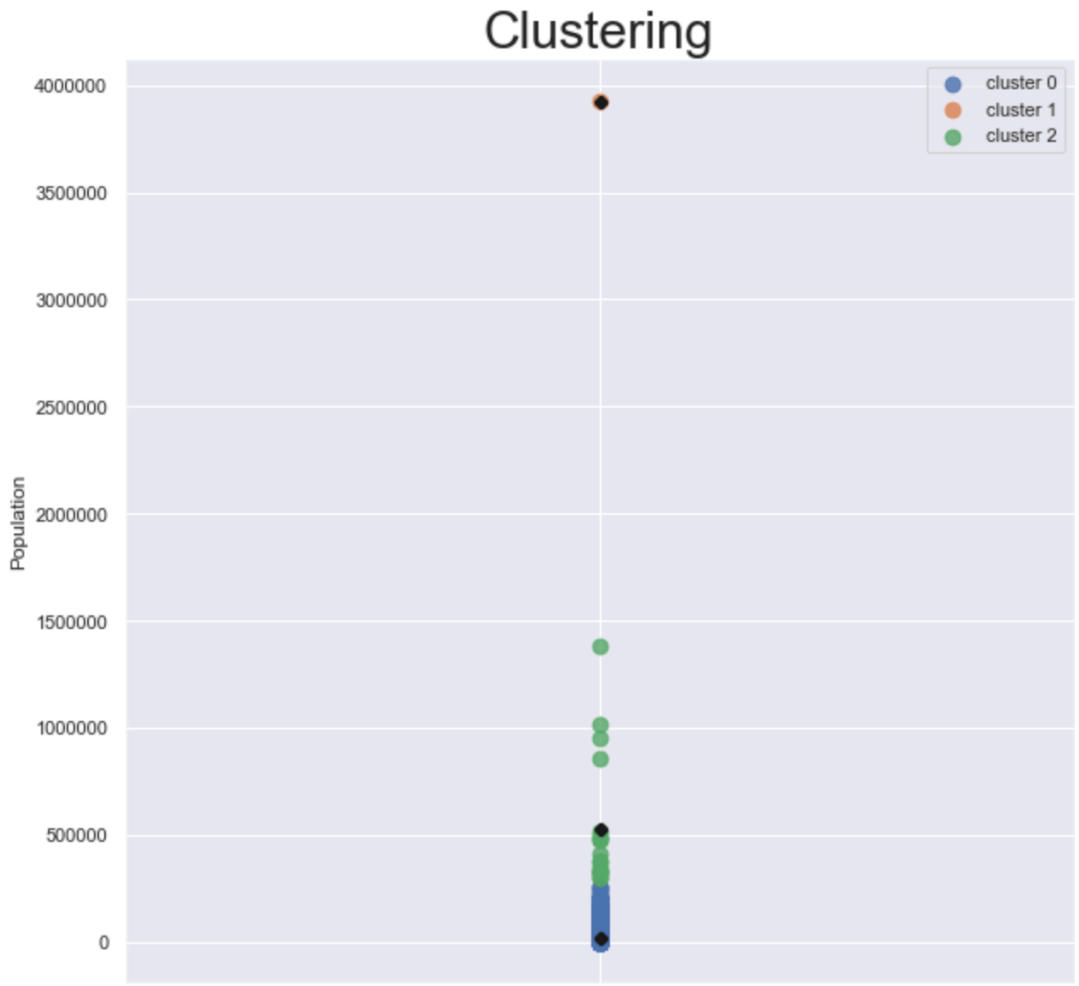
# Zichen Pan Report on excel assignment

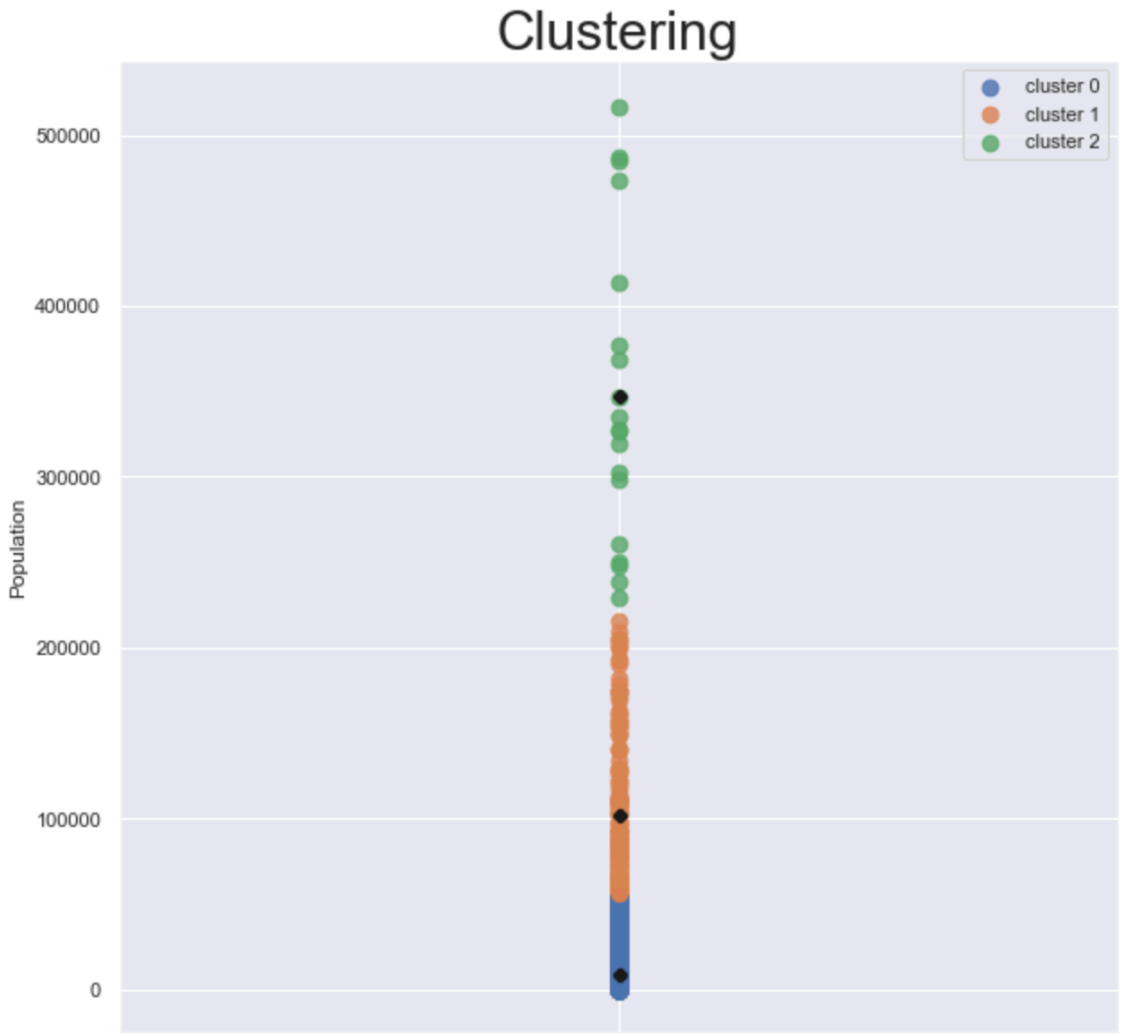
## Task 1: Identify the city types by population data

1. cluster the population data with K-means, k=3




The different colors represent different city types, and the black dot is the cluster center. It is obvious that there are 5 outliers and it must have a great impact on the clustering, especially the highest one. Even so, I saved the cluster results as Population.xslx for back up. In the following report, the five outliers are called ‘extreme big cities’.

