Data Collection

1. Include population
   1.1 Request the API from coordinates
   1.2 Request related figures as json
   1.3 Unpack json and append it as csv format
2. Include deprivation index
   2.1 Download csv file from Otago
   2.2 Using ‘SA1’ column as key to acquire corresponding value from the download csv file

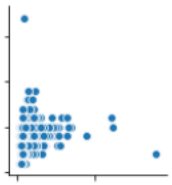
Data Processing

1. Handling NaN issues

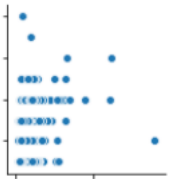
   NaN data existed in “Bathrooms” and “Suburbs” column. As “Suburbs” column will not be used later in the processing, only NaN existed in “Bathrooms” column need to be handled. And I choose to remove them as there’s 2 data points, and it will not be reasonable to give any estimate on this crucial data.

2. Redundancy data / Noise

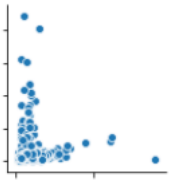
   Using the histogram, the noise is quite obvious.



The diagram on the right is graph for the CV-Bedrooms. It is clear that the left top data is a noise. And should be eliminated before moving into analysis.

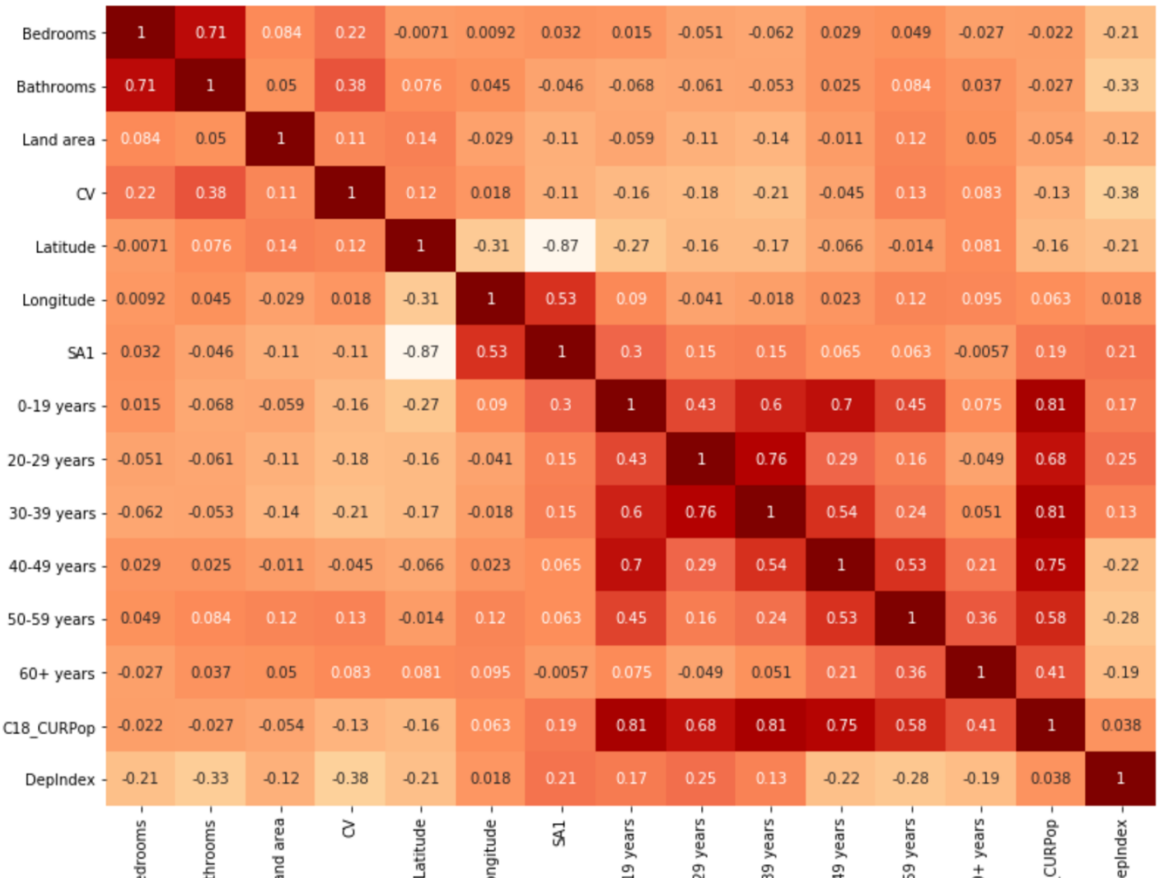


Similarly, the CV-Bathrooms graph demonstrated the unlikeliness that having more than 6 bathrooms.
The CV-Land area figure also indicates 15000 m^2 is probably not a regular range that we should estimate.



The similarity of the three graph above is the three data points having CV price more than 10M. This should also be taken out from our analysis as it will create great influence because its excessive numerical value.

3. Correlation Analysis



By analysing the correlation between CV and other parameter, the three most related figures are number of bedroom, bathrooms and deprivation index.

Data Analysis

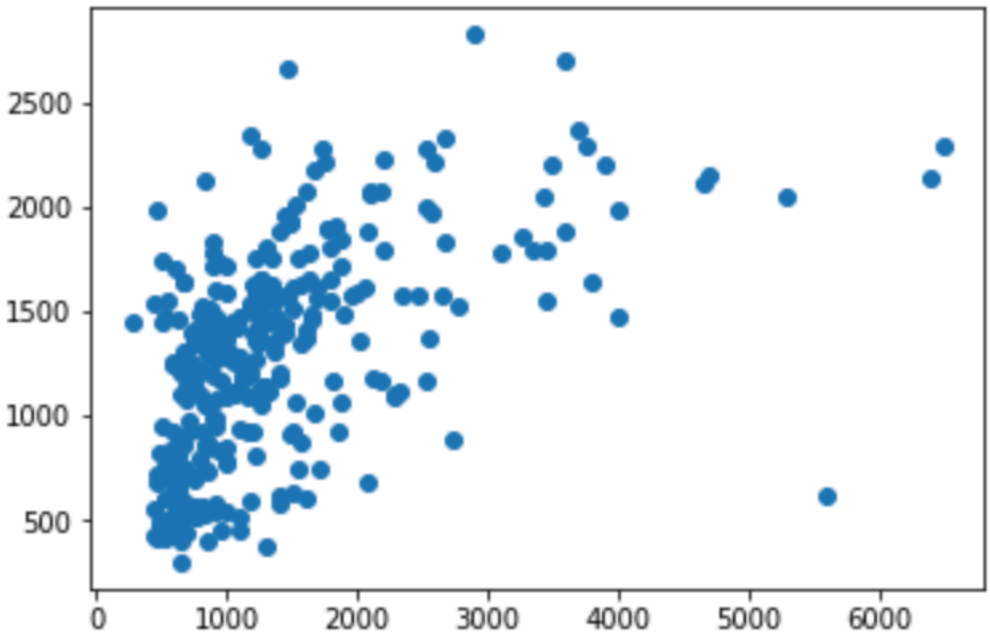
Linear Regression

Linear Regression isn't the best fit model in the situation. As it can hardly justify the impact of area code, such as "SA1", and how longititue and Latitute contributed to the final output. The location has strong impact on people decision to purchase houses but it can hardly to be linearly quantify.

The coefficient of show a different indication as what the heat map indicates.

```
 1.18475309e+04,  2.80552411e+05, -7.31442092e+00, -5.96570762e+04,
 3.88543570e+05, -3.43679564e+01,  3.12183490e+03,  3.94489978e+03,
-6.72674231e+03, -6.20698879e+03,  1.30256811e+04,  2.68116981e+03,
-2.11761880e+03, -9.44175305e+04])
```

The scatter plot show some prediction efficiency between 1000 – 3000 K, but miss out a lot more point in price higher than 3000K. The model generally underestimate the housing price.

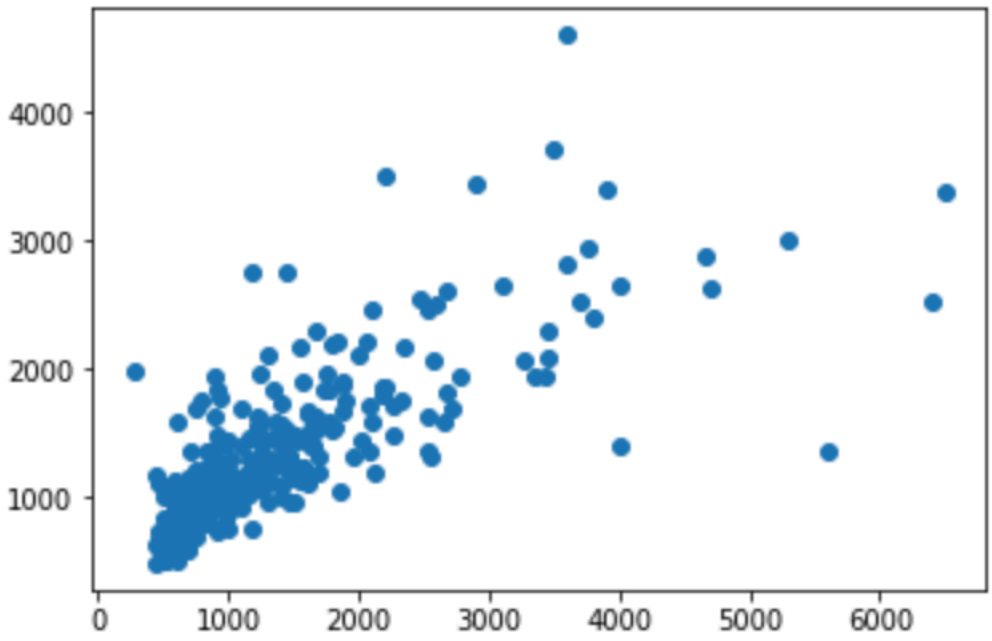



```
linear_model.score(test_x,test_y)
```

```
0.28981868462074956
```

Random Forest Regression

Random Forest Regression will address the problem listing above. Because it has decision trees to help analyse the influence of area code. And the random forest Regression allow a small sample splits, thus, it can better handle extreme training case compare to Linear Regression.



The scatter plot show a more concentration display.

```
RFR_model.score(test_x,test_y)
```

```
0.5629860202264829
```

And it give a more accurate estimate of the housing price.