-0.055	-0.13	-0.16	0.06	0.19	0.81	0.68	0.81	0.75	0.58	0.4	1	0.036	-0.13
NZDep2018	-0.21	-0.33	-0.11	-0.38	-0.21	0.02	0.22	0.17	0.24	0.12	-0.22	-0.28	-0.18	0.036	1	-0.53
CV log	0.33	0.44	0.14	0.85	0.17	-0.0074	-0.17	-0.21	-0.22	-0.24	0.0023	0.18	0.13	-0.13	-0.53	1

We can see that the bedrooms and bathrooms had a high correlation, and so did the longitude and the SA1. The different age groups were also shown to be correlated with each other and with the census results.

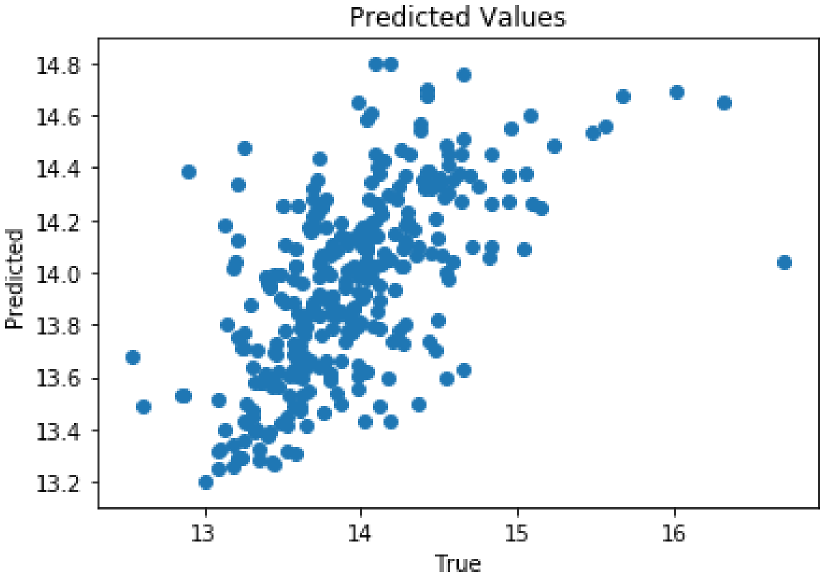
The SA1 and latitude were shown to have a low correlation, as well as the log value of the CV of the house and the deprivation index. The log CV and the bedrooms and bathrooms had a medium correlation.

05

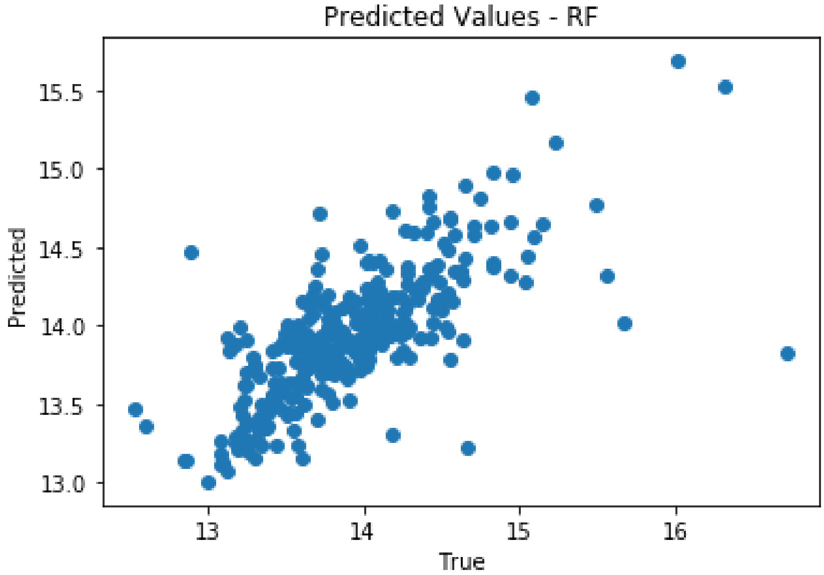
MODELLING

A base linear regression model was created from the SKLearn linear_model module in Python. The model took the CV log attribute as a target and the other attributes (15, excluding the Address and Suburb attributes) as the input features. The model was trained on 70% of the 1051 properties, with the other 30% held for testing.

After the model was fitted to the training data, it then used the test set to predict the log current value of the properties. The accuracy of this model was 0.3717, and a scatterplot of the predicted vs. true values is shown below.



As the accuracy for this model is quite low, I decided to see if it would improve by dropping the features with low correlation. As the NZDep2018 attribute had a low correlation with CV log (the target variable), I first removed this feature. The accuracy worsened, becoming 0.2913.



I then tried a different type of model, the Random Forest (RF) Regression model from the SKLearn ensemble module. This model had the same target and input features as the base model of Linear Regression, and returned a model with an accuracy score of 0.5592. The scatter plot of predicted values are shown above.

Similarly, despite the low correlation between the deprivation index (NZDep2018) and the CV log attributes, removing the deprivation attribute in the RF model also caused the accuracy to worsen, this time to 0.5231.

CONCLUSIONS

In this analysis we cleaned and then examined the dataset and looked at the statistics. We also looked at the attributes and then finally tried to use the dataset with a machine learning model to predict the current value of a property in Auckland based on the attributes.

In conclusion, we can say that it may be possible to use machine learning to predict the current value of house prices in Auckland based on their attributes but more work is required to improve the models and return stronger results that can be relied upon. The random forest model performed better than the linear regression model in terms of accuracy.