**Regression Analysis of House Prices in New Zealand**

**Executive Summary**

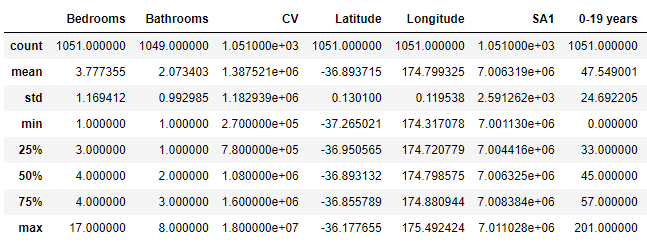
The dataset we currently have is the House prices dataset, which was given for us to use. The dataset contains many attributes regarding each house which includes the number of bedrooms/bathrooms, land area, capital value, latitude and longitude, etc.

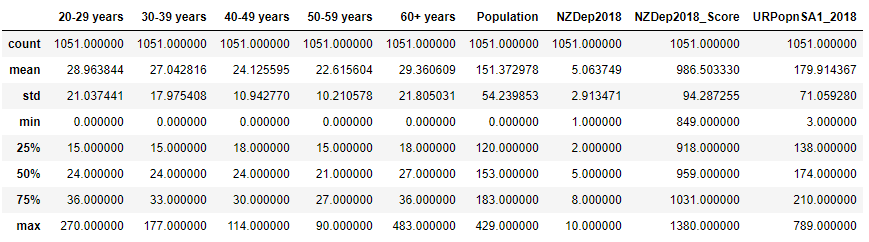
The analysis is based on 1051 observations for each of the 19 variables. The response variable is CV which is the capital value of the property. The rest are explanatory variables which may help us explain the response variable. They contains measurements of the property, For example we have the number of bedrooms and bathrooms as well as land area.

After exploring the data by looking at the descriptive statistics and by creating visualizations of the correlation between each numerical variable, several high correlated variables are found. After exploring the data, two models have been tested for the training dataset and the best model has been chosen based on the r^2 score.

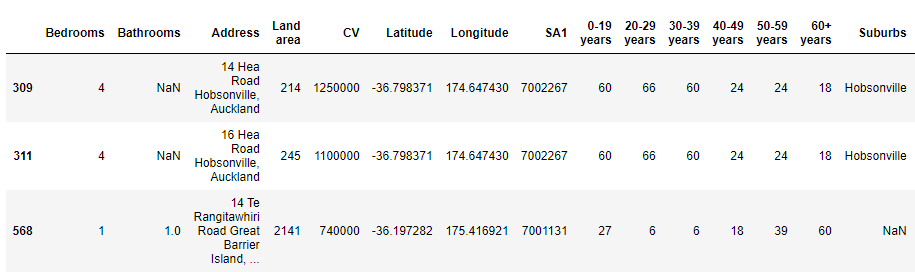
**Initial Data Analysis**

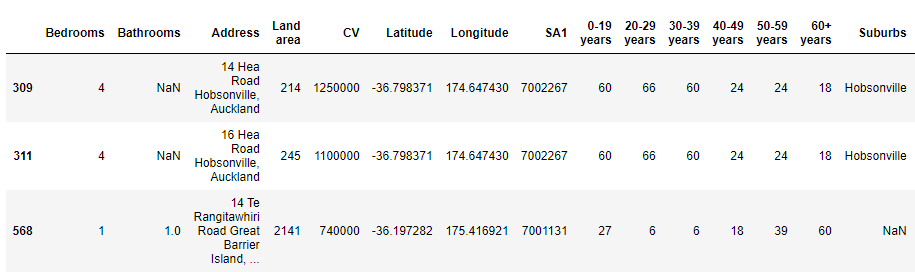
The Initial Data Analysis began with using the describe function in the panda library which gives us summary and descriptive statistics. The describe function allowed us to calculate the minimum, maximum, mean, median, standard deviation for the numerical columns and the results are shown below.





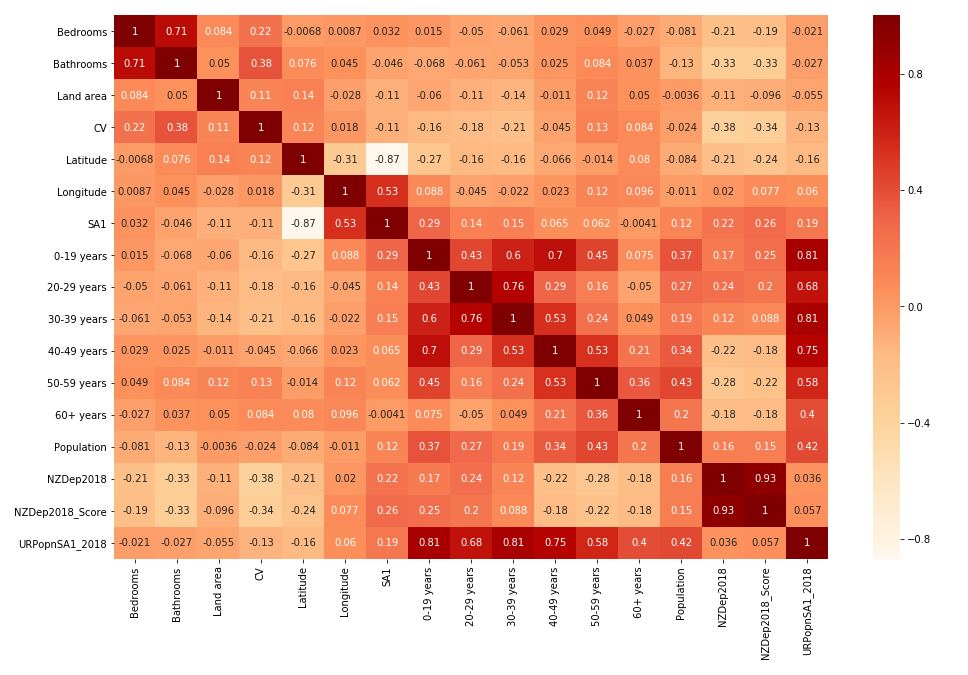
We also noticed that there were missing values in the dataset as shown below in the columns of Bathrooms and Suburbs, so I imputed values in the Suburbs column by manually fitting it in and for the Bathrooms column, I used the mean to fill in the values





**Correlation and Relationships**

We can use correlations between numerical columns to see how if there are any potential relationships between our response variable and our attributes



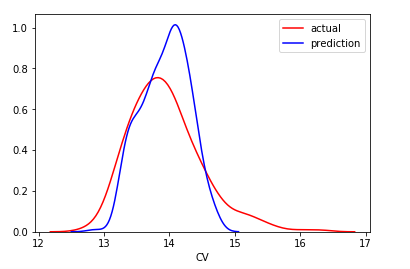
From our plot we can see that there is a strong correlation between bathrooms and bedrooms which may mean that we should include these variables in our machine learning model. Also, we can see that our response variable (CV) seems to have the strongest relationship with Bathroom and then Bedroom and Land area which may support evidence for us to use them in our machine learning model.

**Analysis**

In this Analysis, I used a linear regression and a polynomial linear regression model to analyze our data. I used a 70/30 split between the training and test data. We used a linear regression model because our response variable contains continuous data. I used three metrics to measure how good our model is which was R^2, Mean absolute error and Mean squared error. I removed other attributes as I felt that it didn’t have any effect on the model based on the correlation plot.

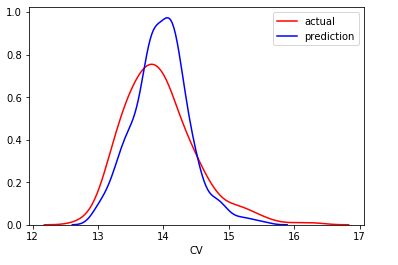
For our Linear regression model, we have the following metrics to assess our model:





For our polynomial regression model, we also have the following metric to assess our model:





As we can see from the screenshots produced above, it seems that Linear regression seems to have a better R^2 and the MSE and MAE are lower so it would be a suitable model to use out of the two

**Conclusion**

Based on the output of the model evaluation, we can see that Linear regression is chosen as the prediction model instead of polynomial regression. We can prove this through the metric scores especially R^2 score. R^2 score tells us the goodness fit of our model. A value of 1 tells that it can predict perfectly our data. So our Linear regression model is able to explain better than our Polynomial regression model so hence it is the most suitable model but this is not a good model to predict future datapoints as this will underfit as it cannot grasp the concepts of the data.