Use Linear Regression to Predict Auckland's House Prices

Jiaming Wan (Jamin), July 2020.
Jwan716@aucklanduni.ac.nz

Executive Summary

Provided by MSA team, the original dataset is about house price in Auckland, and two additional columns have been added by joining with the datasets from Stat.NZ and the University of Otago respectively, which are both reliable sources. It not only contains the basic information of the property such as land area and the number of rooms but also includes population data from census in 2018.

The analysis is based on 1051 observations for each of the 17 variables. The most important attribute is the value of the property (CV), which is an approximation of value I am going to predict in the data analysis. The main house and population data is intuitive, and all of them have statistic measurement such as mean, min, max, standard error, etc. Besides, there are several geographical data that could be used to enhance visualization, providing a better interpretation for the decision-maker. However, it could be redundant because when the dataset has Suburbs, Latitude, and Longitude variables, Address becomes unnecessary, and it can be dropped in the modelling phase. Lastly, the dataset also contains the deprivation index data. It's considered as a valuable attribute since it can be utilized for the social psychology study, and in this case, more related data needs to be collected to fulfil the task.

This report has been divided into three different parts. Firstly, I will explore the data by calculating summary and descriptive statistics, following by the visualization of the correlation between the numerical variable and geographic data by grouping different suburbs within the same value range. Finally, three algorithms have been tested for the training dataset, and the best model has been selected based on the model's accuracy.

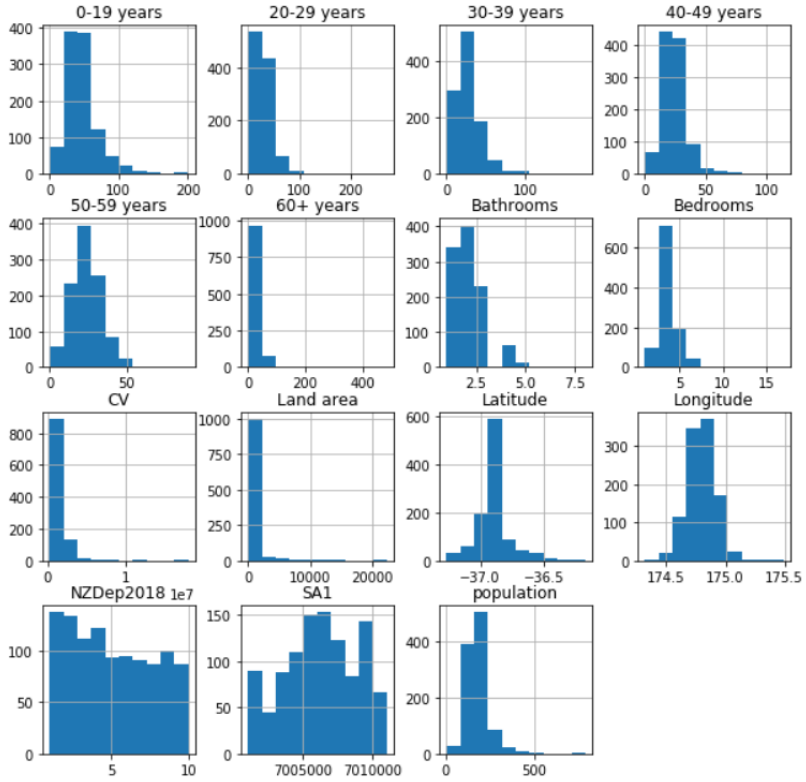
Explore the Initial Data

The count, mean, min and max rows are self-explanatory. The std shows the standard deviation, and the 25%, 50% and 75% rows show the corresponding percentiles. Although there are some houses with 17 bedrooms or 8 bathrooms, I still believe that they are not outliers as it could be the luxury house or for tourist rental purpose.

	Bedrooms	Bathrooms	CV	Latitude	Longitude	SA1	0-19 years	20-2
count	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0
mean	3.7773549	2.0732636	1387520.5518554	-36.8937149	174.7993254	7006319.1826832	47.5490010	28.9
std	1.1694122	0.9920439	1182939.3647126	0.1301004	0.1195384	2591.2617627	24.6922048	21.0
min	1.0000000	1.0000000	270000.0000000	-37.2650209	174.3170782	7001130.0000000	0.0000000	0.0
25%	3.0000000	1.0000000	780000.0000000	-36.9505652	174.7207792	7004415.5000000	33.0000000	15.0
50%	4.0000000	2.0000000	1080000.0000000	-36.8931322	174.7985754	7006325.0000000	45.0000000	24.0
75%	4.0000000	3.0000000	1600000.0000000	-36.8557886	174.8809439	7008383.5000000	57.0000000	36.0
max	17.0000000	8.0000000	18000000.0000000	-36.1776547	175.4924245	7011028.0000000	201.0000000	270.0

20-29 years	30-39 years	40-49 years	50-59 years	60+ years	population	NZDep2018
1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000	1051.0000000
28.9638440	27.0428164	24.1255947	22.6156042	29.3606089	179.9143673	5.0637488
21.0374411	17.9754084	10.9427698	10.2105783	21.8050306	71.0592797	2.9134710
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	3.0000000	1.0000000
15.0000000	15.0000000	18.0000000	15.0000000	18.0000000	138.0000000	2.0000000
24.0000000	24.0000000	24.0000000	21.0000000	27.0000000	174.0000000	5.0000000
36.0000000	33.0000000	30.0000000	27.0000000	36.0000000	210.0000000	8.0000000
270.0000000	177.0000000	114.0000000	90.0000000	483.0000000	789.0000000	10.0000000

To get a better feel of what kind of data I am dealing with, I plot a histogram for each numeric variable. Some of the histograms are left-skewed because for the SA1 unit date, it ideally has a size range of 100-200 residents.



Explore the Price with Suburb Data

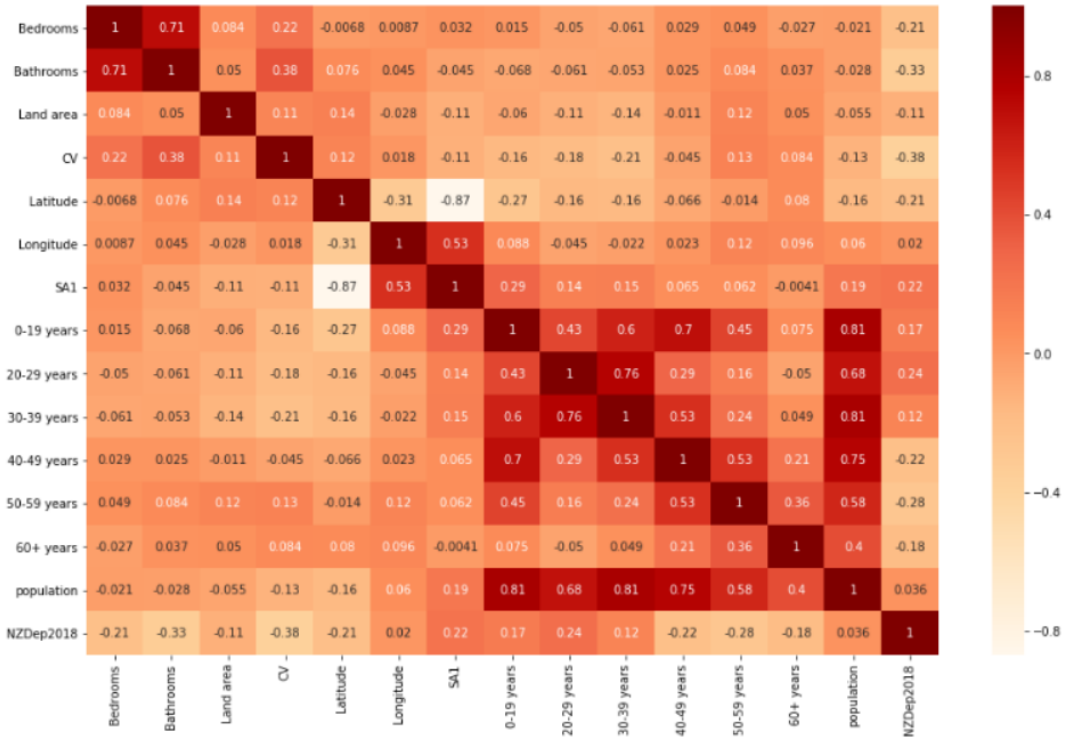
Since each house has a different land area and each suburb has different home prices, I add a new price per square meter variable correlates with house value. Then I can use price per square meter's range to get clusters of suburbs.

```
Index(['Albany Heights', 'Alfriston', 'Army Bay', 'Beach Haven', 'Beachlands',
       'Birkdale', 'Botany Downs', 'Buckland', 'Burswood', 'Clarks Beach',
       'Clendon Park', 'Clover Park', 'Cockle Bay', 'Conifer Grove', 'Drury',
       'East Tamaki Heights', 'Farm Cove', 'Glen Innes', 'Glenfield',
       'Golflands', 'Great Barrier Island',
       'Great Barrier Island (Aotea Island)', 'Green Bay', 'Gulf Harbour',
       'Half Moon Bay', 'Helensville', 'Henderson', 'Highland Park',
       'Hillsborough', 'Huapai', 'Huia', 'Kaipara Flats', 'Karaka',
       'Kawakawa Bay', 'Kelston', 'Laingholm', 'Leigh', 'Lynfield',
       'Manurewa East', 'Maraetai', 'Massey', 'Matakatia', 'Mellons Bay',
       'Murrays Bay', 'Narrow Neck', 'Omiha', 'Oneroa', 'Onetangi', 'Opaheke',
       'Orewa', 'Ostend', 'Otara', 'Pahurehure', 'Pakuranga', 'Palm Beach',
       'Papakura', 'Paremoremo', 'Patumahoe', 'Point Chevalier',
       'Point England', 'Pokeno', 'Pukekohe', 'Ramarama', 'Ranui', 'Red Beach',
       'Red Hill', 'Redvale', 'Rosehill', 'Silverdale', 'Snells Beach',
       'Somerville', 'South Head', 'Stanmore Bay', 'Sunnyhills', 'Surfdale',
       'Swanson', 'Takanini', 'Te Atatu Peninsula', 'The Gardens',
       'Tindalls Beach', 'Titirangi', 'Totara Park', 'Totara Vale', 'Tuakau',
       'Unsworth Heights', 'Waimauku', 'Wainui', 'Waitakere', 'Waitoki',
       'Waiuku', 'Warkworth', 'Wattle Downs', 'Wellsford', 'Weymouth', 'Wiri'],
      dtype='object', name='Suburbs')
```

