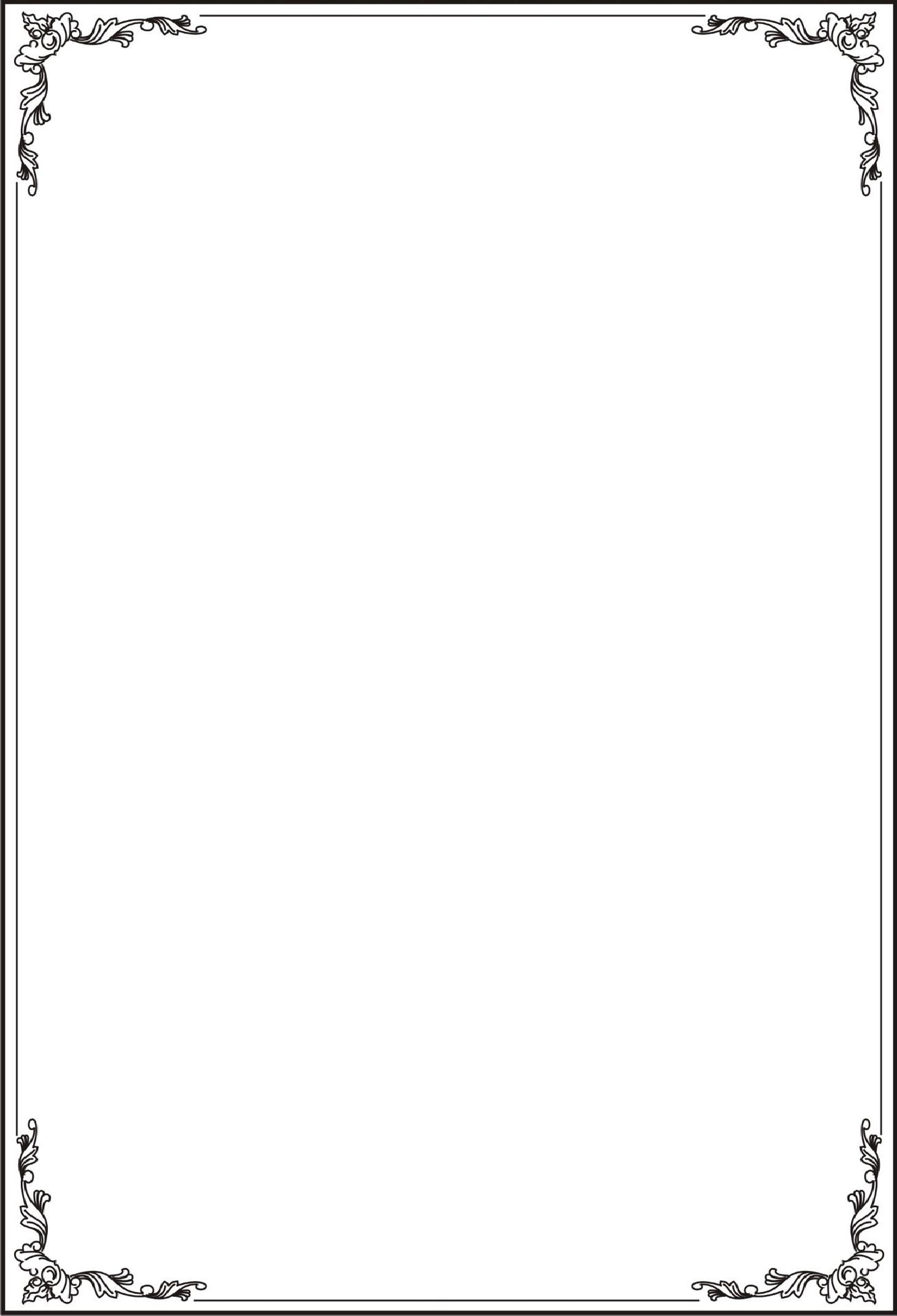
****

**TRƯỜNG ĐẠI HỌC TÀI NGUYÊN VÀ MÔI TRƯỜNG**

**THÀNH PHỐ HỒ CHÍ MINH**

**KHOA HỆ THỐNG THÔNG TIN VÀ VIỄN THÁM**



**BÁO CÁO CUỐI KÌ**

**BẢO MẬT MẠNG MÁY TÍNH VÀ HỆ THỐNG**

**NGHIÊN CỨU KỸ THUẬT PHÂN CỤM K\_MEAN VÀ ỨNG DỤNG CHO BÀI TOÁN PHÂN CỤM DỮ LIỆU TỰ ĐỘNG**

Sinh viên thực hiện: **Tạ Quang Thắng**

**Hoàng Thị Hoa**

**Mai Nguyễn Minh Bảo**

Mã sinh viên : **0950080112**

**0950080106**

**0950080085**

Lớp **: 09\_ĐH\_CNMP3**

Khoá  **: 2020 - 2024**

Giảng viên hướng dẫn : **KS. Phạm Trọng Huynh**

***TP. Hồ Chí Minh, tháng 09 năm 2023***

**TRƯỜNG ĐẠI HỌC TÀI NGUYÊN VÀ MÔI TRƯỜNG**

**THÀNH PHỐ HỒ CHÍ MINH**

**KHOA HỆ THỐNG THÔNG TIN VÀ VIỄN THÁM**

A logo with a person holding a book

Description automatically generated

**BÁO CÁO CUỐI KÌ**

**BẢO MẬT MẠNG MÁY TÍNH VÀ HỆ THỐNG**

**NGHIÊN CỨU KỸ THUẬT PHÂN CỤM K\_MEAN VÀ ỨNG DỤNG CHO BÀI TOÁN PHÂN CỤM DỮ LIỆU TỰ ĐỘNG**

Sinh viên thực hiện: **Tạ Quang Thắng**

**Hoàng Thị Hoa**

**Mai Nguyễn Minh Bảo**

Mã sinh viên : **0950080112**

**0950080106**

**0950080085**

Lớp **: 09\_ĐH\_CNMP3**

Khoá  **: 2020 - 2024**

Giảng viên hướng dẫn : **KS. Phạm Trọng Huynh**

***TP. Hồ Chí Minh, tháng 09 năm 202*3**

**MỤC LỤC**

[**MỤC LỤC** 3](#_Toc164754103)

[**MỞ ĐẦU** 1](#_Toc164754104)

[Chương 1: TỔNG QUAN 2](#_Toc164754105)

[1.1 Mục tiêu chính của đề tài 2](#_Toc164754106)

[1.2 Mô tả dữ liệu 2](#_Toc164754107)

[Chương 2: CƠ SỞ LÝ THUYẾT 3](#_Toc164754108)

[2.1 Bảo mật mạng máy tính 3](#_Toc164754109)

[2.2 Tổng quan về Google Colab 4](#_Toc164754110)

[2.3 Giới thiệu về thuật toán K-mean 5](#_Toc164754111)

[**2.3.1** **Bài toán gom cụm** 5](#_Toc164754112)

[**2.3.2** **Thuật toán K-mean** 5](#_Toc164754113)

[**2.3.3** **Ví dụ minh họa cho thuật toán K-mean** 6](#_Toc164754114)

[CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU 8](#_Toc164754115)

[CHƯƠNG IV: KẾT LUẬN 13](#_Toc164754116)

[TÀI LIỆU THAM KHẢO 14](#_Toc164754117)

**MỞ ĐẦU**

Trong những năm gần đây, cùng với sự nâng cao của chất lượng cuộc sống con người, nhu cầu mua về tinh thần đặc biệt là xem phim để thư giãn càng được đề cao và coi trọng. Giờ đây, chúng ta không chỉ coi phim ảnh là phương tiện để giải trí mà còn được coi như một công cụ để học hỏi những điều hay và có thật trong cuộc sống xung quanh chúng ta.

Nhận thấy được tầm quan trọng, tính thiết yếu của việc xây dựng và triển khai dự án cung cấp dịch vụ phim, chương trình truyền hình trên các quốc gia, chúng em đã quyết định chọn đề tài này để tìm hiểu, nghiên cứu trong bài báo cáo.

Netflix là dịch vụ phát trực tuyến phổ biến cung cấp danh mục phim, chương trình truyền hình và nội dung gốc phong phú. Tập dữ liệu này là phiên bản sạch của phiên bản gốc . Dữ liệu bao gồm các nội dung được thêm vào Netflix từ năm 2008 đến năm 2021. Nội dung cũ nhất là từ năm 1925 và nội dung mới nhất là năm 2021.

# TỔNG QUAN

﻿

## Mục tiêu chính của đề tài

## Mô tả dữ liệu

Nguồn:

https://github.com/codebasics/py/blob/master/ML/13\_kmeans/13\_kmeans\_tutorial.ipynb

* Cấu trúc dữ liệu

Dữ liệu gồm có 23 dòng và 3 cột, các thuộc tính dữ liệu bao gồm

|  |  |  |
| --- | --- | --- |
| STT | Thuộc tính | Mô tả |
| 1 | Name | Tên nhân viên |
| 2 | Age | Tuổi nhân viên |
| 3 | Income | Lương |

# CƠ SỞ LÝ THUYẾT

## Bảo mật mạng máy tính

Bảo mật mạng máy tính là tập hợp các hình thức, công cụ, chương trình được doanh nghiệp sử dụng vào mục đích bảo vệ riêng tư và an toàn cho những thông tin. Bảo mật mạng giúp hạn chế khả năng truy cập, sửa đổi và đánh cắp dữ liệu từ những cá nhân và tổ chức khác.

Một số cách thức mà các doanh nghiệp thường sử dụng để bảo mật mạng:

**Kiểm soát truy cập**

Kiểm soát truy cập hạn chế đối tượng truy cập vào các tài nguyên của doanh nghiệp bằng việc kiểm tra thông tin người truy cập bằng một số phương pháp: đăng nhập bằng tài khoản/mật khẩu được cung cấp, nhận diện gương mặt hoặc dấu vân tay,…

**Tường lửa (Firewall)**

Tường lửa là công cụ giúp cho các doanh nghiệp ngăn chặn chống xâm nhập của người ngoài tổ chức, tạo ra rào chắn giữa mạng nội bộ và mạng khác,…, theo dõi, quản lý lưu lượng truy cập và ngăn chặn ngay khi thấy có điều bất thường

**Hệ thống phát hiện xâm nhập (IDS)**

Hệ thống phát hiện xâm nhập có khả năng theo dõi, phân tích và đánh giá những hoạt động diễn ra trên mạng, sau khi phát hiện hành vi khả nghi sẽ ngay lập tức truyền tín hiệu thông báo đến người quản trị

**VPN (vitual private network)**

VPN là mạng riêng ảo dùng để kết nối các máy tính một cách an toàn bảo mật tới hệ thống mạng của doanh nghiệp qua internet

**Giám sát và phân tích mã độc**

Giúp các loại mã độc đang hiện hữu trên hệ thống hoặc mã độc được gửi đến doanh nghiệp thông qua các thư rác.Từ đó, doanh nghiệp có thể cô lập và ngăn chặn những liên kết có chứa mã độc xâm hại hệ thống doanh nghiệp.

Ngoài ra, đối với máy tính cá nhân ở nhà của bạn cũng nên thiết lập những bảo mật cho nó để tránh những tính trạng bị đột nhập từ bên ngoài gây ra những tổn thất cho bạn. Bạn có thể xem thêm cách bảo mật máy tính cá nhân tại Microsoft.

## Tổng quan về Google Colab

Google Colab là một dịch vụ cung cấp môi trường làm việc dựa trên trình duyệt (cloud-based) cho Python, chủ yếu là sử dụng cho việc phát triển và chạy mã Python thông qua trình duyệt web. Dịch vụ này được cung cấp miễn phí bởi Google và chạy trên cơ sở hạ tầng của Google Cloud.

Dưới đây là một số điểm nổi bật của Google Colab:

* Miễn Phí và Dễ Truy Cập: Google Colab là miễn phí và không yêu cầu cài đặt bất kỳ phần mềm nào. Bạn chỉ cần có tài khoản Google để truy cập và sử dụng.
* Môi Trường Jupyter Notebooks: Google Colab sử dụng môi trường Jupyter Notebooks, giúp bạn tạo và chia sẻ tài liệu kết hợp mã, văn bản, hình ảnh và biểu đồ.
* GPU và TPU: Colab cung cấp khả năng sử dụng GPU (Graphics Processing Unit) và TPU (Tensor Processing Unit) của Google để tăng tốc quá trình huấn luyện mô hình máy học, đặc biệt là trong việc xử lý lớn và nặng.
* Dữ Liệu Trong Đám Mây: Bạn có thể lưu trữ và truy cập dữ liệu của mình trực tiếp từ Google Drive hoặc từ nhiều nguồn dữ liệu khác nhau.
* Kết Nối Với GitHub: Colab tích hợp với GitHub, giúp bạn làm việc với mã nguồn từ kho lưu trữ GitHub một cách thuận lợi.
* Hỗ Trợ Nhiều Ngôn Ngữ Lập Trình: Colab không chỉ hỗ trợ Python mà còn hỗ trợ nhiều ngôn ngữ lập trình khác như R, Scala.

Google Colab thường được sử dụng trong nghiên cứu và phát triển mô hình machine learning, data science và các dự án sử dụng ngôn ngữ lập trình Python.

## Giới thiệu về thuật toán K-mean

### **Bài toán gom cụm**

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection

Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:

- Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn;

- Biology: Phận nhóm động vật và thực vật dựa vào các thuộc tính của chúng;- Libraries: Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…;

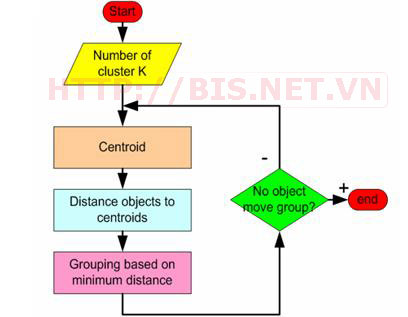
- Insurance, Finance: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds);

- WWW: Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog);…

### **Thuật toán K-mean**

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.

Thuật toán K-Means được mô tả như sau



Thuật toán K-Means thực hiện qua các bước chính sau:

* Bước 1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
* Bước 2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)
* Bước 3. Nhóm các đối tượng vào nhóm gần nhất
* Bước 4. Xác định lại tâm mới cho các nhóm

- Bước 5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

### **Ví dụ minh họa cho thuật toán K-mean**

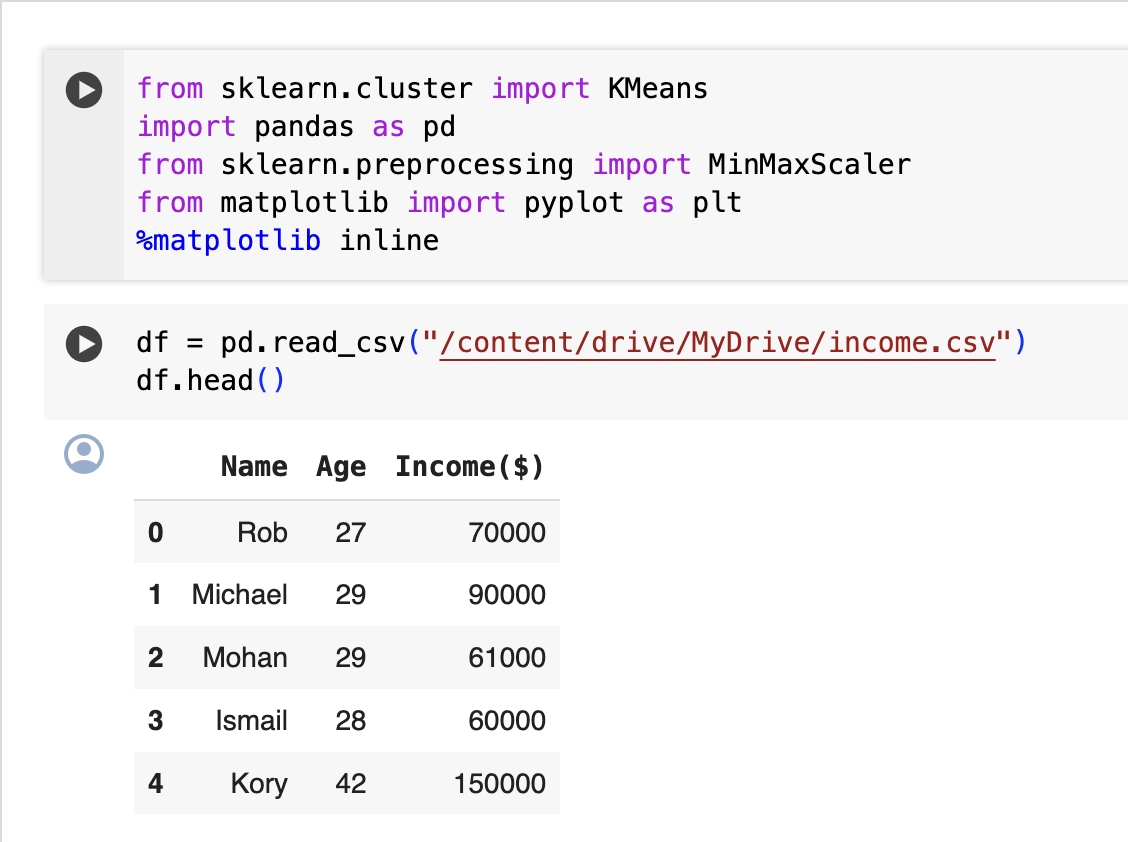
Giả sử ta có 4 loại thuốc A,B,C,D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng

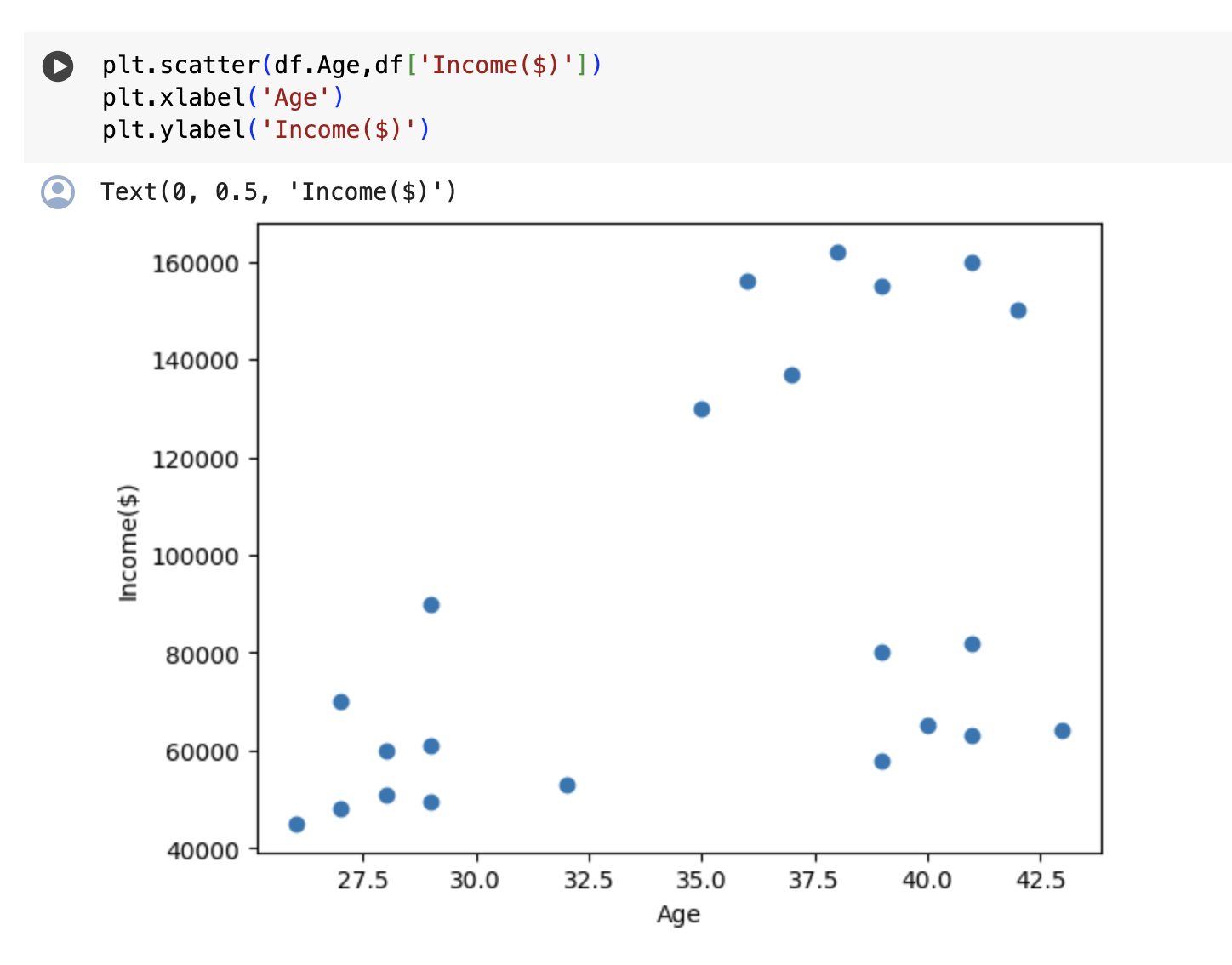
|  |  |  |
| --- | --- | --- |
| Object | Feature 1 (X): weight index | Feature 2 (Y): pH |
| Medicine A | 1 | 1 |
| Medicine B | 2 | 1 |
| Medicine C | 4 | 3 |
| Medicine D | 5 | 4 |

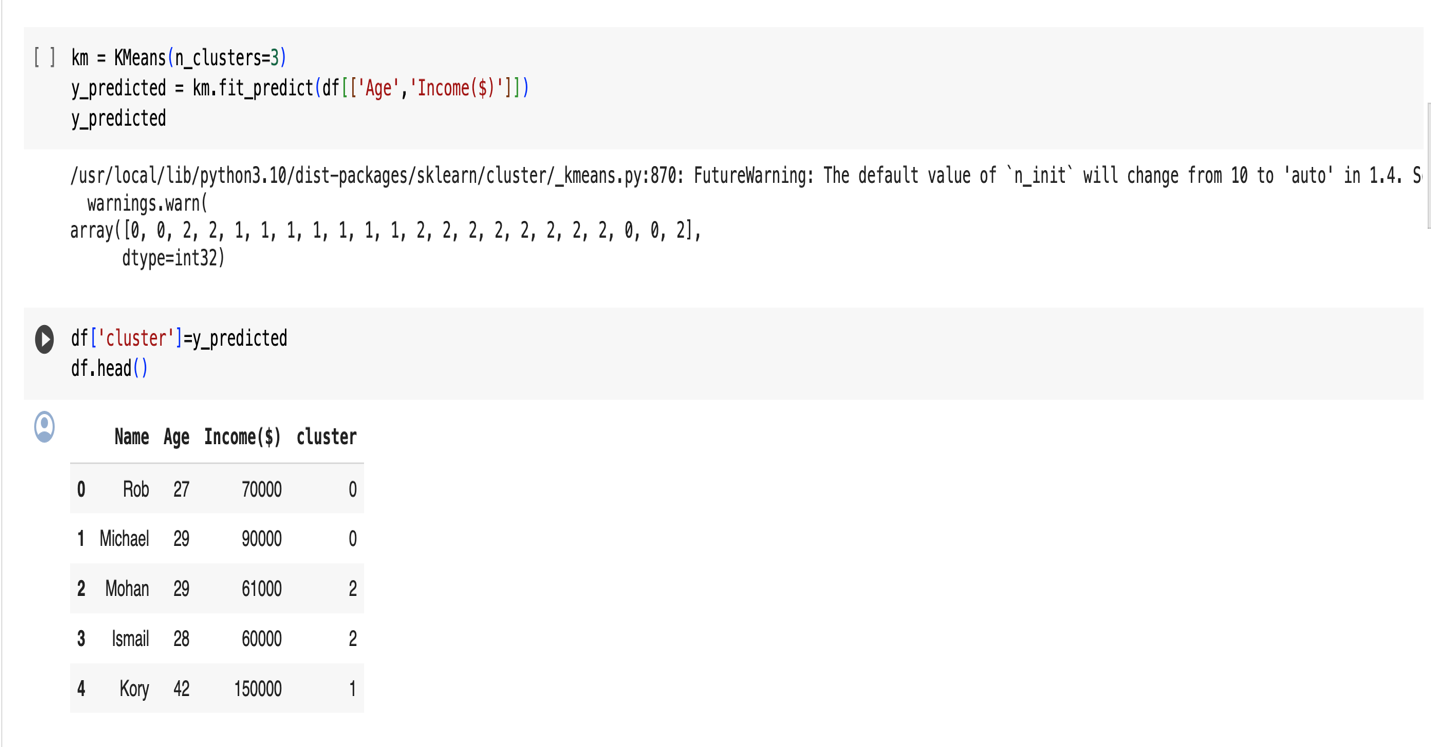
# CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU

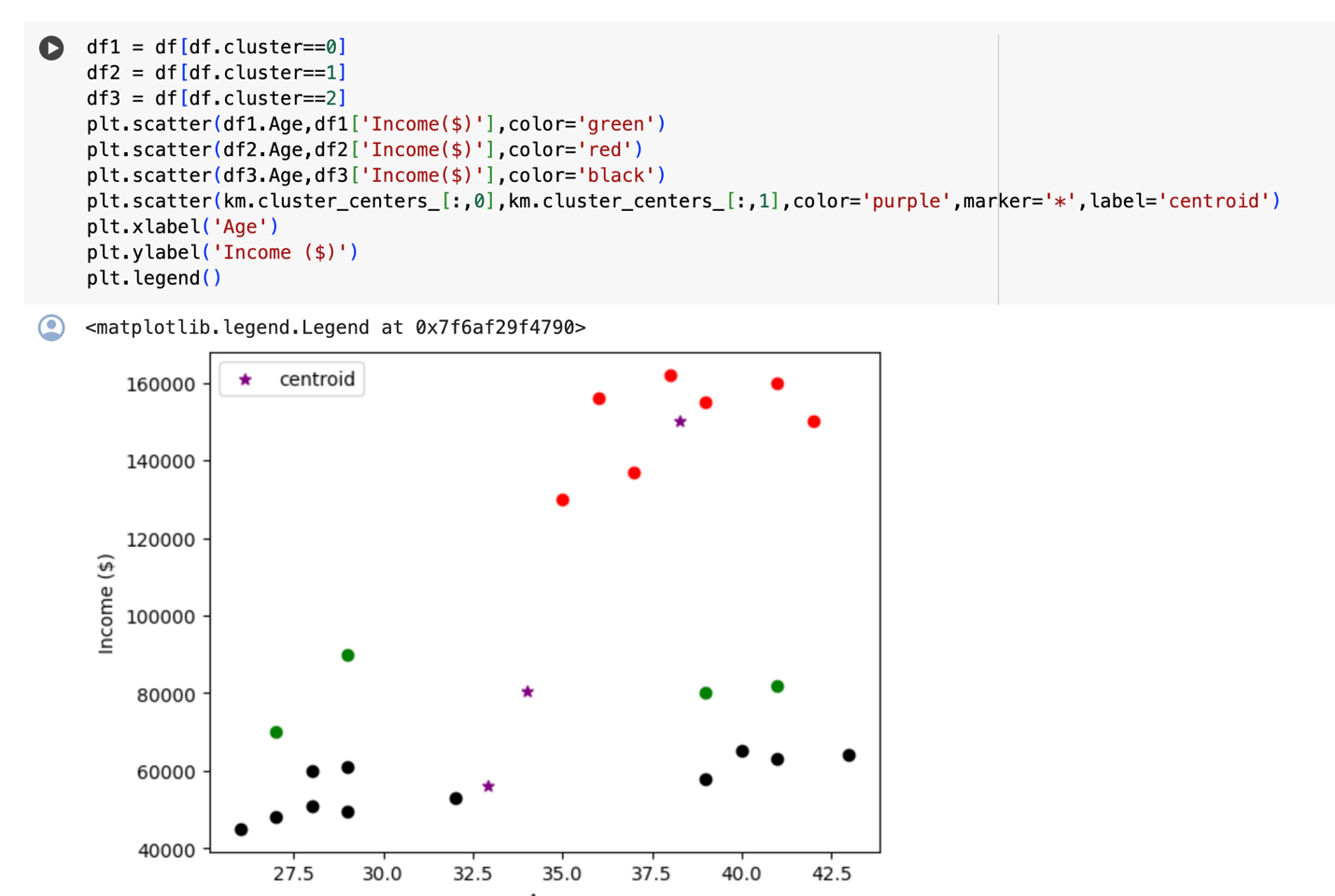
A. Datasheet

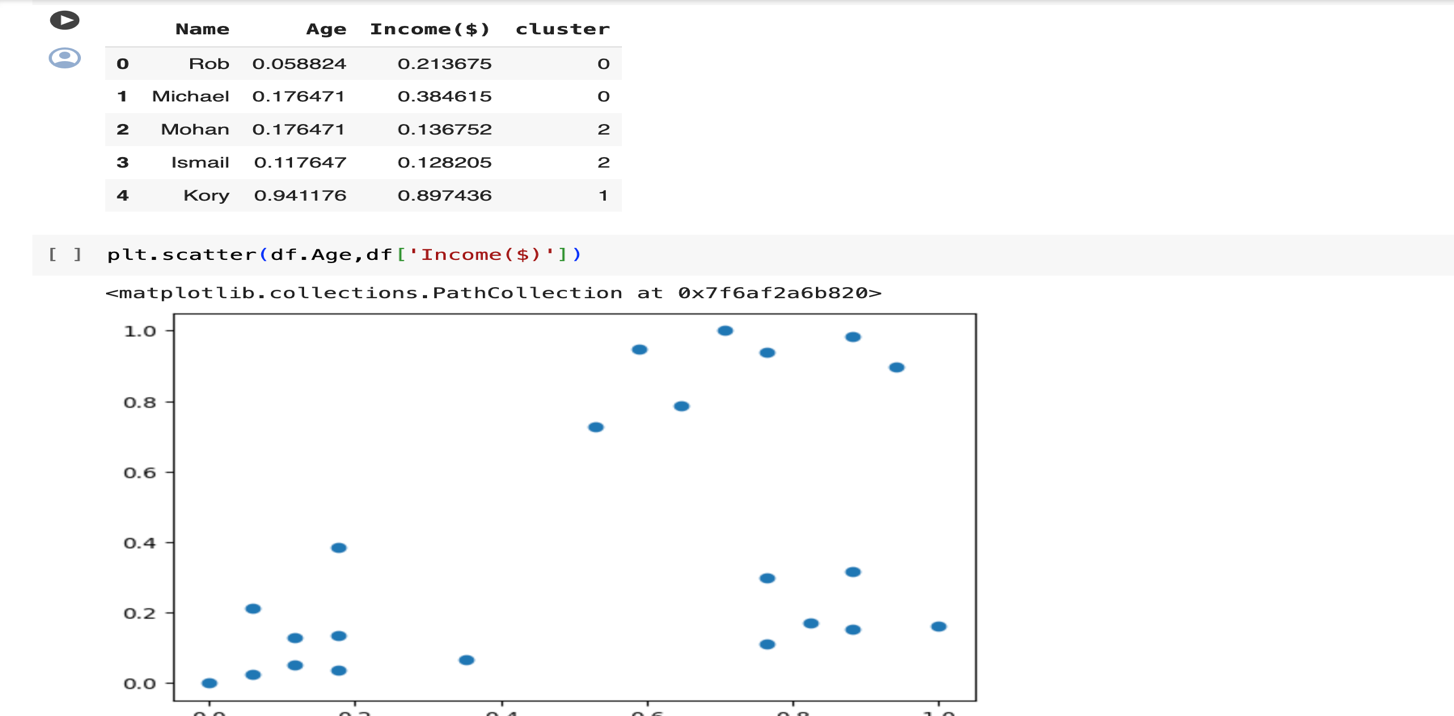


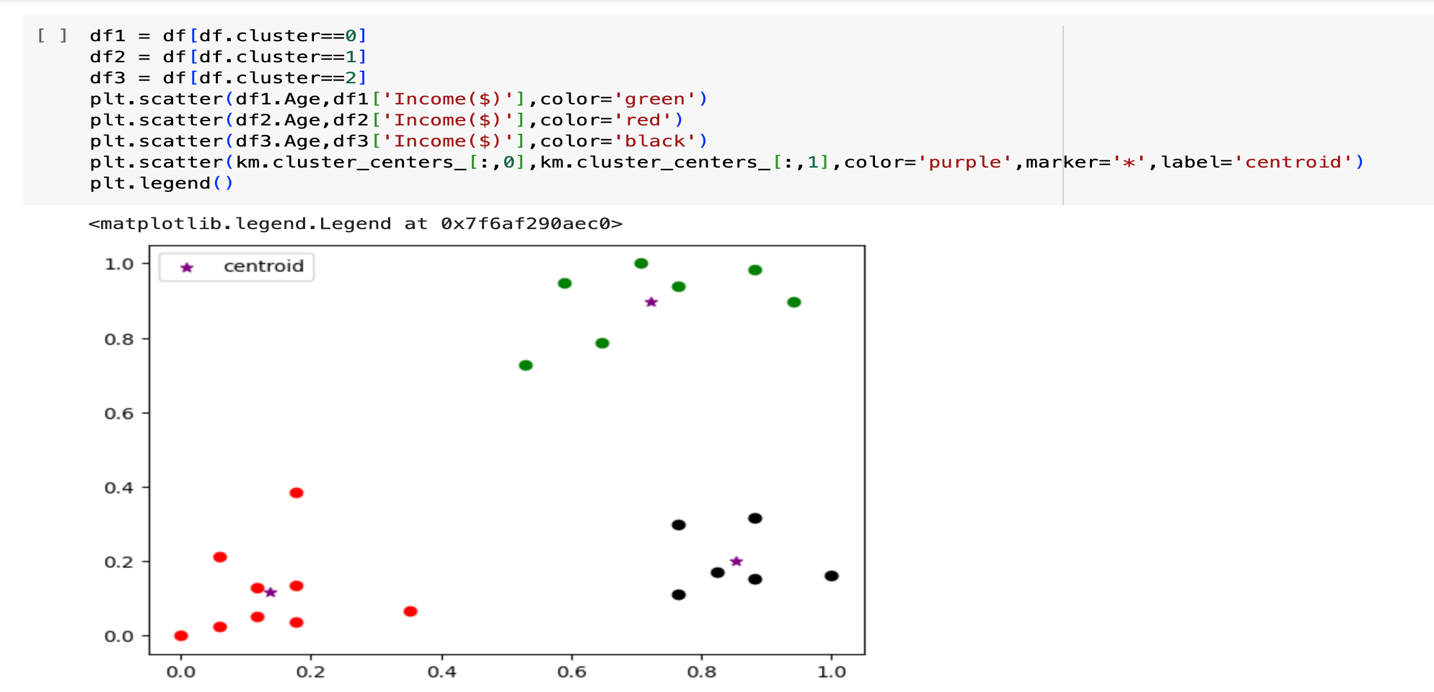
B. Chạy mô hình 

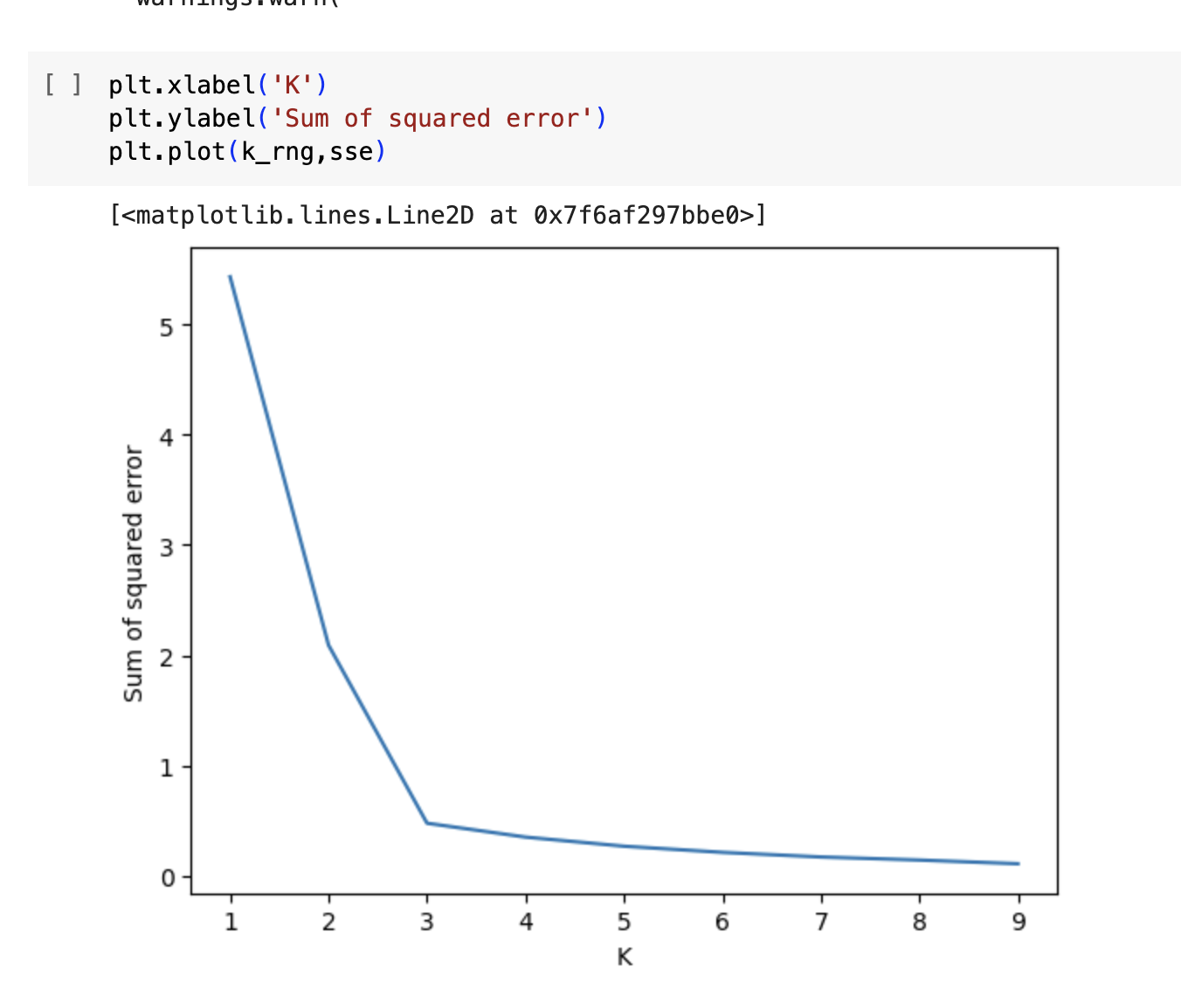












# CHƯƠNG IV: KẾT LUẬN

Tìm kiếm và phân cụm tự động các thông tin phù hợp và có giá trị trên Web là một chủ đề quan trọng và là vấn đề quan trọng của mỗi đơn vị, tổ chức có nhu cầu thu thập và tìm kiếm thông tin trên Internet. Bài báo đã đưa ra cải tiến thuật toán K-means trong phân cụm tài liệu Web. Thay vì chọn số điểm làm trọng tâm thì không chọn số điểm làm trọng tâm cho số cụm mà sẽ tăng số cụm từ 1 lên k cụm bằng cách đưa trung tâm cụm mới vào cụm có mức độ biến dạng lớn nhất và tính lại trọng tâm các cụm. Nhóm tác giả đã cài đặt thử nghiệm trên các bộ cơ sở dữ liệu, cho kết quả bước đầu khá khả quan.

# TÀI LIỆU THAM KHẢO

1. https://github.com/codebasics/py/blob/master/ML/13\_kmeans/13\_kmeans\_tutorial.ipynb