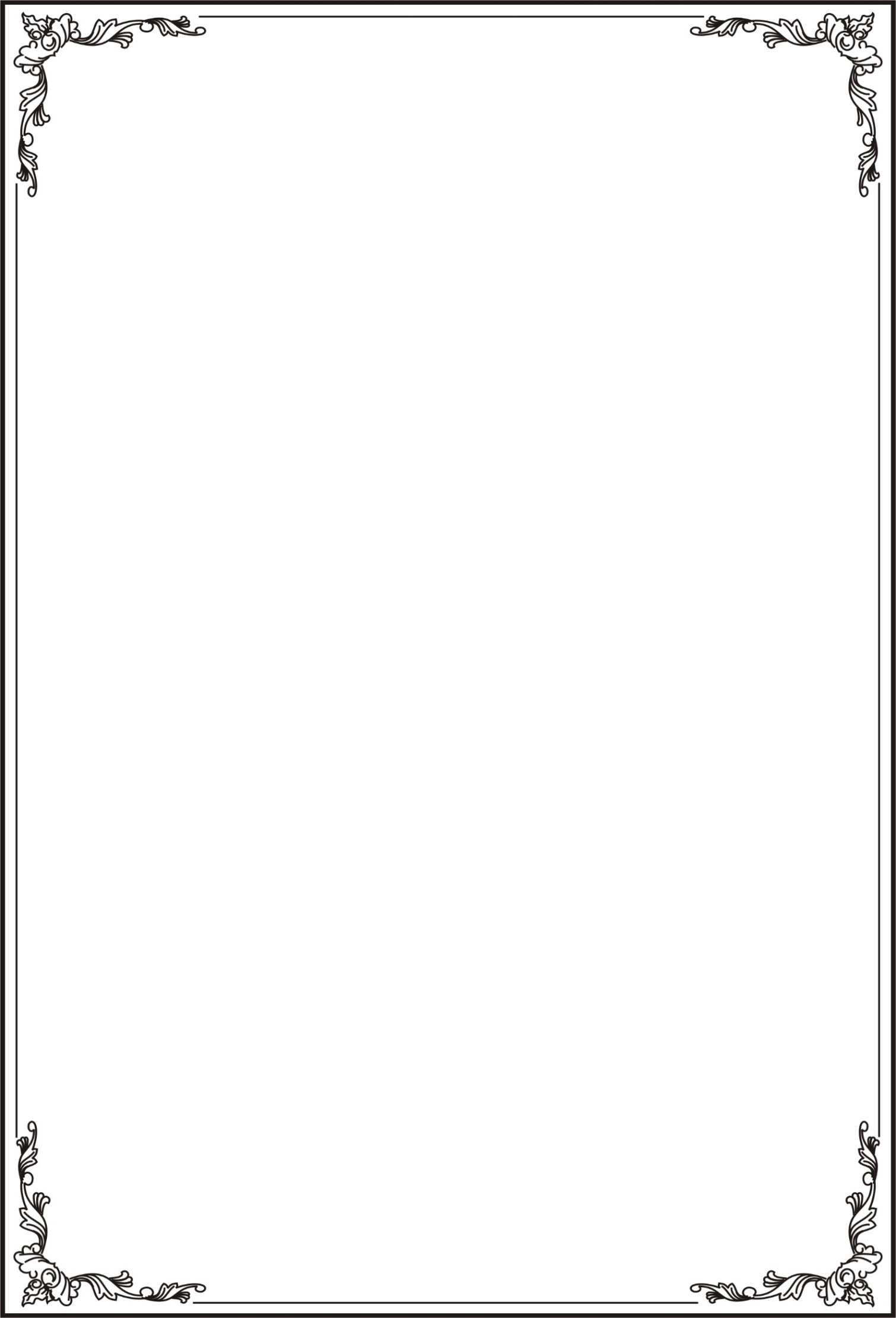
**TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**🙞🕮🙜**

**BÁO CÁO ĐỒ ÁN 1  
TÌM HIỂU MÀ MINH HOẠ THUẬT TOÁN PHÂN CỤM KMEANS**

**GVHD: Ths.Lê Văn Vinh**

**SVTH: Nguyễn Công Thịnh 16110219**

**Nguyễn Tấn Sang 16110197**

**TP. Hồ Chí Minh, tháng 8 năm 2020**

**ĐIỂM SỐ**

|  |  |  |  |
| --- | --- | --- | --- |
| **TIÊU CHÍ** | **NỘI DUNG** | **TRÌNH BÀY** | **TỔNG** |
| **ĐIỂM** |  |  |  |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**Giáo viên hướng dẫn**

(*ký và ghi họ tên*)

…………………….....

**MỤC LỤC**

Table of Contents

[Thuật toán K-means CLUSTERINg 1](#_Toc47381206)

[**1.** **Giới thiệu thuật toán K-Means Clustering** 1](#_Toc47381207)

[**a)** **Clustering là gì?** 1](#_Toc47381208)

[**b)** **Kmeans là gì?** 1](#_Toc47381210)

**[2.](#_Toc47381212)****[Lý thuyết nền tảng của thuật toán K-means Clustering](#_Toc47381212)** [1](#_Toc47381212)

[Tóm tắt thuật toán K-Means Clustering 5](#_Toc47381213)

[3. Tiền xử lí dữ liệu cho giải thuật về phân cụm 6](#_Toc47381214)

[4. Demo code 7](#_Toc47381215)

[TÀI LIỆU THAM KHẢO 10](#_Toc47381216)

# 

# Thuật toán K-means CLUSTERINg

1. **Giới thiệu thuật toán K-Means Clustering**
2. **Clustering là gì?**

`Clustering` là cách nhóm 1 tập hợp các objects dựa trên các đặc điểm, thuộc tính của chúng, tổng hợp chúng lại theo từng nhóm dựa trên ĐỘ GIỐNG NHAU (`similarities`). Liên quan đến data mining, phương pháp này (clustering) phân chia data bằng 1 thuật toán join được chỉ định, thích hợp nhất cho các phân tích thông tin mong muốn.

1. **Kmeans là gì?**

Kmean nó là 1 thuật toán để phân nhóm 1 tập các object dựa trên các thuộc tính của chúng.

Thuật toán clustering được chia làm 2 nhóm chính:

* Distance-based clustering algorithms: thuật toán phân nhóm dựa trên khoảng cách
* Density-based clustering algorithms: thuật toán phân nhóm dựa trên mật độ

Kmean thuộc nhóm distance-based clustering algorithms.

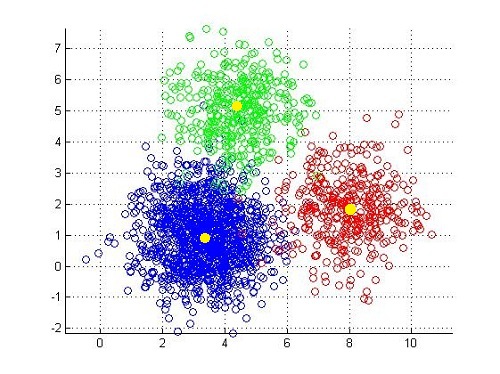
=> K-means Clustering là một thuật toán dùng trong các bài toán phân loại/nhóm n đối tượng thành k nhóm dựa trên đặc tính/thuộc tính của đối tượng (k £n nguyên, dương).  
Về nguyên lý, có n đối tượng, mỗi đối tượng có m thuộc tính, ta phân chia được các đối tượng thành k nhóm dựa trên các thuộc tính của đối tượng bằng việc áp dụng thuật toán này.

