**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

----------🙞🙜🕮🙞🙜----------

**KHAI THÁC DỮ LIỆU**

**Bài thực hành số 5**

***Giảng viên hướng dẫn:***

**ThS. Hà Lê Hoài Trung**

***Sinh viên thực hiện:***

**Hoàng Lê Nam – 18521120**

***Lớp:*** **IS252.L21**

🙡🙢 Tp. Hồ Chí Minh, 6/2021 🙠🙣

***Mục lục***

[1. Bài tập 1: 3](#_Toc74260579)

[2. Bài tập 2: 11](#_Toc74260580)

[3. Bài tập 3: 19](#_Toc74260581)

[4. Bài tập 4: 30](#_Toc74260582)

[5. Bài tập 5: 35](#_Toc74260583)

[6. Bài tập 6: 39](#_Toc74260584)

### Bài tập 1:

* Yêu cầu:

Table

Description automatically generated

* Giải quyết:

#### k-Means:

* Ma trận phân hoạch các điểm thuộc cụm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Vị trí 1 | Vị trí 2 | Vị trí 3 | Vị trí 4 | Vị trí 5 | Vị trí 6 | Vị trí 7 | Vị trí 8 | Vị trí 9 |
| Cụm 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cụm 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cụm 3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

* Vector trọng tâm lúc này:
* Khoảng cách Euclide từ các vị trí đến các cụm lần lượt là:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cụm 1 | Cụm 2 | Cụm 3 |
| Vị trí 1 | 0 | 150.0133 | 188.61066 |
| Vị trí 2 | 150.0133 | 0 | 38.711123 |
| Vị trí 3 | 450.0111 | 300.0033 | 261.44048 |
| Vị trí 4 | 200.0125 | 50.04998 | 11.544059 |
| Vị trí 5 | 120.0375 | 30.21589 | 68.601184 |
| Vị trí 6 | 50.24938 | 100.065 | 138.58616 |
| Vị trí 7 | 400.025 | 250.032 | 211.42978 |
| Vị trí 8 | 100.1249 | 50.16971 | 88.579954 |
| Vị trí 9 | 6.082763 | 150.1233 | 188.59361 |

* Ma trận phân hoạch các điểm thuộc cụm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Vị trí 1 | Vị trí 2 | Vị trí 3 | Vị trí 4 | Vị trí 5 | Vị trí 6 | Vị trí 7 | Vị trí 8 | Vị trí 9 |
| Cụm 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Cụm 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Cụm 3 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |

* Vector trọng tâm lúc này là:
* Khoảng cách Euclide từ các vị trí đến các cụm lần lượt là:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cụm 1 | Cụm 2 | Cụm 3 |
| Vị trí 1 | 17.01633 | 123.3667 | 350.01349 |
| Vị trí 2 | 133.3675 | 26.77063 | 200.01361 |
| Vị trí 3 | 433.34 | 326.6721 | 100.01389 |
| Vị trí 4 | 183.3373 | 76.67029 | 150.0037 |
| Vị trí 5 | 103.3468 | 3.785939 | 230.00966 |
| Vị trí 6 | 33.4149 | 73.37347 | 300.00741 |
| Vị trí 7 | 383.3348 | 276.6719 | 50.02777 |
| Vị trí 8 | 83.35 | 23.43075 | 250.00756 |
| Vị trí 9 | 16.94763 | 123.3896 | 350.02063 |

* Ma trận phân hoạch các điểm thuộc cụm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Vị trí 1 | Vị trí 2 | Vị trí 3 | Vị trí 4 | Vị trí 5 | Vị trí 6 | Vị trí 7 | Vị trí 8 | Vị trí 9 |
| Cụm 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Cụm 2 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| Cụm 3 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

