**Phân khúc khách hàng bằng RFM và Kmean**

*Từ những khách hàng tiêu dùng lớn cho đến những khách hàng rời bỏ doanh nghiệp, tất cả những khách hàng đều có nhu cầu và mong muốn đa dạng. Doanh nghiệp muốn khách hàng chi tiêu nhiều hơn từ những chiến dịch tiếp thị chương trình, sản phẩm mới tới khách hàng theo những cách khác nhau. Tuy nhiên, câu hỏi đặt ra là làm thế nào để đưa ra được các chiến dịch tiếp thị phù hợp với những nhóm khách hàng đang có nhu cầu để từ đó tăng tỷ lệ phản hồi từ khách hàng và từ đó tăng doanh số bán hàng. Bài toán đặt ra là làm thế nào để có thể phân khúc khách hàng một cách tương đối chính xác dựa trên “hành vi giao dịch lịch sử” của khách hàng, thuật toán RFM sẽ giúp chúng ta giải quyết vấn đề này một cách nhanh chóng và hiệu quả.*

**1. RFM (Recently – Frequently – Montery Value)**

Để phân khúc khách hàng, có một vài metrics thường sử dụng như là:

+ Khách hàng mua hàng lần cuối khi nào? (R)

+ Tần suất khách hàng mua hàng nhiều hay ít (F)

+ Khách hàng chi bao nhiêu tiền để mua sản phẩm của công ty (M)

**Recently:** Để tính ngày mua hàng gần nhất: ta lấy ngày hiện tại trừ đi ngày mua hàng gần nhất của khách hàng, ta sẽ được một giá trị gọi là khoảng cách từ ngày mua hàng gần nhất đến hiện tại của khách hàng. Số này càng nhỏ thì khách hàng càng mua hàng gần hơn

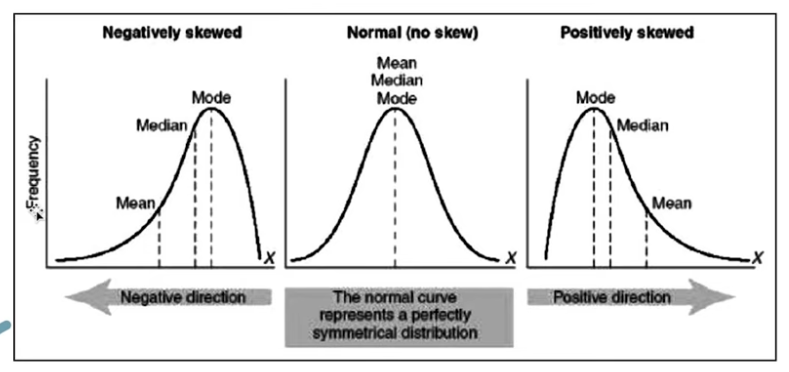
**Frequently:** Đếm số lượt mua hàng của khách hàng trong khoảng thời gian ta trích xuất dữ liệu.

**Montery Value:** tính tổng tiền tất cả mỗi khách hàng. (nhân đơn giá với số lượng và tính tổng)

Từ bảng hóa đơn lịch sử mua sắp khách hàng, ta tính các thông số R, F, M này cho mỗi khách hàng. Coi đó là các Feature để phân nhóm khách hàng.

**2. Chỉ số Skewness**

Chỉ số Skewness dùng để đánh giá dữ liệu có phân phối chuẩn hay không hay bị lệch trái, lệch phải.



Lý tưởng: chỉ số Skewness = 0. Dùng các biện pháp transform để đưa chỉ số Sknewness về càng gần 0 càng tốt.

Một số phương pháp:

+ log transformation

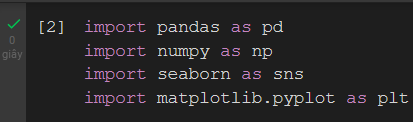
+ square root transformation

+ box-cox transformation

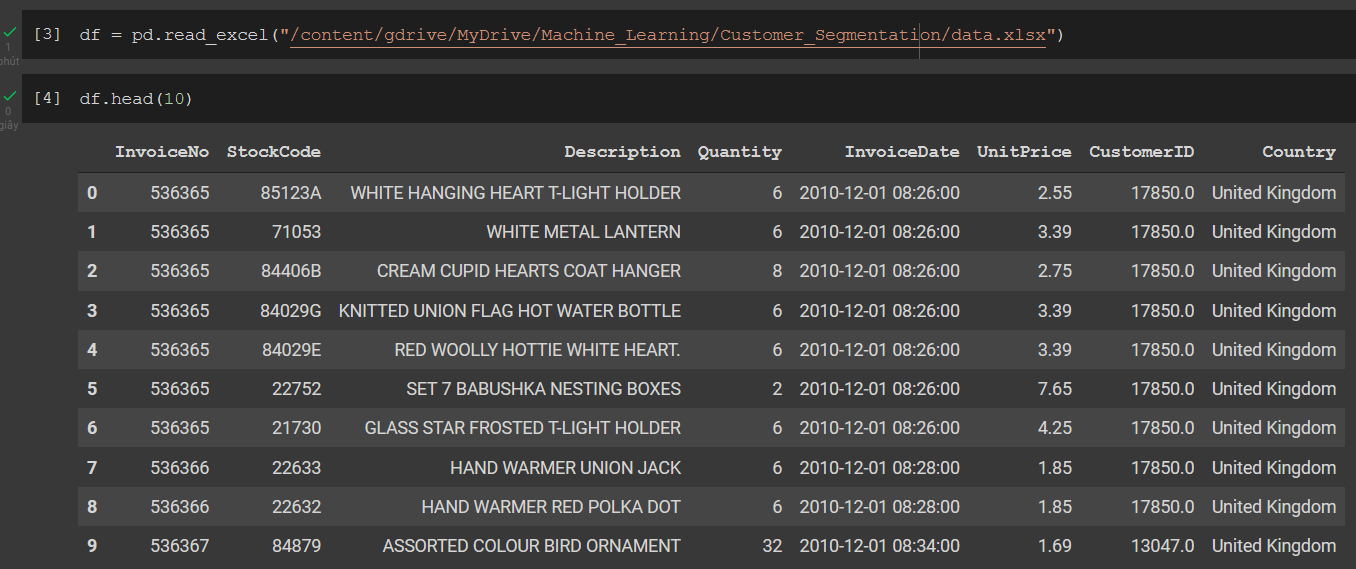
+ cube root transformation (căn bậc 3)

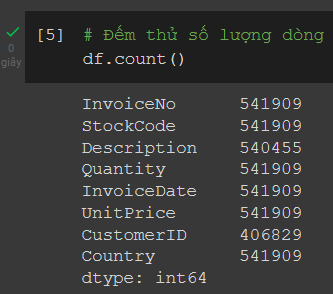
**3. Lập trình**

**1. Import một số thư viện cần thiết**

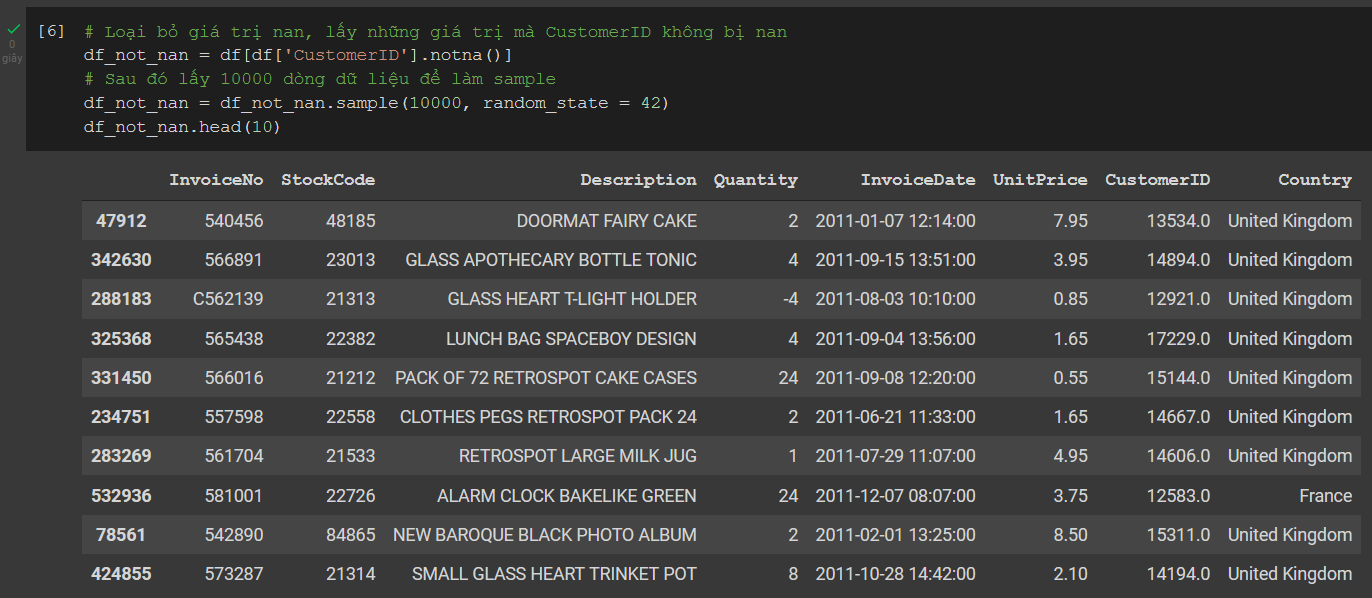


**2. Đọc dữ liệu file excel:**



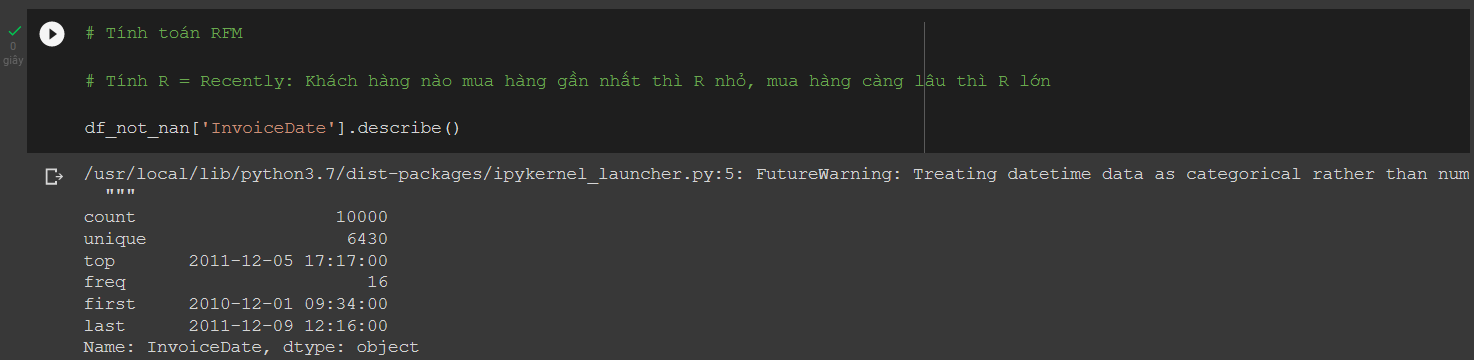


Làm bài toán Sample với khoảng 10000 dòng:

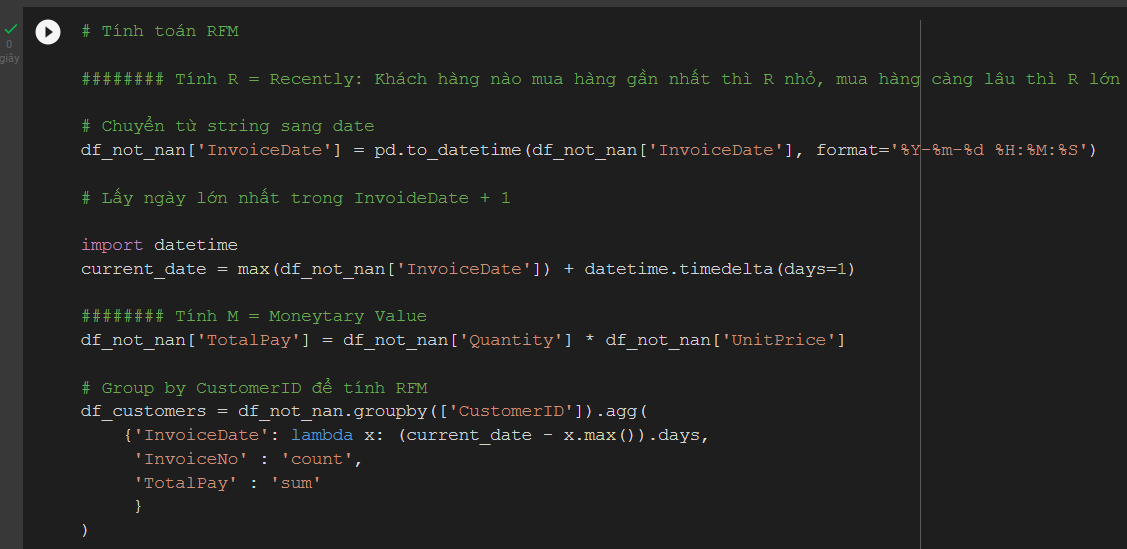


**3. Tính toán RFM**

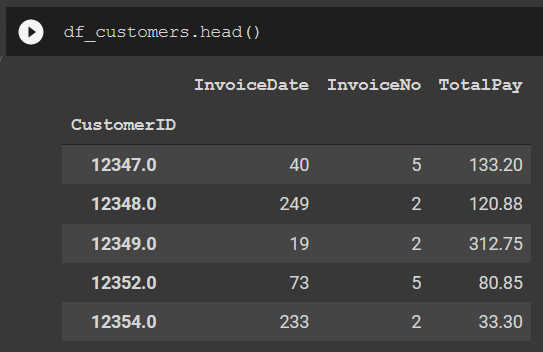
Tính R: Recently: Khách hàng nào mua hàng gần nhất thì R nhỏ, mua hàng càng lâu thì R lớn



Xem mô tả của cột InvoideDate, ta thấy kiểu dữ liệu của nó đang ở dạng Object không phải dạng date.

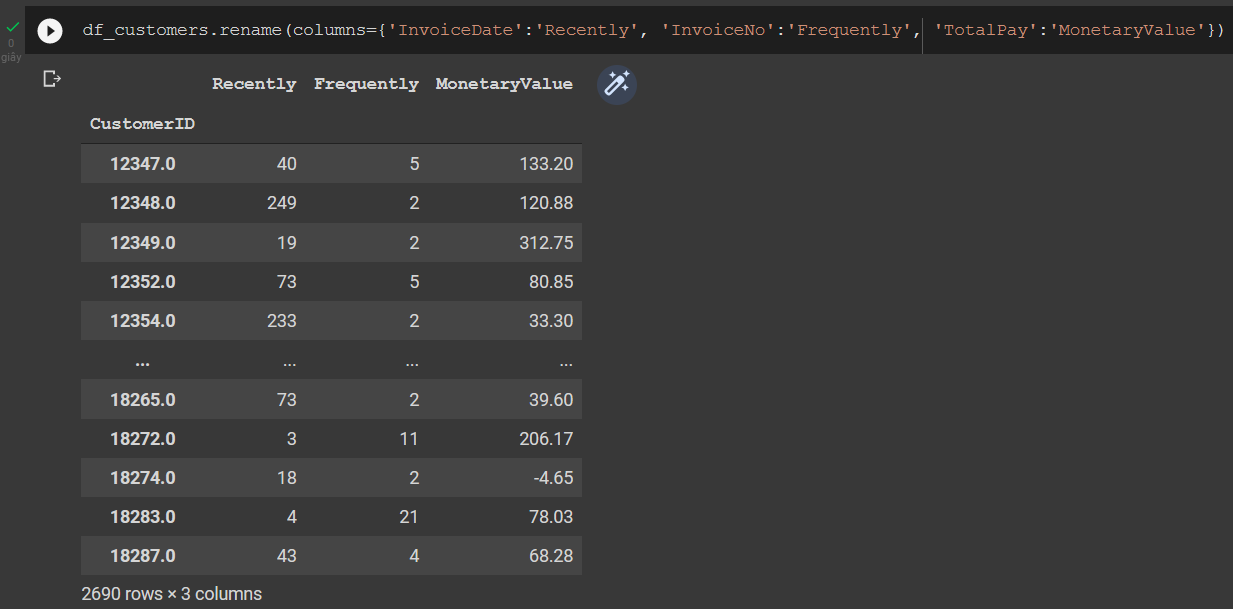


Bây giờ ta có một data frame customer với cột InvoiceDate thể hiện khách hàng mua hàng xa hay gần với thời điểm hiện tại, InvoiceNo thể hiện số lượng đơn hàng khách hàng đã mua, TotalPay là tổng tiền khách hàng đã chi để mua hàng.



Rename các cột lại cho đúng ý nghĩa:



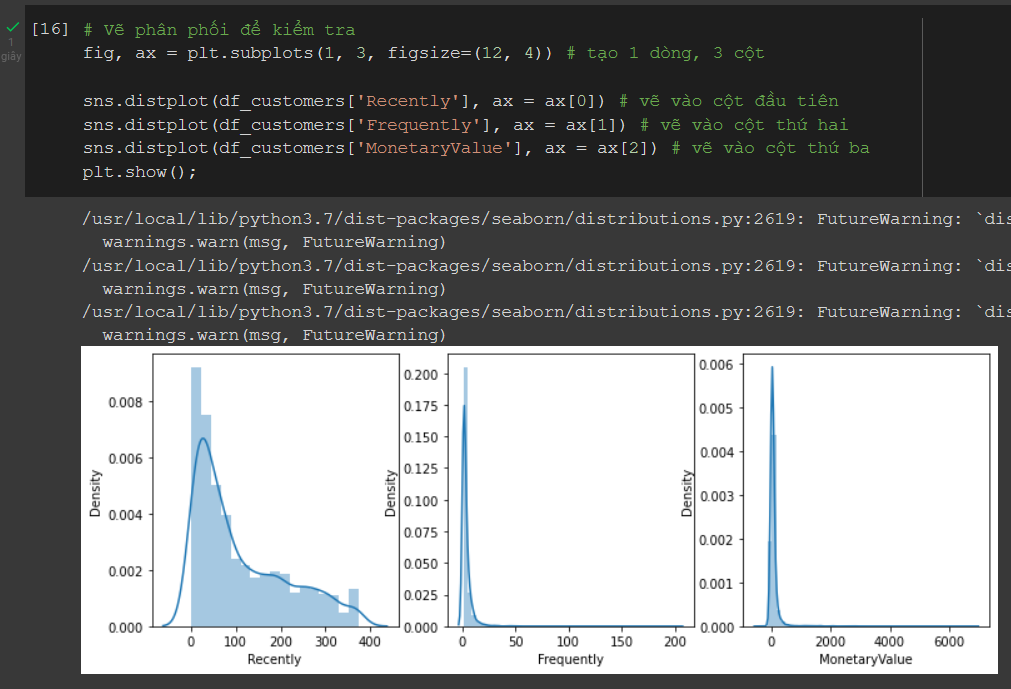


Bây giờ ta có thể dùng để đưa vào model KNN để phân cụm khách hàng, tuy nhiên hiệu quả sẽ không cao vì:

+ Các dữ liệu đang chưa được scale về một khoảng giống nhau.

+ Chưa biết phân phối dữ liệu.

**4. Vẽ phân phối để kiểm tra dữ liệu và áp dụng các biện pháp Transform**



Ta thấy dữ liệu chưa phân bố chuẩn. Ta phải áp dụng các biện pháp Transform.

Hàm kiểm tra xem áp dụng phương pháp nào để ra được phân phối chuẩn:

from scipy import stats

def analyze\_skewness(x):

fig, ax = plt.subplots(2, 2, figsize=(5,5))

sns.distplot(df\_customers[x], ax=ax[0,0])

sns.distplot(np.log(df\_customers[x]), ax=ax[0,1])

sns.distplot(np.sqrt(df\_customers[x]), ax=ax[1,0])

sns.distplot(stats.boxcox(df\_customers[x])[0], ax=ax[1,1])

plt.tight\_layout()

plt.show()

print(df\_customers[x].skew().round(2))

print(np.log(df\_customers[x]).skew().round(2))

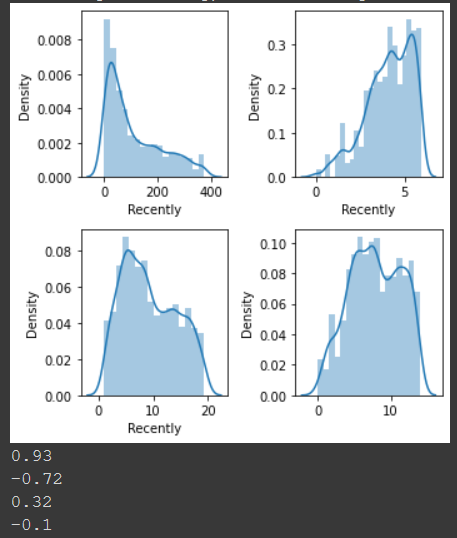
print(np.sqrt(df\_customers[x]).skew().round(2))

print(pd.Series(stats.boxcox(df\_customers[x])[0]).skew().round(2))

**Với cột Recently:** Xem các phân bố khi áp dụng các biện pháp Transformation bằng lệnh:

analyze\_skewness('Recently')

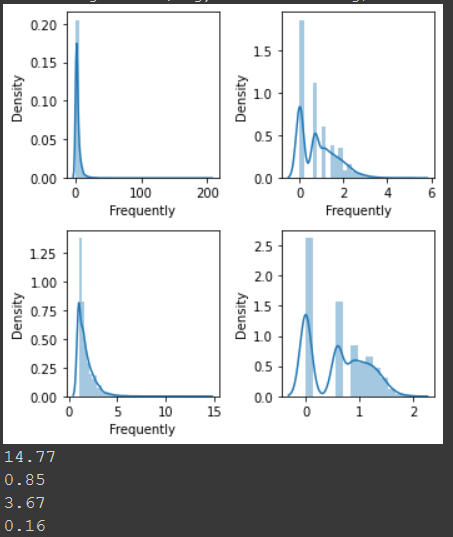
Kết quả: ta thấy box-cox transformation có chỉ số skewness là -0,1 gần với 0 nhất nên nó là tốt nhất.



**Với cột Frequently:**

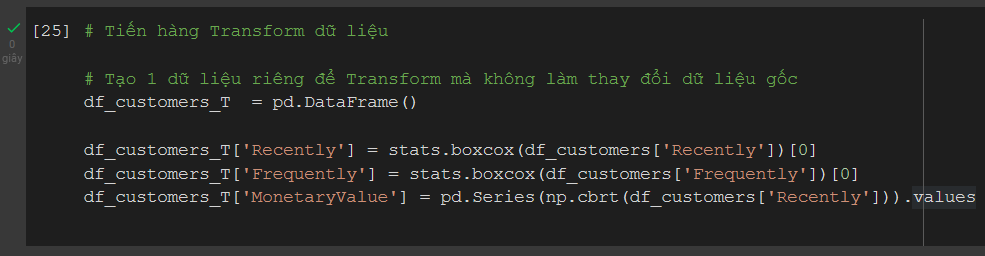
analyze\_skewness('Frequently')

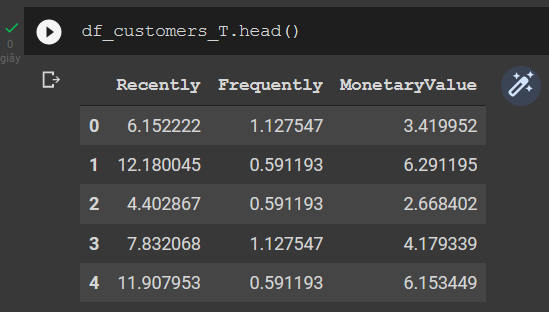
Kết quả: Ta thấy box-cox Transformation vẫn cho kết quả tốt nhất (skewness = 0,16)



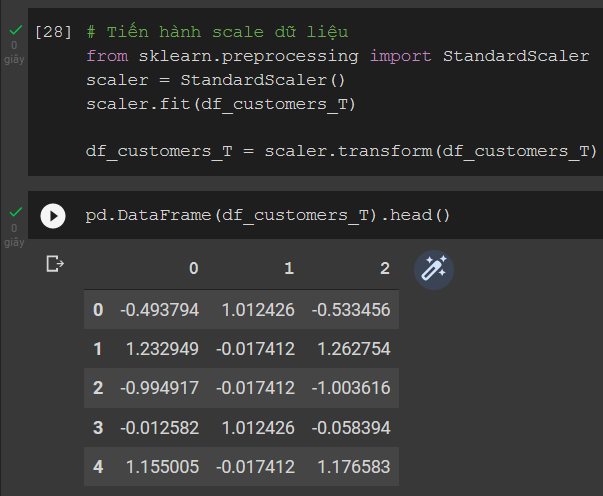
**Với cột MonetaryValue:** vì có giá trị âm nên ta không dùng được các biện pháp Transform như box-cox, sqr root hay log mà ta phải dùng cube root (căn bậc 3).

Tiến hàng Transform dữ liệu:



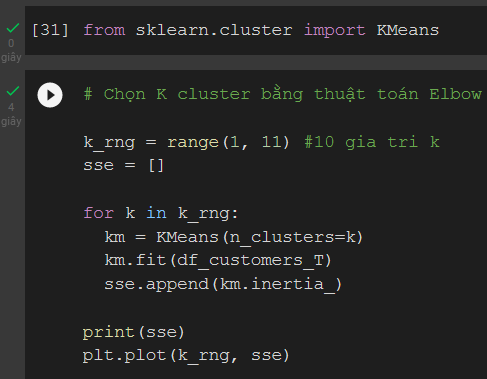


5. Scale dữ liệu

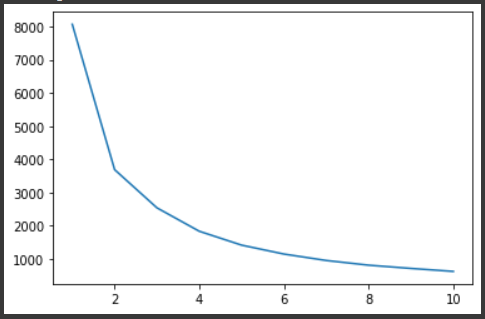


**6. Đưa dữ liệu vào K-means để phân khúc khách hàng**

🡺 *Tiến hành chọn số cụm bằng phương pháp Elbow:*

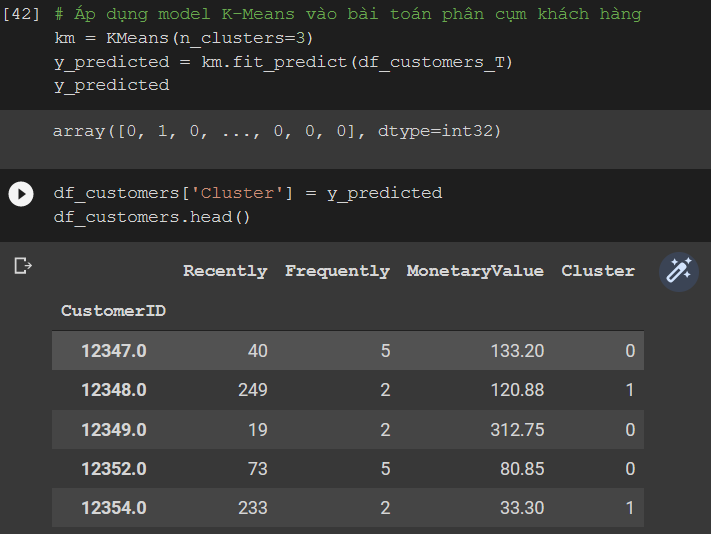


Kết quả:

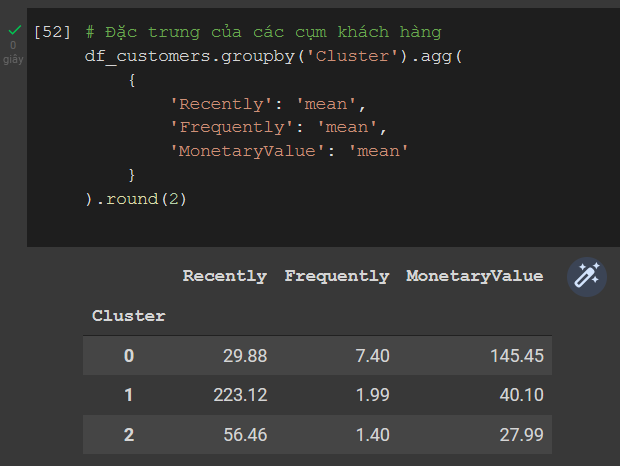


Chọn k sao cho khi k tăng lên 1 thì giá trị SSE không giảm đi quá nhiều. Vì vậy dựa vào đồ thị trên ta chọn k = 3 là phù hợp.

Áp dụng thuật toán phân cụm K-means:



Đặc trưng của từng nhóm khách hàng:



**Nhóm khách hàng 0:** Mua gần đây, số lần mua khá nhiều và tổng chi phí họ mua rất lớn 🡪 Một số khách hàng tiềm năng gần đây, cần có những biện pháp tri ân khách hàng.

**Nhóm khách hàng 1:** Mua hàng rất lâu rồi, số lần mua khá ít và tổng chi phí họ mua ở mức trung bình 🡪 đây có thể là nhóm khách hàng mới sử dụng dịch vụ nên e ngại mua số lượng lớn và họ chỉ dừng ở mức mua để dùng thử.

**Nhóm khách hàng 2:** Mua hàng khá gần, số lần mua ít và chi ít tiền.