Analysis of Auckland House Prices based on Census 2018 data

Doris Lee, 26 July 2020

1. Executive Summary

This analysis is to explore and investigate the correlations between the house prices in Auckland and several factors that may have an impact on it, so as to create a machine learning model to predict the Auckland house values.

The initial dataset is the house prices and other attributes of 1,051 properties in Auckland. It is supplemented with the 2018 Census Population and the New Zealand index of socioeconomic deprivation of the statistical area in which a property is located.

The analysis is based on 1,051 observations for each of the 19 variables. The response variable is the capital value of a property (CV). This is used to calculate payable rates, and thus is an approximation of the house value.

The rest of the variables are explanatory variables. They include the features of a house which are usually provided on the listings of houses for sale, such as the address, suburb, land area and number of bedrooms and bathrooms. The total population, age distribution and the Deprivation index of the statistical area in which the house is located, are collected and merged with the property data to create a dataset for study.

This report first explores the data with descriptive statistics and data visualisation. Second, individual variables are studied by histograms and boxplots for identifying outliers and skewness. Third, data cleaning and preparation are conducted accordingly. Fourth, correlations between variables are investigated. Finally, a machine learning model is built, and the findings are discussed.

2. Initial Data Exploration

The initial exploration of data began with calculating the descriptive statistics. Both metric and categorical (non-metric) variables are found in the dataset.

2.1 Metric Variables

The summary statistics of the metric variables, including minimum, median, mean, maximum and standard deviation, are shown in the table below.

Feature	Min	Median	Mean	Max	Std
CV	270,000	1,080,000	1,387,521	18,000,000	1,182,939.36
Bedrooms	1	4	3.777	17	1.169412
Bathrooms	1	2	2.073	8	0.992985
Land area	40	571	856.99	22,240	1588.156219
0-19 years	0	45	47.55	201	24.692205
20-29 years	0	24	28.96	270	21.037441
30-39 years	0	24	27.04	177	17.975408
40-49 years	0	24	24.13	114	10.94277
50-59 years	0	21	22.62	90	10.210578
60+ years	0	27	29.36	483	21.805031
Population	3	174	179.9	789	71.05928
NZDep2018_Score	849	959	986.5	1380	94.287255

2.2 Categorical Variables

NZDep2018 and Suburbs are the two categorical variables in the dataset. NZDep2018 is the deprivation ordinal scale ranges from 1 to 10, where 1 represents the areas with the least deprived scores and 10 the areas with the most deprived scores. There are 189 Suburbs in the dataset, with 61 properties in Remuera making it the most represented Suburb, whereas 47 Suburbs have only one property representing.