2. To eliminate the impact of outliers, we drop out the extreme big cities and redo the clustering with K-means. This time we have got a more reasonable result.




The symbols and legends are the same as previous picture.

We add the extreme big cities to the big city cluster and get the following result:

| City Type | Small | Medium | Big |
|---|---|---|---|
| Number | 1290 | 168 | 24 |

It is reasonable. And we will use this cluster results for the next task.

**Task 2: Average data after grouping by city type and real estate type**

1. The idea is basically to join the two raw tables: Population and Raw real estate data. But I discover that the name of city in Population dataset is in lowercase while it is in uppercase. Also, some of the city name in Population have a ' , CA ' tail. These are two obstacles in the way of joining tables. Thus we deal with them first. And the result is shown below:

Original Population data:

| | Name | Population | City Type |
|---|---|---|---|
| 0 | Acalanes Ridge | 1137 | small |
| 1 | Acampo | 341 | small |
| 2 | Acton, CA | 7596 | small |
| 3 | Adelanto | 32728 | small |
| 4 | Adin-Lookout | 789 | small |

Population data for joining tables:

| | Name | Population | City Type |
|---|---|---|---|
| 0 | ACALANES RIDGE | 1137 | small |
| 1 | ACAMPO | 341 | small |
| 2 | ACTON | 7596 | small |
| 3 | ADELANTO | 32728 | small |
| 4 | ADIN-LOOKOUT | 789 | small |

2. Left join table Population and Raw real estate data, we get the results as below:

| | street | city | zip | state | beds | baths | sq__ft | type | sale_date | price | latitude | longitude | Name | City Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3526 HIGH ST | SACRAMENTO | 95838 | CA | 2 | 1 | 836 | Residential | Wed May 21 00:00:00 EDT 2008 | 59222 | 38.631913 | -121.434879 | SACRAMENTO | big |
| 1 | 51 OMAHA CT | SACRAMENTO | 95823 | CA | 3 | 1 | 1167 | Residential | Wed May 21 00:00:00 EDT 2008 | 68212 | 38.478902 | -121.431028 | SACRAMENTO | big |
| 2 | 2796 BRANCH ST | SACRAMENTO | 95815 | CA | 2 | 1 | 796 | Residential | Wed May 21 00:00:00 EDT 2008 | 68880 | 38.618305 | -121.443839 | SACRAMENTO | big |
| 3 | 2805 JANETTE WAY | SACRAMENTO | 95815 | CA | 2 | 1 | 852 | Residential | Wed May 21 00:00:00 EDT 2008 | 69307 | 38.616835 | -121.439146 | SACRAMENTO | big |
| 4 | 6001 MCMAHON DR | SACRAMENTO | 95824 | CA | 2 | 1 | 797 | Residential | Wed May 21 00:00:00 EDT 2008 | 81900 | 38.519470 | -121.435768 | SACRAMENTO | big |

3. But when I check the null values, there are five rows with null values, which means the city name does not appear in the population dataset, so we drop the rows out. Also, we drop the duplicate city name to save space. And we get the result:

| | street | city | zip | state | beds | baths | sq__ft | type | sale_date | price | latitude | longitude | City Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3526 HIGH ST | SACRAMENTO | 95838 | CA | 2 | 1 | 836 | Residential | Wed May 21 00:00:00 EDT 2008 | 59222 | 38.631913 | -121.434879 | big |
| 1 | 51 OMAHA CT | SACRAMENTO | 95823 | CA | 3 | 1 | 1167 | Residential | Wed May 21 00:00:00 EDT 2008 | 68212 | 38.478902 | -121.431028 | big |
| 2 | 2796 BRANCH ST | SACRAMENTO | 95815 | CA | 2 | 1 | 796 | Residential | Wed May 21 00:00:00 EDT 2008 | 68880 | 38.618305 | -121.443839 | big |
| 3 | 2805 JANETTE WAY | SACRAMENTO | 95815 | CA | 2 | 1 | 852 | Residential | Wed May 21 00:00:00 EDT 2008 | 69307 | 38.616835 | -121.439146 | big |
| 4 | 6001 MCMAHON DR | SACRAMENTO | 95824 | CA | 2 | 1 | 797 | Residential | Wed May 21 00:00:00 EDT 2008 | 81900 | 38.519470 | -121.435768 | big |

