

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

# **KHÁM PHÁ DỮ LIỆU BÁN HÀNG CỦA DOANH NGHIỆP CHUYÊN NHẬN ORDER**



**NHÓM 5:**  
**NGUYỄN DIỆU PHƯƠNG**  
**NGUYỄN THỊ HUYỀN TRANG**  
**NGUYỄN THỊ MAI TRINH**

# TỔNG QUAN

1 Giới thiệu

2 Mô tả bộ dữ liệu

3 Phương pháp phân tích

4 Khó khăn và thử thách

5 Kết luận

# 1 GIỚI THIỆU

DATA\_MUA HÀNG

Tệp Chỉnh sửa Xem Chèn Định dạng Dữ liệu Công cụ Tiện ích mở rộng Trợ giúp

Chia Sẻ

A1 | fx DATE

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	DATE	NAME	Link	SIZE	MÔ TẢ / MÀU	GIÁ USD	SL	Coupon	Order number	Total	Rate	Total VNĐ	tatus / Tracking
2													
3	thg 4 1	Bank											-15.000.000
4	thg 4 2	macys	<a href="https://www.ma">https://www.ma</a>	XS	BLACK	27,30	2						
5	thg 4 2	macys	<a href="https://www.ma">https://www.ma</a>	S	BLACK	27,3	2						
6	thg 4 3	Bank											-16.000.000
7													
8	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	□6,5	WHITE	\$35,40	3		2690867343	\$106,20	21000	2.230.200	
9													
10	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	□6,5	WHITE	\$35,40	1			\$35,40	21000	743.400	
11	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	6	WHITE	\$35,40	2			\$70,80	21000	1.486.800	
12	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	5	WHITE	\$35,40	2			\$70,80	21000	1.486.800	
13	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	7	WHITE	\$35,40	2			\$70,80	21000	1.486.800	
14	thg 4 6	macys	<a href="https://www.ma">https://www.ma</a>	7,5	WHITE	\$35,40	2		2690900553	\$70,80	21000	1.486.800	
15	thg 4 6	macys		7	Black	\$66,78	1		2690043366	\$66,78	21000	1.402.380	
16	thg 4 7	Bank											-15.000.000
17	thg 4 11	macys	<a href="https://www.ma">https://www.ma</a>	□6,5	WHITE	\$35,40	5			\$177,00		0	
18	thg 4 11	macys	<a href="https://www.ma">https://www.ma</a>	6	WHITE	\$35,40	2			\$70,80		0	

## 2 MÔ TẢ BỘ DỮ LIỆU

- date
- name\_web
- link\_product
- subject
- color
- category
- amount
- purchase\_price\_unit
- sale\_price\_unit

# 3 PHƯƠNG PHÁP PHÂN TÍCH

Tiền xử lý



Phân tích,  
thăm dò dữ liệu



Xây dựng  
mô hình



Đánh giá và  
tối ưu hóa



### 3 Tiết xử lý dữ liệu

- date
- name\_web
- link\_product
- subject
- color
- category
- amount
- purchase\_price\_unit
- sale\_price\_unit

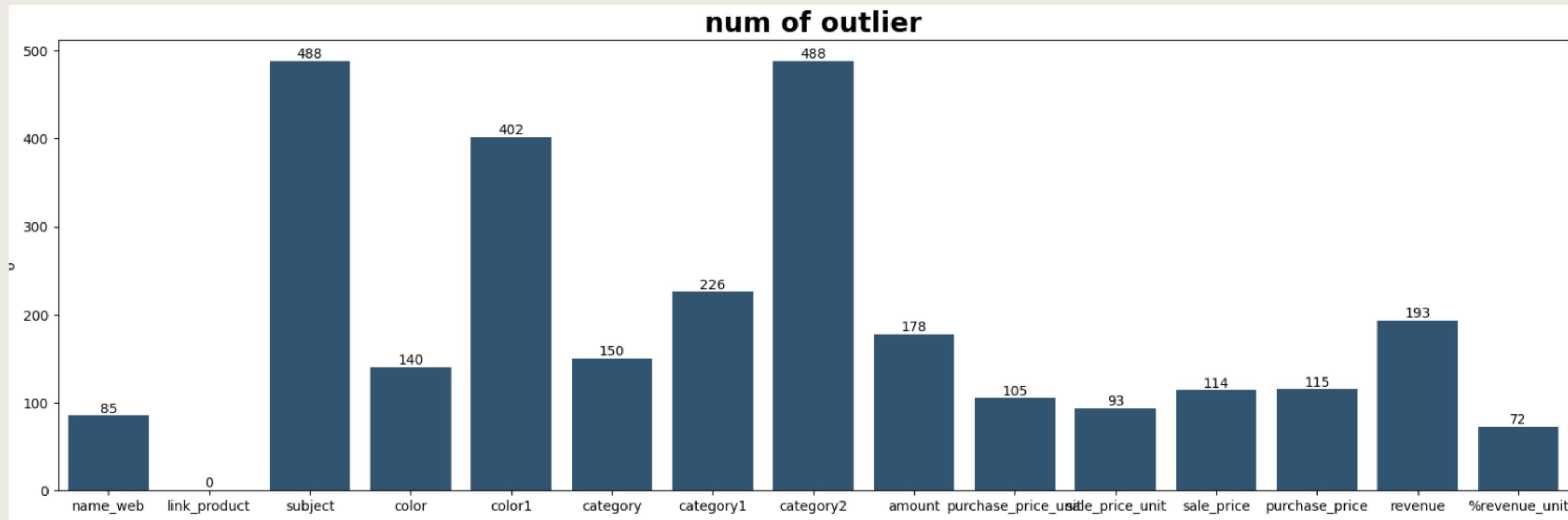
CHUẨN HÓA

- +
- {
- color1
  - category1
  - category2
  - sale\_price
  - purchase\_price
  - revenue
  - %revenue\_unit