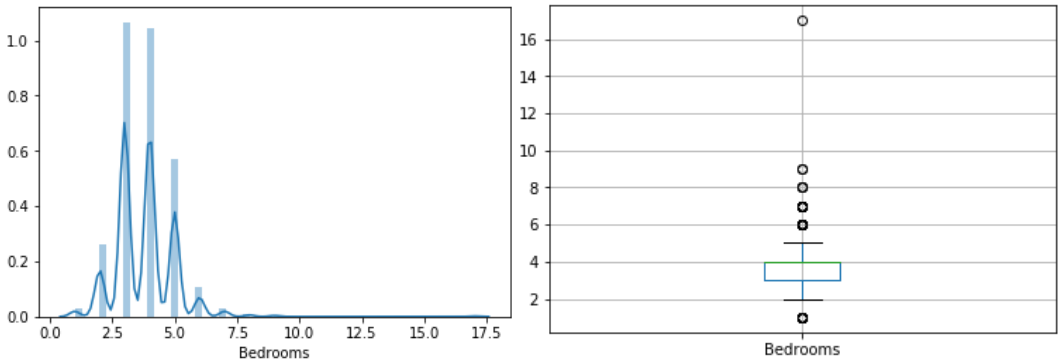
3. Data Cleaning and Preparation

3.1 Missing Values

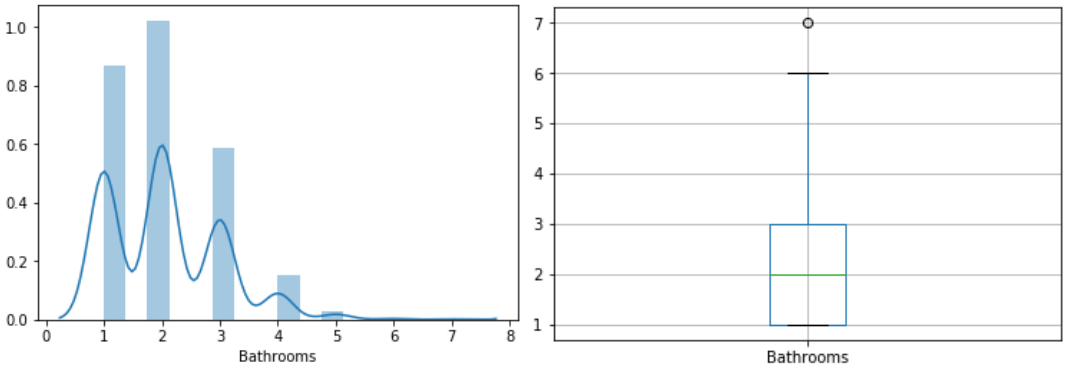
It was found that there were two missing values in the Bathrooms, and one missing value in the Suburb. Since the number of records with missing values in Bathrooms are small, the two rows with NaNs were dropped. The missing value in the Suburb was filled in by retrieving another record in the same statistical area (SA1).

3.2 Outliers

By visualising the distribution of Bedrooms, we can see that the house with 17 bedrooms is a clear outlier. Therefore, this observation is removed.

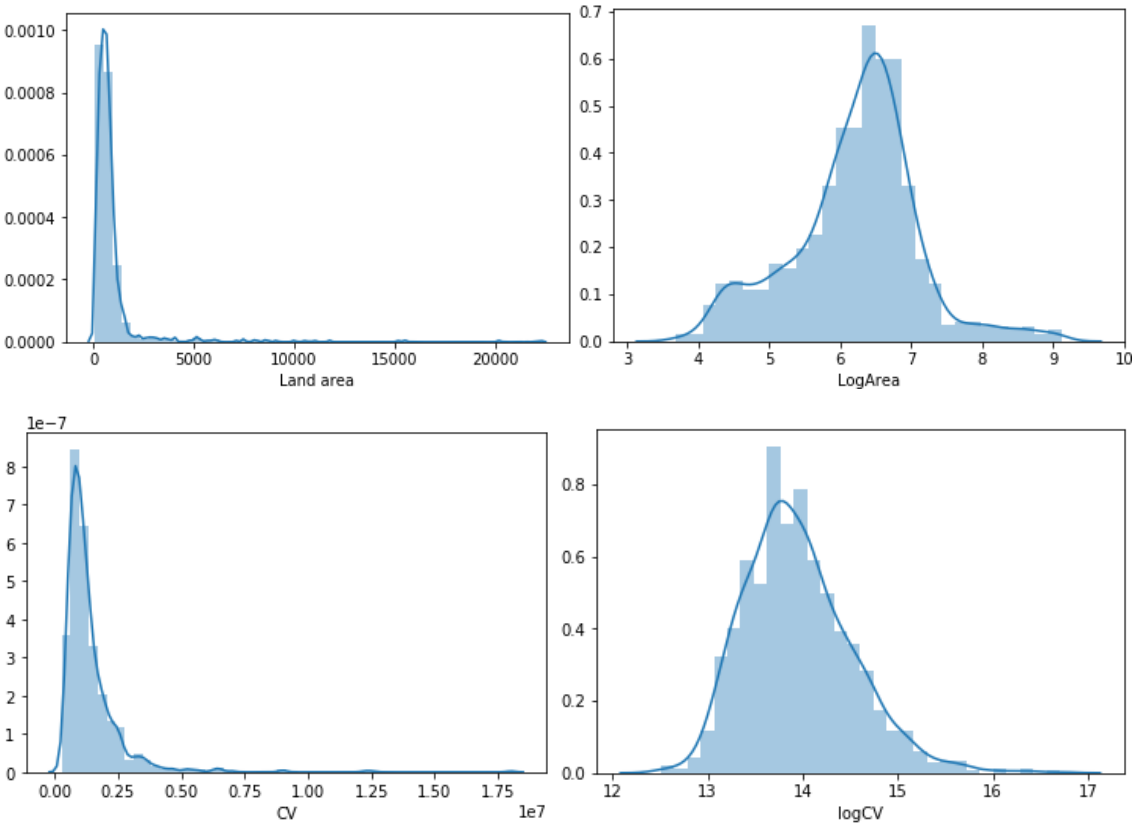


The same goes for the property with 7 bathrooms, so this observation is removed as well.



