GVHD: LÊ NGỌC THÀNH, NGUYỄN NGỌC THẢO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
THÀNH PHỐ HỒ CHÍ MINH

BÀI TẬP 4: CLUSTERING

I. THÔNG TIN SINH VIÊN

Họ và tên: **TRẦN NHẬT HUY**

Mssv: **1612272**

Email: nhathuy13598@gmail.com

Sđt: **0354 878 677**

II. BẢNG BÁO CÁO CÔNG VIỆC

STT	CÁC CÂU HỎI	MỨC ĐỘ HOÀN THÀNH	GHI CHÚ
1	Thử nghiệm với các k từ 3 đến 8	100%	
2	Chúng ta giới thiệu sản phẩm nào và giải thích	100%	
3	Chúng ta giới thiệu sản phẩm nào và giải thích	100%	
4	Kết quả gom cụm có thể nhận diện được hình mẫu người nào hay không	100%	
5	Cụm nào thể hiện sở thích mua hàng cụ thể của người dùng	100%	
6	Nhận diện cụm nào tương ứng với người dùng bị quảng cáo thu hút	100%	
7	Cài đặt chương trình	100%	Link tham khảo

III. CHI TIẾT BÀI LÀM

1. Thử nghiệm với các k khác nhau

k	SSE		Cluster centroids							
			Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
3	128.858	1	0.0938	1	0.375	0.0313	1	0.6563	0.4375	0.2813
		2	0.7838	0.6757	0	0.8919	0.3784	0.2162	0.6757	0.4324
		3	0.9032	0.4839	1	0.6129	0.2903	0.5161	0.7097	0.4516
4	121.7767	1	0.0645	1	0.3548	0.0323	1	0.6774	0.4516	0.2903
		2	0.75	0.7273	0.2273	0.9773	0.3409	0.2955	0.7727	0.5455
		3	1	0.2222	1	1	0.6667	0.7778	0.4444	0.2222
		4	1	0.4375	0.8125	0	0.1875	0.25	0.5625	0.25
5	113.5826	1	0.9615	0.6923	0.6538	0.4615	0.3846	0.5385	0.4615	0
		2	0.6667	0.6667	0	0.963	0.4444	0	0.6296	0.5185
		3	1	0	1	1	0.8	0.8	0.8	0.4
		4	0.8571	0.5714	0.8571	0.7143	0.0714	0.5714	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
6	109.3611	1	0.931	0.7241	0.7241	0.5862	0.3448	0.6552	0.5172	0.1379
		2	0.9583	0.875	0.0833	0.9167	0.3333	0.0833	0.7083	0.5833
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.1667	1	0.1667	0	0.3333	1	0.8333
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1111	0.1111	1	0.5556	0	0.6667	0.5556
7	93.79	1	0.9048	0.9048	0.7143	0.7619	0.381	0.9048	0.6667	0.1905
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.2875	1	0.1429	0	0.2857	1	0.7143
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	1	0.8571
		7	0.8235	0.4118	0.2941	0.6471	0.3529	0	0	0
8	88.9319	1	0.8889	1	0.6667	0.7778	0.3889	0.9444	0.6667	0.2222
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	0.6667	0.8333	0.6667	0.3333
		4	1	0.5	1	0	0	1	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	0	0.8575
		7	0.75	0.5	0	0.8333	0.3333	0	0	0
		8	1	0.2727	1	0.1818	0.2727	0	0.5455	0.2727

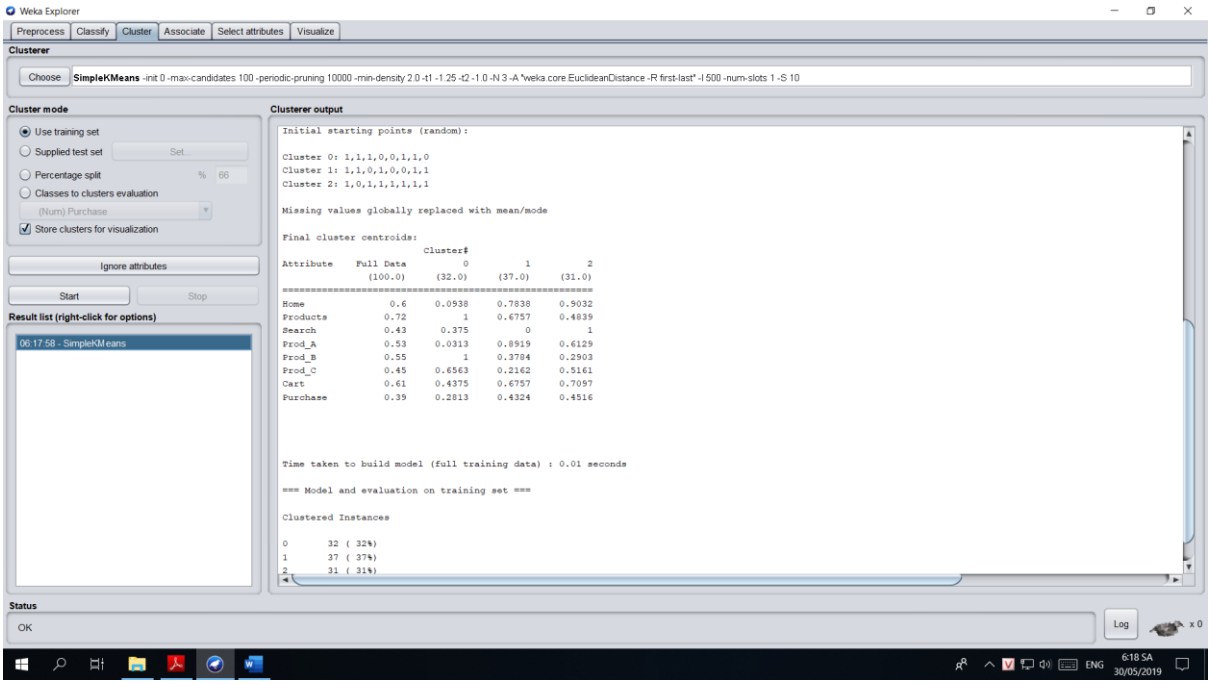


Figure 1: k = 3

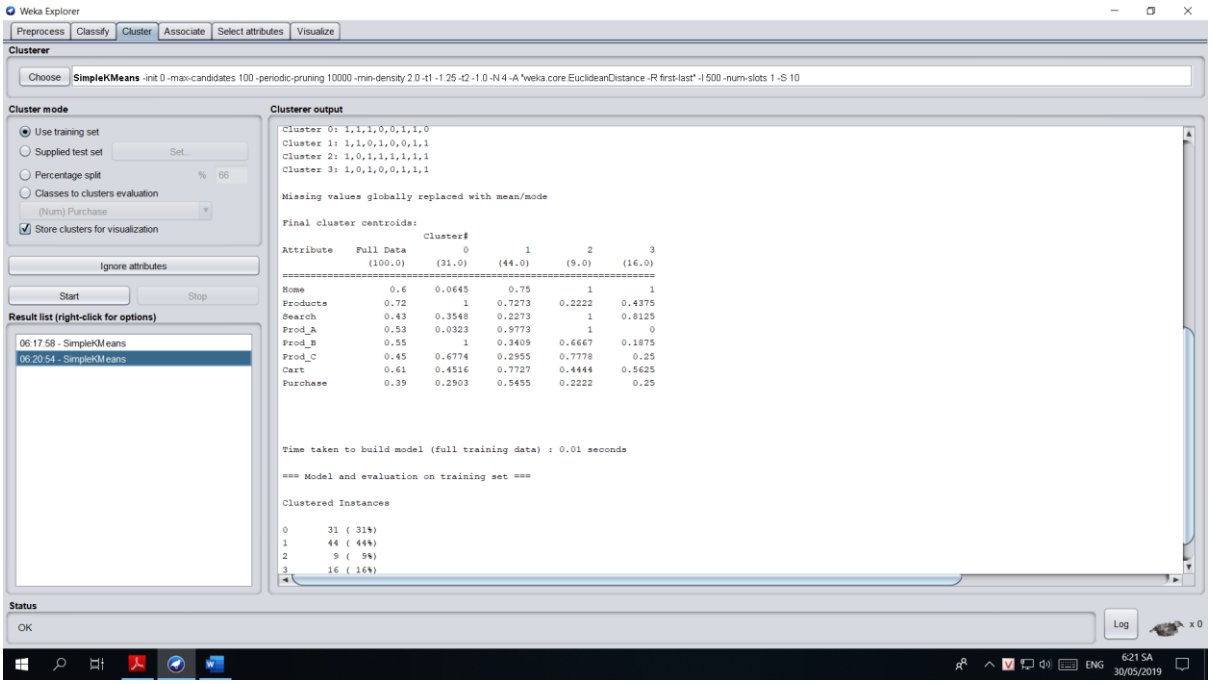


Figure 2: k = 4

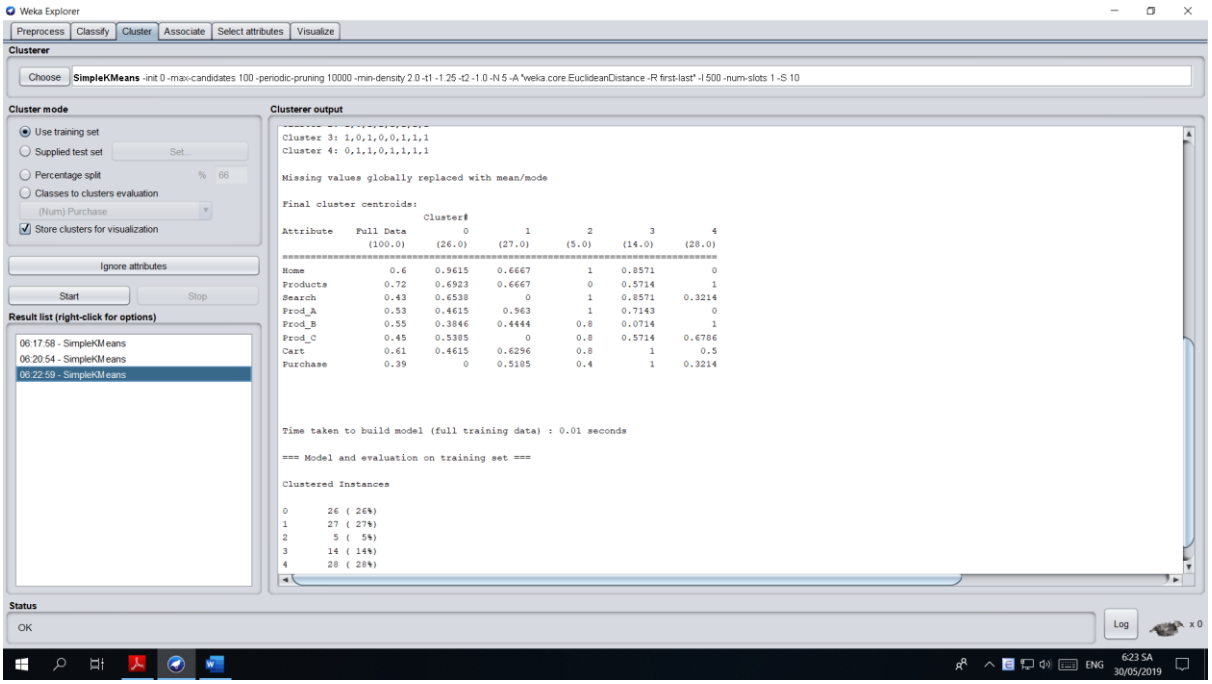


Figure 3: k = 5

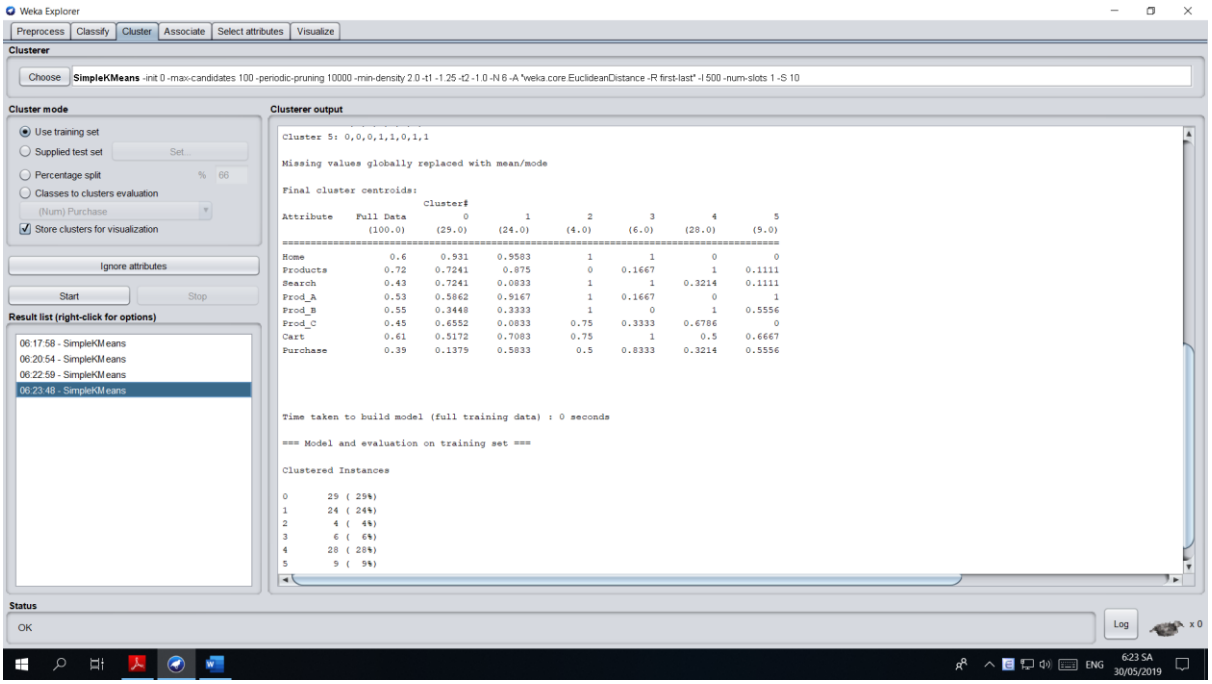


Figure 4: k = 6

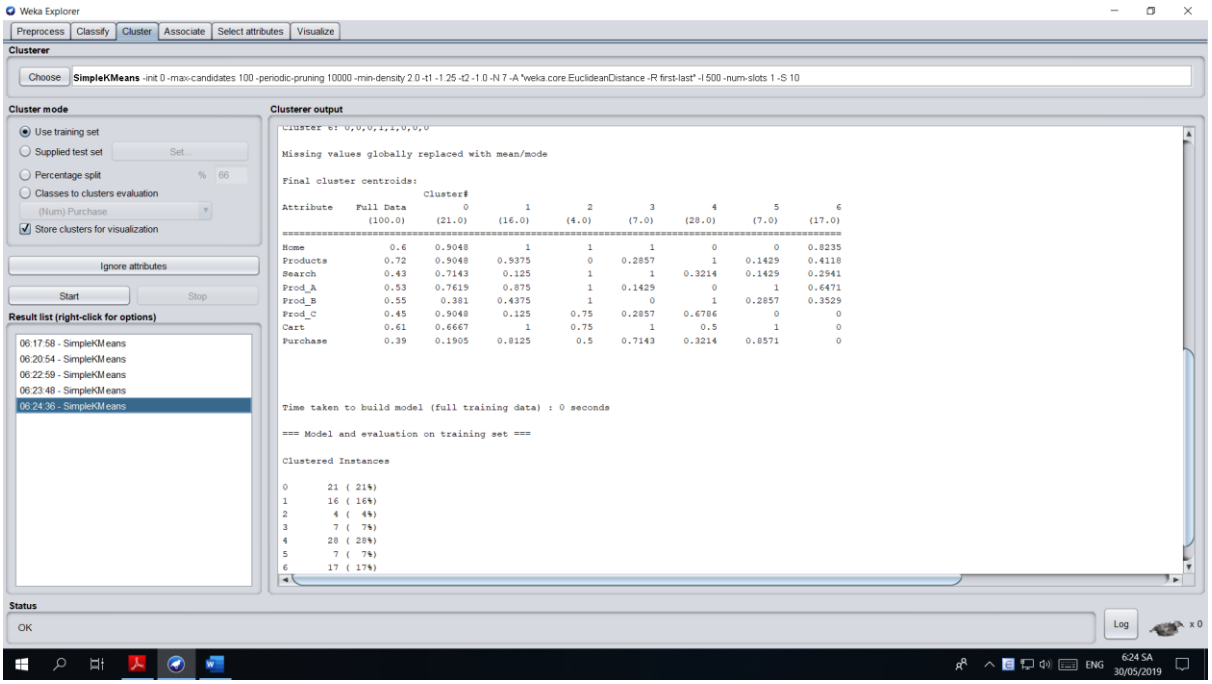


Figure 5: k = 7

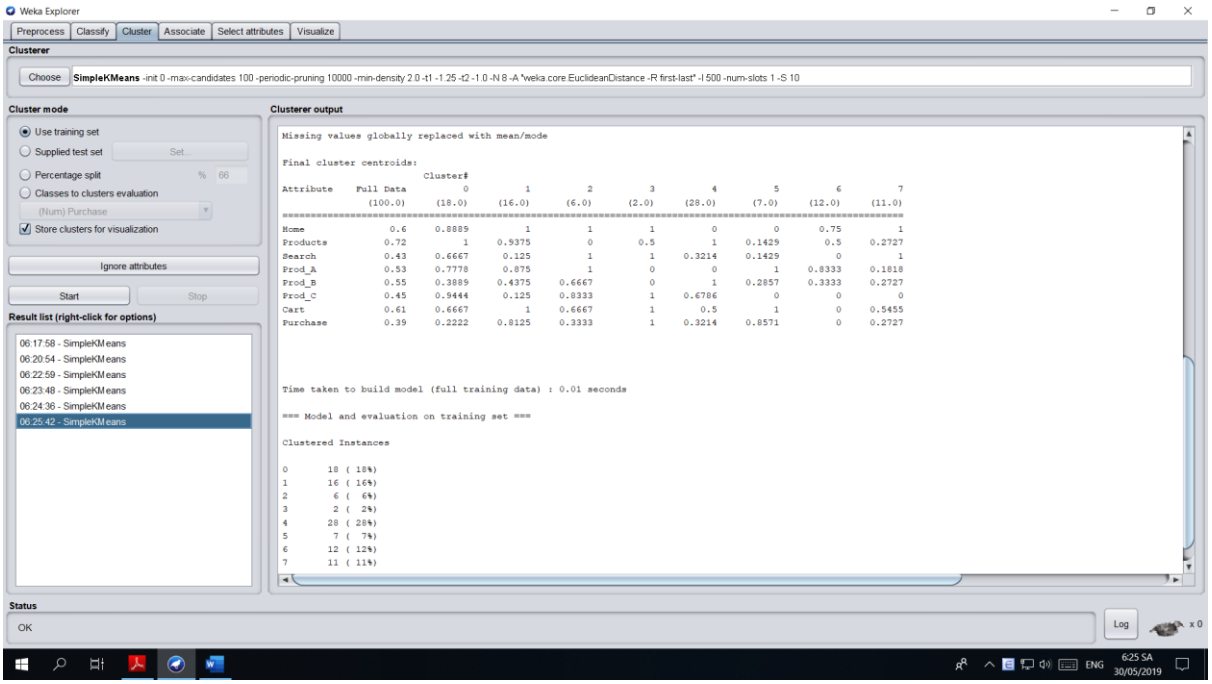


Figure 6: k = 8

2. Kiểm thử

Với mẫu quan sát là Home => Search => Prod_B thì ta có vector input như sau: [1, 0, 1, 0, 1, 0, 0, 0]

Ta sẽ tính khoảng cách từ mẫu quan sát đến tâm của các cluster

Cluster 0: [0.9615, 0.6923, 0.6538, 0.4615, 0.3846, 0.5385, 0.4615, 0]

Cluster 1: [0.6667, 0.6667, 0, 0.963, 0.4444, 0, 0.6296, 0.5185]

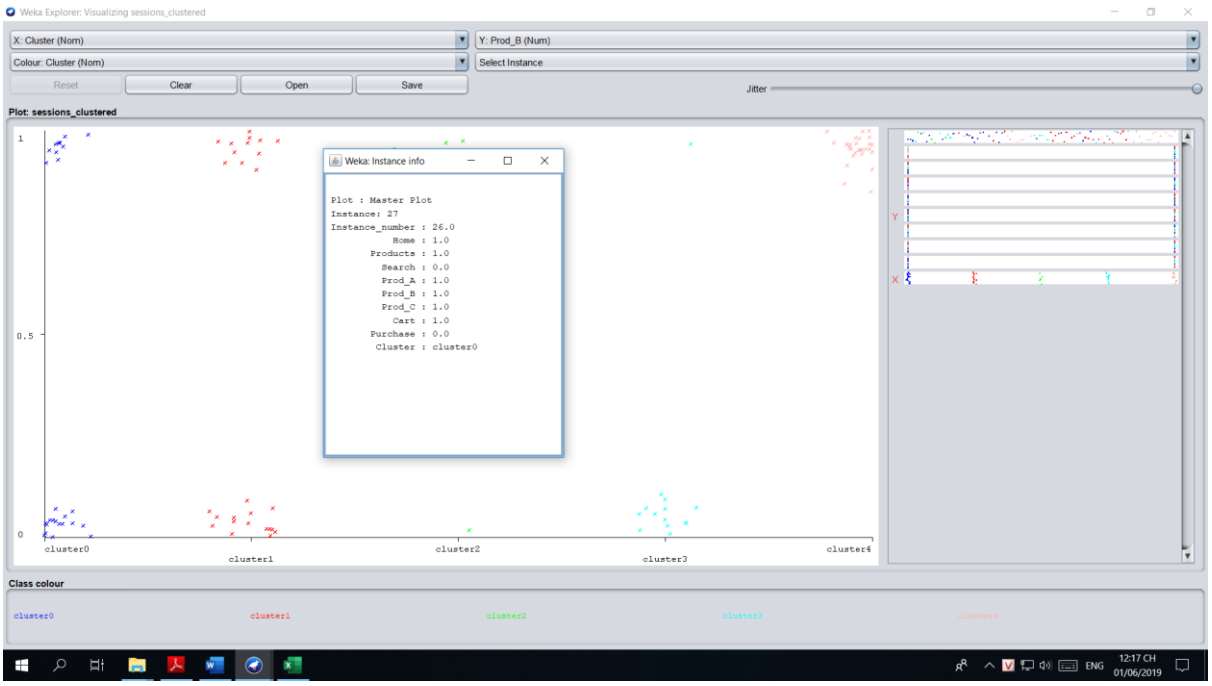
Cluster 2: [**1, 0, 1, 1, 0.8, 0.8, 0.8, 0.4**]

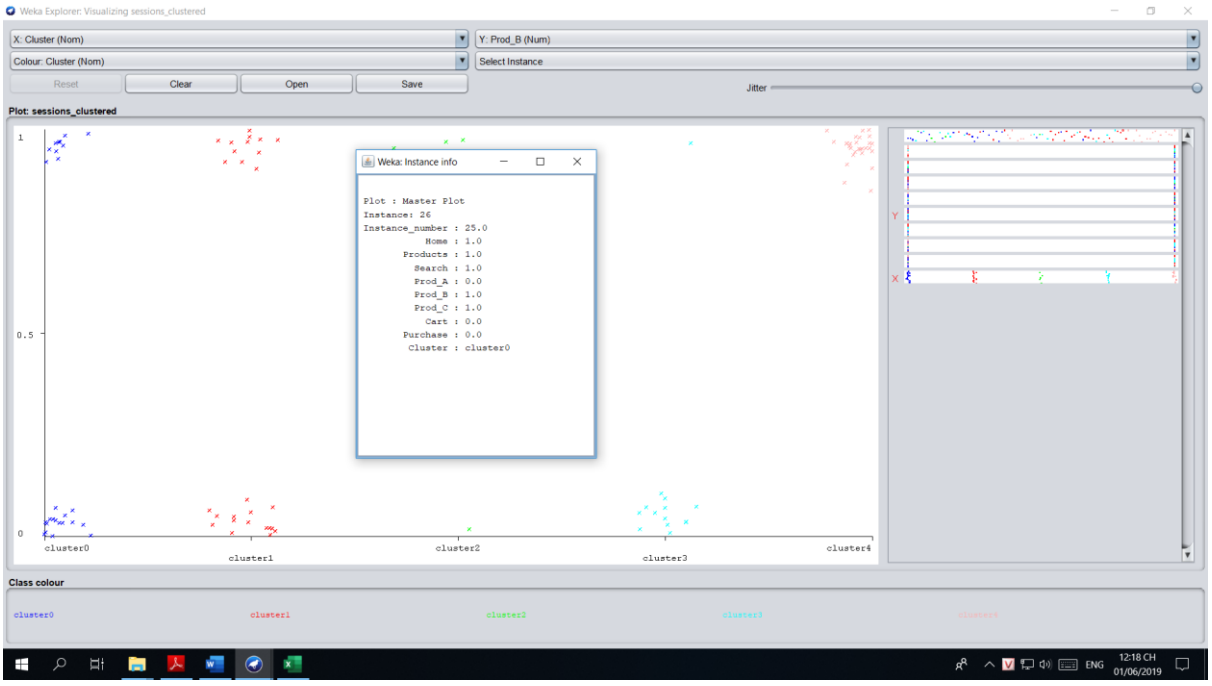
Cluster 3: [**0.8571, 0.5714, 0.8571, 0.7143, 0.0714, 0.5714, 1, 1**]

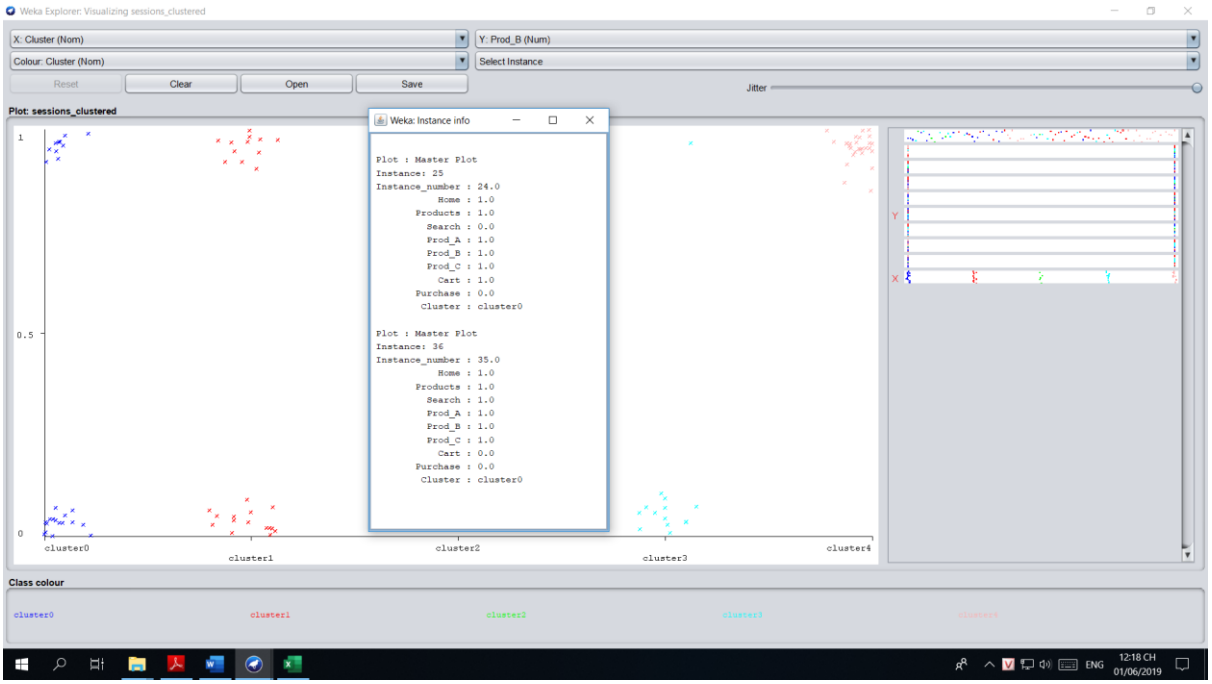
Cluster 4: [**0, 1, 0.3214, 0, 1, 0.6786, 0.5, 0.3214**]

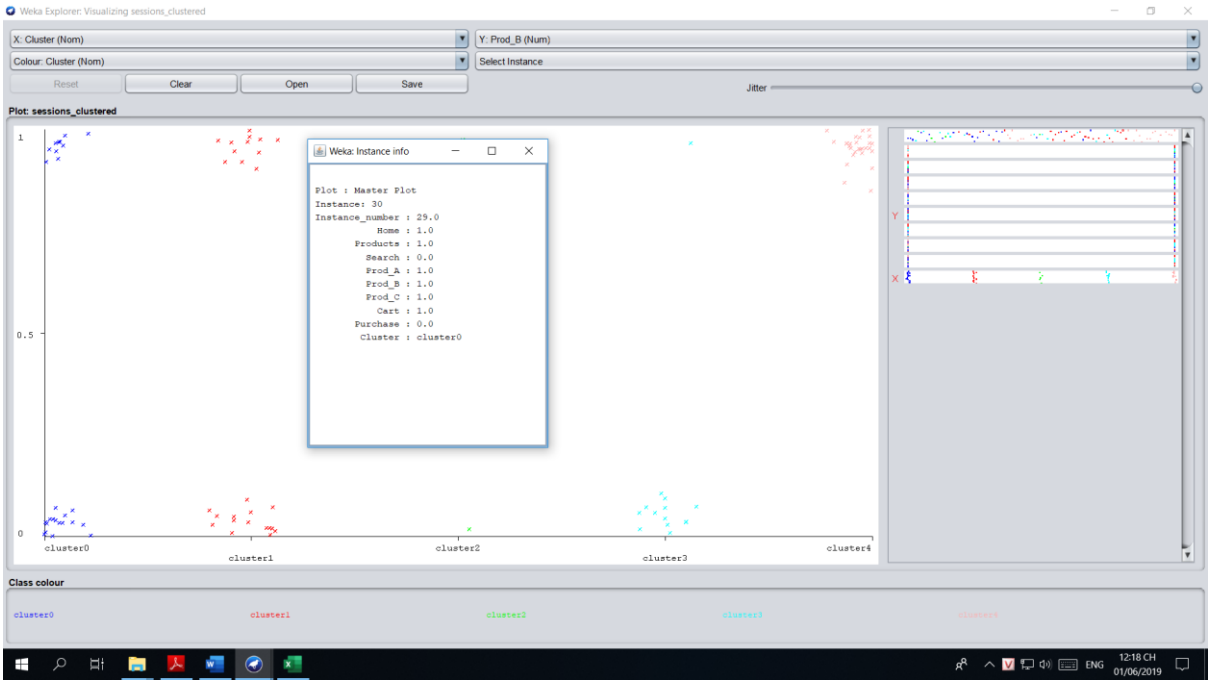
Khoảng cách ta tính được lần lượt là: **1.3020, 1.8593, 1.5748, 2.0165, 1.8094**

Vậy mẫu quan sát này thuộc **Cluster 0**. Do đó ta sẽ gợi ý sản phẩm tiếp theo cho khách hàng này là **Prod_C** và **Prod_A** vì dữ liệu trong cluster này có xu hướng chọn 2 sản phẩm còn lại nếu như chọn **Prod_B** (Ví dụ như trong hình các instance thuộc cluster 0)









3. Kiểm thử

Với mẫu quan sát là Product => Prod_C thì ta có vector input như sau: [**0**, **1**, **0**, **0**, **0**, **1**, **0**, **0**]

Ta sẽ tính khoảng cách từ mẫu quan sát đến tâm của các cluster

Cluster 0: [**0.9615**, **0.6923**, **0.6538**, **0.4615**, **0.3846**, **0.5385**, **0.4615**, **0**]

Cluster 1: [**0.6667**, **0.6667**, **0**, **0.963**, **0.4444**, **0**, **0.6296**, **0.5185**]

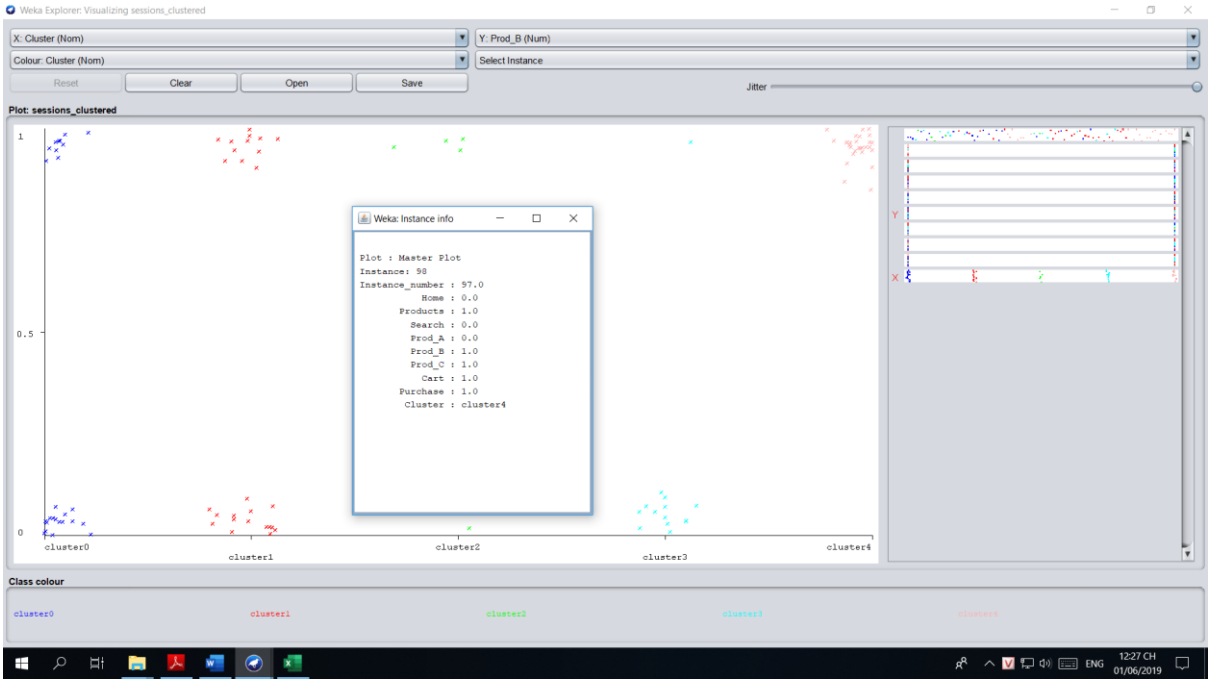
Cluster 2: [**1**, **0**, **1**, **1**, **0.8**, **0.8**, **0.8**, **0.4**]

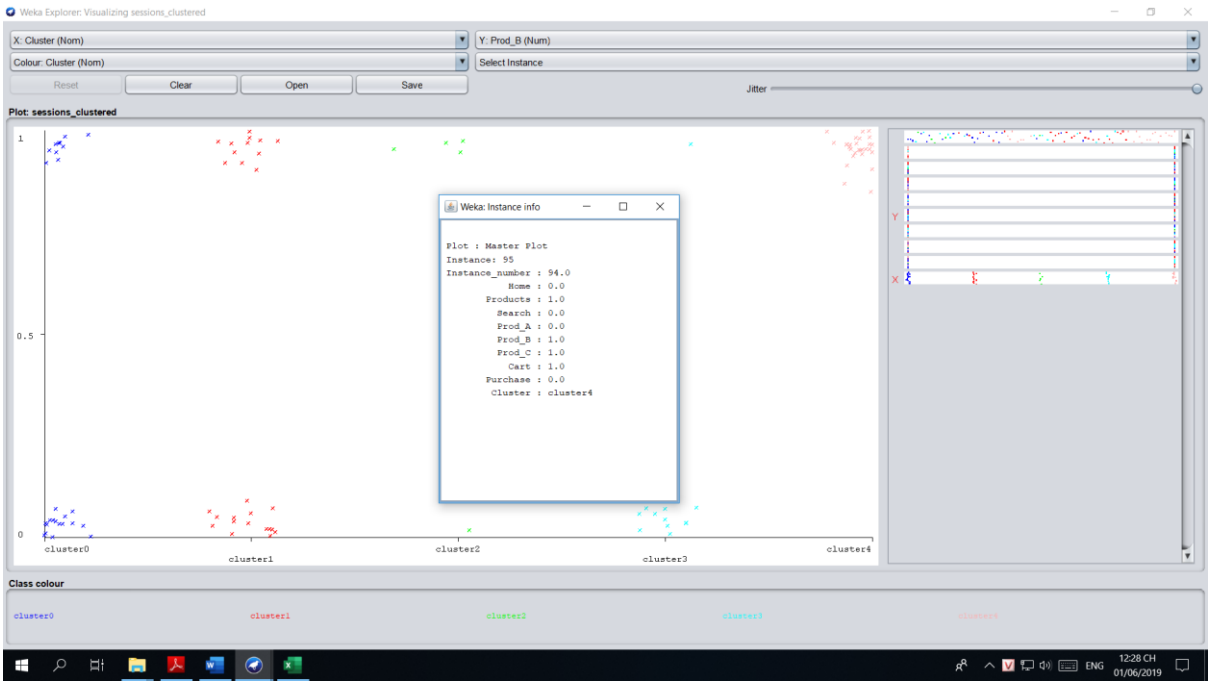
Cluster 3: [**0.8571, 0.5714, 0.8571, 0.7143, 0.0714, 0.5714, 1, 1**]

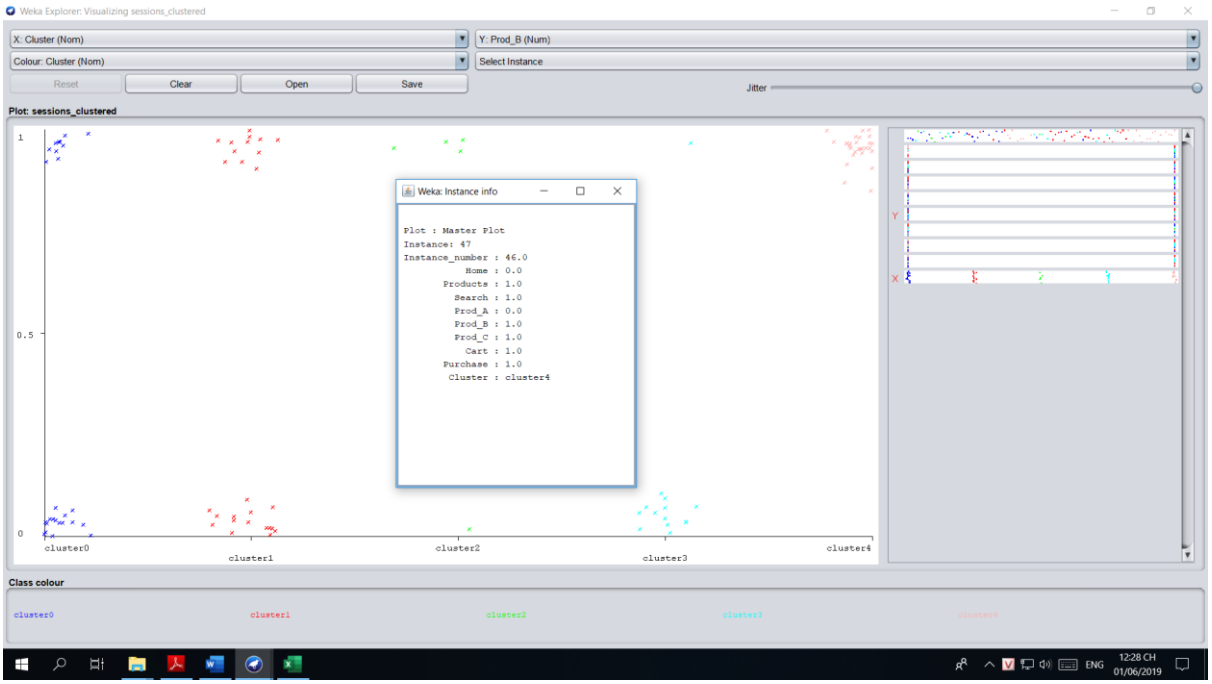
Cluster 4: [**0, 1, 0.3214, 0, 1, 0.6786, 0.5, 0.3214**]

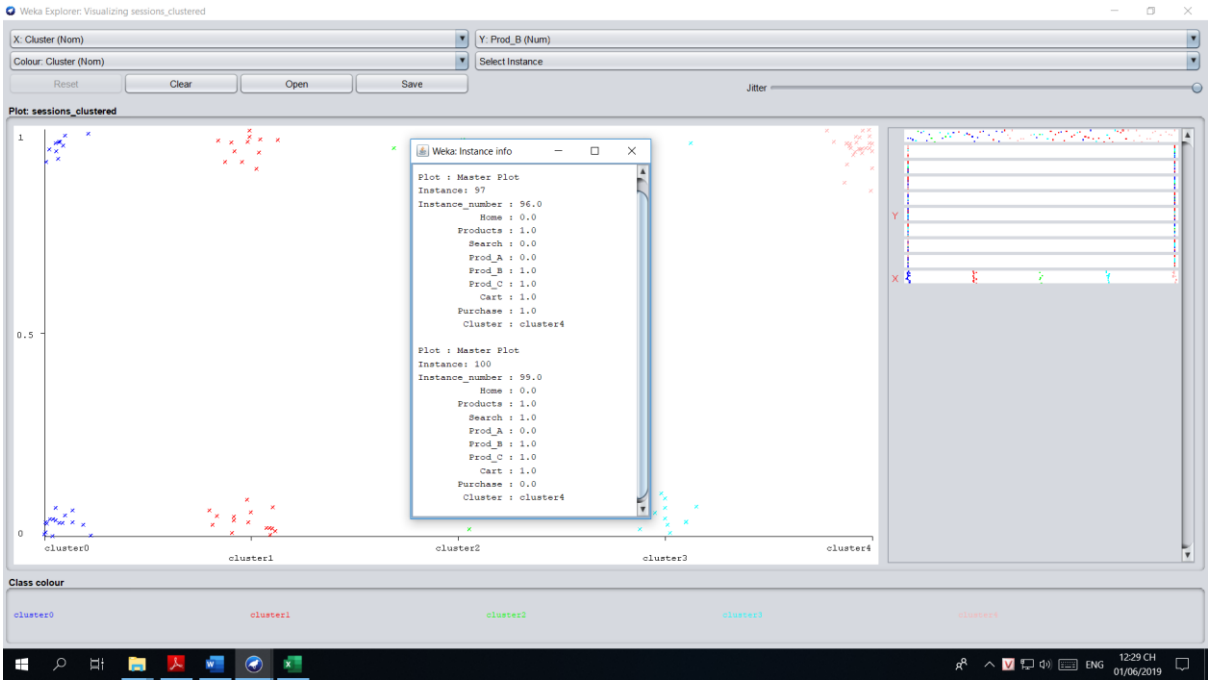
Khoảng cách ta tính được lần lượt là: **1.4945, 1.8291, 2.341, 2.0861, 1.25**

Vậy mẫu quan sát này thuộc **Cluster 4**. Do đó ta sẽ gợi ý sản phẩm tiếp theo cho khách hàng này là **Prod_C** vì dữ liệu trong cluster này có xu hướng chọn sản phẩm **Prod_C** nếu như chọn **Prod_B** (Ví dụ như trong hình các instance thuộc cluster 4)



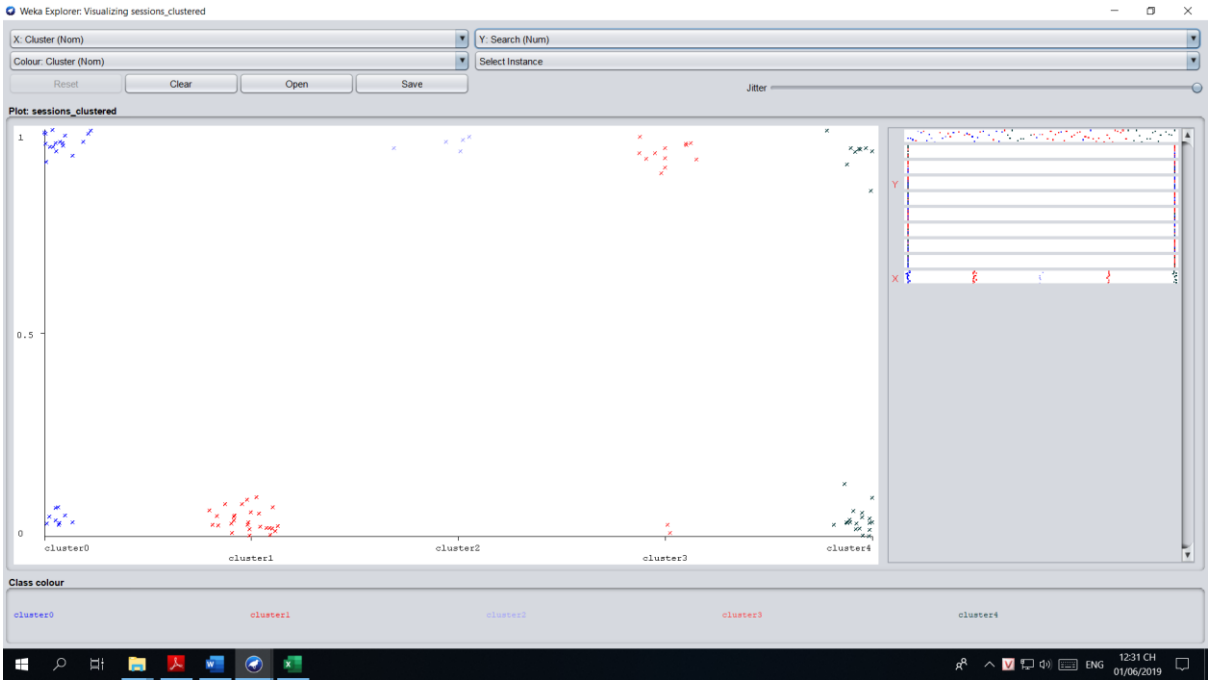






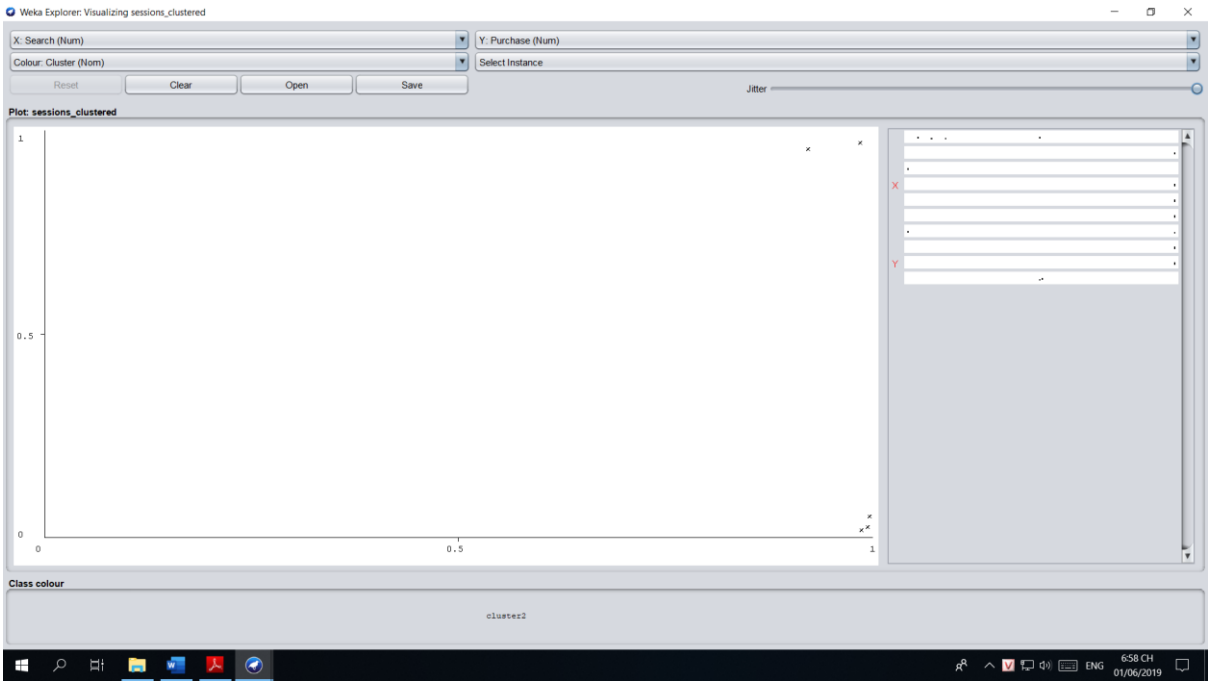
4. Nhận diện mẫu người

Chuyển trục **X thành Cluster** và trục **Y thành Search** (**Người dùng tìm kiếm**)

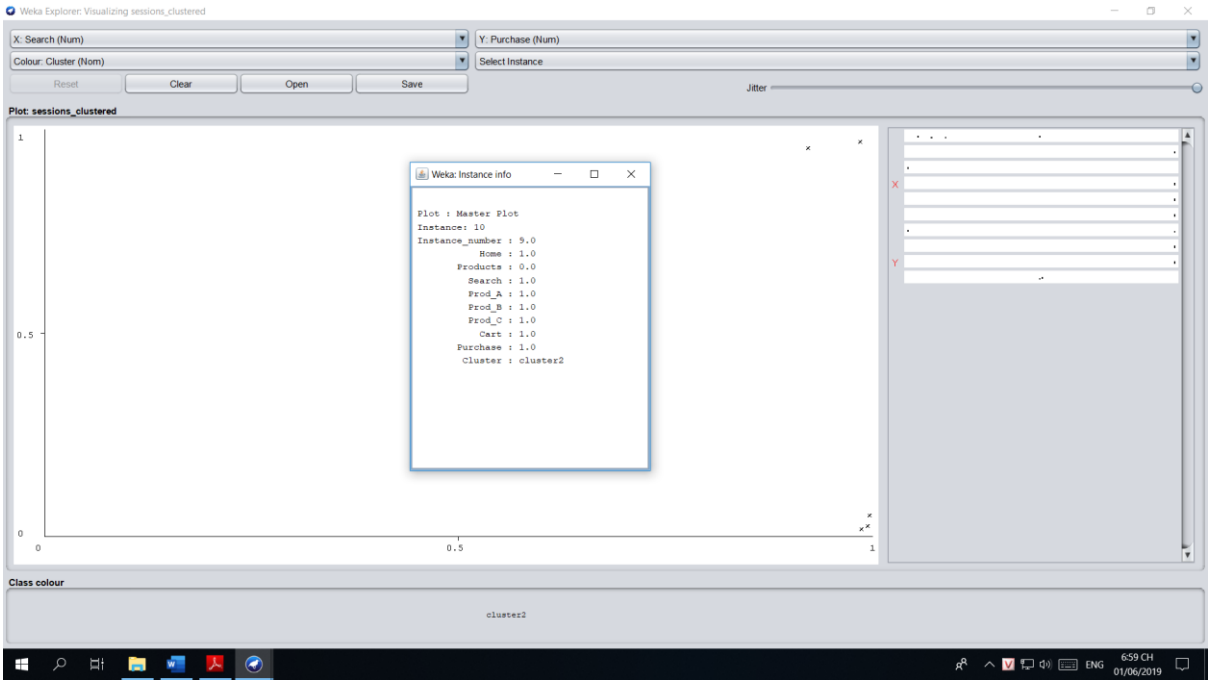


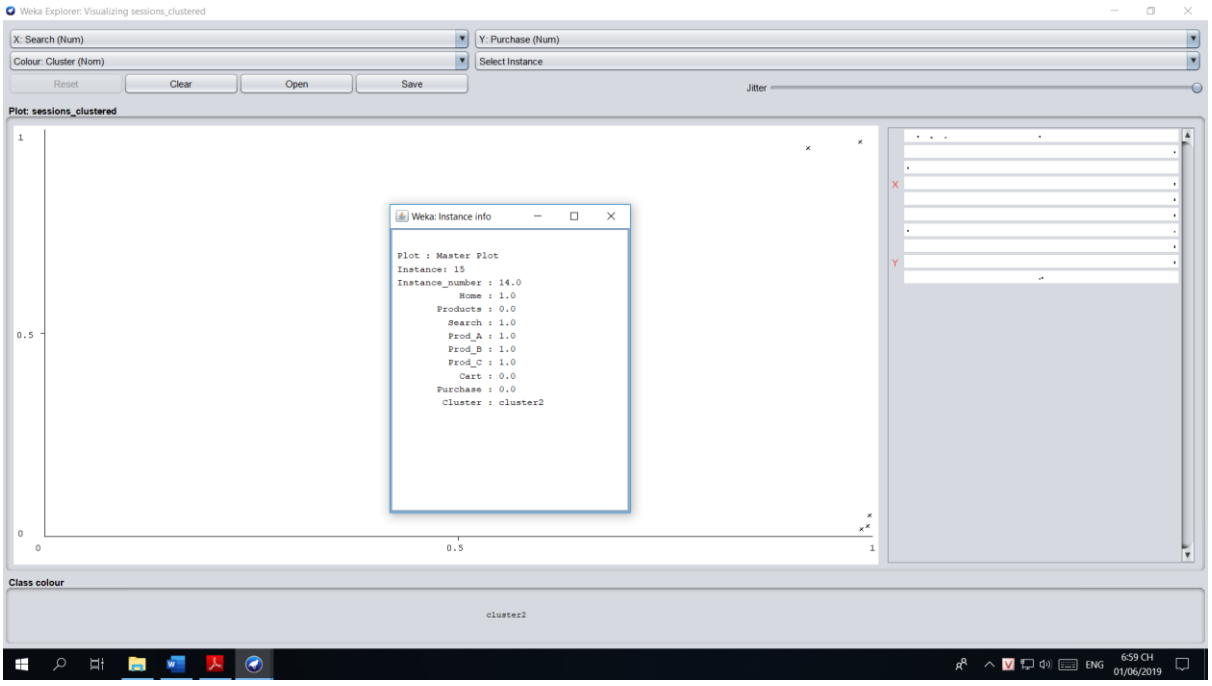
Ta thấy **cluster 2** và **cluster 3** là nhóm người thường sử dụng chức năng tìm kiếm nhiều nhất

Cluster 2 thì số người mua và không mua chia đều nên khó nhận xét. Cụ thể ta chuyển trục **X thành Search**, trục **Y thành Purchase** và màu các cluster khác 2 thành trong suốt.

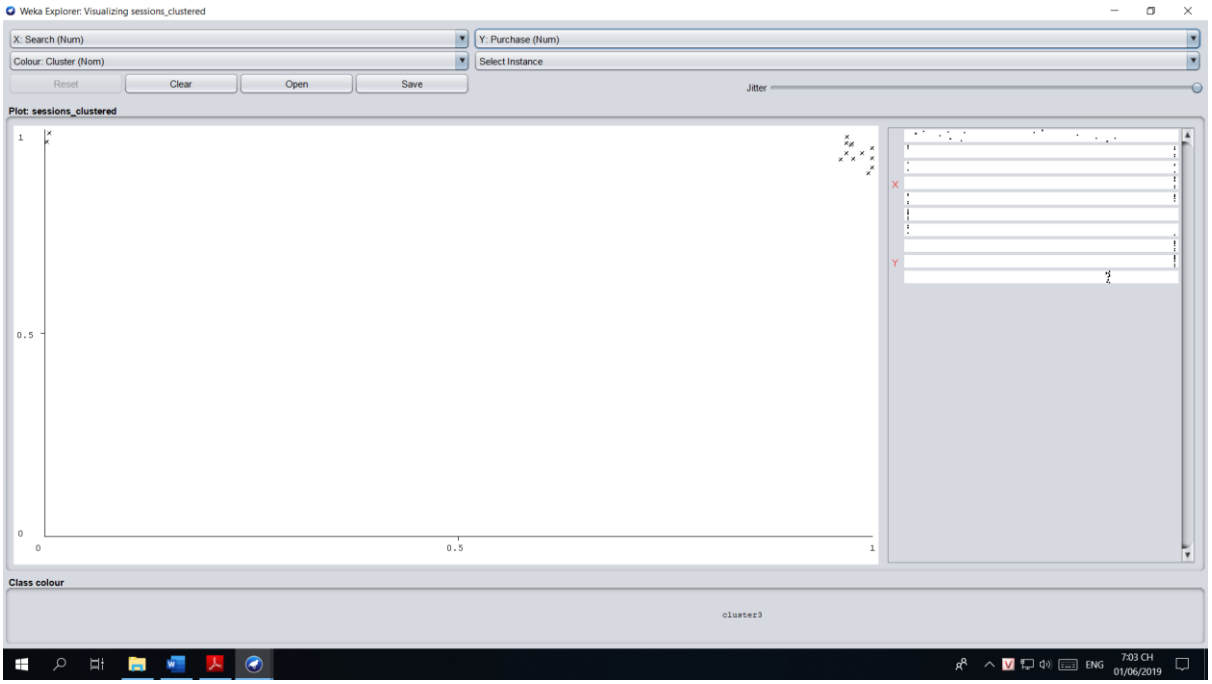


Một số mẫu cụ thể :



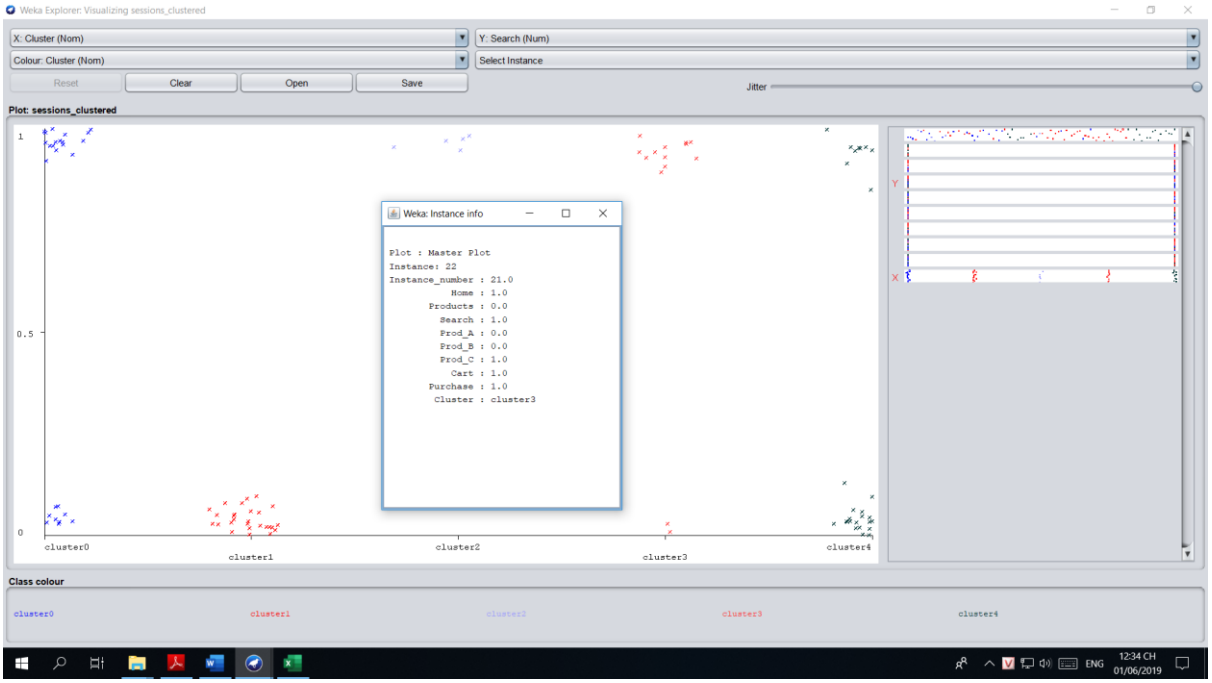


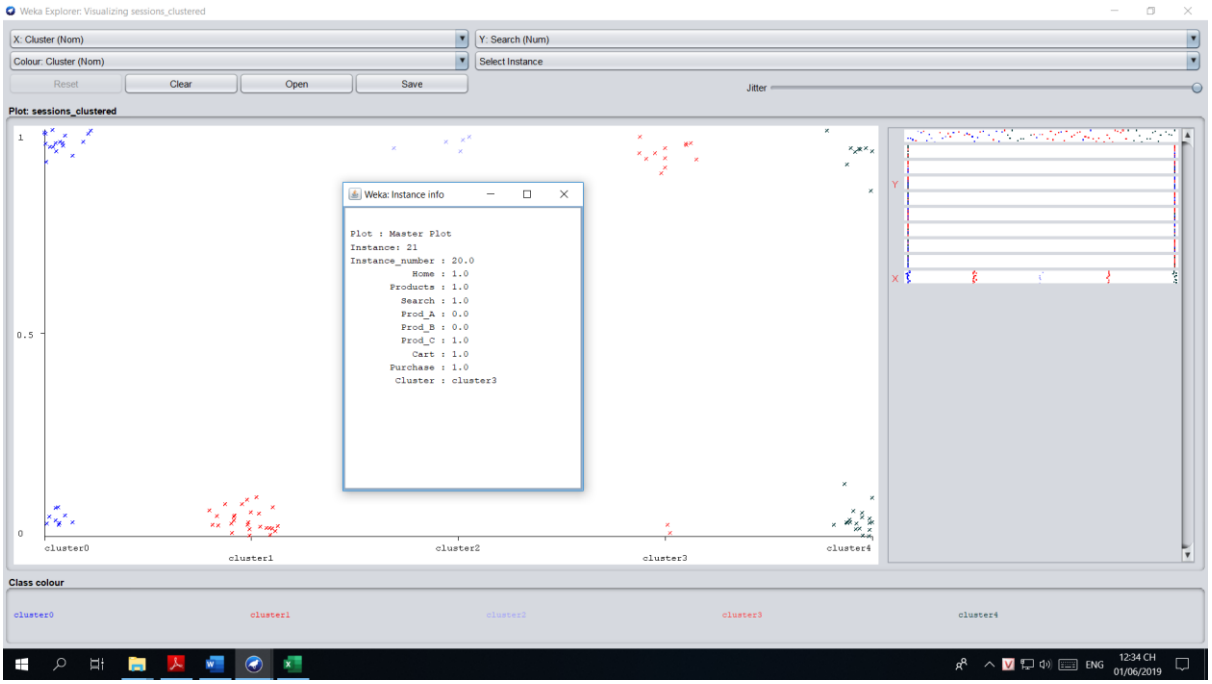
Cluster 3 thì có xu hướng là mua. Cụ thể ta chuyển trục **X thành Search**, trục **Y thành Purchase** và màu các cluster khác 3 thành trong suốt.

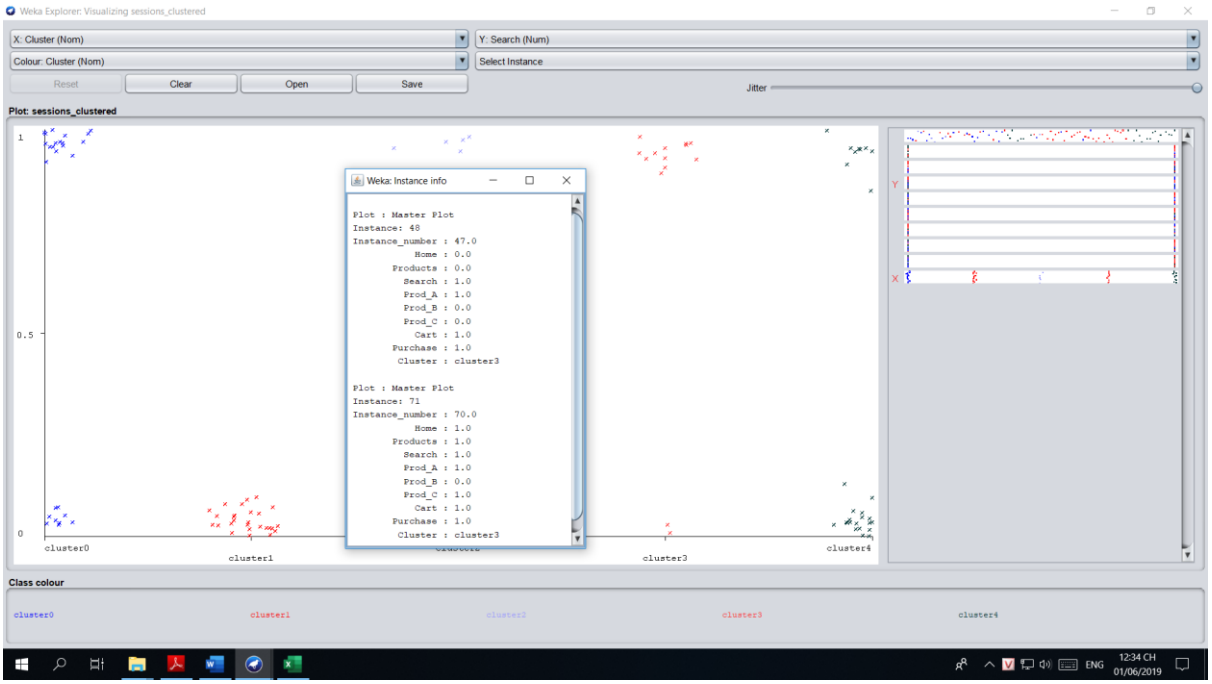


Ta thấy nếu như khách hàng tìm kiếm thì khách hàng đó sẽ mua chiếm đa số trong cluster này

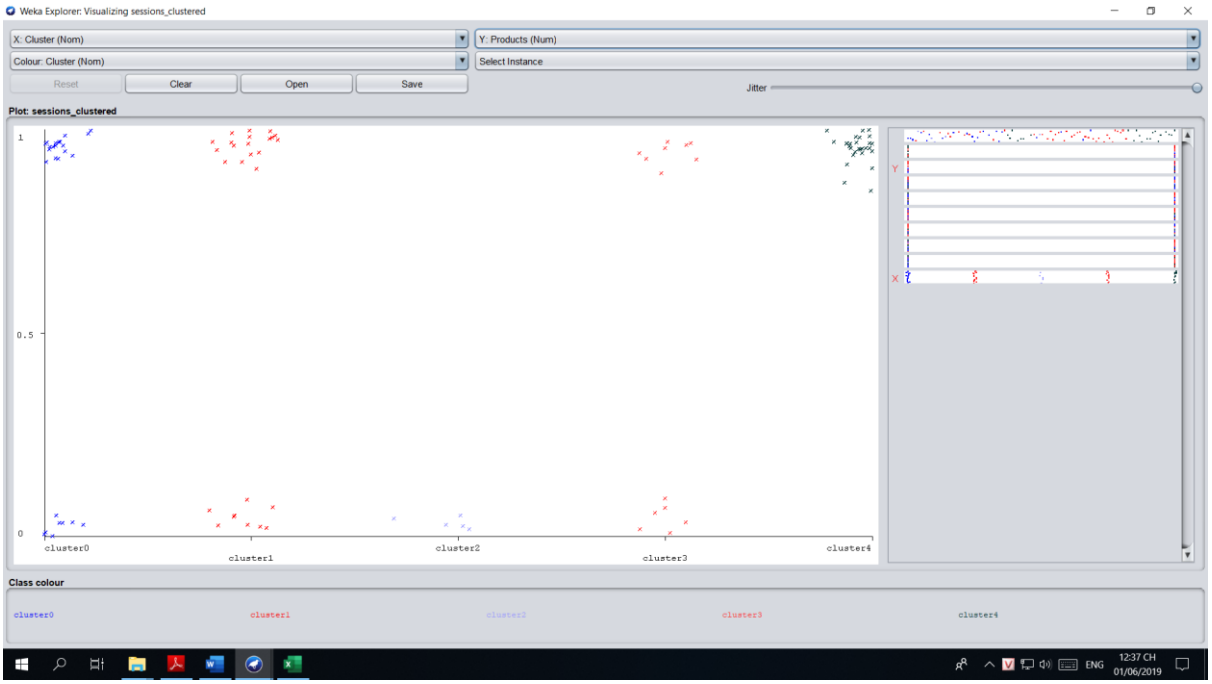
Một số mẫu cụ thể





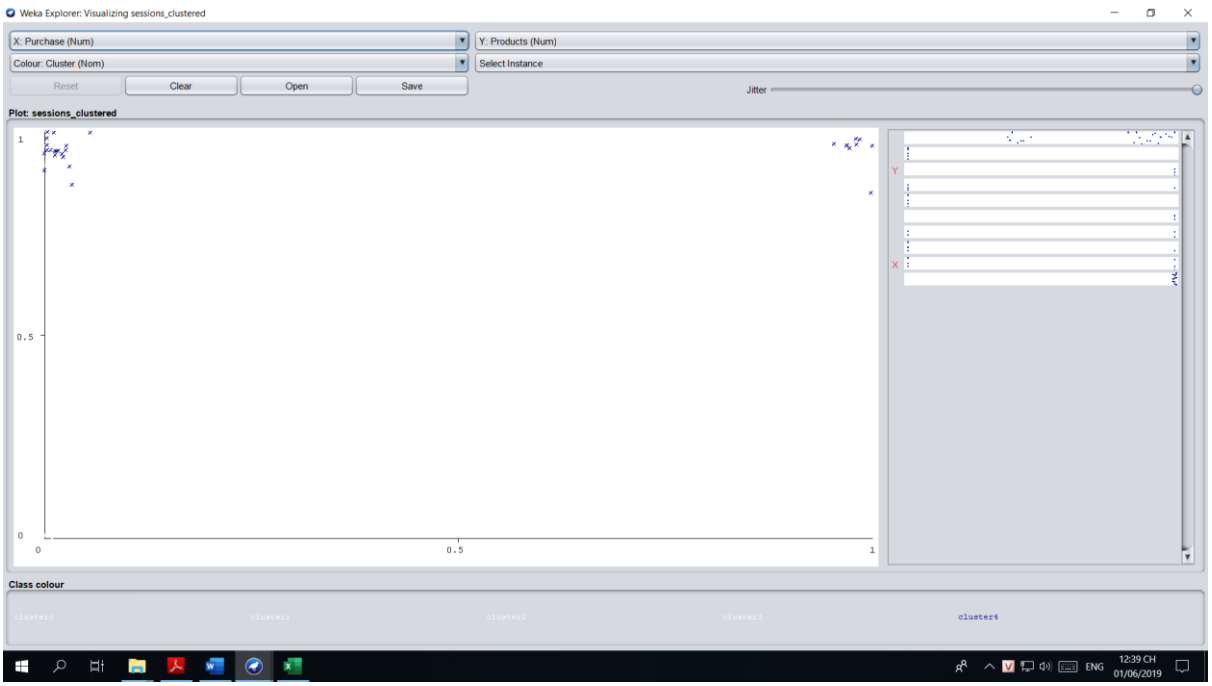


Chuyển trục **X thành Cluster**, **Y thành Products** (Người dùng thông thường xem nhiều sản phẩm)



Ta nhận thấy người dùng thuộc **cluster 4** là có xu hướng luôn vào trang **Product**

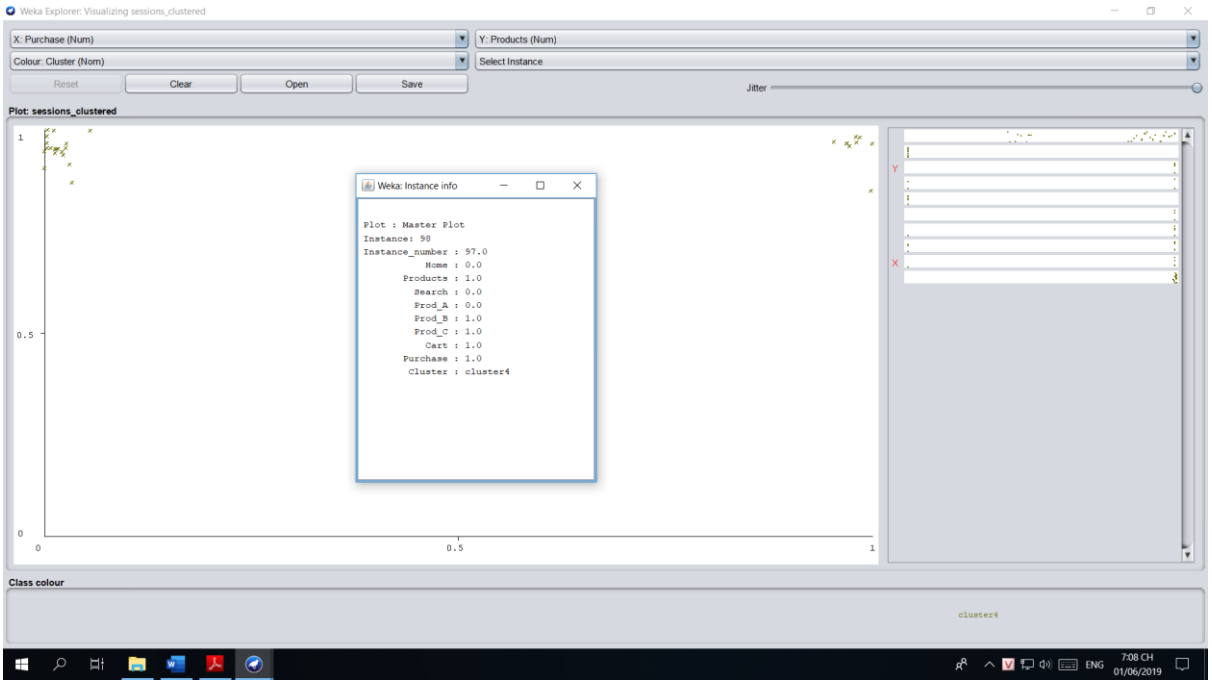
Ta chuyển trục **X thành Purchase**

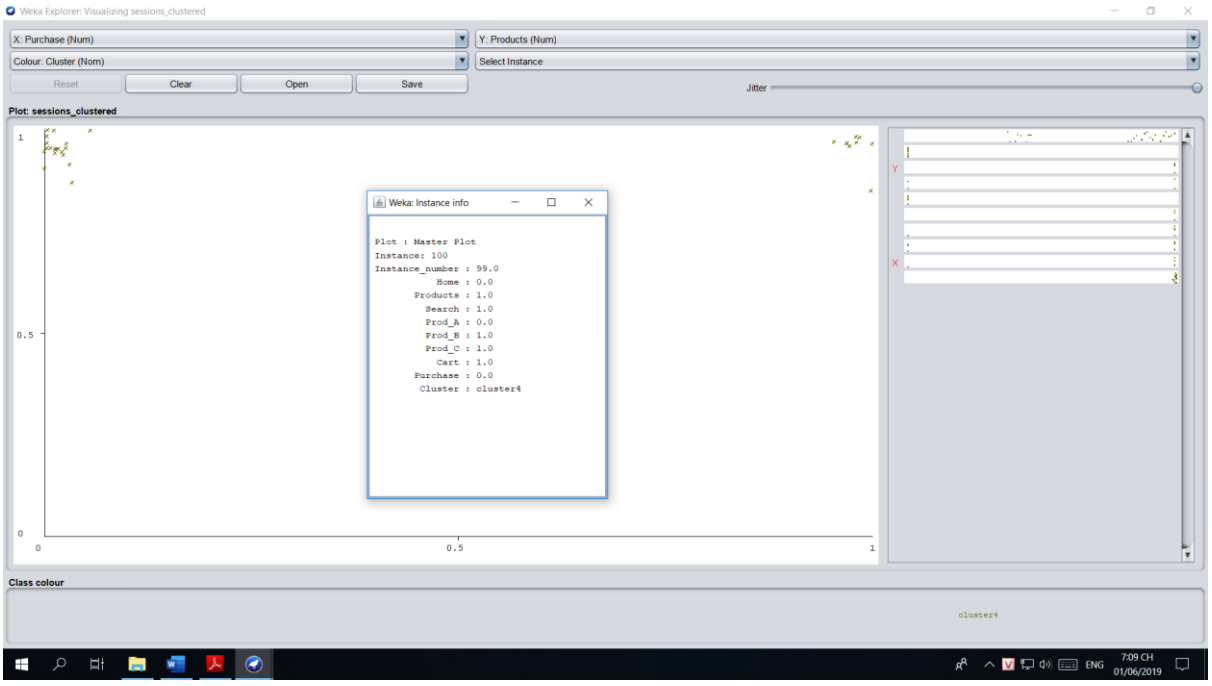


Đổi màu các cluster thành màu trắng chỉ để lại **cluster 4**

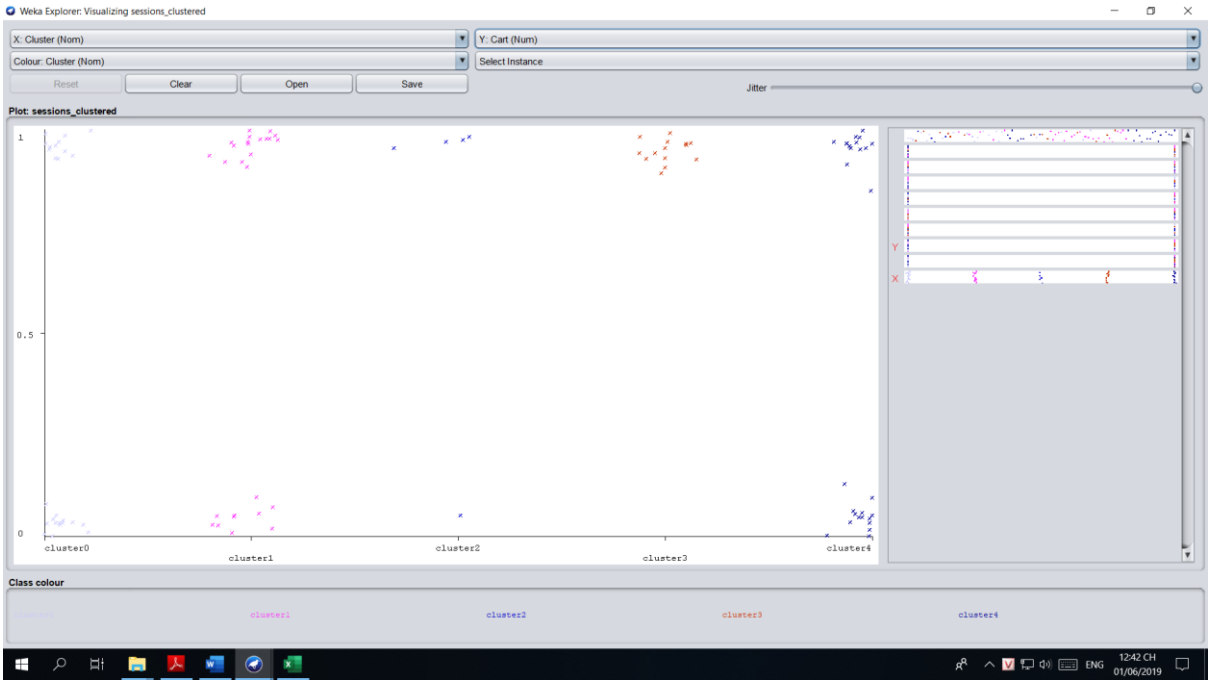
Ta nhận thấy số lượng người không mua hàng nhiều hơn số lượng người mua hàng. Vậy ta có thể tạm kết luận là nhóm khách hàng thông thường thường không mua sản phẩm

Một số mẫu cụ thể



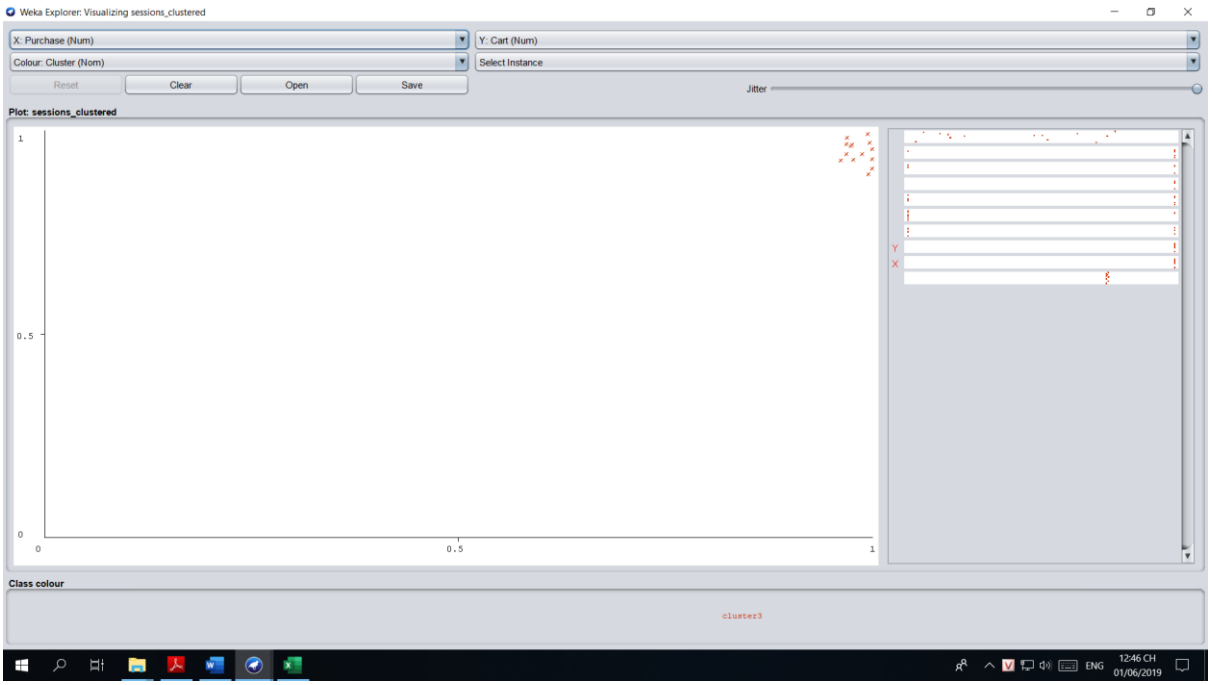


Chuyển trục **X thành Cluster**, **Y thành Cart** (**Người dùng tập trung**)



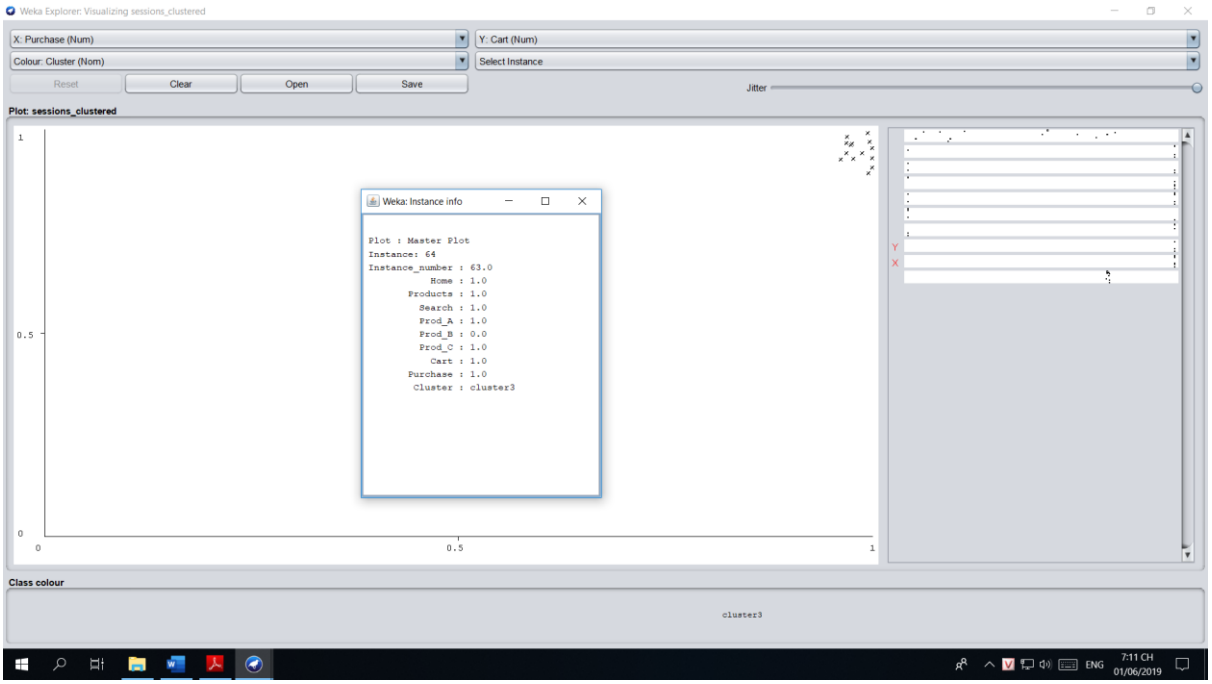
Ta nhận thấy nhóm người thuộc cluster 3 có xu hướng bỏ sản phẩm vào cart

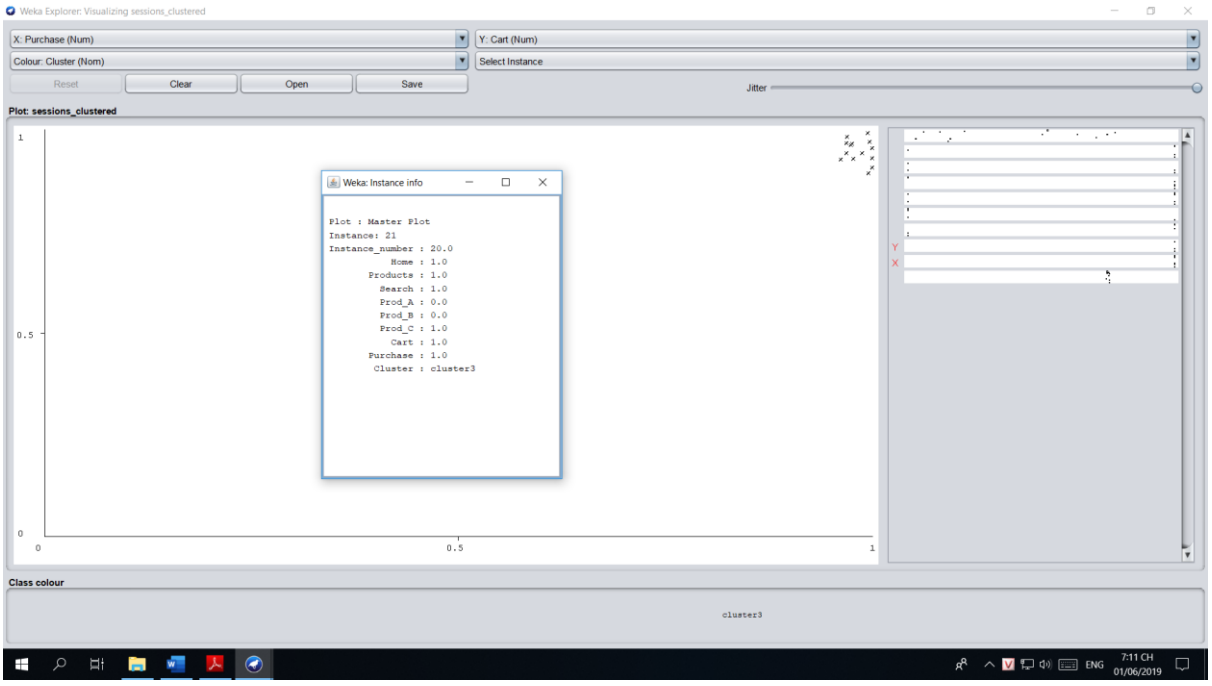
Ta sẽ chuyển trục **X thành Purchase**, **Y thành Cart** và màu của các cluster khác **cluster 3** thành trong suốt



Ta nhận thấy số lượng người thuộc **cluster 3** đều mua sản phẩm sau khi đã bỏ vào cart

Một số mẫu cụ thể





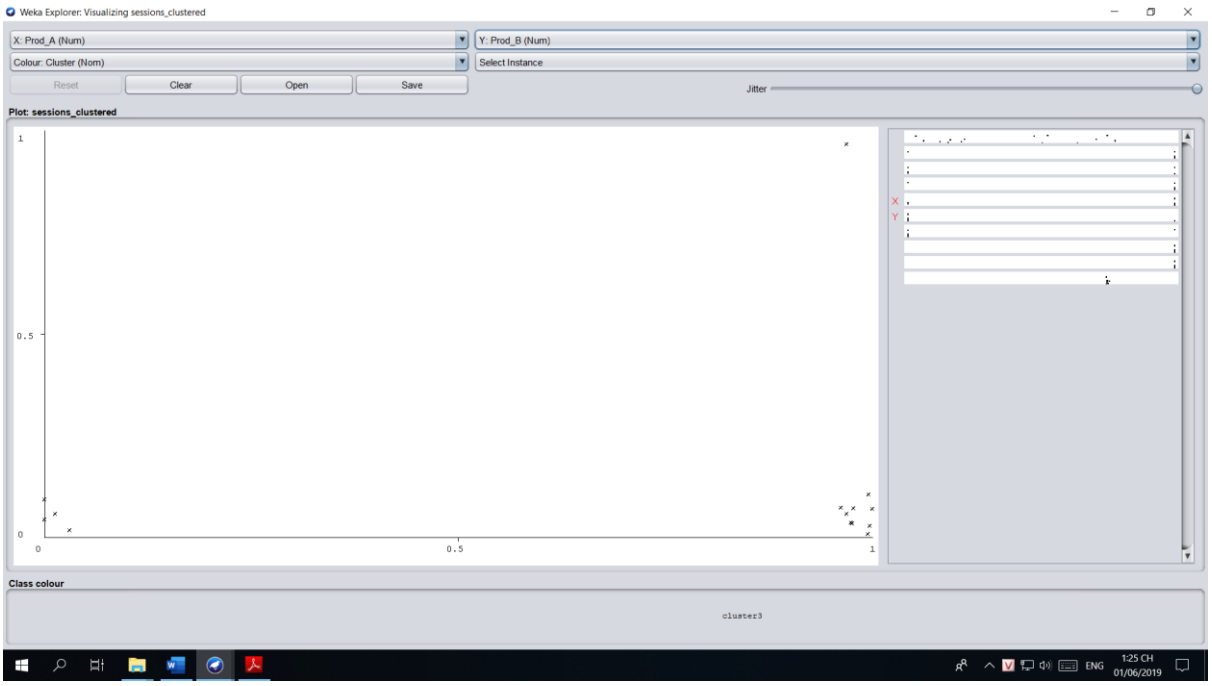
5. Cụm thể hiện sở thích mua hàng

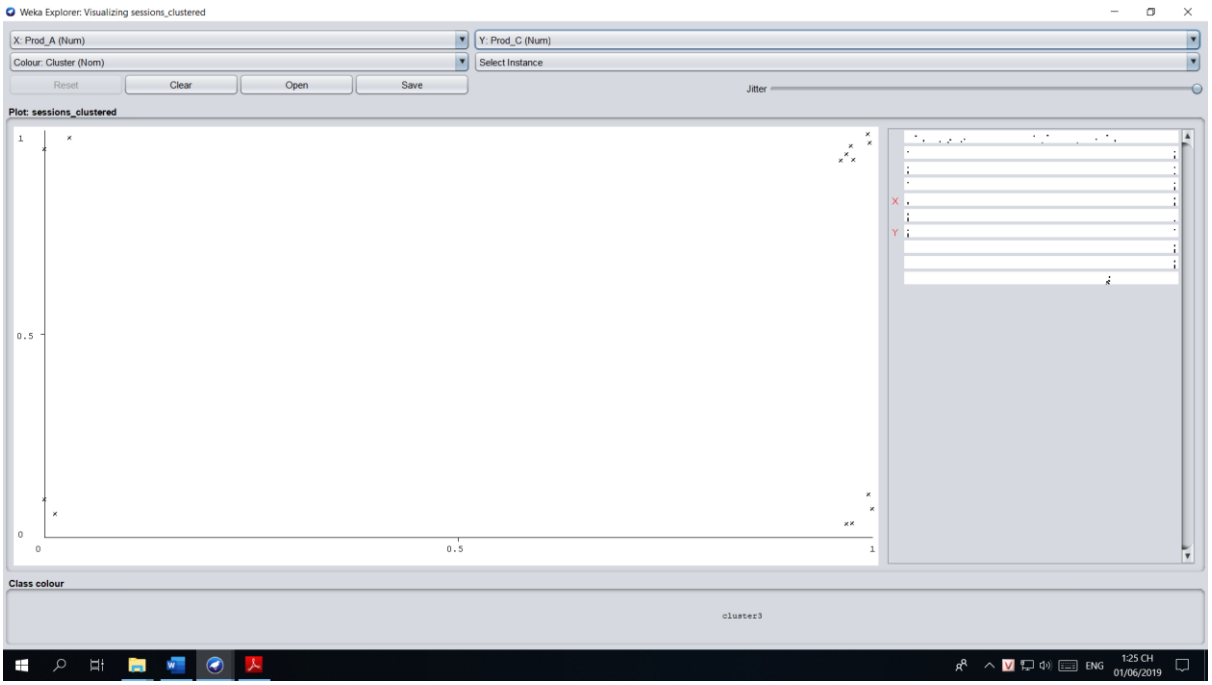
Ta đã biết **cluster 3** là nhóm người thường mua sản phẩm do đó chúng ta sẽ khảo sát trên cluster này để xem là người dùng thường mua theo nhóm sản phẩm hay là đơn lẻ

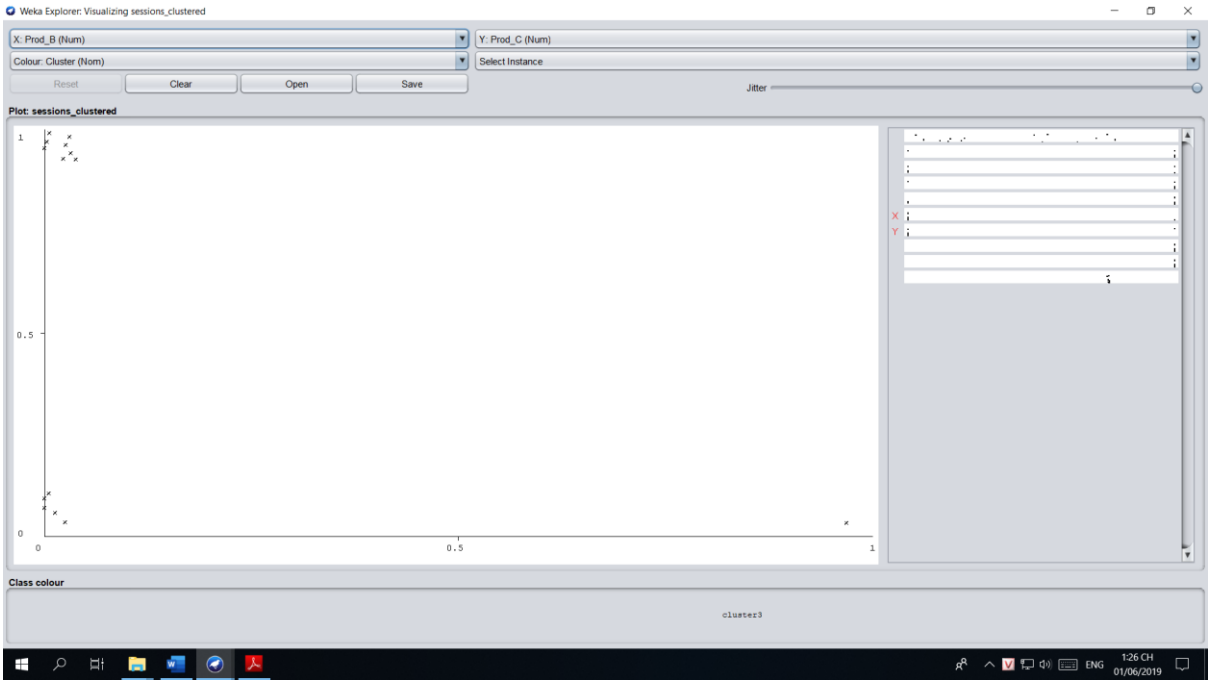
Ta chuyển các màu của các cluster khác **cluster 3** về màu trong suốt để dễ quan sát

Ta tiến hành thay đổi trục **X và Y thành Prod_A, Prod_B, Prod_C** trong **cluster 3**

Ta muốn là các điểm dữ liệu sẽ tập trung ở góc phải trên để kết luận rằng khách hàng thường mua sản phẩm theo nhóm

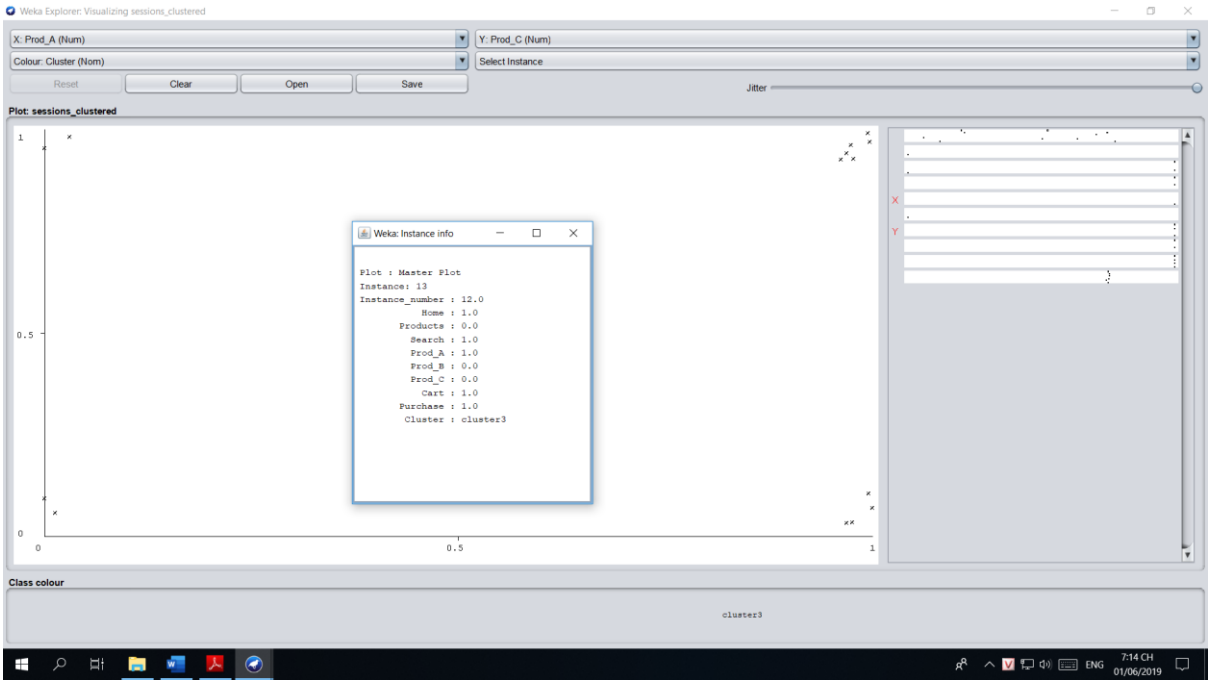


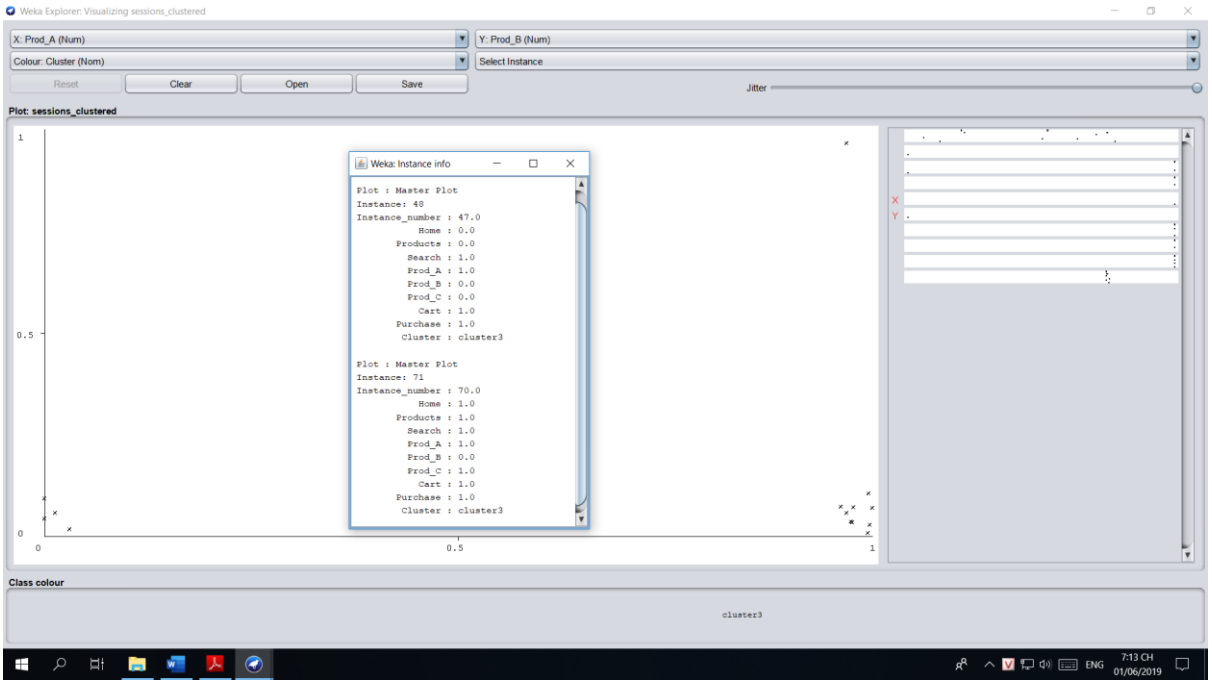




Từ 3 ảnh quan sát trên thì ta có nhận xét: Các điểm dữ liệu không tập trung vào góc phải trên (mua X và mua Y). Do đó, xu hướng mua hàng của khách là theo hướng mua riêng lẻ từng mặt hàng

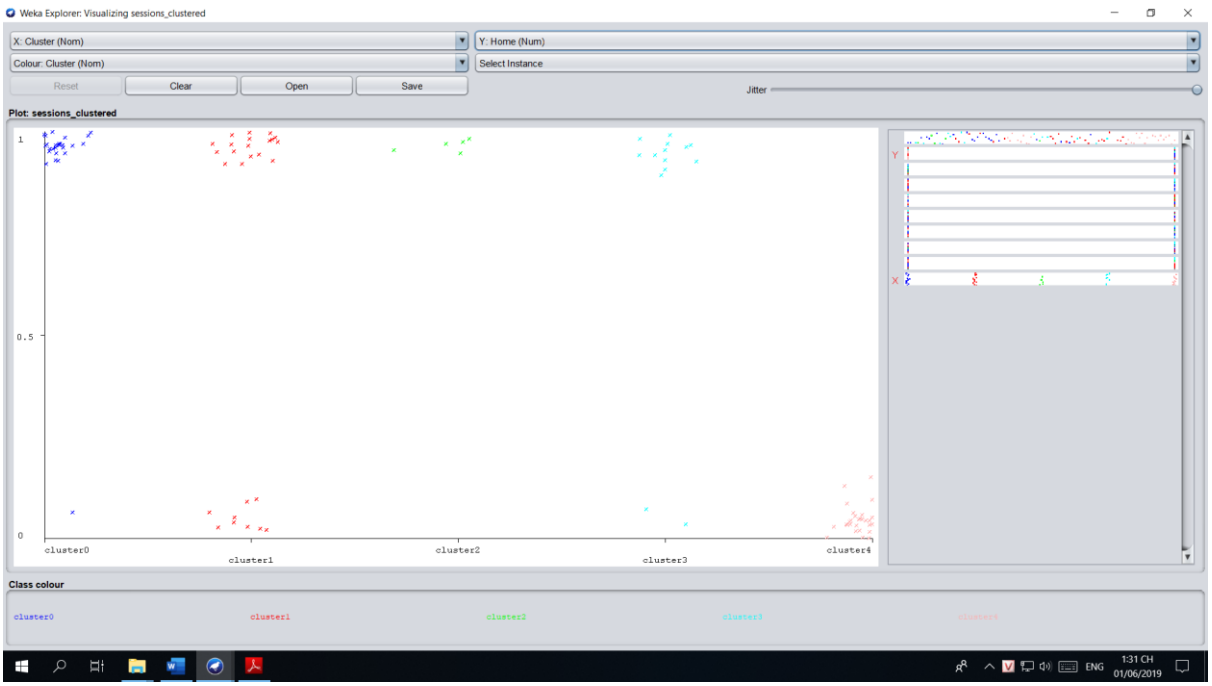
Một số mẫu cụ thể





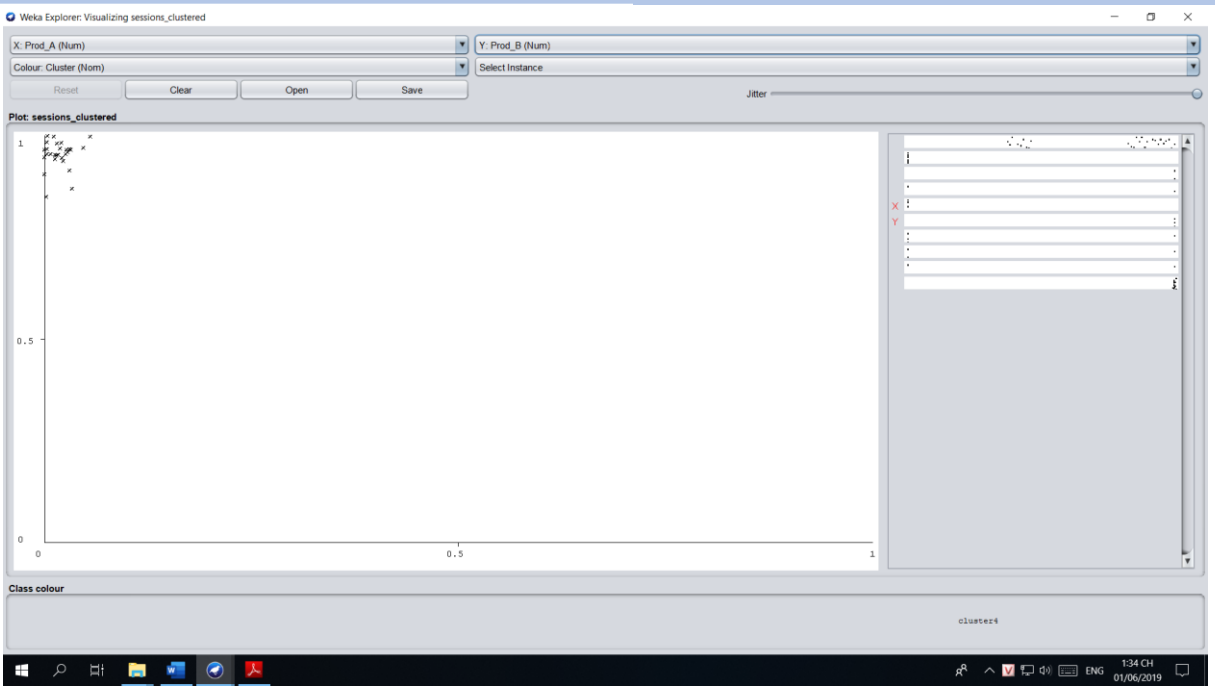
6. Nhận diện nhóm người dùng bị thu hút bởi quảng cáo

Ta đổi trục **X thành Cluster** và **Y thành Home**



Ta nhận thấy ở **cluster 4** là tập trung những khách hàng không ghé thăm trang Home. Do đó, đây là nhóm người bị thu hút bởi quảng cáo

Ta chuyển trục **X thành Prod_A**, **Y thành Prod_B** và màu của các cluster khác 4 thành trong suốt



Ta nhận thấy các điểm dữ liệu tập trung tất cả vào góc trái trên. Do đó, chiến dịch quảng cáo của sản phẩm B là thành công hơn

7. Cài đặt chương trình

Chương trình được cài đặt giống như mẫu trong đề và ta tiến hành chạy bằng command line

Ta tiến hành chạy chương trình cài đặt với k = 3 đến k = 8

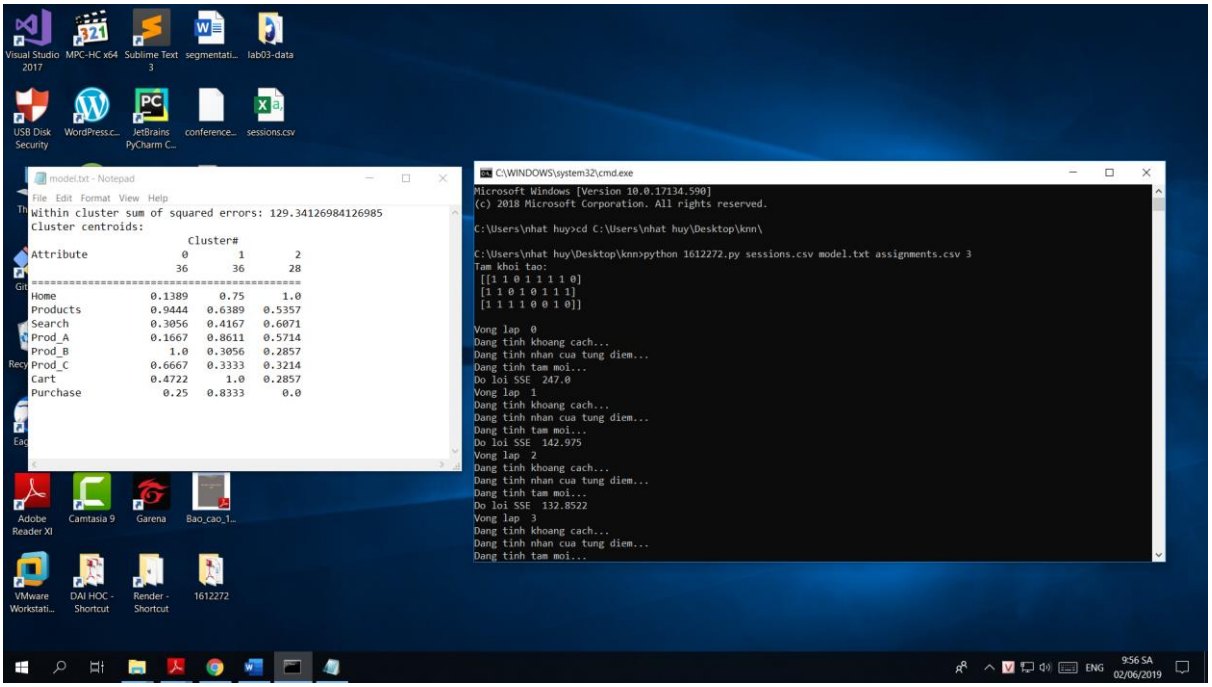


Figure 7: k = 3

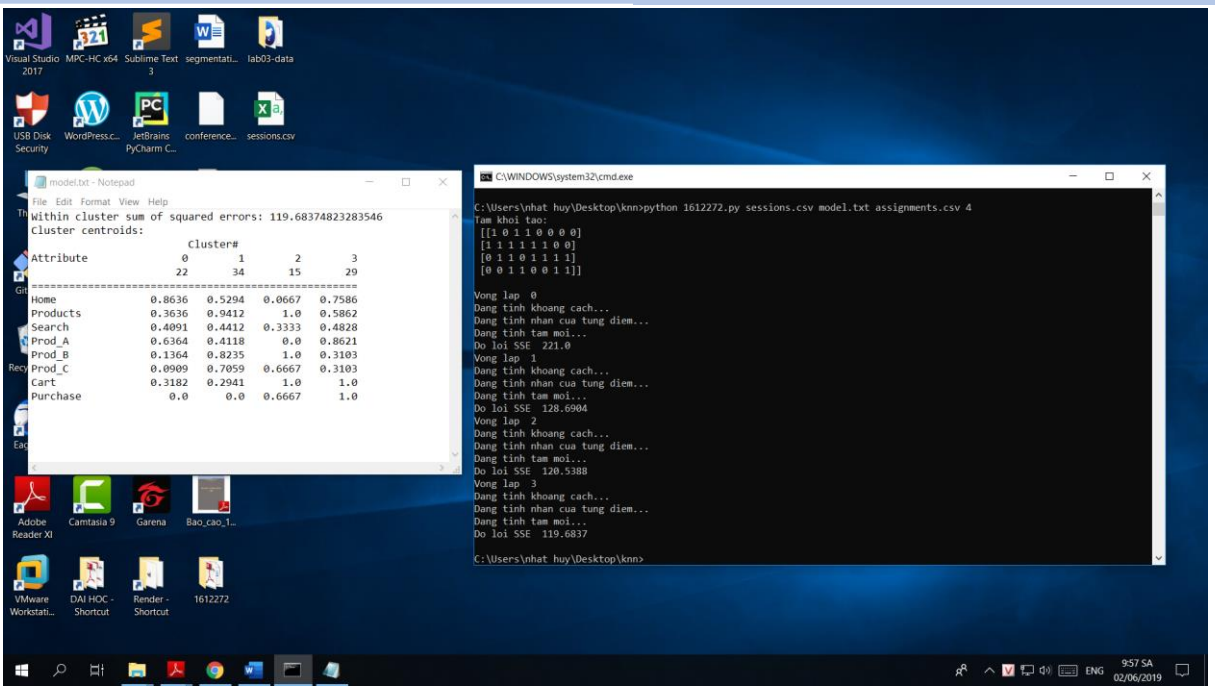


Figure 8: k = 4

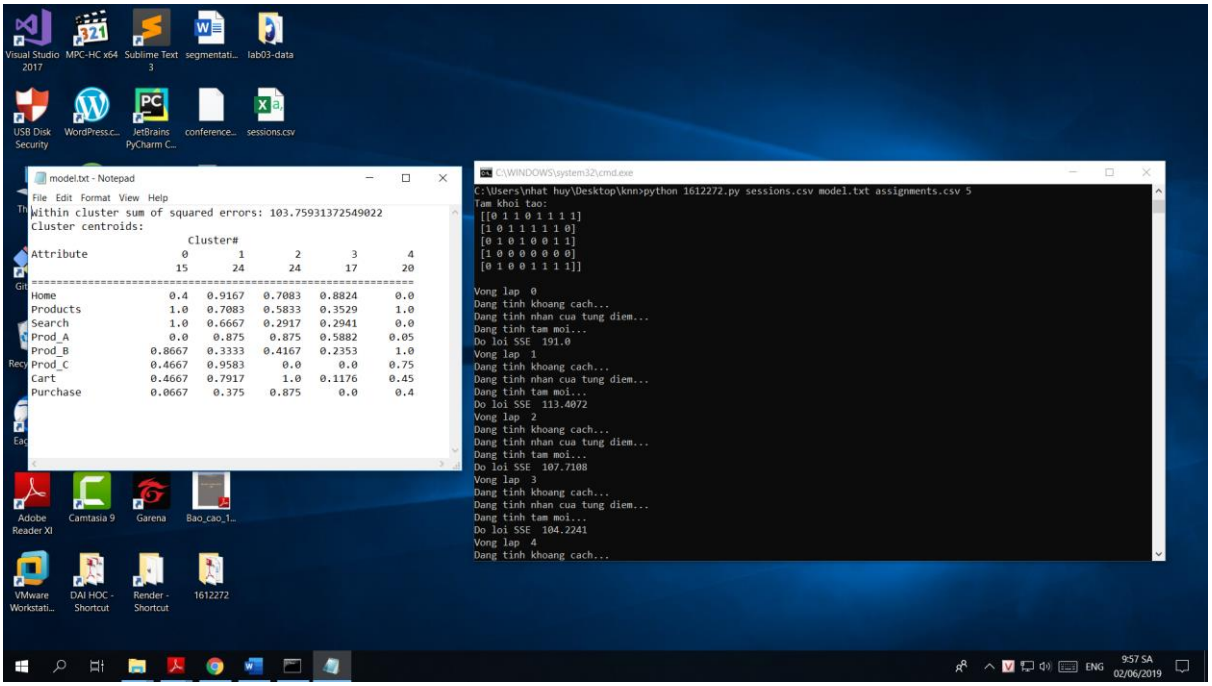


Figure 9: k = 5

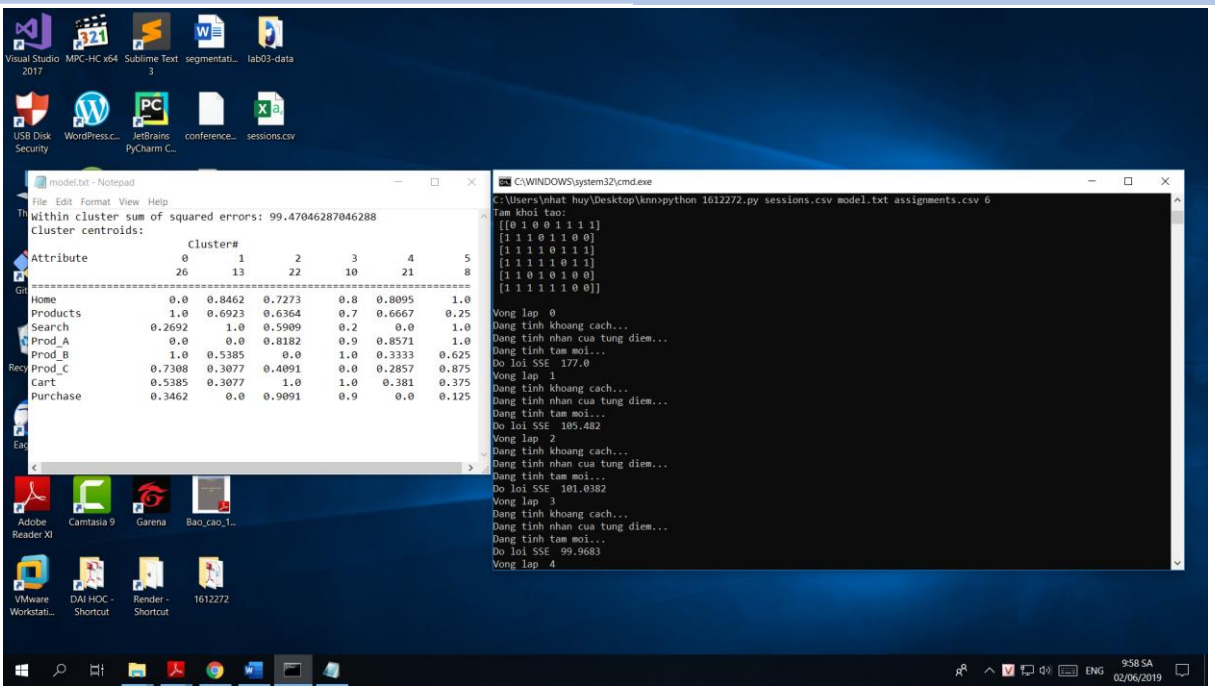


Figure 10: k = 6

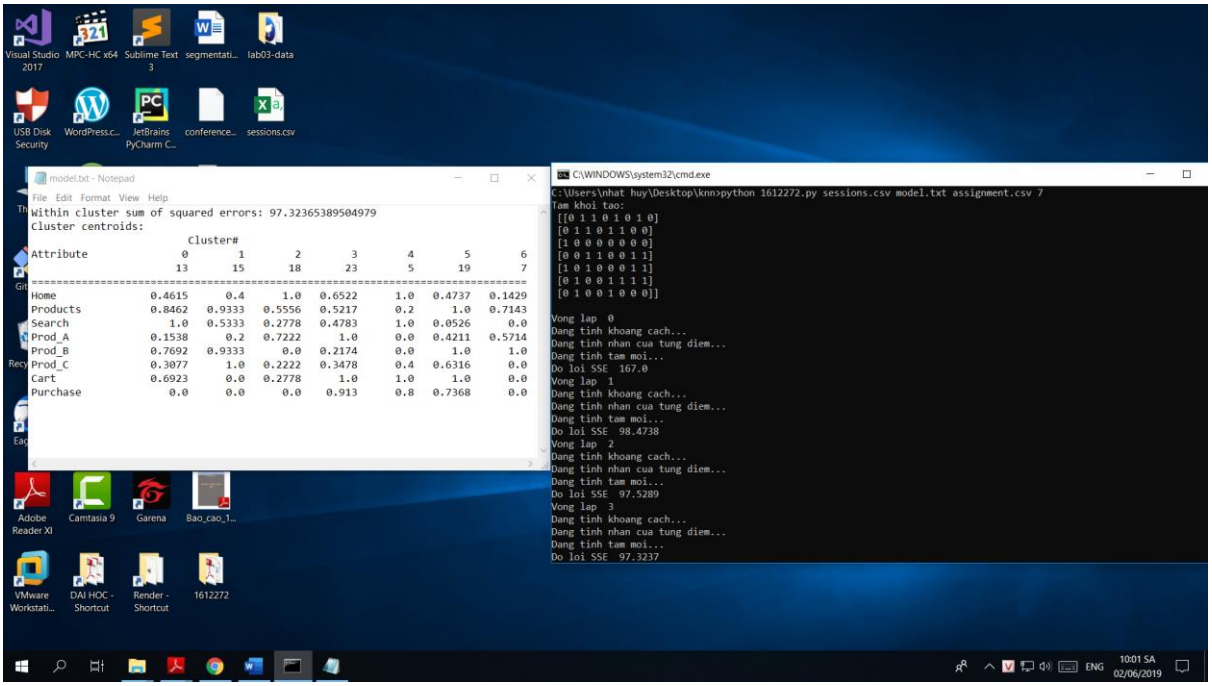


Figure 11: k = 7

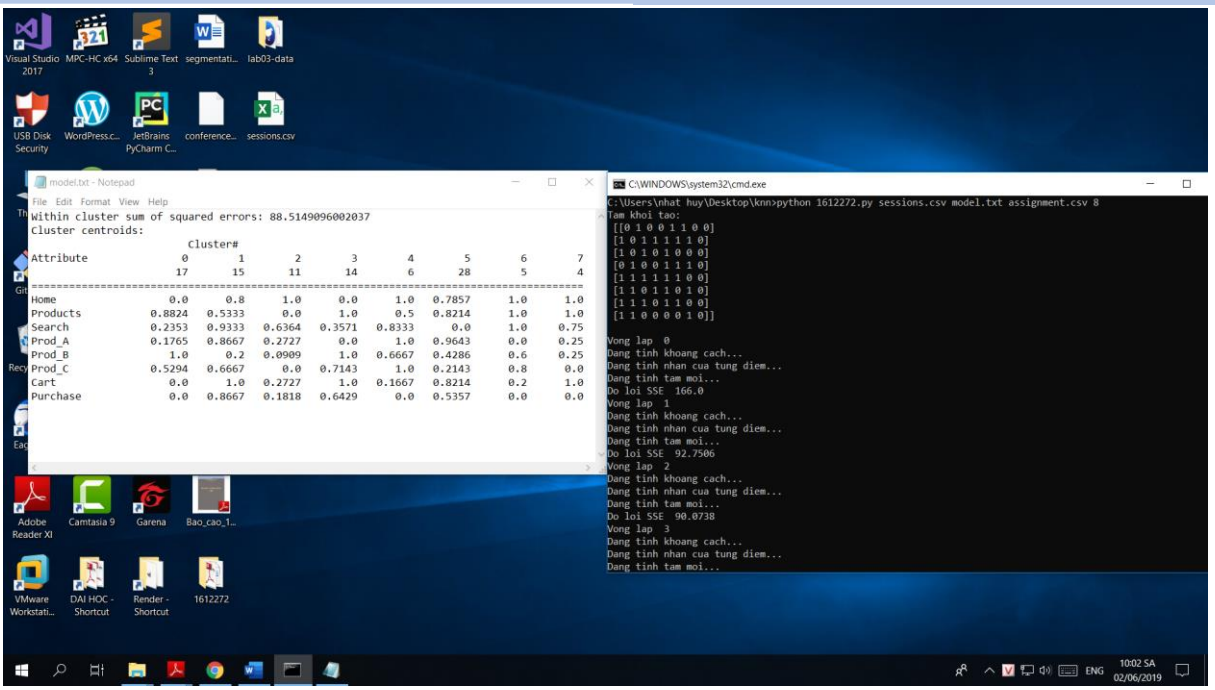


Figure 12: k = 8

Vì chúng ta khởi tạo các điểm là ngẫu nhiên nên việc chạy thuật toán sẽ khác nhau. Tuy nhiên, khi so sánh độ lỗi của từng trường hợp k thì độ lỗi khi chạy bằng code tự cài cũng gần giống với Weka