1. **Lý thuyết nền tảng của thuật toán K-means Clustering**

## Mô hình toán học

Ta gọi điểm tại vị trí trung bình của tất cả các điểm dữ liệu trong một cụm là **trung tâm cụm**. Như vậy, nếu có K cụm thì sẽ có K trung tâm cụm và mỗi trung tâm cụm sẽ nằm gần các điểm dữ liệu trong cụm tương ứng hơn các trung tâm cụm khác.

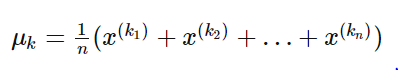
Trong hình dưới đây, K = 3 và ta có 3 trung tâm cụm là các điểm màu vàng.

[](https://3.bp.blogspot.com/-NMrDyo4_ySg/XDqlicW58WI/AAAAAAAAERI/bWXciHyrL_UzlGV2JSkB3D3iCKTtKnyYACLcBGAs/s1600/main.jpg)

Để phân cụm dữ liệu bằng K-Means Clustering, trước hết ta chọn K là số cụm để phân chia và chọn ngẫu nhiên K trong số m dữ liệu ban đầu làm trung tâm cụmμ1,μ2,…,μK.

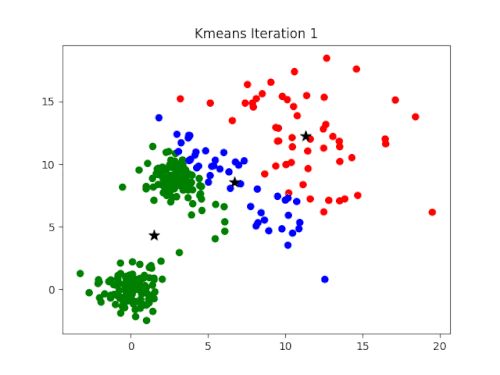
Sau đó, với điểm dữ liệu x(i) ta sẽ gán nó cho cụm c(i) là cụm có trung tâm cụm gần nó nhất.

Khi tất cả các điểm dữ liệu đã được gán về các cụm, bước tiếp theo là tính toán lại vị trí các trung tâm cụm bằng trung bình tọa độ các điểm dữ liệu trong cụm đó.



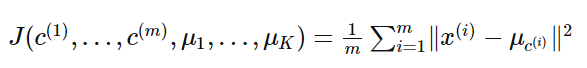
với k1,k2,…,kn là chỉ số các dữ liệu thuộc cụm thứ k.

Các bước trên được lặp lại cho tới khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó.

[](https://3.bp.blogspot.com/-YwhPRp33Icg/XDqrJbA1cFI/AAAAAAAAERU/pp4Wx6weR5IRWC6-05T3Q8jMzKypIbzBwCLcBGAs/s1600/k+means.gif)

## Độ chính xác của thuật toán

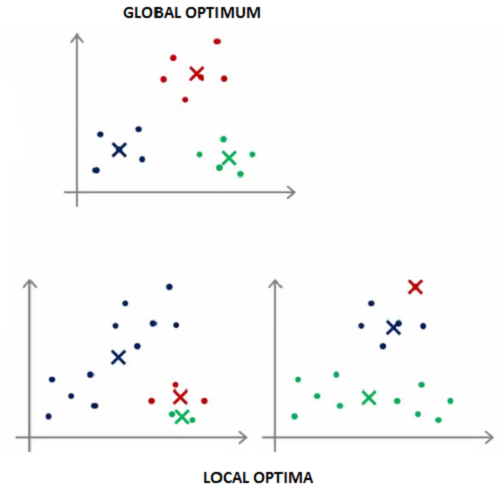
Hàm mất mát của thuật toán K-Means Clustering đặc trưng cho độ chính xác của nó sẽ càng lớn khi khoảng cách từ mỗi điểm dữ liệu tới trung tâm cụm càng lớn.



## Nghiệm của thuật toán

Trong các bước của thuật toán, thực chất bước gán các điểm dữ liệu về trung tâm cụm gần nhất và bước thay đổi trung tâm cụm về vị trí trung bình của các điểm dữ liệu trong cụm đều nhằm mục đích giảm hàm mất mát. Thuật toán kết thúc khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó. Khi đó hàm mất mát đạt giá trị nhỏ nhất.

Khi K càng nhỏ so với m, thuật toán càng dễ đi đến kết quả chưa phải tối ưu. Điều này phụ thuộc vào cách chọn K trung tâm cụm ban đầu.

[](https://3.bp.blogspot.com/-rrkJKQxS2iU/XDqu94rQBCI/AAAAAAAAERg/7ld-oIlnhMoEM89zvDMEdRpsJRXYpX4dACLcBGAs/s1600/Image+%5b10%5d.png)

Để khắc phục điều này, ta cần lặp lại thuật toán nhiều lần và chọn phương án có giá trị hàm mất mát nhỏ nhất.

# Tóm tắt thuật toán K-Means Clustering

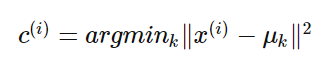
**Yêu Cầu**

Cho trước m bộ dữ liệu x(1),x(2),…,x(m). Nhiệm vụ của ta là phân chia các điểm dữ liệu thành K cụm dựa trên vị trí tương đối của chúng so với nhau.

**Thuật Toán**

**Bước 1:**Chọn ngẫu nhiên K trong số m điểm dữ liệu làm trung tâm cụm μ1,μ2,…,μK.

**Bước 2:** Gán mỗi điểm dữ liệu về cụm có trung tâm cụm gần nhất.



**Bước 3**: Di chuyển các trung tâm cụm về vị trí trung bình của các điểm dữ liệu trong cụm tương ứng.

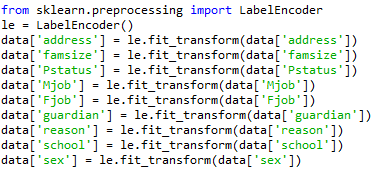
**Bước 4**: Lặp lại hai bước trên tới khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó.

**Bước 5**: Lặp lại tất cả các bước trên một số lần và chọn phương án có hàm mất mát nhỏ nhất.

## Tiền xử lí dữ liệu cho giải thuật về phân cụm

Xử lý dữ liệu chuỗi

Do dataset cô cung cấp một số thuộc tính dữ liệu dạng chuỗi nên ta cần chuyển chúng thành dữ liệu số để thực hiện các giải thuật. Ở đây nhóm đã sử dụng lớp LabelEncoder do thư viện Scikit\_Learn cung cấp chuyển dữ liệu chữ sang số nguyên.



Sử dụng phương thức ***fit\_transform()*** để xử lí và chuyển dữ liệu chữ của các thuộc tính sang số nguyên.

Các thuộc tính được xử lí dữ liệu nhóm sử dụng ở đây là

* address - loại địa chỉ nhà của sinh viên
* famsize - quy mô gia đình
* Pstatus - tình trạng chung sống của cha mẹ
* Medu – nền tảng giáo dục của mẹ
* Fedu – nền tảng giáo dục của cha
* Mjob - công việc của mẹ
* Fjob - công việc của cha
* reason - lý do để chọn trường này
* guardian - người giám hộ của sinh viên
* school - trường học của sinh viên
* sex - giới tính của sinh viên

Xử lý dữ liệu dư thừa

Nhầm đạt độ chính xác cao cho các giải thuật nên nhóm đã loại bỏ một số thuộc tính cho là không liên quan đến giải thuật

