Analysis of New Zealand house pricing in the Statistical Area 1 (SA1) Region

Ka Ming Fei (Ming), July 2020

Executive Summary

This report explores the factors that affect the New Zealand house pricing in the Statistical Area 1 (SA1) Region. The analysis carried out involved statistical, correlation and relationship investigation. Prediction models are still under investigation due to low scores.

The original dataset contains the house prices in the Statistical Area 1 Region given in the 'MSA Phase 1 Data Pathway Course'. The data includes estimated values of homes, location, and the size of the different age groups in their respective SA1 unit area. To better analyse the data, integration with the '2018 SA1 Census Population' from Stats.govt.nz, and '2018 SA1 Deprivation Index' from the University of Otago, was carried out.

From generating visualisations, cost's relationships between population, location, bedroom/bathroom quantity, and depravation index was further investigated.

Initial Data Analysis

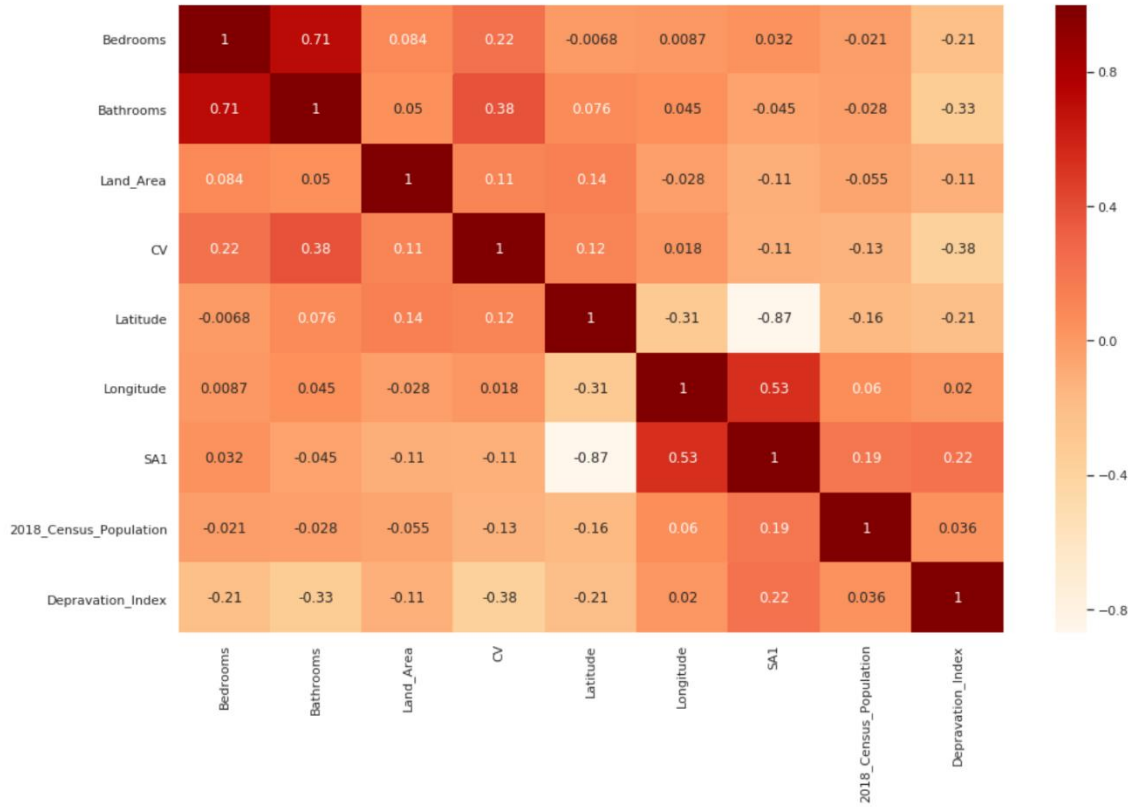
The statistical information was the first information investigated.

	Bedrooms	Bathrooms	Land_Area	CV	Latitude	Longitude	SA1	2018_Census_Population	Depravation_Index
count	1051.000000	1051.000000	1051.000000	1.051000e+03	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	179.914367	5.063749
std	1.169412	0.992044	1588.156219	1.182939e+06	0.130100	0.119538	2.591262e+03	71.059280	2.913471
min	1.000000	1.000000	40.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	3.000000	1.000000
25%	3.000000	1.000000	321.000000	7.800000e+05	-36.950565	174.720779	7.004416e+06	138.000000	2.000000
50%	4.000000	2.000000	571.000000	1.080000e+06	-36.893132	174.798575	7.006325e+06	174.000000	5.000000
75%	4.000000	3.000000	825.000000	1.600000e+06	-36.855789	174.880944	7.008384e+06	210.000000	8.000000
max	17.000000	8.000000	22240.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	789.000000	10.000000

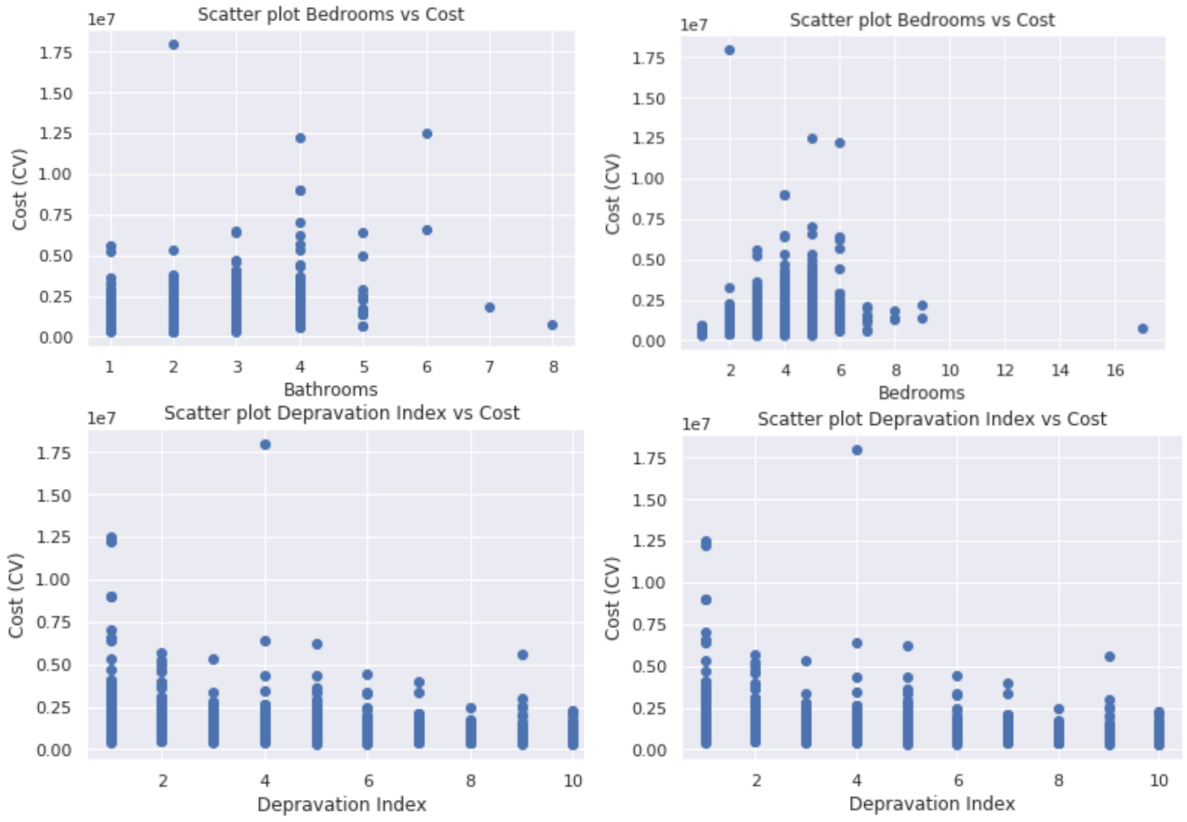
Analysis of Correlations and Patterns

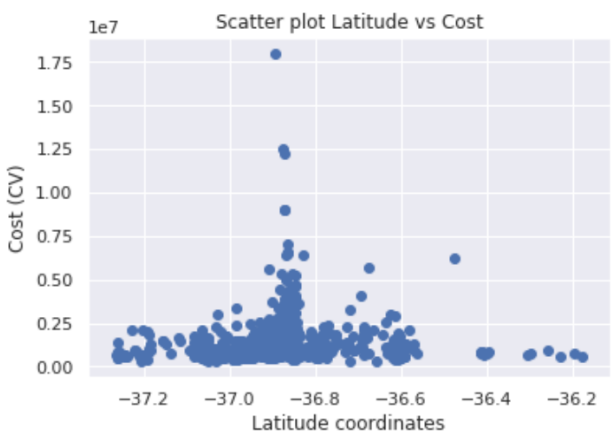
From the visualisations below, noticeable correlations with bathrooms, bedrooms, depravation index, census population and location (geographic co-ordinates) was investigated. The Positive relationships include both bathroom and bedroom quantity. As for Negative relationships, the Depravation Index and Census Population increase resulted on price Decrease. Due the negative relationships, the location was investigated, showing what seemed to be 'centralised' resulted in higher pricing (positive relation)

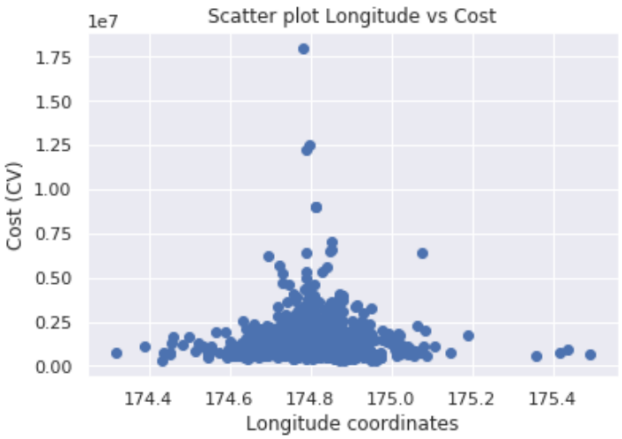
Below is the heat map generated to quickly identify correlations. The heat map utilises dark red and positive number for strong positive correlations, and white/negative numbers for strong negative correlations.



Further analysis of relationships.







Model

To predict the cost of houses with the data provided, a model implementation was attempted. However, the score was too low to be effective and tuning is required. In future tests the following will be adjusted as tuning:

```
predicted = model.predict(test_x)
model.score(test_x, test_y)
```

```
0.04265402843601896
```

```
confusion_matrix(test_y, predicted)
```

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 1, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 1, ..., 0, 0, 0]])
```

- Reducing pool of answers by rounding and grouping costs (e.g. by price range, by suburb)
- Adjust input data

Conclusions

This analysis has shown possibility of price prediction via machine learning. Although the model is inaccurate, the data available along with the visualisations show a promising potential provided appropriate tuning is carried out.