* Vector trọng tâm lúc này là:
* Khoảng cách Euclide từ các vị trí đến các cụm lần lượt là:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cụm 1 | Cụm 2 | Cụm 3 |
| Vị trí 1 | 17.01633 | 142.5257 | 425.01471 |
| Vị trí 2 | 133.3675 | 7.846177 | 275.01182 |
| Vị trí 3 | 433.34 | 307.5062 | 25.04995 |
| Vị trí 4 | 183.3373 | 57.50272 | 225.00556 |
| Vị trí 5 | 103.3468 | 22.56241 | 305.01066 |
| Vị trí 6 | 33.4149 | 92.53682 | 375.00333 |
| Vị trí 7 | 383.3348 | 257.5064 | 25.04995 |
| Vị trí 8 | 83.35 | 42.56245 | 325.00385 |
| Vị trí 9 | 16.94763 | 142.5502 | 425.01706 |

* Ma trận phân hoạch các điểm thuộc cụm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Vị trí 1 | Vị trí 2 | Vị trí 3 | Vị trí 4 | Vị trí 5 | Vị trí 6 | Vị trí 7 | Vị trí 8 | Vị trí 9 |
| Cụm 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Cụm 2 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| Cụm 3 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

* Ta thấy: ma trận phân hoạch lần 3 giống với ma trận phân hoạch lần 2 vậy nên dừng thuộc toán, kết quả là:
* Vị trí 1, 6 và 9 thuộc cụm 1.
* Vị trí 2, 4, 5 và 8 thuộc cụm 2.
* Vị trí 3 và 7 thuộc cụm 3.

#### Kohomen:

(với epochs = 10; R = 0; α = 0.8)

* Khởi tạo giá trị của các vector trọng số:
* Lần lặp 1:
* ***Xét vị trí 1***:

Khoảng cách từ vị trí 1 tới là:

Tương tự như trên, 165.2849; 205.0285. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 1* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 2***:

Khoảng cách từ vị trí 2 tới , , lần lượt là: 130.0107; 15.45647; 55.09882. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 2* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 3***:

Khoảng cách từ vị trí 3 tới , , lần lượt là: 430.0087; 296.9519; 245.0165. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 3* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 4***:

Khoảng cách từ vị trí 4 tới , , lần lượt là: 180.0092; 46.98591; 201.0075. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 4* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 5***:

Khoảng cách từ vị trí 5 tới , , lần lượt là: 100.0371; 70.63237; 281.0167. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 5* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 6***:

Khoảng cách từ vị trí 6 tới , , lần lượt là: 30.3404; 84.20628; 351.0061. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 6* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 7***:

Khoảng cách từ vị trí 7 tới , , lần lượt là: 356.0056; 265.8867; 2.8219. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 7* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 8***:

Khoảng cách từ vị trí 8 tới , , lần lượt là: 56.02183132; 34.2564; 300.2017. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 8* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét vị trí 9***:

Khoảng cách từ vị trí 9 tới , , lần lượt là: 44.2006; 106.8579; 400.2094. Vậy nơron có trọng số là nơron có khoảng cách đến *Vị trí 9* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* Tương tự, lần lặp 2 với từng vị trí thu được vec tor trọng số:
* Tương tự, lần lặp 3 với từng vị trí thu được vec tor trọng số:
* Sau 3 lần lặp nhận xét:
* Các vị trí đã được gom vào vị trí là không thay đổi qua nhiều lần lặp.
* Dừng thuộc toán.
* Vậy kết quả là:
* Vị trí 1, 6 và 9 thuộc Vector 1.
* Vị trí 2, 4, 5 và 8 thuộc Vector 2.
* Vị trí 3 và 7 thuộc Vector 3.

