

## Lab: Data Wrangling (The process of converting raw data into a usable form)

**Nội dung:** Tích hợp dữ liệu từ nhiều nguồn bằng Pandas

**Mục tiêu:** Xây dựng bảng dữ liệu (data frame) dựa trên quá trình trộn dữ liệu từ nhiều tập tin.

**Bài toán:** Hãy phân tích thói quen sử dụng điện thoại di động dựa trên các nhãn hàng

**Dữ liệu đầu vào:**

Bảng thietbi: cho biết danh sách các thiết bị

Bảng nguoidung\_thietbi: cho biết danh sách người dùng đang sở hữu thiết bị

Bảng nguoidung\_sudung: cho biết danh sách quá trình người dùng dùng thiết bị

**Yêu cầu dữ liệu đầu ra:**

Hãy tổng hợp dữ liệu thành bảng dữ liệu mới bao gồm các cột sau để phục vụ phân tích dữ liệu

outgoing\_mins\_per\_month, outgoing\_sms\_per\_month, monthly\_mb, use\_id, platform, device, Branding, Model

**\*Gợi ý:** Sinh viên suy nghĩ về sơ đồ quan hệ của các tập tin dữ liệu (xem mỗi tập tin là một bảng dữ liệu). Sau đó, sinh viên bắt đầu xử lý dữ liệu như hướng dẫn bên dưới.

1. Đọc dữ liệu lên dataframe

```
thietbi = pd.read_csv('/content/thietbi.csv')
nguoidung_thietbi = pd.read_csv('/content/nguoidung_thietbi.csv')
nguoidung_sudung = pd.read_csv('/content/nguoidung_sudung.csv')
```

2. Vẽ mô hình quan hệ (ERD) từ bảng dữ liệu trên

3. Đổi tên các cột dữ liệu cần thiết

```
thietbi.rename(columns={'Retail Branding': 'Branding',
                        'Marketing Name': 'MarketingName'}, inplace=True)
```

4. Trộn bảng nguoidung\_sudung và nguoidung\_thietbi thông qua inner join để được bảng kết quả có các cột sau: [outgoing\_mins\_per\_month, outgoing\_sms\_per\_month, monthly\_mb, use\_id, platform, device]

```
dfKetQua = pd.merge(nguoidung_sudung,
                    nguoidung_thietbi[['use_id', 'platform', 'device']],
                    on='use_id')
```

[16] dfKetQua.head(5)

	outgoing_mins_per_month	outgoing_sms_per_month	monthly_mb	use_id	platform	device
0	21.97	4.82	1557.33	22787	android	GT-I9505
1	1710.08	136.88	7267.55	22788	android	SM-G930F
2	1710.08	136.88	7267.55	22789	android	SM-G930F
3	94.46	35.17	519.12	22790	android	D2303
4	71.59	79.26	1557.33	22792	android	SM-G361F

Hoặc dùng left-join tùy theo nhu cầu

```
# hoặc left-join
dfKetQua = pd.merge(nguoidung_sudung,
                    nguoidung_thietbi[['use_id', 'platform', 'device']],
                    on=[REDACTED], how='left')
```

5. Tiếp tục inner-join bảng thietbi và dfKetQua để có thêm 2 cột Model và Branding trong bảng kết quả cuối cùng. Lưu ý ở đây khóa liên kết sẽ khác tên

```
# inner-join
dfKetQua = pd.merge(dfKetQua,
                    thietbi[['Branding', 'Model']],
                    left_on=[REDACTED],
                    right_on=[REDACTED])
```

Hoặc left-join

```
# left-join
dfKetQua = pd.merge(dfKetQua,
                    thietbi[['Branding', 'Model']],
                    left_on=[REDACTED],
                    right_on=[REDACTED], how='left')
```

6. Hãy liệt kê 5 dòng đầu tiên các mẫu thiết bị (cột device) bắt đầu bằng GT

```
dfKetQua[dfKetQua.device.str.startswith('GT')].head(5)
```

	outgoing_mins_per_month	outgoing_sms_per_month	monthly_mb	use_id	platform	device	Branding	Model
0	21.97	4.82	1557.33	22787	android	GT-I9505	Samsung	GT-I9505
1	69.80	14.70	25955.55	22801	android	GT-I9505	Samsung	GT-I9505
2	249.26	253.22	1557.33	22875	android	GT-I9505	Samsung	GT-I9505
3	249.26	253.22	1557.33	22876	android	GT-I9505	Samsung	GT-I9505
4	83.46	114.06	3114.67	22880	android	GT-I9505	Samsung	GT-I9505

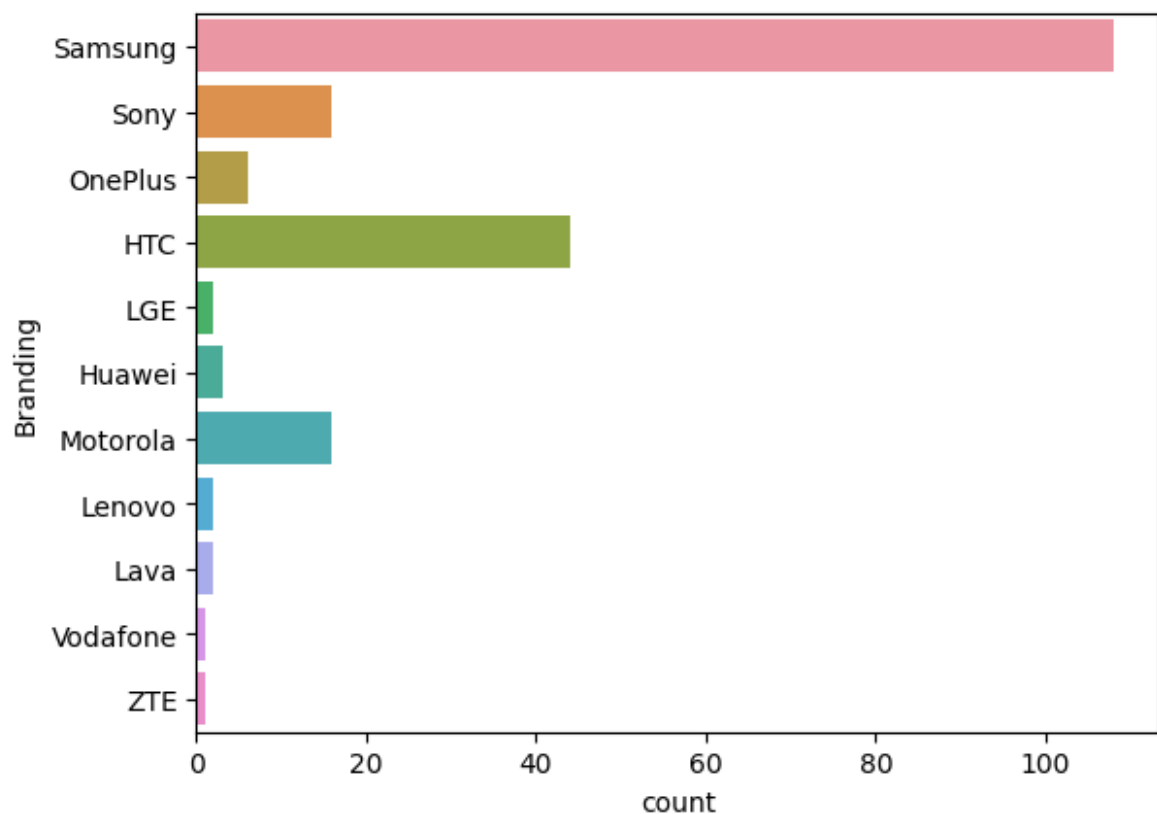
7. Hãy thống kê trung bình số phút hàng tháng, trung bình sms và trung bình data, số lượng sử dụng của các nhãn hàng

```
dfKetQua.groupby('Branding').agg({
    'outgoing_mins_per_month': 'mean',
    'outgoing_sms_per_month': 'mean',
    'monthly_mb': 'mean',
    'use_id': 'count'
})
```

8. Trực quan số lượng người dùng các nhãn hàng

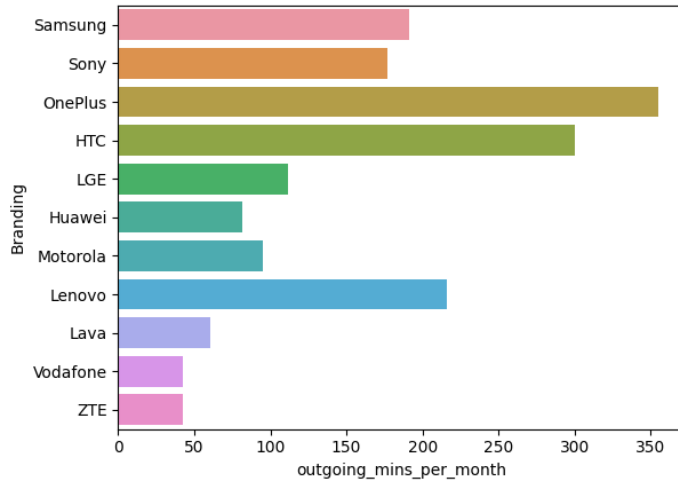
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(y= , data= )
plt.show()
```



9. Trực quan hóa dữ liệu trung bình phút gọi, trung bình sms và dung lượng sử dụng trên từng nhóm nhãn hàng bán lẻ

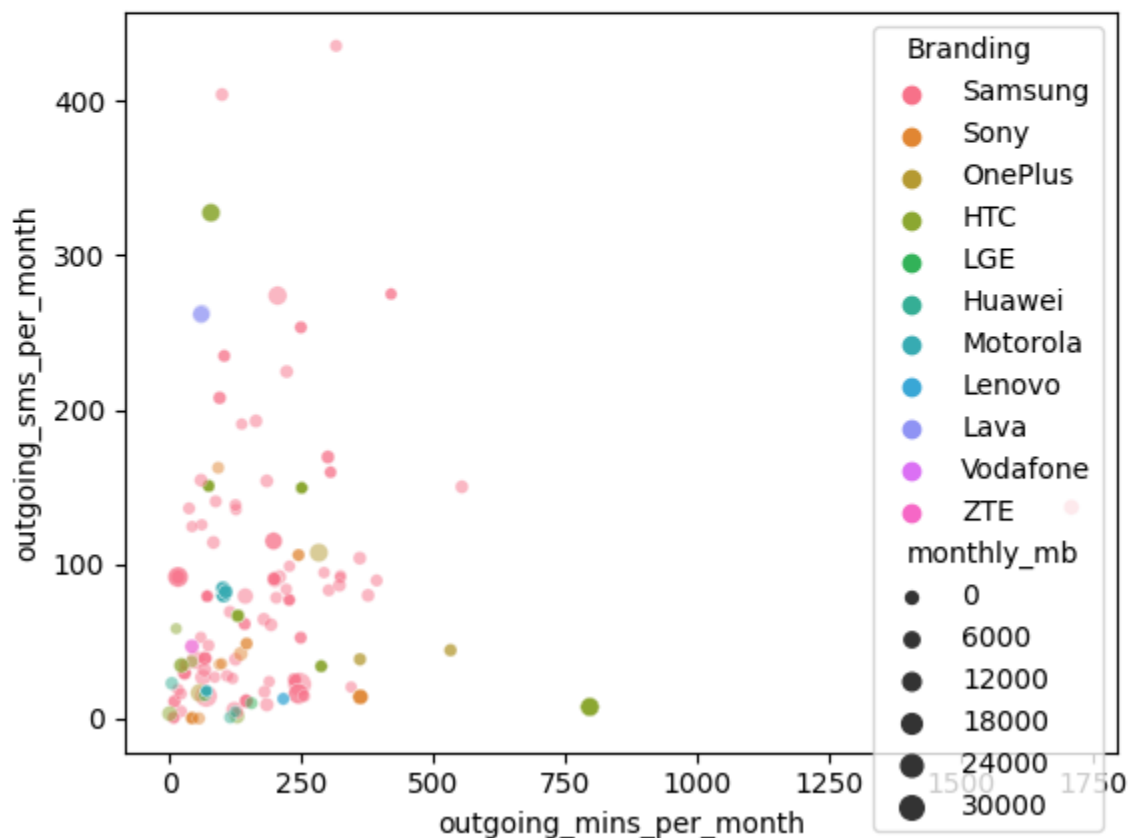
```
import numpy as np
sns.barplot(
plt.show()
```



10. Khảo sát tương quan giữa phút gọi, sms và dung lượng trên từng nhóm nhãn hàng

```
sns.scatterplot()
```

```
plt.show()
```



11. Hãy kiểm định xem trung bình outgoing\_mins\_per\_month có bằng 200 phút
12. Hãy kiểm định xem trung bình outgoing\_sms\_per\_month có bằng 100 tin nhắn
13. Hãy kiểm định xem trung bình monthly\_mb có bằng 2048 mb
14. Hãy tạo ma trận tương quan giữa [outgoing\_mins\_per\_month, outgoing\_sms\_per\_month, monthly\_mb] và sắp xếp tăng dần về mức độ tương quan
15. Kiểm định xem outgoing\_mins\_per\_month, outgoing\_sms\_per\_month có tương quan
16. Hãy kiểm định xem có mối quan hệ nào giữa platform và branding
17. Hãy kiểm định xem có mối quan hệ nào giữa monthly\_mb và platform
18. Hãy kiểm định xem có mối quan hệ nào giữa monthly\_mb và platform theo loại branding

19. Phân tích sự ảnh hưởng của outgoing\_mins\_per\_month, outgoing\_sms\_per\_month đến monthly\_mb dựa trên mô hình hồi quy tuyến tính
20. Dựa trên mô hình hồi quy tuyến tính hãy cho biết nếu outgoing\_mins\_per\_month, outgoing\_sms\_per\_month lần lượt là 50 phút, 75 tin nhắn thì tháng đó người dùng sẽ tiêu hao bao nhiêu dung lượng internet (mb) cho quá trình sử dụng.  
(\*) Lưu ý: Sinh viên xây dựng chương trình phần mềm hoàn chỉnh dạng Console Application với Input là outgoing\_mins\_per\_month, outgoing\_sms\_per\_month và output là giá trị monthly\_mb dự báo được.