4. Group by city type and real estate type and calculate the average:

| | | beds | sq__ft | price |
|---|---|---|---|---|
| City Type | type | | | |
| big | Condo | 1.703704 | 871.629630 | 137690.703704 |
| | Multi-Family | 4.600000 | 2118.300000 | 214189.700000 |
| | Residential | 3.077114 | 1411.350746 | 201359.584577 |
| medium | Condo | 1.900000 | 722.200000 | 156214.450000 |
| | Multi-Family | 5.000000 | 2233.500000 | 246027.000000 |
| | Residential | 3.186047 | 1575.108527 | 296489.941860 |
| small | Condo | 0.857143 | 440.714286 | 180357.142857 |
| | Multi-Family | 2.000000 | 960.000000 | 285000.000000 |
| | Residential | 2.567460 | 979.412698 | 238943.464286 |
| | Unkown | 0.000000 | 0.000000 | 275000.000000 |

5. I find a strange classification as unknown, so I check the number of unknown:

```
In [48]: len(new_info.loc[new_info['type'] == 'Unkown'])
Out[48]: 1
```

Since it only contains one row, I just ignore it.
Save the result to Average Data.xslx.

6. I do the following operations directly in excel:
calculate the average price per bed and price per square footage.
The result is as follows:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | City Type | type | beds | sq__ft | price | price/bed | price/sq__ft |
| 2 | big | Condo | 1.703704 | 871.6296 | 137690.7 | 80818.46 | 157.9693 |
| 3 | | Multi-Family | 4.6 | 2118.3 | 214189.7 | 46562.98 | 101.114 |
| 4 | | Residential | 3.077114 | 1411.351 | 201359.6 | 65437.8 | 142.6715 |
| 5 | medium | Condo | 1.9 | 722.2 | 156214.5 | 82218.13 | 216.3036 |
| 6 | | Multi-Family | 5 | 2233.5 | 246027 | 49205.4 | 110.1531 |
| 7 | | Residential | 3.186047 | 1575.109 | 296489.9 | 93058.89 | 188.2346 |
| 8 | small | Condo | 0.857143 | 440.7143 | 180357.1 | 210416.7 | 409.2382 |
| 9 | | Multi-Family | 2 | 960 | 285000 | 142500 | 296.875 |
| 10 | | Residential | 2.56746 | 979.4127 | 238943.5 | 93066.08 | 243.9661 |
| 11 | | Unkown | 0 | 0 | 275000 | | |
| 12 | | | | | | | |

Assumption: one bed in one bedroom, thus the average beds are the same as average bedrooms.

Conclusions:
(I just ignore some common discoveries and focus on the discoveries that seems counter-intuitive)
1. Comparison between city types:
The average bedrooms in medium cities is slightly more than that of big cities, for each type of real estate. Generally the price in small cities are highest and much higher than the counterpart of medium and big cities.
2. Comparison between real estate types:
Multi-Family type has more bedrooms than Condo and Residential, with Condo the least. The area of Multi-family and Residential is much more than that of Condo. For price per square footage, Condo is the most expensive. Residential in big and medium cities is more expensive than Multi-Family, while it is the opposite situation in small cities.

Suggestions for real estate agent:
If a condo is to be sold, the target customers should be those with relatively high and stable incomes because condo is the most expensive real estate type among all.
Generally speaking, the real estate price in small cities is high, irrespective of the total price or unit price, which reflects the trend that real estate in small cities has the largest market and is most popular among citizens. The relatively low price in big cities indicates the cooling down of real estate in these areas.