#### Nhận xét:

* Từ kết quả của a, b có thể thấy ở 2 thuật toán các vị trí được gom về một cụm đều giống nhau. Các cụm tương ứng là:

|  |  |  |
| --- | --- | --- |
| k-Means | Mạng Kohomen | Vị trí |
| *Cụm 1* | ***Cụm 1*** | ***1, 6, 9*** |
| *Cụm 2* | ***Cụm 2*** | ***2, 4, 5, 8*** |
| *Cụm 3* | ***Cụm 3*** | ***3, 7*** |

#### Dự báo vào cụm nào:

* Từ kết quả của k-Means, ta có:
* Với vị trí tiềm năng mới(11, 6, 450):

+ =

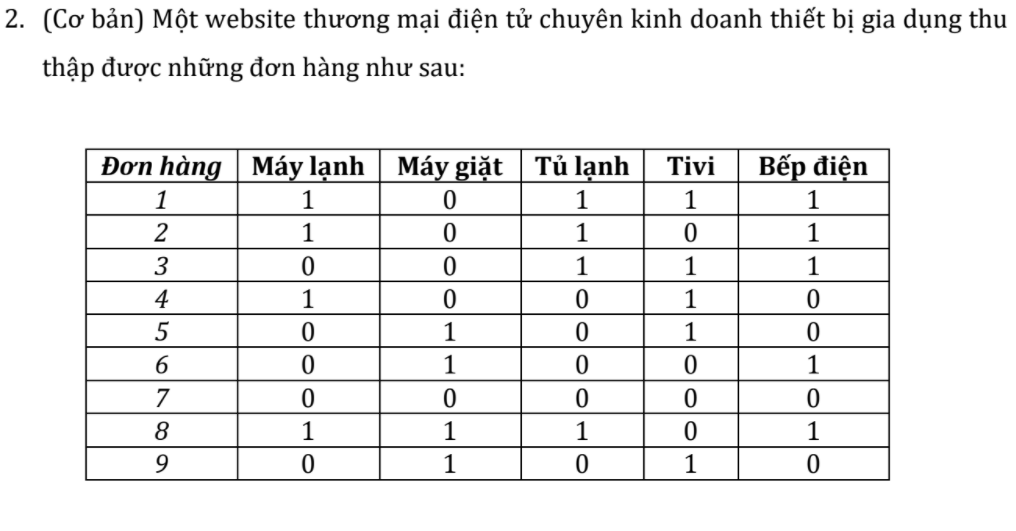
+ =

+ =

Thấy rằng: Khoảng cách từ vị trí tiềm năng đến các cửa hàng thì, cửa hàng 2 phục vụ cho vị trí là tốt nhất vì

### Bài tập 2:

* Yêu cầu:



* Giải quyết:

#### k-Means:

* Ma trận phân hoạch các điểm thuộc nhóm khởi tạo:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ĐH 1 | ĐH 2 | ĐH 3 | ĐH 4 | ĐH 5 | ĐH 6 | ĐH 7 | ĐH 8 | ĐH 9 |
| Nhóm 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nhóm 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nhóm 3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

* Vector trọng tâm:
* Khoảng cách Euclide từ các vị trí đến các nhóm lần lượt là:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nhóm 1 | Nhóm 2 | Nhóm 3 |
| Đơn hàng 1 | 0 | 1 | 1.36277 |
| Đơn hàng 2 | 1 | 0 | 1.414214 |
| Đơn hàng 3 | 1 | 1.414214 | 1.195229 |
| Đơn hàng 4 | 1.414214 | 1.732051 | 1.133893 |
| Đơn hàng 5 | 2 | 2.236068 | 0.845154 |
| Đơn hàng 6 | 2 | 1.732051 | 1 |
| Đơn hàng 7 | 2 | 1.732051 | 1 |
| Đơn hàng 8 | 1.414214 | 1 | 1.36277 |
| Đơn hàng 9 | 2 | 2.236068 | 0.845154 |

* Ma trận phân hoạch các điểm thuộc nhóm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ĐH 1 | ĐH 2 | ĐH 3 | ĐH 4 | ĐH 5 | ĐH 6 | ĐH 7 | ĐH 8 | ĐH 9 |
| Nhóm 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nhóm 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Nhóm 3 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |

* Vector trọng tâm lúc này:
* Khoảng cách Euclide từ các vị trí đến các nhóm lần lượt là:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nhóm 1 | Nhóm 2 | Nhóm 3 |
| Đơn hàng 1 | 0.5 | 1.118034 | 1.67332 |
| Đơn hàng 2 | 1.118034 | 0.5 | 1.732051 |
| Đơn hàng 3 | 0.5 | 1.5 | 1.48324 |
| Đơn hàng 4 | 1.5 | 1.802776 | 1.095445 |
| Đơn hàng 5 | 1.802776 | 2.061553 | 0.632456 |
| Đơn hàng 6 | 1.802776 | 1.5 | 1.095445 |
| Đơn hàng 7 | 1.802776 | 1.802776 | 0.894427 |
| Đơn hàng 8 | 1.5 | 0.5 | 1.67332 |
| Đơn hàng 9 | 1.802776 | 2.061553 | 0.632456 |

* Ma trận phân hoạch các điểm thuộc nhóm:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ĐH 1 | ĐH 2 | ĐH 3 | ĐH 4 | ĐH 5 | ĐH 6 | ĐH 7 | ĐH 8 | ĐH 9 |
| Nhóm 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nhóm 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Nhóm 3 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |

* Ta thấy: ma trận phân hoạch lần 2 giống với ma trận phân hoạch lần 1 vậy nên dừng thuộc toán, kết quả là:
* Vị trí 1 và 3 thuộc nhóm 1.
* Vị trí 2 và 8 thuộc nhóm 2.
* Vị trí 4, 5, 6, 7 và 9 thuộc nhóm 3.

#### Kohomen:

(với epochs = 10; R = 0; α = 0.8)

* Khởi tạo giá trị của các vector trọng số:
* Lần lặp 1:
* ***Xét đơn hàng 1***:

Khoảng cách từ đơn hàng 1 tới là:

Tương tự như trên, 1.042; 1.304. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 1* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 2***:

Khoảng cách từ đơn hàng 2 tới , , lần lượt là: 1.292; 0.94; 1.378. Vậy nơron có trọng số là nơron có khoảng cách đến *đơn hàng 2* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 3***:

Khoảng cách từ đơn hàng 3 tới , , lần lượt là: 1.367; 1.275; 1.140. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 3* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 4***:

Khoảng cách từ đơn hàng 4 tới , , lần lượt là: 1.292; 1.612; 1.566. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 4* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 5***:

Khoảng cách từ đơn hàng 5 tới , , lần lượt là: 1.292; 2.125; 1.540. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 5* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 6***:

Khoảng cách từ đơn hàng 6 tới , , lần lượt là: 1.404; 1.708; 1.540. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 6* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 7***:

Khoảng cách từ đơn hàng 7 tới , , lần lượt là: 1.267; 1.710; 1.553. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 7* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 8***:

Khoảng cách từ đơn hàng 8 tới , , lần lượt là: 1.828; 0.995; 1.603. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 8* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* ***Xét đơn hàng 9***:

Khoảng cách từ đơn hàng 9 tới , , lần lượt là: 1.265; 1.986; 1.540. Vậy nơron có trọng số là nơron có khoảng cách đến *Đơn hàng 9* gần nhất, cập nhật lại trọng số :

Do R = 0 nên không cập nhật vùng lân cận , giữ nguyên giá trị, vector trọng số lúc này:

* Tương tự, lần lặp 2 với từng vị trí thu được vec tor trọng số:
* Tương tự, lần lặp 3 với từng vị trí thu được vec tor trọng số:
* Sau 3 lần lặp nhận xét:
* Các vị trí đã được gom vào vị trí là không thay đổi qua nhiều lần lặp.
* Dừng thuộc toán.
* Vậy kết quả là:
* Vị trí 1 và 3 thuộc Vector 3.
* Vị trí 2 và 8 thuộc Vector 2.
* Vị trí 4, 5, 6, 7 và 9 thuộc Vector 1.

