EDA: Tổng quan data

* info
* check missing values (data.isnull().sum()

1. Age distribution

Age range (most): 40-60 → nhiều nhất: 44

right-skewed

roughly normally distributed

1. check outlier
2. education

→ Bẫy: income của người có 2nd cycle cao hơn người ko có → thực ra ko đúng vì bị chia thành các bậc học với các cột khác nhau → “ko có 2nd cycle” cũng có thể được là người có PhD/..

1. Income theo education

Grad/Master/PhD → income nhỉnh hơn, cao nhất là grad

Basic → income thấp hơn

1. Product vs Acp.Campaign

Overall vẫn là wine, meat nma campaign 5 vẫn được spend nhiều nhất

1. channel preferences : store - catalog - web
2. total purchases per category : store - web - catalog-deal
3. family dynamics kidhome teenhome ảnh hưởng cus spending với meat wine
4. những ng 0 có kid spend wine nhiều
5. kỳ nghỉ đông của những ng đi làm , vacation bao lâu , trẻ con đc nghỉ bao lâu .
6. gộp wines và meat
7. spend on wine and meat by the number of kids at home ( bao gom married , together ,..)

k có trẻ em lượng mua meat thịt nhiều

càng nhiều con càng spend ít

1. gđ có con k dư dả chi tiết => k có con mua đồ đắt tiền
2. campaign

* hâm nóng tcam ( gđ , lứa đôi )
* cuối năm , research , kì nghỉ
* assume : location : US

1. gộp chung together , married => total spending
2. recency by response : những ng khi mà k tới cửa hàng , tgian ng ta k mua hàng càng nhiều thì accept campaign càng it
3. cmp 5đc thể hiện