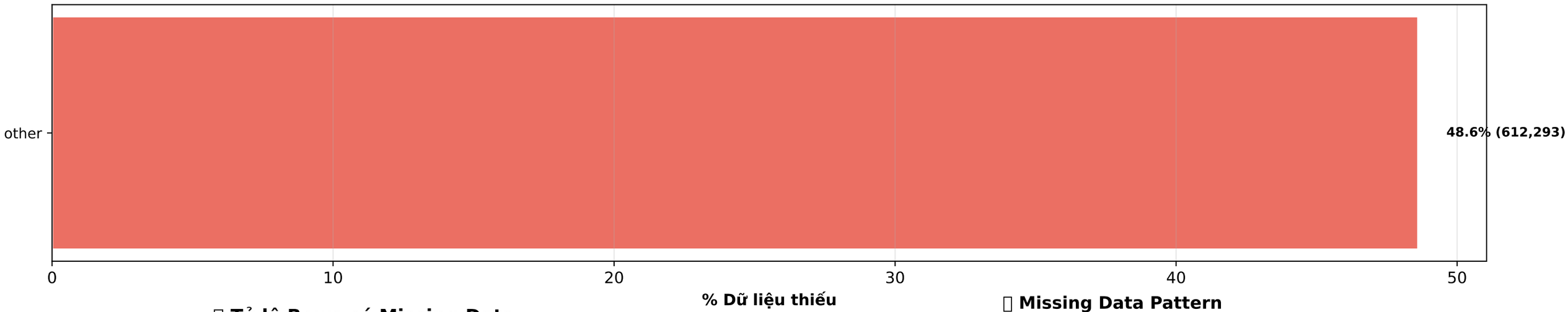
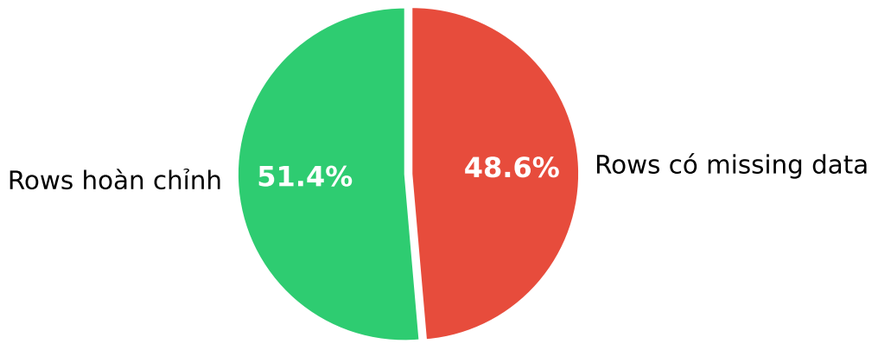
🔍 KHÓ KHĂN 3: MISSING DATA (Dữ liệu thiếu)
Phân tích các trường dữ liệu bị null và tác động đến phân tích

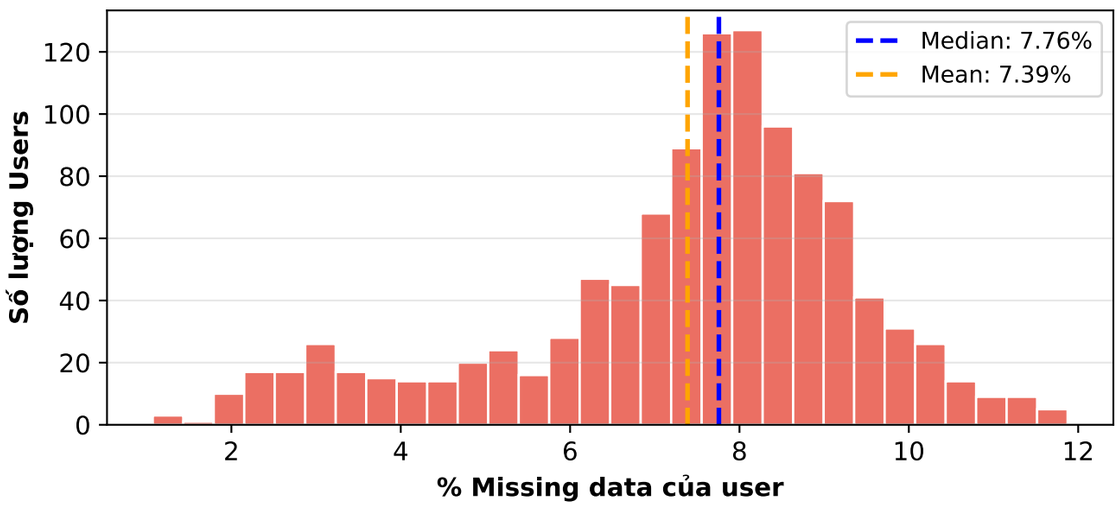
📊 Top Columns có Missing Data
(Tỷ lệ % dữ liệu bị thiếu)



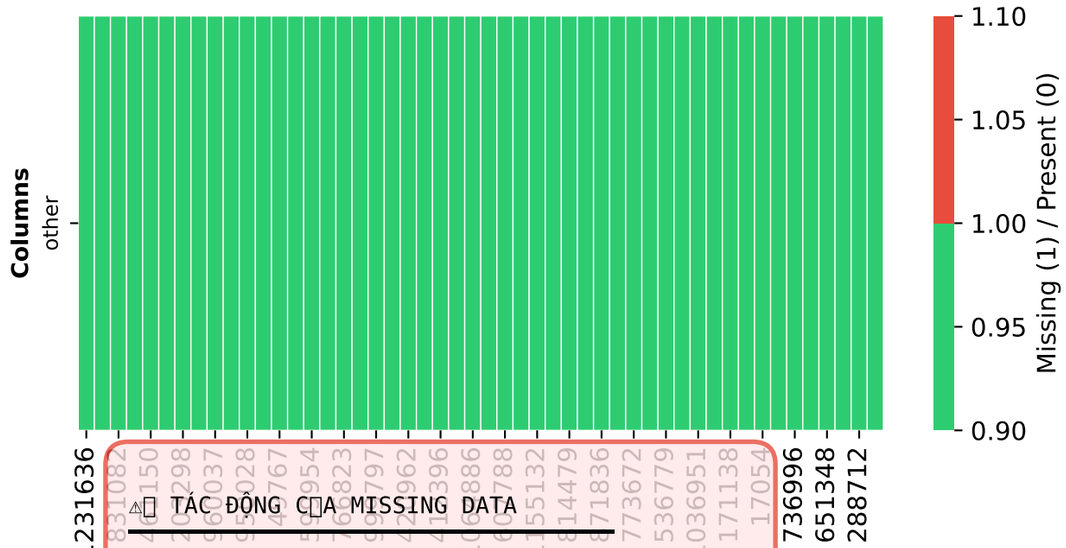
📈 Tỷ lệ Rows có Missing Data



📊 Phân phối Missing Data per User
(Chỉ users có missing)



🔥 Missing Data Pattern
(Mẫu 50 rows)



⚠️ TÁC ĐỘNG CỦA MISSING DATA

📊 Thống kê:
- Tổng cells: 10,075,288
- Missing cells: 612,293 (6.08%)
- Rows bị ảnh hưởng: 612,293 (48.62%)
- Columns bị ảnh hưởng: 1/8

⚠️ Vấn đề:
1. Làm giảm chất lượng feature engineering
2. Nhiều thuật toán ML không xử lý được null
3. Có thể làm sai lệch phân tích thống kê
4. Ảnh hưởng đến clustering và recommendation

💡 Giải pháp:
1. Phân tích pattern: MCAR, MAR, hay MNAR?
2. Imputation: mean, median, mode, KNN
3. Forward/Backward fill cho time series
4. Loại bỏ columns có >50% missing
5. Tạo indicator variable (is_missing)
6. Sử dụng models chấp nhận null (XGBoost)