ĐIỀN  
KHUYẾT

### 3 Phân tích thăm dò

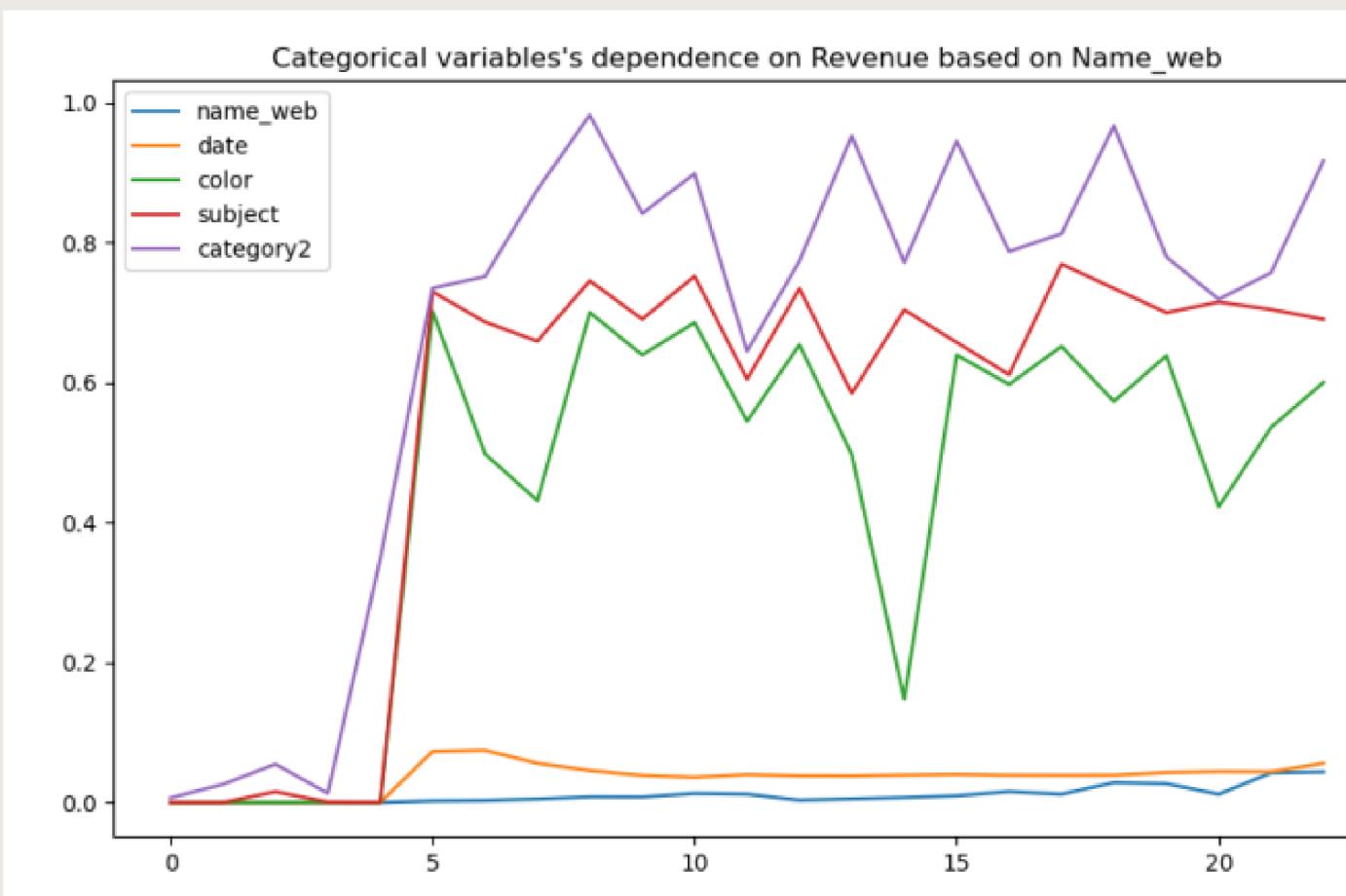


Số lượng outlier trên từng thuộc tính

### 3 Phân tích thăm dò

name_web	date	color	subject	category2	name	total
1.224e-02	0.0373	0.4606	0.6761	0.8831	dataset	2.0693
1.423e-08	0.0010	0.0023	0.0180	0.9674	dataset11	0.9888
7.8720e-38	0.0000	0.0000	0.0110	0.6985	dataset18	0.7096

#### So sánh ĐTQ của biến phân loại



ĐGTQ của biến phân loại khi thay đổi Name\_web

name	dataset	coef	P_value
amount	dataset	0.4388	3.8812e-78
	dataset11	0.4881	7.5706e-89
	dataset18	0.8073	1.7812e-319
purchase_price	dataset	0.4368	2.1564e-77
	dataset11	0.5064	1.3794e-96
	dataset18	0.7038	6.3810e-208
sale_price	dataset	0.7293	2.6347e-272
	dataset11	0.7380	4.5562e-253
	dataset18	0.8165	0.0000e+00
sale_price_unit	dataset	0.3689	4.9556e-54
	dataset11	0.3684	1.8246e-48

So sánh ĐTQ của biến kiểu số

### 3 Phát triển mô hình

Thuộc tính	Mô tả
name_web	Tên trang web đặt mua sản phẩm (adidas, amazon, ashford, macys, saksoff5th, ssense)
color	Màu sắc của sản phẩm (25 màu sắc khác nhau)
subject	Đối tượng sử dụng (M, W, U)
category2	Danh mục sản phẩm (shoes, clothing, bag, accessories, cosmetic)
amount	Số lượng sản phẩm order
sale_price	Tổng giá bán ra của đơn hàng
purchase_price	Tổng giá mua vào của đơn hàng

# 3 Phát triển mô hình

name	coef	p_value
amount	0.9550	1.0504e-73
sale_price	0.9472	4.4214e-69
purchase_price	0.9089	1.6090e-53
amount_w	0.8551	1.2454e-40
amount_black	0.7250	8.7913e-24
amount_shoes	0.6979	1.8545e-21
amount_macys	0.6203	4.91994e-16

Top 7 biến ảnh hưởng nhất đến lợi nhuận mỗi ngày

### 3 Thử nghiệm trên các mô hình dự đoán

Mô hình	Dataset	Dataset11	Dataset18
<b>KNeighborsRegressor</b>	0.4830	0.7864	0.7930
<b>GradientBoostingRegressor</b>	0.6042	0.8693	0.8579
<b>RandomForestRegressor</b>	0.4913	0.8115	0.90467
<b>DecisionTreeRegressor</b>	0.4271	0.3492	0.6930
<b>LinearRegression</b>	0.4958	0.3378	0.8770
<b>LassoRegression</b>	0.6126	0.3427	0.8771
<b>RidgeRegression</b>	0.5239	0.5490	0.9093

So sánh kết quả các mô hình tuyến tính  
trên 3 tập dữ liệu

Mô hình	Không Kiểm chứng chéo	Kiểm chứng chéo
<b>KneighborsRegressor</b>	0.7930	0.8217
<b>GradientBoostingRegressor</b>	0.8579	0.8561
<b>RandomForestRegressor</b>	0.90467	0.9032
<b>DecisionTreeRegressor</b>	0.6930	0.7328
<b>LinearRegression</b>	0.8770	0.8770
<b>LassoRegression</b>	0.8771	0.8771
<b>RidgeRegression</b>	0.9093	0.9093

Kết quả các mô hình khi sử dụng kiểm  
chứng và không kiểm chứng chéo

## 4 KHÓ KHĂN & THỦ THÁCH

- Về phía doanh nghiệp
- Về phía khách hàng
- Về phía bản thân nhóm

## 5 KẾT LUẬN

- Xác định xu hướng mua hàng của nhóm khách hàng
- Xây dựng mô hình dự đoán doanh thu đạt trên 0.9 điểm  $R^2$
- Hỗ trợ doanh nghiệp xây dựng chiến lược kinh doanh

CẢM ƠN THẦY VÀ CÁC BẠN ĐÃ  
CHÚ Ý LẮNG NGHE