**Bùi Quốc Bảo (1412034)**

**Hà Gia Phát (1412382)**

**Vũ Ngọc Minh Hoàng (1412186)**

report lab04

# 1.2: BÁO CÁO KẾT QUẢ THỰC NGHIỆM THAM SỐ K VÀ RANDOM SEED

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Random seed | | | | |  |
| k | 10 | 20 | 30 | 40 | 50 | best |
| 2 | 327.57 | 327.57 | 485.54 | 313.13 | 313.13 | 313.13 |
| 3 | 305.03 | 206.50 | 261.78 | 296.73 | 206.50 | 206.50 |
| 4 | 192.90 | 171.96 | 245.70 | 188.89 | 188.89 | 171.96 |
| 5 | 169.31 | 161.05 | 137.75 | 155.36 | 171.82 | 137.75 |
| 6 | 124.83 | 114.75 | 106.87 | 101.42 | 144.93 | 101.42 |

* Qua bảng thống kê trên ta có thể thấy được rằng số lượng cụm tốt nhất bằng 6

# 1.3: BÁO CÁO KẾT QUẢ THỰC NGHIỆM độ đo khoảng cách

* **Sử dụng khoảng cách Manhattan:**

|  |  |
| --- | --- |
| **K** | **Best** |
| 2 | 1453.29 |
| 3 | 857.50 |
| 4 | 650.90 |
| 5 | 523.81 |
| 6 | 411.74 |

**Kết luận:** Vậy số lượng cụm tốt nhất là bằng 6

* **Sử dụng khoảng cách Cosine:**

|  |  |
| --- | --- |
| **K** | **Best** |
| 2 | 5.64 |
| 3 | 3.24 |
| 4 | 2.52 |
| 5 | 1.99 |
| 6 | 1.83 |

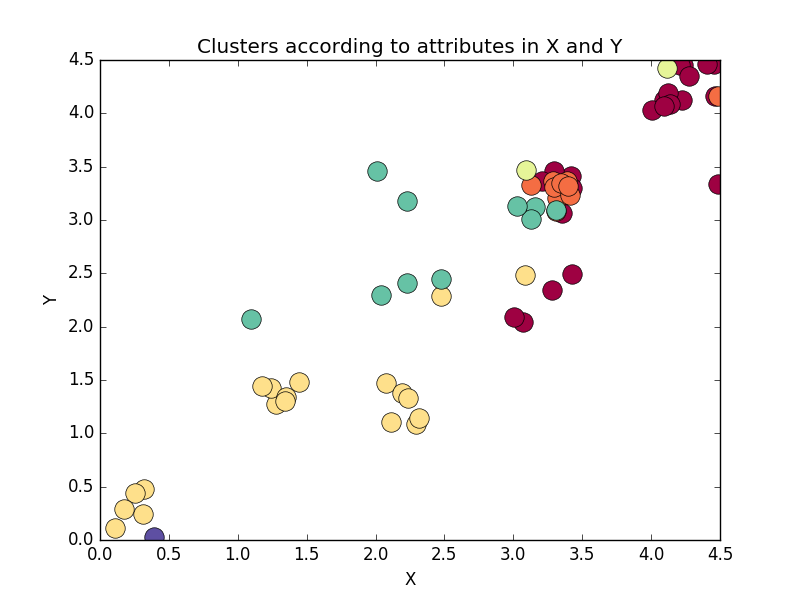
**Kết luận:** Vậy số lượng cụm tốt nhất là bằng 6

* **Nhận xét:**

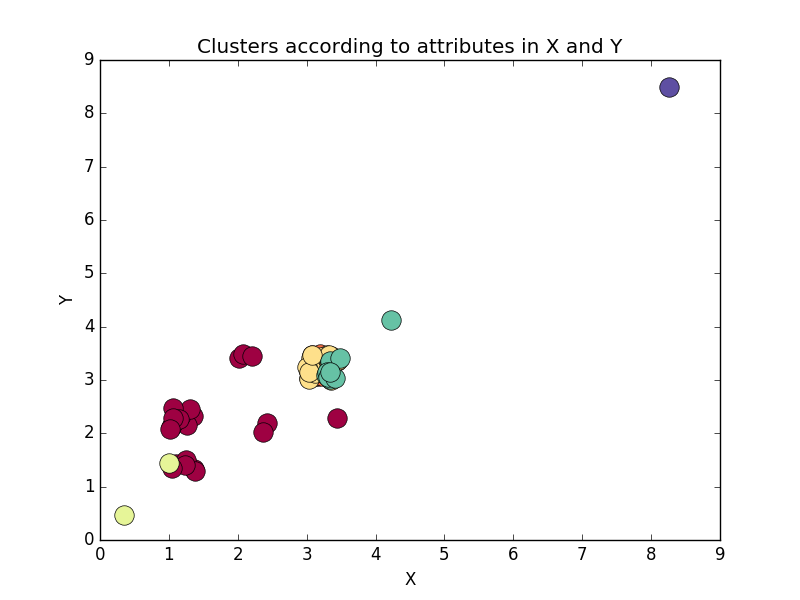
Ta thấy rằng các giá trị khác nhau. Kết quả của có vẻ đáng tin cậy nhất vì thuật toán K-Means chạy bằng cách lặp lại việc tìm các điểm gần với centroid nhất và câp nhật điểm centroid đó qua từng nhóm gom cụm mới. Thuật toán này ngầm sử dụng các công thức tính khoảng cách giữa các cặp điểm (dữ liệu và centroid) vì ta có thể thấy SSE (sum of errors) của K-Means được tính bằng khoảng cách bình phương giữa hai điểm centroid và dữ liệu. Vì thế công thức tính khoảng cách Euclidean trong gom cụm rất phù hợp cho thuật toán K-Means. Ngoài ra nguồn gốc của từ “centroid” là trong hình học Euclidean, mà không gian Euclidean thì ta nên sử dụng khoảng cách Euclidean, các khoảng cách không thuộc về Euclidean (Non-Euclidean distances) thường sẽ không phù hợp khi sử dụng trong không gian Euclidean.

