Carlo Carbonilla

Regression analysis of House Prices

Executive Summary

The dataset is of houses in Auckland which was downloaded from the MSA GitHub. It contains the characteristics of each house and the characteristics based on the statistical area the house located in.

The analysis is based on 1051 observations for each of the 17 variables. The independent variable we're testing for is the house's capital value, it's evaluated price.

The dependent variables are the number of bedrooms and bathrooms of the house, land area in m^2, and its latitude and longitude. The rest of the variables are based on the statistical area (SA1) the house is located in, which includes it's Deprivation Index (a decile rating from Decile 1 representing the least deprived areas to Decile 10, being the most deprived areas), the population of the area and the population of age groups 0-19, 20-29, 30-39, 40-49, 50-59, and 60+ years old in the area.

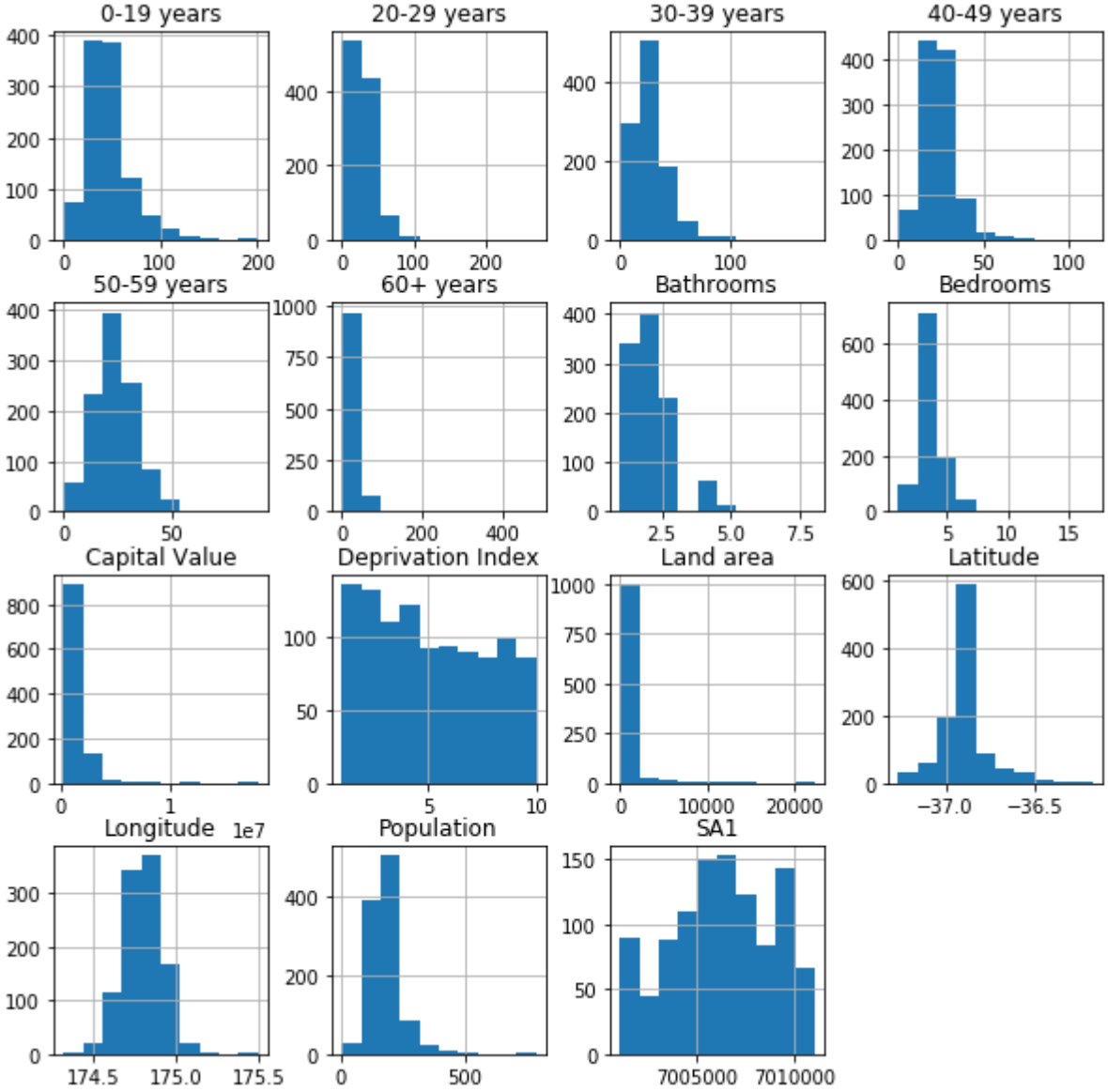
After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the correlation between each numerical variable, a few moderately correlated variables are found. Three regression algorithms were tested to train this dataset and the R^2 value was used to select the best model.

Initial Data Exploration

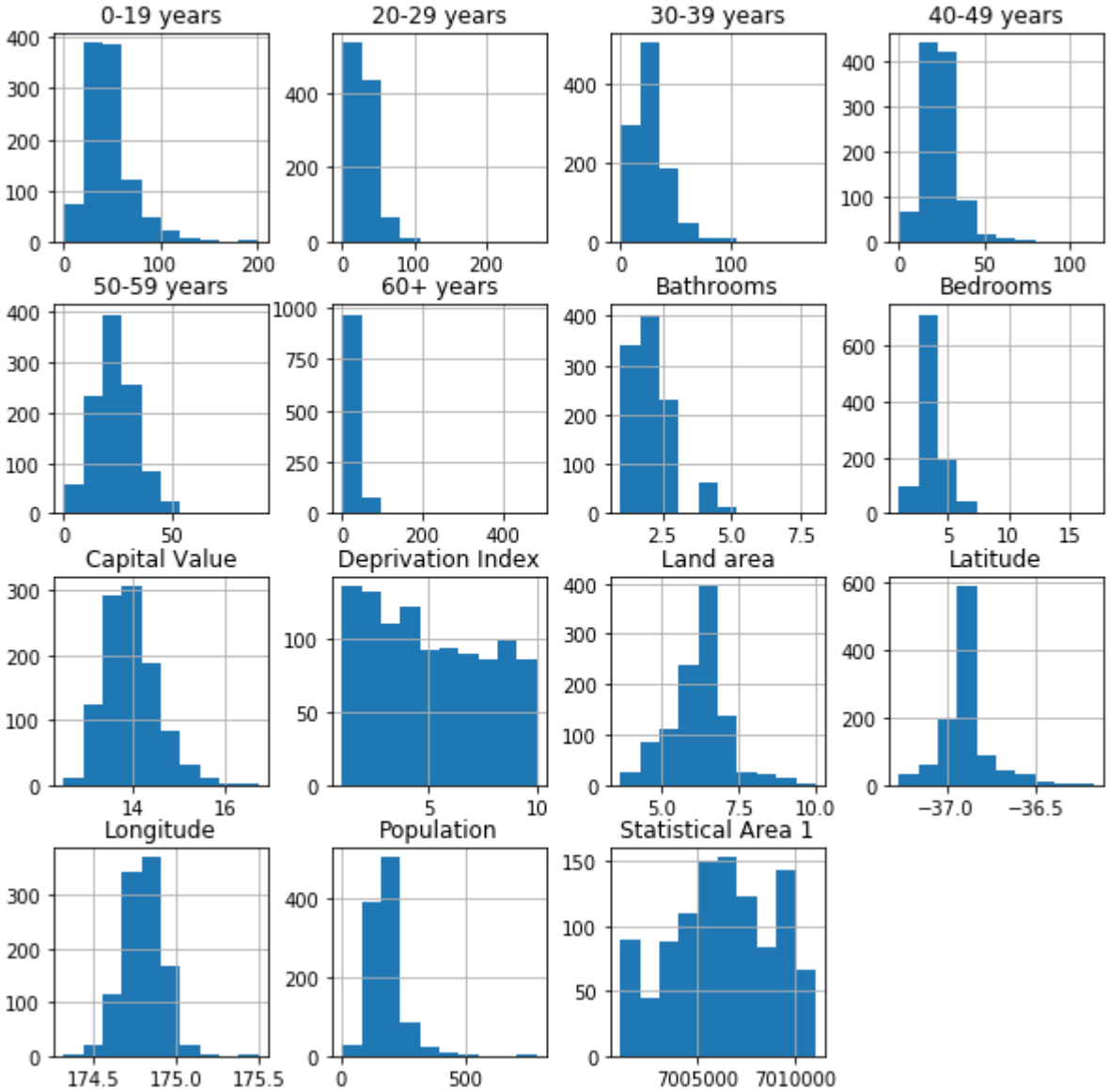
The initial exploration of the data began with some summary and descriptive statistics which include the minimum, maximum, mean, median, and standard deviation of each variable. The distinct count was calculated for the numeric columns and the results were taken from 1051 observations.

Feature	Min	Max	Mean	Median	Std
Bedrooms	1	17	3.777355	4	1.169412
Bathrooms	1	8	2.073264	2	0.992044
Land Area	40	22240	856.989534	571	1588.156219
Capital Value	270,000	18,000,000	1,387,521	1080000	1,182,939
Latitude	-37.265021	-36.177655	-36.893715	-36.893132	0.130100
Longitude	174.317078	175.492424	174.799325	174.798575	0.119538
SA1	7001130	7011028	7006319	7006325	2591.262
0-19 years	0	201	47.549001	45	24.692205
20-29 years	0	270	28.963844	24	21.037441
30-39 years	0	177	27.042816	24	17.975408
40-49 years	0	114	24.125595	24	10.942770
50-59 years	0	90	22.615604	21	10.210578
60+ years	0	483	29.360609	27	21.805031
Population	3	789	179.914367	174	71.059280
Deprivation Index	1	10	5.063749	5	2.913471

The maximum bedrooms is 17 which is far off the mean at around 3.77 bedrooms, this seems to be an outlier from the rest of the data. Land area has a large range, as the minimum area is just 40m² and the maximum area is 22240 m². The minimum capital value is $270,000 which is far off the mean at around 1.4 million dollars, this seems to be an outlier from the rest of the data. The mean population for 0-19 year olds is almost double the mean population of the other age groups, there are some areas with 0 population of certain age groups as well. Population has a large range as the minimum is 3 whereas the maximum is 789, the mean of around 180, being far from these two values as well. The mean deprivation index is around 5, which is expected as it's a score from 1-10.



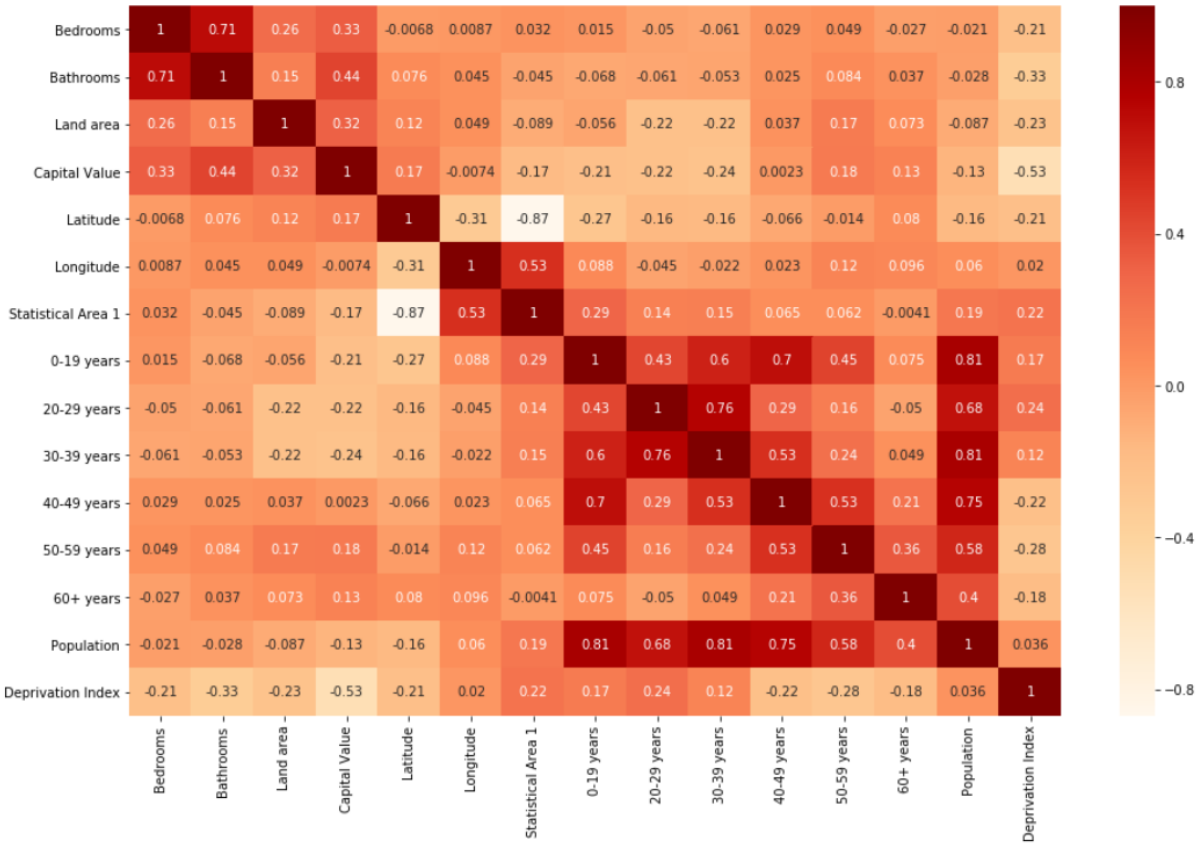
The population histograms are skewed to the right, so most areas have a small population while only some areas have a larger population. This is emphasized by the histogram of the 60+ age group where it is extremely skewed to the right. Most houses seem to have less bathrooms, bedrooms, capital value, and land area while only a few houses have really large values of those columns. Deprivation index has a slight decline as deprivation index increases, meaning there's slightly less deprived areas. Latitude, longitude, and Statistical Area 1 are approximately normally distributed. As capital value and land area seem to be extremely skewed, I'll log transform them to try to improve it.