high price and low frequency suburbs

Correlation and Relationships

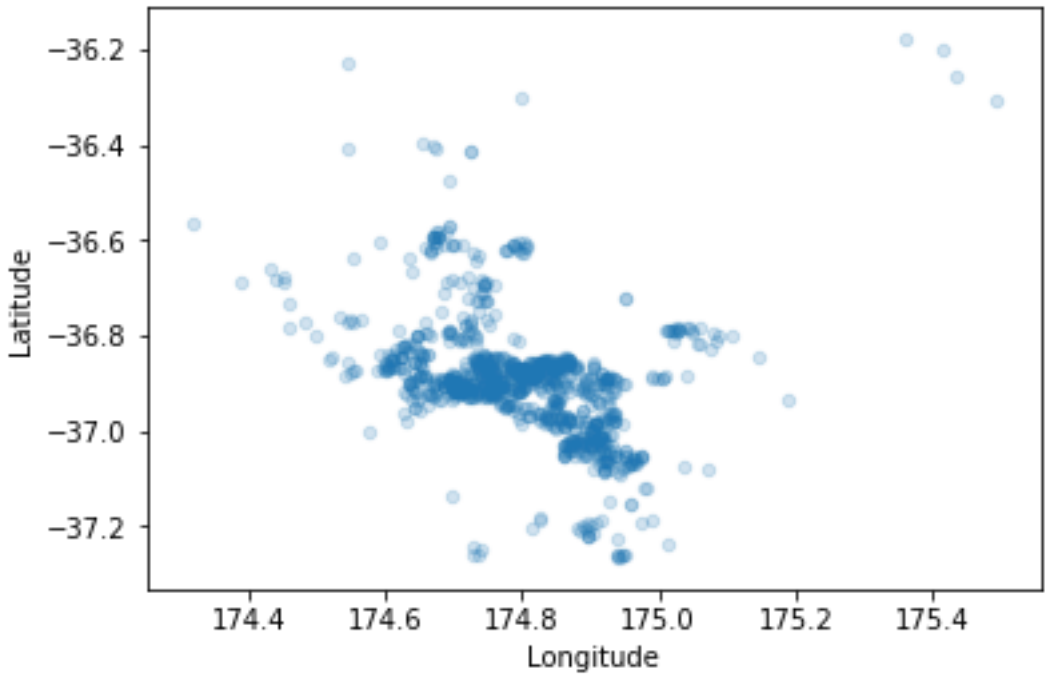
The correlation between the numeric columns has been plotted. The colour and digit numbers on the cell indicate the correlation values which are between -1 and 1. The graph shows that for the house price, the individual variable has little positive correlation, but it could be different when combining multiple variables. The cost of the house increases when the number of bathrooms goes up. In the meantime, you can also see some negative correlation. Coefficients close to zero indicate that there is no linear correlation.