#### Nhận xét:

* Từ kết quả của a, b có thể thấy ở 2 thuật toán các vị trí được gom về một cụm đều giống nhau. Các cụm tương ứng là:

|  |  |  |
| --- | --- | --- |
| k-Means | Mạng Kohomen | Vị trí |
| *Cụm 1* | ***Cụm 3*** | ***1, 3*** |
| *Cụm 2* | ***Cụm 2*** | ***2, 8*** |
| *Cụm 3* | ***Cụm 1*** | ***4, 5, 6, 7, 9*** |

#### Dự báo vào cụm nào:

* Từ kết quả của k-Means, ta có:
* Khách hàng đặt với dữ liệu(1, 0, 1, 0, 0)

Áp dụng, tính khoảng cách giứa các cụm với dữ liệu mới

Thấy rằng: Khoảng cách từ đơn hàng mới đến các nhóm thì, nhóm 2 phù hợp nhất vì

Vậy nhóm 2 sẽ bao gồm đơn hàng 2 và đơn hàng 8 => thường mua tủ lạnh sẽ mua bếp điện và khi thoảng mua máy giặt, không mua Tivi.

### Bài tập 3:

* Yêu cầu:



* Giải quyết:
* import thư viện cần thiết cho bài toán:

Text

Description automatically generated

* Đọc dữ liệu từ file csv:



* Hiển thị 20 phần tử đầu của data mới được đọc từ file csv:

A picture containing table

Description automatically generated

* Hiển thị thông tin của dataset:

Text

Description automatically generated

* Loại bỏ thuộc tính dư thừa:

Graphical user interface, text, application

Description automatically generated

* Dữ liệu sau khi đã được cleaning:

Table

Description automatically generated with medium confidence

* Hiển thị phân phối của thuộc tính age bằng biểu đồ Histogram:

Graphical user interface, text, application, chat or text message

Description automatically generatedChart, line chart

Description automatically generated

* Hiển thị phân phối của thuộc tính Annual Income bằng biểu đồ Histogram:

Chart

Description automatically generatedChart, line chart

Description automatically generated

* Hiển thị phân phối của thuộc tính Spending Score bằng biểu đồ Histogram:

Chart

Description automatically generatedChart, line chart, histogram

Description automatically generated

* Thống kê số mẫu dữ liệu theo Gender:

Chart, bar chart

Description automatically generated

* Hiển thị phân bố của Age và Annual Income theo Gender:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

* Hiển thị phân bố của Annual Income và Spending Score theo Gender:

Chart, scatter chart

Description automatically generated

* Khảo sát việc gom cụm theo thuộc tính Age và Spending Score theo phương pháp khủy tay:

Text

Description automatically generated

* Hiển thị giá trị Intertia trên biểu đồ:

Chart, line chart

Description automatically generated

* Tạo mô hình với số cụm là 4:

Text

Description automatically generated

* Hiển thị số cụm, những phần tử cùng một cụm sẽ có chung 1 màu:

Chart, scatter chart

Description automatically generated

* Hiển thị hệ số dáng điệu để đánh giá kết quả gom cụm:

Graphical user interface, text, application

Description automatically generated

* Hiển thị các phần tử chung một cụm trong không gian 3 chiều: Age, Annual Income, Spending Score

Text

Description automatically generated

Chart

Description automatically generated

### Bài tập 4:

* Yêu cầu:

Text

Description automatically generated

* Giải quyết:
* import thư viện cần thiết cho bài toán:

Text

Description automatically generated

* Đọc dữ liệu từ file csv:



* Hiển thị 20 phần tử đầu trong dataset:

Graphical user interface, table

Description automatically generated

* Hiển thị thông tin của dataset:

Graphical user interface, text, table

Description automatically generated with medium confidence

* Tiến hành cleaning data:

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, table

Description automatically generated with medium confidence

* Data sau khi cleaning:

**Table

Description automatically generated with medium confidence**

* Hiển thị phân bố của fare và age theo sex(1-male, 0-female):