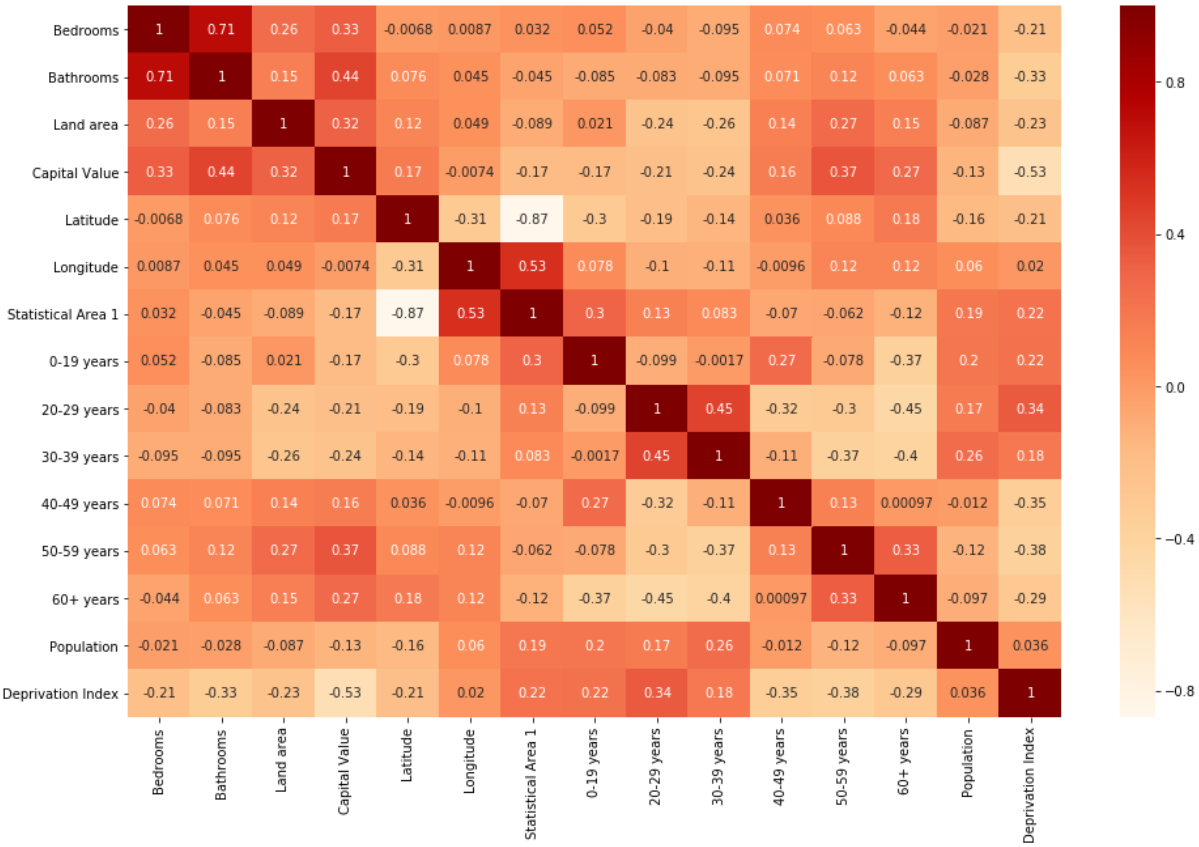
The log transformation helps us show that capital value is right skewed while land area is left skewed.

Correlations and Relationships

The correlation between the numeric columns were calculated and observed in the below correlation plot. The right colour bar indicates the correlation values, dark red means the correlation value is 1 and light yellow means the correlation value is -1.



From the histograms and the correlation matrix, we see that as the number of bedrooms in a house increases, the number of bathrooms also increase. There's a negative correlation between latitude and SA1 and a positive correlation between longitude and SA1, which signals that the Statistical Area is ordered by location. Lastly, there's the obvious positive correlations with population and the population of each age group, as the population increases, the population of each age group also increases. This can be misleading as some areas can have different proportions of age groups, showing this proportion is better than just showing the population of each age group, so I'm changing it to the percentage of the population for each age group. There aren't any significant correlations with capital value which is worrying on trying to make a machine learning model on trying to predict it.



Though changing age groups to percentages has decreased they're correlation with population, it has increased it's correlation with capital value which will help with the machine learning model.

Analysis

In this analysis, three algorithms were tested, lasso, decision tree regressor, and linear regression.

This algorithm was trained with 70% of that data and tested with the remaining 30%, which produced the following results:

Data manipulation	**R^2 value**
Lasso	0.324353175002132
Decision Tree Regressor	0.3669835612804371
Linear Regression	0.41948723305555813

As shown from the above table, Linear Regression has the highest R^2 value to predict house prices.

Decision for model chosen, process and results

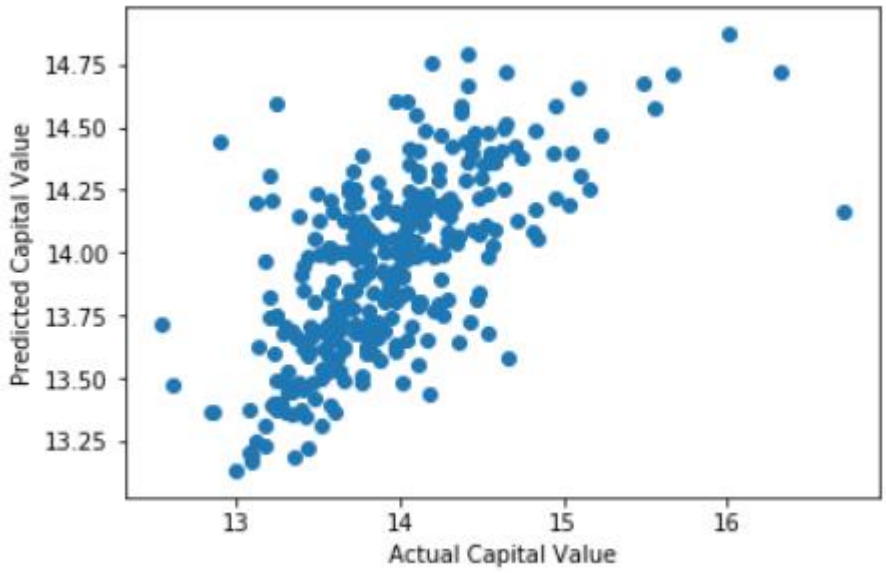
Based on the output of the model evaluation, Linear Regression is chosen as the prediction model.

These were the coefficients of each input variable, how much each variable affected the predicted capital value.

Variable	Coefficient
Bedrooms	-4.30825699e-06
Bathrooms	1.64678901e-01
Land Area	8.05040086e-02
Latitude	1.27948668e-02
SA1	-6.67434852e-06
0-19 years	-5.32176725e-02
20-29 years	5.74711371e-01
30-39 years	6.39935163e-01
40-49 years	5.81816485e-02
50-59 years	1.27376248e+00
60+ years	6.44955000e-01
Population	-4.14015175e-04
Deprivation Index	-6.57548302e-02

Surprisingly, the number of bedrooms, SA1, and population didn't have much effect on the value of the house. What had the most effect seems to be some of the proportion of the age groups in the population, 20-29, 30-39, 50-59, and 60+, along with the number of bathrooms.

A plot comparing the model's predicted capital value of a house against its actual value is shown below.



There are a few outliers at both extremes, but there is a general positive trend. This translates into an R^2 value of 0.42.

Conclusions

This analysis has shown that house prices can't be confidently predicted from this dataset with different house and location statistics. Even though linear regression had the best performance among the three algorithms, it only had an R^2 value of 0.42.

References

Wang, X. (2018). Classification analysis of Breast cancer diagnostic. Retrieved 21 July, 2020, from https://github.com/NZMSA/2020-Phase-1/blob/master/Data%20Science%20and%20Machine%20Learning/SampleReport.pdf