- *20280071*
- *Nguyễn Tiến Phong*

BÁO CÁO

KHAI THÁC DỮ LIỆU – TUẦN 6

Thuật toán K-Medians

- Đọc dữ liệu
- Xác định k cụm và khởi tạo các tâm :
    - Phân làm 6 cụm có các tâm :

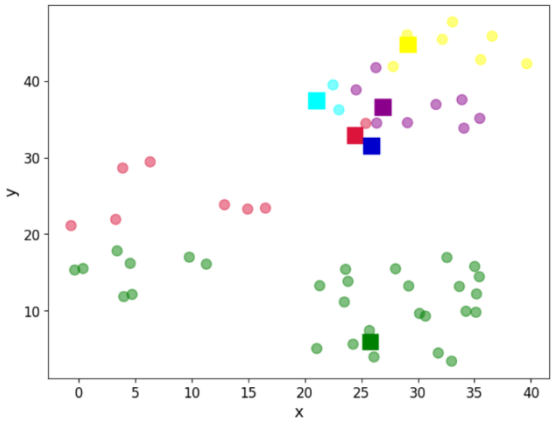
	x	y
0	24.412	32.932
5	25.893	31.515
36	26.878	36.609
45	29.101	44.781
13	25.768	5.967
54	21.034	37.463

- Tính khoảng cách theo công thức Manhattan: $\sum_{i=1}^{k} |x_i - y_i|$
  Và tính sai số phát sinh

```
Error for centroid 0: 6.14
Error for centroid 1: 6.08
Error for centroid 2: 0.00
Error for centroid 3: 10.39
Error for centroid 4: 31.75
Error for centroid 5: 6.70
```

- Gán giá trị các tâm, thêm cột gán tâm và sai số phát sinh :

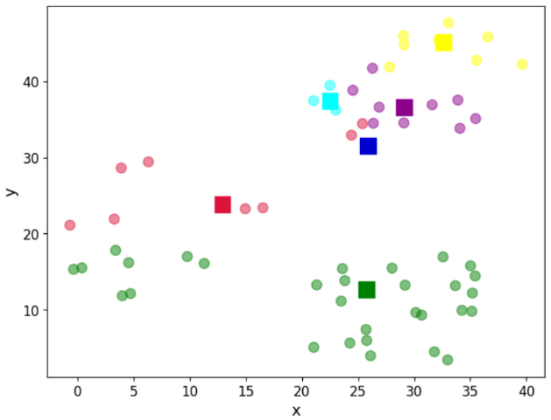
	x	y	centroid	error
0	24.412	32.932	0	0.000
1	35.190	12.189	4	15.644
2	26.288	41.718	2	5.699
3	0.376	15.506	4	34.931
4	26.116	3.963	4	2.352



- *20280071*
- *Nguyễn Tiến Phong*

- Cập nhật vị trí của k tâm

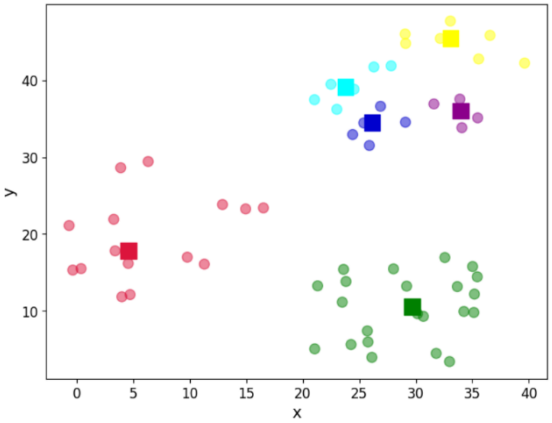
	x	y
0	12.8910	23.832
1	25.8930	31.515
2	29.0810	36.609
3	32.6155	45.101
4	25.7400	12.676
5	22.4930	37.463



Lặp lại các bước trên

- Vị trí tâm cuối cùng

	x	y
0	4.5500	17.8100
1	26.1265	34.4660
2	33.9885	36.0075
3	33.0620	45.4210
4	29.6695	10.5365
5	23.7750	39.1440



- Sử dụng ELBOW để chỉ ra số cụm tối ưu

Nhìn vào đồ thị elbow ta thấy rằng số cụm tăng từ 6 đến 7 thì giá trị sai số giảm chậm hơn so với số cụm tăng từ 3 đến 4.

Số cụm tối ưu là 6.