3.3 Skewed distributions

Land area and CV are observed to be right-skewed by a histogram plot. Therefore, logarithm transformation is performed to make it more normal to help meet the normality assumption of linear regression.

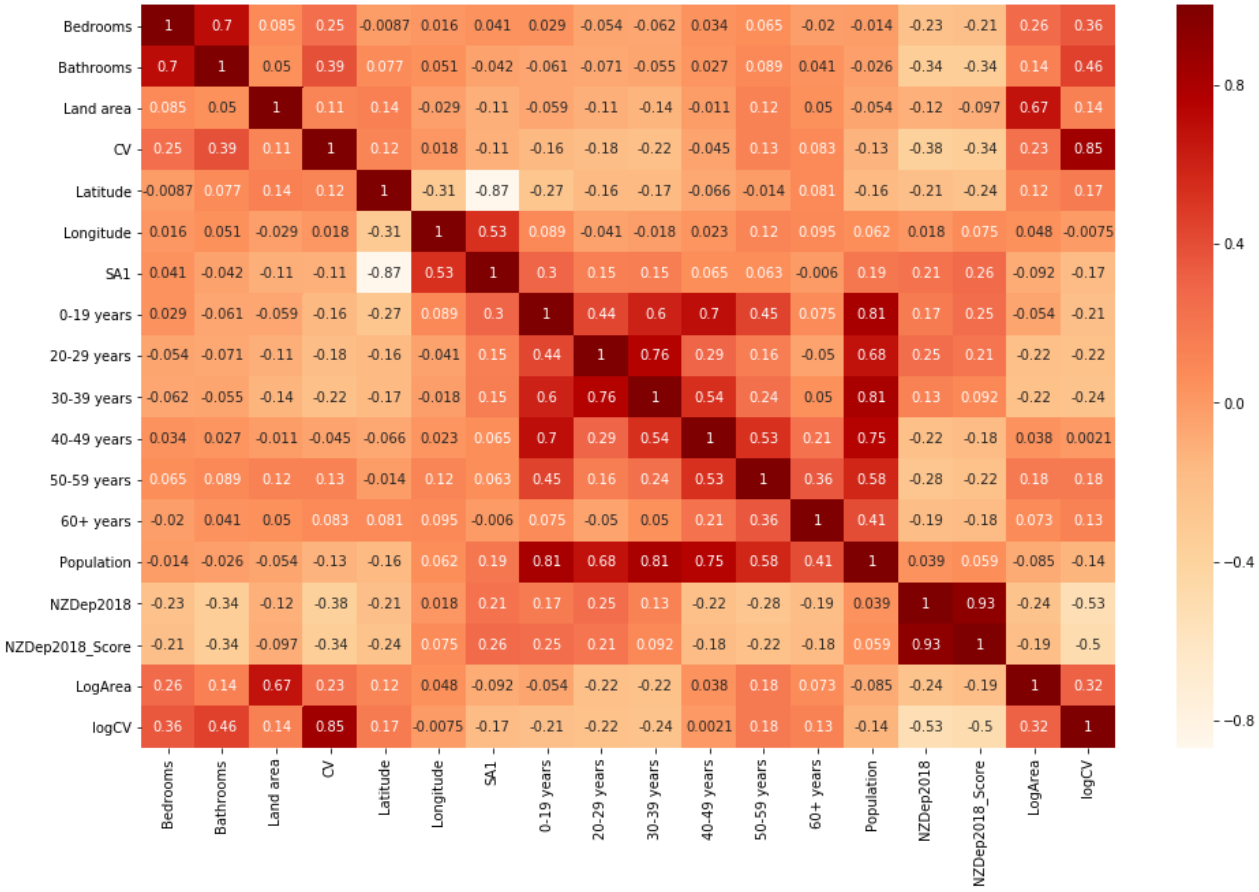


4. Analysis of Correlations and Patterns

4.1 Metric Variables

The correlation matrix and heatmap below shows the correlation between the numeric columns. Some correlations are observed:

- Numbers of Bedrooms and Bathrooms are highly correlated (0.7)
- NZDep2018_Score is the most (negatively) correlated metric variable to log-CV (-0.5)
- Bedrooms (0.36), Bathrooms (0.46), log-Land area (0.32) are all positively correlated to log-CV
- Population is slightly negatively correlated to log-CV (-0.14)
- Population of 0-19 (-0.21), 20-29 (-0.22) and 30-39 (-0.24) years are negatively correlated to log-CV
- Population of 50-59 (0.18) and 60+ (0.13) years are positively correlated to log-CV
- Population of 40-49 (0.0021) years is extremely slightly positively correlated to log-CV



4.2 Categorical Variables

The results of One-way and Two-way ANOVA show that the two categorical variables Suburbs and NZDep2018 both have highly significant influences on CV, but the two variables do not have a significant joint effect on CV.

```
                    Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(Suburbs) 189 5.762e+14 3.049e+12   2.939 <2e-16 ***
Residuals          861 8.931e+14 1.037e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                     Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(NZDep2018)    9 2.408e+14 2.676e+13   22.67 <2e-16 ***
Residuals            1041 1.229e+15 1.180e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                                      Df    Sum Sq   Mean Sq F value
as.factor(Suburbs)                   189 5.762e+14 3.049e+12   2.376
as.factor(NZDep2018)                   9 4.833e+13 5.370e+12   4.185
as.factor(Suburbs):as.factor(NZDep2018) 322 1.647e+14 5.115e+11   0.399
Residuals                            530 6.801e+14 1.283e+12
                                       Pr(>F)
as.factor(Suburbs)                   9.66e-15 ***
as.factor(NZDep2018)                 3.07e-05 ***
as.factor(Suburbs):as.factor(NZDep2018)        1
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

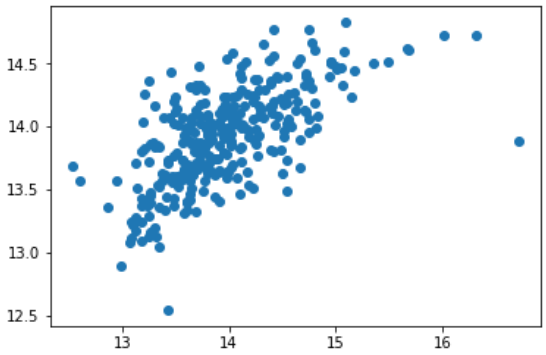
5. Machine Learning Model

A machine learning model was built with linear regression. Several variables were not included. First, Latitude, Longitude, Combined coordinates and SA1 were dropped because they were only used for data collection and was no longer necessary. Second, non-metric variables Address, Suburbs and NZDep2018 were dropped as they were not suitable for performing linear regression. Finally, Land area was dropped as we would use the log-transformed Land area instead.

The linear model below indicates that CV is positively impacted by Bedrooms, Bathrooms, all age groups except 30-39 years and Land area. CV is negatively impacted by 30-39 years, total Population and NZDep2018_Score.

```
array([ 0.03739721,  0.14277485,  0.00098027,  0.00400248, -0.00367907,
        0.00348535,  0.00842448,  0.00369248, -0.00270051, -0.00186615,
        0.09470633])
```

The algorithm was trained with 30% of the data. The model was then tested with the remaining 70% of the data. It yielded a model score of 0.42, which indicates a 42% of accuracy in its prediction. From the scatter plot, we can see that the model works better with lower house prices of less than 1,200,000 (e^{14}), but severely underestimates higher house prices of over 1,200,000.



6. Conclusions

From the correlation analysis, we can see that the Deprivation index has the most explanatory power on house prices, followed by the number of Bathrooms, Bedrooms and Land area. Population and age distribution have a comparatively weak explanatory power on property values.

The correlation analysis indicates that the Deprivation index (interval variable) is the most negatively correlated variable to log-CV. The ANOVA analysis agrees with this finding that the Deprivation index (ordinal scale) has a highly significant impact on the CV. It means that the Deprivation index has a strong negative explanatory power on house values.

The number of Bathrooms, Bedrooms and Land area are all positively correlated to log-CV. Therefore, these house features all have a positive impact on property prices.

Population is slightly negatively correlated to log-CV. It means that population has a negative impact on house prices, but not as much as the Deprivation index and the features of a property.

For the age distribution, younger age groups (0-39 years) are found to have a negative impact on the property values, whereas elder generations (50+) have a positive impact. The impact of 40-49 years is relatively insignificant.

The linear regression machine learning model has a relatively low accuracy (42%) in its prediction. It is also found that the model works better with lower house prices of less than 1,200,000 (e^{14}), but severely underestimates higher house prices of over 1,200,000.

7. Future Work

There is a lot of room for improvement in the accuracy of the machine learning model. Other algorithms can be used to build models for comparison, especially the ones that can take categorical features into account. More data can be collected for training the model so as to improve the accuracy.