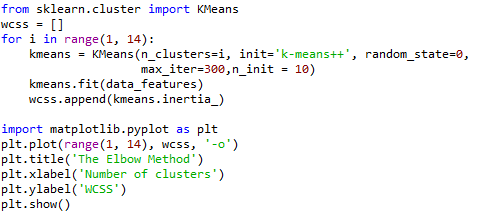


Các thuộc tính không được nhóm sử dụng ở đây là

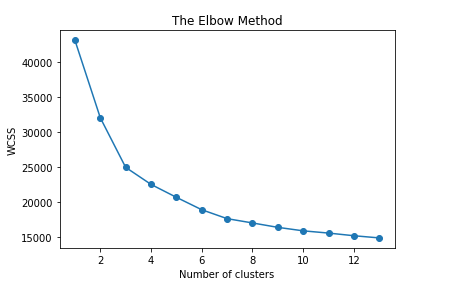
* schoolup - hỗ trợ giáo dục bổ sung
* famsup - hỗ trợ giáo dục gia đình
* paid - lớp học có trả phíhay không
* activities - hoạt động ngoại khóa
* nursery – có học trường mẫu giáo không
* higher - muốn học cao hơn
* internet - Truy cập Internet tại nhà
* romantic - có mối quan hệ bạn trai/bạn gái hay không

## Demo code

Trước khi chạy giải thuật K-Means, nhóm em sẽ tính toán để chọn cụm tốt nhất cho K-Means

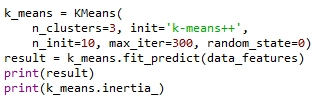


Kết quả sẽ cho ra biểu đồ để nhóm em có thể đánh giá và chọn lựa số cụm phù hợp

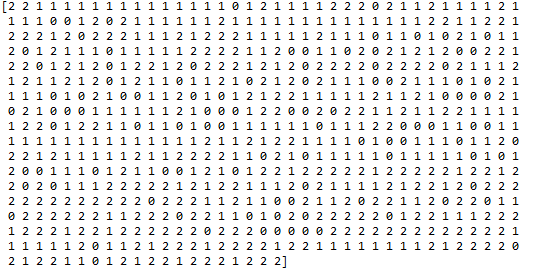


Từ hình vẽ, nhóm quyết định chọn 3 cụm vì biểu đồ gấp khúc nhiều ở vị trí 3

Tiến hành giải thuật với phân **3 cụm**



Sau khi chạy chương trình, nhóm thu được kết quả:



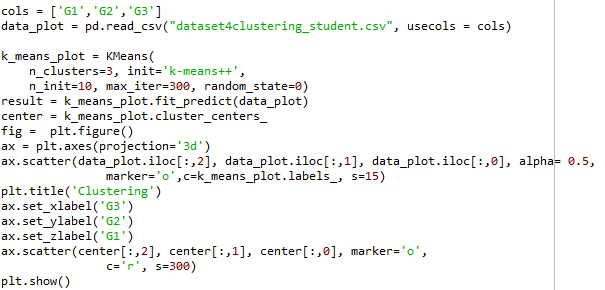
Ma trận gồm 649 giá trị được chia thành 3 cụm là 0, 1 và 2

**Kết quả WCSS**:

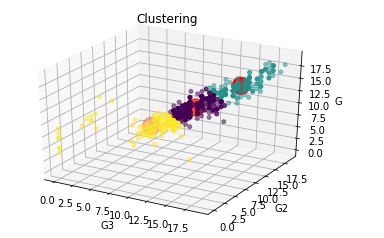


**Thực hiện vẽ đồ thị plot**

Nhóm chọn 3 features để vẽ là: G1( điểm số khi học môn 1), G2( điểm số khi học môn 2) và G3( điểm số khi học môn 3)



Kết quả:



Dữ liệu được chia theo 3 cụm theo các mức điểm của sinh viên

# TÀI LIỆU THAM KHẢO

Giải thuật K-Means:

1. Tìm WCSS: <https://medium.com/@iSunilSV/data-science-python-k-means-clustering-eed68b490e02>
2. Vẽ 3D: <https://jakevdp.github.io/PythonDataScienceHandbook/04.12-three-dimensional-plotting.html>