Analysis of Auckland House Price

Maria Sacramento, July 2020

Executive Summary

The dataset contains the Auckland house price provided by NZMSA for the project. It contains Statistical Area Unit 1 (SA1) along with the Capital Value (CV) of the property which is often used to calculate payable rates and is an approximation of the house value.

The analysis is based on 1051 observations for each of the 17 variables. The variable of interest is the Capital Value (CV). The remaining variables are explanatory variables which contains details such as the number of bedrooms and bathrooms of the property, land area, the number of varying age groups that live in the area based on the 2018 census, mean Capital Value of the suburbs, current population and deprivation index.

After the data exploration through calculation of summary and descriptive statistics as well as by creating visualizations of the correlations between the numerical variables, five highly correlated variables are found. After exploring the data, 3 algorithms have been tested for the train dataset and the best model has been chosen based on how accurately the model predicts the response.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics Summary statistics for distinct count, mean, standard deviation, minimum, maximum and the interquartile range (IQR) which describes the distribution of the data where 50 is the 'median' and 25 and 75 are the upper and lower quarter of the data.

Adding the numbers across all of the age groups will not equal the population number in the Statistical Unit Area 1 (SA1). This is because all this information is based off the 2018 Census, and there are people who did not fill their ages and thus the population is greater than if we sum up the numbers across the age groups.

In [16]:
```
from IPython.display import Image
Image('img/SummaryTable.PNG')
```

Out[16]:

	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	Current Population	Deprivation Index
count	1051.000000	1049.000000	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	1051.000000
mean	3.777355	2.073403	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	5.063749
std	1.169412	0.992985	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	2.913471
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	1.000000
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	2.000000
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	5.000000
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	8.000000
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	10.000000

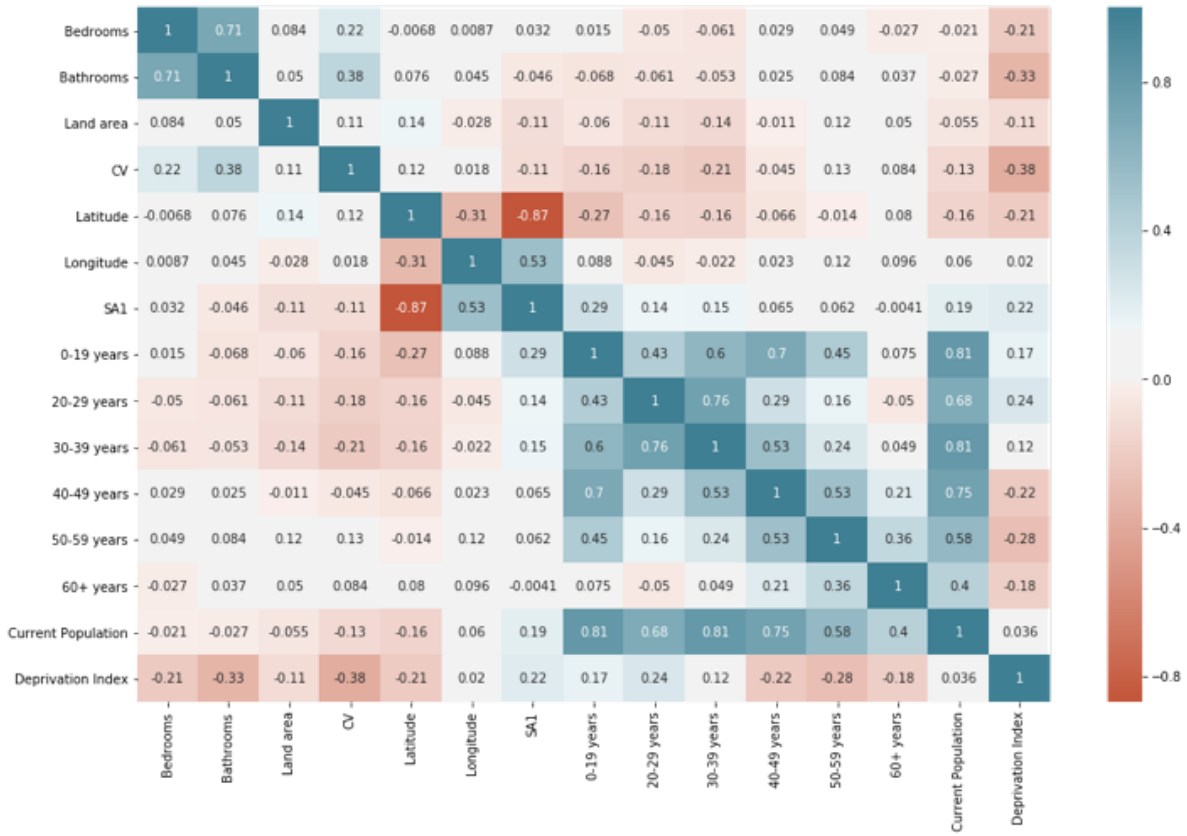
Correlation and Relationships

Numeric Relationships

The correlation between the numeric columns were calculated and observed based in the correlation plot below. The color bar in the right indicates the correlation values. The dark shade of blue shows a correlation value of 1 and the dark shade of red conveys a correlation of negative 1. The stronger the color corresponds to a larger correlation magnitude.

In [21]:
```
Image('img/heatmap.png')
```

Out[21]:



The graph above diplays a strong positive correlation between Current Population and the different age groups (0-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60+ years). The Capital Value (CV) shows weak negative correlation between the Current Population, Deprivation Index, and some weak positive correlation between Bedroom, Bathrooms and Land Area.

Analysis of correlations and patterns in the data

The analysis tests three algorithms, which are Linear Regression, Decision Tree and Random Forest.

All the algorithms are trained with 70% of the data, while the testing model used 30% of the data.

Regression Model:

- R^2 = 33.72%
- RMSE = 1090431.94

Decision Tree:

- Classification Rate = 3.17%
- RMSE = 1617893.82

Random Forest:

- Accuracy = 72.61%
- RMSE = 1537645.75

The Random Forest has the highest Accuracy Precision while the Regression model has the lowest RMSEP.

Model Building

Through model evaluation, Linear Regression is chosen as the best predictive model. The regression model has an R^2 of 0.3372 which means that only 33.72% of the observed variation can be explained by the model's inputs.

The model however had the best Root Mean Square Error out of all the three models that have been built which signifies that out of the three models, the regression model have the closest predicted values to the actual observed data.

Lower RMSE values indicates a better fit, and is generally a good measure to see how well a model predicts the response variable (Capital Value) and is an important criteria if the purpose of the model is for predicting.

The Linear Regression model has the following results:

- Explained Variance: 0.3373
- R^2: 0.3372
- MAE: 449227.0637
- MSE: 1189041826570.668
- RMSE: 1090431.945

Conclusions

The analysis shows that the Capital Value of Auckland houses cannot be confidently predicted due to its low model score. The model can explain 33.72% of the variation in the data. However, the regression model has the lowest Root Mean Square Error (RMSE) out of all the models which is why it's chosen as the best predictive model out of the three.