Chart, scatter chart

Description automatically generated

* Khảo sát việc gom cụm thuộc tính fare và age theo phương pháp khủy tay:

Text

Description automatically generated

* Hiển thị giá trị Inertia trên biểu đồ:

Chart, line chart

Description automatically generated

* Tạo mô hình Kmeans với số cụm là 10:

Text

Description automatically generated

* Biểu diễn kết quả gom cụm bằng biểu đồ:

Chart, scatter chart

Description automatically generated

* In ra giá trị dáng điệu để đánh giá kết quả gom cụm:

Graphical user interface, text, application, Word

Description automatically generated

### Bài tập 5:

* Yêu cầu:

Text

Description automatically generated

* Giải quyết:
* import thư viện cần thiết cho bài toán:

Text

Description automatically generated

* Đọc dữ liệu từ file csv:



* Hiển thị 20 phần tử đầu trong dataset:

Table

Description automatically generated

* Hiển thị thông tin của dataset:

Text

Description automatically generated with low confidence

* Hiển thị phân bố của bán kính(radius\_means) và kết cấu (texture\_means) theo chuẩn đoán(M-ác tính, B-lành tính):

Chart, scatter chart

Description automatically generated

* Khảo sát việc gom cụm thuộc tính bán kính(radius\_means) và kết cấu (texture\_means) theo phương pháp khủy tay:

Graphical user interface, text

Description automatically generated

* Hiển thị giá trị Inertia trên biểu đồ:

Chart, line chart

Description automatically generated

* Tạo mô hình Kmeans với số cụm là 3:

Text

Description automatically generated

* Biểu diễn kết quả gom cụm bằng biểu đồ:

Chart, scatter chart

Description automatically generated

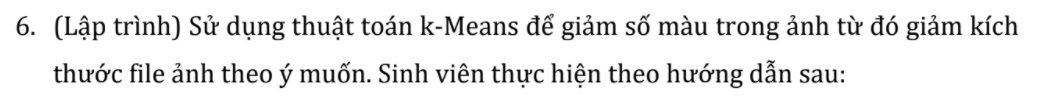
* In ra giá trị dáng điệu để đánh giá kết quả gom cụm:

Graphical user interface, text, application, Word

Description automatically generated

### Bài tập 6:

* Yêu cầu:



* Giải quyết:
* import thư viện cần thiết cho bài toán:

Text

Description automatically generated

* Đọc vào file ảnh cần giảm kích thước và hiển thị ra ngoài màn hình:

Graphical user interface, text, application

Description automatically generated

* Chuyển ảnh đọc được thành mảng số và lấy thông tin chiều rộng, cao của ảnh:

A picture containing text

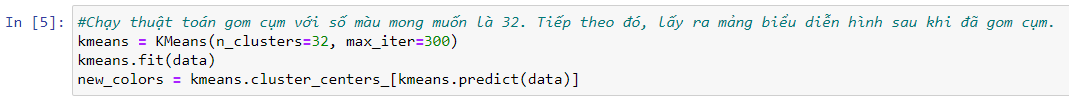
Description automatically generated

* Biến đổi các giá trị màu từ đoạn [0…255] về đoạn [0…1] và chuyển dạng mảng từ 3 chiều(cao, rộng, sâu) thành 2 chiều(cao x rộng, sâu):

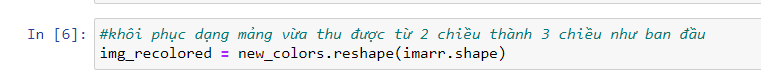
Text

Description automatically generated

* Chạy thuật toán gom cụm với số màu mong muốn là 32. Tiếp theo đó, lấy ra mảng biểu diễn hình sau khi đã gom cụm:



* Khôi phục dạng mảng vừa thu được từ 2 chiều thành 3 chiều như ban đầu:



* So sánh hình ban đầu với ảnh gốc ban đầu:

A screenshot of a dog

Description automatically generated with low confidence

* Lưu kết quả xuống đĩa cứng:

