I. Ý TƯỞNG PROJECT

RFM là một mô hình tương đối đơn giản nhưng vẫn rất thiết thực trong việc phân loại khách hàng.

Tuy nhiên, nếu chỉ dừng ở mức tính toán các chỉ số Recency, Frequency hay Monetary Value thì chưa đủ cơ sở để phân loại.

Chỉ số bằng bao nhiêu thì được coi là cao, bao nhiêu là thấp? Hơn nữa mỗi sản phẩm mỗi phân đoạn thị trường lại có đặc thù khác nhau, nên không thể sử dụng ngành này tham chiếu trong ngành kia. Do đó cần tự so sánh các giá trị của tập khách hàng hiện tại của công ty.

Tuy nhiều, điều khó khăn là không phải tập dữ liệu nào cũng phân bố đều, việc phân chia cũng cần phải linh hoạt và có thể update theo thời gian. Để khắc phục được vấn này, có thể sử dụng đến thuật toán Kmeans để phân loại RFM cho phù hợp với từng bối cảnh cụ thể.

Project này có thể áp dụng cho rất nhiều trường hợp nhỏ khác nhau, ví dụ như tính toán riêng cho từng nhóm sản phẩm (nếu công ty có nhiều sản phẩm), cho từng khu vực kinh doanh,... Vì thời lượng có hạn nên sẽ chỉ thể hiện một case điển hình.

II. CÁC BƯỚC THỰC HIỆN

1. Khám phá dữ liệu

Sử dụng Power Query với tính năng preview

2. Clean tập dữ liệu

- ➢ Đảm bảo các cột được đưa về đúng datatype
- ➢ Bỏ các giá trị Duplicate
- ➢ Xử lý dữ liệu dạng string: Trim & Clean để loại bỏ khoảng trắng thừa
- ➢ Xử lý các dữ liêu không đồng nhất: Lỗi nhập, lỗi đánh máy. United States => USA
- ➢ Xử lý các dữ liệu bị lỗi: Cột quantity và unit price có một số giá trị âm => Loại bỏ

➢ Check các giá trị null. Trong trường hợp này, chỉ có null nằm ở cột Customer_id, số dòng null cũng rất cao ~ 25% tổng số dòng

Vẫn phải loại bỏ vì mục đích là để phân loại khách hàng nhằm có hướng chăm sóc phù hợp. Null có khả năng là do khách hàng không đăng ký thành viên

=> *Cần check kỹ hơn với các bộ phận liên quan. Nếu cần thiết thì phải phân tích riêng với tập khách hàng null này và có hướng xử lý null cho phù hợp.*

3. Phân tách tập dữ liệu thành các trường hợp

Trong tập dữ liệu bao gồm Mỹ và Trung Quốc. Tách thành 2 tập dữ liệu con.

China chiếm phần lớn nên project này sẽ tập trung phân tích khách hàng Trung Quốc.

4. Tính toán RFM cho từng khách hàng

5. Vẽ biểu đồ phân phối

Phân phối của cả 3 giá trị đều lệch rất nhiều. Skewness cách xa 0

=> Cần đưa về dạng phân phối chuẩn

6. Transform dữ liệu về dạng phân phối chuẩn

Thử nghiệm với 3 phương pháp phổ biến:

- Box Cox Transformation
- Log Transformation
- Square Root Transformation

Lựa chọn phương pháp transform để có skewness gần 0 nhất: Box Cox Transformation

7. Lựa chọn số nhóm phân loại

Import ***KMeans*** từ thư viện ***sklearn.cluster***

Phương pháp Elbow => Lựa chọn k=3

8. Tiến hành phân loại

Cluster	Recency	Frequency	MonetaryValue
0	43.06	65.62	1264.88
1	182.54	18.78	449.45
2	19.75	313.42	8964.24

```
#CONCLUSION
# Group 0: Normal Customer
# Group 1: Likely-to-churn customers
# Group 2: Loyal Customer
```

9. Gán nhãn cho từng khách hàng

Dựa trên các group đã được phân loại ở trên để gán và đưa ra các chiến lược cụ thể cho từng nhóm.