Hypotheses:

After analysis by Weka, a Linear Formula was found. The rental cost of Wellington can be calculated with a linear formula.

```
--- Classifier model ---

Scheme:   LinearRegression
Relation: RentalCost-weka.filters.unsupervised.attribute.Remove-R1-weka.filters


Linear Regression Model

Wellington_mean =

      0.0008 * Population +
     -0.3714 * Otago_mean +
      0.9452 * Waikato_mean +
      0.3106 * National_mean +
   -269.8059
```

Fig 1

In fig 1, it shows the linear formula. The population of Wellington plays an important role due to the quantity of the population is very large compared to other attributes. It makes sense, large population means, the big demand for rentable houses.

The population increase will lead to higher rental cost. From the discovery, the rental cost of Wellington also correlates with the rental cost of other cities. The national rental cost is also a factor which affects the rental cost of Wellington. From the observations, the rental cost of Wellington is always higher than the National Rental Cost.

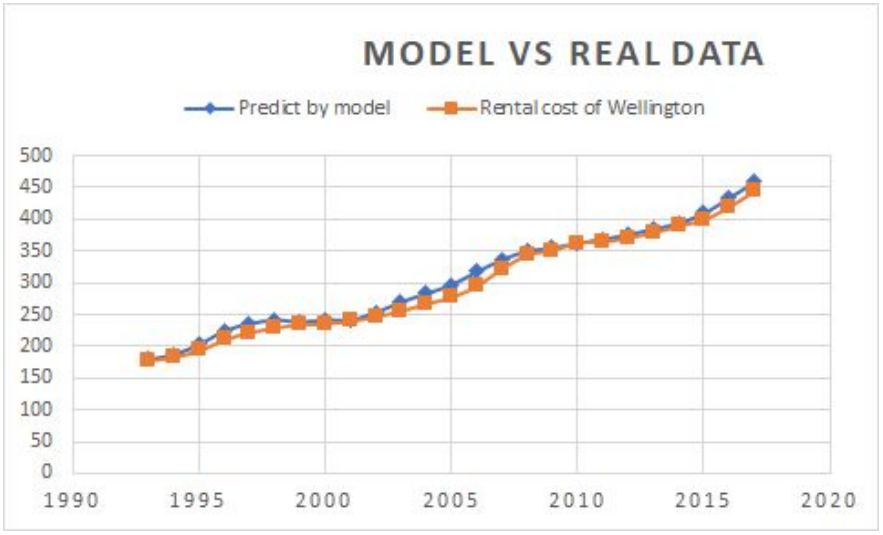


Fig 2

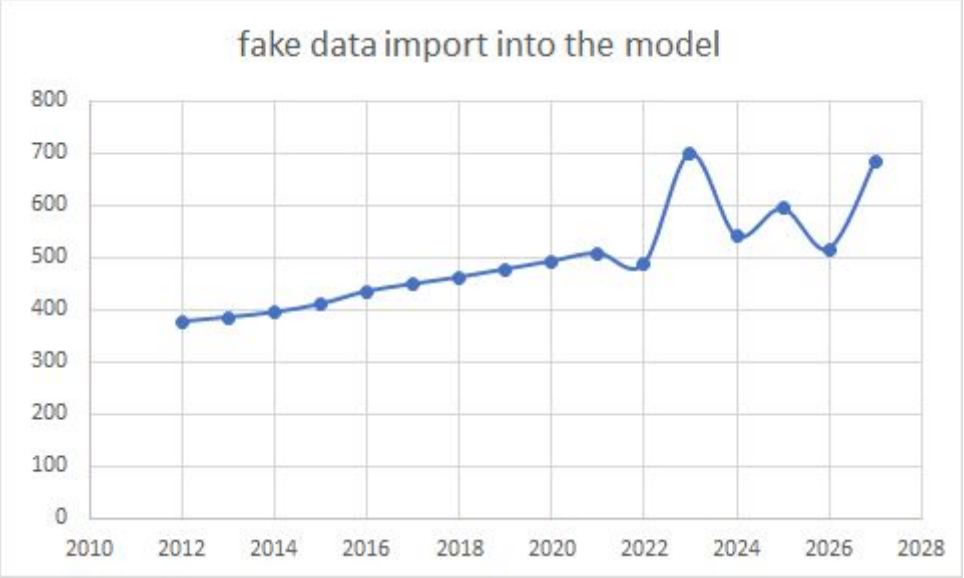
Fig 2, it shows the results obtained by the model are basically consistent with the actual data.

We can predict the rental cost with this model. But maybe it won't be very accurate, because there are only four variables. Say we have a very large outlier in the mean-rental cost Waikato attribute, the rental cost of Wellington will be extremely large.

I created several fake data to try my model.

Year	rental cost of Otago	Rental cost of Waikato	Rental cost across who	Wellington Population	Predict by model
2012	317	283.5833333	347.4166667	485100	376.4908833
2013	327	290.3333333	361.6666667	486700	384.8630333
2014	341.8333333	297.5	376.5833333	491400	394.5209833
2015	359	310.1666667	394.6666667	496900	410.1345
2016	385.1666667	333.6666667	412.9166667	504900	434.69685
2017	415	353	431.75	505900	448.54025
2018	425.6666667	360.35	446.85	512730	461.67993
2019	445.1428572	374.3428572	463.8071429	517920	477.09141
2020	464.6190476	388.3357143	480.7642857	523110	492.5028901
2021	484.0952381	402.3285715	497.7214286	528300	507.9143701
2022	600	416.3214287	514.6785714	533490	487.5122787
2023	523.0476191	600	531.6357143	538680	699.1242671
2024	550	416.3214287	600	543870	540.8871144
2025	550	442.3081634	546.1704082	600000	593.6343048
2026	700	453.3025511	559.4938775	554250	515.8544697
2027	597.3344671	464.2969389	572.8173469	700000	685.1146135

The data marked in red is the real data. And green data was created by excel with the linear regression function in excel. The blue data were the randomly created to find out how reliable the model is. And the black data was calculated with the model just created,



From the graph above, the data of year 2018 to year 2021 followed the slope of the line created by real data. After year 2021 the whole model became unstable. It means the formula found by the model I made is not very good. Because I only made up one data among the attributes each time. Next time if I want to use linear regression I will find more datasets correlated. Also I think I should make more correlated attributes which may can reduce the weight of each variable. It can make the model more stable when meeting outlier values.