Tên: Nguyễn Mạnh Dũng

Lớp: EE003

**BTVN số 5: Phân cụm dữ liệu**

1.

B1: chọn File và tải file excel Supermarket\_Data lên

B2: Chọn cột Tagert là “Outlet\_Location\_Type”

A screenshot of a computer

Description automatically generated

2. Ta sử dụng file vừa được xử lý ở trên và tiền xử lý dữ liệu đó trước khi thực hiện các phương pháp

A picture containing circle, logo, screenshot, graphics

Description automatically generated

* **Phương pháp Hierarchical Clustering**:

B1: Mở Widget Distances và liên kết nó với file trên

B2: Đúp chuột 2 lần để mở Distances A picture containing symbol, logo, font, graphics

Description automatically generated chọn độ đo Euclidean

A screenshot of a computer

Description automatically generated

B3: Chọn Widget Hierarchical Clustering A picture containing logo, font, screenshot, symbol

Description automatically generated nối nó với Distances

B4: Mở Hierarchical Clustering

- Ta chọn Linkage là Average và chia thành 3 cụm

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

B5: Mở để xem chỉ số Sihouette Plot A picture containing logo, graphics, screenshot, design

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

B5: Nhấn giữ phím Ctrl và chọn giá trị cao nhất ở mỗi cụm để xem giá trị

B6: Chọn Data Table để xem giá trị Sihouette Plot

A screenshot of a computer

Description automatically generated

**Nhận xét:** Theo kết quả phương pháp Hierarchical Clustering, ta nên chia thành 3 cụm theo cách tính “ Average” vì giá trị Sihouette Plot cũng khá tốt( các cụm C1, C2, C3 có chỉ số Sihouette Plot lần lượt là 0.829322; 0.888596; 0.405574) vì chỉ số này càng tiến về 1 thì phương pháp này càng tốt.

* **Phương pháp k-Means**

- Vì dữ liệu mà ta cần phân cụm quá lớn nên ta cần phải tạo ra một mẫu dữ liệu nhỏ hơn để có thể áp dụng phương pháp k-Means

B1: chọn Data Sample và chỉnh Fixed Proportion of data xuống 55%

A screenshot of a computer

Description automatically generated with medium confidence

B2: Sau đó ta sử dụng k-Mean A picture containing circle, graphics, clipart, cartoon

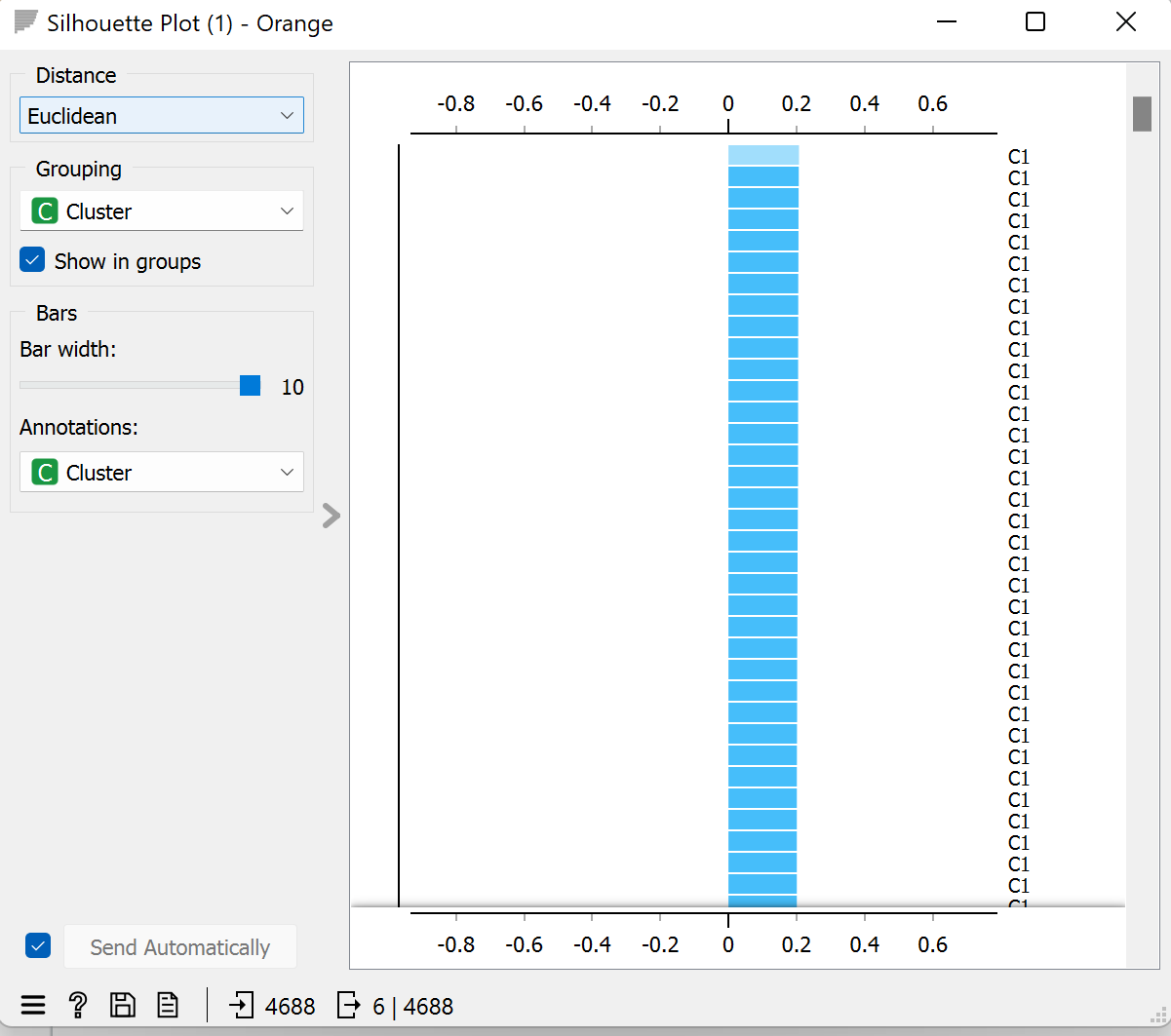
Description automatically generated để xử lý

A screenshot of a computer

Description automatically generated with medium confidence

-Ta thấy rằng chia thành 6 cụm có giá trị Sihouette Plot cao nhất (0.147) nên phân thành 6 cụm là tốt nhất

B3: chọn Sihouette Plot để xem giá trị



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

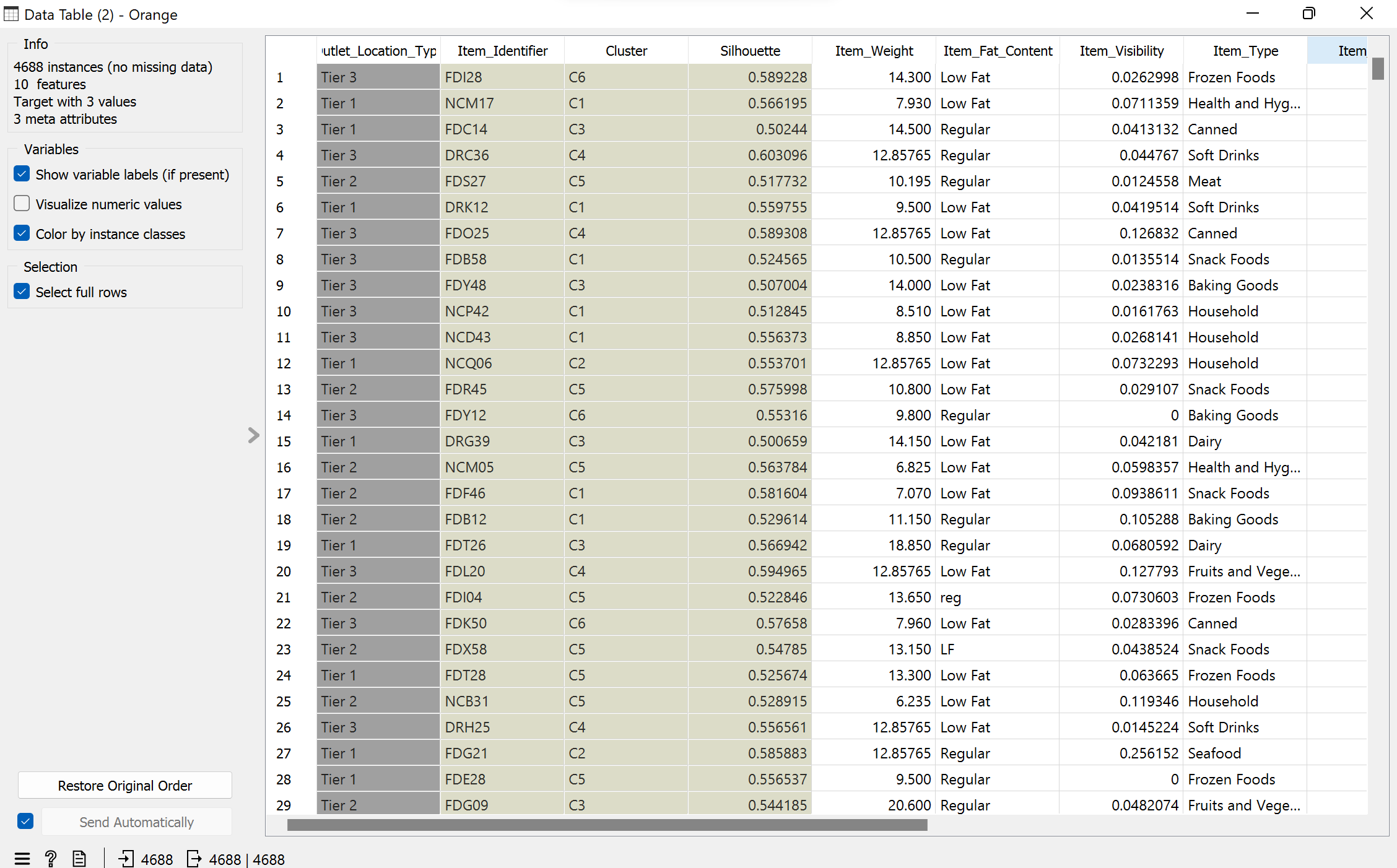
Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence



B4: nhấn giữ Ctrl và chọn giá trị Sihouette Plot cao nhất

B5: chọn Data Table để xem giá trị Sihouette Plot này

A screenshot of a computer

Description automatically generated

**Nhận xét:** Ta thấy rằng dữ liệu này không phù hợp với phương pháp K-means vì giá trị Sihouette Plot( Cluster) khá thấp ( chưa tiến tới quá gần 1) nên phương pháp này chưa phải phương pháp tối ưu nhất.

**Nhận xét chung toàn bài:** Ta thấy rằng trong 2 phương pháp là Hierarchical Clustering và K-means thì giá trị Sihouette Plot( Cluster) của phương pháp Hierarchical Clustering cao hơn nhiều so với phương pháp k-Mean nên ta nên sử dụng phương pháp Hierarchical Clustering để chia cụm dữ liệu cho bài này

**Ảnh bài làm Orange:**

**A computer screen shot of a diagram

Description automatically generated with low confidence**