# 1.4: Phân tích hiệu quả phân tách cụm của các cặp thuộc tính khác nhau

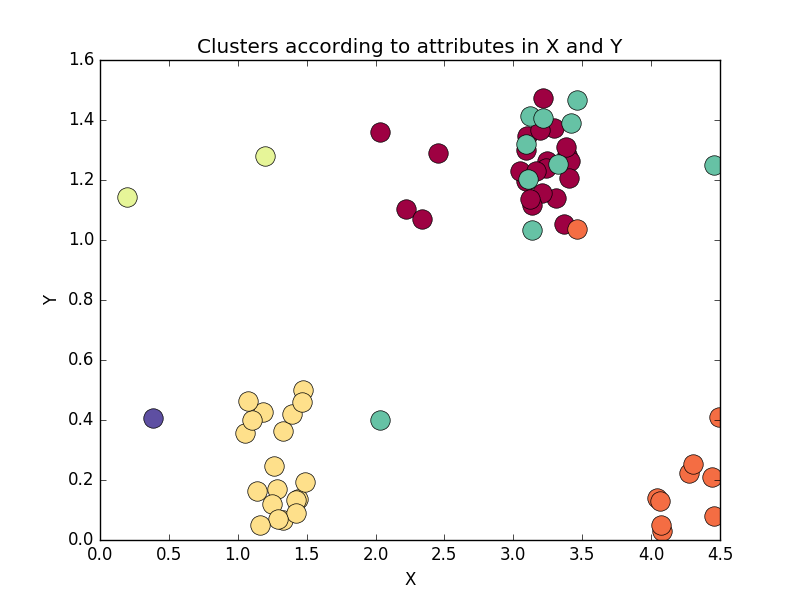
Sau khi chọn số lượng cụm tốt nhất làm dữ liệu đầu vào, ta sẽ được các hình ảnh gom cụm theo thuộc tính X và Y như sau:



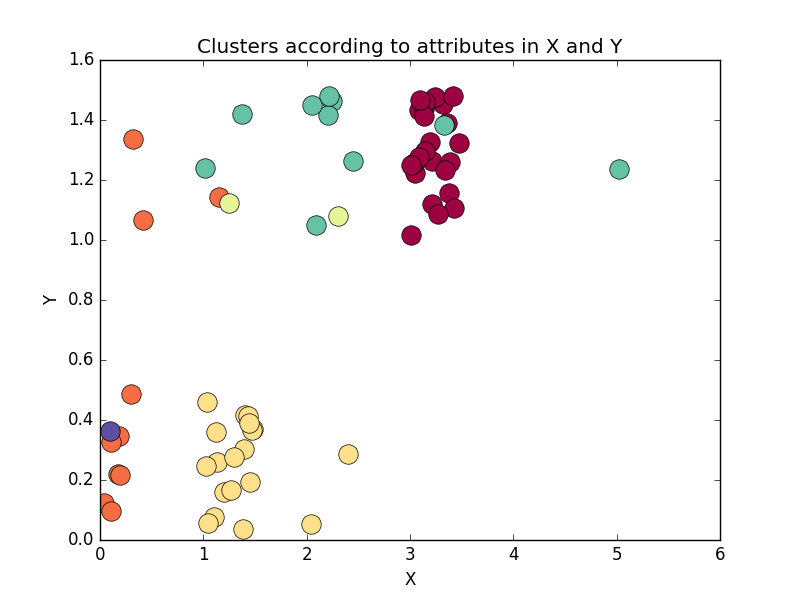
**Hình 1.4.1:** X = răng tiền hàm trên, Y = răng tiền hàm dưới



**Hình 1.4.2:** X = răng hàm trên, Y = răng hàm dưới



**Hình 1.4.3:** X = răng cửa dưới, Y = răng nanh dưới



**Hình 1.4.4:** X = răng cửa trên, Y = răng nanh trên

# 1.5: nhận xét ý nghĩa của cụm dữ liệu