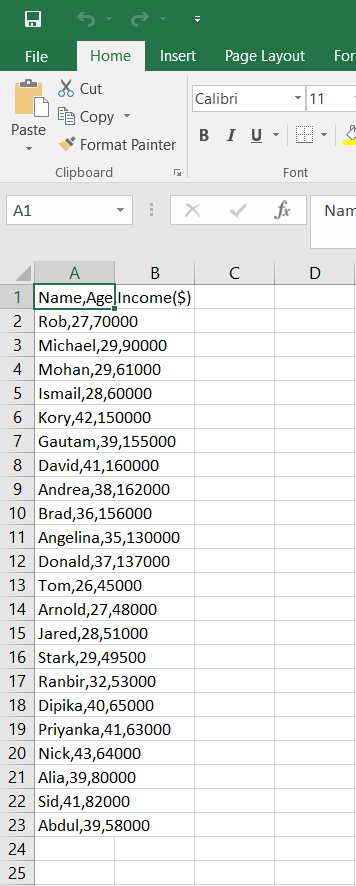
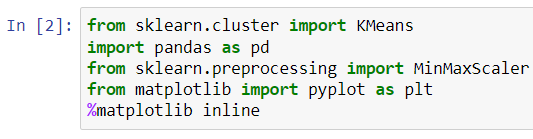
**BÀI TOÁN PHÂN CỤM KHÁCH HÀNG CƠ BẢN**

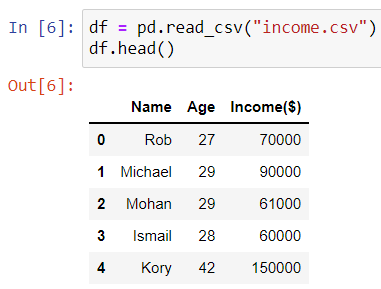
**Dataset:** Thông tin về Tên, tuổi và thu nhập của một số người



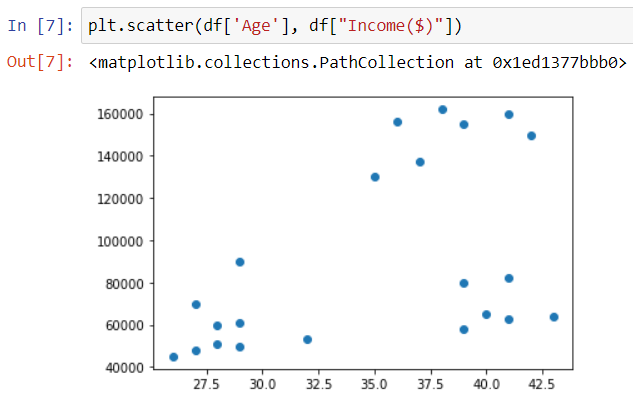
Khai báo một số thư viện cần thiết:



Đọc file income.csv:

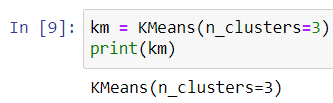


Vẽ đồ thị Scatter về 2 giá trị Age và Income($):

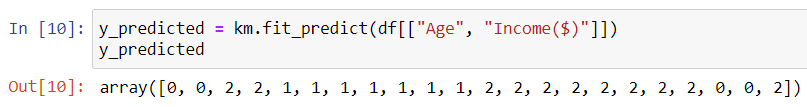


Ta dễ dàng thấy dữ liệu dường như sẽ gồm 3 cụm.

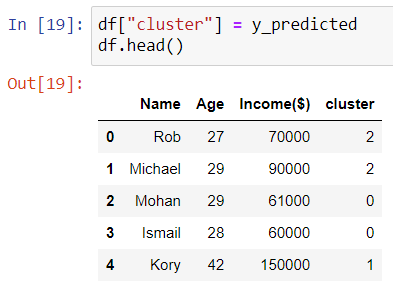
Khai báo Kmeans với số lớp là 3:

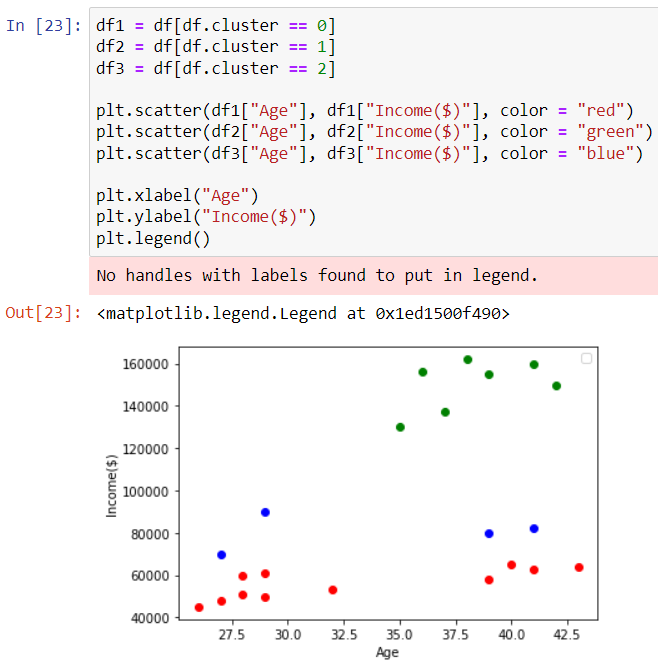


Fit and predict:



Hiển thị kết quả:





**Có thể thấy cluster xanh bị phân nhóm chưa đúng!!**

**Nguyên nhân:** scaling chưa đúng, trục Y (Income) từ 40000 🡪 160000, còn trục x được chia rất hẹp. Đó là lý do ta phải thực hiện việc tiền xử lý dữ liệu và sử dụng MinMaxScaler để scale những datapoint và sau đó mới thực hiện thuật toán Kmeans

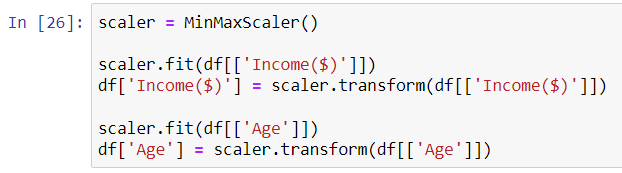
scaler = MinMaxScaler()

scaler.fit(df[['Income($)']])

df['Income($)'] = scaler.transform(df[['Income($)']])

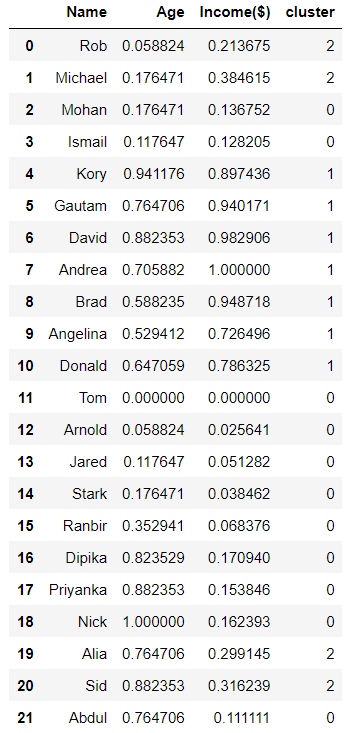
scaler.fit(df[['Age']])

df['Age'] = scaler.transform(df[['Age']])

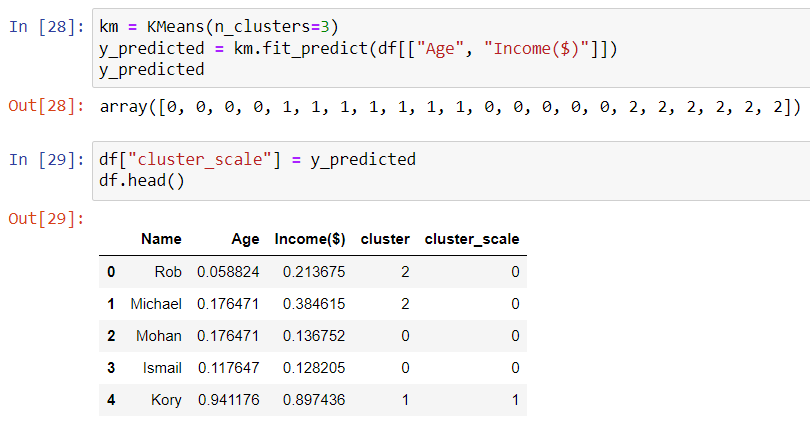


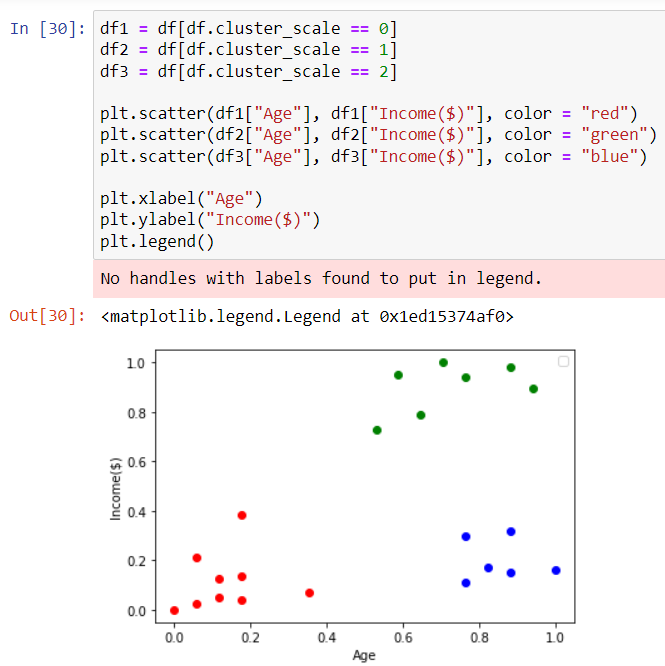
Đoạn code trên thực hiện Scale giá trị Income và Age vào trong khoảng [0, 1]

Có thể thấy kết quả lúc này ở cột Age và Income đã được Scale trong khoảng [0, 1]



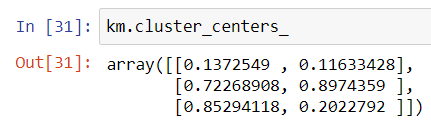
Bây giờ thực hiện thuật toán Kmeans một lần nữa lên các điểm dữ liệu:



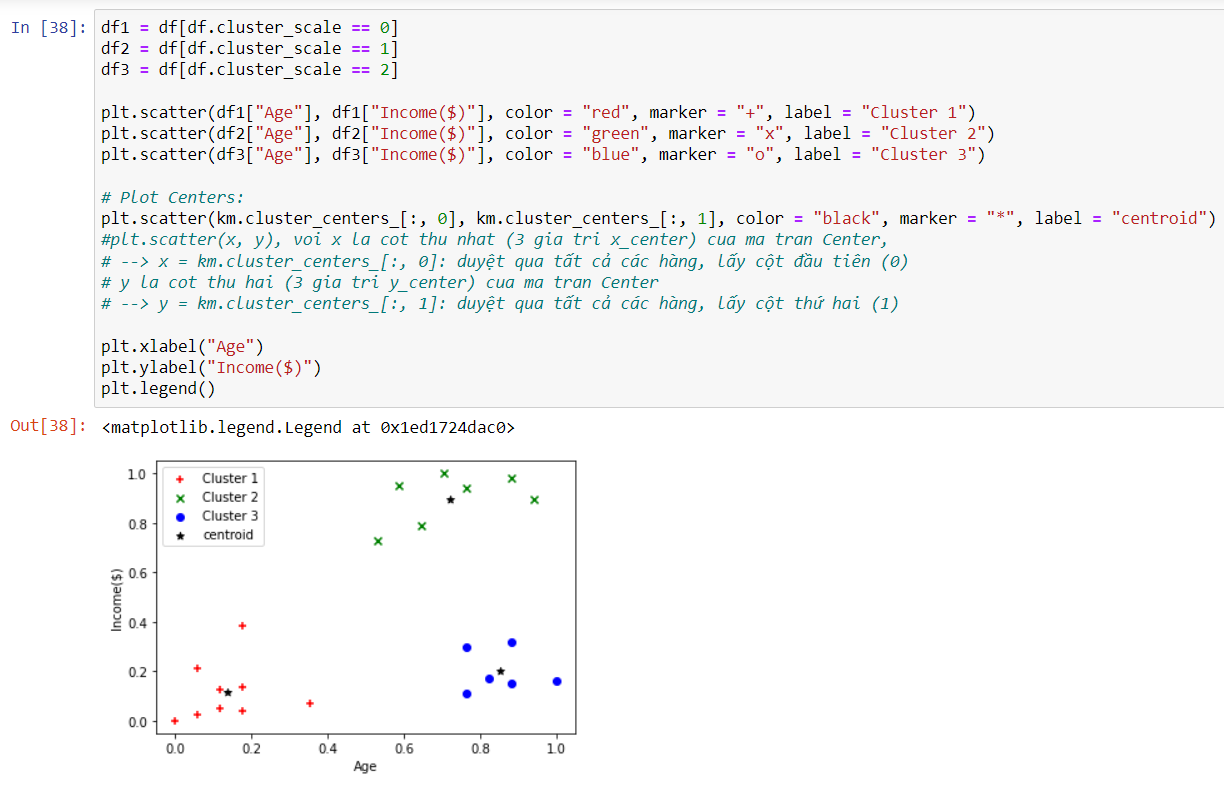


Ta thấy kết quả phân cụm lúc này đã tốt hơn ban đầu rất nhiều.

Ta có thể kiểm tra tâm các cụm bằng lệnh:

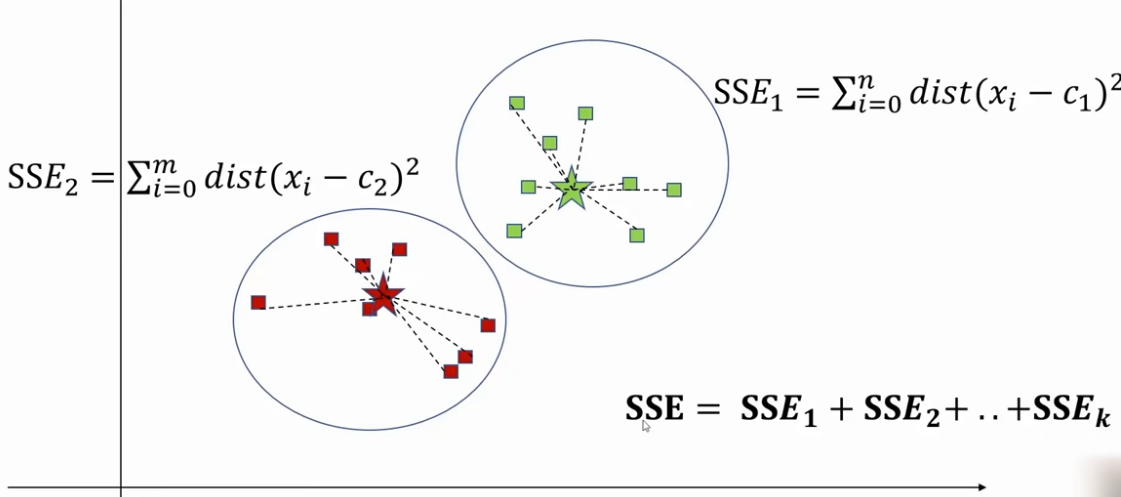


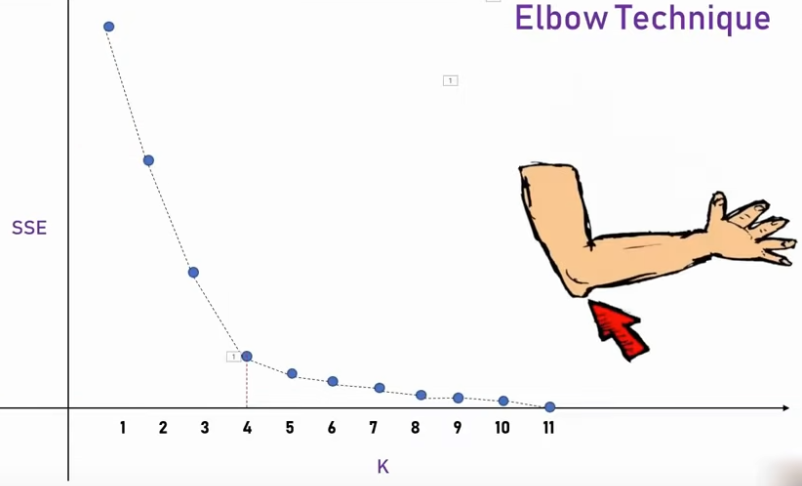
Ta có thể vẽ các tâm này vào các cụm:



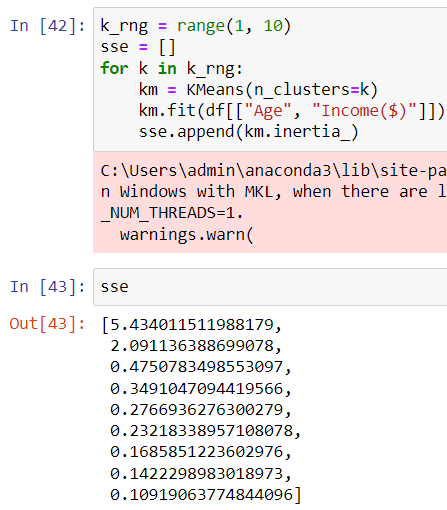
**Thuật toán Elbow:**

Vì dataset ta đang sử dụng là khá ít và đơn giản, có thể vẽ ra để xem trước và dự đoán được bằng trực giác giá trị k dễ dàng. Tuy nhiên ở những dự án thực tế thì ta sẽ gặp những dataset có nhiều đặc trưng hơn (Ví dụ 20 đặc trưng) thì rất khó vẽ ra dữ liệu hiển thị để ta có thể quan sát và tìm ra giá trị k phù hợp bằng trực giác nữa. Do đó ta phải sử dụng thuật toán Elbow để tìm ra giá trị k phù hợp:





Ở mỗi vòng lập ứng với mỗi giá trị k, khi ta gọi hàm km.fit() thì có một thông số là inertia sẽ cho ta the sum of square error, ta chỉ cần thêm giá trị này vào mảng SSE ban đầu.



Hiển thị lên đồ thị:

