MSA 2020 Phase 1 Data Science Pathway Report

Analysis of House Pricing Dataset

Siddharth Jha, July 2020

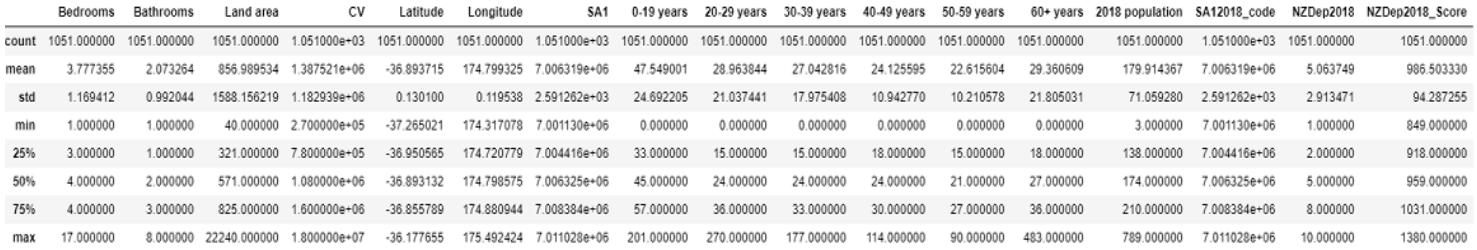
Executive Summary

The dataset being analysed is a house pricing dataset. It is based off the 2018 Census and contains 15 attributes: number of bedrooms and bathrooms of property(2), physical address of property(1), land area of property in meters squared(1), capital value of property(1), coordinates of property(2), SA1, an area unit classification(1), and the number of people of different age groups living in the SA1 unit area(6), and the name of the suburb where property is located. The analysis is based on 1051 observations for each of the 15 variables.

Firstly, an additional column corresponding to the population in 2018 is added, followed by 2 columns corresponding to the 2018 Deprivation Index. After a clean-up of the data by using logistic regression to impute missing values, some initial analysis was performed.

Initial analysis

Using the describe function on the cleaned dataset, the following statistics were revealed:



	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 population	SA12018_code	NZDep2018	NZDep2018_Score
count	1051.000000	1051.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1.051000e+03	1051.000000	1051.000000
mean	3.777355	2.073264	856.989534	1.387521e+06	-36.893715	174.799325	7.006319e+06	47.549001	28.963844	27.042816	24.125595	22.615604	29.360609	179.914367	7.006319e+06	5.063749	986.503330
std	1.169412	0.992044	1588.156219	1.182939e+06	0.130100	0.119538	2.591262e+03	24.692205	21.037441	17.975408	10.942770	10.210578	21.805031	71.059280	2.591262e+03	2.913471	94.287255
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	7.001130e+06	1.000000	849.000000
25%	3.000000	1.000000	321.000000	7.800000e+05	-36.950565	174.720779	7.004416e+06	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	7.004416e+06	2.000000	918.000000
50%	4.000000	2.000000	571.000000	1.080000e+06	-36.893132	174.798575	7.006325e+06	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	174.000000	7.006325e+06	5.000000	959.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.855789	174.880944	7.008384e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	210.000000	7.008384e+06	8.000000	1031.000000
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	7.011028e+06	10.000000	1380.000000

From this we can see the disparity between the maximum and minimum rooms, value, area, etc. for the houses. The numerical data was also visualised using Seaborn, using the Suburbs as a category.

[Unfortunately I have not been able to progress beyond this point due to shortage of time.]