Executive Summary

MSA Phase 1

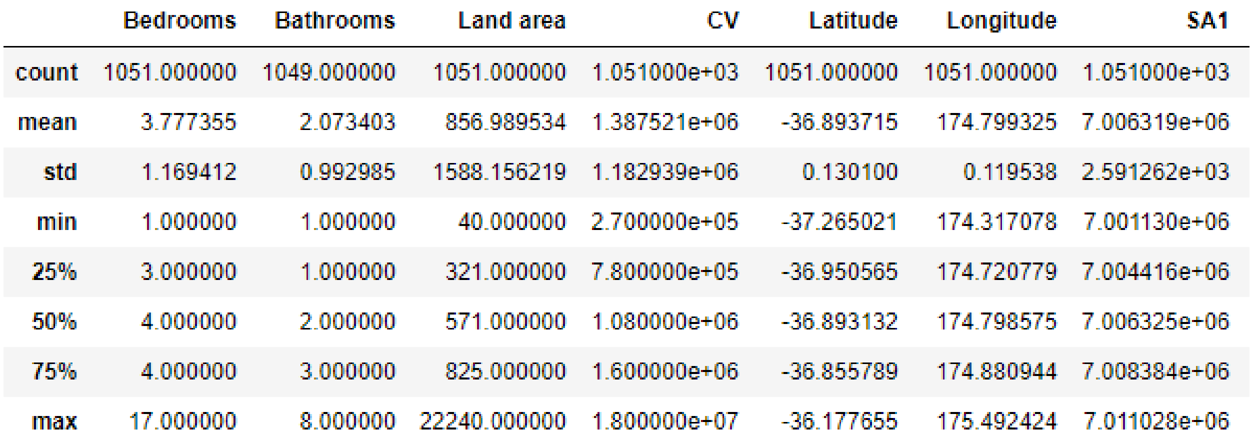
Data Science

David Hoang

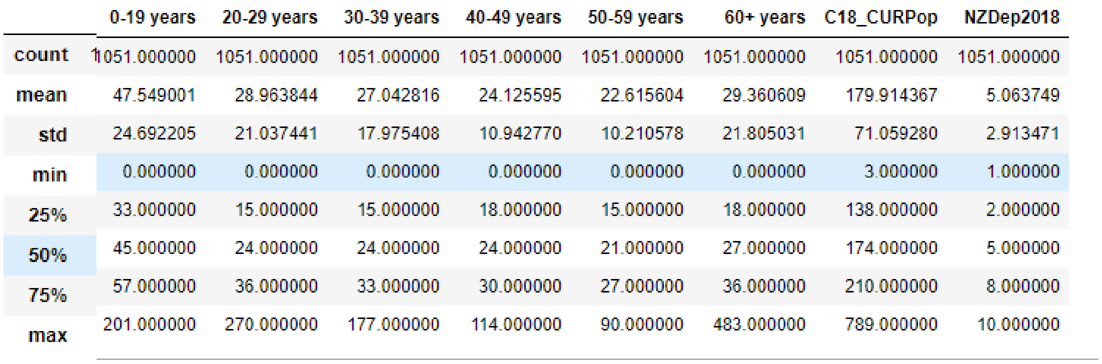
The dataset being used that was given to us contains data about properties, monetary and physical values like physical address, capital value , size of land area, number of bathrooms and so on; but also, contains data about the SA1 area that it resides in and number of people at certain age groups in that SA1 area based of 2018 census. We also extract deprivation index from an Otago University file to further help with this project.

Initial data analysis

Here are some statistics seeing the data we are given, you can see that bathroom has 2 less observation counts, that is because there are missing values in that column. Land area had to be converted from a string to a float to be shown on these tables.



	Bedrooms	Bathrooms	Land area	CV	Latitude	Longitude	SA1
count	1051.000000	1049.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1.051000e+03
mean	3.777355	2.073403	856.989534	1.387521e+06	-36.893715	174.799325	7.006319e+06
std	1.169412	0.992985	1588.156219	1.182939e+06	0.130100	0.119538	2.591262e+03
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06
25%	3.000000	1.000000	321.000000	7.800000e+05	-36.950565	174.720779	7.004416e+06
50%	4.000000	2.000000	571.000000	1.080000e+06	-36.893132	174.798575	7.006325e+06
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.855789	174.880944	7.008384e+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	C18_CURPop	NZDep2018
count	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000
mean	47.549001	28.963844	27.042816	24.125595	22.615604	29.360609	179.914367	5.063749
std	24.692205	21.037441	17.975408	10.942770	10.210578	21.805031	71.059280	2.913471
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000	1.000000
25%	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	2.000000
50%	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000	5.000000
75%	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	210.000000	8.000000
max	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	789.000000	10.000000

Cleaning/Preparing Data for analysis

In the cleaning phase, any missing or NAn values etc. had to be dealt with. 3 were found, one was a missing value in suburbs column which I decided to just drop that data row because I do not know how to obtain the address with only what was given. The other two was number of bathroom values, I replaced the missing values with the median of all the values in the number of bathroom column. Median was a better choice than mean.

For the machine learning prediction model, linear regression would be the most suitable model. I logged the CV values since the data was right skewed, logging it made it more suitable for linear regression as its more uniformly distributed. Columns that seems to not improve or was relevant to the predictions was dropped, such as longitude, SA1 code etc.

Using the .score() function gives the coefficient of R^ 2 of the prediction which was 39%~ which is a low score of how well the data fit the model. However, with RSME the value is low but maybe not low enough, 20% which could indicate that the model could be a good model.

Conclusion

To conclude, the data didn't seem to fit well with our prediction model to predict CV values of houses. The RSME value from our model doesn't seem to be low enough to say with more confidence that the model built is adequate enough.