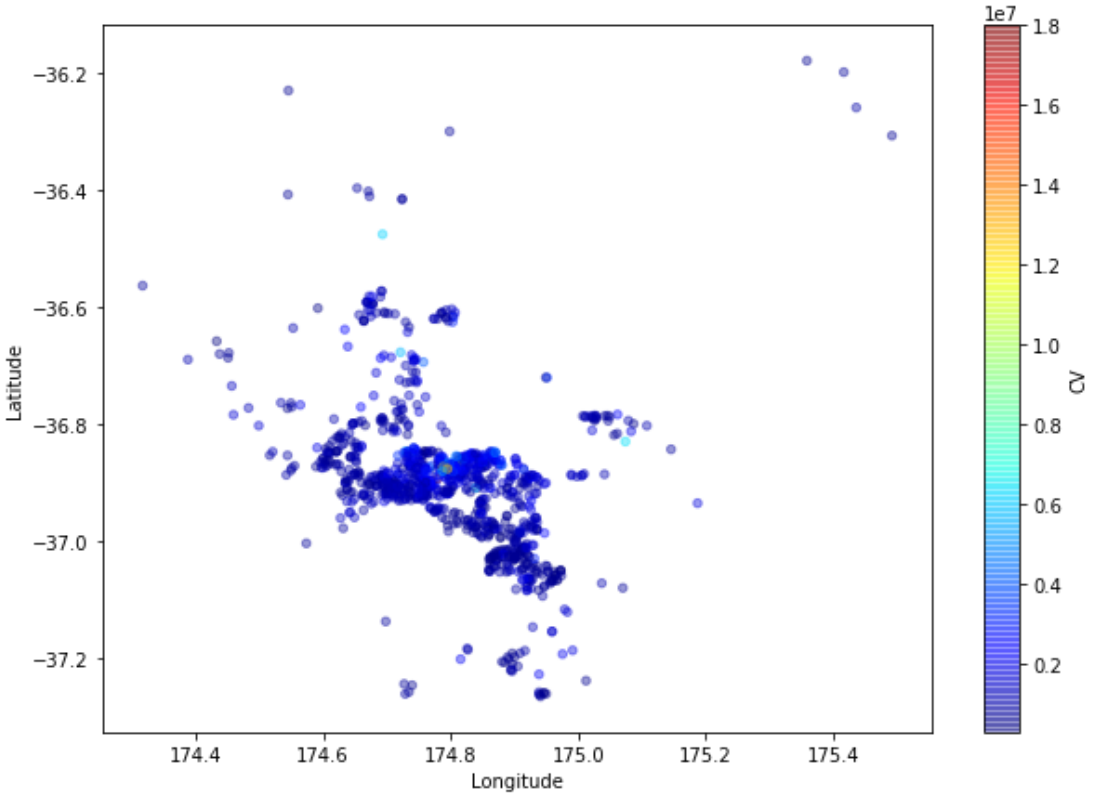

```
CV              1.0000000
Bathrooms       0.3758914
Bedrooms        0.2248361
50-59 years     0.1310294
Latitude        0.1206094
Land area       0.1122202
60+ years       0.0835315
Longitude       0.0183167
40-49 years    -0.0445647
SA1            -0.1099203
population     -0.1284738
0-19 years     -0.1560097
20-29 years    -0.1824583
30-39 years    -0.2143120
NZDep2018      -0.3781205
Name: CV, dtype: float64
```

Geographic Data Visualization

To create a better data visualization, I create a scatter plot with latitude and longitude.



I used colour code from the most expensive to the least expensive areas. If I can add an Auckland map layer, it could be more intuitive.



Analysis

There are three machine learning models been tested to predict the house price, namely Linear Regression, Random Forest Regression and Gradient boosting, and the performance of each model is below. These algorithms were trained with 70% of the data. Testing the model with the remaining 30% of the data.

Name	R squared	RMSE
Linear Regression	0.4365	1978728.8932
Random Forest Regression	0.5728	1978728.8477
Gradient boosting	0.5903	1978728.8512

As we can see from the table, Gradient Boosting has the highest accuracy, and the model predicts the value of the property is around NZD1,978,728.

From the feature importance table, we can see which features are more important in Gradient Boosting model, and the top 3 features are NZDep2018, SA1, Land Area.

Conclusion

The analysis shows that the house price can be predicted by current dataset and variables, but there is still large space to improve because the best model's accuracy is less than 60%. In my opinion, the population data is not usually related to the price. Suppose we could change population data to more house info related data such as type, house of the year, facilities, etc. It could get a better result.