

## 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 bảng thông qua mô hình ERD

**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

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='use_id', how='left')
```

5. Tiếp tục inner-join bảng thiếtbi 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='device',
                    right_on='Model')
```

Hoặc left-join

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

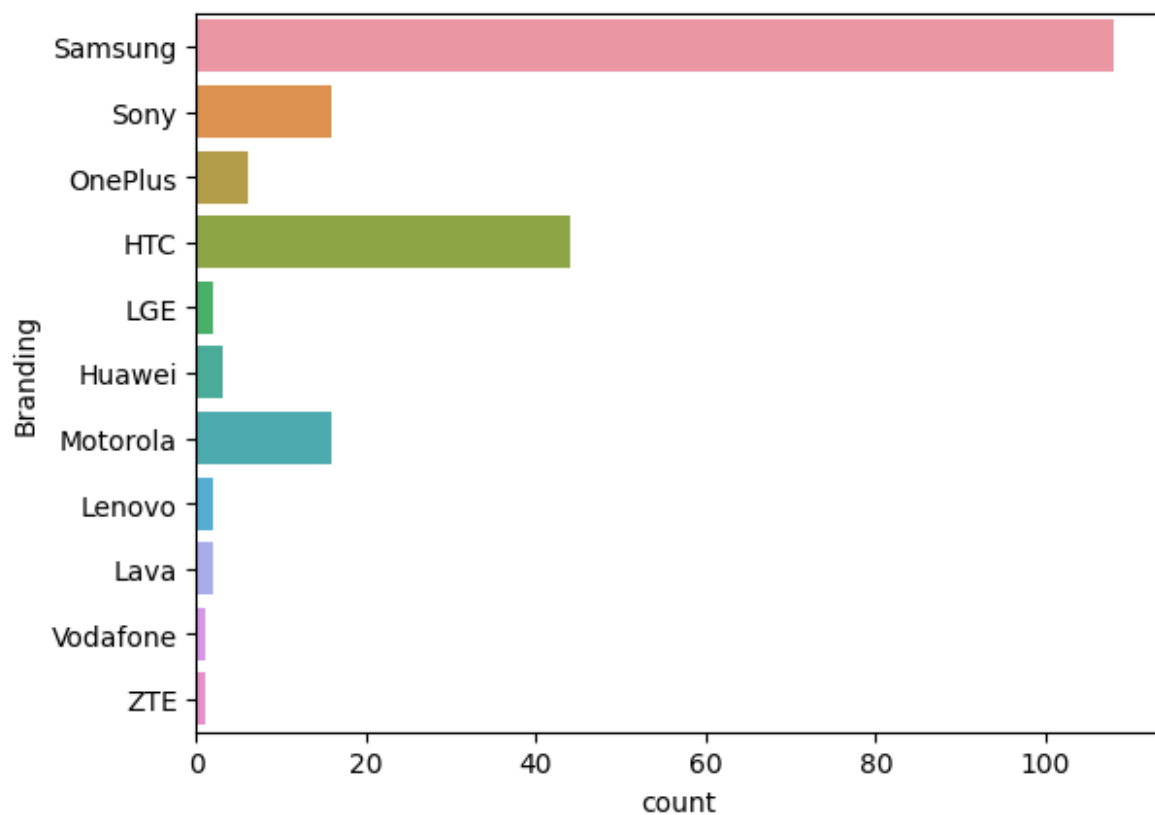
7. Hãy thống kê trung bình số phút hàng tháng, trung bình sms và trung bình data, số lượng sử dụng của các nhà hàng

```
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

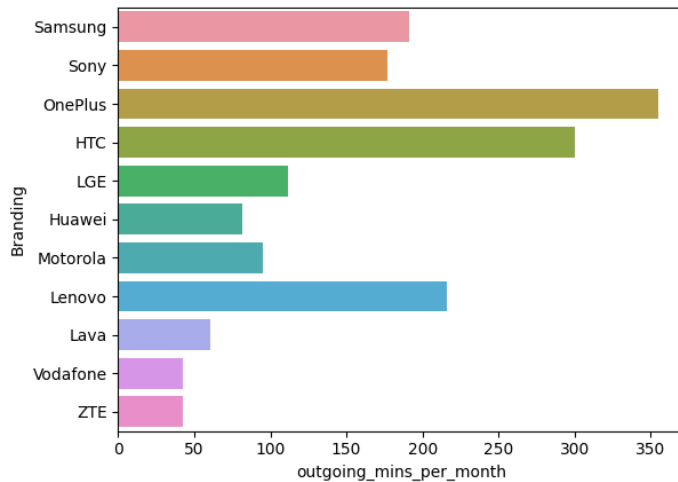
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(y='Branding', data=dfKetQua)
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(y='Branding',x='outgoing_mins_per_month', data=dfKetQua, estimator=np.mean, errorbar=None, orient='horizontal')
plt.show()
```



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

```
sns.scatterplot(x='outgoing_mins_per_month',  
                y='outgoing_sms_per_month',  
                size='monthly_mb',  
                alpha=0.5,  
                hue='Branding',  
                data=dfKetQua)  
  
plt.show()
```

