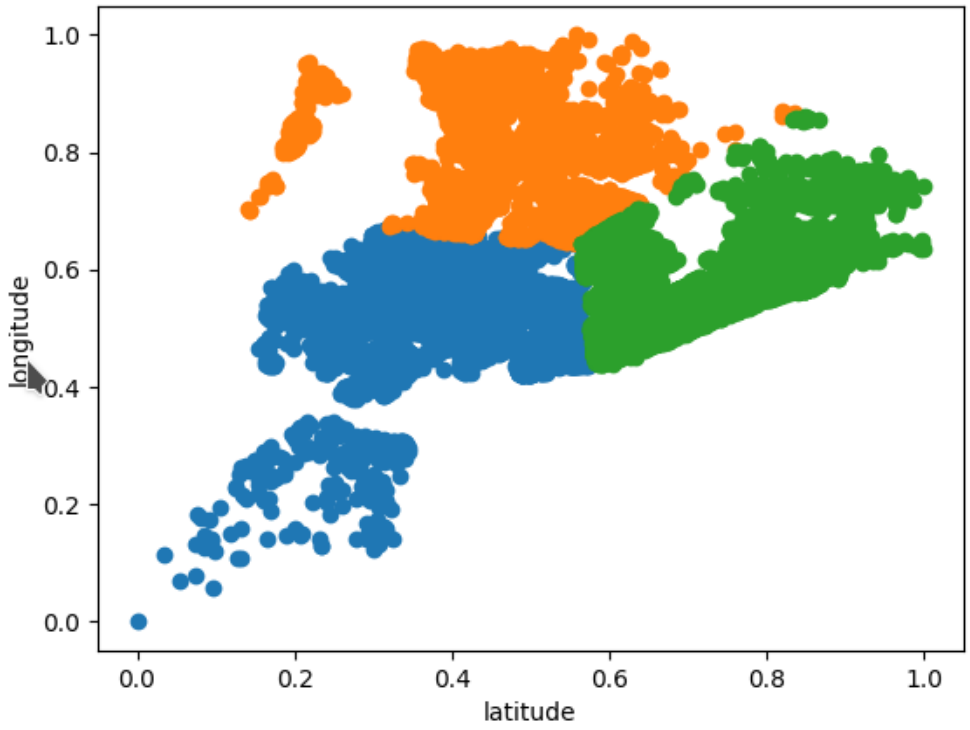
HW 1 Report

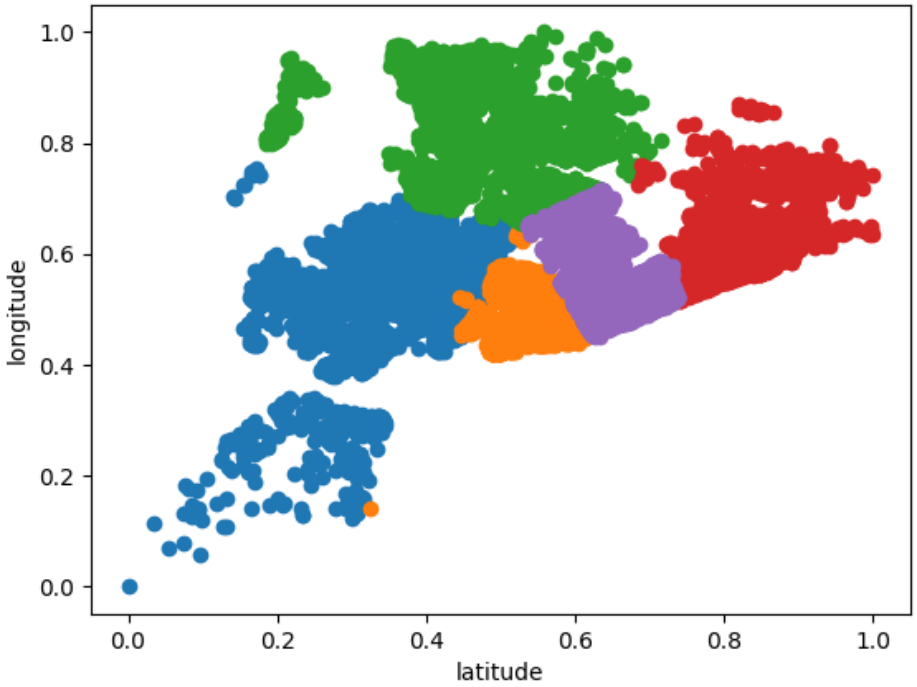
Zhou Shen
10/12/2020

1. Answer code in k_means_clustering.py

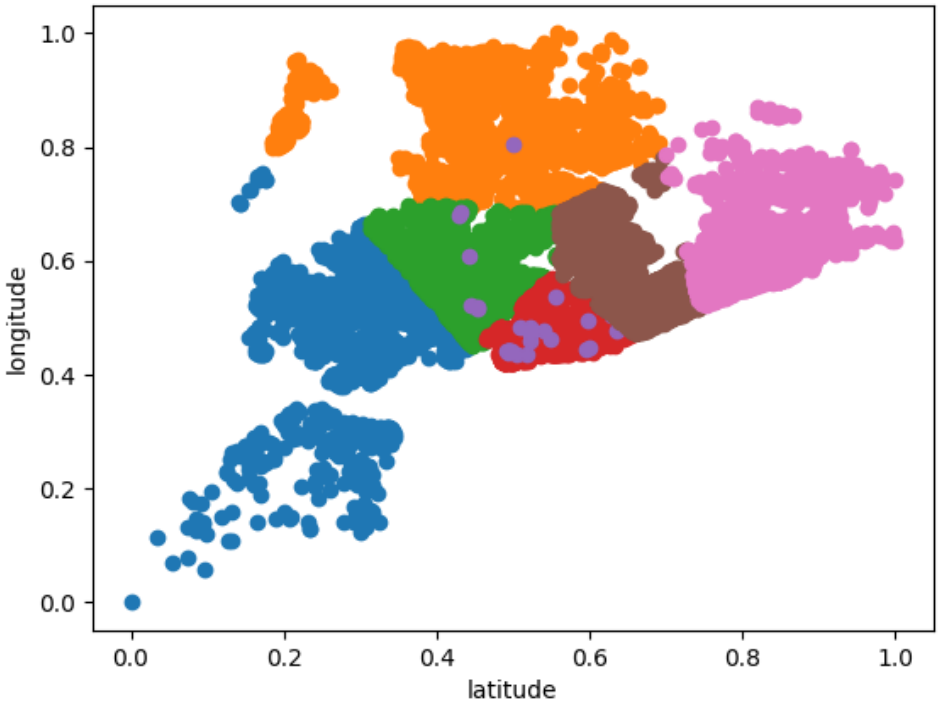
2. Working with the Algorithms, answer code in cluster_zhoushen.py and kmeans_zhou.py files
a) I changed only 1 parameter per time and check how this parameter affect the price.
1. For kmeans++
when k = 3, cluster will look like following graph:



when k = 5, cluster will look like following graph:

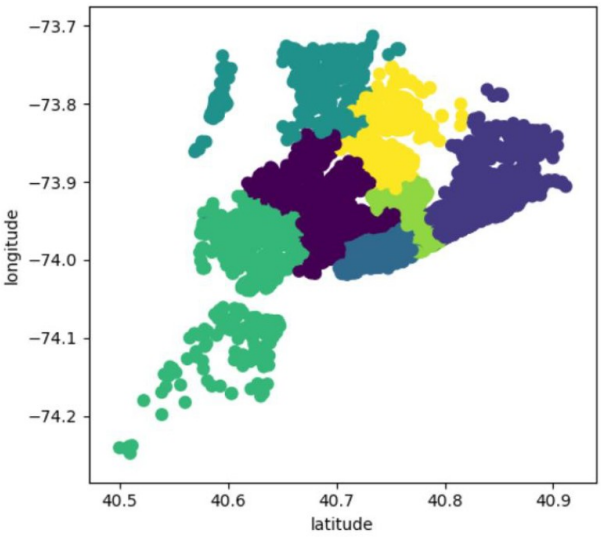


when k = 7, cluster will look like following graph:

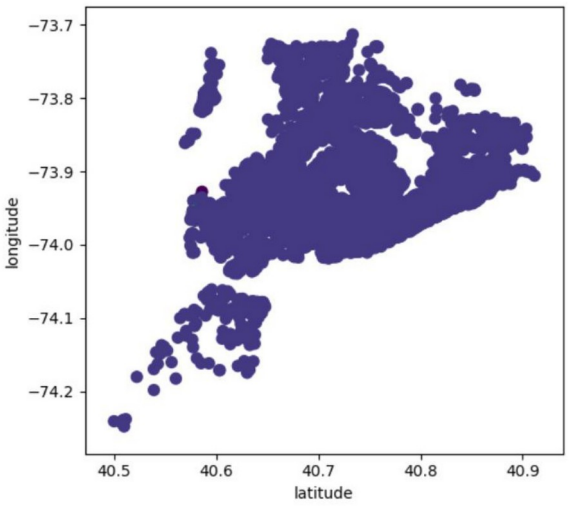


Based on above graphs, I believe when k = 5 is more correct. Since when k = 3, it is vague; when k = 7, it is too precise with several detailed points.

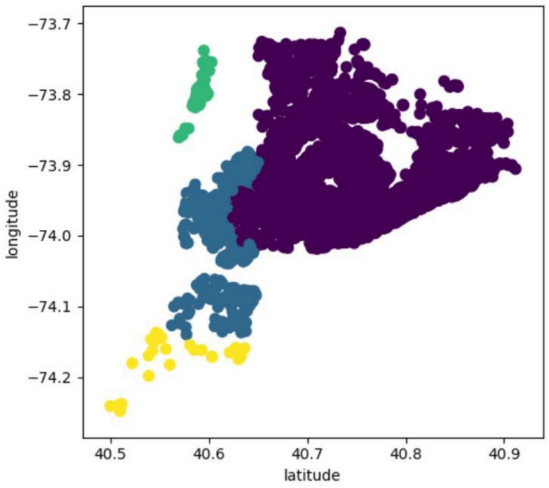
2. Hierarchical
If I chose all data as input, it will run out of memory, so I have to extract 3000 data of them.
When k = 5, linkage is ‘ward’:



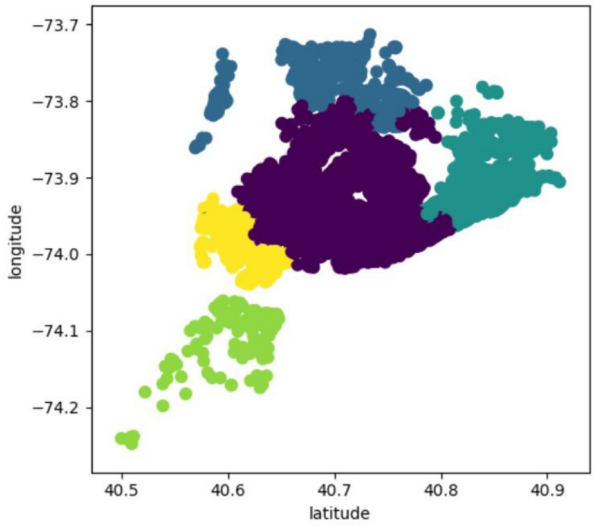
when k = 5, linkage is ‘single’:



when k = 5, linkage is ‘average’:

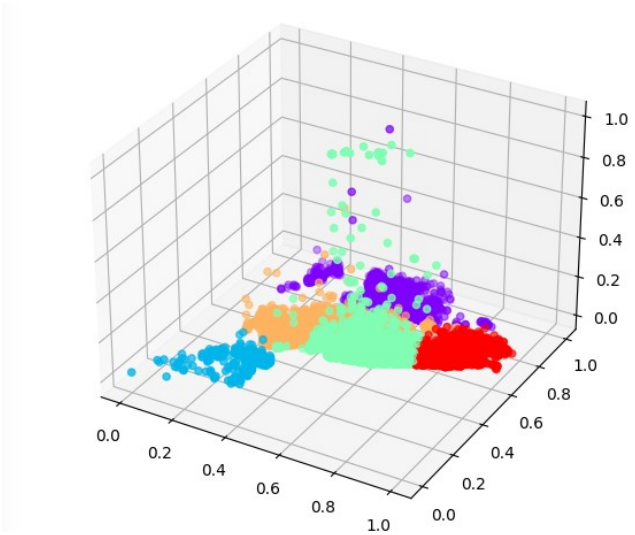


when k = 5, linkage is ‘complete’:

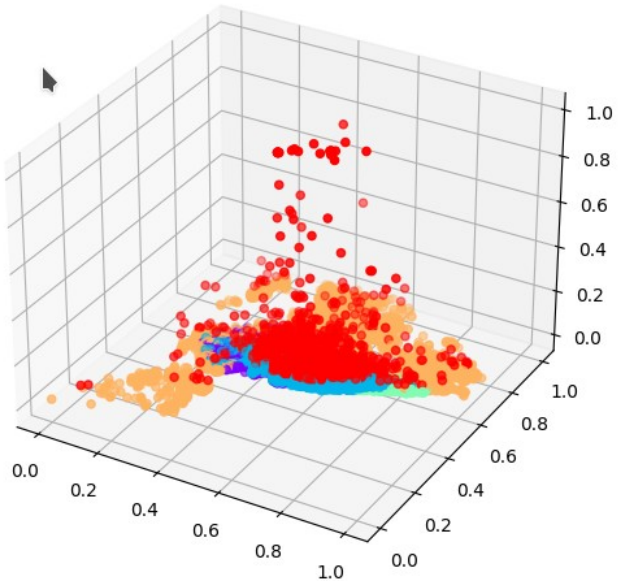


From above graphs, we can find that when linkage is ‘ward’ is better than others since it shows each area clearly.

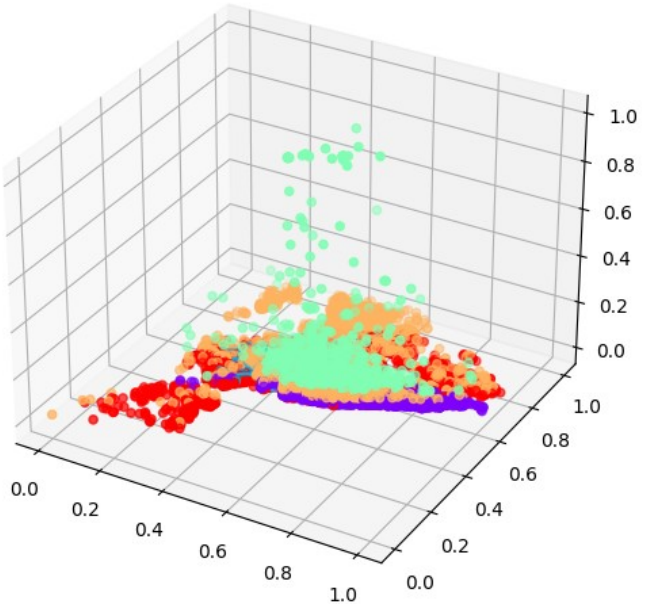
3. GMM
when k = 5, max iteration is 10000, covariance type is ‘tied’:



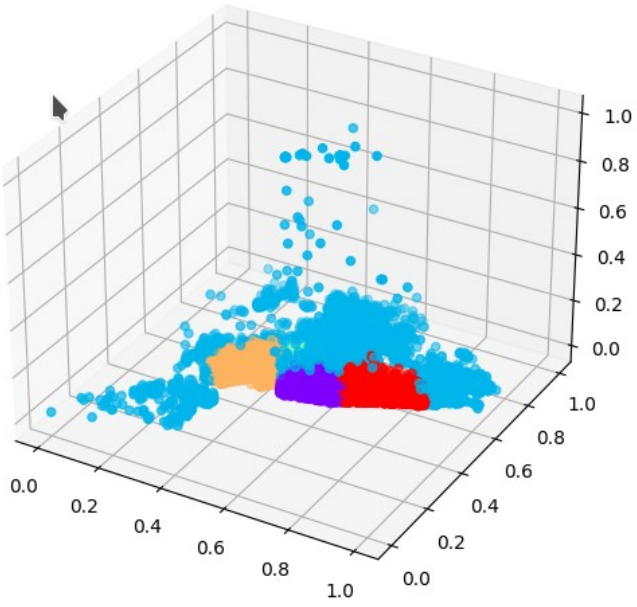
when k = 5, max iteration is 10000, covariance type is ‘diag’:



when k = 5, max iteration is 10000, covariance type is ‘full’:



when k = 5, max iteration is 10000, covariance type is ‘spherical’:

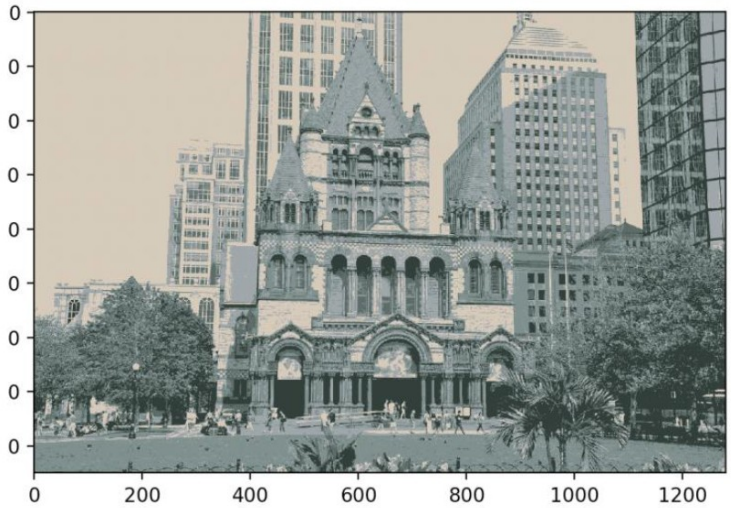


From above diagrams, we can see that ‘tied’ is better than others since areas are more clear.

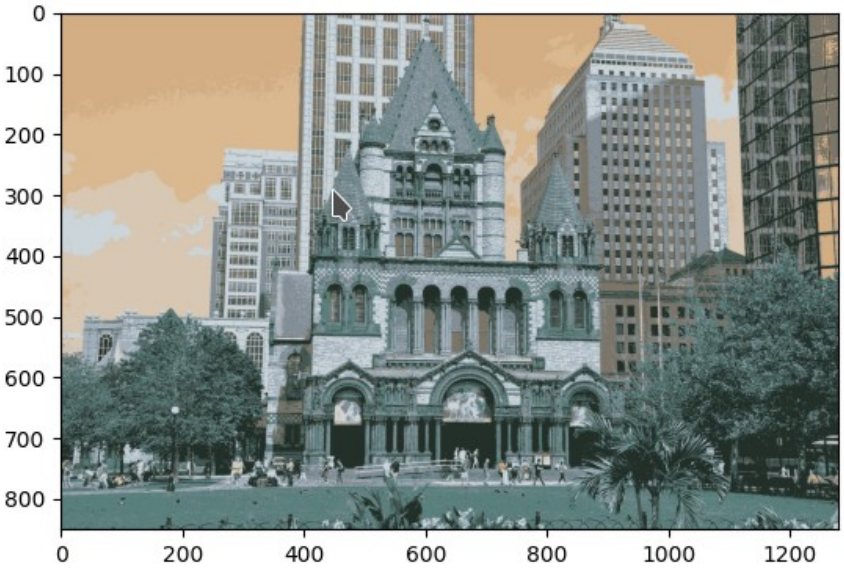
b) Pros and cons
1) K means++:
Pros: simple to implement (basic principle to follow), fast convergence speed.
Cons: user has to compare and decide number of clusters, sensitive to outliers
2) Hierarchical:
Pros: similarity of distance and rules is easy to define and only has few restrictions
Cons: Singular values have great impact, calculation complexity is too high.
3) GMM:
Pros: a point can belong to multiple clusters; cluster shapes are flexible
Cons: will fail to work if the conditionality of the problem is too high

3. Data visualization
a) No, since there is no different color using to represent different price levels in these areas and there is also no exact number about the price of house showing.
b) I did it in above question for each cluster.
c) For k means ++:
`[122.0352269120914, 124.51134710503347, 109.59904999221305, 221.16670145411535, 111.72493224932249, 117.02489914662724, 96.24192991340012]`
For Hierarchical:
`[121.0132232123115, 108.12441212030911, 226.12009317882097, 102.76331229876112, 108.00871226587213, 144.24461297723441, 91.12398784324776]`
For GMM:
`[185.71424281102185, 136.48855165069223, 210.80428479381445, 182.534061458719, 125.44261952087038, 186.15523379223132, 148.944370721327]`
d) bonus point.
e) Based on my understanding of NYC actual price, I believe that GMM’s price is more close.

4. Image Manipulation
Bigger the number of k is, slower the algorithm will run, k=10 spent a lot of time.
when k = 2:



when k = 4:



when k = 10: