**Taipei MRT Clustering**

BU MET CS677 Final Project

Jen-Yin Chao

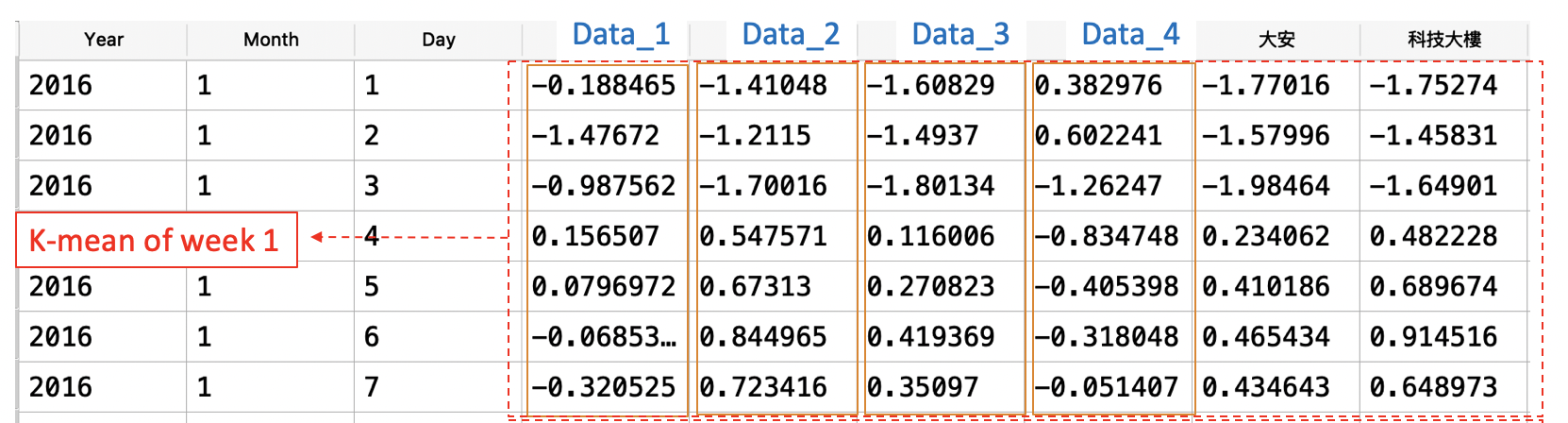
* Introduction

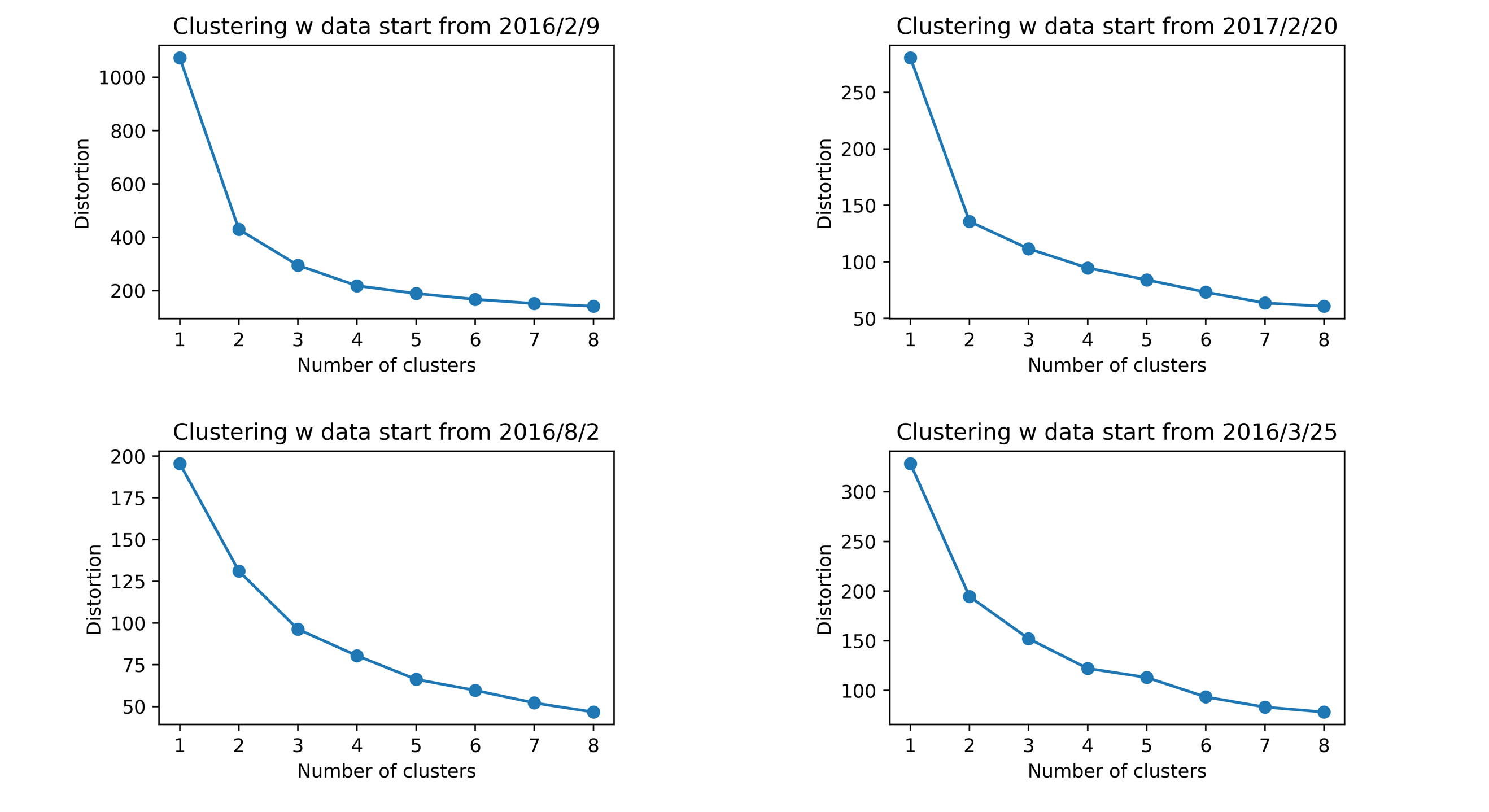
Taipei MRT, i.e. mass repaid transit, is the biggest metro service system in Taiwan. There are five main line with 108 stops and the cumulative rider-ship has been more than 10 billion until now, Apr 2019, while several lines are still under construction.

People take MRT with various purpose, includes going to work, school, traveling, so on and so forth, so some stops are crowed all the times, like Taipei Main Station, which is the biggest transit stop in Taipei, while some may be busy only at some specific time period, like Taipei Zoo which might be busy in the weekend and Software Park which might be busy only during the weekdays. Accordingly, the pattern of each stop’s weekly rider-ship might be different, and theoretically I can separate them into different groups relying on it, which is also the goal that I want to know in this project.

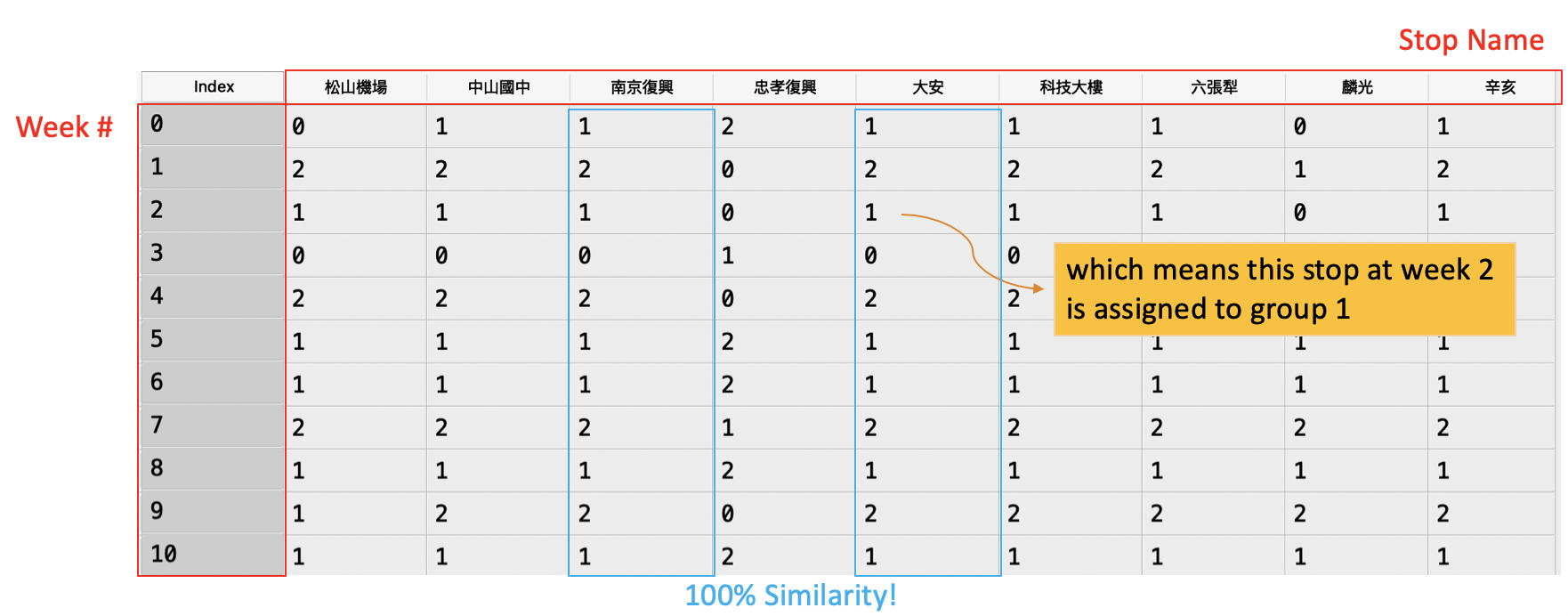
* Approach

After standardizing the rider-ship data, I’ll set the first 7 days row data as new data\_1 of stop 1, also do the same thing to the other stops, then put them in k-means module that I can get the clustering outcome of week 1, which is illustrated in the following picture, and do the same thing to the following weeks.





By choosing the number of clustering 3, depending on the distortion plot, I’ll get the following clustering outcome table, with row name representing the number of week, column name representing stop names, and the values means the clustering outcome of the each stop at each week. And then, I’ll compare the outcome with other stops to see whether they are assigned in the same group and calculate how many times does it happen, which I call it similarity. Take the two stops in the blue box in the following picture for example, they are in the same group all the time that I’ll define they have 100% similarity.



By plotting those paired similarity on the heat map, the picture will look like the following one. The darker the color is, the higher probability the stops are assigned in the same group. And then, I separate them into 3 groups and dive into it to what kind of stops are much more easier to be assigned together.



* Result

I found out that the transit station, tourist attraction, shopping center, and night market are much easier to be assigned in either group 2 or 3, and the stops in group 1 are mostly located at the outer bound that people are usually use it to commute to downtown. As the result, I can tell that those special stops can really be separated from the other stops by their unique weekly rider-ship pattern.



* Conclusion

Different kind of MRT stops do really have different weekly rider-ship pattern, and especially for those special stops, includes transit station, tourist attraction, and shopping center (night market), can really be distinguished by the weekly rider-ship with k-means clustering method.

* Instruction

Main.py is the only script to run. After running it, it will load the csv file from the rawdata file and do the analysis twice, dealing with entry rider-ship data first and then exit rider-ship data. Both distortion plot and similarity heat map will be saved at local file, and the list of stop names of grouping at the last step will be directly shown on the scream.