**BÁO CÁO DỰ ÁN PHÂN TÍCH HÀNH VI KHÁCH HÀNG & RFM SEGMENTATION**

* *Người thực hiện : Phạm Minh Trung*
* *Số điện thoại : 0329129004  
  Email:* [*phamminhtrung1201@gmail.com*](mailto:phamminhtrung1201@gmail.com)
* *Linkedin :pm-trung-data/*
* *GITHUB :MTrung-bot*

**MỤC LỤC**

[**1. Giới thiệu Dự Án** 3](#_Toc201597401)

[**2. Mục Tiêu Dự Án** 3](#_Toc201597402)

[***3. Quy Trình Thực Hiện*** 3](#_Toc201597403)

[**4. Kết Quả** 9](#_Toc201597407)

[**5. Gợi Ý Chiến lược Marketing** 9](#_Toc201597408)

[**6. File Đính Kèm ( upload GitHub)** 10](#_Toc201597409)

[***7. Upload GitHub*** 10](#_Toc201597410)

# **1. Giới thiệu Dự Án**

* ***Tên dự án****: Phân tích hành vi khách hàng & phân khúc theo RFM*
* ***Ngành nghề****: Thương mại điện tử (E-commerce)*
* ***Nguồn dữ liệu****: Kaggle - E-commerce Behavior Data (~14 GB)*
* ***Công cụ dùng****: Python (Pandas), Power BI, GitHub*
* ***Thời gian thực hiện****: Khoảng 1.5 ngày*

# **2. Mục Tiêu Dự Án**

* *Hiểu và phân nhóm khách hàng dựa trên hành vi mua sắm*
* *Đo lường 3 yếu tố RFM:*
  1. ***Recency****: Giao dịch gần nhất*
  2. ***Frequency****: Tần suất giao dịch*
  3. ***Monetary****: Số tiền chi tiêu*
* *Hỗ trợ marketing nhắm đúng đối tượng*

# ***3. Quy Trình Thực Hiện***

***Bước 1: Tiền xử dữ liệu (Python)***

* *Gộp 2 file 2019-Oct.csv và 2019-Nov.csv*
* *Chuyển đổi event\_time, lọc event\_type = "purchase"*
* *Loại NA ở user\_id, product\_id, price*

**Bước 2: Tính chỉ số RFM**

* *Dùng groupby theo user\_id*
* *Recency: Tính theo sự chênh lệch tính từ max(event\_time)*
* *Frequency: Đếm số giao dịch mua*
* *Monetary: Cộng tổng chi tiêu (sum price)*
* *Tạo cột event\_date dạng Date*

**Bước 3: Gán điểm RFM**

* *Dùng qcut() chia 5 nhóm cho mỗi chỉ số*
* *Tạo cột RFM\_Score (VD: 555 = R5 F5 M5)*

**Bước 4: Phân loại nhóm khách hàng**

* *Gán nhãn theo logic:*
  + *VIP: R, F, M đều cao*
  + *Chi tiêu cao: M cao*
  + *Mới mua nhiều: R & F cao*
  + *Nguy cơ rời bỏ: R thấp*
  + *Khác: Trung tính*

import pandas as pd

# Chuẩn bị list để lưu dữ liệu đã lọc

chunks = []

chunksize = 1\_000\_000  # Đọc mỗi lần 1 triệu dòng

# Đọc file theo khối và lọc những dòng là purchase

for chunk in pd.read\_csv('D:/VINAMILK/2019-Nov.csv', chunksize=chunksize):

    purchase\_chunk = chunk[chunk['event\_type'] == 'purchase']

    chunks.append(purchase\_chunk)

# Gộp tất cả các purchase lại

df\_nov = pd.concat(chunks, ignore\_index=True)

# Lặp lại với file tháng 10

chunks\_oct = []

for chunk in pd.read\_csv('D:/VINAMILK/2019-Oct.csv', chunksize=chunksize):

    purchase\_chunk = chunk[chunk['event\_type'] == 'purchase']

    chunks\_oct.append(purchase\_chunk)

df\_oct = pd.concat(chunks\_oct, ignore\_index=True)

# Gộp 2 tháng lại

df\_purchase = pd.concat([df\_nov, df\_oct], ignore\_index=True)

# Xem thử

print("Số dòng purchase:", df\_purchase.shape)

df\_purchase.head()

import pandas as pd

# Làm lại: Đảm bảo 'event\_time' là datetime

df\_purchase['event\_time'] = pd.to\_datetime(df\_purchase['event\_time'])

# Bỏ dòng không có giá

df\_purchase = df\_purchase.dropna(subset=['price'])

# Kiểm tra lần nữa

print("Số dòng còn lại sau làm sạch:", df\_purchase.shape)

# Ngày phân tích = ngày cuối cùng trong dữ liệu

NOW = df\_purchase['event\_time'].max()

# Nhóm theo user\_id

rfm = df\_purchase.groupby('user\_id').agg({

    'event\_time': lambda x: (NOW - x.max()).days,       # Recency

    'user\_session': 'nunique',                          # Frequency

    'price': 'sum'                                      # Monetary

}).reset\_index()

# Đổi tên cột

rfm.columns = ['user\_id', 'Recency', 'Frequency', 'Monetary']

# Kiểm tra kết quả

rfm.head()

# Chấm điểm theo phân vị (quantile)

rfm['R\_score'] = pd.qcut(rfm['Recency'], 5, labels=[5,4,3,2,1])

rfm['F\_score'] = pd.qcut(rfm['Frequency'].rank(method='first'), 5, labels=[1,2,3,4,5])

rfm['M\_score'] = pd.qcut(rfm['Monetary'], 5, labels=[1,2,3,4,5])

# Gộp lại thành segment

rfm['RFM\_Segment'] = rfm['R\_score'].astype(str) + rfm['F\_score'].astype(str) + rfm['M\_score'].astype(str)

# Phân loại nhóm khách hàng

def segment\_customer(segment):

    if segment == '555':

        return 'VIP'

    elif segment.startswith('5'):

        return 'Mới mua nhiều'

    elif segment.endswith('5'):

        return 'Chi tiêu cao'

    elif segment.startswith('1'):

        return 'Nguy cơ rời bỏ'

    else:

        return 'Khác'

rfm['Customer\_Type'] = rfm['RFM\_Segment'].apply(segment\_customer)

segment\_summary = rfm['Customer\_Type'].value\_counts(normalize=True) \* 100

print(segment\_summary.round(2))

# Gộp thông tin RFM vào df\_purchase

df\_merged = df\_purchase.merge(rfm[['user\_id', 'Customer\_Type']], on='user\_id', how='left')

# Đếm số lần mua mỗi loại sản phẩm theo nhóm khách

category\_by\_segment = df\_merged.groupby(['Customer\_Type', 'category\_code']).size().reset\_index(name='count')

# Top 10 sản phẩm mỗi nhóm

top\_categories = category\_by\_segment.sort\_values(['Customer\_Type', 'count'], ascending=[True, False]).groupby('Customer\_Type').head(5)

print(top\_categories)

# Nhóm theo loại khách và thương hiệu

brand\_by\_segment = df\_merged.groupby(['Customer\_Type', 'brand']).agg({

    'price': 'sum',

    'user\_id': 'nunique'

}).reset\_index().rename(columns={'price': 'total\_spent', 'user\_id': 'unique\_customers'})

# Top 5 thương hiệu mỗi nhóm

top\_brands = brand\_by\_segment.sort\_values(['Customer\_Type', 'total\_spent'], ascending=[True, False]).groupby('Customer\_Type').head(5)

print(top\_brands)

import seaborn as sns

import matplotlib.pyplot as plt

# Gộp RFM vào df\_purchase

df\_viz = df\_purchase.merge(rfm[['user\_id', 'Customer\_Type']], on='user\_id', how='left')

# Bỏ NaN

df\_viz = df\_viz.dropna(subset=['Customer\_Type', 'category\_code', 'brand'])

plt.figure(figsize=(14, 6))

top\_cat = (

    df\_viz.groupby(['Customer\_Type', 'category\_code'])

    .size()

    .reset\_index(name='count')

    .sort\_values(['Customer\_Type', 'count'], ascending=[True, False])

    .groupby('Customer\_Type')

    .head(5)

)

sns.barplot(

    data=top\_cat,

    x='category\_code', y='count',

    hue='Customer\_Type'

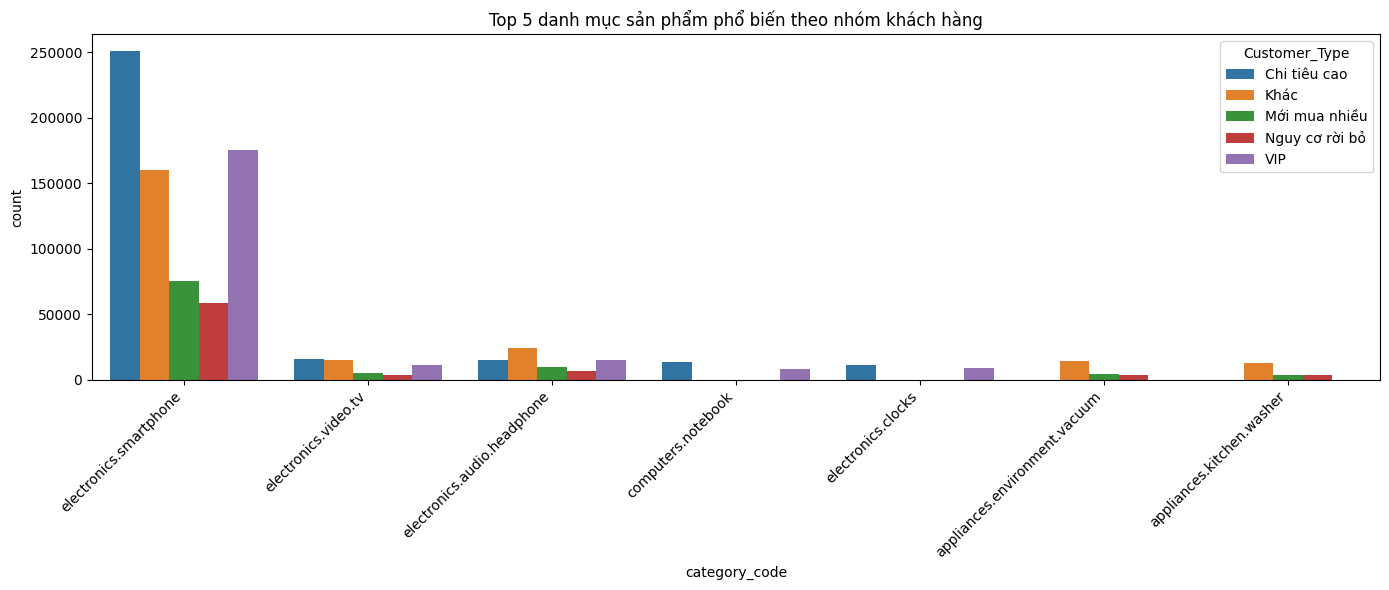
)

plt.title("Top 5 danh mục sản phẩm phổ biến theo nhóm khách hàng")

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()

**

plt.figure(figsize=(14, 6))

top\_brand = (

    df\_viz.groupby(['Customer\_Type', 'brand'])

    .agg({'price': 'sum'})

    .reset\_index()

    .sort\_values(['Customer\_Type', 'price'], ascending=[True, False])

    .groupby('Customer\_Type')

    .head(5)

)

sns.barplot(

    data=top\_brand,

    x='brand', y='price',

    hue='Customer\_Type'

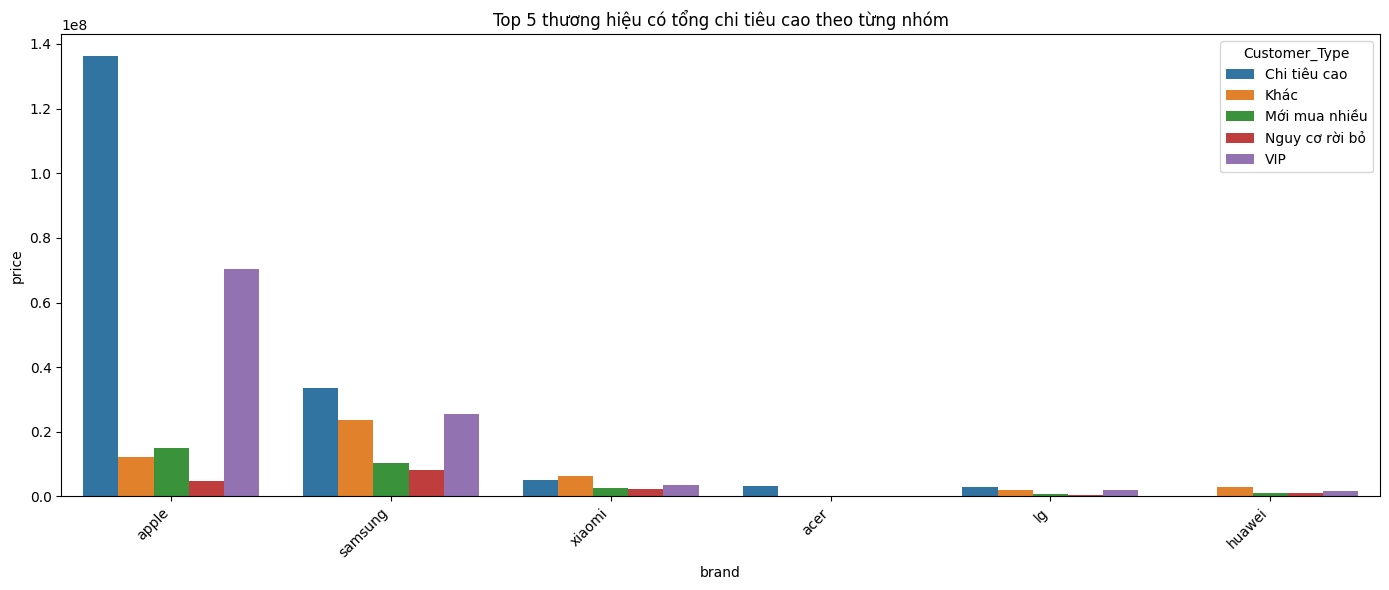
)

plt.title("Top 5 thương hiệu có tổng chi tiêu cao theo từng nhóm")

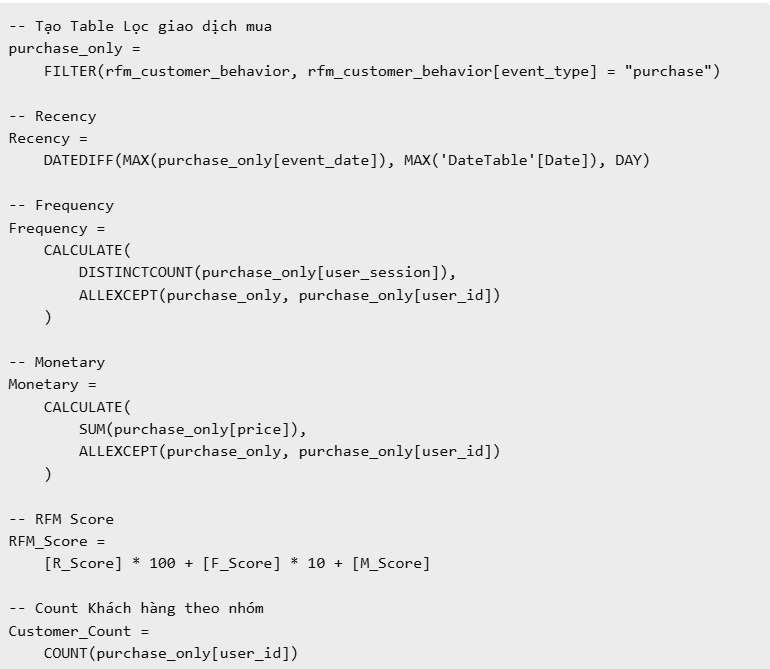
plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()

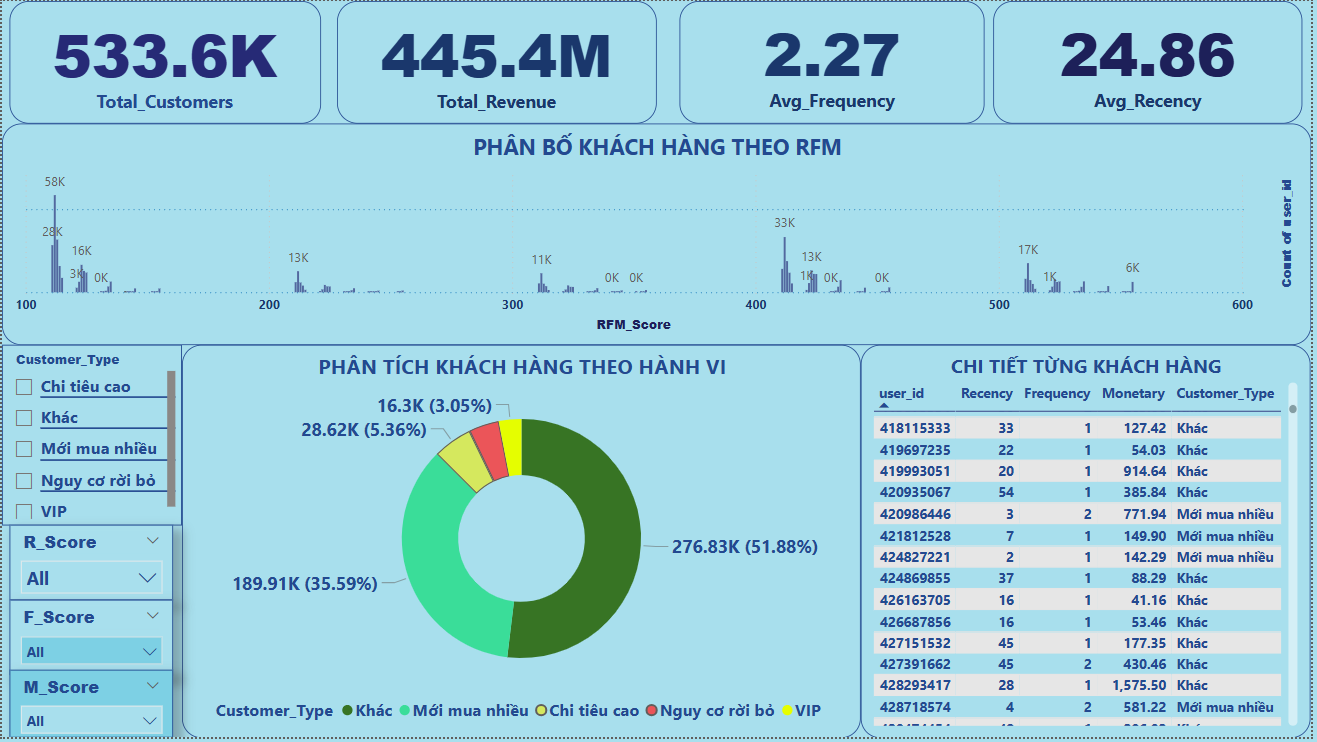
**

***Bước 5: DAX Measures chính***

**

***Bước 6: Xây dựng Power BI Dashboard***

* *Tạo Date Table, liên kết với event\_date*
* *Visuals:*
  1. *Bar chart: Số KH theo RFM\_Score, Customer\_Type*
  2. *Matrix: user\_id, R, F, M, Customer\_Type*
  3. *KPI Card: Đếm VIP, doanh thu trung bình*
  4. *Slicer: Customer\_Type, R\_Score, F\_Score, M\_Score*

******

# **4. Kết Quả**

* *VIP chiếm ~3%*
* *Nguy cơ rời bỏ ~4.11%*
* *Chi tiêu cao ~5.36%*
* *Mới mua nhiều ~35.59%*
* *Khác ~51.88% cần được nuôi dưỡng*

# **5. Gợi Ý Chiến lược Marketing**

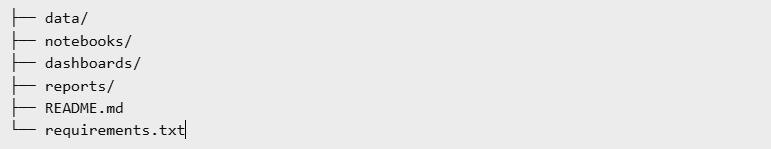
|  |  |
| --- | --- |
| ***Nhóm KH*** | ***Chiến lược gợi ý*** |
| *VIP* | *Loyalty, ưu đãi cao cấp* |
| *Rời bỏ* | *Email winback, re-targeting* |
| *Mới mua* | *Upsell + Cross-sell* |
| *Chi tiêu cao* | *Combo cao cấp, bundling* |
| *Khác* | *Chăm sóc, remarketing* |

# **6. File Đính Kèm ( upload GitHub)**

* *rfm\_analysis\_notebook.ipynb: Xử lý Python*
* *rfm\_model.pbix: Power BI Dashboard*
* *rfm\_report.pdf: Báo cáo nhanh*
* *rfm\_presentation.pptx: Slide Portfolio*
* *README.md: Mô tả GitHub*

# ***7. Upload GitHub***

1. *Tạo repo mới: rfm-customer-analysis*
2. *Cố định các file trong thư mục:*

**

*3. Dùng Git push hoặc upload trực tiếp*